EURANDOM PREPRINT SERIES

2019-015

December 2, 2019

Scaling analysis of an extended machine-repair model

L. van Kreveld, J.-P. Dorsman, O. Boxma, M. Mandjes
ISSN 1389-2355

Scaling analysis of an extended machine-repair model

L.R. van Kreveld
l.r.vankreveld@uva.nl
University of Amsterdam

O.J. Boxma
o.j.boxma@tue.nl
Eindhoven University of Technology

J.L. Dorsman
j.l.dorsman@uva.nl
University of Amsterdam

M.R.H. Mandjes
m.r.h.mandjes@uva.nl
University of Amsterdam

ABSTRACT

We consider an extension of the classic machine-repair model, where we explicitly model the fact that machines, apart from requiring service from a single repairer, also supply service themselves to products. Due to this dual role of the machines, the system exhibits an intricate relation between the processing rate of products and the performance of the repairer. To characterize this relation, we analyze this model under a Halfin-Whitt inspired scaling regime, where we amplify the arrival rate of products, the repair speed of the repairer and the number of machines appropriately. The resulting limiting stationary distribution is elegant, allows for a closed-form expression and provides intuition on the system's behavior resulting from the machines' dual role. With numerical results we illustrate the convergence, and assess under which conditions the limiting distributions lead to accurate approximations. Next to this valuable insight, the analysis in this paper can be viewed as a first step towards a unifying scaling analysis for more general queueing networks.

ACM Reference Format:

L.R. van Kreveld, O.J. Boxma, J.L. Dorsman, and M.R.H. Mandjes. 2019. Scaling analysis of an extended machine-repair model. In *Proceedings of Valuetools*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this paper we consider a manufacturing model, where machines are subject to breakdown. When broken down, machines wait in a single-server repair queue to be repaired. This model is known as the machine-repair model, which has been studied extensively; see e.g. [11, 15] and references therein. In the literature, the feature that machines may also process products themselves is oftentimes ignored. Therefore, we extend the machine-repair model by adding a queue of products, which is served by the machines when they do not reside in the repair queue. This allows us to study the impact of breakdowns on the processing of the products. The resulting model, which is depicted in Figure 1, can be interpreted as a layered queueing network (cf. [1, 2, 7] and references therein), where the machines fulfill a dual role. They act as customers in the second layer, which forms the machine-repair part of the model. In the first layer, however, they are servers to a single queue of products. We assume that products do not enter the system when there is no machine available, hence the first layer can be interpreted as an Erlang loss model. While both the Erlang loss model and the machine-repair model have incited much research effort, layered queueing networks such as these have received fairly little attention. This is mainly due to the intricate interaction between the layers of these networks, making exact analysis involved even for small models.

Our model, though, possesses a helpful structural property. As will be pointed out later, the blocking mechanism of the first layer ensures that the states of all machines, being either idle, occupied (i.e. processing a product), or broken, can be viewed as objects in a closed queueing network. This gives access to the large body of literature on conventional queueing networks (see e.g. [5] for an overview), from which it follows that under relatively mild conditions, the joint stationary distribution of our model has a known product form. Although this distribution allows for numerical evaluation, explicitly deriving the corresponding marginal distributions from it is not straightforward. Therefore, it is challenging to get a handle on quantifying the correlation between the lengths of the product queue and the repair queue, or ultimately, to identify the interaction between the two layers of the model. A particular complication concerns the evaluation of the normalization constant, which is often done using asymptotic methods (e.g. [4, 13]).

To overcome the difficulties posed above, we consider the behavior of the stationary distribution under a specific scaling. The one we propose, which amplifies the arrival rate, the repair speed, and the number of machines, is reminiscent of the Halfin-Whitt scaling regime [10]. Our scaling allows the explicit calculation of performance measures, which are otherwise hard to obtain. In particular, we derive elegant expressions for the limiting distributions, namely that the number of occupied and broken machines follow a normal and exponential distribution respectively, where one of the two

Valuetools, May 18–20, 2020, Tsukuba, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

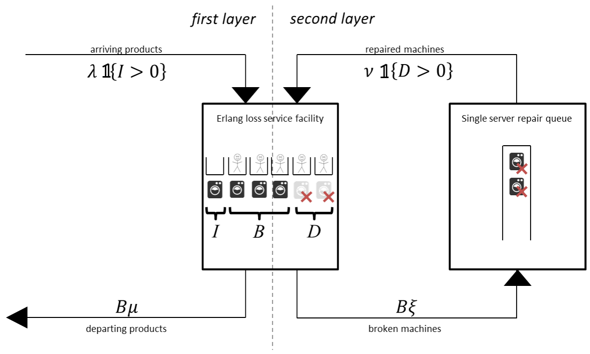

Figure 1: The extended machine-repair model.

