Scaling limits for closed product-form queueing networks

van Kreveld, L.R.; Boxma, O.J.; Dorsman, J.L.; Mandjes, M.R.H.

DOI
10.1016/j.peva.2021.102220
Publication date
2021
Document Version
Final published version
Published in
Performance Evaluation

Performance Evaluation 151 (2021) 102220

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Scaling limits for closed product-form queueing networks

L.R. van Kreveld [b,*], O.J. Boxma [a], J.L. Dorsman [b], M.R.H. Mandjes [b]

[a] Eurandom and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[b] Korteweg–de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 10 July 2021

Keywords:
Queueing theory
Queueing networks
Scaling limits

ABSTRACT

We consider a general class of closed product-form queueing networks, consisting of single-server queues and infinite-server queues. Even if a network is of product-form type, performance evaluation tends to be difficult due to the potentially large state space and the dependence between the individual queues. To remedy this, we analyze the model in a Halfin–Whitt inspired scaling regime, where we jointly blow up the traffic loads of all queues and the number of customers in the network. This leads to a closed-form limiting stationary distribution, which provides intuition on the impact of the dependence between the queues on the network's behavior. We assess the practical applicability of our results through a series of numerical experiments, which illustrate the convergence and show how the scaling parameters can be chosen to obtain accurate approximations.

1. Introduction

Queues are often part of larger systems. Aiming to evaluate their performance, a substantial research effort has focused on the analysis of queueing networks, a prominent complication being that the individual queues in the network are often dependent. A central role is played by the class of networks obeying a product-form stationary distribution, where its components correspond to the numbers of customers in the individual queues (also referred to as stations). These product-form networks were first studied in the seminal papers by R. Jackson [1] and J. Jackson [2] in the 1950s, triggering much research in this area. Most notably, a large class of product-form networks, so-called BCMP networks [3], was identified in the 1970s, covering queueing networks consisting of single-server, multi-server and infinite-server stations. Since the discovery of the BCMP class, many further results have been obtained. On one hand, it has been shown that introducing features such as batch routing [4], loss dynamics [5], discrete-time dynamics [6, Chapter 6] and negative customers [7] does not break the product-form nature of the stationary distribution. On the other hand, general properties of product-form networks have been revealed, such as the arrival theorem [8] and aggregation theorems [9]. For an overview of the queueing-network literature we refer to [6].

Within the study of queueing networks a distinction has been made between open and closed networks. In open networks, i.e. networks with external arrivals and departures, if the stationary distribution factorizes into components corresponding to individual queues, then the queue lengths are mutually independent. Closed networks, however, have the additional constraint that the sum of the queue lengths must equal the population size at any point in time, rendering the individual queue lengths dependent. As a consequence, analytical and numerical difficulties arise when one aims at

* Corresponding author.
E-mail addresses: l.r.vankreveld@uva.nl (L.R. van Kreveld), o.j.boxma@tue.nl (O.J. Boxma), j.l.dorsman@uva.nl (J.L. Dorsman), m.r.h.mandjes@uva.nl (M.R.H. Mandjes).

https://doi.org/10.1016/j.peva.2021.102220

evaluating performance measures. Closed-form expressions are often beyond reach because the population-size constraint complicates the evaluation of terms summing over all possible queue-length vectors, which appear in for instance the normalization constant. Additionally, for large networks numerical approaches face computational challenges, such as the need to evaluate summations over a large set of states. In addition, there is the risk of running into computer-precision related problems, a challenge that has been addressed by Lam [10], but only in relation to the evaluation of the normalization constant.

To remedy the above-mentioned issues arising when analyzing closed product-form queueing networks, in various papers one has advocated the use of scaling limits. Here the objective is to obtain closed-form distributional results in specific asymptotic regimes. A prominent approach relies on integral representations for the generating function of the stationary distribution, which has been used to asymptotically evaluate the normalization constant [11,12]. In [13] strong approximation theory is applied to produce limit theorems for a large class of queueing networks. We also refer to the exact-order asymptotic analysis developed in [14].

Our paper can be seen as part of the research area discussed in the previous paragraph, in that it addresses the analytical and numerical difficulties of closed queueing networks by proposing a scaling method. The general idea of such a method is to make some of the model parameters depend on a number n in a certain way, and let n tend to infinity. When done in a suitable way, one can obtain insightful asymptotic results, revealing a tractable approximation for the behavior of the more complex unscaled system. A prime example of the effectiveness of scaling is the widely-recognized Halfin–Whitt regime [15]. For an Erlang loss queue, this regime scales the workload and the number of servers in a quality-and-efficiency driven way: the utilization of the servers approaches 100%, while the blocking probability remains close to zero. Halfin and Whitt prove that the number of customers in the scaled system tends to a truncated normal random variable in the limit.

For our model, i.e. a general closed product-form network consisting of both single-server stations and infinite-server stations, we define a new scaling regime inspired by the Halfin–Whitt scaling. Indeed, we extend the Halfin–Whitt scaling in such a way that it is applicable to queueing networks rather than individual stations in isolation. This we do by letting the traffic load at all stations become large and choosing the population size in such a way that the joint queue-length distribution has a non-degenerate limit. We remark that for finite-capacity open networks, our regime has the same quality-and-efficiency property as the Halfin–Whitt regime. This study can be considered as an extension of our previous work [16], where a similar approach has been followed for a specific three-station closed network representing an extended machine-repair model. We substantially generalize the results from [16], in that we establish similar asymptotic results for a more general class of closed product-form networks.

The contributions of this paper are the following. Under the Halfin–Whitt inspired scaling, we obtain the asymptotic stationary joint distribution of all queue lengths in the closed product-form network. This specifically entails that, appropriately normalized, the queue lengths of the single-server stations behave as (possibly truncated) exponential random variables, whereas the queue lengths of the infinite-server stations behave as (possibly truncated) normal random variables. Whether the truncation needs to be imposed, depends on whether the queue under consideration is a dominant queue, i.e. the station with largest queue-length variance. In the typical case that there is a single dominant station, the queue lengths are asymptotically independent. Importantly, although the pre-limit stationary distribution is relatively involved, it considerably simplifies under our scaling. Furthermore, we observe by means of numerical experiments that for a reasonably sized system, the queue-length distributions are well approximated by their limit distributions.

The paper is organized as follows. In Section 2 we describe our model in detail, analyze the normalization constant and introduce our scaling regime. The main results are then stated and discussed in Section 3. The proof of our main theorem is given in Section 4, where we leave some technical details for Appendix A. Subsequently, the practical relevance of our model is discussed in greater detail in Section 5, while numerical results in Section 6 show that the limiting queue-length distributions are able to yield accurate approximations. We conclude and provide pointers for further research in Section 7.

2. Model and preliminaries

This section presents the model description and a number of key concepts that play an important role in this paper. First, Section 2.1 introduces the product-form stationary distribution and describes which networks satisfy it. We discuss the normalization constant of this stationary distribution in Section 2.2. Subsequently, Section 2.3 gives the precise definition of the scaling regime that we study in this paper.

2.1. Model description

We consider a closed queueing network with C customers. Each station can be of two types: R stations are infinite-server queues, while the remaining $K+1$ are single-server queues. At a later stage we omit one single-server station, as its queue length equals C minus the sum of the other queue lengths. It is worth noting that the total service rate provided to all customers in any of the R infinite-server stations is linear in the number of customers present, since all customers can be served simultaneously in an infinite-server queue. In any of the single-server stations, however, the service rate provided is constant whenever the number of customers present is positive, and zero otherwise. See Fig. 1 for an example of such a network. Let $B_1, \ldots, B_R$ be the stationary numbers of customers at the infinite-server stations, and $D_1, \ldots, D_{K+1}$ their counterparts at the single-server stations.

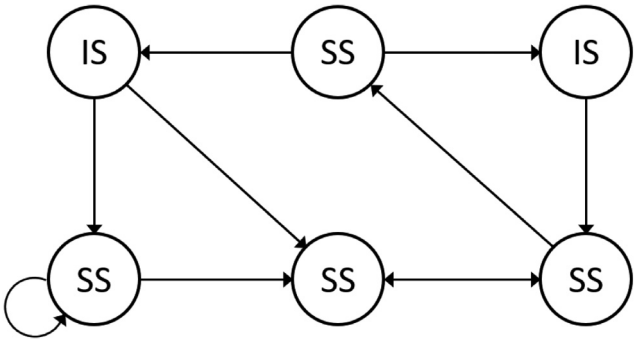

Fig. 1. Closed queueing network with infinite-server (IS) stations and single-server (SS) stations.

The only further assumption we impose on our model is that it has the following product-form stationary distribution: for $b_1, \ldots, b_R, d_1, \ldots, d_{K+1}$ such that $b_1 + \cdots + b_R + d_1 + \cdots + d_{K+1} = C$,

$$\mathbb{P}(B_1 = b_1, \ldots, B_R = b_R, D_1 = d_1, \ldots, D_{K+1} = d_{K+1}) = \tilde{p}_0 \prod_{r=1}^{R} \frac{\eta_r^{b_r}}{b_r!} \prod_{k=1}^{K+1} \theta_k^{d_k}. \tag{1}$$

Here $\eta_r, \theta_k \geqslant 0$ are parameters representing traffic loads of individual stations, and $\tilde{p}_0$ is the normalization constant, which ensures that all probabilities sum to 1.

The stationary distribution (1) applies under fairly broad conditions. A precise specification providing all instances that yield this particular product form is rather challenging (see [6, Section 5.7] for a detailed discussion). However, an important sufficient network property for this is *quasi-reversibility*. This property, concerning individual stations, states that if a station has a Poisson arrival process, its departure process is also Poisson and the queue length is independent of past departures (see e.g. [17, Section 6] for more background). If all stations of the network are quasi-reversible when considered in isolation, then the stationary distribution is guaranteed to obey the product form (1). To give an example, quasi-reversibility holds for infinite-server stations under arbitrary service time distributions, and for single-server stations when service times are exponential under FCFS, or when the server applies processor sharing or the LCFS pre-emptive resume discipline. Quasi-reversibility is a sufficient condition for a product-form stationary distribution, but it is not necessary (cf. [18]).

Although the network described above is assumed to be closed, there is also a class of finite-capacity open networks with Poisson arrivals satisfying (1). Such networks are technically open, but behave exactly like a closed network. This concept is explained in great detail in [19], in the context of computer systems with window flow control. A similar reasoning applies to an open queueing network with $R + K$ stations, where external arrivals are blocked when there are already C customers present in the network. To see why this system can be interpreted as a closed network, suppose that all departures out of the system join an artificial single-server station (which we identify with station $K + 1$), and that all external admitted arrivals form the departure process of this station. The system is now closed, and with the total number of customers being equal to C, the behavior at the original stations is the same. Indeed, the original situation where a customer is blocked is equivalent to the situation where no customers are present in station $K + 1$.

We proceed with a few notational issues. Throughout this paper we will denote vectors by bold symbols. This for instance means that we denote by $\boldsymbol{b}$ and $\boldsymbol{d}$ the vectors $(b_1, \ldots, b_R)$ and $(d_1, \ldots, d_K)$, respectively. We also introduce specific notation related to the truncation of such vectors to their first entries: for instance for $r \leqslant R$ we mean by $\boldsymbol{b}_r$ the vector $(b_1, \ldots, b_r)$, so that $\boldsymbol{b}_R = \boldsymbol{b}$. For the sum of the entries of vectors we use the well-known norm notation, e.g., $\|\boldsymbol{b}\| := \sum_{r=1}^{R} b_r$. Furthermore, we use the convention that a geometric random variable has support $\{0, 1, 2, \ldots\}$; if the success probability is p, we write $\mathscr{G}(p)$. With $\mathscr{P}(\mu)$ we denote a Poisson random variable with mean μ. For sequences f_n and g_n, we write $f_n \sim g_n$ if $f_n/g_n \to 1$ as $n \to \infty$. Additionally, '$=_{\mathrm{d}}$' and '$\to_{\mathrm{d}}$', respectively, denote equality in distribution and convergence in distribution.

In this paper, we will work with a normalized version of (1). We assume without loss of generality that $\theta_{K+1} = \max_{k=1,\ldots,K+1}\{\theta_k\}$. Due to the closed nature of the network, rescaling all station parameters through division by θ_{K+1} leads to a different normalization constant, but otherwise this has no effect on the stationary joint distribution. Therefore, (1) can be rewritten as, with $\|\boldsymbol{b}\| + \|\boldsymbol{d}\| \leqslant C$,

$$p_{\boldsymbol{b},\boldsymbol{d}} := \mathbb{P}(\boldsymbol{B} = \boldsymbol{b}, \boldsymbol{D} = \boldsymbol{d}) = p_0 \prod_{r=1}^{R} \frac{\rho_r^{b_r}}{b_r!} \prod_{k=1}^{K} \sigma_k^{d_k}, \tag{2}$$

where $p_0 = \theta_{K+1}^{C} \tilde{p}_0$, $\rho_r = \eta_r/\theta_{K+1}$ for all r and $\sigma_k = \theta_k/\theta_{K+1} \leqslant 1$ for all k. The parameters $\rho_1, \ldots, \rho_R$ and $\sigma_1, \ldots, \sigma_K$ can be interpreted as the traffic loads of the corresponding stations. The joint stationary distribution (2) is the starting point

of the scaling analysis presented in this paper, and we view $\rho_1, \ldots, \rho_R$ and $\sigma_1, \ldots, \sigma_K$ as system parameters. The results in the rest of the paper are valid for every queueing network that satisfies (2).

Remark 1. In this paper we consider the system's behavior in a specific regime in which the total number of customers C grows, according to a scaling that we will specify later. By (2), there is dependence between the stations: the individual stationary queue lengths are correlated due to the constraint $\|\boldsymbol{B}\| + \|\boldsymbol{D}\| \leqslant C$. However, it also implies that when C grows large, this dependence becomes weaker, and in the limit as $C \to \infty$, vanishes.

For the infinite-server station indexed by $r \leqslant R$, the probability of b_r customers at the station is proportional to $\rho_r^{b_r}/b_r!$. For this reason, the queue length at infinite-server station r is approximately distributed as $\mathscr{P}(\rho_r)$ as C grows large.

Likewise, for a single-server station index by $k \leqslant K$, the probability of d_k customers at the station is proportional to $\sigma_k^{d_k}$. If $\sigma_k < 1$ this implies that the queue-length distribution approximately behaves as $\mathscr{G}(\sigma_k)$ as C grows large.

The remaining station, single-server station $K+1$, has the highest traffic load among the single-server stations and thus acts as a bottleneck of the network. This means that its queue length becomes arbitrarily large with C. $\diamond$

2.2. Normalization constant

We now turn our attention to the calculation of the normalization constant p_0 in (2). Note that since $\|\boldsymbol{B}\| + \|\boldsymbol{D}\| \leqslant C$, we have

$$p_0^{-1} = \sum_{\boldsymbol{b},\boldsymbol{d}\,:\,\|\boldsymbol{b}\|+\|\boldsymbol{d}\|\leqslant C} \prod_{r=1}^{R} \frac{\rho_r^{b_r}}{b_r!} \prod_{k=1}^{K} \sigma_k^{d_k}. \tag{3}$$

Observe that the normalization constant involves summation over terms that are products of R Poisson-type factors, and K geometric-type factors. Because of the $R+K$ indices, the number of terms in the summation in (3) is $\binom{C+R+K}{R+K}$, making direct calculation of the normalization constant beyond reach for large networks. Alternative approaches have been proposed, such as the method of generating functions [20], for efficient algorithms computing the normalization constant and marginal queue-length distributions. More recently, a generalized method of moments has been studied for closed product-form networks [21].

To provide a different way of decreasing its numerical complexity, we now focus on a simplified representation of the normalization constant, given in Lemma 2. The proofs of the two lemmas in this subsection are not essential for understanding the main results of this paper, and could be skipped at first reading. We still include them however, since we use the same approach in the proof of our main theorem.

To obtain the different representation for p_0, we evaluate the summation over the K geometric-type indices, and subsequently use a probabilistic argument for the summation over the R Poisson-type factors. For this purpose, it is useful to define

$$S_j(x) := \sum_{\boldsymbol{b}\,:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/x)^{b_r}}{b_r!} \sum_{\boldsymbol{d}\,:\,\|\boldsymbol{d}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j} \left(\frac{\sigma_k}{x}\right)^{d_k}, \tag{4}$$

so that $p_0^{-1} = S_K(1)$. To evaluate the inner geometric sum, we wish to express $p_0^{-1} = S_K(1)$ in terms of $S_0(x)$ for certain x. The following recursion is a key element in this derivation.

Lemma 1. *For $x \neq \sigma_j$, $S_j(x)$ satisfies the recursion*

$$S_j(x) = \frac{1}{1-\sigma_j/x} S_{j-1}(x) - \frac{(\sigma_j/x)^{C+1}}{1-\sigma_j/x} S_{j-1}(\sigma_j), \qquad j = 1, \ldots, K.$$

Proof. The recursion follows from an evaluation of the geometric series. Taking the sum over d_j, we see that for $x \neq \sigma_j$,

$$\begin{aligned} S_j(x) &= \sum_{\boldsymbol{b}\,:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/x)^{b_r}}{b_r!} \sum_{\boldsymbol{d}_{j-1}\,:\,\|\boldsymbol{d}_{j-1}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j-1} \left(\frac{\sigma_k}{x}\right)^{d_k} \frac{1-(\sigma_j/x)^{C-\|\boldsymbol{b}\|-\|\boldsymbol{d}_{j-1}\|+1}}{1-\sigma_j/x} \\ &= \frac{1}{1-\sigma_j/x} \left(\sum_{\boldsymbol{b}\,:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/x)^{b_r}}{b_r!} \sum_{\boldsymbol{d}_{j-1}\,:\,\|\boldsymbol{d}_{j-1}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j-1} \left(\frac{\sigma_k}{x}\right)^{d_k} \right. \\ &\qquad \left. - (\sigma_j/x)^{C+1} \sum_{\boldsymbol{b}\,:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/\sigma_j)^{b_r}}{b_r!} \sum_{\boldsymbol{d}_{j-1}\,:\,\|\boldsymbol{d}_{j-1}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j-1} \left(\frac{\sigma_k}{\sigma_j}\right)^{d_k} \right) \\ &= \frac{1}{1-\sigma_j/x} S_{j-1}(x) - \frac{(\sigma_j/x)^{C+1}}{1-\sigma_j/x} S_{j-1}(\sigma_j), \end{aligned} \tag{5}$$

thus proving the claim. $\square$

The lemma shows that $S_j(x)$ can be split into two terms, each involving $S_{j-1}(\cdot)$. We exploit this recursion to derive an alternative expression for p_0, which is presented in the following lemma. As an aside, we remark that here and in the sequel, the cases $\sigma_k = 1$ for some k and $\sigma_j = \sigma_l$ for some j, l can be resolved using L'Hôpital's rule.

