EURANDOM PREPRINT SERIES
2013-005
March 6, 2013

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

J.L. Dorsman, M. Vlasiou, B. Zwart
ISSN 1389-2355

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

J.L. Dorsman [*] [†] j.l.dorsman@tue.nl
M. Vlasiou [*] [†] m.vlasiou@tue.nl
B. Zwart [†] [‡] [*] [§] Bert.Zwart@cwi.nl

March 6, 2013

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.

Keywords: Layered queueing networks, machine-repair model, functional central limit theorem, semi-martingale reflected Brownian motion.

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. Even if one were interested in marginal processes, one would run into the problem that the service speed process does not have independent increments, complicating the analysis considerably. 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 e.g. [18]).

The study of this general network is motivated in part by the fact that it captures a large class of so-called layered queueing networks (LQNs). LQNs are queueing networks that are characterised by simultaneous or separate phases where entities are no longer necessarily classified in the traditional roles of ‘servers’ and ‘customers’, but may also have a dual role of being either a server to higher-layer entities or a customer to lower-layer entities. Recent applications in engineering, business, and the public sector led to systems with complex, often layered, service architectures. For example, this phenomenon occurs naturally in various computer-science problems; see [20] and references therein for an overview. Another important example of an LQN that we will refer to later is a network inspired by a manufacturing application. This network consists of machines, that each process their own queue of products in the role of upper-layer servers, but break down from time to time so that they require service

Funded in the framework of the STAR-project “Multilayered queueing systems” by the Netherlands Organization for Scientific Research (NWO). The research of M. Vlasiou is also partly supported by an NWO individual grant through project 632.003.002.
[*] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands
[†] Stochastics, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
[‡] Department of Mathematics, Faculty of Sciences, VU University Amsterdam, Amsterdam, The Netherlands
[§] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

from a repairman. At moments of breakdown, the machines take the role of customers at the lower layer, where the repairman acts as the server. This model can be interpreted as an extension of the well-known machine-repair problem (cf. [44, Chapter 5]). Since the number of machines is larger than the number of repairmen, the machines compete with each other for access to the repairman. As a result, consecutive downtimes of a single machine are correlated. These dynamics in the lower layer make exact analysis of the queues in the upper layer notoriously difficult, so that one has to resort to approximations (see [16, 17, 18]). The extended machine-repair model fits the network studied in this paper.

It is interesting to note that the layers of an LQN may interact significantly. For instance, we will observe in the sequel that under heavy-traffic assumptions, the workload, virtual waiting time and queue length processes for a single-server queue in isolation exhibit so-called state-space collapse (cf. [39]). However, in the limit these processes are still dependent on characteristics of the service-speed processes pertaining to the servers. In the LQN-setting, this means that the lower layer (modelled by the single continuous-time Markov chain) significantly affects the dynamics of the upper-layer queues. For example, the marginal distributions of the workload, virtual waiting time and queue length processes will turn out to be exponential with parameters that involve the asymptotic mean and variance of the service speed process pertaining to the corresponding queue. As a result, the formulation and study of LQNs is important, as analysis of each of the layers separately appears to be insufficient.

Another important feature of the model is the fact that the service speeds vary over time. In many classical queueing models, service rates are assumed to be constant. This assumption, however, may not always be appropriate. For example, in telecommunication systems with congestion control mechanisms or systems where the servers represent human beings, the service speed may be influenced by factors such as the workload present in the system. This leads to the formulation of queues with state-dependent service rates; see e.g. [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 (e.g. [2, 21, 43, 47]) or to optimise a trade-off between service quality and service speed (e.g. [26]) 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. Several single-server queueing models with Markov-modulated service speeds have been studied in the literature. The case where the server alternates between two service speeds has been analysed in [5, 49]. In [22, 37], models are considered where the service speed of the server is governed by a birth-and-death process. Results for the case where the service speed is governed by an arbitrary continuous-time Markov process can be found in [38], which analyses the busy period of the server and stability conditions, and in [34], where matrix geometric methods are used to approximate performance measures. In [45], exact results are derived for a system where arrivals occur only at transition epochs of the modulating Markov process. 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.

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 [31]; see [32] 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 [27]. Less straightforward computations might involve studying the singularities of the characterising matrix exponent pertaining to the reflected MAP (cf. [27]). In the past, stationary distributions for special cases of reflected MAPs have also been analysed by studying its spectral expansion (e.g. [35]) or by determining the boundary probabilities in terms of the solution of a generalised eigenvalue problem (e.g. [46]).

For our heavy-traffic analysis, we will use a functional central limit theorem approach mainly developed by Iglehart and Whitt; see [48] for an overview. This approach requires a continuous mapping argument, and the interchange of steady-state and heavy-traffic limits. As will also turn out for our case, this is not always trivial; see for example [14, 33].

As we study networks with general service speeds, our model also captures a class of queues with service interruptions. Single-server queues with service interruptions have received some interest in the heavy-traffic literature. In particular, in [30], a single-server queue is considered where the durations and the frequency of the vacations, which occur at moments the queue empties, do not scale with the traffic intensity. Its heavy-traffic asymptotics are shown to be equivalent to those for similar queues without service interruptions, but have different rates. This paper also considers queues with rare long service interruptions, i.e., queues where the durations and frequency scale with the traffic intensity appropriately. Following this paper, queueing networks with rare long service interruptions were studied in [8] and [48, Section 14.7]. As opposed to these models, our model incorporates the possibility for the durations of consecutive service interruptions to be interdependent through the Markovian random environment; see also [10]. Furthermore, the start of a service interruption in our model is not restricted to a point in time the queue empties, and the durations do not depend on the traffic intensity.

For the network 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 $\mathbb{R}_+^N$. Such an SRBM behaves like a standard N-dimensional Brownian motion in the non-negative orthant $\mathbb{R}_+^N$, but is pushed back at the $(N-1)$-dimensional boundaries of the orthant in a direction specified by the reflection matrix.

In many queueing networks, SRBMs arise as the heavy-traffic limit of the workload process, see e.g. [6]. As a result, approximations for queueing networks have been proposed by replacing the workload process with an SRBM, as these so-called Brownian models require less restrictive assumptions than the classical results for queueing networks and work particularly well when the system is heavily utilised (see e.g. [23]). Regarding the stability of an SRBM, necessary and sufficient conditions are derived in [24] for a unique stationary distribution to exist under certain assumptions of the reflection matrix. For general reflection matrices, necessary and sufficient stability conditions are obtained in [7, 19] for the cases $N = 2$ and $N = 3$. As for the stationary distribution itself, even when positive conclusions can be drawn about its existence, the computation of it is a hard problem when $N \geq 2$. It is shown in [25] that under rather strict assumptions on the reflection matrix and the covariance matrix of the underlying Brownian motion, the stationary distribution has a product form, each marginal being exponential. For $N = 2$, tail asymptotics for the stationary distribution are derived in [12, 13]. Conjectures on the tail asymptotics for higher dimensions are given in [36]. For two-dimensional SRBMs in a wedge, necessary and sufficient conditions are defined in [15] for the stationary density to be written as a finite sum of terms of exponential product form.

In our case, the reflection matrix is an identity matrix, so that positive conclusions about the existence of a stationary distribution can be drawn. However, computing this distribution is challenging. The conditions needed for the stationary distribution to have a product form do not generally apply to our model, and results such as those of [15] seem hard to translate to our setting. In this paper, we therefore use the numerical methods developed in [11] 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 Section 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 Section 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.

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, $\mathbf{0}$ and $\mathbf{1}$ represent vectors of appropriate size where each of the elements are equal to zero and one respectively.

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. 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 $\sigma^2_{V,i} = \text{Var}[V_i(1)] = \mathbb{E}[N_i(1)]\text{Var}[B_i] + \text{Var}[N_i(1)]\mathbb{E}[B_i]^2 = \text{Var}[B_i] + \mathbb{E}[B_i]^2 = \mathbb{E}[B_i^2]$. Note that the arrival process has stationary and independent increments, so that $t^{-1}\mathbb{E}[V_i(t)] = m_{V,i}$ and $t^{-1}\text{Var}[V_i(t)] = \sigma^2_{V,i}$ for any $t > 0$.

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

$$C_i(t) = \int_{s=0}^{t} \phi_i(\Phi(s))ds.$$

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

$$m_{C,i} = \mathbb{E}[C_i(1)] = \int_{s=0}^{1} \sum_{\omega \in \boldsymbol{\mathcal{S}}} \phi_i(\omega)\mathbb{P}(\Phi(s) = \omega)ds = \sum_{\omega \in \boldsymbol{\mathcal{S}}} \phi_i(\omega)\pi_\omega.$$

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 $\lim_{t\to\infty} t^{-1}\text{Var}[C_i(t)]$ by $\sigma^2_{C,i}$. Similarly, the long-run time-averaged covariance between the service speed processes of the servers at Q_i and Q_j is represented by $\gamma^C_{i,j} = \lim_{t\to\infty} \frac{1}{t}\text{Cov}[C_i(t), C_j(t)]$. Computing expressions for $\sigma^2_{C,i}$ and $\gamma^C_{i,j}$ is not trivial. We focus on this problem in Section 5.2.

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 $m_{C,i}$ 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 $\lim_{r\to\infty} f(r)/g(r) = 0$.