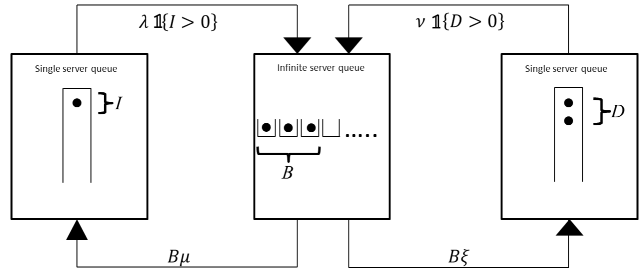

Figure 2: An equivalent closed queueing network.

(depending on the specific values of the scaling parameters) is truncated. Furthermore, we identify the dependence structure between the number of occupied and broken machines. This dependence structure reveals that under our scaling, for nearly all parameter values, the two layers of the model become uncorrelated, which allows for the derivation of approximations of the complete system. Lastly, by means of numerical experiments, we assess the accuracy of such approximations. Although many variants of the Halfin-Whitt regime have been studied for single-station queues, less is known about larger queueing networks [12]. Hence, the analysis presented in this paper can be seen as a step towards developing a scaling analysis in more general product-form networks.

The paper is organized as follows: in Section 2 we describe our model in more detail and discuss the general behavior of the model as well as our scaling framework. Our main theorem and its proof are then given in Sections 3 and 4, respectively. Sections 5 and 6 focus on two special cases of our scaling that require special attention. Lastly, in Section 7 we discuss numerical experiments.

2 MODEL AND PRELIMINARIES

In this section we describe the model, present some preliminary results and introduce the scaling that we impose.

2.1 Model

As depicted in Figure 1, the extended machine-repair model that we consider constitutes a manufacturing facility with C machines. Products arrive according to a Poisson(λ) process and require an exponentially(μ) distributed amount of service. If upon arrival of a product all C machines are occupied (i.e. processing another product) or broken, then the product is rejected and immediately leaves the system. In the remaining case, it is immediately taken into service by an idle machine. When machines are processing products, they are subject to breakdown. Such a breakdown occurs at an exponential rate ξ, after which the machine is immediately sent to a single-server repair queue. The product processing is halted, and the product waits until its machine is repaired. The time it takes for the repairer to repair any machine is exponentially(ν) distributed. Once repaired, the machine immediately returns to the product it was processing at the moment of breakdown.

In the sequel, we let the number of occupied (busy) machines be B, and D denotes the number of broken (down) machines. Apart from Figure 1, there is a different way of viewing our model. Because of the blocking mechanism in the first layer, we may as well keep track of the number of idle machines I, rather than the number of products in progress. Note that at all times $I + B + D = C$, so that the system can be viewed as a closed queueing network depicted in Figure 2. Because of this constraint, it suffices to keep track of only two of the three variables, in our case B and D. It is also worth noting from Figure 2 that the system is highly symmetric: when $\lambda = \nu$ and $\mu = \xi$, it is immediate that I and D have the same stationary distribution. Finally, $\rho_1 := \lambda/\mu$ and $\rho_2 := \nu/\xi$ represent the 'workloads' in each of the two layers, and $\sigma := \rho_1/\rho_2$ their ratio.

