Heavy-traffic asymptotics for networks of parallel queues with Markov-modulated service speeds

Queueing Syst (2015) 79:293–319
DOI 10.1007/s11134-014-9422-x

Heavy-traffic asymptotics for networks of parallel queues with Markov-modulated service speeds

Jan-Pieter L. Dorsman · Maria Vlasiou · Bert Zwart

Received: 6 March 2013 / Revised: 26 August 2014 / Published online: 2 October 2014

Abstract We study a network of parallel single-server queues, where the speeds of the servers are varying over time and governed by a single continuous-time Markov chain. We obtain heavy-traffic limits for the distributions of the joint workload, waiting-time and queue length processes. We do so by using a functional central limit theorem approach, which requires the interchange of steady-state and heavy-traffic limits. The marginals of these limiting distributions are shown to be exponential with rates that can be computed by matrix-analytic methods. Moreover, we show how to numerically compute the joint distributions, by viewing the limit processes as multi-dimensional semi-martingale reflected Brownian motions in the non-negative orthant.

J.-P. L. Dorsman (✉) · M. Vlasiou · B. Zwart
EURANDOM and Department of Mathematics and Computer Science,
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: j.l.dorsman@tue.nl

M. Vlasiou
e-mail: m.vlasiou@tue.nl

J.-P. L. Dorsman · M. Vlasiou · B. Zwart
Stochastics, Centrum Wiskunde & Informatica (CWI),
Amsterdam, The Netherlands

B. Zwart
Department of Mathematics, Faculty of Sciences, VU University Amsterdam,
Amsterdam, The Netherlands
e-mail: Bert.Zwart@cwi.nl

B. Zwart
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, GA 30332, USA

Keywords Functional central limit theorem · Layered queueing networks · Machine-repair model · Semi-martingale reflected Brownian motion

Mathematics Subject Classification 60K25 · 68M20 · 90B22

1 Introduction

In this paper, we consider a parallel network of N single-server queues. The speeds of the servers vary over time and are in addition mutually dependent. More specifically, we assume that these service speeds are governed by a single, irreducible, continuous-time Markov chain with a finite state space. For this network, we are interested in both the marginal and the joint workload processes for each of the queues, as well as the processes describing the virtual waiting time and the queue length. Stationary distributions for these processes are difficult to obtain, since the workload process pertaining to one queue as well as the virtual waiting-time and the queue length processes are correlated with the corresponding processes of the other queues. Our goal in this paper is to derive the heavy-traffic behaviour of the network by obtaining the limiting stationary distributions of the aforementioned processes. These results can serve as simple and accurate approximations when the network is heavily utilised or can be combined with known light-traffic results to obtain approximations for arbitrarily loaded systems (see, for example, [14]).

The study of this general network is motivated by the fact that multi-queue performance models with time-varying and mutually dependent service speeds find a wide variety of applications. An example is the field of *wireless networks*, where multiple users transmit data packets through a wireless medium at speeds that are typically varying over time and mutually dependent, for example due to phenomena such as 'shadow fading' (cf. [38]). Another such application constitutes an *I/O subsystem* of an application server (see, for example, [40]), in which the content of multiple I/O buffers is transferred to clients at varying and mutually dependent speeds, due to the varying level of congestion of the application server's network connection. A final example is given by the phenomenon of *garbage collection* in multi-threaded computer systems (cf. [33]). Typically, when the total memory utilisation in such a system exceeds a certain threshold, the processing speeds of the threads are temporarily reduced and are as such mutually dependent.

Queueing models with service speeds that vary over time have received attention in multiple settings in the literature. In practice, service speeds may be dependent on factors such as the workload present in the system, which leads to the formulation of queues with state-dependent service rates; see, for example, [3] for an overview. Another branch of work on time-varying service speeds is that of service rate control, where the aim is to minimise waiting and capacity costs (for example [2,16,35,41]) or to optimise a trade-off between service quality and service speed (for example [20]) based on the state of the system by dynamically varying the service speed. In our case, the service speeds depend on an external environment that is governed by a Markov process. Analyses of single-server queueing models with Markov-modulated service speeds can be found in [17,27,29,30,37]. However, none of these papers

concern themselves with the derivation of heavy-traffic asymptotics. In this paper, we focus on a queueing network where the service speeds of *all* servers in the network are simultaneously governed by a *single* continuous-time Markov chain. This allows us to incorporate mutual dependencies between the service speeds into the model. Conceptually, there are no additional challenges in obtaining heavy-traffic results for the queueing network with multiple queues compared to the single-queue case, although deriving the results for the multi-queue case is more cumbersome at times.

We are mainly interested in the heavy-traffic asymptotics of the network of queues. The study of queues in heavy traffic was initiated by Kingman with a series of papers in the 1960s, starting with [24]; see [25] for an overview of these early results. These papers were largely focused on the use of Laplace transforms. In our case, however, Laplace transforms for the stationary distribution of the total workload process or even the workload process for a queue in isolation are hard to obtain. The workload process of a queue in isolation can in principle be modelled as a reflected Markov additive process (MAP). For the definition and an overview of the standard theory on MAPs, see [1, Section XI.2]. However, the stationary distribution of the workload process is not easily derived from that. For example, standard techniques such as relating the Laplace transforms of the stationary workload conditional on the states of the modulator to each other typically lead to a linear system with a number of equations smaller than the number of unknowns, defying straightforward solutions, as shown in [21]. Less straightforward computations might involve studying the singularities of the characterising matrix exponent pertaining to the reflected MAP (cf. [21]). In the past, stationary distributions for special cases of reflected MAPs have also been analysed by studying their spectral expansion (for example [28]) or by determining the boundary probabilities in terms of the solution of a generalised eigenvalue problem (for example [39]).

As it is not clear that the approach via Laplace transforms will work in our case, we will use a functional central limit theorem approach mainly developed by Iglehart and Whitt; see [43] for an overview. This is not always trivial; see for example [10,26]. Heavy-traffic approximations for generalised Jackson networks were studied in [5,15]. However, the model that we consider does not fall in the framework of generalised Jackson networks. Instead, we tailor more classical arguments for single-node systems to our setting. An advantage of our approach is that it can be extended to allow for variations or generalisations of our model. For example, it is assumed that the workload input processes of the queues are compound Poisson processes. As we will see in the sequel, however, our heavy-traffic analysis still works through completely under relaxed assumptions if Lemma 3.2 can be proved for this more general setting.

As we study networks with general service speeds, our model also captures a class of queues with service interruptions. Heavy-traffic asymptotics for single-server queues with vacations have been studied in [23]. Related but different problems are networks with interruptions, of which durations and frequency scale with the traffic intensity, and have been studied in [6,23] and [43, Section 14.7]. As opposed to these models, our model allows the durations of consecutive service interruptions, which we assume to be independent of the traffic intensity, to be interdependent through the Markovian random environment (see also [8]), and the interruptions are not restricted to a point in time the queue empties.

For the network that we study in this paper, we find that the marginal workload, virtual waiting-time and queue length processes pertaining to a queue in isolation exhibit state-space collapse under heavy-traffic assumptions and have exponential limiting distributions. Moreover, we show that the limiting distribution of the joint workload process (as well as that of the virtual waiting-time and the queue length processes) corresponds to the stationary distribution of an N-dimensional semi-martingale reflected Brownian motion (SRBM) with state space $\mathbb{R}_+^N$ (see, for example, [7, Theorem 6.2] for a definition). The reflection matrix corresponding to this SRBM is an identity matrix, so that positive conclusions about the existence of a stationary distribution can be drawn (cf. [18]). However, computing this distribution is challenging. The conditions needed for the stationary distribution to have a product form do not apply to our model, and results such as those of [11] seem hard to translate to our setting. In this paper, we therefore show how to use the numerical methods developed in [9] for steady-state analysis of multi-dimensional SRBMs to analyse the joint limiting distribution of the stationary workload process. This allows us to compute quantities such as the correlation coefficients between the marginal components.

The rest of this paper is organised as follows. Section 2 describes the model in detail, gives the necessary notation and gives several preliminary results. In Sect. 3, we derive the heavy-traffic limit for a properly scaled workload process and observe that the stationary distribution of the marginal workload processes converges to an exponential distribution. Section 4 extends these results to heavy-traffic limits for the virtual waiting-time and queue length processes. Finally, in Sect. 5, we study how one can compute the joint distribution of the limiting processes pertaining to the workloads, virtual waiting times and the queue lengths, by viewing these as SRBMs. By means of simulation results, we also show that the obtained heavy-traffic results give rise to accurate approximations for considerably loaded systems, which mark the usefulness of the heavy-traffic analysis that we perform from an application perspective.

2 Notation and preliminaries

In this section, we introduce the notation used in this paper, and we present several preliminary results. In the remainder of this paper, vectors and matrices are printed in bold face. Furthermore, **0** and **1** represent vectors of appropriate size where each of the elements are equal to zero and one, respectively.

2.1 Arrival processes

We study the heavy-traffic asymptotics of a network consisting of N parallel single-server queues $Q_1, \ldots, Q_N$, each with its own dedicated arrival stream. Type-i customers arrive at Q_i according to a Poisson process with rate λ_i and have a service requirement distributed according to a random variable B_i with finite first two moments $\mathbb{E}[B_i]$ and $\mathbb{E}[B_i^2]$. In particular, we represent by $B_{i,j}$ the service requirement of the j-th arriving type-i customer. We assume the service requirements of all customers to be mutually independent. Further, we denote by $\{N_i(t), t > 0\}$ a unit-rate Poisson process. Then, the cumulative workload that enters Q_i during the time interval $[0, t)$ is given by

$$V_i(\lambda_i t) = \sum_{j=1}^{N_i(\lambda_i t)} B_{i,j},$$

where the arrival rate is left as part of the argument, as this will prove to be useful for heavy-traffic scaling purposes in the sequel. In the remainder of this paper, we will refer to $\{V_i(t), t \geq 0\}$ as the arrival process of Q_i. The mean corresponding to this arrival process is given by $m_{V,i} = \mathbb{E}[V_i(1)] = \mathbb{E}[B_i]$. Similarly, the variance is given by $\sigma^2_{V,i} = \mathrm{Var}[V_i(1)] = \mathbb{E}[N_i(1)]\mathrm{Var}[B_i] + \mathrm{Var}[N_i(1)]\mathbb{E}[B_i]^2 = \mathrm{Var}[B_i] + \mathbb{E}[B_i]^2 = \mathbb{E}[B_i^2]$. Note that the arrival process has stationary and independent increments, so that $t^{-1}\mathbb{E}[V_i(t)] = m_{V,i}$ and $t^{-1}\mathrm{Var}[V_i(t)] = \sigma^2_{V,i}$ for any $t > 0$.

2.2 Cumulative service processes

The service speeds of the N servers serving $Q_1, \ldots, Q_N$ may vary over time and are mutually dependent. More specifically, the joint process of these service speeds is modulated by a single irreducible, stationary, continuous-time Markov chain $\{\Phi(t), t \geq 0\}$ with finite state space $\boldsymbol{\mathcal{S}}$ and invariant probability measure $\boldsymbol{\pi} = (\pi_i)_{i \in \boldsymbol{\mathcal{S}}}$. When this Markov chain resides in the state $\omega \in \boldsymbol{\mathcal{S}}$, the server of Q_i drains its queue at service rate $\phi_i(\omega)$. We have as a consequence that the workload that the server of Q_i has been capable of processing during the time interval $[0, t)$ is represented by

$$C_i(t) = \int_0^t \phi_i(\Phi(s))\mathrm{d}s.$$

We will also refer to the process $\{C_i(t), t \geq 0\}$ as the cumulative service process of Q_i. Note that, as the Markov process $\{\Phi(t), t \geq 0\}$ is in stationarity, the increments of the process $\{C_i(t), t \geq 0\}$ are also stationary. The mean corresponding to the process $\{C_i(t), t \geq 0\}$ is given by

$$m_{C,i} = \mathbb{E}[C_i(1)] = \int_0^1 \sum_{\omega \in \boldsymbol{\mathcal{S}}} \phi_i(\omega)\mathbb{P}\left(\Phi(s) = \omega\right) \mathrm{d}s = \sum_{\omega \in \boldsymbol{\mathcal{S}}} \phi_i(\omega)\pi_\omega.$$

Since the C_i-process has stationary increments, it holds that $t^{-1}\mathbb{E}[C_i(t)] = m_{C,i}$ for any $t > 0$. We denote the asymptotic variance $\lim_{t\to\infty} t^{-1}\mathrm{Var}[C_i(t)]$ by $\sigma^2_{C,i}$. Similarly, the long-run time-averaged covariance between the cumulative service processes of the servers at Q_i and Q_j is represented by $\gamma^C_{i,j} = \lim_{t\to\infty} \frac{1}{t}\mathrm{Cov}[C_i(t), C_j(t)]$. Computing expressions for $\sigma^2_{C,i}$ and $\gamma^C_{i,j}$ is not trivial. We focus on this problem in Sect. 5.2.