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 $W_{i,r}(0) = 0$:

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

As the joint process $\{(C_1(t), \ldots, C_N(t)), t \geq 0\}$ has stationary increments, we have that the vector $\Big(C_1(t) - C_1(s), \ldots, C_N(t) - C_N(s)\Big)$ is in distribution equal to $\Big(C_1(t-s), \ldots, C_N(t-s)\Big)$. By noting that the joint process $\{(V_1(\lambda_{1,r}t), \ldots, V_N(\lambda_{N,r}t)), t \geq 0\}$ has reversible increments, substituting $u = t - s$ and subsequently taking the limit $u \to \infty$ (the steady-state limit), we obtain

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

where $\stackrel{d}{=}$ means equality in distribution. 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 (1) that the scaled workload can be written in terms of the similarly scaled net-input process. After scaling time by a factor r^2, we have

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

Due to the time scaling by r^2, we can obtain heavy-traffic limits for the joint scaled net-input process involved in (2) using the functional central limit theorem (cf. [48]). In particular, we have that

$$\Big\{ \Big(\frac{V_1(\lambda_{1,r}r^2 t) - \mathbb{E}[V_1(\lambda_{1,r}r^2 t)]}{\sqrt{\lambda_{1,r}}r}, \ldots, \frac{V_N(\lambda_{N,r}r^2 t) - \mathbb{E}[V_N(\lambda_{N,r}r^2 t)]}{\sqrt{\lambda_{N,r}}r} \Big), t \geq 0 \Big\} \stackrel{d}{\to} \{\boldsymbol{Z}_V(t), t \geq 0\} \tag{3}$$

and

$$\Big\{ \Big(\frac{C_1(r^2 t) - \mathbb{E}[C_1(r^2 t)]}{r}, \ldots, \frac{C_n(r^2 t) - \mathbb{E}[C_N(r^2 t)]}{r} \Big), t \geq 0 \Big\} \stackrel{d}{\to} \{\boldsymbol{Z}_C(t), t \geq 0\}, \tag{4}$$

as $r \to \infty$, where $\{\boldsymbol{Z}_V(t), t \geq 0\}$ and $\{\boldsymbol{Z}_C(t), t \geq 0\}$ are N-dimensional Brownian motions. As the arrival processes $\{V_i(t), t \geq 0\}$, $i = 1, \ldots, N$ are independent, $\{\boldsymbol{Z}_V(t), t \geq 0\}$ has zero drift and covariance matrix $\boldsymbol{\Gamma}^V = diag(\sigma^2_{V,1}, \ldots, \sigma^2_{V,N})$. The Brownian motion $\{\boldsymbol{Z}_C(t), t \geq 0\}$ has zero drift, and covariance matrix $\boldsymbol{\Gamma}^C$ with elements $\Gamma^C_{i,j} = \gamma^C_{i,j}$. To derive a heavy-traffic limit for the joint scaled net-input process based on (3) and (4), note that $\mathbb{E}[V_i(\lambda_{i,r}r^2 t)] = \lambda_{i,r}r^2\mathbb{E}[B_i]t$ and $\mathbb{E}[C_i(r^2 t)] = m_{C,i}r^2 t$, so that

$$\frac{\mathbb{E}[C_i(r^2 t)] - \mathbb{E}[V_i(\lambda_{i,r}r^2 t)]}{r} = \frac{m_{C,i}r^2 t - \lambda_{i,r}r^2\mathbb{E}[B_i]t}{r} = \beta_i m_{C,i}t, \tag{5}$$

where the last equality follows from that fact that $r = \beta_i(1 - \frac{\lambda_{i,r}\mathbb{E}[B_i]}{m_{C,i}})^{-1}$. By combining (3) and (4) with (5), it then follows that, as $r \to \infty$,

$$\Big\{ \Big(\frac{V_1(\lambda_{1,r}r^2 t) - C_1(r^2 t)}{r}, \ldots, \frac{V_N(\lambda_{N,r}r^2 t) - C_N(r^2 t)}{r} \Big), t \geq 0 \Big\} \stackrel{d}{\to} \{\boldsymbol{Z}(t), t \geq 0\}, \tag{6}$$

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

$$\boldsymbol{\Gamma} = diag(\frac{m_{C,1}}{\mathbb{E}[B_1]}\sigma^2_{V,1}, \ldots, \frac{m_{C,N}}{\mathbb{E}[B_N]}\sigma^2_{V,N}) + \boldsymbol{\Gamma}^C. \tag{7}$$

For the sake of notational convenience, we write

$$\overline{\boldsymbol{Z}} = (\sup_{t \geq 0}\{Z_1(t)\}, \ldots, \sup_{t \geq 0}\{Z_N(t)\}), \tag{8}$$

and we denote its i-th element by $\overline{Z}_i$. It is tempting to conclude from a combination of (2) and (6) that $\widetilde{\boldsymbol{W}}_r$ converges to $\overline{\boldsymbol{Z}}$ in distribution as $r \to \infty$ by use of a continuous mapping argument. However, complications arise since the supremum applied to càdlàg functions on the infinite domain $[0, \infty)$ is not necessarily a continuous functional. To overcome this, we have to justify the interchange of the heavy-traffic and the steady-state limits. This forms the main result of the next section.

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*

$$\widetilde{\boldsymbol{W}}_r \xrightarrow{d} \overline{\boldsymbol{Z}},$$

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

In order to prove this theorem, we need some auxiliary results. As mentioned before, Theorem 3.1 cannot be proved directly by the use of the continuous mapping theorem, as the supremum of càdlàg functions on an infinite domain $[0, \infty)$ is not necessarily a continuous functional. However, it is continuous in case of a finite domain $[0, M)$, $M \in \mathbb{R}_+$; see e.g. [48]. The proof uses this fact in combination with an additional result stated in Lemma 3.4. To prove Lemma 3.4, we start with two auxiliary results in Lemmas 3.2 and 3.3 that establish upper bounds for the tail probabilities

$$\mathbb{P}(\sup_{t\in[0,T)} \{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t\} \geq x) \text{ and } \mathbb{P}(\sup_{t\in[0,T)} \{\mathbb{E}[C_i(1)]t - C_i(t)\} \geq x)$$

respectively, for any $i \in \{1, \dots, N\}$ and $r, x, T \in \mathbb{R}_+$.

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

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

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

Proof. As the process $\{V_i(\lambda_1 t) - \mathbb{E}[V_i(\lambda_{i,r})]t, t \geq 0\}$ is a right-continuous martingale, we have that

$$\begin{aligned}
\mathbb{P}(\sup_{t\in[0,T)} \{V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t\} \geq x) &\leq \mathbb{P}(\sup_{t\in[0,T)} \{|V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t|\} \geq x) \\
&\leq \frac{\sup_{t\in[0,T)}\{\mathbb{E}[(V_i(\lambda_{i,r}t) - \mathbb{E}[V_i(\lambda_{i,r})]t)^2]\}}{x^2} \\
&= \frac{\sup_{t\in[0,T)}\{\text{Var}[V_i(\lambda_{i,r}t)]\}}{x^2},
\end{aligned}$$

where the second inequality follows from Doob's inequality (cf. [40, Theorem II.1.7]). Since $\text{Var}[V_i(\lambda_{i,r}t)] = \lambda_{i,r}\sigma_{V,i}^2 t$ is strictly increasing in t, the lemma follows. □

Lemma 3.3. *For the service speed 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*

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

Proof. The lemma is a consequence of Proposition 1 in [28]. To apply this proposition, define $h = \max_{\omega\in\boldsymbol{\mathcal{S}}} \phi_i(\omega)$, $H(t) = ht - C_i(t)$ and $b = \mathbb{E}[H(1)] = h - \mathbb{E}[C_i(1)]$, so that $\mathbb{P}(\sup_{t\in[0,T)}\{\mathbb{E}[C_i(1)]t - C_i(t)\} > x) = \mathbb{P}(\sup_{t\in[0,T)}\{H(t) - bt\} > x)$. Note that $\{H(t), t \geq 0\}$ represents increments of the regenerative process $\{h - \phi_i(\Phi(t)), t \geq 0\}$. This process 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 $\nu_1, \nu_2, \dots$ 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 [28] 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

$$\mathbb{P}(\sup_{t\in[0,T)} \{\mathbb{E}[C_i(1)]t - C_i(t)\} > x) \leq d_1(e^{-d_2\frac{x^2}{T}} + e^{-d_3 T} + Te^{-d_4\sqrt{x}}), \tag{9}$$

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 and $Q(x)$ in [28, Proposition 1] by $H(t)$ as defined above and $\sqrt{x}$ respectively. The lemma is a consequence from (9) by noting that $e^{-x} < x^{-1}$ for all $x > 0$ and taking $c_1 = d_1 d_2^{-1}, c_2 = d_1 d_3^{-1}, c_3 = d_1$ and $c_4 = d_4$, if the necessary conditions mentioned hold. To show that this is the 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 $\epsilon > 0$. Similarly, as $\gamma_n^* \leq h\nu_n$ for any $n > 0$, we have that $\mathbb{E}[e^{\sqrt{\gamma_n^*}}] \leq \mathbb{E}[e^{\sqrt{h\nu_n}}] = \mathbb{E}[e^{\sqrt{hY}}] < \mathbb{E}[e^{\epsilon hY+\epsilon^{-1}}] = e^{\epsilon^{-1}}\mathbb{E}[e^{\epsilon hY}]$ for all $n \in \mathbb{N}$ and any $\epsilon > 0$.