Lemma 2. *The normalization constant equals*

$$p_0 = \left(\prod_{k=1}^{K} \frac{1}{1-\sigma_k} \left(S_0(1) - \sum_{l=1}^{K} \sigma_l^{C+1} \prod_{j=1, j\neq l}^{K} \frac{1-\sigma_j}{1-\sigma_j/\sigma_l} S_0(\sigma_l) \right) \right)^{-1}, \tag{6}$$

where

$$S_0(x) = \sum_{i=0}^{C} \frac{(\|\boldsymbol{\rho}\|/x)^i}{i!}.$$

Proof. Note that by (3) and Lemma 1,

$$p_0^{-1} = S_K(1) = \frac{1}{1-\sigma_K} S_{K-1}(1) - \frac{\sigma_K^{C+1}}{1-\sigma_K} S_{K-1}(\sigma_K).$$

Applying Lemma 1 another $K-1$ times leads to an expression of the form

$$p_0^{-1} = a\, S_0(1) + \sum_{l=1}^{K} u_l\, S_0(\sigma_l), \tag{7}$$

where a and $u_1, \ldots, u_K$ are coefficients depending on $\sigma_1, \ldots, \sigma_K$. To find a, observe that the only term with $S_0(1)$ results from the first term of Lemma 1 of all K iterations. Therefore, $a = \prod_{k=1}^{K}(1-\sigma_k)^{-1}$. Similarly, observe that the only term with $S_0(\sigma_K)$ follows from the second term in the first iteration and then the first term in all remaining iterations. Therefore, $u_K = -\sigma_K^{C+1}(1-\sigma_K)^{-1}\prod_{j=1}^{K-1}(1-\sigma_j/\sigma_K)^{-1}$. Note that the single-server stations $1, \ldots, K$ are identical in (3) up to their parameters $\sigma_1, \ldots, \sigma_K$. By symmetry, we conclude that, for any $l = 1, \ldots, K$,

$$u_l = -\frac{\sigma_l^{C+1}}{1-\sigma_l} \prod_{j=1, j\neq l}^{K} \frac{1}{1-\sigma_j/\sigma_l}.$$

Thus, it holds that

$$\begin{aligned} p_0^{-1} &= S_0(1) \prod_{k=1}^{K} \frac{1}{1-\sigma_k} - \sum_{l=1}^{K} S_0(\sigma_l) \frac{\sigma_l^{C+1}}{1-\sigma_l} \prod_{j=1, j\neq l}^{K} \frac{1}{1-\sigma_j/\sigma_l} \\ &= \left(\prod_{k=1}^{K} \frac{1}{1-\sigma_k} \right) \left(S_0(1) - \sum_{l=1}^{K} S_0(\sigma_l)\, \sigma_l^{C+1} \prod_{j=1, j\neq l}^{K} \frac{1-\sigma_j}{1-\sigma_j/\sigma_l} \right). \end{aligned} \tag{8}$$

To prove Lemma 2, it remains to show that $S_0(x) = \sum_{i=0}^{C}(\|\boldsymbol{\rho}\|/x)^i/i!$. This is done by expressing $S_0(x)$ in terms of cumulative Poisson probabilities. That is, using (4),

$$\begin{aligned} S_0(x) &= \sum_{\boldsymbol{b}:\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/x)^{b_r}}{b_r!} \\ &= e^{\|\boldsymbol{\rho}\|/x} \sum_{\boldsymbol{b}:\|\boldsymbol{b}\|\leqslant C} \mathbb{P}\left(\mathscr{P}\left(\frac{\rho_1}{x}\right) = b_1, \ldots, \mathscr{P}\left(\frac{\rho_R}{x}\right) = b_R \right) \\ &= e^{\|\boldsymbol{\rho}\|/x} \mathbb{P}\left(\sum_{r=1}^{R} \mathscr{P}\left(\frac{\rho_r}{x}\right) \leqslant C \right) = e^{\|\boldsymbol{\rho}\|/x} \mathbb{P}\left(\mathscr{P}\left(\frac{\|\boldsymbol{\rho}\|}{x}\right) \leqslant C \right) = \sum_{i=0}^{C} \frac{(\|\boldsymbol{\rho}\|/x)^i}{i!}, \end{aligned}$$

which concludes the proof. □

The number of numerical operations needed to evaluate the normalization constant using (6) is $O(K^2 + KC)$. The algorithm of [20] has a complexity of $O(RC^2 + KC)$, and has the benefit that marginal queue lengths follow without much additional computational effort. However, when knowledge of the joint distribution is required, one cannot avoid computing the individual probabilities of all $\binom{C+R+K}{R+K}$ states, which is tractable only for very small networks. We resolve this issue by working in a scaling regime, that will be introduced in the next subsection, and in which the stationary distribution exhibits easy-to-interpret behavior. The asymptotic findings can be used to devise approximations for the unscaled system, as will be pointed out in Section 6.

2.3. The scaling regime

When distributions do not allow a closed-form analysis, a commonly used approach in applied probability is to resort to scaling limits. The main idea is to parametrize (a subset of) the system parameters by n, with the objective to arrive at an explicit limiting distribution as $n \to \infty$. It is often not *a priori* clear how this parametrization should be done; finding a scaling that leads to useful and meaningful results in the limit is an art on its own. The resulting limiting distribution can be used to produce approximations for the pre-limit system.

In their celebrated 1981 paper, Halfin and Whitt [15] introduced an important new scaling for many-server queues. The asymptotic regime considered corresponds to letting the workload ρ and the number of servers C grow to infinity in such a way that $(C-\rho)/\sqrt{\rho}$ converges to a constant $\bar{\beta} > 0$. In the specific context of the Erlang loss model, an appropriately normalized version of the queue length then asymptotically behaves as a normal random variable truncated at $\bar{\beta}$ [15, Theorem 3]. The scaling we impose in our network setting is inspired by the Halfin–Whitt regime, in that we also scale the parameter ρ and the total number of customers C.

We now give a precise definition of the scaling we impose in this paper. Let $\nu_1, \ldots, \nu_R, \alpha_1, \ldots, \alpha_K \in \mathbb{R}$ and $w_1, \ldots, w_R, c_1, \ldots, c_K > 0$ be scaling parameters for individual stations. It proves useful to assume without loss of generality that $\nu_1 \geqslant \cdots \geqslant \nu_R$ and that $\alpha_1 \leqslant \cdots \leqslant \alpha_K$. We scale the system parameters as follows:

- we replace ρ_r by $\rho_r^{(n)}$ for each $r \in \{1, \ldots, R\}$,
- we replace σ_k by $\sigma_k^{(n)}$ for each $k \in \{1, \ldots, K\}$,
- we replace C by C_n,

where

$$\rho_r^{(n)} = w_r\, n^{\nu_r}, \quad \sigma_k^{(n)} := \frac{n}{n + c_k n^{\alpha_k}},$$

which can be interpreted as scaled traffic loads, and where for $\beta > 0$ and $\gamma := \max\{1-\alpha_1, \frac{1}{2}\nu_1\}$, we let the total number of customers be defined as

$$C_n := \left\lfloor \left\| \boldsymbol{\rho}^{(n)} \right\| + \beta n^{\gamma} \right\rfloor. \tag{9}$$

The definitions for $\rho_r^{(n)}$, $\sigma_k^{(n)}$ and C_n may seem restrictive, but any network satisfying (2) can be constructed with the right choices of the scaling parameters. For a detailed discussion on fitting these parameters to a system in practice, we refer to Section 6.2.

Observe that in this scaling regime, the traffic loads of the infinite-server stations become arbitrarily large as $n \to \infty$ (provided $\nu_1, \ldots, \nu_R > 0$), and the traffic loads of the single-server stations tend to 1 as $n \to \infty$ (provided $\alpha_1, \ldots, \alpha_K < 1$). To account for the large queue lengths that are inherent for these traffic loads, also the population size C_n grows (at a suitable pace) as $n \to \infty$.

Our precise choice (9) for C_n can be motivated as follows. It turns out that, to get non-degenerate limits, the total number of customers should be picked such that it equals the mean of $\|\boldsymbol{B}\|$ increased by a constant β times the largest of the standard deviations of all queue lengths. As argued in Remark 1, when the total number of customers is large, B_r behaves as $\mathscr{P}(\rho_r) = \mathscr{P}(\rho_r^{(n)})$, which has standard deviation $(\rho_r^{(n)})^{\frac{1}{2}} = \sqrt{w_r}\, n^{\frac{1}{2}\nu_r}$. In addition, when $\sigma_k < 1$, D_k behaves as $\mathscr{G}(1-\sigma_k) = \mathscr{G}(1-\sigma_k^{(n)})$, which has standard deviation

$$\frac{\sqrt{\sigma_k^{(n)}}}{1-\sigma_k^{(n)}} \sim \frac{1}{c_k} n^{1-\alpha_k}.$$

Since $\nu_1 \geqslant \cdots \geqslant \nu_R$ and $\alpha_1 \leqslant \cdots \leqslant \alpha_K$, the largest asymptotic standard deviation is attained by either $B_1^{(n)}$ or $D_1^{(n)}$. Note that the first case applies if $1-\alpha_1 < \frac{1}{2}\nu_1$, and the second if $1-\alpha_1 > \frac{1}{2}\nu_1$. These observations, and the fact that the population size must be an integer, intuitively explain our choice (9) for C_n.

In line with the observations above, we say that a station is *dominant* if its asymptotic queue-length variance has the largest power of n out of all the stations (excluding single-server station $K+1$). Thus, infinite-server station $r \in \{1, \ldots, R\}$ is dominant if $\frac{1}{2}\nu_r = \gamma$, and single-server station $k \in \{1, \ldots, K\}$ is dominant if $1-\alpha_k = \gamma$. Later it will turn out that in the limit as $n \to \infty$, it is precisely the dominant stations that are affected by the population size constraint.

Under our scaling the queue lengths $\boldsymbol{B}$ at the infinite-server stations and the queue lengths $\boldsymbol{D}$ at the single-server stations depend on n. In the sequel we let $\boldsymbol{B}^{(n)}$ and $\boldsymbol{D}^{(n)}$, respectively, denote the corresponding random vectors. Our objective is to analyze their behavior as $n \to \infty$. Since their means may become arbitrarily large with n, we consider normalized versions: for $r \in \{1, \ldots, R\}$ and $k \in \{1, \ldots, K\}$,

$$\bar{B}_r^{(n)} := \frac{B_r^{(n)} - \rho_r^{(n)}}{\sqrt{\rho_r^{(n)}}}, \quad \text{and} \quad \bar{D}_k^{(n)} := (1-\sigma_k^{(n)}) D_k^{(n)}. \tag{10}$$

To help the reader understand the key concepts in this paper, a table of the most important scaling parameters is included in Appendix D.

3. Results

In this section, we derive the asymptotic joint distribution of $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$. The remaining queue length, $D_{K+1}^{(n)}$, can then be obtained from the identity

$$\left\|\boldsymbol{B}^{(n)}\right\| + \left\|\boldsymbol{D}^{(n)}\right\| + D_{K+1}^{(n)} = C_n. \tag{11}$$

Our main result, which holds for $\nu_1, \ldots, \nu_R > 0$ and $\alpha_1, \ldots, \alpha_K < 1$, is presented in Section 3.1. Section 3.2 presents an adaptation of our main result for networks with only single-server stations. The remaining cases, where $\nu_R \leqslant 0$ or $\alpha_K \geqslant 1$, are investigated in Section 3.3.

3.1. Main result

We first consider the case where the traffic loads are large at all stations. That is, we assume that $\nu_1, \ldots, \nu_R > 0$ (the traffic loads at infinite-server stations tend to infinity) and $\alpha_1, \ldots, \alpha_K < 1$ (the traffic loads at single-server stations tend to 1). We study the normalized queue lengths $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$ by means of their joint Laplace–Stieltjes transform (LST). We define this LST using (10):

$$\begin{aligned} P_n(\boldsymbol{s},\boldsymbol{t}) &:= \mathbb{E}\left(\prod_{r=1}^{R} e^{-s_r \bar{B}_r^{(n)}} \prod_{k=1}^{K} e^{-t_k \bar{D}_k^{(n)}}\right) \\ &= \sum_{\boldsymbol{b},\boldsymbol{d}\,:\,\|\boldsymbol{b}\|+\|\boldsymbol{d}\|\leqslant C_n} \left(\prod_{r=1}^{R} e^{-s_r \frac{b_r-\rho_r^{(n)}}{\sqrt{\rho_r^{(n)}}}}\right)\left(\prod_{k=1}^{K} e^{-t_k(1-\sigma_k^{(n)})d_k}\right)\mathbb{P}(\boldsymbol{B}^{(n)}=\boldsymbol{b},\boldsymbol{D}^{(n)}=\boldsymbol{d}). \end{aligned} \tag{12}$$

It is noted that strictly speaking $P_n(\boldsymbol{s},\boldsymbol{t})$ is not an LST, as the random variables $\bar{B}_r^{(n)}$ may attain negative values. This feature does not affect the upcoming analysis, including the application of Lévy's convergence theorem, and hence we will stick to the term LST.

Our main theorem, to be proved in Section 4, gives an explicit expression for the limit of $P_n(\boldsymbol{s},\boldsymbol{t})$ as $n \to \infty$, from which we can directly derive the asymptotic distribution of $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$.

A bit of notation is necessary for the statement of the main theorem. First, for a standard-normally distributed random variable, we write

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

for its density function and $\Psi(x) := \Phi(x)/\phi(x)$ for its *Mills ratio*, i.e. its distribution function divided by its density function. Secondly, concerning the highest-loaded stations, let $R^- \leqslant R$ be the largest integer such that $\nu_1 = \nu_2 = \cdots = \nu_{R^-}$ and let $K^- \leqslant K$ be the largest integer such that $\alpha_1 = \alpha_2 = \cdots = \alpha_{K^-}$. Observe that the number of dominant stations can now be expressed as

$$R^- \mathbb{1}_{\{1-\alpha_1 \leqslant \frac{1}{2}\nu_1\}} + K^- \mathbb{1}_{\{1-\alpha_1 \geqslant \frac{1}{2}\nu_1\}}.$$

We then denote

$$W := \sum_{r=1}^{R^-} w_r \quad \text{and} \quad \lambda(\boldsymbol{s}) := \frac{\beta + \sum_{r=1}^{R^-} s_r \sqrt{w_r}}{\sqrt{W}}.$$

In addition, define

$$\kappa_{jl}(\boldsymbol{t}) := \frac{c_j(1+t_j)}{c_j(1+t_j) - c_l(1+t_l)}, \quad \zeta(\boldsymbol{t}) := 1 - \sum_{l=1}^{K^-}\left(\prod_{j=1,\,j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{t})\right) e^{-\beta c_l(1+t_l)},$$

$$\eta(\boldsymbol{s},\boldsymbol{t}) := \Psi\left(\lambda(\boldsymbol{s})\right) - \sum_{l=1}^{K^-}\left(\prod_{j=1,\,j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{t})\right) \Psi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right),$$

$$\xi(\boldsymbol{s}) := \Phi\left(\lambda(\boldsymbol{s})\right) \quad \text{and} \quad \chi(\boldsymbol{s}) := \phi(\lambda(\boldsymbol{s})).$$

With this notation, we can state the main theorem as follows.

Theorem 1. *Consider a queueing network with stationary distribution (2). Assume that $\nu_R > 0$ and that $\alpha_K < 1$. Then the joint LST $P_n(\boldsymbol{s},\boldsymbol{t})$ of $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$ satisfies*

$$\lim_{n\to\infty} P_n(\boldsymbol{s},\boldsymbol{t}) = \left(\prod_{r=1}^{R} e^{\frac{1}{2}s_r^2}\right)\left(\prod_{k=1}^{K} \frac{1}{1+t_k}\right) U(\boldsymbol{s},\boldsymbol{t}), \tag{13}$$

where

$$U(\boldsymbol{s},\boldsymbol{t}) := \begin{cases} \dfrac{\zeta(\boldsymbol{t})}{\zeta(\boldsymbol{0})} & \text{if } 1-\alpha_1 > \frac{1}{2}\nu_1, \\ \dfrac{\chi(\boldsymbol{s})}{\chi(\boldsymbol{0})}\cdot\dfrac{\eta(\boldsymbol{s},\boldsymbol{t})}{\eta(\boldsymbol{0},\boldsymbol{0})} & \text{if } 1-\alpha_1 = \frac{1}{2}\nu_1, \\ \dfrac{\xi(\boldsymbol{s})}{\xi(\boldsymbol{0})} & \text{if } 1-\alpha_1 < \frac{1}{2}\nu_1. \end{cases}$$

In all three cases we recognize known distributions from the joint LST. Let $\mathscr{N}_1,\ldots,\mathscr{N}_R,\mathscr{E}_1,\ldots,\mathscr{E}_K$ be independent random variables, the first R having standard-normal distributions and the last K having unit-rate exponential distributions. In the corollary below we claim that the right-hand side of (13) is the LST of the $(R+K)$-tuple $(\mathscr{N},\mathscr{E})$ conditioned on $Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta$, where

$$Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-}) := \mathbb{1}_{\{1-\alpha_1\leqslant\frac{1}{2}\nu_1\}}\sum_{r=1}^{R^-}\sqrt{w_r}\mathscr{N}_r + \mathbb{1}_{\{1-\alpha_1\geqslant\frac{1}{2}\nu_1\}}\sum_{k=1}^{K^-}\frac{1}{c_k}\mathscr{E}_k. \tag{14}$$

To provide some intuition why this condition makes sense, consider the population size constraint $\|\boldsymbol{B}^{(n)}\|+\|\boldsymbol{D}^{(n)}\|\leqslant C_n$. Subtracting $\|\boldsymbol{\rho}^{(n)}\|$ and dividing by n^γ on both sides, we have

$$\frac{\|\boldsymbol{B}^{(n)}\|-\|\boldsymbol{\rho}^{(n)}\|}{n^\gamma}+\frac{\|\boldsymbol{D}^{(n)}\|}{n^\gamma}\leqslant\frac{C_n-\|\boldsymbol{\rho}^{(n)}\|}{n^\gamma}. \tag{15}$$

Recalling the definition $\gamma := \max\{1-\alpha_1,\frac{1}{2}\nu_1\}$ and the scaled queue lengths (10), if $\bar{B}_r^{(n)}\to_{\mathrm{d}}\mathscr{N}_r$ for each r and $\bar{D}_k^{(n)}\to_{\mathrm{d}}\mathscr{E}_k$ for each k, the inequality (15) would converge, as $n\to\infty$, to $Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta$.