2.3 Scaling

A queue Q_i is said to be 'stable' if the expected amount of arriving work $\lambda_i\mathbb{E}[B_i]$ per time unit is smaller than the average workload $m_{C,i}$ that its server is capable of

processing per time unit. Equivalently, Q_i is stable if its load, defined as $\rho_i = \frac{\lambda_i \mathbb{E}[B_i]}{m_{C,i}}$, is less than one. We are interested in the performance of the network of queues in heavy traffic; i.e. the case for which the arrival rates $\lambda_1, \ldots, \lambda_N$ are scaled so that $(\rho_1, \ldots, \rho_N) \to \mathbf{1}$. For this purpose, it is convenient to introduce the index r. In the r-th system, each arrival rate λ_i is taken so that $\beta_i(1-\rho_i)^{-1} = r$, where the β_i-parameters control the rate at which the arrival rates are scaled by r, while the series of service requirements $B_{i,1}, B_{i,2}, \ldots$ and the C_i-processes are not scaled by r. The heavy-traffic limit for any performance measure of the system corresponds to the limit $r \to \infty$. We denote by $\lambda_{i,r}$ the arrival rate of type-i customers corresponding to the r-th system, so that $\lambda_{i,r} \to \frac{m_{C,i}}{\mathbb{E}[B_i]}$ when $r \to \infty$. For notational convenience, we write for two functions $f(r)$ and $g(r)$ that $f(r) = o(g(r))$ if $\lim_{r\to\infty} f(r)/g(r) = 0$.

2.4 Functional central limit theorems for primitive processes

For purposes that will become clear in the sequel, we now state heavy-traffic limits for the primitive processes that are scaled in time by a factor r^2. First, for the scaled arrival processes, we observe that $\mathbb{E}[V_i(\lambda_{i,r} r^2 t)] = \lambda_{i,r} r^2 \mathbb{E}[B_i] t$. As the arrival processes constitute independent renewal reward processes, the functional central limit theorem for renewal reward processes (see, for example, [43, Theorem 7.4.1]) implies that

$$\left\{\left(\frac{V_1(\lambda_{1,r} r^2 t) - \lambda_{1,r} r^2 \mathbb{E}[B_1] t}{\sqrt{\lambda_{1,r} r}}, \ldots, \frac{V_N(\lambda_{N,r} r^2 t) - \lambda_{N,r} r^2 \mathbb{E}[B_N] t}{\sqrt{\lambda_{N,r} r}}\right), t \geq 0\right\} \xrightarrow{d} \{\mathbf{Z}_V(t), t \geq 0\} \tag{1}$$

as $r \to \infty$, where $\{\mathbf{Z}_V(t), t \geq 0\}$ is an N-dimensional Brownian motion with zero drift and covariance matrix $\mathbf{\Gamma}^V = \mathrm{diag}(\sigma^2_{V,1}, \ldots, \sigma^2_{V,N})$.

Similarly, after observing that $\mathbb{E}[C_i(r^2 t)] = m_{C,i} r^2 t$, it follows from results in [42] that the time-scaled cumulative service processes satisfy

$$\left\{\left(\frac{C_1(r^2 t) - m_{C,1} r^2 t}{r}, \ldots, \frac{C_n(r^2 t) - m_{C,N} r^2 t}{r}\right), t \geq 0\right\} \xrightarrow{d} \{\mathbf{Z}_C(t), t \geq 0\} \tag{2}$$

as $r \to \infty$, where $\{\mathbf{Z}_C(t), t \geq 0\}$ is an N-dimensional Brownian motion with zero drift and covariance matrix $\mathbf{\Gamma}^C$ with elements $\Gamma^C_{i,j} = \gamma^C_{i,j}$. Alternatively, this result follows from the functional central limit theorem for MAPs obtained in [34, Theorem 3.4]. Using the results of [34], we will show how to obtain expressions for $\gamma^C_{i,j}$ in Sect. 5.2.

A heavy-traffic limit for the joint scaled net-input process now follows by combining (1) and (2) with the observation that $\frac{\lambda_{i,r} r^2 \mathbb{E}[B_i] t - m_{C,i} r^2 t}{r} = \beta_i m_{C,i} t$. In particular, this leads to

$$\left\{\left(\frac{V_1(\lambda_{1,r} r^2 t) - C_1(r^2 t)}{r}, \ldots, \frac{V_N(\lambda_{N,r} r^2 t) - C_N(r^2 t)}{r}\right), t \geq 0\right\} \xrightarrow{d} \{\mathbf{Z}(t), t \geq 0\} \tag{3}$$

as $r \to \infty$, where $\{\boldsymbol{Z}(t) = (Z_1(t), \ldots, Z_N(t)), t \geq 0\}$ is an N-dimensional Brownian motion with drift vector $\boldsymbol{\mu} = (-\beta_1 m_{C,1}, \ldots, -\beta_N m_{C,N})$ and covariance matrix

$$\boldsymbol{\Gamma} = \operatorname{diag}\left(\frac{m_{C,1}}{\mathbb{E}[B_1]}\sigma^2_{V,1}, \ldots, \frac{m_{C,N}}{\mathbb{E}[B_N]}\sigma^2_{V,N}\right) + \boldsymbol{\Gamma}^C. \tag{4}$$

2.5 Representations

Let $\{\boldsymbol{W}_r(t) = (W_{1,r}(t), \ldots, W_{N,r}(t)), t \geq 0\}$ be the process that describes the workload in each queue of the r-th system at time t and let $\boldsymbol{W}_r = (W_{1,r}, \ldots, W_{N,r}) = \boldsymbol{W}_r(\infty)$ denote the workload in the system in steady state. The processes $\{\boldsymbol{D}_r(t), t \geq 0\}$ and $\{\boldsymbol{L}_r(t), t \geq 0\}$ as well as $\boldsymbol{D}_r$ and $\boldsymbol{L}_r$ are similarly defined for the virtual waiting time (the delay faced by an imaginary customer arriving at time t) and the queue length (excluding the customer in service), respectively.

The workload $W_{i,r}(t)$ present in Q_i at time t can be represented by the one-sided reflection of the net-input process $\{V_i(\lambda_{i,r}t) - C_i(t), t \geq 0\}$, under the assumption that $W_{i,r}(0) = 0$:

$$\begin{aligned} W_{i,r}(t) &= V_i(\lambda_{i,r}t) - C_i(t) - \inf_{s\in[0,t]}\left\{V_i(\lambda_{i,r}s) - C_i(s)\right\} \\ &= \sup_{s\in[0,t]}\left\{V_i(\lambda_{i,r}t) - V_i(\lambda_{i,r}s) - (C_i(t) - C_i(s))\right\}. \end{aligned} \tag{5}$$

As the joint cumulative service process $\{(C_1(t), \ldots, C_N(t)), t \geq 0\}$ has stationary increments, it holds that $\Big(C_1(t) - C_1(s), \ldots, C_N(t) - C_N(s)\Big) \stackrel{d}{=} \Big(C_1(t - s), \ldots, C_N(t - s)\Big)$, where $\stackrel{d}{=}$ means equality in distribution. Furthermore, since the arrival processes are independent, and compound Poisson processes have time-reversible increments, we also have that $\Big(V_1(\lambda_{1,r}t) - V_1(\lambda_{1,r}s), \ldots, V_N(\lambda_{N,r}t) - V_N(\lambda_{N,r}s)\Big) \stackrel{d}{=} \Big(V_1(\lambda_{1,r}(t - s)), \ldots, V_N(\lambda_{N,r}(t - s))\Big)$. Due to this, we have by (5) that $\boldsymbol{W}_r(t)$ satisfies

$$\begin{aligned} \boldsymbol{W}_r(t) \stackrel{d}{=} \Big(&\sup_{s\in[0,t]}\left\{V_1(\lambda_{1,r}(t - s)) - C_1(t - s)\right\}, \ldots, \sup_{s\in[0,t]}\left\{V_N(\lambda_{N,r}(t - s))\right. \\ &\left. - C_N(t - s)\right\}\Big) \\ = \Big(&\sup_{s\in[0,t]}\left\{V_1(\lambda_{1,r}(s)) - C_1(s)\right\}, \ldots, \sup_{s\in[0,t]}\left\{V_N(\lambda_{N,r}(s)) - C_N(s)\right\}\Big). \end{aligned}$$

By letting $t \to \infty$, this results in

$$\boldsymbol{W}_r \stackrel{d}{=} \Big(\sup_{s\geq 0}\left\{V_1(\lambda_{1,r}s) - C_1(s)\right\}, \ldots, \sup_{s\geq 0}\left\{V_N(\lambda_{N,r}s) - C_N(s)\right\}\Big). \tag{6}$$

In this study, we are particularly interested in the distribution of the scaled workload $\widetilde{\boldsymbol{W}}_r = \frac{\boldsymbol{W}_r}{r}$ (as well as the similarly defined scaled virtual waiting time $\widetilde{\boldsymbol{D}}_r$ and scaled queue length $\widetilde{\boldsymbol{L}}_r$) in heavy traffic, i.e. as $r \to \infty$. It is easily seen from (6) that the scaled workload can be written in terms of the similarly scaled net-input process. That is, after scaling time by a factor r^2, we have

$$\widetilde{\boldsymbol{W}}_r \stackrel{d}{=} \left(\sup_{t\geq 0} \left\{ \frac{V_1(\lambda_{1,r} r^2 t) - C_1(r^2 t)}{r} \right\}, \ldots, \sup_{t\geq 0} \left\{ \frac{V_N(\lambda_{N,r} r^2 t) - C_N(r^2 t)}{r} \right\} \right). \quad (7)$$

3 Heavy-traffic asymptotics of the workload

In this section, we derive the following heavy-traffic asymptotic result for the scaled workload $\widetilde{\boldsymbol{W}}_r$.

Theorem 3.1 *For the scaled workload vector $\widetilde{\boldsymbol{W}}_r$, we have*

$$\widetilde{\boldsymbol{W}}_r \stackrel{d}{\to} \overline{\mathbf{Z}},$$

as $r \to \infty$, where $\overline{\mathbf{Z}} = (\overline{Z}_1, \ldots, \overline{Z}_N)$, $\overline{Z}_i = \sup_{t\geq 0}\{Z_i(t)\}$, and $Z_i(t)$ is as introduced in Sect. 2.

In order to prove this theorem, observe that, as opposed to the infinite-domain case, the supremum of càdlàg functions on a finite domain $[0, M)$, $M \in \mathbb{R}_+$, is a continuous functional; see, for example, [43]. The proof uses this fact in combination with an additional result stated in Lemma 3.4. To prove Lemma 3.4, we first establish upper bounds of the tail probabilities of the suprema of the processes $\{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t, t \geq 0\}$ and $\{\mathbb{E}[C_i(1)]t - C_i(t), t \geq 0\}$ in Lemmas 3.2 and 3.3, respectively.

Lemma 3.2 *For the arrival process $\{V_i(\lambda_{i,r}), t \geq 0\}$ of Q_i, we have that*

$$\mathbb{P}\left(\sup_{t\in[0,T)} \{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t\} \geq x \right) \leq \frac{\lambda_{i,r}\mathbb{E}[B_i^2]T}{x^2}$$

for any $r, x, T \in \mathbb{R}_+$.