It is thus left to show that there exists a value $\epsilon > 0$ for which $\mathbb{E}[e^{\epsilon Y}] < \infty$. For this purpose, note that the regeneration time Y constitutes the return time of state ω in the Markov chain $\{\Phi(t), t \geq 0\}$. Thus, Y can be decomposed into the period of time between the entry into state ω at the start of the regeneration period and the subsequent departure from state ω, which we denote by Y_1, and the period of time between this departure and the next entry into state s, which we denote by Y_2. The former period Y_1 is exponentially distributed with a rate α that equals the total outgoing rate of state ω in the Markov process $\{\Phi(t), t \geq 0\}$, so that $\mathbb{E}[e^{\epsilon Y_1}] = \frac{\alpha}{\alpha-\epsilon}$ for $\epsilon < \alpha$. The latter period Y_2 is easily seen to be stochastically smaller than a geometrically distributed random variable, denoted by G, with success parameter $q = \min_{\omega'\in\boldsymbol{\mathcal{S}}\setminus\omega} \mathbb{P}(\Phi(t+1) = \omega \mid \Phi(t) = \omega')$, $t > 0$. As the Markov process $\{\Phi(t), t \geq 0\}$ is irreducible and has a finite state space, q must be positive. Therefore, $\mathbb{E}[e^{\epsilon Y_2}] \leq \mathbb{E}[e^{\epsilon G}] = \frac{qe^{\epsilon}}{1-(1-q)e^{\epsilon}}$ for $\epsilon < -\log(1-q)$. Summarising, as Y_1 and Y_2 are mutually independent, we have that

$$\mathbb{E}[e^{\epsilon Y}] = \mathbb{E}[e^{\epsilon Y_1}]\mathbb{E}[e^{\epsilon Y_2}] \leq \frac{\alpha}{\alpha-\epsilon}\frac{qe^{\epsilon}}{1-(1-q)e^{\epsilon}} < \infty$$

for $0 < \epsilon < \min\{\alpha, -\log(1-q)\}$. This concludes the proof. □

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

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

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

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

Proof. The first part of the proof is inspired by the proof of (20) in [41]. 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

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

As t runs over $[2^jr^2M, 2^{j+1}r^2M]$ in the last expression, we have that the negative terms $-\frac{1}{2}\beta_i m_{C,i}r^{-1}t$ have a value of at most $-\frac{1}{2}\beta_i m_{C,i}r^{-1}2^jr^2M = -2^{j-1}\beta_i m_{C,i}rM$. Replacing the negative terms by these upper bounds, moving them to the right-hand sides of the inequalities, and consequently enlarging the intervals of the suprema to also include $[0, 2^j r^2 M)$, we obtain

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

for certain positive constants c_1, c_2, c_3 and c_4. The last inequality follows from Lemmas 3.2 and 3.3. Simplifying this expression leads to

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

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 trivially from (10) by taking the limit $r\to\infty$ and subsequently the limit $M\to\infty$, if $\lim_{r\to\infty}\sum_{j=0}^{\infty}f_{i,j}(r,M)=0$.

We now show that this condition holds. 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 $h_{i,j}(M) := c_4\sqrt{2^{j-1}\beta_i m_{C,i}M}$. Note that $\frac{\partial}{\partial r}f_{i,j}(r,M)$ is negative if and only if $4-h_{i,j}(M)\sqrt{r}$ is negative. Due to the monotonicity of $h_{i,j}(M)$ and $\sqrt{r}$ in j and r respectively, there exist positive constants j_0 and r_0, such that for any $j\geq j_0$ and $r\geq r_0$ the latter statement holds true. Thus, there exist positive constants j_0 and r_0, so that $\sup_{r\geq r_*}f_{i,j}(r,M)=f_{i,j}(r_*,M)$ for every $r_*\geq r_0$. This leads to an upper bound for $\sum_{j=0}^{\infty}f_{i,j}(r,M)$ when $r\geq r_*$:

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

In the limiting case of $r\to\infty$, we can apply (11) with r_* taken arbitrarily large so that the condition $r_0\leq r_*\leq r$ still holds. By doing this, we obtain

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

Combining this inequality with the fact that $\lim_{r\to\infty}f_{i,j}(r,M)=0$ results in $\lim_{r\to\infty}\sum_{j=0}^{\infty}f_{i,j}(r,M)\leq 0$. We also trivially have that $\lim_{r\to\infty}\sum_{j=0}^{\infty}f_{i,j}(r,M)\geq 0$ due to the non-negativity of $f_{i,j}(r,M)$ for any $i\in\{1,\ldots,n\}$, $j\in\mathbb{N}_+$ and $r,M\in\mathbb{R}_+$. This results in the fact 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. Using the representation of the distribution of $\widetilde{\boldsymbol{W}}_r$ given in (2), it is clear that it is enough to show that the tail probability of the right-hand side of (2) in the heavy-traffic limit $r\to\infty$ coincides with the tail probability of $\overline{\boldsymbol{Z}}$, i.e.:

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

for all $x_1, \dots, x_N > 0$. First, we obtain a lower bound for the left-hand side of (12):

$$\lim_{r\to\infty} \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i \Big\}) \geq \lim_{r\to\infty} \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\in[0,M)} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i \Big\}) = \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\in[0,M)} \{Z_i(t)\} \geq x_i \Big\}) \quad (13)$$

for all $M \in \mathbb{R}_+$, where the inequality follows from the monotonicity property of the supremum functional, and the equality follows from (6) 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 $[0, M)$.

Second, we derive an upper bound for the left-hand side of (12). Denote by $E_{M,i}$ the event that

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

or colloquially speaking, the event that the scaled net-input process of Q_i attains its largest value before time $t = M$. Furthermore, we denote by $E^c_{M,i}$ its complementary event. By De Morgan's law, we have that

$$\mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i \Big\}) = \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i; E_{M,i} \Big\}) + \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i \Big\}; \bigcup_{i=1}^{N} E^c_{M,i}). \quad (14)$$

An upper bound for the first term of the right-hand side in (14) is given by

$$\mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i; E_{M,i} \Big\}) \leq \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\in[0,M)} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i \Big\}) \quad (15)$$

for all $M \in \mathbb{R}_+$. For the second term of the right-hand side in (14), we have that

$$\mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i \Big\}; \bigcup_{i=1}^{N} E^c_{M,i}) \leq \sum_{i=1}^{N} \mathbb{P}(\sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i; E^c_{M,i}) \leq \sum_{i=1}^{N} \mathbb{P}(\sup_{t\geq M} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i), \quad (16)$$

for all $M \in \mathbb{R}_+$. Thus, by combining (14)–(16) and taking the limit $r \to \infty$, we obtain the following upper bound for the right-hand side of (14):

$$\lim_{r\to\infty} \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\geq 0} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{r} \Big\} \geq x_i \Big\}) \leq \mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\in[0,M)} \{Z_i(t)\} \geq x_i \Big\}) + \sum_{i=1}^{N} \mathbb{P}(\sup_{t\geq M} \Big\{ \frac{V_i(\lambda_{i,r} r^2 t) - C_i(r^2 t)}{R} \Big\} \geq x_i). \quad (17)$$

When taking the limit $M \to \infty$, we have that the lower bound for the left-hand side of (12) established in (13) converges to $\mathbb{P}(\bigcap_{i=1}^{N} \Big\{ \sup_{t\in[0,\infty)} \{Z_i(t)\} \geq x_i \Big\})$. The upper bound found in (17) also converges to this expression, as the second term in the right-hand side of (17) vanishes due to Lemma 3.4. From this, (12) immediately follows, which proves the theorem. □

Remark 3.1. The joint distribution of the vector $\overline{\boldsymbol{Z}}$ is not straightforward to derive explicitly. As a result, it is hard to give an explicit characterisation of the distribution of the joint workload vector in heavy traffic. However, explicit expressions for the marginal distribution of $\overline{Z}_i$ are easier 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 $r\to\infty$. We will study the derivation of the joint distribution of $\widetilde{\boldsymbol{W}}_r$ as $r\to\infty$ in Section 5.3.

4 Extension to virtual waiting times and queue lengths

In Section 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 regarding 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 Section 4.1 and Section 4.2 respectively. It turns out that, when $r\to\infty$, both $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{L}}_r$ are elementwise equal to $\widetilde{\boldsymbol{W}}_R$ 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, using the simple fact that the event $\{D_{i,r}(u)>s_i\}$ is tantamount to the event $\{W_{i,r}(u)>C_i(s_i)-C_i(u)\}$, as explained below. 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*

$$\begin{aligned}&\lim_{r\to\infty}\mathbb{P}(\widetilde{D}_{1,r}\geq s_1,\dots,\widetilde{D}_{N,r}\geq s_N,\widetilde{W}_{1,r}\geq t_1,\dots,\widetilde{W}_{N,r}\geq t_N)\\&\qquad=\mathbb{P}(\overline{Z}_1\geq\max\{m_{C,1}s_1,t_1\},\dots,\overline{Z}_N\geq\max\{m_{C,N}s_N,t_N\})\end{aligned}$$

with $\overline{Z}_1,\dots,\overline{Z}_N$ defined in Section 2.

Proof. To derive this result, we first study the relation between $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{W}}_r$. If 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 $C_i(s_i)-C_i(u)$. This is evident, since the latter number represents the amount of work the server of Q_i is able to process in the s_i time units following time u. In other words, $\{D_{i,r}(u)>s_i\}$ is tantamount to the event $\{W_{i,r}(u)>C_i(s_i)-C_i(u)\}$ for $i=1,\dots,N$. In terms of tail probabilities, this leads to

$$\begin{aligned}&\mathbb{P}(D_{1,r}(u)>s_1,\dots,D_{N,r}(u)>s_N,W_{1,r}(u)>t_1,\dots,W_{N,r}(u)>t_N)\\&=\mathbb{P}(W_{1,r}(u)>\max\{C_1(s_1)-C_1(u),t_1\},\dots,W_{N,r}(u)>\max\{C_N(s_N)-C_N(u),t_N\}).\end{aligned}$$

Thus, in steady state (i.e., $u\to\infty$), we have

$$\begin{aligned}&\mathbb{P}(D_{1,r}>s_1,\dots,D_{N,r}>s_N,W_{1,r}>t_1,\dots,W_{N,r}>t_N)\\&=\mathbb{P}(W_{1,r}>\max\{C_1(s_1),t_1\},\dots,W_{N,r}>\max\{C_N(s_N),t_N\}).\end{aligned}\tag{18}$$

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

$$\begin{aligned}&\mathbb{P}(\widetilde{D}_{1,r}\geq s_1,\dots,\widetilde{D}_{N,r}\geq s_N,\widetilde{W}_{1,r}\geq t_1,\dots,\widetilde{W}_{N,r}\geq t_N)\\&=\mathbb{P}(D_{1,r}\geq rs_1,\dots,D_{N,r}\geq rs_N,W_{1,r}\geq rt_1,\dots,W_{N,r}\geq rt_N)\\&=\mathbb{P}(W_{1,r}\geq\max\{C_1(rs_1),rt_1\}\dots,W_{N,r}\geq\max\{C_N(rs_N),rt_N\})\\&=\mathbb{P}(\widetilde{W}_{1,r}\geq\max\Big\{\frac{C_1(rs_1)}{r},t_1\Big\},\dots,\widetilde{W}_{N,r}\geq\max\Big\{\frac{C_N(rs_N)}{r},t_N\Big\}),\end{aligned}\tag{19}$$

where we used (18) in the second equality.