To summarize, Theorem 1 leads to the following corollary.

Corollary 1. *As $n\to\infty$,*

$$(\bar{\boldsymbol{B}}^{(n)},\bar{\boldsymbol{D}}^{(n)})\to_{\mathrm{d}}\Big(\mathscr{N},\mathscr{E}\,\Big|\,Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta\Big). \tag{16}$$

Consequently, the random variables $\bar{B}_1^{(n)},\ldots,\bar{B}_R^{(n)},\bar{D}_1^{(n)},\ldots,\bar{D}_K^{(n)}$ are asymptotically independent if the number of dominant stations is one.

Proof. With standard integration techniques one can check that the joint Laplace–Stieltjes transform of the tuple $(\mathscr{N},\mathscr{E}\,|\,Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta)$ is precisely as given in Theorem 1 (see Appendix C), so that the stated follows by Lévy's convergence theorem. The independence statement follows from the fact that if there is only one dominant station, then Z depends on just one random variable. □

It can be seen from (14) that the condition $Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta$, relating to the condition that there are at most C_n customers at the $R+K$ stations, only involves indices corresponding to the dominant stations. Hence in the limit, the population size constraint only affects the dominant stations. In addition, suppose that the number of dominant stations is one, which happens precisely if $\frac{1}{2}\nu_1>\max\{\nu_2,1-\alpha_1\}$ or $1-\alpha_1>\max\{\nu_1,1-\alpha_2\}$. In that case, the condition applies to only one random variable, which amounts to a truncation of that variable.

Remark 2. We mention the consequences of our main result for single-server station $K+1$. As described earlier, this is the single-server station with the highest traffic load, the queue length of which follows from the remaining queue lengths by the population size constraint. Corollary 1 thus implicitly provides the asymptotic distribution of $D_{K+1}^{(n)}$. Dividing the population size constraint (11) by n^γ, we have that

$$\frac{D_{K+1}^{(n)}}{n^\gamma}=\frac{C_n-\|\boldsymbol{\rho}^{(n)}\|}{n^\gamma}-\frac{\|\boldsymbol{B}^{(n)}\|-\|\boldsymbol{\rho}^{(n)}\|}{n^\gamma}-\frac{\|\boldsymbol{D}^{(n)}\|}{n^\gamma}.$$

By (9), (10) and Corollary 1, it thus holds that

$$\frac{D_{K+1}^{(n)}}{n^\gamma}\to_{\mathrm{d}}\beta-(Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\,|\,Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta)$$

as $n\to\infty$. ◇

3.2. Scaling result for $R=0$

So far we have omitted networks consisting of single-server stations only, because Theorem 1 relies on the value of ν_1. In the case that $R=0$, however, this parameter does not exist. With a slight modification, we can establish the counterpart of Corollary 1 for single-server networks.

Corollary 2. *Suppose $R = 0$ and $\alpha_K < 1$. As $n \to \infty$,*

$$\bar{\boldsymbol{D}}^{(n)} \to_{\mathrm{d}} \left(\boldsymbol{\mathscr{E}} \,\middle|\, Z\left(\mathbf{0}, \boldsymbol{\mathscr{E}}_{K^-}\right) \leqslant \beta \right). \tag{17}$$

The variables $\bar{D}_1^{(n)}, \ldots, \bar{D}_K^{(n)}$ are thus asymptotically independent if $K^- = 1$.

Proof. The result follows from Corollary 1 by setting $\nu_1 = -\infty$. □

3.3. Scaling results for $\nu_R \leqslant 0$ and/or $\alpha_K \geqslant 1$

Since Theorem 1 only covers the case where $\nu_R > 0$ and $\alpha_K < 1$, it remains to analyze its complement in which $\nu_R \leqslant 0$ and/or $\alpha_K \geqslant 1$. Recall that in Section 2.3, we introduced normalized versions of $\boldsymbol{B}^{(n)}$ and $\boldsymbol{D}^{(n)}$ in order to preserve finite mean. Note however that for infinite-server stations r with $\nu_r \leqslant 0$ and for single-server stations k with $\alpha_k \geqslant 1$, the *unnormalized* queue length converges to a finite-mean random variable. Because of this, it is no longer necessary to normalize in these cases. For all $r \in \{1, \ldots, R\}$ for which $\nu_r \leqslant 0$, we will therefore consider the distribution of the random variable $B_r^{(n)}$ instead of $\bar{B}_r^{(n)}$. Likewise, for all $k \in \{1, \ldots, K\}$ for which $\alpha_k \geqslant 1$, we will consider the distribution of the random variable $D_k^{(n)}$ instead of $\bar{D}_k^{(n)}$.

In this regime, a statement similar to Corollary 1 holds, which is given in the following corollary.

Corollary 3. *Assume that $\nu_1 > 0$ or $\alpha_1 < 1$. Let I be the smallest integer such that $\nu_I \leqslant 0$, and let J be the smallest integer such that $\alpha_J \geqslant 1$. As $n \to \infty$,*

$$\left(\bar{B}_1^{(n)}, \ldots, \bar{B}_{I-1}^{(n)}, \bar{D}_1^{(n)}, \ldots, \bar{D}_{J-1}^{(n)} \right) \to_{\mathrm{d}} \left(\mathscr{N}_1, \ldots, \mathscr{N}_{I-1}, \mathscr{E}_1, \ldots, \mathscr{E}_{J-1} \,\middle|\, Z\left(\boldsymbol{\mathscr{N}}_{R^-}, \boldsymbol{\mathscr{E}}_{K^-}\right) \leqslant \beta \right).$$

Proof. See Appendix B. □

Remark 3. The remaining random variables, i.e. $B_I^{(n)}, \ldots, B_R^{(n)}$ and $D_J^{(n)}, \ldots, D_K^{(n)}$, all have finite mean because $\nu_I, \ldots, \nu_R \leqslant 0$ and $\alpha_J, \ldots, \alpha_K \geqslant 1$. In the proof of Corollary 3 we will see that they behave as Poisson random variables with means $\rho_I^{(n)}, \ldots, \rho_R^{(n)}$ and geometric random variables with parameters $1-\sigma_J^{(n)}, \ldots, 1-\sigma_K^{(n)}$, respectively. This implies in particular that, for each r such that $\nu_r < 0$ and for each k such that $\alpha_k > 1$, the random variables $B_r^{(n)}$ and $D_k^{(n)}$ become degenerate with value 0 as $n \to \infty$. ◇

Remark 4. Corollary 3 assumes that $\nu_1 > 0$ or $\alpha_1 < 1$ because our scaling would not make sense otherwise. If $\nu_1 \leqslant 0$ and $\alpha_1 \geqslant 1$, the stations' traffic loads no longer increase with n. ◇

4. Proof of Theorem 1

In this section we present a proof of our main theorem, Theorem 1, which consists of two parts. First, in Section 4.1, we derive a structured expression for $P_n(\boldsymbol{s}, \boldsymbol{t})$ (Lemma 3) relying on techniques similar to those used in the proof of Lemma 2 (the derivation of the normalization constant). Then, we discuss this expression piece by piece, already recognizing some known LSTs and providing intuition.

In the second part of the proof (Section 4.2), we asymptotically analyze in Lemmas 5–8 all parts of the expression obtained from Lemma 3. In most cases, we can build on a version of the central limit theorem (Lemma 4) to find the asymptotics. One particular case, however, requires more subtle asymptotic bounds, and this case is treated in Lemma 8. We finish the proof by substituting the asymptotically analyzed parts back into the expression for $P_n(\boldsymbol{s}, \boldsymbol{t})$.

In the proofs, some mathematical expressions will repeatedly appear in our calculations. To keep these calculations readable, we use the following notation.

- The adapted traffic load for infinite-server station r:

$$\zeta_r^{(n)}(s_r) := \rho_r^{(n)} \exp\left(-\frac{s_r}{\sqrt{\rho_r^{(n)}}} \right), \quad \zeta^{(n)}(\boldsymbol{s}) := \sum_{r=1}^{R} \zeta_r^{(n)}(s_r).$$

- The adapted traffic load for single-server station l:

$$\delta_l^{(n)}(t_l) := \sigma_l^{(n)} e^{-t_l(1-\sigma_l^{(n)})}.$$

- A frequently occurring quantity related to the single-server stations j and l:

$$y_{jl}^{(n)}(t_j, t_l) := \frac{1 - \delta_j^{(n)}(t_j)}{1 - \delta_j^{(n)}(t_j)/\delta_l^{(n)}(t_l)}.$$

- A Poisson probability related to $S_0(1)$:
$$f^{(n)}(\boldsymbol{s}) := \mathbb{P}\left(\mathscr{P}\left(\zeta^{(n)}(\boldsymbol{s})\right) \leqslant C_n\right).$$
- A Poisson probability related to $S_0(\delta_l^{(n)}(t_l))$:
$$g_l^{(n)}(\boldsymbol{s}, t_l) := \mathbb{P}\left(\mathscr{P}\left(\zeta^{(n)}(\boldsymbol{s})/\delta_l^{(n)}(t_l)\right) \leqslant C_n\right).$$
- A quantity appearing in $P_n(\boldsymbol{s}, \boldsymbol{t})$:
$$h_l^{(n)}(\boldsymbol{s}, t_l) := \exp\left(\zeta^{(n)}(\boldsymbol{s})\left(\frac{1}{\delta_l^{(n)}(t_l)} - 1\right)\right)\left(\delta_l^{(n)}(t_l)\right)^{C_n+1}.$$

4.1. Structured form of LST

The following lemma gives an exact expression for the LST $P_n(\boldsymbol{s}, \boldsymbol{t})$ in terms of the new notation that was introduced above, and forms the backbone of the proof of Theorem 1.

Lemma 3. *The LST of* $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$ *satisfies*

$$\begin{aligned} P_n(\boldsymbol{s}, \boldsymbol{t}) = &\left(\prod_{r=1}^R e^{-\rho_r^{(n)} + s_r\sqrt{\rho_r^{(n)}} + \zeta_r^{(n)}(s_r)}\right) \times \left(\prod_{k=1}^K \frac{1-\sigma_k^{(n)}}{1-\delta_k^{(n)}(t_k)}\right) \\ &\times \frac{f^{(n)}(\boldsymbol{s}) - \sum_{l=1}^K \left(\prod_{j=1, j\neq l}^K y_{jl}^{(n)}(t_j, t_l)\right) g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l)}{f^{(n)}(\boldsymbol{0}) - \sum_{l=1}^K \left(\prod_{j=1, j\neq l}^K y_{jl}^{(n)}(0, 0)\right) g_l^{(n)}(\boldsymbol{0}, 0)\, h_l^{(n)}(\boldsymbol{0}, 0)}. \end{aligned} \tag{18}$$

Proof. Denote by $p_{\boldsymbol{b},\boldsymbol{d}}^{(n)}$ and $p_0^{(n)}$, respectively, the stationary distribution and normalization constant of the scaled system. Then, we can rewrite the joint LST of $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$ in (12) as

$$\begin{aligned} P_n(\boldsymbol{s}, \boldsymbol{t}) &= \sum_{\boldsymbol{b},\boldsymbol{d}:\, \|\boldsymbol{b}\|+\|\boldsymbol{d}\| \leqslant C_n} \left(\prod_{r=1}^R e^{-s_r \frac{b_r - \rho_r^{(n)}}{\sqrt{\rho_r^{(n)}}}}\right)\left(\prod_{k=1}^K e^{-t_k(1-\sigma_k^{(n)})d_k}\right) p_{\boldsymbol{b},\boldsymbol{d}}^{(n)} \\ &= p_0^{(n)}\left(\prod_{r=1}^R e^{s_r\sqrt{\rho_r^{(n)}}}\right) \sum_{\boldsymbol{b}:\, \|\boldsymbol{b}\| \leqslant C_n} \left(\prod_{r=1}^R \frac{\left(\zeta_r^{(n)}(s_r)\right)^{b_r}}{b_r!}\right) \sum_{\boldsymbol{d}:\, \|\boldsymbol{d}\| \leqslant C_n - \|\boldsymbol{b}\|} \left(\prod_{k=1}^K \left(\delta_k^{(n)}(t_k)\right)^{d_k}\right) \\ &= p_0^{(n)}\left(\prod_{r=1}^R e^{s_r\sqrt{\rho_r^{(n)}}}\right) S_K^{(n)}(1), \end{aligned}$$

where $S_j^{(n)}(x)$ is obtained from $S_j(x)$ when (ρ_r, σ_k, C) is replaced by $(\zeta_r^{(n)}(s_r), \delta_k^{(n)}(t_k), C_n)$. Therefore we have by (8) that

$$\begin{aligned} P_n(\boldsymbol{s}, \boldsymbol{t}) &= p_0^{(n)}\left(\prod_{r=1}^R e^{s_r\sqrt{\rho_r^{(n)}}}\right)\left(\prod_{k=1}^K \frac{1}{1-\delta_k^{(n)}(t_k)}\right)\left(S_0^{(n)}(1) - \sum_{l=1}^K \left(\delta_l^{(n)}(t_l)\right)^{C_n+1}\left(\prod_{j=1, j\neq l}^K y_{jl}^{(n)}(t_j, t_l)\right) S_0^{(n)}\left(\delta_l^{(n)}(t_l)\right)\right) \\ &= p_0^{(n)}\left(\prod_{r=1}^R e^{s_r\sqrt{\rho_r^{(n)}}}\right)\left(\prod_{k=1}^K \frac{1}{1-\delta_k^{(n)}(t_k)}\right)\Bigg(e^{\zeta^{(n)}(\boldsymbol{s})}\mathbb{P}\left(\mathscr{P}\left(\zeta^{(n)}(\boldsymbol{s})\right) \leqslant C_n\right) \\ &\quad - \sum_{l=1}^K \left(\delta_l^{(n)}(t_l)\right)^{C_n+1}\left(\prod_{j=1, j\neq l}^K y_{jl}^{(n)}(t_j, t_l)\right) e^{\frac{1}{\delta_l^{(n)}(t_l)}\zeta^{(n)}(\boldsymbol{s})}\,\mathbb{P}\left(\mathscr{P}\left(\frac{1}{\delta_l^{(n)}(t_l)}\zeta^{(n)}(\boldsymbol{s})\right) \leqslant C_n\right)\Bigg) \\ &= p_0^{(n)}\left(\prod_{r=1}^R e^{s_r\sqrt{\rho_r^{(n)}} + \zeta_r^{(n)}(s_r)}\right)\left(\prod_{k=1}^K \frac{1}{1-\delta_k^{(n)}(t_k)}\right)\left(f^{(n)}(\boldsymbol{s}) - \sum_{l=1}^K\left(\prod_{j=1, j\neq l}^K y_{jl}^{(n)}(t_j, t_l)\right) g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l)\right). \end{aligned} \tag{19}$$

Using $P_n(\boldsymbol{0}, \boldsymbol{0}) = 1$ we find that $p_0^{(n)}$ equals

$$\left(\prod_{r=1}^R e^{-\rho_r^{(n)}}\right)\left(\prod_{k=1}^K \left(1-\sigma_k^{(n)}\right)\right)\left(f^{(n)}(\boldsymbol{0}) - \sum_{l=1}^K\left(\prod_{j=1, j\neq l}^K y_{jl}^{(n)}(0, 0)\right) g_l^{(n)}(\boldsymbol{0}, 0)\, h_l^{(n)}(\boldsymbol{0}, 0)\right)^{-1},$$

and after substituting this back in (19), the proof is completed. □

The expression for $P_n(\boldsymbol{s},\boldsymbol{t})$ in Lemma 3 is a product of three factors (separated by the $\times$-symbols). These factors, say $u_1^{(n)}$, $u_2^{(n)}$, and $u_3^{(n)}$, each play an intuitively appealing role in relation to Corollary 1. More specifically, our analysis below reveals that as $n\to\infty$ the first two factors $u_1^{(n)}$ and $u_2^{(n)}$ correspond to the transforms of the normal and exponential distribution, respectively. In addition, we show that as $n\to\infty$ the factor $u_3^{(n)}$ (which is the second line of (18)) immediately relates to the condition $Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta$. As will become clear in the proofs, the factor $u_3^{(n)}$ is significantly more subtle to analyze than the factors $u_1^{(n)}$ and $u_2^{(n)}$.

Let us start with $u_1^{(n)}$. By applying a standard Taylor expansion to $\exp(-s_r/\sqrt{\rho_r}^{(n)})$ around zero, we obtain

$$\begin{aligned}&\exp\left(-\rho_r^{(n)}+s_r\sqrt{\rho_r^{(n)}}+\zeta_r^{(n)}(s_r)\right)\\&\quad=\exp\left(-\rho_r^{(n)}+s_r\sqrt{\rho_r^{(n)}}+\rho_r^{(n)}\left(1-\frac{s_r}{\sqrt{\rho_r^{(n)}}}+\frac{s_r^2}{2\rho_r^{(n)}}+o\left(\frac{1}{\rho_r^{(n)}}\right)\right)\right).\end{aligned}\tag{20}$$

As $n\to\infty$, Expression (20) converges to $\exp(\frac{1}{2}s_r^2)$, which can be recognized as the transform $\mathbb{E}(\exp(-s_r\mathscr{N}))$ of a standard-normal random variable $\mathscr{N}$. From this we can conclude that, as $n\to\infty$, $u_1^{(n)}$ converges to a product of R standard-normal LSTs.