Proof As $\{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t, t \geq 0\}$ is a right-continuous martingale, we have by Doob's inequality (cf. [31, Theorem II.1.7]) that $\mathbb{P}(\sup_{t\in[0,T)}\{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t\} \geq x) \leq x^{-2}\sup_{t\in[0,T)}\{\mathrm{Var}[V_i(\lambda_{i,r}t)]\}$. Since $\mathrm{Var}[V_i(\lambda_{i,r}t)] = \lambda_{i,r}\sigma^2_{V,i}t$ is strictly increasing in t, the lemma follows. □

Lemma 3.3 *For the cumulative service process $\{C_i(t), t \geq 0\}$ pertaining to the server of Q_i, there exist for every $x, T \in \mathbb{R}_+$ a set of positive real constants c_1, c_2, c_3 and c_4 such that*

$$\mathbb{P}\left(\sup_{t\in[0,T)} \{\mathbb{E}[C_i(1)]t - C_i(t)\} \geq x \right) \leq \frac{c_1 T}{x^2} + \frac{c_2}{T} + \frac{c_3 T}{e^{c_4\sqrt{x}}}.$$

Proof The lemma is a consequence of Proposition 1 in [22]. Define $h = \max_{\omega\in\boldsymbol{\mathcal{S}}}\{\phi_i(\omega)\}$ and $H(t) = ht - C_i(t)$. The process $\{H(t), t \geq 0\}$ represents increments of the regenerative process $\{h - \phi_i(\Phi(t)), t \geq 0\}$ and regenerates for example every time the Markov process $\{\Phi(t), t \geq 0\}$ enters the reference state $\omega = \Phi(0)$. We denote the n-th of such regeneration times by T_n. Furthermore, we define $\gamma_n^* = \sup_{T_{n-1}\leq t\leq T_n}\{H(t) - H(T_{n-1})\}$ and $\nu_n = T_n - T_{n-1}$. Note that $\nu_1, \nu_2, \ldots$ can be seen as i.i.d. samples from a random variable Y, and represent return times of state ω in the Markov chain $\{\Phi(t), t \geq 0\}$. Proposition 1 in [22] now implies that, for all $x, T \in \mathbb{R}_+$, there exist positive real constants d_1, d_2, d_3 and d_4 such that

$$\mathbb{P}\left(\sup_{t\in[0,T)}\{\mathbb{E}[C_i(1)]t - C_i(t)\} > x\right) \leq d_1\left(e^{-d_2\frac{x^2}{T}} + e^{-d_3T} + Te^{-d_4\sqrt{x}}\right), \quad (8)$$

if $\mathbb{E}[e^{\sqrt{\sup_{0\leq t\leq Y}\{H(t)\}}}] < \infty$ and $\mathbb{E}[e^{\sqrt{\gamma_n^*}}] < \infty$ for any $n \in \mathbb{N}_+$. This statement follows by substituting the variables B_t, b and $Q(x)$ in [22, Proposition 1] by $H(t)$, $h - \mathbb{E}[C_i(1)]$ and $\sqrt{x}$, respectively. To show that the necessary conditions hold in our case, observe that $H(t)$ is non-decreasing in t and takes values from $[0, ht]$. By combining this with the fact that $\sqrt{x} < \epsilon x + \frac{1}{\epsilon}$ for any $x \geq 0$ and $\epsilon > 0$, we have that $\mathbb{E}[e^{\sqrt{\sup_{0\leq t\leq Y}\{H(t)\}}}] = \mathbb{E}[e^{\sqrt{H(Y)}}] \leq \mathbb{E}[e^{\sqrt{hY}}] < \mathbb{E}[e^{\epsilon hY+\epsilon^{-1}}] = e^{\epsilon^{-1}}\mathbb{E}[e^{\epsilon hY}]$ for any $\epsilon > 0$. As $\gamma_n^* \leq h\nu_n$ for any $n > 0$, similar computations yield that $\mathbb{E}[e^{\sqrt{\gamma_n^*}}] < e^{\epsilon^{-1}}\mathbb{E}[e^{\epsilon hY}]$ for all $n \in \mathbb{N}$ and any $\epsilon > 0$. Subsequently, note that the regeneration time Y, which constitutes the return time of state ω in the Markov chain $\{\Phi(t), t \geq 0\}$, can be decomposed into a period of time Y_1 until the transition away from ω, and the following period Y_2 until re-entry into state ω. The former period Y_1 is exponentially distributed with a certain rate α, so that $\mathbb{E}[e^{\epsilon hY_1}] = \frac{\alpha}{\alpha-\epsilon h}$ for $\epsilon < h^{-1}\alpha$. The latter period Y_2 is easily seen to be stochastically smaller than a geometrically distributed random variable with the positive success parameter $q = \min_{\omega'\in\boldsymbol{\mathcal{S}}\setminus\{\omega\}}\{\mathbb{P}(\Phi(1) = \omega \mid \Phi(0) = \omega')\}$. Hence, $\mathbb{E}[e^{\epsilon hY_2}] \leq \frac{qe^{\epsilon h}}{1-(1-q)e^{\epsilon h}}$ for $\epsilon < -h^{-1}\log(1-q)$. As Y_1 and Y_2 are mutually independent, we thus have for $0 < \epsilon < h^{-1}\min\{\alpha, -\log(1-q)\}$ that $e^{\epsilon^{-1}}\mathbb{E}[e^{\epsilon hY}] \leq e^{\epsilon^{-1}}\frac{\alpha}{\alpha-\epsilon h}\frac{qe^{\epsilon h}}{1-(1-q)e^{\epsilon h}} < \infty$, so that the necessary conditions are satisfied. The lemma now follows from (8) by noting that $e^{-T} < T^{-1}$ for all $T > 0$ and taking $c_1 = d_1d_2^{-1}$, $c_2 = d_1d_3^{-1}$, $c_3 = d_1$ and $c_4 = d_4$. □

Based on the results obtained in Lemmas 3.2 and 3.3, we now establish the final auxiliary result needed to prove Theorem 3.1. This result is summarised in the following lemma.

Lemma 3.4 *The scaled net-input process* $\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}, t > 0\}$ *corresponding to* Q_i *satisfies*

$$\lim_{M\to\infty}\lim_{r\to\infty}\mathbb{P}\left(\sup_{t\geq M}\left\{\frac{V_i(\lambda_{i,r}r^2t) - C_i(r^2t)}{r}\right\} \geq x\right) = 0$$

for all $x, M \in \mathbb{R}_+$.

Proof The first part of the proof is inspired by the proof of (20) in [32]. For any r, let $b_{i,r} = \frac{\mathbb{E}[V_i(\lambda_{i,r})]+\mathbb{E}[C_i(1)]}{2}$, so that $b_{i,r} - \mathbb{E}[V_i(\lambda_{i,r})] = \mathbb{E}[C_i(1)] - b_{i,r} = \frac{m_{C,i}-\lambda_{i,r}\mathbb{E}[B_i]}{2} = \frac{1}{2}\beta_i m_{C,i} r^{-1}$. Due to the subadditivity property of the supremum operator, we have for any $M > 0$ that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{t\geq M}\left\{\frac{V_i(\lambda_{i,r}r^2t) - C_i(r^2t)}{r}\right\} \geq x\right)\\
&\leq \mathbb{P}\left(\sup_{t\geq M}\left\{\frac{V_i(\lambda_{i,r}r^2t) - b_{i,r}r^2t}{r}\right\} + \sup_{t\geq M}\left\{\frac{b_{i,r}r^2t - C_i(r^2t)}{r}\right\} \geq x\right)\\
&\leq \mathbb{P}\left(\sup_{t\geq M}\{V_i(\lambda_{i,r}r^2t) - b_{i,r}r^2t\} \geq 0\right) + \mathbb{P}\left(\sup_{t\geq M}\{b_{i,r}r^2t - C_i(r^2t)\} \geq 0\right)\\
&\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in[2^jM,2^{j+1}M)}\{V_i(\lambda_{i,r}r^2t) - b_{i,r}r^2t\} \geq 0\right)\\
&\quad + \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in[2^jM,2^{j+1}M)}\{b_{i,r}r^2t - C_i(r^2t)\} \geq 0\right)\\
&= \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in[2^jr^2M,2^{j+1}r^2M)}\{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t - \frac{1}{2}\beta_i m_{C,i}r^{-1}t\} \geq 0\right)\\
&\quad + \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in[2^jr^2M,2^{j+1}r^2M)}\{\mathbb{E}[C_i(1)]t - C_i(t) - \frac{1}{2}\beta_i m_{C,i}r^{-1}t\} \geq 0\right)\\
&\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in[0,2^{j+1}r^2M)}\{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t\} \geq 2^{j-1}\beta_i m_{C,i}rM\right)\\
&\quad + \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in[0,2^{j+1}r^2M)}\{\mathbb{E}[C_i(1)]t - C_i(t)\} \geq 2^{j-1}\beta_i m_{C,i}rM\right)\\
&\leq \sum_{j=0}^{\infty}\frac{\lambda_{i,r}\mathbb{E}[B_i^2]2^{j+1}r^2M}{2^{2j-2}\beta_i^2 m_{C,i}^2 r^2M^2} + \sum_{j=0}^{\infty}\left(\frac{c_1 2^{j+1}r^2M}{2^{2j-2}\beta_i^2 m_{C,i}^2 r^2M^2} + \frac{c_2}{2^{j+1}m_{C,i}r^2M}\right.\\
&\quad \left. + \frac{c_3 2^{j+1}r^2M}{e^{c_4\sqrt{2^{j-1}\beta_i m_{C,i}rM}}}\right) \qquad (9)
\end{aligned}
$$

for certain positive constants c_1, c_2, c_3 and c_4. The second-to-last inequality follows by observing that the maximum value of $-\frac{1}{2}\beta_i m_{C,i} r^{-1}t$ in the domain $t \in [2^jr^2M, 2^{j+1}r^2M]$ equals $-2^{j-1}\beta_i m_{C,i}rM$ and by enlarging the intervals of the suprema to also include $[0, 2^jr^2M)$. The last inequality follows from Lemmas 3.2 and 3.3. Simplifying (9) leads to

$$\mathbb{P}\left(\sup_{t\geq M}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\}\geq x\right)$$
$$\leq \frac{16(\lambda_{i,r}\mathbb{E}[B_i^2]+c_1)}{\beta_i^2 m_{C,i}^2 M}+\frac{c_2}{m_{C,i}r^2M}+\sum_{j=0}^{\infty} f_{i,j}(r,M), \tag{10}$$

where $f_{i,j}(r,M)=c_3 2^{j+1}r^2Me^{-c_4\sqrt{2^{j-1}\beta_i m_{C,i}rM}}$. The lemma now follows from (10) by taking the limit $r\to\infty$ and subsequently the limit $M\to\infty$, if $\lim_{r\to\infty}\sum_{j=0}^{\infty} f_{i,j}(r,M)=0$. To show that this condition holds, observe that the derivative of $f_{i,j}$ with respect to r reads $\frac{\partial}{\partial r}f_{i,j}(r,M)=c_3 2^j rMe^{-h_{i,j}(M)\sqrt{r}}(4-h_{i,j}(M)\sqrt{r})$, where $h_{i,j}(M):=c_4\sqrt{2^{j-1}\beta_i m_{C,i}M}$. As a result, $\frac{\partial}{\partial r}f_{i,j}(r,M)<0$ if and only if $4-h_{i,j}(M)\sqrt{r}<0$. Due to the monotonicity of $h_{i,j}(M)$ and $\sqrt{r}$ in j and r, respectively, there thus exist positive constants j_0 and r_0, so that $\frac{\partial}{\partial r}f_{i,j}(r,M)<0$ for any $j\geq j_0$ and $r\geq r_0$. This results in the fact that $\sup_{r\geq r_*}f_{i,j}(r,M)=f_{i,j}(r_*,M)$ for every $r_*\geq r_0$. Hence, an upper bound for $\sum_{j=0}^{\infty}f_{i,j}(r,M)$ when $r\geq r_*\geq r_0$ is given by

$$\sum_{j=0}^{\infty} f_{i,j}(r,M)=\sum_{j=0}^{j_0-1} f_{i,j}(r,M)+\sum_{j=j_0}^{\infty} f_{i,j}(r,M)\leq \sum_{j=0}^{j_0-1} f_{i,j}(r,M)+\sum_{j=j_0}^{\infty} f_{i,j}(r_*,M). \tag{11}$$

When $r\to\infty$, we can use (11) with r_* taken arbitrarily large so that

$$\lim_{r\to\infty}\sum_{j=0}^{\infty} f_{i,j}(r,M)\leq \lim_{r\to\infty}\sum_{j=0}^{j_0-1} f_{i,j}(r,M)+\sum_{j=j_0}^{\infty}\lim_{r_*\to\infty} f_{i,j}(r_*,M).$$

By observing that $\lim_{r\to\infty}f_{i,j}(r,M)=0$, this inequality reduces to $\lim_{r\to\infty}\sum_{j=0}^{\infty}f_{i,j}(r,M)\leq 0$. Since $f_{i,j}(r,M)\geq 0$, it thus must hold that $\lim_{r\to\infty}\sum_{j=0}^{\infty}f_{i,j}(r,M)=0$, which concludes the proof. □

Using these auxiliary results, we can now prove Theorem 3.1.