In the remainder of the proof, we focus on showing that

$$\lim_{r\to\infty} \mathbb{P}(\widetilde{W}_{1,r} \geq \max\Big\{\frac{C_1(rs_1)}{r}, t_1\Big\}, \ldots, \widetilde{W}_{N,r} \geq \max\Big\{\frac{C_N(rs_N)}{r}, t_N\Big\}) \\ = \mathbb{P}(\overline{Z}_1 \geq \max\{m_{C,1}s_1, t_1\}, \ldots, \overline{Z}_N \geq \max\{m_{C,N}s_N, t_N\}), \tag{20}$$

which combined with (19) directly implies the result to be proved. To this end, we observe that by viewing $\{C_i(t), t \geq 0\}$ as a renewal-reward process with the times where $\{\Phi(t), t \geq 0\}$ enters a certain reference state as renewal epochs, we have that $r^{-1}C_i(rs_i) \to m_{C,i}s_i$ 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 $\frac{1}{r}C_i(rs_i) \in [m_{C,i}s_i - \epsilon, m_{C,i}s_i + \epsilon]$ and let $F^{\epsilon,c}_{i,r}$ be its complementary event. Thus, $\lim_{r\to\infty} \mathbb{P}(F^{\epsilon}_{i,r}) = 1$. Similarly to the proof of Theorem 3.1, we now partition all combinations of events into $\bigcap_{i=1}^N F^{\epsilon}_{i,r}$, the case where each of the events $F^{\epsilon}_{1,r}, \ldots, F^{\epsilon}_{N,r}$ holds true, and $\bigcup_{i=1}^N F^{\epsilon,c}_{i,r}$, the case where at least one of these events does not hold true. Then, we have as a result of De Morgan's law that

$$\mathbb{P}(\widetilde{W}_{1,r} \geq \max\Big\{\frac{C_1(rs_1)}{r}, t_1\Big\}, \ldots, \widetilde{W}_{N,r} \geq \max\Big\{\frac{C_N(rs_N)}{r}, t_N\Big\}) \\ = \mathbb{P}(\widetilde{W}_{1,r} \geq \max\Big\{\frac{C_1(rs_1)}{r}, t_1\Big\}, \ldots, \widetilde{W}_{N,r} \geq \max\Big\{\frac{C_N(rs_N)}{r}, t_N\Big\}; \bigcap_{i=1}^N F^{\epsilon}_{i,r}) + o(1).$$

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 (20):

$$\lim_{r\to\infty} \mathbb{P}(\widetilde{W}_{1,r} \geq \max\Big\{\frac{C_1(rs_1)}{r}, t_1\Big\}, \ldots, \widetilde{W}_{N,r} \geq \max\Big\{\frac{C_N(rs_N)}{r}, t_1\Big\}) \\ \geq \mathbb{P}(\overline{Z}_1 \geq \max\{m_{C,1}s_1 + \epsilon, t_1\}, \ldots, \overline{Z}_N \geq \max\{m_{C,N}s_N + \epsilon, t_N\}). \tag{21}$$

Similarly, an upper bound for the left-hand side of (20) is given by

$$\lim_{r\to\infty} \mathbb{P}(\widetilde{W}_{1,r} \geq \max\Big\{\frac{C_1(rs_1)}{r}, t_1\Big\}, \ldots, \widetilde{W}_{N,r} \geq \max\Big\{\frac{C_N(rs_N)}{r}, t_1\Big\}) \\ \leq \mathbb{P}(\overline{Z}_1 \geq \max\{m_{C,1}s_1 - \epsilon, t_1\}, \ldots, \overline{Z}_N \geq \max\{m_{C,N}s_N - \epsilon, t_N\}). \tag{22}$$

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 points. In particular, there is 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 (21) and (22), we obtain (20), 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*

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

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

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

Remark 4.1. Similar to the observations of Remark 3.1, explicit expressions for the joint distribution of $\widetilde{\boldsymbol{D}}_r$ as $r \to \infty$ are hard to derive. However, again an explicit characterisation for the marginal distribution of the scaled virtual waiting time in a single queue as $r \to \infty$ is easier to obtain. By Theorem 3.1 and Corollary 4.2, the heavy-traffic distributions of $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{W}}_r$ only differ elementwise by the multiplicative factors $\frac{1}{m_{C,i}}$, $i = 1, \ldots, N$. Due to this, it follows from Remark 3.1 that the distribution of $\widetilde{D}_{i,r}$ converges to an exponential distribution with rate $2\beta_i\left(\frac{m_{C,i}\sigma^2_{V,i}}{\mathbb{E}[B_i]} + \sigma^2_{C,i}\right)^{-1}$ as $r \to \infty$ for $i = 1, \ldots, N$. We will study the derivation of the joint distribution of $\widetilde{\boldsymbol{D}}_r$ as $r \to \infty$ in Section 5.3.

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 $B_{i,r}^R$ 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

$$\boldsymbol{W}_r = (B_{1,r}^R, \ldots, B_{N,r}^R) + \Big(\sum_{j=1}^{L_{1,r}} \widehat{B}_{1,j}, \ldots, \sum_{j=1}^{L_{N,r}} \widehat{B}_{N,j} \Big) \tag{23}$$

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 Section 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 (23) is given by

$$\widetilde{\boldsymbol{W}}_r = (\widetilde{B}_{1,r}^R, \ldots, \widetilde{B}_{N,r}^R) + \frac{1}{r}\Big(\sum_{j=1}^{r\widetilde{L}_{1,r}} \widehat{B}_{1,j}, \ldots, \sum_{j=1}^{r\widetilde{L}_{N,r}} \widehat{B}_{N,j} \Big), \tag{24}$$

where $\widetilde{B}_{i,r}^R = \frac{1}{r} B_{i,r}^R$ for $i = 1, \ldots, N$. It is intuitively tempting to conclude that $(\widetilde{B}_{1,r}^R, \ldots, \widetilde{B}_{N,r}^R) \to \mathbf{0}$ 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 $(\widetilde{B}_{1,r}^R, \ldots, \widetilde{B}_{N,r}^R)$ are not independent. We make these results rigorous in this section. Inspired by [50, 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 $(\widetilde{L}_{i,r}, \widetilde{W}_{i,r}, \widetilde{B}_{i,r}^R)$ in Lemma 4.4, which imply that $\widetilde{B}_{i,r}^R \to 0$ as $r \to \infty$. We subsequently conclude that $(\widetilde{B}_{1,r}^R, \ldots, \widetilde{B}_{N,r}^R) \to \mathbf{0}$ 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 $W_{i,n}^r$ and $L_{i,n}^r$ 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, $A_{i,j}^r$ refers to the time between the j-th and the $j+1$-st arriving type-i customer in the r-th system, so that $S_{i,n}^{A,r} = \sum_{j=1}^n A_{i,j}^r$ and $S_{i,n}^B = \sum_{j=1}^n B_{i,j}$ represent the cumulative series of interarrival times and service requirements of type-i customers. By construction of the heavy-traffic scaling, $A_{i,j}^r \xrightarrow{d} A_{i,j}$ and $\mathbb{E}[A_{i,j}^r] \to \mathbb{E}[A_{i,j}]$ as $r \to \infty$, where $A_{i,j}$ are i.i.d. samples from an exponential $\left(\frac{m_{C,i}}{\mathbb{E}[B_i]}\right)$ distribution. Finally, we define $S_{i,n}^r = S_{i,n}^B - C_i(S_{i,n}^{A,r})$. The needed 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*

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

Proof. The proof is inspired by [50, Proposition 1]. Observe that, for any $k \geq 1$ and $n \geq 1$, the event $\{L_{i,n+k}^r \geq k\}$ 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, $C_i(S_{i,n+k-1}^{A,r}) - C_i(S_{i,n-1}^{A,r})$, does not exceed the sum $W_{i,n}^r + B_{i,n}$ of the workload present in Q_i just before the arrival of the n-th customer and the service requirement of this customer. Hence, we have that

$$\{L_{i,n+k}^r \geq k\} = \{W_{i,n}^r + B_{i,n} \geq C_i(S_{i,n+k-1}^{A,r}) - C_i(S_{i,n-1}^{A,r})\}. \tag{25}$$

Moreover, due to Lindley's recursion $W_{i,n+1}^r = \max\{W_{i,n}^r + S_{i,n}^r - S_{i,n-1}^r, 0\}$, or $W_{i,n+k}^r = \max\{W_{i,n}^r + S_{i,n+k-1}^r - S_{i,n-1}^r, \max_{j \in \{0,\ldots,k-1\}} \{S_{i,n+k-1} - S_{i,n+j}\}\}$, we have for the event $\{W_{n+k}^r \geq y\}$ for any $y \geq 0$ that

$$\{W_{n+k}^r \geq y\} = \Big\{ \max\Big\{ W_{i,n}^r + S_{i,n+k-1}^r - S_{i,n-1}^r, \max_{j \in \{0,\ldots,k-1\}} \{S_{i,n+k-1}^r - S_{i,n+j}^r\} \Big\} \geq y \Big\}. \tag{26}$$

By combining (25) and (26), taking probabilities, 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

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

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 $i \in \{1, \dots, N\}$, with $\overline{Z}_i$ defined in Section 2.