For $u_2^{(n)}$, we can apply the same strategy: for each k and $\alpha_k<1$, as $n\to\infty$, we have $\sigma_k^{(n)}\to 1$ so that we can apply a Taylor expansion to $\exp(-t_k(1-\sigma_k^{(n)}))$ around zero. Therefore,

$$\begin{aligned}\frac{1-\sigma_k^{(n)}}{1-\delta_k^{(n)}(t_k)}&=\frac{1-\sigma_k^{(n)}}{1-\sigma_k^{(n)}(1-t_k(1-\sigma_k^{(n)})+o(1-\sigma_k^{(n)}))}\\&=\frac{1-\sigma_k^{(n)}}{(1-\sigma_k^{(n)})(1+\sigma_k^{(n)}t_k)+o(1-\sigma_k^{(n)})}.\end{aligned}\tag{21}$$

As $n\to\infty$, Expression (21) converges to $(1+t_k)^{-1}$, which is the LST of an exponentially distributed random variable with rate 1. This implies that, as $n\to\infty$, $u_2^{(n)}$ converges to a product of K unit-rate exponential LSTs.

With the asymptotic behavior of $u_1^{(n)}$ and $u_2^{(n)}$ at hand, to prove Theorem 1 it remains to analyze $u_3^{(n)}$. To this end, we inspect the behavior of the functions $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s},t_l)$ and $h_l^{(n)}(\boldsymbol{s},t_l)$ in the limiting regime for $n\to\infty$, distinguishing different values of α_1 and α_l. This analysis is covered by the next subsection.

In addition to $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s},t_l)$ and $h_l^{(n)}(\boldsymbol{s},t_l)$, the sequence $u_3^{(n)}$ also contains the coefficients $\prod_{j=1,j\neq l}^K y_{jl}^{(n)}(t_j,t_l)$ for each single-server station $l=1,\ldots,K$. Multiplying the numerator and denominator of

$$y_{jl}^{(n)}(t_j,t_l)=\frac{1-\delta_j^{(n)}(t_j)}{1-\delta_j^{(n)}(t_j)/\delta_l^{(n)}(t_l)}$$

by $n/\delta_j^{(n)}(t_j)$, it follows that

$$y_{jl}^{(n)}(t_j,t_l)=\frac{n\left(e^{t_j(1-\sigma_j^{(n)})}-1\right)+c_jn^{\alpha_j}e^{t_j(1-\sigma_j^{(n)})}}{n\left(e^{t_j(1-\sigma_j^{(n)})}-e^{t_l(1-\sigma_l^{(n)})}\right)+c_jn^{\alpha_j}e^{t_j(1-\sigma_j^{(n)})}-c_ln^{\alpha_l}e^{t_l(1-\sigma_l^{(n)})}}.\tag{22}$$

As $\alpha_j,\alpha_l<1$, we have

$$y_{jl}^{(n)}(t_j,t_l)\sim\frac{c_j(1+t_j)}{c_j(1+t_j)-c_ln^{\alpha_l-\alpha_j}(1+t_l)}\to\begin{cases}0 & \text{if}\quad \alpha_j<\alpha_l,\\ \dfrac{c_j(1+t_j)}{c_j(1+t_j)-c_l(1+t_l)} & \text{if}\quad \alpha_j=\alpha_l,\\ 1 & \text{if}\quad \alpha_l<\alpha_j,\end{cases}\tag{23}$$

as $n\to\infty$. In particular, it holds that

$$\lim_{n\to\infty}\prod_{j=1,j\neq l}^K y_{jl}^{(n)}(t_j,t_l)=0\iff\alpha_l>\alpha_1.$$

This fact has an intuitive backing: we expect the condition in Corollary 1 to apply only to the stations with largest variability in queue lengths. For single-server stations, these are the stations l for which α_l is minimal.

4.2. Asymptotic analysis of $f^{(n)}(\cdot)$, $g_l^{(n)}(\cdot,\cdot)$ and $h_l^{(n)}(\cdot,\cdot)$

In Lemmas 5–8 we explicitly analyze the asymptotic behavior of the functions $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s},t_l)$ and $h_l^{(n)}(\boldsymbol{s},t_l)$ for all values of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$. To evaluate the cumulative Poisson probabilities appearing in $f^{(n)}(\boldsymbol{s})$ and $g_l^{(n)}(\boldsymbol{s},t_l)$, we require an additional central-limit type result given in Lemma 4. All lemmas in this subsection are proved in Appendix A.

Lemma 4. *Suppose $x_n \to \infty$ as $n \to \infty$. If $(C_n - x_n)/\sqrt{x_n} \to Q$ with $Q \in [-\infty,\infty]$, then $\mathbb{P}(\mathscr{P}(x_n) \leqslant C_n) \to \Phi(Q)$ as $n \to \infty$.*

We proceed by determining the asymptotic behavior of the functions $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s},t_l)$ and $h_l^{(n)}(\boldsymbol{s},t_l)$ as $n \to \infty$. This behavior is highly dependent on the values of ν_1, α_1 and α_l, so that it is necessary to distinguish various cases. In most cases standard asymptotic methods suffice (Lemmas 5–7), but one particular case requires a more refined approach (Lemma 8).

Lemma 5. *As $n \to \infty$,*

$$f^{(n)}(\boldsymbol{s}) \to \begin{cases} 1 & \text{if} \quad 1-\alpha_1 > \frac{1}{2}\nu_1, \\ \Phi(\lambda(\boldsymbol{s})) & \text{if} \quad 1-\alpha_1 \leqslant \frac{1}{2}\nu_1. \end{cases}$$

Lemma 6. *As $n \to \infty$,*

$$g_l^{(n)}(\boldsymbol{s},t_l) \to \begin{cases} 1 & \text{if} \quad 1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1, \\ \Phi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right) & \text{if} \quad 1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1. \end{cases}$$

Lemma 7. *As $n \to \infty$,*

$$h_l^{(n)}(\boldsymbol{s},t_l) \to \begin{cases} 0 & \text{if} \quad 1-\alpha_1 > 1-\alpha_l \text{ and } 1-\alpha_1 \geqslant \nu_1, \\ 0 & \text{if} \quad \nu_1 > 1-\alpha_1 > 1-\alpha_l \geqslant \frac{1}{2}\nu_1, \\ \exp(-\beta c_l(1+t_l)) & \text{if} \quad 1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1, \\ \dfrac{\phi(\lambda(\boldsymbol{s}))}{\phi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right)} & \text{if} \quad 1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1. \end{cases}$$

The above lemmas treat all cases where either $1-\alpha_1 \geqslant \nu_1$ or $1-\alpha_l \geqslant \frac{1}{2}\nu_1$. Although $g_l^{(n)}(\boldsymbol{s},t_l)$ is not evaluated in all these cases, observe that in (18) this function only occurs as the product $g_l^{(n)}(\boldsymbol{s},t_l)\,h_l^{(n)}(\boldsymbol{s},t_l)$. Since $g_l^{(n)}(\boldsymbol{s},t_l) \in [0,1]$, it follows that its specific value is irrelevant as long as $h_l^{(n)}(\boldsymbol{s},t_l) \to 0$. We conclude that only the case where both $1-\alpha_1 < \nu_1$ and $1-\alpha_l < \frac{1}{2}\nu_1$ remains.

This last case requires a more subtle reasoning. Since $g_l^{(n)}(\boldsymbol{s},t_l) \to 0$ and $h_l^{(n)}(\boldsymbol{s},t_l) \to \infty$ as $n \to \infty$, we must analyze the product of the two functions before taking the limit. The proof of Lemma 8 relies on a change-of-measure argument.

Lemma 8. *If $1-\alpha_1 < \nu_1$ and $1-\alpha_l < \frac{1}{2}\nu_1$, then $g_l^{(n)}(\boldsymbol{s},t_l)\,h_l^{(n)}(\boldsymbol{s},t_l) \to 0$ as $n \to \infty$.*

We have now collected all the ingredients to establish the asymptotic expression for $P_n(\boldsymbol{s},\boldsymbol{t})$ as presented in Theorem 1.

Proof of Theorem 1. The result is a consequence of Lemma 3 when substituting Eqs. (20), (21), and (23), in combination with the functions that we asymptotically evaluated in Lemmas 5–8 (both for general $\boldsymbol{s},\boldsymbol{t}$ and for $\boldsymbol{s}=\boldsymbol{t}=\boldsymbol{0}$, that is). □

5. Applications

Corollary 1 describes the asymptotic joint queue-length distribution under our scaling. This result may serve as the basis for approximations of the pre-limit distribution, which can be used e.g. when designing the network. Closed queueing networks can be broadly applied, as they can be used to represent for instance hospital units, computer systems, communication networks and manufacturing systems. In this section we discuss two illustrative examples.

Example 1 (*Extended Machine-Repair Model*)**.** In the extended machine-repair model, products that require processing arrive at a facility with C machines. If all machines are occupied, products are blocked and immediately leave the system upon arrival. An occupied machine may break down, and resumes processing only after it has been repaired by a single repairer. It is hereby assumed that a product remains assigned to the same machine for the duration of its service, even if the machine breaks down intermediately. An in-depth analysis of this system can be found in [16].

The queueing dynamics of this facility are visualized in Fig. 2. The network is open, but by the discussion in Section 2.1 it is equivalent to a closed network with two single-server stations (the external station and the repair station) and an infinite-server station (processing station). This closed network is depicted in Fig. 3, and under the conditions described in Section 2 it obeys a product-form distribution (2).

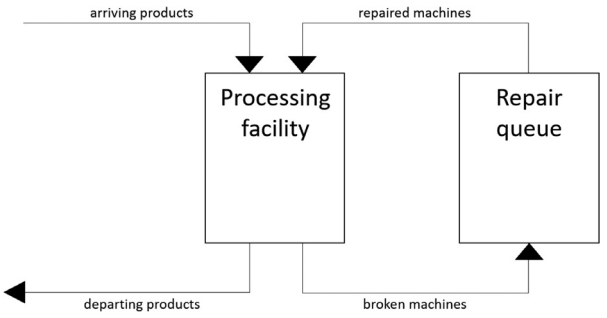

Fig. 2. Extended machine-repair model.

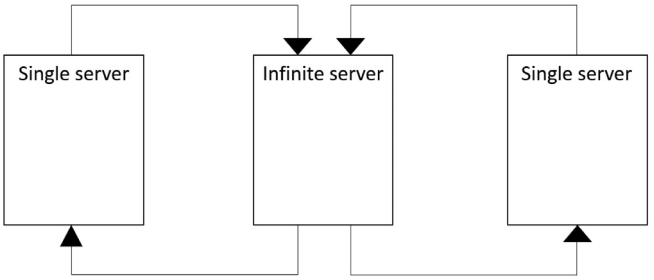

Fig. 3. Equivalent closed network.

Corollary 1 then states that the normalized numbers of occupied and broken machines tend to a normal and exponential distribution respectively, with a condition depending on the values of the chosen scaling parameters. If $1-\alpha_1 > \frac{1}{2}\nu_1$, the limiting distribution of the number of broken machines is truncated at $c_1\beta$. If $1-\alpha_1 < \frac{1}{2}\nu_1$, the limiting distribution of the number of occupied machines is truncated at $\beta/\sqrt{w_1}$. Finally, if $1-\alpha_1 = \frac{1}{2}\nu_1$, the condition amounts to $\sqrt{w_1}\mathcal{N}_1 + \frac{1}{c_1}\mathcal{E}_1 \leqslant \beta$, which in particular implies dependence between the queue lengths. $\diamond$

Example 2 (*Vehicle Sharing System*)**.** In modern society the demand for flexible transportation has led to the development of vehicle sharing systems. In such systems, a number of vehicles is scattered among a fixed number of locations. Users may pick up a vehicle at any location (if available) and drop it off at any, possibly different, location. To accurately describe the behavior of such a system, a well-fitting model is important.

Closed queueing networks are often used in modeling vehicle sharing systems, see e.g. George and Xia [22]. In this model, the population size C is the total number of vehicles across the network. The pick-up (and drop-off) locations are modeled by single-server queues, and between each ordered pair of pick-up locations, an infinite-server queue is used to describe the time spent by a user between these locations. See Fig. 4 for an example with three pick-up locations.

Notice that the number of stations used to model the network grows quadratically in the number of pick-up locations. For this reason, vehicle-sharing systems quickly become analytically and numerically intractable when the number of pick-up locations increases. The result of Corollary 1, however, does not become more complex as the number of stations grows. Under typical circumstances R^- and K^- are low and the asymptotic queue-length distributions are therefore (asymptotically) tractable. $\diamond$

6. Numerical illustrations

In Section 4 we have established a convergence-in-distribution result for the random vector $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$. In this section we will discuss the performance of approximations based on this scaling limit. In Section 6.1, we assess the pre-limit distributions by means of numerical experiments, and compare them to the limiting distributions of Corollary 1. Importantly, the number of scaling parameters (relating to the vectors $\boldsymbol{w}$, $\boldsymbol{\nu}$, $\boldsymbol{c}$ and $\boldsymbol{\alpha}$ and the scalar β) exceeds the number of parameters of our pre-limit model (i.e., the vectors $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ and the scalar C). This leaves us with some freedom to choose the scaling parameters; using an example network, we show in Section 6.2 how this can be done.

Computation of individual queue-length distributions directly from the stationary distribution (2) is hard for networks with many stations, as the state space grows quickly with the size of the network. We therefore choose to use an acceptance–rejection simulation [23] for all numerical experiments in this section. This entails sampling from the stationary distribution without the population size constraint (which comes down to separately sampling from Poisson

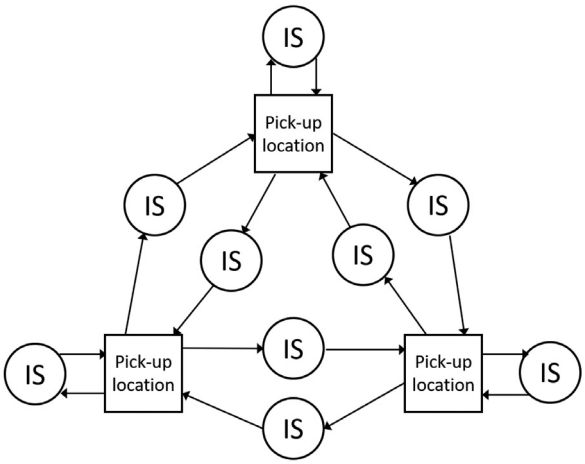

Fig. 4. Vehicle sharing system illustrated as a closed queueing network with single-server stations (pick-up locations) and infinite-server stations (IS).

Table 1
Five cases for the values of $\nu_1, \alpha_1, \alpha_2$ and the corresponding condition on the joint distribution of $(\bar{B}_1^{(n)}, \bar{D}_1^{(n)}, \bar{D}_2^{(n)})$.

	Values of $\nu_1, \alpha_1, \alpha_2$	Condition on the joint distribution of $(\bar{B}_1^{(n)}, \bar{D}_1^{(n)}, \bar{D}_2^{(n)})$
Case 1	$1-\alpha_1 > 1-\alpha_2,\ 1-\alpha_1 > \frac{1}{2}\nu_1$	$\frac{1}{c_1}\bar{D}_1^{(n)} \leqslant \beta$
Case 2	$1-\alpha_1 = 1-\alpha_2 > \frac{1}{2}\nu_1$	$\frac{1}{c_1}\bar{D}_1^{(n)} + \frac{1}{c_2}\bar{D}_2^{(n)} \leqslant \beta$
Case 3	$1-\alpha_1 = 1-\alpha_2 = \frac{1}{2}\nu_1$	$\sqrt{w_1}\bar{B}_1^{(n)} + \frac{1}{c_1}\bar{D}_1^{(n)} + \frac{1}{c_2}\bar{D}_2^{(n)} \leqslant \beta$
Case 4	$1-\alpha_1 = \frac{1}{2}\nu_1 > 1-\alpha_2$	$\sqrt{w_1}\bar{B}_1^{(n)} + \frac{1}{c_1}\bar{D}_1^{(n)} \leqslant \beta$
Case 5	$1-\alpha_1 < \frac{1}{2}\nu_1$	$\sqrt{w_1}\bar{B}_1^{(n)} \leqslant \beta$

and geometric distributions), and rejecting the samples that fail to satisfy the population size constraint. The set of accepted samples is then stochastically equivalent to a set of equally many independent samples from (2). An estimate for the marginal queue-length distributions can thus be found by counting the number of accepted samples with each possible queue-length value. Throughout this section we constantly use 10^7 samples for all simulation results (of which at least half gets accepted).

6.1. Accuracy of approximation

We start by considering networks consisting of a large number of stations — for instance, one can think of a vehicle-sharing system from Section 5 with ten pick-up locations, which has more than a hundred stations. We consider a setting with $R^- = 1$ and $K^- \leqslant 2$ (such that there are at most three dominant stations). This entails by Corollary 1 that the variables $\bar{B}_2^{(n)}, \ldots, \bar{B}_R^{(n)}$ converge to independent standard-normal random variables, and that $\bar{D}_3^{(n)}, \ldots, \bar{D}_K^{(n)}$ converge to independent unit-rate exponential random variables. The asymptotic distributions of $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ are less trivial because they may be affected by the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$. Fig. 5 therefore focuses on these random variables: it shows their density functions, estimated by simulation. Strictly speaking, the variables $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ have probability masses instead of densities for finite n since they are discrete. However, we consider the scaled mass functions of these variables and refer to them as densities in the sequel, so as to facilitate comparison with their limits as $n \to \infty$.

How the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$ impacts the distributions of $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ depends mainly on the values of ν_1, α_1 and α_2, see Table 1. The five different cases are visible in the rows of Fig. 5, in which simulation results are shown for a network with $R = 6$ and $K = 7$. In Cases 1 and 5, the condition applies to only one random variable, which causes the associated density function to be truncated at β.

In all of the cases the density of $\bar{B}_1^{(n)}$ resembles the normal density, whereas the densities of $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ resemble the exponential density. At a more detailed level, Fig. 5 also shows the impact of the different conditions (i.e. the five cases that were displayed in Table 2). In cases where $\bar{B}_1^{(n)}$ (or $\bar{D}_1^{(n)}, \bar{D}_2^{(n)}$) is not part of the condition, its density function is simply a slightly perturbed version of that of a standard normal (or unit-rate exponential). On the other hand, in cases where

Table 2
Scaling parameter values for the plots in Fig. 5.