Proof of Theorem 3.1 By (7), it is enough to show that

$$\lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\sup_{t\geq 0}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\}\geq x_i\right\}\right)=\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\sup_{t\geq 0}\{Z_i(t)\}\geq x_i\right\}\right) \tag{12}$$

for all $x_1,\ldots,x_N\geq 0$. We first obtain a lower bound for the left-hand side of (12):

$$\lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\sup_{t\geq 0}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\}\geq x_i\right\}\right)$$
$$\geq \lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\sup_{t\in[0,M)}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\}\geq x_i\right\}\right)$$

$$= \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\sup_{t\in[0,M)}\{Z_i(t)\}\geq x_i\right\}\right) \tag{13}$$

for all $M \in \mathbb{R}_+$, where the equality follows from (3) together with a combination of the continuous mapping theorem and the continuity property of the supremum operator applied to càdlàg-functions on the finite domain $[0, M)$. Next, to derive an upper bound for the left-hand side of (12), denote by $E_{M,i}$ the event that

$$\sup_{t\in[0,M)}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\} = \sup_{t\geq 0}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\},$$

and let $E^c_{M,i}$ be its complementary event. It is trivial to see that $\mathbb{P}(\bigcap_{i=1}^N\{\sup_{t\in[0,M)}\{Z_i(t)\}\geq x_i\})$ is an upper bound for $\lim_{r\to\infty}\mathbb{P}(\bigcap_{i=1}^N\{\sup_{t\geq 0}\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\}\geq x_i; E_{M,i}\})$ for all $M \in \mathbb{R}_+$. Furthermore, we have that $\sum_{i=1}^N\mathbb{P}(\sup_{t\geq M}\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\}\geq x_i)$ is an upper bound for $\mathbb{P}(\bigcap_{i=1}^N\{\sup_{t\geq 0}\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\}\geq x_i\}; \bigcup_{i=1}^N E^c_{M,i})$. Therefore, we obtain by using De Morgan's law that

$$\begin{aligned}
&\lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\sup_{t\geq 0}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\}\geq x_i\right\}\right)\\
&\quad\leq \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\sup_{t\in[0,M)}\{Z_i(t)\}\geq x_i\right\}\right)\\
&\qquad+\lim_{r\to\infty}\sum_{i=1}^{N}\mathbb{P}\left(\sup_{t\geq M}\left\{\frac{V_i(\lambda_{i,r}r^2t)-C_i(r^2t)}{r}\right\}\geq x_i\right).
\end{aligned} \tag{14}$$

When $M \to \infty$, the lower bound established in (13) converges to $\mathbb{P}(\bigcap_{i=1}^N\{\sup_{t\in[0,\infty)}\{Z_i(t)\}\geq x_i\})$. The upper bound found in (14) also converges to this expression, as the second term in the right-hand side of (14) vanishes due to Lemma 3.4. From this, (12) immediately follows, which proves the theorem. □

Remark 3.1 The joint distribution of $\overline{\boldsymbol{Z}}$ is not straightforward to derive explicitly. However, explicit expressions for the marginal distribution of $\overline{Z}_i$ are not hard to obtain. Note that $\overline{Z}_i = \sup_{t\geq 0} Z_i(t)$ is the all-time supremum of a one-dimensional Brownian motion with negative drift $-\beta_i m_{C,i}$ and variance $\frac{m_{C,i}}{\mathbb{E}[B_i]}\sigma^2_{V,i}+\sigma^2_{C,i}$. It is well known that the all-time supremum of a Brownian motion with negative drift $-a$ and variance b is exponentially $(\frac{2a}{b})$ distributed. Therefore, the distribution of the steady-state scaled workload $\widetilde{W}_{i,r}$ present in Q_i converges to an exponential distribution with rate $2\beta_i\left(\frac{\sigma^2_{V,i}}{\mathbb{E}[B_i]}+\frac{\sigma^2_{C,i}}{m_{C,i}}\right)^{-1}$ as $r\to\infty$. In the next section, we will see that the limiting distributions of $\widetilde{D}_{i,r}$ and $\widetilde{L}_{i,r}$ only differ from the limiting distribution of $\widetilde{W}_{i,r}$ by a multiplicative factor $m^{-1}_{C,i}$ and $\mathbb{E}[B_i]^{-1}$, respectively. As a result,

the distributions of the steady-state delay $\widetilde{D}_{i,r}$ and the steady-state queue length $\widetilde{L}_{i,r}$ also converge to exponential distributions with rates $2\beta_i m_{C,i}\left(\frac{\sigma^2_{V,i}}{\mathbb{E}[B_i]}+\frac{\sigma^2_{C,i}}{m_{C,i}}\right)^{-1}$ and $2\beta_i \mathbb{E}[B_i]\left(\frac{\sigma^2_{V,i}}{\mathbb{E}[B_i]}+\frac{\sigma^2_{C,i}}{m_{C,i}}\right)^{-1}$, respectively. We will study the derivation of the complete distribution of $\overline{\boldsymbol{Z}}$ in Sect. 5.3.

4 Extension to virtual waiting times and queue lengths

In Sect. 3, we derived a heavy-traffic limit theorem for the scaled workload vector $\widetilde{\boldsymbol{W}}_r$. In this section, we extend this result to heavy-traffic limits for the distributions of the virtual waiting-time vector $\widetilde{\boldsymbol{D}}_r$ and the queue length vector $\widetilde{\boldsymbol{L}}_r$ by considering the joint distribution of $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{W}}_r$ as well as that of $\widetilde{\boldsymbol{L}}_r$ and $\widetilde{\boldsymbol{W}}_r$ in Sects. 4.1 and 4.2, respectively. It turns out that, when $r \to \infty$, the distributions of both $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{L}}_r$ are elementwise equal to the distribution of $\widetilde{\boldsymbol{W}}_r$ up to a multiplicative constant.

4.1 Heavy-traffic asymptotics of the virtual waiting time

We now study the distribution of the scaled virtual waiting time in heavy traffic. First, we obtain the tail probability of the joint distribution of $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{W}}_r$ as $r \to \infty$ in Proposition 4.1. Based on this, we obtain an extension of Theorem 3.1 for the scaled virtual waiting time in Corollary 4.2.

Proposition 4.1 *The tail probability of the limiting joint distribution of $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{W}}_r$ satisfies*

$$\lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{D}_{i,r}\geq s_i;\,\widetilde{W}_{i,r}\geq t_i\right\}\right)=\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\overline{Z}_i\geq\max\{m_{C,i}s_i,t_i\}\right\}\right)$$

with $\overline{Z}_1,\ldots,\overline{Z}_N$ as defined in Theorem 3.1.

Proof Observe that since the waiting time faced by an imaginary type-i customer arriving at time u is longer than s_i time units, the workload present in Q_i just before u is larger than $C_i(u+s_i)-C_i(u)$. This is evident, since the latter number represents the amount of work that the server of Q_i is able to process in the s_i time units following time u. In other words, the event $\{D_{i,r}(u) > s_i\}$ is tantamount to the event $\{W_{i,r}(u) > C_i(u+s_i)-C_i(u)\}$ for $i=1,\ldots,N$, so that in steady state (i.e. $u\to\infty$) we have

$$\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{D_{i,r}>s_i;\,W_{i,r}>t_i\right\}\right)=\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{W_{i,r}>\max\{C_i(s_i),t_i\}\right\}\right). \tag{15}$$

Based on this, we obtain an expression for the tail probability of the joint distribution of $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{W}}_r$:

$$\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{D}_{i,r}\geq s_i;\,\widetilde{W}_{i,r}\geq t_i\right\}\right)=\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{W_{i,r}\geq\max\{C_i(rs_i),rt_i\}\right\}\right)$$
$$=\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r}\geq\max\left\{\frac{C_i(rs_i)}{r},t_i\right\}\right\}\right),\qquad(16)$$

where we used (15) in the first equality. We now focus on showing that

$$\lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r}\geq\max\left\{\frac{C_i(rs_i)}{r},t_i\right\}\right\}\right)=\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\overline{Z}_i\geq\max\{m_{C,i}s_i,t_i\}\right\}\right),\qquad(17)$$

which, combined with (16), directly implies the result to be proved. To this end, we observe that, since $\{C_i(t),t\geq 0\}$ is a renewal reward process, $r^{-1}C_i(rs_i)\to m_{C,i}s_i$ almost surely as $r\to\infty$ due to standard results in renewal theory. Denote by $F^{\epsilon}_{i,r}$ for any $\epsilon>0$ the event that $r^{-1}C_i(rs_i)\in[m_{C,i}s_i-\epsilon,m_{C,i}s_i+\epsilon]$ and let $F^{\epsilon,c}_{i,r}$ be its complementary event. Thus, $\lim_{r\to\infty}\mathbb{P}(F^{\epsilon}_{i,r})=1$. As a result, we have, due to De Morgan's law, that

$$\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r}\geq\max\left\{\frac{C_i(rs_i)}{r},t_1\right\}\right\}\right)$$
$$=\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r}\geq\max\left\{\frac{C_i(rs_i)}{r},t_i\right\};\,F^{\epsilon}_{i,r}\right\}\right)+o(1).$$

Letting $r\to\infty$ in this expression, using the definition of the event $F^{\epsilon}_{i,r}$ and applying Theorem 3.1, we obtain the following lower bound for the left-hand side of (17):

$$\lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r}\geq\max\left\{\frac{C_i(rs_i)}{r},t_i\right\}\right\}\right)\geq\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\overline{Z}_i\geq\max\{m_{C,i}s_i+\epsilon,t_i\}\right\}\right).\qquad(18)$$

Similarly, an upper bound for the left-hand side of (17) is given by

$$\lim_{r\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r}\geq\max\left\{\frac{C_i(rs_i)}{r},t_i\right\}\right\}\right)\leq\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\overline{Z}_i\geq\max\{m_{C,i}s_i-\epsilon,t_i\}\right\}\right).\qquad(19)$$

In Remark 3.1, we found that $\overline{Z}_i$ is exponentially distributed for $i=1,\ldots N$, so that the joint distribution of $\overline{\boldsymbol{Z}}$ has no discontinuity in the point $(m_{C,1}s_1,\ldots,m_{C,N}s_N)$. As a consequence, by taking the limit $\epsilon\to 0$ in the right-hand sides of (18) and (19), we obtain (17), which, as explained above, proves the proposition. □

From Proposition 4.1, the heavy-traffic limit for the virtual waiting time follows in the following corollary.

Corollary 4.2 *For the scaled virtual waiting-time vector* $\widetilde{\boldsymbol{D}}_r$*, it holds that*

$$\widetilde{\boldsymbol{D}}_r \xrightarrow{d} \left(\frac{1}{m_{C,1}}, \ldots, \frac{1}{m_{C,N}}\right)\overline{\boldsymbol{Z}},$$

as $r \to \infty$*, with* $\overline{\boldsymbol{Z}}$ *defined in Theorem* 3.1.