Proof. Again, the proof is inspired by [50, Proposition 1]. We first focus on the joint distribution of $\widetilde{L}_{i,r}$ and $\widetilde{W}_{i,r}$. Note that 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, we have already seen in the proof of Proposition 4.1 that $t^{-1}C_i(t) \to m_{C,i}$ almost surely as $t \to \infty$. Consequently, we have that

$$\frac{C_i(S^{A,r}_{i,\lceil rx \rceil})}{r} = \frac{C_i(S^{A,r}_{i,\lceil rx \rceil})}{S^{A,r}_{i,\lceil rx \rceil}} \frac{S^{A,r}_{i,\lceil rx \rceil}}{r} \to \mathbb{E}[B_i]x \tag{27}$$

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 $r^{-1}\max_{j\in\{1,\dots,\lceil rx \rceil\}}\{S^r_{i,\lceil rx \rceil} - S^r_{i,j}\} \to 0$ as $r \to \infty$. Let, for any $\epsilon > 0$, $G^\epsilon_{i,R}$ denote the event

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

Due to the convergence results above, $\lim_{r\to\infty} \mathbb{P}(G^\epsilon_{i,r}) = 1$, so that, because of the law of total probability,

$$\mathbb{P}(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y) = \mathbb{P}(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y; G^\epsilon_{i,r}) + o(1).$$

A combination with Lemma 4.3 leads by taking the limit $r \to \infty$ to, since $\widetilde{B}_i \to 0$ as $r \to \infty$,

$$\begin{aligned}&\lim_{r\to\infty} \mathbb{P}(\widetilde{W}_{i,r} \geq \max\{\mathbb{E}[B_i]x + \epsilon, y + \epsilon\})\\ &\quad\leq \lim_{r\to\infty} \mathbb{P}(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y) \leq \lim_{r\to\infty} \mathbb{P}(\widetilde{W}_{i,r} \geq \max\{\mathbb{E}[B_i]x - \epsilon, y - \epsilon\}).\end{aligned}$$

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

$$\lim_{r\to\infty} \mathbb{P}(\widetilde{L}_{i,r} \geq x; \widetilde{W}_{i,r} \geq y) = \mathbb{P}(\overline{Z}_i \geq \max\{\mathbb{E}[B_i]x, y\}). \tag{28}$$

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 (24), we have that

$$\mathbb{P}(\widetilde{B}^R_{i,r} > \delta) = \mathbb{P}(\widetilde{W}_{i,r} > \frac{1}{r}\sum_{j=1}^{r\widetilde{L}_{i,r}} \widehat{B}_{i,j} + \delta). \tag{29}$$

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

$$\mathbb{P}(\widetilde{B}^R_{i,r} > \delta) = \mathbb{P}(\widetilde{W}_{i,r} > \frac{1}{r}\sum_{j=1}^{r\widetilde{L}_{i,r}}\widehat{B}_{i,j} + \delta; H^{\epsilon}_{i,r}) + o(1) = \mathbb{P}(\widetilde{W}_{i,r} > \widetilde{L}_{i,r}\frac{1}{r\widetilde{L}_{i,r}}\sum_{j=1}^{r\widetilde{L}_{i,r}}\widehat{B}_{i,j} + \delta; H^{\epsilon}_{i,r}) + o(1).$$

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

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

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

$$\lim_{r\to\infty}\mathbb{P}(\widetilde{B}^R_{i,r} > \delta) = \lim_{r\to\infty}\mathbb{P}(\widetilde{W}_{i,r} > \widetilde{L}_{i,r}\mathbb{E}[B_i]+\delta) = 0,$$

where the second equality follows from (28) 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*

$$\begin{aligned}&\lim_{r\to\infty}\mathbb{P}(\widetilde{L}_{1,r} \geq s_1, \ldots, \widetilde{L}_{N,r} \geq s_N, \widetilde{W}_{1,r} \geq t_1, \ldots, \widetilde{W}_{N,r} \geq t_N)\\ &= \mathbb{P}(\overline{Z}_1 \geq \min\{\mathbb{E}[B_1]s_1, t_1\}, \ldots, \overline{Z}_N \geq \min\{\mathbb{E}[B_N]s_N, t_N\})\end{aligned} \tag{30}$$

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

Proof. Equation (24) 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

$$\begin{aligned}&\mathbb{P}(\widetilde{L}_{1,r} \geq s_1, \ldots, \widetilde{L}_{N,r} \geq s_N, \widetilde{W}_{1,r} \geq t_1, \ldots, \widetilde{W}_{N,r} \geq t_N)\\ &= \mathbb{P}(\widetilde{W}_{1,r} \geq \max\{\widetilde{B}^R_{1,r} + \frac{1}{r}\sum_{j=1}^{rs_1}\widehat{B}_{1,j}, t_1\}, \ldots, \widetilde{W}_{N,r} \geq \max\{\widetilde{B}^R_{N,r} + \frac{1}{r}\sum_{j=1}^{rs_N}\widehat{B}_{N,j}, t_N\}).\end{aligned}$$

Let $H^{\epsilon}_{i,r}$ be defined as in the proof of Lemma 4.4. 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,

$$\begin{aligned}&\mathbb{P}(\widetilde{L}_{1,r} \geq s_1, \ldots, \widetilde{L}_{N,r} \geq s_N, \widetilde{W}_{1,r} \geq t_1, \ldots, \widetilde{W}_{N,r} \geq t_N)\\ &= \mathbb{P}(\widetilde{W}_{1,r} \geq \max\{\widetilde{B}^R_{1,r} + s_1\frac{1}{rs_1}\sum_{j=1}^{rs_1}\widehat{B}_{1,j}, t_1\}, \ldots, \widetilde{W}_{N,r} \geq \max\{\widetilde{B}^R_{N,r} + s_N\frac{1}{rs_N}\sum_{j=1}^{rs_N}\widehat{B}_{N,j}, t_N\}; \bigcap_{i=1}^{N}H^{\epsilon}_{i,r})\\ &\quad + o(1).\end{aligned}$$

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 $r \to \infty$. Letting $r \to \infty$ and exploiting the definition of $H^{\epsilon}_{i,r}$, we obtain

$$\begin{aligned}&\lim_{r\to\infty}\mathbb{P}(\widetilde{W}_{1,r} \geq \max\{\mathbb{E}[B_i]+\epsilon, t_1\}, \ldots, \widetilde{W}_{N,r} \geq \max\{\mathbb{E}[B_N]+\epsilon, t_N\})\\ &\leq \lim_{r\to\infty}\mathbb{P}(\widetilde{L}_{1,r} \geq s_1, \ldots, \widetilde{L}_{N,r} \geq s_N, \widetilde{W}_{1,r} \geq t_1, \ldots, \widetilde{W}_{N,r} \geq t_N)\\ &\leq \lim_{r\to\infty}\mathbb{P}(\widetilde{W}_{1,r} \geq \max\{\mathbb{E}[B_1]-\epsilon, t_1\}, \ldots, \widetilde{W}_{N,r} \geq \max\{\mathbb{E}[B_N]-\epsilon, t_N\}).\end{aligned}$$

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 yields the desired result. □

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

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

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

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

Remark 4.2. In line with the observations in Remarks 3.1 and 4.1, Corollary 4.6 does not straightforwardly lead to explicit expressions for the limiting joint distribution of $\widetilde{\boldsymbol{L}}_r$. However, explicit expressions for the limiting marginal distribution of the scaled steady-state queue length of a single queue are available. Note that Lemma 4.4 implies that, for $i = 1, \dots, N$, $\widetilde{L}_{i,r}$ and $\widetilde{W}_{i,r}$ only differ elementwise up to a multiplicative constant $\frac{1}{\mathbb{E}[B_i]}$ as $r \to \infty$. It then follows immediately from the findings in Remark 3.1 that the distribution of $\widetilde{L}_{i,r}$ converges to an exponential distribution with rate $2\beta_i \mathbb{E}[B_i] \left(\frac{\sigma^2_{V,i}}{\mathbb{E}[B_i]} + \frac{\sigma^2_{C,i}}{m_{C,i}}\right)^{-1}$ as $r \to \infty$. Note that this result can also be found by an application of the distributional form of Little's law (cf. [29]) on the distribution found for $\boldsymbol{D}_{i,r}$ in Remark 4.1. We will study the derivation of the joint distribution of $\widetilde{\boldsymbol{L}}_r$ as $r \to \infty$ in Section 5.3.

5 Application to a two-layered network

In this section, we apply the results obtained so far in this paper to the manufacturing example of the LQN mentioned in Section 1. As this particular LQN consists of two layers, we will also refer to this example as the two-layered network. We first describe this two-layered network in more detail in Section 5.1 and show that this particular model fits naturally in the general framework described in Section 2. Then, in Section 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 Markov additive processes. Finally, we 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 Section 5.3.

5.1 Description of the two-layered network

The two-layered network is an extension of the machine-repair model and consists of N machines $M_1, \dots, M_N$ as well as a single repairman R, see Figure 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 requires a stochastic amount of repair time from the repairman. For this purpose, it moves to the repair buffer, where it will wait if the repairman is busy repairing, otherwise repair will start instantly. Note that each machine can have its own lifetime and repair-time distribution. 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. The products arriving at Q_i do so according to a Poisson (λ_i) process, and their individual service requirement B_i is generally distributed with finite first two moments $\mathbb{E}[B_i]$ and $\mathbb{E}[B_i^2]$. The products are served by their machine in the order of arrival. This forms the first layer of the layered network. Observe that the downtimes of the machines are mutually correlated, since the machines compete with each other for repair facilities in the second layer. Due to this correlation, exact analysis for the queue lengths of arbitrarily loaded queues in the first layer is difficult.