	ν_1	$\nu_2, \ldots, \nu_6$	α_1	α_2	$\alpha_3, \ldots, \alpha_7$	$w_1, \ldots, w_6$	c_1	c_2	$c_3, \ldots, c_7$	β	n
Case 1	1	0.5	-1	0	0.5	1	1	1	1	1	25
Case 2	1	0.5	0	0	0.5	1	1	2	1	1	100
Case 3	1	0.5	0.5	0.5	0.9	1	1	2	1	1	100
Case 4	1	0.5	0.5	0.8	0.9	1	1	1	1	1	100
Case 5	2	0.5	0.9	0.9	0.9	1	1	1	1	1	100

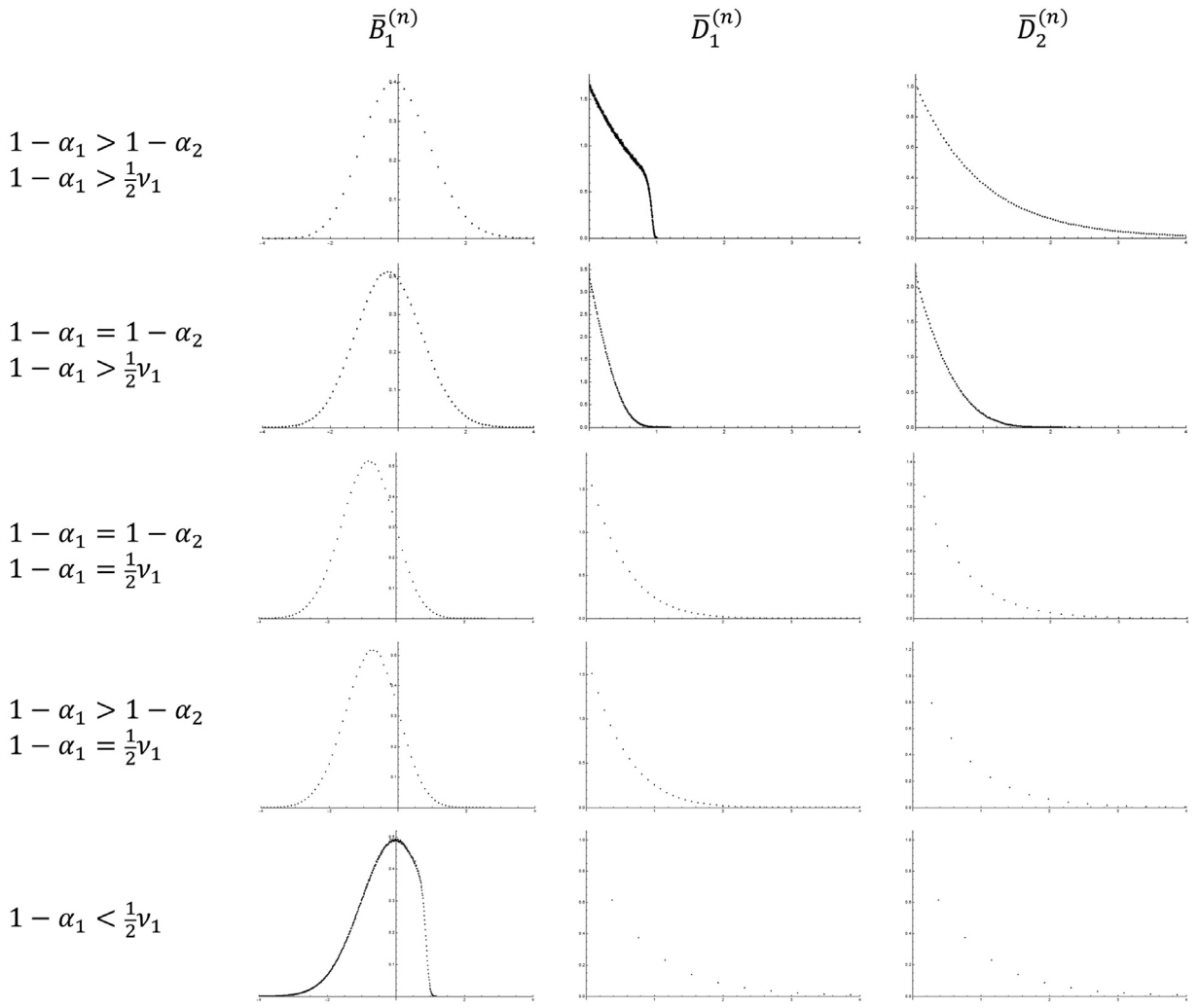

Fig. 5. Density functions of $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$, estimated by simulation, depending on the values of the scaling parameters $\nu_1, \alpha_1, \alpha_2$. The exact parameter values for the five cases (top to bottom) can be found in Table 2.

$\bar{B}_1^{(n)}$ (or $\bar{D}_1^{(n)}, \bar{D}_2^{(n)}$) *is* part of the condition, we see that the corresponding random variable is less likely to assume larger values. We conclude that the structure of the limit distributions, as identified in Corollary 1, carries over to the pre-limit setting.

6.2. Fitting scaling parameters

In the remainder of this section we show how to use our scaling regime in a concrete queueing network model. Particularly, we show how the scaling parameters can be chosen to appropriately reflect the model at hand. The extended

Table 3
Three parameter sets of the extended machine-repair model.

	C	ρ_1	σ_1
Parameter set 1	100	40	0.99
Parameter set 2	100	90	0.90
Parameter set 3	100	90	0.66

machine-repair model described in Section 5 will serve as an example. For this system we compare the actual queue-length distributions (obtained by acceptance–rejection simulation) to the limiting distributions in the scaling regime (as stated in Corollary 1). Although the acceptance–rejection simulation gives sufficiently accurate results and is consistent with the previous subsection, we remark that for the small machine-repair network, direct calculation of (2) is also numerically tractable.

To compare the behavior of the queue lengths under a given set of model parameters with our limit results, we have to choose appropriate scaling parameters. Since we have $2R + 2K + 1$ scaling parameters (the entries of the vectors $\boldsymbol{w}$, $\boldsymbol{\nu}$, $\boldsymbol{c}$, and $\boldsymbol{\alpha}$ and the scalar β) compared to only $R + K + 1$ model parameters (the entries of the vectors $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ and the scalar C), this can be done in many different ways. Choosing appropriate scaling parameters is important, because not all choices lead to accurate approximations. The following intuitive procedure may serve as a guideline.

(1) First select which stations are dominant, i.e. the stations whose queue lengths will be incorporated in the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$. These should correspond to the queue lengths with largest variance, as these are affected most by the population size constraint. These variances are, respectively, $\rho_1, \ldots, \rho_R, \sigma_1/(1-\sigma_1)^2, \ldots, \sigma_K/(1-\sigma_K)^2$ (if we drop the population size constraint). There is some freedom in choosing the number of dominant stations. Working with more dominant stations yields a limit result that is more accurate but whose evaluation is more complex.
(2) Choose values for $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$ such that the dominant stations are indeed affected by the condition, and the remaining stations are not.
(3) Fix n, and choose values for $\boldsymbol{w}, \boldsymbol{c}, \beta$ such that the scaled network coincides with the network under consideration.

With these underlying ideas, one may choose scaling parameters as follows.

(1) Let A denote the set of dominant stations, which we construct as follows. Order the variances $\rho_1, \ldots, \rho_R, \sigma_1/(1-\sigma_1)^2, \ldots, \sigma_K/(1-\sigma_K)^2$ from high to low. Include in A the smallest number of stations with highest variance, such that

$$\sum_{r\in A} \rho_r + \sum_{k\in A} \frac{\sigma_k}{(1-\sigma_k)^2} \geqslant T \cdot \left(\sum_{r=1}^{R} \rho_r + \sum_{k=1}^{K} \frac{\sigma_k}{(1-\sigma_k)^2} \right)$$

for some threshold fraction $T \leqslant 1$. That is, choose the largest variances such that the sum of these variances makes up for at least a fraction T of the total variance. The best value of T may depend on the model at hand and the trade-off between accuracy and complexity discussed above. We empirically observed that picking $T = 0.8$ provides rather accurate approximations in most cases.
(2) For $r, k \in A$ and some $\gamma > 0$, choose values for ν_r, α_k such that $\frac{1}{2}\nu_r = 1 - \alpha_k = \gamma := \max\{\frac{1}{2}\nu_1, 1-\alpha_1\}$. It turns out that the value γ may be chosen arbitrarily, since each value leads to the same limit result (see the discussion below). For $r, k \notin A$, choose values for ν_r, α_k such that $\frac{1}{2}\nu_r < \gamma$ and $1 - \alpha_k < \gamma$ (the exact values are again irrelevant).
(3) Pick any value for n. Then choose values for $w_1, \ldots, w_R$ such that $\boldsymbol{\rho}^{(n)} = \boldsymbol{\rho}$, values for $c_1, \ldots, c_K$ such that $\boldsymbol{\sigma}^{(n)} = \boldsymbol{\sigma}$ and β such that $C_n = C$.

Despite the freedom in choosing scaling parameters, we underline that the decision for the set of dominant stations A completely determines the limit result. This can be verified with the following argument. For non-dominant stations, observe that the queue lengths converge to standard-normal and unit-rate exponential variables regardless of the scaling parameters. The queue-length distributions of the dominant stations on the other hand, depend on the condition $\sum_{r\in A} \sqrt{w_r}\mathscr{N}_r + \sum_{k\in A} \frac{1}{c_k}\mathscr{E}_k \leqslant \beta$. It therefore seems like the distributions of the dominant queue lengths depend on the values of $\boldsymbol{w}$, $\boldsymbol{c}$ and β. However, the identities $\boldsymbol{\rho}^{(n)} = \boldsymbol{\rho}$, $\boldsymbol{\sigma}^{(n)} = \boldsymbol{\sigma}$ and $C_n = C$ imply that $w_r = \rho_r n^{-2\gamma}$ for $r \in A$, $c_k = (\sigma_k^{-1} - 1)n^{\gamma}$ for $k \in A$ and $\beta = (C - \|\boldsymbol{\rho}\|)n^{-\gamma}$. With these scaling parameter values, the factor $n^{-\gamma}$ cancels out of the condition, which makes the values of n and γ irrelevant for the limit result.

We now move to the extended machine-repair model (described in Section 5) for a concrete numerical example of the steps above. Recall that in this model, the queue lengths of interest are the number of occupied machines and the number of broken machines. We denote these random variables by B_1 and D_1 respectively. The model has three parameters: the total number of machines C, the traffic load of the processing station ρ_1, and the traffic load of the repair station σ_1. Consider the three sets of parameter values shown in Table 3.

Table 4
Scaling parameter sets generating the same system as the parameter sets in Table 3.

	ν_1	α_1	w_1	c_1	β	n
Scaling parameter set 1	1	−0.24567	1	1	0.61	40
Scaling parameter set 2	1	0.5	1	1.054	1.1	90
Scaling parameter set 3	1	0.8526	1	1	1.1	90

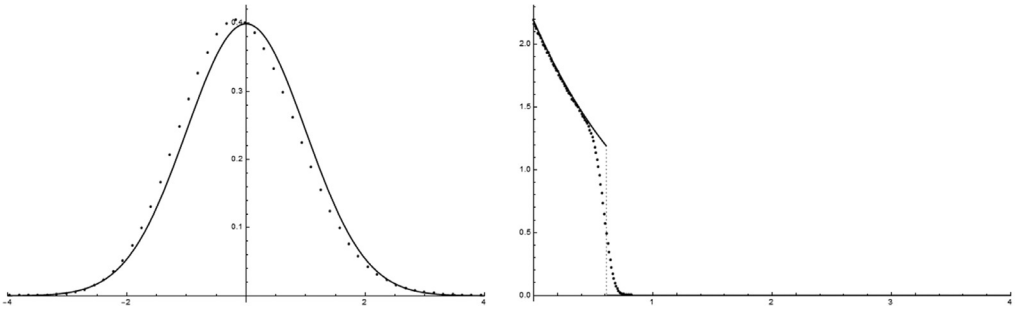

Fig. 6. Density functions of $\bar{B}_1^{(n)}$ (left) and $\bar{D}_1^{(n)}$ (right) for parameter set 1, both for $n = 40$ (dots) and for $n \to \infty$ (line).

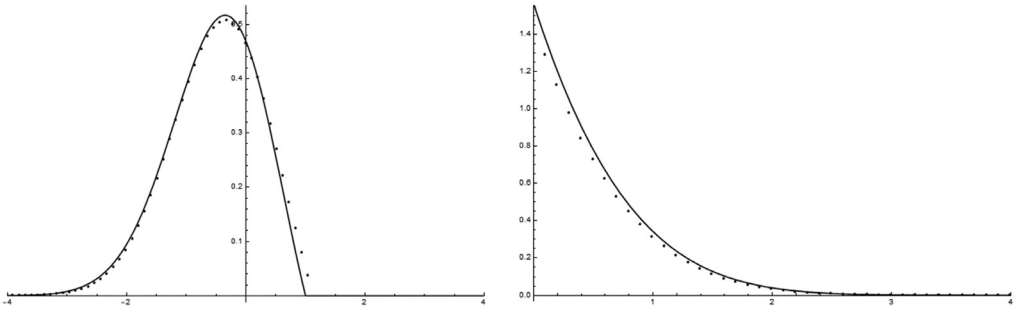

Fig. 7. Density functions of $\bar{B}_1^{(n)}$ (left) and $\bar{D}_1^{(n)}$ (right) for parameter set 2, both for $n = 90$ (dots) and for $n \to \infty$ (line).

Following the steps above, we find the scaling parameters given in Table 4. Note that by definition of our scaling regime, taking scaling parameter set 1 induces precisely the machine-repair model with parameter set 1. Hence, for this parameter set, comparing the actual queue-length distributions and our limit results amounts to comparing $n = 40$ and $n \to \infty$. The same holds for $n = 90$ and parameter sets 2 and 3.

In Figs. 6–8 we show plots of the density functions of $\bar{B}_1^{(n)}$ and $\bar{D}_1^{(n)}$ for each parameter set, obtained by simulation. For comparison, the densities are plotted against the limit results of Corollary 1.

Parameter set 1. Observe that $1 - \alpha_1 > \frac{1}{2}\nu_1$. Corollary 1 states in this case that

$$\bar{B}_1^{(n)} \to_{\mathrm{d}} \mathscr{N}_1, \qquad \bar{D}_1^{(n)} \to_{\mathrm{d}} (\mathscr{E}_1 \mid \mathscr{E}_1 \leqslant \beta)$$

as $n \to \infty$. Therefore, we have plotted in Fig. 6 the densities of $\bar{B}_1^{(40)}$ and $\bar{D}_1^{(40)}$ (obtained through simulation) against respectively a standard-normal density and a unit-rate exponential density truncated at β.

Parameter set 2. Observe that $1 - \alpha_1 = \frac{1}{2}\nu_1$. Corollary 1 states for $1 - \alpha_1 = \frac{1}{2}\nu_1$ that

$$(\bar{B}_1^{(n)}, \bar{D}_1^{(n)}) \to_{\mathrm{d}} (\mathscr{N}_1, \mathscr{E}_1 | \mathscr{N}_1 + \mathscr{E}_1 \leqslant \beta)$$

as $n \to \infty$. Therefore, Fig. 7 plots the densities of $\bar{B}_1^{(90)}$ and $\bar{D}_1^{(90)}$ against respectively the densities of $(\mathscr{N}_1 \mid \mathscr{N}_1 + \mathscr{E}_1 \leqslant \beta)$ and $(\mathscr{E}_1 \mid \mathscr{N}_1 + \mathscr{E}_1 \leqslant \beta)$.

Parameter set 3. Observe that $1 - \alpha_1 < \frac{1}{2}\nu_1$. Corollary 1 states in this case that

$$\bar{B}_1^{(n)} \to_{\mathrm{d}} (\mathscr{N}_1 \mid \mathscr{N}_1 \leqslant \beta), \qquad \bar{D}_1^{(n)} \to_{\mathrm{d}} \mathscr{E}_1$$

as $n \to \infty$. Therefore, Fig. 8 plots the densities of $\bar{B}_1^{(90)}$ and $\bar{D}_1^{(90)}$ against respectively a standard-normal density truncated at β and a unit-rate exponential density.

Figs. 6–8 show that for a network with $C = 100$, Corollary 1 provides rather accurate approximations of the queue-length densities. In relatively small networks there is the obvious alternative of direct evaluation of the product-form

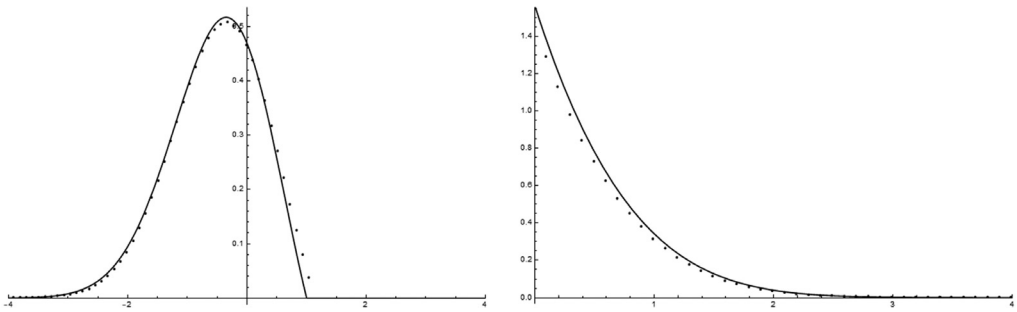

Fig. 8. Density functions of $\bar{B}_1^{(n)}$ (left) and $\bar{D}_1^{(n)}$ (right) for parameter set 3, both for $n = 90$ (dots) and for $n \to \infty$ (line).

density. For larger networks this will lead to computational issues, whereas the complexity of our asymptotic results is just mildly affected by the network size.

7. Discussion & further research

For a broad class of queueing networks, such as those of BCMP type, the joint queue-length distribution has a product-form structure. It may seem to lend itself well to numerical evaluation, but in case of closed networks the population size constraint makes this a non-trivial task. To overcome such computational issues, we have proposed a scaling regime, inspired by the Halfin–Whitt scaling. The corresponding limiting joint stationary queue-length distribution is transparent, numerically tractable and provides insight into the dependencies between the individual queue lengths. We point out how to map our scaling parameters on those of the queueing network under consideration. A series of numerical experiments shows that the resulting approximations are close to the true (pre-limit, that is) values.