Proof This is an immediate result from Proposition 4.1 by taking $t_1 = \ldots = t_N = 0$. □

4.2 The joint queue-length distribution

In this section, we obtain an extension of Theorem 3.1 for the scaled steady-state queue length $\widetilde{\boldsymbol{L}}_r$ in heavy traffic. Let $B^R_{i,r}$ be the remaining service requirement of a type-i customer in service in the r-th system if $L_{i,r} > 0$, and zero otherwise. It is then trivially seen that

$$\boldsymbol{W}_r = \left(B^R_{1,r}, \ldots, B^R_{N,r}\right) + \left(\sum_{j=1}^{L_{1,r}} \widehat{B}_{1,j}, \ldots, \sum_{j=1}^{L_{N,r}} \widehat{B}_{N,j}\right) \tag{20}$$

for all $i > 0$, where $\widehat{B}_{i,j}$ represents the service requirement of the waiting customer in the j-th waiting position of Q_i and is distributed according to B_i. These service requirements are mutually independent as well as independent from $\boldsymbol{W}_r$ and $\boldsymbol{L}_r$. Note that $\widehat{B}_{i,j}$ is defined differently from $B_{i,j}$, which we defined in Sect. 2 to be the service requirement of the j-th arriving type-i customer since the start of the queueing process. The scaled version of (20) is given by

$$\widetilde{\boldsymbol{W}}_r = \left(\widetilde{B}^R_{1,r}, \ldots, \widetilde{B}^R_{N,r}\right) + \frac{1}{r}\left(\sum_{j=1}^{r\widetilde{L}_{1,r}} \widehat{B}_{1,j}, \ldots, \sum_{j=1}^{r\widetilde{L}_{N,r}} \widehat{B}_{N,j}\right), \tag{21}$$

where $\widetilde{B}^R_{i,r} = \frac{1}{r} B^R_{i,r}$ for $i = 1, \ldots, N$. It is intuitively tempting to conclude that $(\widetilde{B}^R_{1,r}, \ldots, \widetilde{B}^R_{N,r}) \to \boldsymbol{0}$ as $r \to \infty$, and based on that, conclude that $\widetilde{\boldsymbol{W}}_r$ and $\widetilde{\boldsymbol{L}}_r$ are equal elementwise up to a multiplicative constant. However, this is not straightforward, since, for example, $\widetilde{\boldsymbol{L}}_r$ and $(\widetilde{B}^R_{1,r}, \ldots, \widetilde{B}^R_{N,r})$ are not independent. We make these results rigorous in this section. Inspired by [44, Proposition 1], we first obtain another representation for the joint distribution of $\widetilde{L}_{i,r}$ and $\widetilde{W}_{i,r}$ for a single queue Q_i in Lemma 4.3. Based on this result, we derive the heavy-traffic asymptotics for $(\widetilde{L}_{i,r}, \widetilde{W}_{i,r}, \widetilde{B}^R_{i,r})$ in Lemma 4.4, which imply that $\widetilde{B}^R_{i,r} \to 0$ as $r \to \infty$. We subsequently conclude that $(\widetilde{B}^R_{1,r}, \ldots, \widetilde{B}^R_{N,r}) \to \boldsymbol{0}$ as $r \to \infty$ and derive the joint distribution of $\widetilde{\boldsymbol{L}}_r$ and $\widetilde{\boldsymbol{W}}_r$

as $r \to \infty$ in Proposition 4.5. From this, an extension of Theorem 3.1 for the scaled queue length $\widetilde{\boldsymbol{L}}_r$ follows in Corollary 4.6.

In order to construct an additional representation for the joint distribution of $\widetilde{L}_{i,r}$ and $\widetilde{W}_{i,r}$, we need to introduce some additional notation. Denote by $W^r_{i,n}$ and $L^r_{i,n}$ the workload present in Q_i and the queue length of Q_i, respectively in the r-th system, just before the n-th arrival of a type-i customer. Furthermore, $A^r_{i,j}$ refers to the time between the j-th and the $(j+1)$-st arriving type-i customer in the r-th system, so that $S^{A,r}_{i,n} = \sum_{j=1}^{n} A^r_{i,j}$ and $S^B_{i,n} = \sum_{j=1}^{n} B_{i,j}$ represent the cumulative series of interarrival times and service requirements of type-i customers. By construction of the heavy-traffic scaling, $A^r_{i,j} \xrightarrow{d} A_{i,j}$ and $\mathbb{E}[A^r_{i,j}] \to \mathbb{E}[A_{i,j}]$ as $r \to \infty$, where $A_{i,j}$ are i.i.d. samples from an exponential $\left(m_{C,i}/\mathbb{E}[B_i]\right)$ distribution. Finally, we define $S^r_{i,n} = S^B_{i,n} - C_i(S^{A,r}_{i,n})$. The required representation is now given in the following lemma.

Lemma 4.3 *For any $x, y > 0$ and $i = 1, \ldots, N$, the joint distribution of $\widetilde{L}_{i,r}$ and $\widetilde{W}_{i,r}$ satisfies*

$$\mathbb{P}\left(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y\right) = \mathbb{P}\Big(W_{i,r} + B_i \geq C_i(S^{A,r}_{i,\lceil rx \rceil}); \\ r^{-1} \max\left\{W_{i,r} + S^r_{i,\lceil rx \rceil}, \max_{j \in \{1,\ldots,\lceil rx \rceil\}} \{S^r_{i,\lceil rx \rceil} - S^r_{i,j}\}\right\} \geq y\Big).$$

Proof The proof is inspired by [44, Proposition 1]. Observe that, for any $k \geq 1$ and $n \geq 1$, the event $\{L^r_{i,n+k} \geq k\}$ coincides with the event that the workload the server at Q_i was capable of processing between the arrival of the n-th and $(n+k)$-th customer, $C_i(S^{A,r}_{i,n+k-1}) - C_i(S^{A,r}_{i,n-1})$, does not exceed the amount $W^r_{i,n} + B_{i,n}$ of work present in Q_i just after the arrival of the n-th customer. Hence, we have that

$$\{L^r_{i,n+k} \geq k\} = \{W^r_{i,n} + B_{i,n} \geq C_i(S^{A,r}_{i,n+k-1}) - C_i(S^{A,r}_{i,n-1})\}. \tag{22}$$

Moreover, due to Lindley's recursion, which is given by $W^r_{i,n+1} = \max\{W^r_{i,n} + S^r_{i,n} - S^r_{i,n-1}, 0\}$ or $W^r_{i,n+k} = \max\{W^r_{i,n} + S^r_{i,n+k-1} - S^r_{i,n-1}, \max_{j \in \{0,\ldots,k-1\}}\{S_{i,n+k-1} - S_{i,n+j}\}\}$, we have for any $y \geq 0$ that

$$\{W^r_{n+k} \geq y\} = \left\{\max\left\{W^r_{i,n} + S^r_{i,n+k-1} - S^r_{i,n-1}, \max_{j \in \{0,\ldots,k-1\}} \{S^r_{i,n+k-1} - S^r_{i,n+j}\}\right\} \geq y\right\}. \tag{23}$$

By combining (22) and (23), taking the probabilities of these events, letting $n \to \infty$ and observing that the vector $(L^r_{i,n}, W^r_{i,n})$ weakly converges to $(L_{i,r}, W_{i,r})$, we obtain

$$\mathbb{P}\left(L_{i,r} \geq k; W_{i,r} \geq y\right) \\ = \mathbb{P}\left(W_{i,r} + B_i \geq C_i(S^{A,r}_{i,k}); \max\left\{W_{i,r} + S^r_{i,k}, \max_{j \in \{1,\ldots,k\}} \{S^r_{i,k} - S^r_{i,j}\}\right\} \geq y\right),$$

for any $k \geq 1$, $y \geq 0$. By noting that $\mathbb{P}(\widetilde{L}_{i,r} \geq x, \widetilde{W}_{i,r} \geq y) = \mathbb{P}(L_{i,r} \geq \lceil rx \rceil, r^{-1} W_{i,r} \geq y)$, the desired statement follows immediately. □

Based on Lemma 4.3, we derive the heavy-traffic asymptotics of $(\widetilde{L}_{i,r}, \widetilde{W}_{i,r}, \widetilde{B}^R_{i,r})$ in the following lemma. This lemma directly implies that $\widetilde{B}^R_{i,r} \to 0$ as $r \to \infty$.

Lemma 4.4 *For any queue, the scaled steady-state queue length, workload and remaining service requirement exhibit state-space collapse under heavy-traffic assumptions. In particular, we have that*

$$(\widetilde{L}_{i,r}, \widetilde{W}_{i,r}, \widetilde{B}^R_{i,r}) \xrightarrow{d} \left(\frac{1}{\mathbb{E}[B_i]}, 1, 0\right) \overline{Z}_i$$

as $r \to \infty$ for any $i \in \{1, \ldots, N\}$, with $\overline{Z}_i$ defined in Sect. 2.

Proof Again, the proof is inspired by [44, Proposition 1]. We first focus on the joint distribution of $\widetilde{L}_{i,r}$ and $\widetilde{W}_{i,r}$. Due to the strong law of large numbers, $r^{-1} S^{A,r}_{i,\lceil rx \rceil} \to \mathbb{E}[A_{i,j}]x = \frac{\mathbb{E}[B_i]x}{m_{C,i}}$ almost surely as $r \to \infty$. Moreover, $t^{-1} C_i(t) \to m_{C,i}$ almost surely as $t \to \infty$, so that

$$\frac{C_i(S^{A,r}_{i,\lceil rx \rceil})}{r} = \frac{C_i(S^{A,r}_{i,\lceil rx \rceil})}{S^{A,r}_{i,\lceil rx \rceil}} \frac{S^{A,r}_{i,\lceil rx \rceil}}{r} \to \mathbb{E}[B_i]x \tag{24}$$

in probability as $r \to \infty$. We further have, due to the weak law of large numbers, that $r^{-1} S^B_{i,\lceil rx \rceil} \to \mathbb{E}[B_i]x$, so that $r^{-1} S^r_{i,\lceil rx \rceil} \to 0$ and $r^{-1} \max_{j \in \{1,\ldots,\lceil rx \rceil\}} \{S^r_{i,\lceil rx \rceil} - S^r_{i,j}\} \to 0$ as $r \to \infty$. Let, for any $\epsilon > 0$, $G^\epsilon_{i,r}$ denote the event

$$\{r^{-1} C_i(S^{A,r}_{i,\lceil rx \rceil}) \in [\mathbb{E}[B_i]x - \epsilon, \mathbb{E}[B_i]x + \epsilon]; r^{-1} S^B_{i,\lceil rx \rceil} \in [\mathbb{E}[B_i]x - \epsilon, \mathbb{E}[B_i]x + \epsilon];$$
$$r^{-1} S^r_{i,\lceil rx \rceil} \in [-\epsilon, \epsilon]; r^{-1} \max_{j \in \{1,\ldots,\lceil rx \rceil\}} \{S^r_{i,\lceil rx \rceil} - S^r_{i,j}\} \in [0, \epsilon]\}.$$

Due to the convergence results above, $\lim_{r\to\infty} \mathbb{P}(G^\epsilon_{i,r}) = 1$ so that $\mathbb{P}(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y) = \mathbb{P}(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y; G^\epsilon_{i,r}) + o(1)$. After combining this with Lemma 4.3 and consequently taking the limit $r \to \infty$, we obtain

$$\lim_{r\to\infty} \mathbb{P}\left(\widetilde{W}_{i,r} \geq \max\{\mathbb{E}[B_i]x + \epsilon, y + \epsilon\}\right)$$
$$\leq \lim_{r\to\infty} \mathbb{P}\left(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y\right) \leq \lim_{r\to\infty} \mathbb{P}\left(\widetilde{W}_{i,r} \geq \max\{\mathbb{E}[B_i]x - \epsilon, y - \epsilon\}\right),$$

since $\widetilde{B}_i \to 0$ as $r \to \infty$. By first applying Theorem 3.1 on the left-hand side and the right-hand side, next noting that the distribution of $\overline{Z}_i$ has no discontinuity points (cf. Remark 3.1), and finally letting $\epsilon \to 0$, we obtain

$$\lim_{r\to\infty} \mathbb{P}\left(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y\right) = \mathbb{P}\left(\overline{Z}_i \geq \max\{\mathbb{E}[B_i]x, y\}\right). \tag{25}$$

It remains to consider the convergence of $\widetilde{B}^R_{i,r}$. We show that $\lim_{r\to\infty} \mathbb{P}(\widetilde{B}^R_{i,r} > \delta) = 0$ for all $\delta > 0$, which finalises the proof of the desired statement. Note that due

to representation (21), we have that $\mathbb{P}(\widetilde{B}^R_{i,r} > \delta) = \mathbb{P}(\widetilde{W}_{i,r} > \frac{1}{r}\sum_{j=1}^{r\widetilde{L}_{i,r}} \widehat{B}_{i,j} + \delta)$. Let $H^\epsilon_{i,r}$ denote the event $\{\frac{1}{n}\sum_{j=1}^{n} \widehat{B}_{i,j} \in (\mathbb{E}[B_i] - \epsilon, \mathbb{E}[B_i] + \epsilon)$ for all $n \geq \sqrt{r}\}$. By using the law of total probability and noting that $\lim_{r\to\infty} \mathbb{P}(H^\epsilon_{i,r}) = 1$ due to the weak law of large numbers, we thus have, similar to earlier calculations, that

$$\begin{aligned}\mathbb{P}\left(\widetilde{B}^R_{i,r} > \delta\right) &= \mathbb{P}\left(\widetilde{W}_{i,r} > \frac{1}{r}\sum_{j=1}^{r\widetilde{L}_{i,r}} \widehat{B}_{i,j} + \delta;\, H^\epsilon_{i,r}\right) + o(1)\\ &= \mathbb{P}\left(\widetilde{W}_{i,r} > \widetilde{L}_{i,r}\frac{1}{r\widetilde{L}_{i,r}}\sum_{j=1}^{r\widetilde{L}_{i,r}} \widehat{B}_{i,j} + \delta;\, H^\epsilon_{i,r}\right) + o(1).\end{aligned}$$