The two-layered network fits the general model given in Section 2, provided that the lifetimes and repair times of each machine follow a phase-type distribution. The equivalence between the first layer of the two-layered network and the parallel single-server queues in the general model is immediate. To also fit the second layer in the general framework, observe that the availability of the machines can be modelled naturally as a continuous-time Markov chain, due to the phase-type nature of lifetimes and repair times. To reduce complexity of upcoming calculations, we assume for the remainder of Section 5 that $N = 2$ and that the lifetime and repair-time distributions of machine M_i are exponentially distributed with rate σ_i and ν_i respectively. In this case, the state of the machines M_1 and M_2 is modelled by the continuous-time Markov chain $\{\Phi(t), t \geq 0\}$ operating on the state space $\boldsymbol{\mathcal{S}} = \{(U,U), (U,R), (R,U), (W,R), (R,W)\}$. A state $\boldsymbol{\omega} = (\omega_1, \omega_2) \in \boldsymbol{\mathcal{S}}$ represents for each

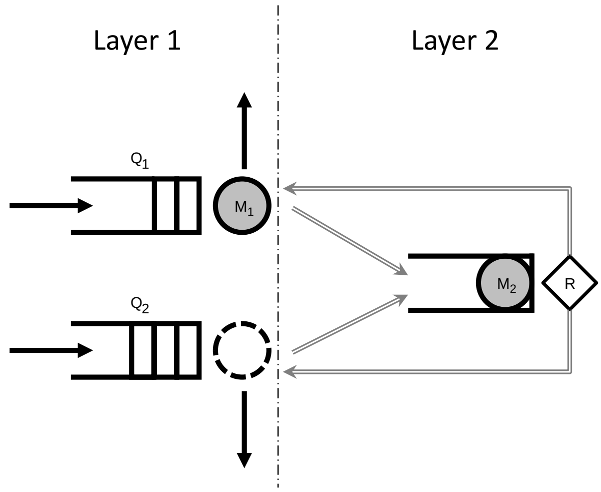

Figure 1: The two-layered model under consideration.

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 $q_{i,j}$, $i, j \in \boldsymbol{\mathcal{S}}$ is given by

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

The continuous-time Markov chain $\{\Phi(t), t \geq 0\}$ is irreducible and aperiodic, so that its invariant probability measure $\boldsymbol{\pi}$ is uniquely determined by the equations $\boldsymbol{\pi Q} = 0$ and $\boldsymbol{\pi 1} = 1$ 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 Section 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 Section 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's parameters. We do this based on the example of the two-layered network described in Section 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 Section 2 without any conceptual complications. By (7), it remains to compute expressions for the covariance terms $\gamma_{i,j}^C = \lim_{t\to\infty} \frac{1}{t}\text{Cov}[C_i(t), C_j(t)]$ 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. A functional-central limit theorem for MAPs obtained in [42] leads to expressions for $\sigma_{C,i}^2$, $\sigma_{C,j}^2$ and $\lim_{t\to\infty} \frac{1}{t}\text{Var}[C_i(t) + C_j(t)]$, i.e., the variance parameters of the limits of the scaled Markov additive processes. From these variance parameters, expressions for $\gamma_{i,j}^C$ immediately follow.

To state the results of [42], we first introduce some preliminary notation. Let $\boldsymbol{\omega}_{\text{ref}} \in \boldsymbol{\mathcal{S}}$ be an arbitrary reference state. Furthermore, denote by τ_k the time of the k-th jump of $\{\Phi(t), t \geq 0\}$ for $k = 1, 2, \ldots$. Let $T_0 = \inf\{t > 0 : \Phi(t) = \boldsymbol{\omega}_{\text{ref}}, \Phi(t-) \neq \boldsymbol{\omega}_{\text{ref}}\}$ be the first time $\{\Phi(t), t \geq 0\}$ enters the reference state, and let $T_1, T_2, \ldots$ be the subsequent entrance times into the reference state. The instantaneous drift and variance parameters of a process $\{Y(t), t \geq 0\}$ that is modulated by $\{\Phi(t), t \geq 0\}$, are given by

$$d_i^Y = \mathbb{E}[\frac{Y(\tau_k + w) - Y(\tau_k)}{w} \mid \Phi(z) = i \text{ for } \tau_k \leq z \leq \tau_k + w]$$

and

$$v_i^Y = \mathbb{E}[\frac{(Y(\tau_k + w) - Y(\tau_k))^2 - (d_i^Y w)^2}{w} \mid \Phi(z) = i \text{ for } \tau_k \leq z \leq \tau_k + w].$$

The vector $\boldsymbol{\varphi}^Y$ representing the second moment of Y is given by

$$\varphi_i^Y = \frac{\mathbb{E}[(Y(\tau_k) - Y(\tau_{k-1}))^2 \mid \Phi(\tau_{k-1}) = i]}{\mathbb{E}[\tau_k - \tau_{k-1} \mid \Phi(\tau_{k-1}) = i]}.$$

The matrix $\boldsymbol{M}^Y = (M_{i,j}^Y)_{i,j \in \boldsymbol{\mathcal{S}}}$ is defined to be a $|\boldsymbol{\mathcal{S}}| \times |\boldsymbol{\mathcal{S}}|$ matrix with elements $M_{i,i}^Y = M_{i,\omega_{\text{ref}}}^Y = 0$ and $M_{i,j}^Y = -\frac{q_{i,j}}{q_{i,i}} d_i^Y$ for $j \in \boldsymbol{\mathcal{S}} \backslash \{\{i\} \cup \{\omega_{\text{ref}}\}\}$. Finally, the vector $\boldsymbol{f}^Y$ is given by $f_i^Y = \mathbb{E}[Y(T_0) - Y(0) \mid \Phi(0) = i]$. Using this additional notation, the following lemma, which is directly implied by the work of [42], holds.

Lemma 5.1. *Let $\{(\Phi(t), Y(t)), t \geq 0\}$ be a Markov additive process, where $\{Y(t), t \geq 0\}$ is the additive part modulated by the continuous time Markov chain $\{\Phi(t), t \geq 0\}$ and has an average drift of zero and no jumps. Furthermore, assume that d_i^Y and v_i^Y are well-defined for all $i \in \boldsymbol{\mathcal{S}}$. 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 $\boldsymbol{\pi}\boldsymbol{\varphi}^Y + 2\boldsymbol{\pi}\boldsymbol{M}^Y \boldsymbol{f}^Y$. In particular, we have that*

$$\lim_{t \to \infty} \frac{1}{t} \text{Var}[Y(t)] = \boldsymbol{\pi}\boldsymbol{\varphi}^Y + 2\boldsymbol{\pi}\boldsymbol{M}^Y \boldsymbol{f}^Y.$$

Proof. The convergence in distribution immediately follows from [42, 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, let $M(t) = \max_{k: T_k \leq t}\{k\}$ count the number of times the Markov chain returned to the reference state up till time t, so that $\{M(t), t \geq 0\}$ can be interpreted as a (delayed) renewal process. Then, we have that

$$\begin{aligned}
\lim_{t \to \infty} \frac{\text{Var}[Y(t)]}{t} &= \lim_{t \to \infty} \frac{\text{Var}[Y(\sum_{i=1}^{N(t)} (T_i - T_{i-1}))] + o(t)}{t} \\
&= \lim_{t \to \infty} \frac{\mathbb{E}[M(t)]\text{Var}[Y(T_1 - T_0)] + \text{Var}[M(t)]\mathbb{E}[Y(T_1 - T_0)]^2}{t} \\
&= \text{Var}[Y(T_1 - T_0)] \lim_{t \to \infty} \frac{\mathbb{E}[M(t)]}{t} \\
&= \frac{\text{Var}[Y(T_1 - T_0)]}{\mathbb{E}[T_1 - T_0]},
\end{aligned}$$

where the second equality follows from the fact that the summands of $Y(\sum_{i=1}^{N(t)} (T_i - T_{i-1})) = \sum_{i=1}^{N(t)} (Y(T_i) - Y(T_{i-1}))$ are independent and identically distributed to $Y(T_1 - T_0)$, so that $Y(\sum_{i=1}^{N(t)} (T_i - T_{i-1}))$ can be seen as a compound Poisson process. The third equality holds because the modulated process has an average drift of zero, so that $\mathbb{E}[Y(T_1 - T_0)] = 0$. The fourth equality follows from standard results on renewal theory. Section 3 in [42] shows that $\text{Var}[Y(T_1 - T_0)] = \mathbb{E}[(Y(T_1 - T_0))^2] = (\boldsymbol{\pi}\boldsymbol{\varphi}^Y + 2\boldsymbol{\pi}\boldsymbol{M}^Y \boldsymbol{f}^Y)\mathbb{E}[T_1 - T_0]$, which concludes the proof. □

We now apply this lemma to obtain the covariance matrix for the two-layered model with $N = 2$. More specifically, we compute $\sigma_{C,1}^2$, $\sigma_{C,2}^2$ and $\lim_{t \to \infty} \frac{1}{t} \text{Var}[C_1(t) + C_2(t)]$, out of which expressions for $\gamma_{1,2}^C$ will follow.

To derive an expression for $\sigma_{C,1}^2$, let $Y_1(t) = \frac{1}{t} C_1(t) - \mathbb{E}[C_1(t)] = \frac{1}{t} C_1(t) - (\pi_{(U,U)} + \pi_{(U,R)})t$. It is easily seen that the drift of $Y_1(t)$ equals $1 - (\pi_{(U,U)} + \pi_{(U,R)})$ when the modulator $\boldsymbol{\Phi}$ resides in the states (U, U) and (U, R), and $-(\pi_{(U,U)} + \pi_{(U,R)})$ otherwise. The drift vector $\boldsymbol{d}^{Y_1}$ is thus given by

$$d_i^{Y_1} = \mathbb{1}_{\{i \in \{(U,U),(U,R)\}\}} - (\pi_{(U,U)} + \pi_{(U,R)}).$$

Due to the Markov nature of the process $\{\Phi(t), t \geq 0\}$, we have that $\mathbb{E}[\tau_k - \tau_{k-1} \mid \Phi(\tau_{k-1}) = i] = -q_{i,i}$. Moreover, since Y_1 locally behaves like a pure drift process, it holds that $\mathbb{E}[(Y(\tau_k) - Y(\tau_{k-1}))^2 \mid \Phi(\tau_{k-1}) = i] = \mathbb{E}[(d_i^{Y_1})^2 (\tau_k - \tau_{k-1})^2 \mid \boldsymbol{\Phi}(\tau_{k-1}) = i] = 2\left(\frac{d_i^{Y_1}}{-q_{i,i}}\right)^2$. The vector $\boldsymbol{\varphi}^{Y_1}$ is thus given by $\varphi_i^{Y_1} = -2\left(\frac{d_i^{Y_1}}{q_{i,i}}\right)$.