Scaling methods in queueing networks form a rich research area in which there is still ample room to extend our current results. One option is to include multi-server stations in the network. As the queue lengths become very large in our scaling regime, we expect that such a station would effectively behave as a single-server station, with the service rate multiplied by the number of servers. A formal proof may be challenging.

Another model extension preserving product form relates to multiclass networks. In these models customers may be of different classes, where each class may have its specific routing and service requirements. The product form of the stationary distribution is preserved under class-dependent routing probabilities and, for certain station types, under class-dependent service requirements. Many queueing network results apply to multiclass networks, but scaling analysis becomes more involved, primarily because each customer class now has its own population size.

Further research efforts could focus on exploiting our scaling results for design and optimization purposes. In addition, as we have indicated, our scaling method provides freedom in relation to the choice of the scaling parameters, which raises the question how to choose the entries of $\boldsymbol{w}$, $\boldsymbol{\nu}$, $\boldsymbol{c}$, $\boldsymbol{\alpha}$ and β so as to maximally accurately represent the underlying queueing network.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research in this paper is supported by the Netherlands Organisation for Scientific Research (NWO) through Gravitation-grant NETWORKS-024.002.003.

Appendix A. Proofs of Section 4

Lemma 4. *Suppose $x_n \to \infty$ as $n \to \infty$. If $(C_n - x_n)/\sqrt{x_n} \to Q$ with $Q \in [-\infty, \infty]$, then $\mathbb{P}(\mathscr{P}(x_n) \leqslant C_n) \to \Phi(Q)$ as $n \to \infty$.*

Proof. Observe that a Poisson random variable with mean $m \in \mathbb{N}$ can be written as a sum of m Poisson random variables with mean 1. Therefore, with $(X_i)_{i\in\mathbb{N}}, (Y_i)_{i\in\mathbb{N}}$ i.i.d. copies of $\mathscr{P}(1)$,

$$\sum_{i=1}^{\lfloor x_n \rfloor} X_i =_{\mathrm{d}} \mathscr{P}(\lfloor x_n \rfloor) \leqslant_{\mathrm{st}} \mathscr{P}(x_n) \leqslant_{\mathrm{st}} \mathscr{P}(\lceil x_n \rceil) =_{\mathrm{d}} \sum_{i=1}^{\lceil x_n \rceil} Y_i.$$

Subtracting x_n and dividing by $\sqrt{x_n}$ yields

$$\frac{1}{\sqrt{x_n}}\sum_{i=1}^{\lfloor x_n\rfloor}(X_i-1)-\frac{x_n-\lfloor x_n\rfloor}{\sqrt{x_n}}\leqslant_{\rm st}\frac{\mathscr{P}(x_n)-x_n}{\sqrt{x_n}}\leqslant_{\rm st}\frac{1}{\sqrt{x_n}}\sum_{i=1}^{\lceil x_n\rceil}(Y_i-1)+\frac{\lceil x_n\rceil-x_n}{\sqrt{x_n}}.$$

Appropriately rewritten as

$$\frac{1}{\sqrt{\lfloor x_n\rfloor}+O(1)}\sum_{i=1}^{\lfloor x_n\rfloor}(X_i-1)-\frac{O(1)}{\sqrt{x_n}}\leqslant_{\rm st}\frac{\mathscr{P}(x_n)-x_n}{\sqrt{x_n}}\leqslant_{\rm st}\frac{1}{\sqrt{\lceil x_n\rceil}-O(1)}\sum_{i=1}^{\lceil x_n\rceil}(Y_i-1)+\frac{O(1)}{\sqrt{x_n}},$$

we may apply the central limit theorem to conclude that $(\mathscr{P}(x_n)-x_n)/\sqrt{x_n}$ converges to a standard-normal random variable as $x_n\to\infty$. Using this observation the result immediately follows from the fact that

$$\mathbb{P}\left(\mathscr{P}(x_n)\leqslant C_n\right)=\mathbb{P}\left(\frac{\mathscr{P}(x_n)-x_n}{\sqrt{x_n}}\leqslant\frac{C_n-x_n}{\sqrt{x_n}}\right)\to\Phi(Q)$$

as $x_n\to\infty$. □

In the following proofs we write ρ_r, ζ_r, σ_l and δ_l for $\rho_r^{(n)}$, $\zeta_r^{(n)}(s_r)$, $\sigma_l^{(n)}$ and $\delta_l^{(n)}(t_l)$ to simplify the notation.

Lemma 5. *As $n\to\infty$,*

$$f^{(n)}(\boldsymbol{s})\to\begin{cases}1 & \text{if}\quad 1-\alpha_1>\frac{1}{2}\nu_1,\\ \Phi\left(\lambda(\boldsymbol{s})\right) & \text{if}\quad 1-\alpha_1\leqslant\frac{1}{2}\nu_1.\end{cases}$$

Proof. We start with the case $1-\alpha_1>\frac{1}{2}\nu_1$. Note that in this case

$$C_n=\left\lfloor\sum_{r=1}^{R}\rho_r+\beta n^{1-\alpha_1}\right\rfloor.$$

To prove $f^{(n)}(\boldsymbol{s})\to 1$, we apply Lemma 4 with

$$x_n=\sum_{r=1}^{R}\zeta_r=\sum_{r=1}^{R}\left(\rho_r-s_r\sqrt{\rho_r}+o\left(\sqrt{\rho_r}\right)\right)$$

(so that $Q=\infty$).

For the limit of $f^{(n)}(\boldsymbol{s})$ in case $1-\alpha_1\leqslant\frac{1}{2}\nu_1$, an application of Lemma 4 with

$$x_n=\sum_{r=1}^{R}\zeta_r=\sum_{r=1}^{R}\left(\rho_r-s_r\sqrt{\rho_r}+o\left(\sqrt{\rho_r}\right)\right)$$

and

$$C_n=\left\lfloor\sum_{r=1}^{R}\rho_r+\beta n^{\frac{1}{2}\nu_1}\right\rfloor$$

leads to

$$Q=\lim_{n\to\infty}\frac{C_n-x_n}{\sqrt{x_n}}=\lim_{n\to\infty}\frac{\beta n^{\frac{1}{2}\nu_1}+\sum_{r=1}^{R}\left(s_r\sqrt{\rho_r}+o\left(\sqrt{\rho_r}\right)\right)}{\sqrt{\sum_{r=1}^{R}\left(\rho_r+o\left(\rho_r\right)\right)}}=\frac{\beta+\sum_{r=1}^{R^-}s_r\sqrt{w_r}}{\sqrt{\sum_{r=1}^{R^-}w_r}}=\lambda(\boldsymbol{s}).$$

Recall that R^- is defined as the largest integer such that $\nu_1=\cdots=\nu_{R^-}$. Hence, if $1-\alpha_1\leqslant\frac{1}{2}\nu_1$, then $f^{(n)}(\boldsymbol{s})\to\Phi\left(\lambda(\boldsymbol{s})\right)$. □

Lemma 6. *As $n\to\infty$,*

$$g_l^{(n)}(\boldsymbol{s},t_l)\to\begin{cases}1 & \text{if}\quad 1-\alpha_1=1-\alpha_l>\frac{1}{2}\nu_1,\\ \Phi\left(\lambda(\boldsymbol{s})-c_l(1+t_l)\sqrt{W}\right) & \text{if}\quad 1-\alpha_1=1-\alpha_l=\frac{1}{2}\nu_1.\end{cases}$$

Proof. The proof for $g_l^{(n)}(\boldsymbol{s},t_l)$ is similar to the proof for $f^{(n)}(\boldsymbol{s})$. Suppose first that $1-\alpha_1=1-\alpha_l>\frac{1}{2}\nu_1$. Note that with

$$x_n=\sum_{r=1}^{R}\zeta_r/\delta_l=\sum_{r=1}^{R}\rho_r\frac{n+c_ln^{\alpha_l}}{n}\exp(-s_r/\sqrt{\rho_r})\,e^{t_l(1-\sigma_l)},$$

we have

$$\frac{C_n - x_n}{\sqrt{x_n}} = \frac{\sum_{r=1}^R \rho_r + \frac{\beta}{c_1} n^{1-\alpha_1} - \sum_{r=1}^R \left(\rho_r + O(n^{\nu_r+\alpha_l-1}) - O(n^{\frac{1}{2}\nu_r})\right)}{\sqrt{\sum_{r=1}^R (\rho_r + o(\rho_r))}}.$$

When $1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1$ we take $Q=\infty$ in Lemma 4, concluding that $g_l^{(n)}(\boldsymbol{s}, t_l) \to 1$.

Next, for $g_l^{(n)}(\boldsymbol{s}, t_l)$ as $1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1$, an application of Lemma 4 with

$$x_n = \sum_{r=1}^R \zeta_r/\delta_l = \sum_{r=1}^R \rho_r \frac{n + c_l n^{\alpha_l}}{n} \exp(-s_r/\sqrt{\rho_r})\, e^{t_l(1-\sigma_l)}$$

and

$$C_n = \left\lfloor \sum_{r=1}^R \rho_r + \beta n^{\frac{1}{2}\nu_1} \right\rfloor$$

leads to

$$Q = \lim_{n\to\infty} \frac{C_n - x_n}{\sqrt{x_n}} = \lim_{n\to\infty} \frac{\sum_{r=1}^R \rho_r + \beta n^{\frac{1}{2}\nu_1} - \sum_{r=1}^R \left(\rho_r + c_l(1+t_l)n^{\alpha_l-1}\rho_r - s_r\sqrt{\rho_r}\right) + o(\sqrt{\rho_r})}{\sqrt{\sum_{r=1}^R (\rho_r + o(\rho_r))}}$$

$$= \frac{\beta + \sum_{r=1}^{R^-} \left(s_r\sqrt{w_r} - c_l(1+t_l)w_r\right)}{\sqrt{\sum_{j=1}^{R^-} w_r}} = \lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}.$$

We conclude that if $1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1$, then $g_l^{(n)}(\boldsymbol{s}, t_l) \to \Phi(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W})$. □

Lemma 7. *As $n\to\infty$,*

$$h_l^{(n)}(\boldsymbol{s}, t_l) \to \begin{cases} 0 & \text{if} \quad 1-\alpha_1 > 1-\alpha_l \text{ and } 1-\alpha_1 \geqslant \nu_1, \\ 0 & \text{if} \quad \nu_1 > 1-\alpha_1 > 1-\alpha_l \geqslant \frac{1}{2}\nu_1, \\ \exp\left(-\beta c_l(1+t_l)\right) & \text{if} \quad 1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1, \\ \dfrac{\phi(\lambda(\boldsymbol{s}))}{\phi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right)} & \text{if} \quad 1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1. \end{cases}$$

Proof. Let $H_l^{(n)}(\boldsymbol{s}, t_l) := \ln(h_l^{(n)}(\boldsymbol{s}, t_l))$ and observe that $H_l^{(n)}(\boldsymbol{s}, t_l)$ is a sum of the two components $\zeta^{(n)}(\boldsymbol{s})(1/\delta_l(t_l) - 1)$ and $(C_n+1)\ln(\delta_l(t_l))$. We explicitly consider these two components separately. In the following calculations, terms irrelevant as $n\to\infty$ will be dealt with using the '$\sim$' symbol. From the first component of $H_l^{(n)}(\boldsymbol{s}, t_l)$ we extract the leading terms by applying Taylor expansions. As $n\to\infty$, this component can be rewritten as

$$\begin{aligned} \sum_{r=1}^R \rho_r e^{-s_r/\sqrt{\rho_r}} \left(\frac{e^{t_l(1-\sigma_l)}}{\sigma_l} - 1\right) &\sim \sum_{r=1}^R \rho_r(1 - s_r/\sqrt{\rho_r}) \left(\frac{1}{\sigma_l} - 1 + \frac{t_l(1-\sigma_l)}{\sigma_l} + \frac{\frac{1}{2}t_l^2(1-\sigma_l)^2}{\sigma_l}\right) \\ &\sim \sum_{r=1}^R w_r n^{\nu_r}(1 - s_r/\sqrt{w_r n^{\nu_r}}) \left(c_l n^{\alpha_l-1} + t_l c_l n^{\alpha_l-1} + \frac{1}{2}c_l^2 t_l^2 n^{2\alpha_l-2}\right) \\ &\sim \sum_{r=1}^R \left(w_r c_l(1+t_l) n^{\alpha_l-1+\nu_r} - \sqrt{w_r} c_l(1+t_l) s_r n^{\alpha_l-1+\frac{1}{2}\nu_r} + \frac{1}{2} w_r c_l^2 t_l^2 n^{2\alpha_l-2+\nu_r}\right). \end{aligned} \tag{24}$$

We continue by considering the second component of $H_l^{(n)}(\boldsymbol{s}, t_l)$. Defining $\tau_l(t_l) := 1 - \delta_l(t_l) = 1 - \sigma_l e^{-t_l(1-\sigma_l)} = (1 - e^{-t_l(1-\sigma_l)}) + (1-\sigma_l)e^{-t_l(1-\sigma_l)}$ and using that $1-\sigma_l \sim c_l n^{\alpha_l-1} - c_l^2 n^{2\alpha_l-2}$, we have for this component that

$$(C_n+1)\ln(1-\tau_l(t_l)) = -(C_n+1)\left(\tau_l(t_l) + \frac{1}{2}\tau_l(t_l)^2 + o(\tau_l(t_l)^2)\right) = -C_n\left(\tau_l(t_l) + \frac{1}{2}\tau_l(t_l)^2 + o(\tau_l(t_l)^2)\right) + o(1).$$

Observing that $\tau_l(t_l) \sim (1+t_l)(1-\sigma_l) - (\frac{1}{2}t_l^2 + t_l)(1-\sigma_l)^2$, the second component is thus asymptotically equivalent to

$$\begin{aligned} -C_n\left((1+t_l)(1-\sigma_l) + \frac{1}{2}(1-\sigma_l)^2 + o(1-\sigma_l)^2\right) &\sim -C_n\left((1+t_l)(c_l n^{\alpha_l-1} - c_l^2 n^{2\alpha_l-2}) + \frac{1}{2}c_l^2 n^{2\alpha_l-2}\right) \\ &= -C_n\left((1+t_l)c_l n^{\alpha_l-1} - (\frac{1}{2} + t_l)c_l^2 n^{2\alpha_l-2}\right). \end{aligned} \tag{25}$$

When adding the two components (24) and (25), we conclude that $h_l^{(n)}(\boldsymbol{s}, t_l) = \exp(H_l^{(n)}(\boldsymbol{s}, t_l))$ as $n \to \infty$, with

$$\begin{aligned}
H_l^{(n)}(\boldsymbol{s}, t_l) &\sim (1+t_l)c_l n^{\alpha_l - 1}\left(\sum_{r=1}^{R}\left(w_r n^{\nu_r} - \sqrt{w_r}s_r n^{\frac{1}{2}\nu_r}\right) - C_n\right) \\
&\quad + c_l^2 n^{2\alpha_l - 2}\left(\sum_{r=1}^{R}\frac{1}{2}w_r t_l^2 n^{\nu_r} + C_n(t_l + \frac{1}{2})\right) \\
&= -(1+t_l)c_l n^{\alpha_l - 1}\left(\sum_{r=1}^{R}\sqrt{w_r}s_r n^{\frac{1}{2}\nu_r} + \left(C_n - \sum_{r=1}^{R} w_r n^{\nu_r}\right)\right) \\
&\quad + c_l^2 n^{2\alpha_l - 2}\left(\frac{1}{2}(1+t_l)^2\sum_{r=1}^{R} w_r n^{\nu_r} + (\frac{1}{2} + t_l)\left(C_n - \sum_{r=1}^{R} w_r n^{\nu_r}\right)\right) \\
&= -(1+t_l)c_l n^{\alpha_l - 1}\left(\sum_{r=1}^{R}\sqrt{w_r}s_r n^{\frac{1}{2}\nu_r} + \beta n^{\gamma}\right) \\
&\quad + c_l^2 n^{2\alpha_l - 2}\left(\frac{1}{2}(1+t_l)^2\sum_{r=1}^{R} w_r n^{\nu_r} + (\frac{1}{2} + t_l)\beta n^{\gamma}\right). \qquad (26)
\end{aligned}$$

We consider (26) as a reference point from now on, and distinguish four cases:

(1) Suppose that $1-\alpha_1 > 1-\alpha_l$ and $1-\alpha_1 \geqslant \nu_1$. Then $\gamma = 1-\alpha_1 \geqslant \nu_1$, so (26) has leading term $-c_l(1+t_l)\beta n^{\alpha_l - 1 + \gamma}$, which tends to $-\infty$ as $n \to \infty$. Hence, in this case $h_l^{(n)}(\boldsymbol{s}, t_l) \to 0$ as $n \to \infty$.