By taking the limit $r \to \infty$ and using the established convergence of $\widetilde{L}_{i,r}$, we obtain

$$\begin{aligned}\lim_{r\to\infty} \mathbb{P}\left(\widetilde{W}_{i,r} > \widetilde{L}_{i,r}(\mathbb{E}[B_i] + \epsilon) + \delta\right) &\leq \lim_{r\to\infty} \mathbb{P}\left(\widetilde{B}^R_{i,r} > \delta\right)\\ &\leq \lim_{r\to\infty} \mathbb{P}\left(\widetilde{W}_{i,r} > \widetilde{L}_{i,r}(\mathbb{E}[B_i] - \epsilon) + \delta\right).\end{aligned}$$

By letting $\epsilon \to 0$ and noting, as before, that the limiting distribution of $\widetilde{W}_{i,r}$ has no discontinuity points, this leads to $\lim_{r\to\infty} \mathbb{P}(\widetilde{B}^R_{i,r} > \delta) = \lim_{r\to\infty} \mathbb{P}(\widetilde{W}_{i,r} > \widetilde{L}_{i,r}\mathbb{E}[B_i] + \delta)$ for any $\delta > 0$. Observe that (25) implies that $\lim_{r\to\infty} \mathbb{P}(\widetilde{W}_{i,r} > \widetilde{L}_{i,r}\mathbb{E}[B_i] + \delta) = 0$ for any $\delta > 0$, which completes the proof. □

Based on the previous results, we now obtain the limiting joint distribution of $\widetilde{\boldsymbol{L}}_r$ and $\widetilde{\boldsymbol{W}}_r$ in the following proposition.

Proposition 4.5 *The tail probability of the limiting joint distribution of* $\widetilde{\boldsymbol{L}}_r$ *and* $\widetilde{\boldsymbol{W}}_r$ *satisfies*

$$\lim_{r\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{L}_{i,r} \geq s_i;\, \widetilde{W}_{i,r} \geq t_i\right\}\right) = \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\overline{Z}_i \geq \min\{\mathbb{E}[B_i]s_i, t_i\}\right\}\right) \tag{26}$$

with $\overline{Z}_1, \ldots, \overline{Z}_N$ *defined in Sect. 2.*

Proof Equation (21) implies that the event $\{\widetilde{L}_{i,r} \geq s_i\}$ coincides with the event $\{\widetilde{W}_{i,r} \geq \widetilde{B}^R_{i,r} + \frac{1}{r}\sum_{j=1}^{rs_i} \widehat{B}_{i,j}\}$, as the $\widehat{B}_{i,j}$ can only take non-negative values. Thus, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{L}_{i,r} \geq s_i;\, \widetilde{W}_{i,r} \geq t_i\right\}\right) = \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r} \geq \max\{\widetilde{B}^R_{i,r} + \frac{1}{r}\sum_{j=1}^{rs_i} \widehat{B}_{i,j}, t_i\}\right\}\right).$$

Let $H^\epsilon_{i,r}$ be defined as before and recall that $\lim_{r\to\infty} \mathbb{P}(\bigcap_{i=1}^{N} H^\epsilon_{i,r}) = 1$, so that, due to the law of total probability,

$$\mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{L}_{i,r} \geq s_i;\, \widetilde{W}_{i,r} \geq t_i\right\}\right)$$
$$= \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r} \geq \max\{\widetilde{B}^R_{i,r} + s_i \frac{1}{rs_i}\sum_{j=1}^{rs_i}\widehat{B}_{i,j}, t_i\};\, H^{\epsilon}_{i,r}\right\}\right) + o(1).$$

Note that, according to Lemma 4.4, $\widetilde{B}^R_{i,r} \to 0$ as $r \to \infty$ for $i = 1, \ldots, N$, so that also $(\widetilde{B}^R_{1,r}, \ldots, \widetilde{B}^R_{N,r}) \to \mathbf{0}$ as $r \to \infty$. We thus obtain

$$\lim_{r\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r} \geq \max\{\mathbb{E}[B_i] + \epsilon, t_i\}\right\}\right) \leq \lim_{r\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{L}_{i,r} \geq s_i;\, \widetilde{W}_{i,r} \geq t_i\right\}\right)$$
$$\leq \lim_{r\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{N}\left\{\widetilde{W}_{i,r} \geq \max\{\mathbb{E}[B_i] - \epsilon, t_i\}\right\}\right).$$

By taking the limit $\epsilon \to 0$, an application of Theorem 3.1 and the notion that the distribution of $\overline{\boldsymbol{Z}}$ has no discontinuity points yield the desired result. □

Corollary 4.6 *For the scaled queue length vector $\widetilde{\boldsymbol{L}}_r$, it holds that*

$$\widetilde{\boldsymbol{L}}_r \xrightarrow{d} \left(\frac{1}{\mathbb{E}[B_1]}, \ldots, \frac{1}{\mathbb{E}[B_N]}\right)\overline{\boldsymbol{Z}},$$

as $r \to \infty$, with $\overline{\boldsymbol{Z}}$ defined in Sect. 2.

Proof The desired statement follows immediately from Proposition 4.5 by taking $t_1 = \ldots = t_N = 0$. □

5 Application to a two-layered network

In this section, we apply the results obtained so far in this paper to a network that is inspired by a manufacturing application and fits the class of so-called layered queueing networks (see e.g. [12–14]). We will also refer to this network as the *two-layered network*. We first describe the network in more detail in Sect. 5.1 and show that this particular model fits naturally in the general framework described in Sect. 2. Then, in Sect. 5.2, we study the question of how to compute the covariance matrix $\boldsymbol{\Gamma}$ of the N-dimensional Brownian motion $\boldsymbol{Z}$ based on this example. More specifically, we obtain expressions for the covariance terms $\gamma^C_{i,j}$, by using results from the literature on MAPs. We also compute the limiting distributions of $\widetilde{\boldsymbol{W}}_r$, $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{L}}_r$. Doing so in an exact fashion turns out to be hard. Therefore, we study how to numerically obtain the limiting distributions, by viewing $\overline{\boldsymbol{Z}}$ as an N-dimensional SRBM in Sect. 5.3. Finally, in Sect. 5.4, we conclude by means of simulation that the distribution of $\widetilde{\boldsymbol{W}}_r$ converges quickly to the distribution of $\overline{\boldsymbol{Z}}$ as $r \to \infty$, and therefore, that the heavy-traffic asymptotics constitute useful approximations for stable systems with a considerable load.

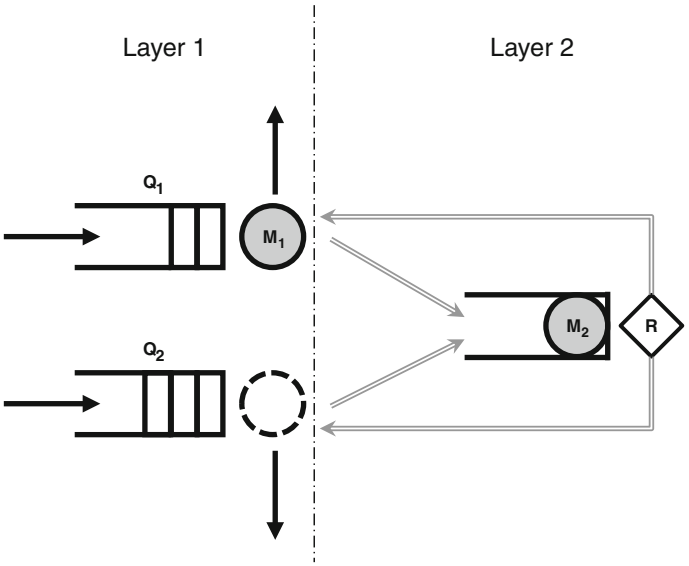

Fig. 1 The two-layered model under consideration

5.1 Description of the two-layered network

The two-layered network that we consider in this section is an extension of the machine-repair model (cf. [36, Chapter 5]) and consists of N machines $M_1, \ldots, M_N$ as well as a single repairman R, see Fig. 1. The second layer of this network constitutes the classical machine-repair model, where each machine breaks down after a stochastic lifetime, and the repairman repairs the machines in the order of breakdown. In the event of a breakdown, the machine moves to the repair buffer, where it will wait if the repairman is busy repairing, otherwise repair will start instantly. Contrary to the classical machine-repair model, we assume that each machine M_i also processes its own queue Q_i of products at a service speed of one when it is operational, which forms the first layer of the two-layered network.

When lifetimes and repair times follow a phase-type distribution, this networks fits the general model given in Sect. 2, as the availability of the Markov chains can then be modelled by a continuous-time Markov chain $\{\Phi(t), t \geq 0\}$. For the sake of brevity, we will assume in the remainder of Sect. 5 that $N = 2$ and that the lifetime and repair-time distributions of M_i are exponentially distributed with rate σ_i and ν_i, respectively. Then, $\{\Phi(t), t \geq 0\}$ operates on the state space $\mathcal{S} = \{(U, U), (U, R), (R, U), (W, R), (R, W)\}$. A state $\boldsymbol{\omega} = (\omega_1, \omega_2) \in \mathcal{S}$ represents for each machine M_i its condition of being up ($\omega_i = U$), in repair ($\omega_i = R$), or waiting in the repair buffer for repair ($\omega_i = W$) at time t. The generator matrix $\boldsymbol{Q}$ with elements $q_{i,j}$, $i, j \in \mathcal{S}$, that corresponds to this Markov chain is given by

$$\boldsymbol{Q} = \begin{pmatrix} -\sigma_1 - \sigma_2 & \sigma_2 & \sigma_1 & 0 & 0 \\ \nu_2 & -\nu_2 - \sigma_1 & 0 & \sigma_1 & 0 \\ \nu_1 & 0 & -\nu_1 - \sigma_2 & 0 & \sigma_2 \\ 0 & 0 & \nu_2 & -\nu_2 & 0 \\ 0 & \nu_1 & 0 & 0 & -\nu_1 \end{pmatrix},$$

and we let $q_i = -q_{i,i}$ be the sum of the outgoing rates of state i. The continuous-time Markov chain $\{\Phi(t), t \geq 0\}$ is irreducible and aperiodic, so that its invariant probability measure $\boldsymbol{\pi} = (\pi_j)_{j \in \mathcal{S}}$ is uniquely determined by the equations $\boldsymbol{\pi}\boldsymbol{Q} = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{1} = 1$ and can be obtained explicitly in terms of the model parameters $\sigma_1, \sigma_2, \nu_1$ and ν_2. Since the machines drain their queues of products at service rate one if they are operational (and zero otherwise), the connection with the general framework in Sect. 2 is completed by choosing the state-dependent service speeds as $\phi_i(\boldsymbol{\omega}) = \mathbb{1}_{\{\omega_i = U\}}$, where $\mathbb{1}_{\{A\}}$ denotes the indicator function on the event A.

5.2 Derivation of the covariance matrix

Now that the two-layered network is cast as a special instance of the general model given in Sect. 2, we show how to compute expressions for the covariance matrix $\boldsymbol{\Gamma}$ of the N-dimensional Brownian motion $\boldsymbol{Z}$ completely in terms of the model parameters. We do this based on the example of the two-layered network described in Sect. 5.1. However, the following methods can also be used to find the covariance matrix $\boldsymbol{\Gamma}$ for any instance of the model given in Sect. 2 without any conceptual complications. By (4), it remains to compute expressions for the covariance terms $\gamma^C_{i,j} = \lim_{t \to \infty} \frac{1}{t}\mathrm{Cov}[C_i(t), C_j(t)]$ for all $i, j \in \{1, \ldots, N\}$. In order to compute these, observe that the increments of $\{C_i(t), t \geq 0\}$ and $\{C_j(t), t \geq 0\}$ are conditionally independent given $\{\Phi(t), t \geq 0\}$. Therefore, we can view $\{(\Phi(t), C_i(t)), t \geq 0\}$, $\{(\Phi(t), C_j(t)), t \geq 0\}$ and $\{(\Phi(t), C_i(t) + C_j(t)), t \geq 0\}$ as MAPs. As a consequence, a functional central limit theorem for MAPs obtained in [34] can be applied to compute $\gamma^C_{i,j}$ for all $i, j \in \{1, \ldots, N\}$. Let $\omega_{\mathrm{ref}} \in \mathcal{S}$ be an arbitrary reference state and let T_k be the k-th time after $t = 0$ that the Markov chain $\{\Phi(t), t \geq 0\}$ enters this state. Then, the results of [34] imply the following lemma.