When taking $\omega_{\text{ref}} = (R, W)$ as the reference state, the elements $f_i^{Y_1} = \mathbb{E}[Y(T_1) - Y(0) \mid \Phi(0) = i]$ of the vector f^{Y_1} are easily seen to satisfy the set of equations

$$f_i^{Y_1} = -\frac{d_i^{Y_1}}{q_{i,i}} - \sum_{j \in \boldsymbol{S} \setminus \{(R,W)\}} \frac{q_{i,j}}{q_{i,i}} f_i^{Y_1},$$

since $\mathbb{E}[Y(\tau_k) - Y(\tau_{k-1}) \mid \boldsymbol{\Phi}(\tau_{k-1}) = i] = -\frac{d_i^{Y_1}}{q_{i,i}}$. This system of equations leads to a unique, explicit solution for the vector $\boldsymbol{f}^{Y_1}$. The matrix M^{Y_1} pertaining to Y_1 has elements $M_{i,j}^{Y_1} = -\mathbb{1}_{\{j \notin \{i\} \cup \{(R,W)\}\}} \frac{q_{i,j}}{q_{i,i}} d_i^{Y_1}$. An application of Lemma 5.1 then leads to

$$\sigma_{C,1}^2 = \lim_{t \to \infty} \frac{1}{t} \text{Var}[C_1(t)] = \lim_{t \to \infty} \frac{1}{t} \text{Var}[Y_1(t)] = \boldsymbol{\pi} \boldsymbol{\varphi}^{Y_1} + 2\boldsymbol{\pi} \boldsymbol{M}^{Y_1} \boldsymbol{f}^{Y_1}.$$

When studying $Y_2(t) = C_2(t) - \mathbb{E}[C_2(t)] = \frac{C_2(t) - (\pi_{(U,U)} + \pi_{(R,U)})t}{t}$, an expression for $\sigma_{C,2}^2$ can be found similarly to the computations above. Alternatively, interchanging the indices of the model parameters in the expression of $\sigma_{C,1}^2$ also leads to this expression.

Finally, an expression for $\lim_{t \to \infty} \frac{1}{t} \text{Var}[C_1(t) + C_2(t)]$ can be found by considering

$$Y_{1,2}(t) = C_1(t) + C_2(t) - (\mathbb{E}[C_1(t) + C_2(t)]) = C_1(t) + C_2(t) - (2\pi_{(U,U)} + \pi_{(U,R)} + \pi_{(R,U)})t.$$

The process $\{(\Phi(t), Y_{1,2}(t)), t \geq 0\}$ is then again a MAP that satisfies the assumptions of Lemma 5.1. It is easily seen that the vector $\boldsymbol{d}^{Y_{1,2}}$ with elements $d_i^{Y_{1,2}} = \mathbb{1}_{\{i \in \{(U,U),(U,R)\}\}} + \mathbb{1}_{\{i \in \{(U,U),(R,U)\}\}} - (2\pi_{(U,U)} + \pi_{(U,R)} + \pi_{(R,U)})$ specifies the conditional drift of the modulated process $Y_{1,2}$ when the modulator Φ resides in state i. Analogous to the computations in the previous paragraph, we obtain the vectors $\boldsymbol{\varphi}^{Y_{1,2}}$ and the matrix $\boldsymbol{M}^{Y_{1,2}}$ with elements $\varphi_i^{Y_{1,2}} = -2\frac{(d_i^{Y_{1,2}})^2}{q_{i,i}}$, and $M_{i,j}^{Y_{1,2}} = -\mathbb{1}_{\{j \notin \{i\} \cup \{(R,W)\}\}} \frac{q_{i,j}}{q_{i,i}} d_i^{Y_{1,2}}$ respectively. The vector $f^{Y_{1,2}}$ is uniquely and explicitly determined by the system of equations $f_i^{Y_{1,2}} = -\frac{d_i^{Y_{1,2}}}{q_{i,i}} - \sum_{j \in \boldsymbol{S} \setminus \{(R,W)\}} \frac{q_{i,j}}{q_{i,i}} f_i^{Y_{1,2}}$ for all $i \in \boldsymbol{S}$. Applying Lemma 5.1 now yields

$$\lim_{t \to \infty} \frac{1}{t} \text{Var}[C_1(t) + C_2(t)] = \lim_{t \to \infty} \frac{1}{t} \text{Var}[Y_{1,2}(t)] = \boldsymbol{\pi} \boldsymbol{\varphi}^{Y_{1,2}} + 2\boldsymbol{\pi} \boldsymbol{M}^{Y_{1,2}} \boldsymbol{f}^{Y_{1,2}}.$$

After these preliminary computations, the covariance matrix $\boldsymbol{\Gamma}$ can be expressed explicitly in terms of the model parameters. The covariance parameters $\gamma_{1,1}^C$ and $\gamma_{2,2}^C$ are by definition equal to $\sigma_{C,1}^2$ and $\sigma_{C,2}^2$, for which we have already derived explicit expressions. As for the remaining parameters, we have that both $\gamma_{1,2}^C$ and $\gamma_{2,1}^C$ are equal to

$$\lim_{t \to \infty} \frac{1}{t} \text{Cov}[C_1(t), C_2(t)] = \frac{1}{2} \left(\lim_{t \to \infty} \frac{1}{t} \text{Var}[C_1(t) + C_2(t)] - \lim_{t \to \infty} \frac{1}{t} \text{Var}[C_1(t)] - \lim_{t \to \infty} \frac{1}{t} \text{Var}[C_2(t)] \right).$$

Since we already computed the three terms between brackets in the right-hand side, expressions for all of the covariance parameters $\gamma_{i,j}^C$ are now available in terms of the model parameters $\sigma_1, \sigma_2, \nu_1$ and ν_2. As the rest of the terms appearing in (7) were already expressed in terms of the model's parameters, the covariance matrix $\boldsymbol{\Gamma}$ 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 $\overline{\boldsymbol{Z}}$, the limiting distribution of the scaled workload $\widetilde{\boldsymbol{W}}_r$, in stationarity. We do this by viewing this distribution as the stationary distribution of an SRBM. We obtain numerical results for the example of the two-layered network. Since the limiting distributions of $\widetilde{\boldsymbol{D}}_r$ or $\widetilde{\boldsymbol{L}}_r$ equal the distribution $\overline{\boldsymbol{Z}}$ 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 $\overline{\boldsymbol{Z}}$, we observe that this distribution is the stationary distribution of the process $\{\overline{\boldsymbol{Z}}(t), t \geq 0\}$, where

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

In this expression $\boldsymbol{Z}(t)$ is the N-dimensional Brownian motion defined in Section 2, $\boldsymbol{R}$ 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 $\mathbf{0}$, of which the elements Y_i can only increase at times t when $Z_i(t) = 0$ ($i = 1, \ldots, N$). A process with such a representation is known to be an SRBM (see e.g. [9, Section 7.4]). As briefly mentioned in the introduction, such a process evolves like a Brownian motion in the interior of the positive orthant $\mathbb{R}_+^N$, but is pushed back when it reaches a boundary face $\{z \in \mathbb{R}_+^N : z_j = 0\}$ in a direction determined by the j-th column of the reflection matrix $\boldsymbol{R}$, $j = 1, \ldots, N$. The j-th element of the regulator process $\{\boldsymbol{Y}(t), t \geq 0\}$ indicates the cumulative amount of 'effort' spent in pushing back at the j-th boundary face. An SRBM is thus identified by the drift vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Gamma}$ of the underlying Brownian motion $\{\boldsymbol{Z}(t), t \geq 0\}$, together with the reflection matrix $\boldsymbol{R}$.

The stationary distribution of an SRBM is known to be the solution of a partial differential equation problem called the basic adjoint relationship (BAR). For a one-dimensional SRBM, the BAR can be solved, and the stationary distribution turns out to be exponential, provided that the drift pertaining to the underlying Brownian motion is negative (see, e.g., [9, Theorem 6.2]). Observe that $\{\overline{Z}_i(t), t \geq 0\}$ can be written as a one-dimensional SRBM similar to the computations above, so that the limiting distributions of $\widetilde{W}_{i,r}$, $\widetilde{D}_{i,r}$ and $\widetilde{L}_{i,r}$ are indeed exponential distributions in line with Remarks 3.1, 4.1 and 4.2. For the multi-dimensional case, it is shown in [24] that if $\boldsymbol{R}$ is an M-matrix in the definition of [4, Chapter 6], a stationary distribution exists iff the reflection matrix satisfies $\boldsymbol{R}^{-1}\boldsymbol{\mu} < 0$. Under this condition, the stationary distribution is also shown to be unique. However, determining an exact solution to the BAR is generally a hard problem. In the special case where the reflection matrix $\boldsymbol{R}$ and the covariance matrix $\boldsymbol{\Gamma}$ satisfy a so-called skew-symmetry condition, the density of the stationary distribution is known to be of product form, of which each marginal is exponential (see [25]).

Numerical algorithms for solving the BAR however exist, so that the stationary distribution of SRBMs can be computed numerically. In [11] an algorithm has been developed to compute the stationary distribution, by exploiting a certain orthogonality property of the solution to the basic adjoint relationship. By taking a well-chosen reference density such as the product form density mentioned above, and introducing a reference measure, this algorithm computes in an iterative manner an unknown vector that can be thought of as some adjusting factor of how far the actual density of the stationary distribution is from the reference density. The computed unknown vector and the reference density then together form the desired solution.

For the model as given in Section 2, a unique stationary distribution for $\overline{\boldsymbol{Z}} = \lim_{t\to\infty} \overline{\boldsymbol{Z}}(t)$ exists, as the reflection matrix $\boldsymbol{R}$ and the drift vector $\boldsymbol{\mu}$ of the N-dimensional Brownian Motion $\boldsymbol{Z}$ satisfy the conditions mentioned. The skew-symmetry conditions only hold in our setting when $\boldsymbol{\Gamma}$ is a diagonal matrix, but this is only the case for very specific choices of the service speed functions $\phi_i(\cdot)$ and/or the Markov chain $\{\Phi(t), t \geq 0\}$. In the application of the two-layered network for instance, we have that $\gamma_{1,2}^C > 0$ by the expressions found in the previous section, so that the skew-symmetry condition is violated. The numerical algorithm developed in [11], however, can be applied generally to the model described in Section 2.