For the stationary distribution presented below, we assume that repair times are unknown, so that the repairer cannot serve machines in an order based on individual repair times. Then we have by standard results on closed queueing networks (e.g. [3]) that the stationary distribution of the model satisfies, for $b + d \leqslant C$,

$$p_{b,d} := \mathbb{P}(B = b, D = d) = p_0 \cdot \frac{\rho_1^b}{b!} \cdot \sigma^d, \tag{1}$$

where p_0 is a normalization constant. While the earlier-mentioned exponential assumptions were made for ease of exposition, it is worth emphasizing that (1) holds more generally. In fact, the product form of this equation is also guaranteed when the three queues in isolation are so-called quasi-reversible (see e.g. [14, Chapter 10] for more background). However, even quasi-reversibility is not necessary. For example, the model considered in [8] does not consist of quasi-reversible queues, since it assumes non-exponential service times of products that are resumed upon the return of a machine, rather than repeated. For this case, [8] shows that (1) still holds. In the scaling analysis of this paper, (1) functions as a starting point. Hence, the results in the rest of the paper are valid for every extended machine-repair model that satisfies (1), even non-exponential ones.

The normalization constant p_0 in (1) can be computed as follows. Note that since $B + D \leqslant C$, we have

$$p_0^{-1} = \sum_{b=0}^{C} \frac{\rho_1^b}{b!} \sum_{d=0}^{C-b} \sigma^d. \tag{2}$$

We recognize the right sum as a geometric series. Writing out its explicit solution (assuming for now that $\rho_1 \neq \rho_2$), we obtain

$$p_0^{-1} = \sum_{b=0}^{C} \frac{\rho_1^b}{b!} \frac{1-\sigma^{C-b+1}}{1-\sigma} = \frac{1}{1-\sigma}\left(\sum_{b=0}^{C} \frac{\rho_1^b}{b!} - \sigma^{C+1} \sum_{b=0}^{C} \frac{(\rho_1/\sigma)^b}{b!}\right)$$
$$= \frac{1}{1-\sigma}\left(e^{\rho_1} \sum_{b=0}^{C} e^{-\rho_1}\frac{\rho_1^b}{b!} - e^{\rho_2}\sigma^{C+1} \sum_{b=0}^{C} e^{-\rho_2}\frac{\rho_2^b}{b!}\right). \tag{3}$$

Therefore, the calculation of p_0 requires the evaluation of two Poisson distribution functions:

$$p_0^{-1} = \frac{1}{1-\sigma}\left(e^{\rho_1} Q(C,\rho_1) - e^{\rho_2}\sigma^{C+1} Q(C,\rho_2)\right), \tag{4}$$

where $Q(C,x)$ is the distribution function of a Poisson random variable with mean $x \in \mathbb{R}_{\geq 0}$ evaluated at $C \in \mathbb{N}$:

$$Q(C,x) := \sum_{k=0}^{C} e^{-x}\frac{x^k}{k!}.$$

The case $\rho_1 = \rho_2$ can be dealt with by letting $\sigma \to 1$ in (3) using L'Hôpital's rule.

In the sequel we use the convention that a geometric random variable has support $\{0, 1, 2, \ldots\}$; if the success probability is p, we write $\mathscr{G}(p)$. By $\mathscr{P}(\mu)$ we denote a Poisson random variable with mean μ. Finally, for sequences f_n and g_n, we write $f_n \sim g_n$ if $f_n/g_n \to 1$ as $n \to \infty$. Also, '$=_{\rm d}$' and '$\to_{\rm d}$', respectively, mean equality in distribution and convergence in distribution.

2.2 System behavior with many machines

We proceed by considering the system's behavior as the number of machines C grows. By (1), there is dependence between the layers: B and D are correlated due to the constraint $B + D \leqslant C$. However, it also suggests that when C grows large, this dependence becomes weaker. We comment on this phenomenon, distinguishing between $\rho_1 < \rho_2$, $\rho_2 < \rho_1$, and $\rho_1 = \rho_2$.

For $\rho_1 < \rho_2$, we conclude from (3) that $p_0 \to (1-\sigma)e^{-\rho_1}$ as $C \to \infty$. Inserting this into (1), we indeed conclude that B and D become independent, with $B =_{\rm d} \mathscr{P}(\rho_1)$ and $D =_{\rm d} \mathscr{G}(1-\sigma)$. Since $I = C - B - D$, the number of idle machines I will tend to infinity.

Using the symmetry of the model, the queue size distributions for $\rho_1 > \rho_2$ follow similarly: as $C \to \infty$, we find that B and D are asymptotically independent. More particularly, $B =_{\rm d} \mathscr{P}(\rho_2)$ and $I =_{\rm d} \mathscr{G}(1-1/\sigma)$, and the number of broken machines D will tend to infinity.