Lemma 5.1 *Suppose that $\{Y(t), t \geq 0\}$ is a Markov-modulated drift process, of which the drift equals d_k when the Markov chain $\{\Phi(t), t \geq 0\}$ is in state $k \in \mathcal{S}$. Furthermore, suppose that $|d_k| < \infty$ for each $k \in \mathcal{S}$ and that $\sum_{k \in \mathcal{S}} \pi_k d_k = 0$. Then, $\{\frac{1}{\sqrt{s}} Y(st), t \geq 0\}$ converges in distribution, as $s \to \infty$, to a driftless Brownian motion starting at 0 with variance parameter*

$$\sigma_Y^2 = 2 \sum_{k \in \mathcal{S}} \pi_k \left(\frac{d_k^2}{q_k} + \sum_{l \in \mathcal{S} \setminus \{\{k\} \cup \{\omega_{ref}\}\}} \frac{q_{k,l} d_k f_l}{q_k} \right), \tag{27}$$

where the f_l-parameters are the unique solution to the set of linear equations

$$f_m = \frac{d_m}{q_m} + \sum_{n \in \mathcal{S} \setminus \{\{m\} \cup \{\omega_{ref}\}\}} \frac{q_{m,n}}{q_m} f_n.$$

In particular, we have that $\lim_{t\to\infty} \frac{1}{t}\mathrm{Var}[Y(t)] = \sigma_Y^2$.

Proof The convergence in distribution immediately follows from [34, Theorem 3.4] by taking $X(t) = \Phi(t)$ and $D_{i,j} = V_{i,j} = \upsilon_i = 0$ for all i, j in the notation of that paper. To show the result for the asymptotic variance of the modulated process Y, observe that $M(t) = \max_{k:T_k\le t}\{k\}$ counts the number of times the Markov chain returned to the reference state until time t, so that $\{M(t), t \ge 0\}$ can be interpreted as a (delayed) renewal process. As a consequence,

$$\begin{aligned}\lim_{t\to\infty}\frac{\mathrm{Var}[Y(t)]}{t} &= \lim_{t\to\infty}\frac{\mathrm{Var}[Y(\sum_{i=1}^{M(t)}(T_{i+1}-T_i))]+o(t)}{t}\\ &= \lim_{t\to\infty}\frac{\mathbb{E}[M(t)]\mathrm{Var}[Y(T_2-T_1)]+\mathrm{Var}[M(t)]\mathbb{E}[Y(T_2-T_1)]^2}{t}\\ &= \mathrm{Var}[Y(T_2-T_1)]\lim_{t\to\infty}\frac{\mathbb{E}[M(t)]}{t} = \frac{\mathrm{Var}[Y(T_2-T_1)]}{\mathbb{E}[T_2-T_1]}.\end{aligned}$$

Section 3 in [34] shows that $\mathrm{Var}[Y(T_2-T_1)] = \mathbb{E}[(Y(T_2-T_1))^2] = \sigma_Y^2\mathbb{E}[T_2-T_1]$, which concludes the proof. □

We now apply this lemma to obtain the covariance matrix for the two-layered model with $N = 2$. In particular, to compute $\sigma^2_{C,1}$, we study the process $Y(t) = C_1(t) - \mathbb{E}[C_1(t)] = C_1(t) - (\pi_{(U,U)}+\pi_{(U,R)})t$ with conditional drift $d_k = \mathbb{1}_{\{k\in\{(U,U),(U,R)\}\}} - (\pi_{(U,U)}+\pi_{(U,R)})$ when the modulator Φ resides in state k. As $\mathrm{Var}[Y(t)] = \mathrm{Var}[C_1(t)]$ for any $t \ge 0$, an expression for $\sigma^2_{C,1}$ is then readily given in Lemma 5.1 by (27). An expression for $\sigma^2_{C,2}$ can be found similarly to the computations above or simply by interchanging the indices in the expression of $\sigma^2_{C,1}$. Observe that an expression for $\lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[C_1(t)+C_2(t)]$ can also be found using the same technique, but now considering the process $Y(t) = C_1(t)+C_2(t)-(\mathbb{E}[C_1(t)+C_2(t)]) = C_1(t)+C_2(t)-(2\pi_{(U,U)}+\pi_{(U,R)}+\pi_{(R,U)})t$ instead with $d_k = \mathbb{1}_{\{k\in\{(U,U),(U,R)\}\}}+\mathbb{1}_{\{k\in\{(U,U),(R,U)\}\}} - (2\pi_{(U,U)}+\pi_{(U,R)}+\pi_{(R,U)})$. Again, it then holds that an expression for $\lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[C_1(t)+C_2(t)]$ is given in (27). After these computations, the covariance matrix $\mathbf{\Gamma}$ can be expressed explicitly in terms of the model parameters. The covariance parameters $\gamma^C_{1,1}$ and $\gamma^C_{2,2}$ are by definition equal to $\sigma^2_{C,1}$ and $\sigma^2_{C,2}$, for which we have already derived explicit expressions. As for the remaining parameters, we have that both $\gamma^C_{1,2}$ and $\gamma^C_{2,1}$ are equal to

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\mathrm{Cov}[C_1(t),C_2(t)]\\ &\quad= \frac{1}{2}\left(\lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[C_1(t)+C_2(t)] - \lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[C_1(t)] - \lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[C_2(t)]\right),\end{aligned}$$

where all of the terms between the brackets in the right-hand side are now known. As the rest of the terms appearing in (4) were already expressed in terms of the model parameters, the covariance matrix $\mathbf{\Gamma}$ is now explicitly known.

5.3 Numerical evaluation of the limiting distribution of $\overline{\boldsymbol{Z}}$

Now that $\boldsymbol{\Gamma}$ can be computed explicitly, we investigate in this section the joint distribution of $\overline{\boldsymbol{Z}}$, i.e. the limiting distribution of the scaled workload $\widetilde{\boldsymbol{W}}_r$, in stationarity. Since the limiting distributions of $\widetilde{\boldsymbol{D}}_r$ or $\widetilde{\boldsymbol{L}}_r$ equal the distribution of $\overline{\boldsymbol{Z}}$ up to a scalar as observed in Corollaries 4.2 and 4.6, the results also directly relate to the limiting distributions of the scaled virtual waiting time and the scaled queue length.

To study the joint distribution of $\overline{\boldsymbol{Z}}$ as defined in Theorem 3.1, we first observe that this distribution equals the stationary distribution of an N-dimensional SRBM. In particular, by the definitions of $\boldsymbol{Z}(t)$ and $\overline{Z}_i(t)$ in Sect. 2 and Theorem 3.1, respectively, we have that the process $\overline{\boldsymbol{Z}}(t) = \{\overline{Z}_1(t), \ldots, \overline{Z}_N(t)\}$ satisfies

$$\begin{aligned}
\overline{\boldsymbol{Z}}(t) &= \left(\sup_{s\in[0,t]} \{Z_1(s)\}, \ldots, \sup_{s\in[0,t]} \{Z_N(s)\} \right) \\
&\overset{d}{=} \left(\sup_{s\in[0,t]} \{Z_1(t) - Z_1(t-s)\}, \ldots, \sup_{s\in[0,t]} \{Z_N(t) - Z_N(t-s)\} \right) \\
&= \left(Z_1(t) - \inf_{s\in[0,t]} \{Z_1(s)\}, \ldots, Z_N(t) - \inf_{s\in[0,t]} \{Z_N(s)\} \right) \\
&= \boldsymbol{Z}(t) + \boldsymbol{R}\boldsymbol{Y}(t),
\end{aligned}$$

where the equality in distribution follows since multi-dimensional Brownian motions are time-reversible [4, Lemma II.2]. In this representation, $\boldsymbol{R}$ is the $N \times N$ identity matrix, and $\boldsymbol{Y}(t) = (Y_1(t), \ldots, Y_N(t)) = (-\inf_{s\in[0,t]}\{Z_1(s)\}, \ldots, -\inf_{s\in[0,t]}\{Z_N(s)\})$. Observe that $\{\boldsymbol{Y}(t), t \geq 0\}$ is a continuous, non-decreasing process starting in $\boldsymbol{0}$, of which the elements Y_i can only increase at times t when $\overline{Z}_i(t) = 0$. A process with such a representation is known to be an SRBM on the state space $\mathbb{R}_+^N$ (see, for example, [7, Section 7.4]). By letting $t \to \infty$, it is now clear that the joint distribution of $\overline{\boldsymbol{Z}}$ coincides with the stationary distribution of an SRBM on the non-negative orthant with drift vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Gamma}$ and reflection matrix $\boldsymbol{R}$.

Computing the stationary distribution of a multi-dimensional SRBM is in general a challenging problem. Although the SRBM corresponding to our model satisfies the conditions derived in [18] for a unique stationary distribution to exist, it does not necessarily satisfy the necessary requirements found in [19] for this distribution to have a product form. A numerical approach obtained in [9] to compute the stationary distribution is, however, applicable to our setting.

We now apply this numerical algorithm to the two-layered network and observe several parameter effects. Note that for the two-layered network, $\boldsymbol{R}$ resolves to a 2×2 identity matrix, and the underlying Brownian motion $\{\boldsymbol{Z}(t), t \geq 0\}$ has a drift vector $\boldsymbol{\mu} = \big(-\beta_1(\pi_{(U,U)} + \pi_{(U,R)}), -\beta_2(\pi_{(U,U)} + \pi_{(R,U)})\big)$ and a covariance matrix $\boldsymbol{\Gamma} = \operatorname{diag}\left(\frac{\mathbb{E}[B_1^2]}{\mathbb{E}[B_1]}(\pi_{(U,U)} + \pi_{(U,R)}), \frac{\mathbb{E}[B_2^2]}{\mathbb{E}[B_2]}(\pi_{(U,U)} + \pi_{(R,U)}) \right) + \boldsymbol{\Gamma}^C$, where $\boldsymbol{\Gamma}^C$ is a 2×2 matrix consisting of the elements $\gamma_{i,j}^C$ computed in Sect. 5.2. For a number of instances of the two-layered network, we have computed several characteristics of the stationary distribution, such as the first two moments and the cross-moment of $\overline{Z}_1$

Table 1 Numerical results for several instances of the two-layered network

Instance no.	β_1	β_2	$\mathbb{E}[B_1]$	$\mathbb{E}[B_1^2]$	$\mathbb{E}[B_2]$	$\mathbb{E}[B_2^2]$	σ_1	σ_2	ν_1	ν_2	$\mathbb{E}[\overline{Z}_1]$	$\mathbb{E}[\overline{Z}_2]$	Corr $[\overline{Z}_1, \overline{Z}_2]$
1	1	1	1	2	1	2	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	4.33	4.33	0.274
2	$\frac{1}{2}$	1	1	2	1	2	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	8.67	4.33	0.228
3	1	1	1	5	1	5	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	5.83	5.83	0.195
4	1	1	$\frac{1}{2}$	$\frac{1}{2}$	2	8	$\frac{1}{5}$	$\frac{1}{20}$	$\frac{1}{5}$	$\frac{1}{20}$	3.84	7.18	0.446
5	1	1	1	2	1	2	1	1	1	1	1.33	1.33	0.080
6	1	1	1	2	1	2	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{5}$	$\frac{1}{5}$	2.06	2.06	0.124

and $\overline{Z}_2$. The results are summarised in Table 1, where for each of the instances the calculated values for $\mathbb{E}[\overline{Z}_1]$, $\mathbb{E}[\overline{Z}_2]$ and the correlation coefficient $\mathrm{Corr}[\overline{Z}_1, \overline{Z}_2] = \frac{\mathbb{E}[\overline{Z}_1\overline{Z}_2]-\mathbb{E}[\overline{Z}_1]\mathbb{E}[\overline{Z}_2]}{\sqrt{\mathbb{E}[\overline{Z}_1^2]-\mathbb{E}[\overline{Z}_1]^2}\sqrt{\mathbb{E}[\overline{Z}_2^2]-\mathbb{E}[\overline{Z}_2]^2}}$ are given. Recall that the marginal distribution of $\overline{Z}_i$ is exponential, so that $\mathbb{E}[\overline{Z}_i^2] = 2\mathbb{E}[\overline{Z}_i]^2$. Observe also that the limiting distributions of $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{L}}_r$ are equal to the distribution of $\overline{\boldsymbol{Z}}$ up to a scalar, so that $\mathrm{Corr}[\overline{Z}_1, \overline{Z}_2]$ does not only represent the correlation coefficient pertaining to the limiting distribution of the scaled workload $\widetilde{\boldsymbol{W}}_r$, but also to that of the scaled virtual waiting time and the scaled queue length. It follows from Table 1 that the competition between the machines of the repair facilities can be of such a level, that the correlation coefficient pertaining to the queue lengths is significant. Moreover, by taking the first instance as a reference, we observe that the correlation coefficient is highly influenced by the relative convergence speed of the arrival rates (instance no. 2), the variability of the service times (instance no. 3), the level of asymmetry in the model parameters (instance no. 4), the frequency of machine breakdowns and speed of machine repairs with respect to the arrivals and services of products (instance no. 5), and the duration of the machine lifetimes with respect to that of their repairs (instance no. 6).