We end this section by applying the numerical algorithm to the two-layered network given in Section 5.1 and observing several parameter effects. Note that the limiting distribution $\overline{\boldsymbol{Z}}$ coincides with the stationary distribution of an SRBM with parameters $\boldsymbol{R}$ being a 2×2 identity matrix, $\boldsymbol{\mu} = \left(-\beta_1(\pi_{(U,U)} + \pi_{(U,R)}), -\beta_2(\pi_{(U,U)} + \pi_{(R,U)})\right)$ and $\boldsymbol{\Gamma} = diag\left(\frac{\mathbb{E}[B_1^2]}{\mathbb{E}[B_1]}(\pi_{(U,U)} + \pi_{(U,R)}), \frac{\mathbb{E}[B_2^2]}{\mathbb{E}[B_2]}(\pi_{(U,U)} + \pi_{(R,U)})\right) + \boldsymbol{\Gamma}^C$, where $\boldsymbol{\Gamma}^C$ is a 2x2 matrix consisting of the elements $\gamma_{i,j}^C$ computed in Section 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 $\overline{Z}_1$ and $\overline{Z}_2$. The results are summarised in Table 1, where for each of the instances the found values for $\mathbb{E}[\overline{Z}_1]$, $\mathbb{E}[\overline{Z}_2]$ and the correlation coefficient $\text{Corr}[\overline{Z}_1, \overline{Z}_2] = \frac{\mathbb{E}[\overline{Z}_1\overline{Z}_2] - \mathbb{E}[\overline{Z}_1]\mathbb{E}[\overline{Z}_2]}{\sqrt{\mathbb{E}[\overline{Z}_1^2] - \mathbb{E}[\overline{Z}_1]^2}\sqrt{\mathbb{E}[\overline{Z}_2^2] - \mathbb{E}[\overline{Z}_2]^2}}$ are given. Recall that the marginal distribution of $\overline{Z}_i$ is exponential, so that $\mathbb{E}[Z_i^2] = 2\mathbb{E}[Z_i]^2$. Observe also that the limiting distributions of $\widetilde{\boldsymbol{D}}_r$ and $\widetilde{\boldsymbol{L}}_r$ are equal to the distribution of

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	0.445
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

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

$\overline{\boldsymbol{Z}}$ up to a scalar, so that 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 frequence 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).

Acknowledgements

The authors wish to thank Sem Borst and Onno Boxma for providing valuable comments on earlier drafts of this paper.

References

[1] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003.

[2] B. Ata and S. Shneorson. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science*, 52:1778–1791, 2006.

[3] R. Bekker. *Queues with State-Dependent Rates*. PhD thesis, Eindhoven University of Technology, 2005.

[4] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979.

[5] O.J. Boxma and I.A. Kurkova. The M/G/1 queue with two service speeds. *Advances in Applied Probability*, 33:520–540, 2001.

[6] M. Bramson and J.G. Dai. Heavy traffic limits for some queueing networks. *The Annals of Applied Probability*, 11:49–90, 2001.

[7] M. Bramson, J.G. Dai, and J.M. Harrison. Positive recurrence of reflecting Brownian motion in three dimensions. *The Annals of Applied Probability*, 20:753–783, 2010.

[8] H. Chen and W. Whitt. Diffusion approximations for open queueing networks with service interruptions. *Queueing Systems*, 13:335–359, 1993.

[9] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks*. Springer-Verlag, New York, 2001.

[10] G. L. Choudhury, A. Mandelbaum, M.I. Reiman, and W. Whitt. Fluid and diffusion limits for queues in slowly changing environments. *Communications in Statistics. Stochastic Models*, 13:121–146, 1997.

[11] J.G. Dai and J.M. Harrison. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, 2:65–86, 1992.

[12] J.G. Dai and M. Miyazawa. Reflecting Brownian motion in two dimensions: exact asymptotics for the stationary distribution. *Stochastic Systems*, 1:146–208, 2011.

[13] J.G. Dai and M. Miyazawa. Stationary distribution of a two-dimensional SRBM: geometric views and boundary measures. *Queueing Systems*, 2013. To appear.

[14] K. Debicki, K.M. Kosiński, and M. Mandjes. Gaussian queues in light and heavy traffic. *Queueing Systems*, 71:137–149, 2012.

[15] A.B. Dieker and J. Moriarty. Reflected Brownian motion in a wedge: sum-of-exponential stationary densities. *Electronic Communications in Probability*, 14:1–16, 2009.

[16] J.L. Dorsman, S. Bhulai, and M. Vlasiou. Dynamic server assignment in an extended machine-repair model. Technical Report 2012-020, Eurandom Preprint Series, 2012.

[17] J.L. Dorsman, O.J. Boxma, and M. Vlasiou. Marginal queue length approximations for a two-layered network with correlated queues. *Queueing Systems*, 2013. To appear.

[18] J.L. Dorsman, R.D. Van der Mei, and M. Vlasiou. Analysis of a two-layered network with correlated queues by means of the power-series algorithm. Technical Report 2012-005, Eurandom Preprint Series, 2012.

[19] A. El Kharroubi, A.B. Tahar, and A. Yaacoubi. Sur la récurrence positive du mouvement Brownien réflechi dans l'orthant positif de $\mathbb{R}^n$. *Stochastics and Stochastics Reports*, 68:229–253, 2000.

[20] G. Franks, T. Al-Omari, M. Woodside, O. Das, and S. Derisavi. Enhanced modeling and solution of layered queuing networks. *IEEE Transactions on Software Engineering*, 35:148–161, 2009.

[21] J.M. George and J.M. Harrison. Dynamic control of a queue with adjustable service rate. *Operations Research*, 49:720–731, 2001.

[22] S. Halfin. Steady-state distribution for the buffer content of an M/G/1 queue with varying service rate. *SIAM Journal on Applied Mathematics*, 23:356–363, 1972.

[23] J.M. Harrison and V. Nguyen. Brownian models of multiclass queueing networks: current status and open problems. *Queueing Systems*, 13:5–40, 1993.

[24] J.M. Harrison and R.J. Williams. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, 22:77–115, 1987.

[25] J.M. Harrison and R.J. Williams. Multidimensional reflected Brownian motions having exponential stationary distributions. *The Annals of Probability*, 15:115–137, 1987.

[26] W.J. Hopp, S.M.R. Iravani, and G.J. Yuen. Operations systems with discretionary task completion. *Management Science*, 53:61–77, 2006.

[27] J. Ivanovs, O.J. Boxma, and M.R.H. Mandjes. Singularities of the matrix exponent of a Markov additive process with one-sided jumps. *Stochastic Processes and their Applications*, 120:1776–1794, 2010.

[28] P.R. Jelenković, P. Momčilović, and B. Zwart. Reduced load equivalence under subexponentiality. *Queueing Systems*, 46:97–112, 2004.

[29] J. Keilson and L.D. Servi. A distributional form of Little's law. *Operations Research Letters*, 7:223–227, 1988.

[30] O. Kella and W. Whitt. Diffusion approximations for queues with server vacations. *Advances in Applied Probability*, 22:706–729, 1990.

[31] J.F.C. Kingman. The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, 57:902–904, 1961.

[32] J.F.C. Kingman. The heavy traffic approximation in the theory of queues. In W.L. Smith and W.E. Wilkinson, editors, *Proceedings of the Symposium on Congestion Theory*, pages 137–159. Chapel Hill: University of North Carolina Press, 1965.

[33] K.M. Kosiński, O.J. Boxma, and B. Zwart. Convergence of the all-time supremum of a Lévy process in the heavy-traffic regime. *Queueing Systems*, 67:295–304, 2011.

[34] S.R. Mahabhashyam and N. Gautam. On queues with Markov modulated service rates. *Queueing Systems*, 51:89–113, 2005.

[35] D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability*, 20:646–676, 1988.

[36] M. Miyazawa and M. Kobayashi. Conjectures on tail asymptotics of the marginal stationary distribution for a multidimensional SRBM. *Queueing Systems*, 68:251–260, 2011.

[37] R. Núñez-Queija. A queueing model with varying service rate for ABR. In R. Puigjaner, N.N. Savino, and B. Serra, editors, *Proceedings of the 10th International Conference on Computer Performance Evaluation: Modelling Techniques and Tools*, pages 93–104. Springer, Berlin, 1998.

[38] P. Purdue. The M/M/1 queue in a Markovian environment. *Operations Research*, 22:562–569, 1974.

[39] M.I. Reiman. Some diffusion approximations with state space collapse. In *Modelling and Performance Evaluation Methodology (Paris, 1983)*, Lecture Notes in Control and Information Sciences, pages 209–240. Springer, Berlin, 1984.

[40] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Springer, New York, 1999.

[41] S. Shneer and V. Wachtel. A unified approach to the heavy-traffic analysis of the maximum of random walks. *Theory of Probability and Its Applications*, 55:332–341, 2011.

[42] J.L. Steichen. A functional central limit theorem for Markov additive processes with an application to the closed Lu-Kumar network. *Stochastic Models*, 17:459–489, 2001.

[43] S. Stidham Jr. and R.R. Weber. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations Research*, 37:611–625, 1989.

[44] L. Takács. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.

[45] T. Takine. Single-server queues with Markov-modulated arrivals and service speed. *Queueing Systems*, 49:7–22, 2005.

[46] E.I. Tzenova, I.J.B.F. Adan, and V.G. Kulkarni. Fluid models with jumps. *Stochastic Models*, 21:37–55, 2005.

[47] R.R. Weber and S. Stidham, Jr. Optimal control of service rates in networks of queues. *Advances in Applied Probability*, 19:202–218, 1987.

[48] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.

[49] U. Yechiali and P. Naor. Queueing problems with heterogeneous arrival and service. *Operations Research*, 19:722–734, 1971.

[50] B. Zwart. Heavy-traffic asymptotics for the single-server queue with random order of service. *Operations Research Letters*, 33:511–518, 2004.