Finally, when $\rho_1 = \rho_2$, it also follows from (1) that $B =_{\rm d} \mathscr{P}(\rho_1) = \mathscr{P}(\rho_2)$ as $C \to \infty$. Moreover, by the symmetry of the model and the fact that the stationary distribution (1) only depends on the system parameters through ρ_1 and ρ_2, we have that I and D must have the same distribution. Since B remains finite in the limit, this means that I and D will tend to infinity as $C \to \infty$.

So far we have not discussed stability of the model. To this end, realize that it permits only a finite number of products, making it stable without imposing any conditions. As $C \to \infty$ this is not the case anymore. As in our scaling the number of machines tends to infinity, we henceforth assume that $\rho_1 < \rho_2$.

2.3 Scaling

As motivated in Section 1, we consider our model under a specific scaling, which we now introduce. Scaling limits are a common tool to analyze stochastic systems in an asymptotic regime, and are particularly useful when non-tractable systems in the limit simplify. Loosely speaking, scaling means that a subset of the model parameters is parametrized by n, after which n is sent to ∞. 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.

In their celebrated 1981 paper, Halfin and Whitt [10] 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 $\beta > 0$. For the Erlang loss model in particular, a normalized version of the queue size then asymptotically behaves as a normal random variable truncated at β [10, Thm. 3].

The scaling we impose is inspired by the Halfin-Whitt regime, in that we scale the parameter ρ_1 and the number of machines C. Our system, however, distinguishes itself from a multi-server queue by having an additional layer with breakdowns and repairs. In order to preserve stability, we should therefore scale ρ_2 as well.

We now give a precise definition of the scaling we impose in this paper. Let $\gamma_n := \frac{n^\alpha}{n+n^\alpha}$. Then, for $\alpha \in \mathbb{R}$, $\beta > 0$,

- we replace λ by μn (such that ρ_1 becomes n);
- we replace ν by $\xi\,(n + n^\alpha)$ (such that ρ_2 becomes $n + n^\alpha$);
- we replace C by C_n, where we define C_n as $\lfloor n + \beta/\gamma_n \rfloor$ if $\alpha < \frac{1}{2}$, and as $\lfloor n + \beta\sqrt{n} \rfloor$ if $\alpha \geq \frac{1}{2}$.

Along with the system parameters, the numbers of occupied machines B and broken machines D depend on n. Let B_n and D_n respectively denote these random variables under the scaling regime. We are now interested in their behavior as $n \to \infty$. Since their means may become arbitrarily large with n, we consider their appropriately normalized versions

$$\bar{B}_n := \frac{B_n - n}{\sqrt{n}} \qquad \text{and} \qquad \bar{D}_n := \gamma_n D_n.$$

Our choice for C_n can be motivated as follows. It turns out that, to get non-degenerate limits, C should be picked such that it equals the mean of B increased by a constant β times the largest of the standard deviations of B and D. As argued in Section 2.2, for C large and $\rho_1 < \rho_2$, B behaves as $\mathscr{P}(\rho_1) = \mathscr{P}(n)$, which has standard deviation $\sqrt{n}$. In addition, D behaves as $\mathscr{G}(1-\sigma) = \mathscr{G}(\gamma_n)$, which has standard deviation $\sqrt{1-\gamma_n}/\gamma_n$. Note that the standard deviation of B is larger than that of D when $\alpha > \frac{1}{2}$, and vice versa if $\alpha < \frac{1}{2}$. In the latter case, observe that $\sqrt{1-\gamma_n}/\gamma_n \sim 1/\gamma_n$. These observations lead to our choice of C_n.

Before concluding this section, we discuss the intuition behind each value of the scaling parameter α.

- Case $\alpha < \frac{1}{2}$ (*slow repair regime*): ρ_2 is only slightly larger than ρ_1, causing the repair queue size (right queue in Figure 2) to be large, and in particular, to fluctuate much more than the product queue size (middle queue in Figure 2).
- Case $\alpha = \frac{1}{2}$ (*moderate repair regime*): ρ_2 is larger than ρ_1 in such a way that the repair queue size fluctuates just as much as the product queue size.
- Case $\frac{1}{2} < \alpha < 1$ (*fast repair regime*): ρ_2 is much larger than ρ_1, causing the repair queue size to fluctuate much less than the product queue size.
- Case $\alpha \geqslant 1$ (*nearly-instantaneous repair regime*): similar to the fast repair regime, but with very small repair queue size.