(2) If $\nu_1 > 1-\alpha_1 > 1-\alpha_l \geqslant \frac{1}{2}\nu_1$, then $n^{\alpha_l - 1 + \frac{1}{2}\nu_r} = O(1)$, hence

$$\begin{aligned}
h_l^{(n)}(\boldsymbol{s}, t_l) &= \exp\left(-(1+t_l)\left(O(1) + \beta c_l n^{1-\alpha_1} n^{\alpha_l - 1}\right)\right) \cdot \exp\left(O(1) + \beta n^{1-\alpha_1}(\frac{1}{2} + t_l)c_l^2 n^{2\alpha_l - 2}\right) \\
&= \exp\left(-(1+t_l)\beta c_l n^{\alpha_l - \alpha_1} + O(1)\right) \to 0.
\end{aligned}$$

(3) If $1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1$, then $n^{\alpha_l - 1 + \frac{1}{2}\nu_1} = o(1)$, so with (26) we have

$$\begin{aligned}
h_l^{(n)}(\boldsymbol{s}, t_l) &= \exp\left(-(1+t_l)\left(O(n^{\alpha_l - 1 + \frac{1}{2}\nu_1}) + \beta n^{1-\alpha_1} c_l n^{\alpha_l - 1}\right)\right) \cdot \exp\left(O(n^{2\alpha_l - 2 + \nu_1}) + O(n^{2\alpha_l - 2 - (\alpha_1 - 1)})\right) \\
&= \exp\left(-(1+t_l)\beta c_l + o(1)\right) \to \exp\left(-\beta c_l(1+t_l)\right).
\end{aligned}$$

(4) Finally, if $1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1$, then with (26),

$$\begin{aligned}
h_l^{(n)}(\boldsymbol{s}, t_l) &= \exp\left(-c_l(1+t_l)\left(\sum_{r=1}^{R}\sqrt{w_r}s_r n^{\alpha_l - 1 + \frac{1}{2}\nu_r} + \beta\right) + \frac{1}{2}c_l^2(1+t_l)^2\sum_{r=1}^{R} w_r n^{2\alpha_l - 2 + \nu_r} + O(n^{2\alpha_l - 2 + \frac{1}{2}\nu_1})\right) \\
&\to \exp\left(-c_l(1+t_l)\left(\beta + \sum_{r=1}^{R^-} s_r\sqrt{w_r}\right) + \frac{1}{2}c_l^2\,(1+t_l)^2\sum_{r=1}^{R^-} w_r\right) \\
&= \exp\left(-c_l(1+t_l)\sqrt{W}\lambda(\boldsymbol{s}) + \frac{1}{2}c_l^2\,(1+t_l)^2\,W\right) \\
&= \exp\left(\frac{1}{2}\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right)^2 - \frac{1}{2}\lambda(\boldsymbol{s})^2\right) \\
&= \frac{\phi\left(\lambda(\boldsymbol{s})\right)}{\phi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right)}.
\end{aligned}$$

This completes the proof of Lemma 7. □

Lemma 8. *If $1-\alpha_1 < \nu_1$ and $1-\alpha_l < \frac{1}{2}\nu_1$, then $g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l) \to 0$ as $n \to \infty$.*

Proof. Let

$$x_l^{(n)} = \zeta^{(n)}(\boldsymbol{s})/\delta_l^{(n)}(t_l) = \sum_{r=1}^{R} \frac{\rho_r^{(n)}}{\sigma_l^{(n)}} \exp(-s_r/\sqrt{\rho_r^{(n)}})\, e^{t_l(1-\sigma_l^{(n)})};$$

in the sequel we write just x_n for brevity. In this proof, our first objective is to identify the asymptotics of $g_l^{(n)}(\boldsymbol{s}, t_l) = \mathbb{P}(\mathscr{P}(x_n) \leqslant C_n)$. To this end, let $P_n =_{\mathrm{d}} \mathscr{P}(x_n)$, and let $\mathbb{Q}$ be an alternative measure under which this Poisson random variable has mean C_n, such that

$$g_l^{(n)}(\boldsymbol{s}, t_l) = \mathbb{E}_{\mathbb{P}}\left(\mathbb{1}\{P_n \leqslant C_n\}\right) = \mathbb{E}_{\mathbb{Q}}\left(L\,\mathbb{1}\{P_n \leqslant C_n\}\right),$$

with L denoting the likelihood ratio or Radon–Nikodym derivative

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \left(\frac{e^{-x_n}(x_n)^{P_n}}{P_n!}\right) \Big/ \left(\frac{e^{-C_n}(C_n)^{P_n}}{P_n!}\right) = e^{C_n - x_n}\left(\frac{x_n}{C_n}\right)^{P_n}.$$

We thus arrive at

$$g_l^{(n)}(\boldsymbol{s}, t_l) = e^{C_n - x_n}\mathbb{E}_{\mathbb{Q}}\left((x_n/C_n)^{P_n}\ \mathbb{1}\{P_n \leqslant C_n\}\right).$$

Define $\bar{P}_n := (P_n - C_n)/\sqrt{C_n}$ and recall that, by the central limit theorem, the distribution of $\bar{P}_n$ converges to a standard-normal distribution. In terms of this new random variable, we have

$$g_l^{(n)}(\boldsymbol{s}, t_l) = e^{C_n - x_n}\left(\frac{x_n}{C_n}\right)^{C_n} q_n, \tag{27}$$

where

$$q_n := \mathbb{E}_{\mathbb{Q}}\left(\left((x_n/C_n)^{\sqrt{C_n}}\right)^{\bar{P}_n}\ \mathbb{1}\{\bar{P}_n \leqslant 0\}\right) = \int_{-\infty}^{0}\left(\left(\frac{x_n}{C_n}\right)^{\sqrt{C_n}}\right)^{y} \mathrm{d}F_{\bar{P}_n}(y), \tag{28}$$

with $F_{\bar{P}_n}(y)$ being the distribution function of $\bar{P}_n$. The idea is to show that $F_{\bar{P}_n}(y)$ behaves as a standard-normal distribution for sufficiently large C_n, and hence that

$$q_n \sim \int_{-\infty}^{0}\left(\left(\frac{x_n}{C_n}\right)^{\sqrt{C_n}}\right)^{y} \phi(y)\,\mathrm{d}y,$$

where $\phi(y)$ is the standard-normal density function in y. To formally achieve this, we bound $F_{\bar{P}_n}(y)$ using the Berry–Esseen theorem. This states that for all C_n large enough,

$$\sup_y \left|F_{\bar{P}_n}(y) - \Phi(y) - \frac{m_3}{6\sqrt{C_n}}(1-y^2)\phi(y) - \phi(y)l(y)\right| = O\left(\frac{1}{\sqrt{C_n}}\right),$$

where m_3 is the third moment of a Poisson(1) random variable and $l(\cdot)$ is a function that is bounded by a constant times $1/\sqrt{C_n}$.

We proceed by analyzing q_n using the Berry–Esseen theorem. Observe that (28) contains the density of $\bar{P}_n$, whereas 'Berry–Esseen' concerns a bound in terms of the corresponding distribution function. Therefore, we apply integration by parts, yielding

$$\begin{aligned} q_n &= \int_{-\infty}^{0}\left(\left(\frac{x_n}{C_n}\right)^{\sqrt{C_n}}\right)^{y} \mathrm{d}F_{\bar{P}_n}(y) = \int_{-\infty}^{0} e^{a_n y}\frac{\mathrm{d}}{\mathrm{d}y}\left(F_{\bar{P}_n}(y) - F_{\bar{P}_n}(0)\right)\mathrm{d}y \\ &= -\int_{-\infty}^{0} a_n e^{a_n y}\left(F_{\bar{P}_n}(y) - F_{\bar{P}_n}(0)\right)\mathrm{d}y, \end{aligned} \tag{29}$$

where $a_n := \sqrt{C_n}\ln(x_n/C_n)$. Now applying the Berry–Esseen bound in (29),

$$\begin{aligned} q_n = &-\int_{-\infty}^{0} a_n e^{a_n y}\left(\Phi(y) - \Phi(0)\right)\mathrm{d}y + \int_{-\infty}^{0} a_n e^{a_n y}\frac{m_3}{6\sqrt{C_n}}\left((1-y^2)\phi(y) - \phi(0)\right)\mathrm{d}y \\ &+ \int_{-\infty}^{0} a_n e^{a_n y}\left(\phi(y)l(y) - \phi(0)l(0)\right)\mathrm{d}y + \int_{-\infty}^{0} a_n e^{a_n y}\cdot O\left(\frac{1}{\sqrt{C_n}}\right)\mathrm{d}y. \end{aligned}$$

Recall that $l(y) = O\left(1/\sqrt{C_n}\right)$, so the last three integrals contain a term of that order. For the first integral, we integrate by parts once more to obtain

$$q_n = \int_{-\infty}^{0} e^{a_n y}\phi(y)\,dy + \frac{1}{\sqrt{C_n}}\int_{-\infty}^{0} a_n e^{a_n y} r(y)\,dy,$$

where $r(y)$ is bounded by a quadratic function. Observe that the second integral converges, so the second term is $O(1/\sqrt{C_n})$. Therefore, completing the square in the exponent,

$$\begin{aligned} q_n &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(a_n y - \frac{y^2}{2}\right) \mathrm{d}y + O\left(\frac{1}{\sqrt{C_n}}\right) \\ &= \exp\left(\frac{a_n^2}{2}\right) \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-a_n)^2\right) \mathrm{d}y + O\left(\frac{1}{\sqrt{C_n}}\right) \\ &= \exp\left(\frac{a_n^2}{2}\right) \Phi(-a_n) + O\left(\frac{1}{\sqrt{C_n}}\right) \\ &= \exp\left(\frac{a_n^2}{2}\right) (1-\Phi(a_n)) + O\left(\frac{1}{\sqrt{C_n}}\right) \end{aligned} \tag{30}$$

using the symmetry of the normal distribution. A known property of the tail of the normal distribution is

$$e^{x^2/2}\ (1-\Phi(x)) \sim \frac{1}{x\sqrt{2\pi}} \tag{31}$$

as $x \to \infty$ (cf. [24, p. 175]). To apply this property to the first term on the right-hand side of (30), it is necessary to verify that a_n goes to ∞ as $n \to \infty$. This can be seen by relying on a Taylor expansion, and recalling that $1-\alpha_1 < \nu_1$ and $1-\alpha_l < \frac{1}{2}\nu_1$:

$$\begin{aligned} a_n = \sqrt{C_n} \ln \frac{x_n}{C_n} &= \sqrt{\sum_{r=1}^{R} w_r n^{\nu_r} + o(n^{\nu_1})} \cdot \ln\left(\frac{\sum_{r=1}^{R} w_r n^{\nu_r}(1+c_l n^{\alpha_l-1}) e^{-\frac{s_r}{\sqrt{w_r n^{\nu_r}}}} e^{t_l(1-\sigma_l^{(n)})}}{\sum_{r=1}^{R} w_r n^{\nu_r} + o(n^{\nu_1})}\right) \\ &= \Omega(n^{\frac{1}{2}\nu_1}) \ln\left(1+\Omega\left(n^{\alpha_l-1}\right)\right) = \Omega(n^{\frac{1}{2}\nu_1-(1-\alpha_l)}) \to \infty \end{aligned}$$

as $n \to \infty$, where $\Omega(u_n)$ denotes a sequence v_n such that for some constant $c > 0$ it holds that $\lim_{n\to\infty} v_n/u_n \geqslant c$. Using property (31) in (30), and substituting the result in (27), we thus obtain that, as $n \to \infty$,

$$g_l^{(n)}(\boldsymbol{s}, t_l) = e^{C_n - x_n}\left(\frac{x_n}{C_n}\right)^{C_n}\left(\frac{1}{a_n\sqrt{2\pi}} + O\left(\frac{1}{\sqrt{C_n}}\right)\right) = e^{C_n-x_n}\left(\frac{x_n}{C_n}\right)^{C_n} \cdot o(1).$$

Multiplying with $h_l^{(n)}(\boldsymbol{s}, t_l)$, and using $\delta_l^{(n)}(t_l)x_n = \zeta^{(n)}(\boldsymbol{s})$, it holds that

$$g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l) = e^{C_n-x_n}\left(\frac{x_n}{C_n}\right)^{C_n} \cdot o(1) \cdot e^{x_n-\zeta^{(n)}(\boldsymbol{s})}\left(\delta_l^{(n)}(t_l)\right)^{C_n+1} = o(1) \cdot e^{C_n-\zeta^{(n)}(\boldsymbol{s})}\left(\frac{\zeta^{(n)}(\boldsymbol{s})}{C_n}\right)^{C_n}.$$

The stated result now follows from writing all terms as exponentials and applying the Taylor expansion to the logarithm:

$$\begin{aligned} g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l) &= o(1) \exp\left(C_n - \zeta^{(n)}(\boldsymbol{s}) + C_n \ln\left(1+\left(\frac{\zeta^{(n)}(\boldsymbol{s})-C_n}{C_n}\right)\right)\right) \\ &= o(1) \exp\left(C_n - \zeta^{(n)}(\boldsymbol{s}) + C_n \frac{\zeta^{(n)}(\boldsymbol{s})-C_n}{C_n} + O(1)\right) \to 0, \end{aligned}$$

as $n \to \infty$. □

Appendix B. Proof of Corollary 3

Proof of Corollary 3. This proof mimics the proof of Theorem 1. For convenience we write

$$T_n(\boldsymbol{s}, \boldsymbol{t}) := \mathbb{E}\left(\prod_{r=1}^{I-1} e^{-s_r \bar{B}_r^{(n)}} \prod_{m=I}^{R} e^{-s_m B_m^{(n)}} \prod_{k=1}^{J-1} e^{-t_k \bar{D}_k^{(n)}} \prod_{l=J}^{K} e^{-t_l D_l^{(n)}}\right),$$

which differs from $P_n(\boldsymbol{s}, \boldsymbol{t})$ in the fact that $B_I^{(n)}, \ldots, B_R^{(n)}$ and $D_J^{(n)}, \ldots, D_K^{(n)}$ are unscaled. We now follow the line of the proof of Lemma 3, with a few adjustments for the unscaled random variables:

- the factors $e^{s_m\sqrt{\rho_m^{(n)}}}$ are removed, for $m = I, \ldots, R$,
- the variables $\zeta_m^{(n)}(s_m)$ are defined as $\rho_m^{(n)} e^{-s_m}$ rather than $\rho_m^{(n)} e^{-s_m/\sqrt{\rho_m^{(n)}}}$, for $m = I, \ldots, R$,
- $t_l(1-\sigma_l)$ is replaced by t_l, for $l = J, \ldots, K$.

Hence, $T_n(\boldsymbol{s},\boldsymbol{t})$ equals the right-hand side of (18) subject to the adjustments above. The first term of $T_n(\boldsymbol{s},\boldsymbol{t})$ then equals

$$\prod_{r=1}^{I-1} e^{-\rho_r^{(n)}+s_r\sqrt{\rho_r^{(n)}}+\zeta_r^{(n)}(s_r)} \prod_{m=I}^{R} e^{\rho_m^{(n)}(e^{-s_m}-1)}.$$

In this expression, the first $I-1$ factors are as in (20) and converge to standard-normal LSTs as $n\to\infty$. We recognize the latter $R-I+1$ factors as LSTs of Poisson random variables with mean $\rho_m^{(n)}$.

The second term of $T_n(\boldsymbol{s},\boldsymbol{t})$ equals

$$\prod_{k=1}^{J-1} \frac{1-\sigma_k^{(n)}}{1-\delta_k^{(n)}(t_k)} \prod_{l=J}^{K} \frac{1-\sigma_l^{(n)}}{1-\sigma_l^{(n)}e^{-t_l}}.$$

With (21), the first $J-1$ factors of this expression converge to LSTs of unit-rate exponential random variables as $n\to\infty$, and the second $K-J+1$ are easily identified as geometric LSTs with parameter $1-\sigma_l^{(n)}$.

Following the proofs of Lemmas 5–8, it can be seen that the adjustments do not change the asymptotic behavior of the last term of $T_n(\boldsymbol{s},\boldsymbol{t})$. Therefore, the result follows from Theorem 1 and recognizing known LSTs in the first two terms of $T_n(\boldsymbol{s},\boldsymbol{t})$ as described above. □

Appendix C. Proof of Corollary 1

Proof of Corollary 1. This proof amounts to verifying that the LST corresponding to $(\mathscr{N},\mathscr{E}\,|\,Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta)$ equals the right-hand side of (13). This can be done with standard integration techniques. In this section we illustrate the proof for the case that $1-\alpha_1=\frac{1}{2}\nu_1$. We leave out the other two cases, as these can be verified using the precise same steps.

Rather than the LST of $(\mathscr{N},\mathscr{E}\,|\,Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta)$, we consider in this section the LST of $(\mathscr{N}_{R^-},\mathscr{E}_{K^-}\,|\,Z(\mathscr{N}_{R^-},\mathscr{E}_{K^-})\leqslant\beta)$, which we call $Q(\boldsymbol{s},\boldsymbol{t})$. The former LST can simply be obtained through multiplying $Q(\boldsymbol{s},\boldsymbol{t})$ by $R-R^-$ standard-normal LSTs and $K-K^-$ unit-rate exponential LSTs. For $1-\alpha_1=\frac{1}{2}\nu_1$, observe that the right-hand side of (13) equals

$$\left(\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2}\right)\left(\prod_{k=1}^{K^-}\frac{1}{1+t_k}\right)\cdot\frac{\phi(\lambda(\boldsymbol{s}))}{\phi(\lambda(\boldsymbol{0}))}\cdot\frac{\Psi\left(\lambda(\boldsymbol{s})\right)-\sum_{l=1}^{K^-}\left(\prod_{j=1,\,j\neq l}^{K^-}\kappa_{jl}(\boldsymbol{t})\right)\Psi\left(\lambda(\boldsymbol{s})-c_l(1+t_l)\sqrt{W}\right)}{\Psi\left(\lambda(\boldsymbol{0})\right)-\sum_{l=1}^{K^-}\left(\prod_{j=1,\,j\neq l}^{K^-}\kappa_{jl}(\boldsymbol{0})\right)\Psi\left(\lambda(\boldsymbol{0})-c_l\sqrt{W}\right)}. \tag{32}$$

Our objective is to show that $Q(\boldsymbol{s},\boldsymbol{t})$ equals (32).

To simplify the notation, we write

$$p:=\mathbb{P}\left(\sum_{r=1}^{R^-}\sqrt{w_r}\mathscr{N}_r+\sum_{k=1}^{K^-}\frac{1}{c_k}\mathscr{E}_k\leqslant\beta\right),\quad \widehat{b}:=\sum_{r=1}^{R^-}b_r\sqrt{w_r},\quad\text{and}\quad \widehat{d}_j:=\sum_{k=1}^{j}d_k/c_k.$$

By definition of the LST, and with $\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_k)$ denoting an abbreviation for $\mathrm{d}b_1\ldots\mathrm{d}b_{R^-}\,\mathrm{d}d_1\ldots\mathrm{d}d_k$, it follows that

$$\begin{aligned}
Q(\boldsymbol{s},\boldsymbol{t}) &= \mathbb{E}\left(\left(\prod_{r=1}^{R^-}e^{-s_r\mathscr{N}_r}\right)\left(\prod_{k=1}^{K^-}e^{-t_k\mathscr{E}_k}\right)\,\Big|\,\sum_{r=1}^{R^-}\sqrt{w_r}\mathscr{N}_r+\sum_{k=1}^{K^-}\frac{1}{c_k}\mathscr{E}_k\leqslant\beta\right)\\
&= \int_{\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-}:\,\widehat{b}+\widehat{d}_{K^-}\leqslant\beta}\left(\prod_{r=1}^{R^-}e^{-s_rb_r}\phi(b_r)\right)\left(\prod_{k=1}^{K^-}e^{-t_kd_k}e^{-d_k}\right)\cdot\frac{1}{p}\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-})\\
&= \frac{1}{p}\left(\prod_{r=1}^{R^-}e^{\frac{1}{2}s_r^2}\right)\int_{\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-}:\,\widehat{b}+\widehat{d}_{K^-}\leqslant\beta}\left(\prod_{r=1}^{R^-}\phi\left(b_r+s_r\right)\right)\left(\prod_{k=1}^{K^-}e^{-(t_k+1)d_k}\right)\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-}).
\end{aligned}$$

This expression may be compared to Expression (3), where we encountered a large summation containing products of Poisson-type and geometric-type factors. Here, we have its continuous version: an integral containing products of normal and exponential densities. This effectively means that the proof steps are similar to those of Lemmas 1 and 2: we give a recursive argument to evaluate the integrals over exponential densities, and a probabilistic approach is used for the integrals over normal densities.