5.4 Comparison with simulation results

We end this section with an assessment of the quality of the distribution of $\overline{\boldsymbol{Z}}$ as an approximation for the joint workload distribution in systems with a considerable load. In Table 2, simulation results for the scaled workload $\widetilde{\boldsymbol{W}}_r$ corresponding to the values $r = 5, 10, 20$ are given for each of the instances given in Table 1. Recall that $\rho_i = 1 - \frac{\beta_i}{r}$, so that $r = 5, 10, 20$ corresponds to $\rho_i = 0.8, 0.9, 0.95$ if $\beta_i = 1$. Thus, the values $r = 5, 10, 20$ represent systems that operate under a high load, as is often the case in practice.

As expected, Tables 1 and 2 suggest that the distribution of $\overline{\boldsymbol{Z}}$ generally approximates the distribution of $\widetilde{\boldsymbol{W}}_r$ well in terms of marginal means and the correlation coefficient. In particular, the tables confirm that $\mathbb{E}[\widetilde{W}_{i,r}]$ converges to $\mathbb{E}[\overline{Z}_i]$ from below as $r \to \infty$ at a fast rate, so that $\mathbb{E}[\overline{Z}_i]$ is a provably useful upper bound close to the actual value of $\mathbb{E}[\widetilde{W}_{i,r}]$ for large r (i.e. significantly loaded systems). Surprisingly,

Table 2 Simulation results for $\widetilde{\boldsymbol{W}}_5$, $\widetilde{\boldsymbol{W}}_{10}$ and $\widetilde{\boldsymbol{W}}_{20}$

Instance no.	$\mathbb{E}[\widetilde{W}_{1,5}]$	$\mathbb{E}[\widetilde{W}_{1,10}]$	$\mathbb{E}[\widetilde{W}_{1,20}]$	$\mathbb{E}[\widetilde{W}_{2,5}]$	$\mathbb{E}[\widetilde{W}_{2,10}]$	$\mathbb{E}[\widetilde{W}_{2,20}]$	$\mathrm{Corr}[\widetilde{W}_{1,5}, \widetilde{W}_{2,5}]$	$\mathrm{Corr}[\widetilde{W}_{1,10}, \widetilde{W}_{2,10}]$	$\mathrm{Corr}[\widetilde{W}_{1,20}, \widetilde{W}_{2,20}]$
1	3.46	3.90	4.12	3.46	3.90	4.12	0.262	0.271	0.273
2	7.80	8.23	8.45	3.46	3.90	4.12	0.217	0.225	0.228
3	4.42	5.11	5.47	4.42	5.11	5.47	0.180	0.189	0.192
4	3.08	3.46	3.65	5.72	6.46	6.82	0.466	0.460	0.453
5	1.07	1.20	1.27	1.07	1.20	1.27	−0.053	0.001	0.044
6	1.64	1.85	1.95	1.64	1.85	1.95	0.121	0.126	0.125

the rate at which $\mathbb{E}[\widetilde{W}_{i,r}]$ converges to $\mathbb{E}[\overline{Z}_i]$ does not seem to differ much between the model instances. The slowest convergence occurs in the third model instance due to the high variability of the service times, but it does not deviate much from the other instances. The only outlying rate of convergence can be found in the expected scaled waiting time of the first queue in the second model instance, where convergence is a lot faster. However, this is obvious by the nature of our scaling, since $\beta_1 = 1/2$ for that model instance instead of $\beta_1 = 1$. Furthermore, the values of $\mathrm{Corr}[\overline{Z}_1, \overline{Z}_2]$ turn out to be accurate approximations of the values $\mathrm{Corr}[\widetilde{W}_{1,r}, \widetilde{W}_{2,r}]$, $r = 5, 10, 20$, for almost all of the model instances. Thus, the limiting distribution seems to capture the correlation structure between the queue lengths in the stable case rather well. One can argue that the fifth model instance is an exception to this. However, due to the high frequency of machine breakdowns and repairs, there hardly is any correlation between the queues, making correlation coefficients hard to approximate accurately.

Acknowledgments The authors are indebted to Sem Borst and Onno Boxma for providing valuable comments on earlier drafts of this paper. Furthermore, the authors wish to thank an anonymous referee for providing constructive criticism and for making several suggestions that led to an improved exposition of the contents in this paper. Funded in the framework of the STAR-project 'Multilayered queueing systems' by the Netherlands Organization for Scientific Research (NWO). The research of Maria Vlasiou is also partly supported by an NWO individual Grant through Project 632.003.002. The research of Bert Zwart is partly supported by an NWO VIDI grant.

References

1. Asmussen, S.: Applied Probability and Queues. Springer, New York (2003)
2. Ata, B., Shneorson, S.: Dynamic control of an M/M/1 service system with adjustable arrival and service rates. Manag. Sci. **52**, 1778–1791 (2006)
3. Bekker, R.: Queues with State-Dependent Rates. PhD thesis, Eindhoven University of Technology (2005)
4. Bertoin, J.: Lévy Processes. Cambridge University Press, Cambridge (1996)
5. Budhiraja, A., Lee, C.: Stationary distribution convergence for generalized Jackson networks in heavy traffic. Math. Oper. Res. **34**, 45–56 (2009)
6. Chen, H., Whitt, W.: Diffusion approximations for open queueing networks with service interruptions. Queueing Syst. **13**, 335–359 (1993)
7. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks. Springer, New York (2001)
8. Choudhury, G.L., Mandelbaum, A., Reiman, M.I., Whitt, W.: Fluid and diffusion limits for queues in slowly changing environments. Commun. Stat. Stoch. Models **13**, 121–146 (1997)
9. Dai, J.G., Harrison, J.M.: Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. Ann. Appl. Probab. **2**, 65–86 (1992)
10. Debicki, K., Kosiński, K.M., Mandjes, M.: Gaussian queues in light and heavy traffic. Queueing Syst. **71**, 137–149 (2012)
11. Dieker, A.B., Moriarty, J.: Reflected Brownian motion in a wedge: sum-of-exponential stationary densities. Electron. Commun. Probab. **14**, 1–16 (2009)
12. Dorsman, J.L., Bhulai, S., Vlasiou, M.: Dynamic server assignment in an extended machine-repair model. IIE Trans. (2014). doi:10.1080/0740817X.2014.928962
13. Dorsman, J.L., Boxma, O.J., Vlasiou, M.: Marginal queue length approximations for a two-layered network with correlated queues. Queueing Syst. **75**, 29–63 (2013)
14. Dorsman, J.L., van der Mei, R.D., Vlasiou, M.: Analysis of a two-layered network by means of the power-series algorithm. Perform. Eval. **70**, 1072–1089 (2013)
15. Gamarnik, D., Zeevi, A.: Validity of heavy traffic steady-state approximation in generalized Jackson networks. Ann. Appl. Probab. **16**, 56–90 (2006)

16. George, J.M., Harrison, J.M.: Dynamic control of a queue with adjustable service rate. Oper. Res. **49**, 720–731 (2001)
17. Halfin, S.: Steady-state distribution for the buffer content of an M/G/1 queue with varying service rate. SIAM J. Appl. Math. **23**, 356–363 (1972)
18. Harrison, J.M., Williams, R.J.: Brownian models of open queueing networks with homogeneous customer populations. Stochastics **22**, 77–115 (1987)
19. Harrison, J.M., Williams, R.J.: Multidimensional reflected Brownian motions having exponential stationary distributions. Ann. Probab. **15**, 115–137 (1987)
20. Hopp, W.J., Iravani, S.M.R., Yuen, G.J.: Operations systems with discretionary task completion. Manag. Sci. **53**, 61–77 (2006)
21. Ivanovs, J., Boxma, O.J., Mandjes, M.R.H.: Singularities of the matrix exponent of a Markov additive process with one-sided jumps. Stoch. Process. Their Appl. **120**, 1776–1794 (2010)
22. Jelenković, P.R., Momčilović, P., Zwart, B.: Reduced load equivalence under subexponentiality. Queueing Syst. **46**, 97–112 (2004)
23. Kella, O., Whitt, W.: Diffusion approximations for queues with server vacations. Adv. Appl. Probab. **22**, 706–729 (1990)
24. Kingman, J.F.C.: The single server queue in heavy traffic. Math. Proc. Camb. Philos. Soc. **57**, 902–904 (1961)
25. Kingman, J.F.C.: The heavy traffic approximation in the theory of queues. In: Smith, W.L., Wilkinson, W.E. (eds.) Proceedings of the Symposium on Congestion Theory, pp. 137–159. University of North Carolina Press, Chapel Hill (1965)
26. Kosiński, K.M., Boxma, O.J., Zwart, B.: Convergence of the all-time supremum of a Lévy process in the heavy-traffic regime. Queueing Syst. **67**, 295–304 (2011)
27. Mahabhashyam, S.R., Gautam, N.: On queues with Markov modulated service rates. Queueing Syst. **51**, 89–113 (2005)
28. Mitra, D.: Stochastic theory of a fluid model of producers and consumers coupled by a buffer. Adv. Appl. Probab. **20**, 646–676 (1988)
29. Núñez-Queija, R.: A queueing model with varying service rate for ABR. In: Puigjaner, R., Savino, N.N., Serra, B. (eds.) Proceedings of the 10th International Conference on Computer Performance Evaluation: Modelling Techniques and Tools, pp. 93–104. Springer, Berlin (1998)
30. Purdue, P.: The M/M/1 queue in a Markovian environment. Oper. Res. **22**, 562–569 (1974)
31. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion. Springer, New York (1999)
32. Shneer, S., Wachtel, V.: A unified approach to the heavy-traffic analysis of the maximum of random walks. Theory Probab. Appl. **55**, 332–341 (2011)
33. Siebert, F.: Real-time garbage collection in multi-threaded systems on a single processor. In: Proceedings of the 20th IEEE Real-Time Systems Symposium, pp. 277–278 (1999)
34. Steichen, J.L.: A functional central limit theorem for Markov additive processes with an application to the closed Lu-Kumar network. Stoch. Models **17**, 459–489 (2001)
35. Stidham Jr, S., Weber, R.R.: Monotonic and insensitive optimal policies for control of queues with undiscounted costs. Oper. Res. **37**, 611–625 (1989)
36. Takács, L.: Introduction to the Theory of Queues. Oxford University Press, New York (1962)
37. Takine, T.: Single-server queues with Markov-modulated arrivals and service speed. Queueing Syst. **49**, 7–22 (2005)
38. Tse, D., Viswanath, P.: Fundamentals of Wireless Communication. Cambridge University Press, Cambridge (2005)
39. Tzenova, E.I., Adan, I.J.B.F., Kulkarni, V.G.: Fluid models with jumps. Stoch. Models **21**, 37–55 (2005)
40. van der Mei, R.D., Hariharan, R., Reeser, P.K.: Web server performance modeling. Telecommun. Syst. **16**, 361–378 (2001)
41. Weber, R.R., Stidham Jr, S.: Optimal control of service rates in networks of queues. Adv. Appl. Probab. **19**, 202–218 (1987)
42. Whitt, W.: Asymptotic formulas for Markov processes with applications to simulation. Oper. Res. **40**, 279–291 (1992)
43. Whitt, W.: Stochastic-Process Limits. Springer, New York (2002)
44. Zwart, B.: Heavy-traffic asymptotics for the single-server queue with random order of service. Oper. Res. Lett. **33**, 511–518 (2004)