As such, the parameter α controls the level at which the queue sizes fluctuate (deviate from their mean). This in turn has impact on the parameter C_n.

3 MAIN RESULTS

In this section, we derive the asymptotic joint distribution of $\bar{B}_n$ and $\bar{D}_n$ for $\alpha < 1$, by means of their joint Laplace-Stieltjes transform (LST). The case $\alpha \geq 1$ will be studied in Section 6.

Define, for $s, t \geqslant 0$,

$$P_n(s,t) = \mathbb{E}\left(e^{-s\bar{B}_n} e^{-t\bar{D}_n}\right) = \sum_{b+d \leqslant C_n} e^{-s\frac{b-n}{\sqrt{n}}} e^{-t\gamma_n d} p_{b,d}^{(n)} \quad (5)$$

as the joint Laplace-Stieltjes transform of $\bar{B}_n$ and $\bar{D}_n$. Furthermore, for a normally distributed random variable, we write

$$\Psi(s) = \frac{\Phi(s)}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2}}$$

for its *Mills ratio*, i.e. its distribution function divided by its density function. The following theorem, to be proven in Section 4, gives an explicit expression for the limit of $P_n(s,t)$ as $n \to \infty$, from which we can directly derive the asymptotic distribution of $\left(\bar{B}_n, \bar{D}_n\right)$.

Theorem 1. *The two-dimensional LST of $\left(\bar{B}_n, \bar{D}_n\right)$ satisfies, as $n \to \infty$,*

$$P_n(s,t) \to \begin{cases} e^{\frac{1}{2}s^2} \frac{1}{1+t} \frac{1-e^{-\beta(t+1)}}{1-e^{-\beta}} & \text{if } \alpha < \frac{1}{2}, \\ e^{-s\beta} \frac{1}{1+t} \frac{\Psi(\beta+s)-\Psi(\beta+s-t-1)}{\Psi(\beta)-\Psi(\beta-1)} & \text{if } \alpha = \frac{1}{2}, \\ e^{\frac{1}{2}s^2} \frac{1}{1+t} \frac{\Phi(\beta+s)}{\Phi(\beta)} & \text{if } \frac{1}{2} < \alpha < 1. \end{cases}$$

Unless $\alpha = \frac{1}{2}$, the asymptotic distributions of $\bar{B}_n$ and $\bar{D}_n$ immediately follow from Theorem 1. This is summarised in the following corollary. The interpretation of the joint LST in case $\alpha = \frac{1}{2}$ is more intricate, and will be studied in Section 5. To state the next corollary, we let $\mathscr{N}$ denote a standard-normal random variable, while $\mathscr{E}$ stands for an exponential random variable with rate 1.

Corollary 1. *For $\alpha < \frac{1}{2}$, as $n \to \infty$,*

$$\bar{B}_n \to_{\mathrm{d}} \mathscr{N} \quad \text{and} \quad \bar{D}_n \to_{\mathrm{d}} (\mathscr{E} \mid \mathscr{E} \leqslant \beta). \quad (6)$$

For $\frac{1}{2} < \alpha < 1$, as $n \to \infty$,

$$\bar{B}_n \to_{\mathrm{d}} (\mathscr{N} \mid \mathscr{N} \leqslant \beta) \quad \text{and} \quad \bar{D}_n \to_{\mathrm{d}} \mathscr{E}. \quad (7)$$

In both of these two cases, $\bar{B}_n$ and $\bar{D}_n$ are asymptotically independent.

Proof. We recognize the expressions of Theorem 1 as known Laplace-Stieltjes transforms. Let X be a random variable and let $f(u) := \mathbb{E}(e^{-uX})$ be its LST. Then we can check by straightforward calculations that

- $f(u) = e^{\frac{1}{2}u^2}$ if $X =_{\mathrm{d}} \mathscr{N}$,
- $f(u) = e^{\frac{1}{2}u^2} \frac{\Phi(\beta+u)}{\Phi(\beta)}$ if $X =_{\mathrm{d}} (\mathscr{N} \mid \mathscr{N} \leqslant \beta)$,
- $f(u) = \frac{1}{1+u}$ if $X =_{\mathrm{d}} \mathscr{E}$, and
- $f(u) = \frac{1}{1+u} \frac{1-e^{-\beta(u+1)}}{1-e^{-\beta}}$ if $X =_{\mathrm{d}} (\mathscr{E} \mid \mathscr{E} \leqslant \beta)$.