For intermediate steps where $j\leqslant K^-$ integrals over exponential densities are left, define

$$\begin{aligned}
V_j(x) &= e^{-\lambda(\boldsymbol{s})\sqrt{W}x+\frac{1}{2}Wx^2}\int_{\boldsymbol{b}_{R^-}:\,\widehat{b}\leqslant\beta}\left(\prod_{r=1}^{R^-}\phi\left(b_r+s_r-\sqrt{w_r}x\right)\right)\\
&\times\int_{d_1=0}^{c_1(\beta-\widehat{b})}e^{\left(\frac{x}{c_1}-(t_1+1)\right)d_1}\cdots\int_{d_j=0}^{c_j(\beta-\widehat{b}-\widehat{d}_{j-1})}e^{\left(\frac{x}{c_j}-(t_j+1)\right)d_j}\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_j),
\end{aligned}\tag{33}$$

and notice that $Q(\boldsymbol{s},\boldsymbol{t}) = p^{-1}\left(\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2}\right) V_{K^-}(0)$.

Lemma 9. *$V_j(x)$ satisfies the recursion*

$$V_j(x) = \frac{c_j}{c_j(1+t_j)-x}\left(V_{j-1}(x) - V_{j-1}(c_j(1+t_j))\right).$$

Proof. Integrating (33) over d_j yields

$$\begin{aligned} V_j(x) = e^{-\lambda(\boldsymbol{s})\sqrt{W}x+\frac{1}{2}Wx^2} \int_{\boldsymbol{b}_{R^-}:\widehat{b}\leqslant\beta} \left(\prod_{r=1}^{R^-}\phi\left(b_r+s_r-\sqrt{w_r}x\right)\right) \\ \times \int_{d_1=0}^{c_1(\beta-\widehat{b})} e^{\left(\frac{x}{c_1}-(1+t_1)\right)d_1}\cdots\int_{d_{j-1}=0}^{c_{j-1}(\beta-\widehat{b}-\widehat{d}_{j-2})} e^{\left(\frac{x}{c_{j-1}}-(1+t_{j-1})\right)d_{j-1}}\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_{j-1}) \\ \times \frac{c_j}{c_j(1+t_j)-x}\left(1-e^{(x-c_j(1+t_j))(\beta-\widehat{b}-\widehat{d}_{j-1})}\right). \end{aligned}$$

Observe that the last exponential contains the indices $b_1,\ldots,b_{R^-}$ and $d_1,\ldots,d_{j-1}$. Carefully distributing these indices over the corresponding integrals gives

$$\begin{aligned} V_j(x) &= \frac{c_j}{c_j(1+t_j)-x}\Bigg(V_{j-1}(x) \\ &\quad - e^{-\lambda(\boldsymbol{s})\sqrt{W}c_j(1+t_j)+\frac{1}{2}\left(c_j(1+t_j)\sqrt{W}\right)^2}\int_{\boldsymbol{b}_{R^-}:\widehat{b}\leqslant\beta}\left(\prod_{r=1}^{R^-}\phi\left(b_r+s_r-\sqrt{w_r}c_j(1+t_j)\right)\right) \\ &\quad \int_{d_1=0}^{c_1(\beta-\widehat{b})} e^{\left(\frac{c_j(1+t_j)}{c_1}-(1+t_1)\right)d_1}\cdots\int_{d_{j-1}=0}^{c_{j-1}(\beta-\widehat{b}-\widehat{d}_{j-2})} e^{\left(\frac{c_j(1+t_j)}{c_{j-1}}-(1+t_{j-1})\right)d_{j-1}}\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_{j-1})\Bigg) \\ &= \frac{c_j}{c_j(1+t_j)-x}\left(V_{j-1}(x)-V_{j-1}(c_j(1+t_j))\right), \end{aligned}$$

yielding the stated. □

We are finally ready to show that $Q(\boldsymbol{s},\boldsymbol{t})$ is given by (32), which proves Corollary 1. We proceed as follows: first, we use Lemma 9 to write $Q(\boldsymbol{s},\boldsymbol{t})$ in terms of $V_0(x)$, for certain x. A probabilistic argument subsequently gives an expression for the integrals in $V_0(x)$. Some rearrangements then lead to the equality of $Q(\boldsymbol{s},\boldsymbol{t})$ and (32).

Since $Q(\boldsymbol{s},\boldsymbol{t}) = p^{-1}\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2} V_{K^-}(0)$, we are interested in the value of $V_{K^-}(0)$. For this variable, Lemma 9 implies

$$V_{K^-}(0) = \frac{1}{1+t_{K^-}}\left(V_{K^--1}(0) - V_{K^--1}(c_{K^-}(1+t_{K^-}))\right).$$

Iterating K^- times gives an expression of the form

$$V_{K^-}(0) = a\,V_0(0) + \sum_{l=1}^{K^-} u_l\,V_0\left(c_l(1+t_l)\right),$$

where a and $u_1,\ldots,u_{K^-}$ are coefficients depending on $c_1,\ldots,c_{K^-}$ and $t_1,\ldots,t_{K^-}$. To find a, observe that the only term with $V_0(0)$ results from repeatedly taking the left term of all K^- iterations. Therefore, $a = \prod_{k=1}^{K^-}(1+t_k)^{-1}$. Similarly, observe that the only term with $V_0\left(c_{K^-}(1+t_{K^-})\right)$ results from taking the right term in the first iteration and then repeatedly taking the left term of the remaining iterations. Therefore,

$$u_{K^-} = -\frac{1}{1+t_{K^-}}\prod_{j=1}^{K^--1}\frac{c_j}{c_j(1+t_j)-c_{K^-}(1+t_{K^-})}.$$

By symmetry, we conclude that, for any $l=1,\ldots,K^-$,

$$u_l = -\frac{1}{1+t_l}\prod_{j=1,\,j\neq l}^{K^-}\frac{c_j}{c_j(1+t_j)-c_l(1+t_l)}.$$

Thus, it holds that

$$\begin{aligned} V_{K^-}(0) &= \left(\prod_{k=1}^{K^-}\frac{1}{1+t_k}\right)V_0(0) - \sum_{l=1}^{K^-}\frac{1}{1+t_l}\left(\prod_{j=1,j\neq l}^{K^-}\frac{c_j}{c_j(1+t_j)-c_l(1+t_l)}\right)V_0\left(c_l(1+t_l)\right) \\ &= \left(\prod_{k=1}^{K^-}\frac{1}{1+t_k}\right)\left(V_0(0) - \sum_{l=1}^{K^-}\left(\prod_{j=1,j\neq l}^{K^-}\kappa_{jl}(\boldsymbol{t})\right)V_0\left(c_l(1+t_l)\right)\right). \end{aligned} \tag{34}$$

With expression (34) at hand, $V_0(x)$ still needs to be analyzed. Using (33) and observing that its integral can be written as a probability involving R^- normal random variables, we have

$$\begin{aligned} V_0(x) &= e^{-\lambda(\boldsymbol{s})\sqrt{W}x+\frac{1}{2}Wx^2}\int_{\boldsymbol{b}_{R^-}:\widehat{b}\leqslant\beta}\left(\prod_{r=1}^{R^-}\phi\left(b_r+s_r-\sqrt{w_r}x\right)\right)\mathrm{d}\boldsymbol{b}_{R^-} \\ &= e^{\frac{1}{2}\left(\lambda(\boldsymbol{s})-\sqrt{W}x\right)^2}e^{-\frac{1}{2}\lambda(\boldsymbol{s})^2}\mathbb{P}\left(\sum_{r=1}^{R^-}\sqrt{w_r}\mathcal{N}\left(\sqrt{w_r}x-s_r,1\right)\leqslant\beta\right) \\ &= \frac{\phi(\lambda(\boldsymbol{s}))}{\phi\left(\lambda(\boldsymbol{s})-\sqrt{W}x\right)}\mathbb{P}\left(\mathcal{N}\left(w_rx-\sqrt{w_r}s_r,W\right)\leqslant\beta\right) \\ &= \frac{\phi(\lambda(\boldsymbol{s}))}{\phi\left(\lambda(\boldsymbol{s})-\sqrt{W}x\right)}\Phi\left(\lambda(\boldsymbol{s})-\sqrt{W}x\right) = \phi(\lambda(\boldsymbol{s}))\Psi\left(\lambda(\boldsymbol{s})-\sqrt{W}x\right). \end{aligned} \tag{35}$$

Substituting (35) into (34), we conclude that

$$\begin{aligned} Q(\boldsymbol{s},\boldsymbol{t}) = \frac{1}{p}\left(\prod_{r=1}^{R^-}e^{\frac{1}{2}s_r^2}\right)V_{K^-}(0) &= \frac{1}{p}\left(\prod_{r=1}^{R^-}e^{\frac{1}{2}s_r^2}\right)\left(\prod_{k=1}^{K^-}\frac{1}{1+t_k}\right)\phi(\lambda(\boldsymbol{s})) \\ &\times\left(\Psi\left(\lambda(\boldsymbol{s})\right)-\sum_{l=1}^{K^-}\left(\prod_{j=1,j\neq l}^{K^-}\kappa_{jl}(\boldsymbol{t})\right)\Psi\left(\lambda(\boldsymbol{s})-c_l(1+t_l)\sqrt{W}\right)\right). \end{aligned}$$

We now find the value of p by using that $Q(\boldsymbol{0},\boldsymbol{0})=1$. We then indeed have that $Q(\boldsymbol{s},\boldsymbol{t})$ equals (32) , which completes the proof. □

Appendix D. Table of definitions

Table D.5
Definitions of the main scaling parameters and random variables.

Object	Definition
w_r $(r=1,\ldots,R)$	Positive parameter affecting the traffic load at station r
c_k $(k=1,\ldots,K)$	Positive parameter affecting the traffic load at station k
ν_r $(r=1,\ldots,R)$	Real parameter affecting the traffic load at station r, where $\nu_1\geqslant\ldots\geqslant\nu_R$
α_k $(k=1,\ldots,K)$	Real parameter affecting the traffic load at station k, where $\alpha_1\leqslant\cdots\leqslant\alpha_K$
β	Positive parameter affecting the population size
γ	Parameter affecting the population size, defined as $\gamma=\max\{1-\alpha_1,\frac{1}{2}\nu_1\}$
$\rho_r^{(n)}$ $(r=1,\ldots,R)$	Traffic load at station r, defined as $\rho_r^{(n)}=w_rn^{\nu_r}$
$\sigma_k^{(n)}$ $(k=1,\ldots,K)$	Traffic load at station k, defined as $\sigma_k^{(n)}=n/(n+c_kn^{\alpha_k})$
C_n	Population size of the network, defined as $C_n=\lfloor\|\boldsymbol{\rho}^{(n)}\|+\beta n^{\gamma}\rfloor$
$\overline{B}_r^{(n)}$ $(r=1,\ldots,R)$	Scaled queue length at station r, defined as $\overline{B}_r^{(n)}=(B_r^{(n)}-\rho_r^{(n)})/\sqrt{\rho_r^{(n)}}$
$\overline{D}_k^{(n)}$ $(k=1,\ldots,K)$	Scaled queue length at station k, defined as $\overline{D}_k^{(n)}=(1-\sigma_k^{(n)})D_k^{(n)}$
R^-	Number of heaviest loaded infinite-server stations, defined as the largest integer such that $\nu_1=\nu_2=\cdots=\nu_{R^-}$
K^-	Number of heaviest loaded single-server stations, defined as the largest integer such that $\alpha_1=\alpha_2=\cdots=\alpha_{K^-}$

References

[1] R. Jackson, Queueing systems with phase type service, J. Oper. Res. Soc. 5 (1954) 109–120.
[2] J. Jackson, Networks of waiting lines, Oper. Res. 5 (1957) 518–521.
[3] F. Baskett, K. Chandy, R. Muntz, F. Palacios, Open, closed, and mixed networks of queues with different classes of customers, J. ACM 22 (1975) 248–260.
[4] R. Boucherie, N. van Dijk, Product forms for queueing networks with state-dependent multiple job transitions, Adv. Appl. Probab. 23 (1991) 152–187.
[5] S. Balsamo, V. de Nitto Personè, A survey of product form queueing networks with blocking and their equivalences, Ann. Oper. Res. 48 (1994) 31–61.
[6] R. Boucherie, N. van Dijk, Queueing Networks: A Fundamental Approach, Springer, 2010.
[7] E. Gelenbe, Product-form queueing networks with negative and positive customers, J. Appl. Probab. 28 (1991) 656–663.
[8] S. Lavenberg, M. Reiser, Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers, J. Appl. Probab. 17 (1980) 1048–1061.
[9] R. Boucherie, Norton's equivalent for queueing networks comprised of quasireversible components linked by state-dependent routing, Perform. Eval. 32 (1998) 83–99.
[10] S. Lam, Dynamic scaling and growth behavior of queuing network normalization constants, J. ACM 29 (1982) 492–513.
[11] A. Birman, Y. Kogan, Error bounds for asymptotic approximations of the partition function, Queueing Syst. 23 (1996) 217–234.
[12] D. Mitra, J. McKenna, Asymptotic expansions for closed Markovian networks with state-dependent service rates, J. ACM 33 (1986) 568–592.
[13] A. Mandelbaum, W. Massey, M. Reiman, Strong approximations for Markovian service networks, Queueing Syst. 30 (1998) 149–201.
[14] D. George, C. Xia, M. Squillante, Exact-order asymptotic analysis for closed queueing networks, J. Appl. Probab. 49 (2012) 503–520.
[15] S. Halfin, W. Whitt, Heavy-traffic limits for queues with many exponential servers, Oper. Res. 29 (1981) 567–587.
[16] L. van Kreveld, O. Boxma, J. Dorsman, M. Mandjes, Scaling analysis of an extended machine-repair model, in: Proceedings of the 13th EAI International Conference on Performance Evaluation Methodologies and Tools (ValueTools 2020), 2020, pp. 172–179.
[17] F. Kelly, Networks of queues, Adv. Appl. Probab. 8 (1976) 416–432.
[18] X. Chao, M. Miyazawa, R. Serfozo, H. Takada, Markov network processes with product form stationary distributions, Queueing Syst. 28 (1998) 377–401.
[19] M. Reiser, A queueing network analysis of computer communication networks with window flow control, IEEE Trans. Commun. 27 (1979) 1199–1209.
[20] M. Reiser, H. Kobayashi, Queuing networks with multiple closed chains: theory and computational algorithms, IBM J. Res. Dev. 19 (3) (1975) 283–294.
[21] G. Casale, A generalized method of moments for closed queueing networks, Perform. Eval. 68 (2011) 180–200.
[22] D. George, C. Xia, Fleet-sizing and service availability for a vehicle rental system via closed queueing networks, European J. Oper. Res. 211 (2011) 198–207.
[23] C. Harvey, C. Hills, Determining grades of service in a network, in: Ninth International Teletraffic Conference, Torremolinos, Spain, 1979.
[24] W. Feller, An Introduction to Probability Theory and its Applications, Vol. 1, Wiley, 1968.

Lucas van Kreveld obtained his Bachelor's degree (2016) and Master's degree (2018) in mathematics at Utrecht University. From November 2018, Lucas is a PhD student at the University of Amsterdam, supervised by Jan-Pieter Dorsman, Michel Mandjes and Onno Boxma. Lucas's research is focused on queueing theory with emphasis on asymptotic methods, scheduling and queueing networks.

Onno Boxma is emeritus professor of Stochastic Operations Research in the Department of Mathematics and Computer Science of Eindhoven University of Technology. His main research interests are in queueing theory and its applications to the performance analysis of computer-, communication-, production- and traffic systems, and in insurance risk. During 2004-2009 he was editor-in-chief of Queueing Systems, and from 2005 till 2011 he acted as scientific director of Eurandom. Onno Boxma has received honorary doctorates from the University of Haifa and Heriot-Watt University (Edinburgh), and was recipient of the 2011 ACM SIGMETRICS Achievement Award and the 2014 Arne Jensen Lifetime Award of ITC.

Jan-Pieter Dorsman received his MSc degree in Business Mathematics and Informatics from the VU University Amsterdam in 2010. Subsequently, in 2015, he received his Ph.D. degree from the Eindhoven University of Technology, after working on a PhD project joint with the Centrum Wiskunde & Informatica. From 2015 to 2017, he was a postdoctoral researcher at Leiden University. He is now an assistant professor at the Korteweg-de Vries Institute for Mathematics of the University of Amsterdam. Jan-Pieter's main research interests center around the performance evaluation of stochastic systems and several topics in applied probability and stochastic processes.

Michel Mandjes received M.Sc. degrees in mathematics and econometrics from Vrije Universiteit Amsterdam (the Netherlands) in 1993, and a Ph.D. in Operations Research from the same university in 1996. Currently he is a full professor at the University of Amsterdam, after having been a member of technical staff at Bell Labs (Murray Hill, New Jersey), a full professor of stochastic operations research at the University of Twente, and department head at CWI in Amsterdam. He was visiting professor at Stanford University, Columbia University and New York University. His main research interests are in stochastic processes, queueing processes, efficient simulation techniques, and applications in communication networks and finance. He is the author of two books (a single-authored book on Gaussian queues, published by Wiley, and a coauthored book on Levy Fluctuation Theory and Queues, published by Springer), and he has published about 250 papers in journals and proceedings of conferences. Michel Mandjes was program chair of several leading conferences (INFORMS Applied Probability, Stochastic Networks), and he serves on the editorial board of five journals. He is the main applicant and programme leader of the Dutch research programme NETWORKS.