The statement now follows from Theorem 1 in combination with Lévy's convergence theorem [16, Thm. 18.1]. □

This corollary generalizes the Halfin-Whitt result [10, Thm. 3], which states that under a similar regime the number of customers in an Erlang loss system has a truncated normal distribution. That result is covered by ours by letting $\alpha \to \infty$. Repairs are then instantaneous, so that the system behaves as if breakdowns never occur.

Since the normal and exponential distributions can be seen as the continuous counterparts of the Poisson and geometric distributions, the limiting distributions of $\bar{B}_n$ and $\bar{D}_n$ have a natural connection with the Poisson and geometric distributions of Section 2.2. Specifically, suppose that $B =_{\mathrm{d}} \mathscr{P}(\rho_1) = \mathscr{P}(n)$ and $D =_{\mathrm{d}} \mathscr{G}(1-\sigma) = \mathscr{G}(\gamma_n)$ as was the case for $\rho_1 < \rho_2$ if $C \to \infty$. In that case, we would have $\bar{B}_n \to_{\mathrm{d}} \mathscr{N}$ and $\bar{D}_n \to_{\mathrm{d}} \mathscr{E}$ as $n \to \infty$. However, in Section 2.2 the total number of machines C was infinitely large compared to B and D, which is not the case under our scaling. It is essentially the limitation of $B_n + D_n$ to $\{0, \dots, C_n\}$ that causes the truncations in Corollary 1. If $\alpha < \frac{1}{2}$, then D_n has the largest variance so its support is noticeably limited by C_n. The same goes for B_n when $\alpha > \frac{1}{2}$.

4 PROOF OF THEOREM 1

To prove Theorem 1, we first rewrite the right-hand side of (5) in Lemma 1. The resulting expression is then split into three factors: the LST of a normal distribution, the LST of an exponential distribution, and a term related to the truncation. For each of these terms separately, we derive an asymptotic expression. This is straightforward for the first two terms, but the last term requires careful asymptotic analysis, where we must distinguish between various regimes of α. Lemmas 3 and 4 are devoted to this, and build on an important observation presented in Lemma 2.

In order to keep expressions concise throughout, we introduce the following notation. Recalling that $Q(C,x) = \mathbb{P}(\mathscr{P}(x) \leq C)$,

$$f_n(s) := Q\left(C_n, ne^{-\frac{s}{\sqrt{n}}}\right), \qquad s \geq 0,$$

$$g_n(s,t) := Q\left(C_n, (n+n^{\alpha})e^{-\frac{s}{\sqrt{n}}} e^{t\gamma_n}\right), \qquad s,t \geq 0, \quad \text{and}$$

$$h_n(s,t) := \exp\left((n+n^{\alpha})\, e^{-\frac{s}{\sqrt{n}}} e^{t\gamma_n} - ne^{-\frac{s}{\sqrt{n}}}\right) \times \left((1-\gamma_n)\, e^{-t\gamma_n}\right)^{C_n+1}, \qquad s,t \geq 0.$$

Using this notation, Lemma 1 provides an alternative expression for the LST of $\left(\bar{B}_n, \bar{D}_n\right)$.

Lemma 1. *The two-dimensional LST of $\left(\bar{B}_n, \bar{D}_n\right)$ satisfies*

$$P_n(s,t) = e^{-n+s\sqrt{n}+ne^{-\frac{s}{\sqrt{n}}}} \frac{\gamma_n}{1-(1-\gamma_n)e^{-t\gamma_n}} \times \frac{f_n(s) - g_n(s,t)h_n(s,t)}{f_n(0) - g_n(0,0)h_n(0,0)}. \tag{8}$$

Proof. The ideas underlying this proof resemble those underlying the calculation of the normalization constant; cf. Equations (2) and (3). Noting that $B_n + D_n \leqslant C_n$, and using (1), we obtain

$$P_n(s,t) = \sum_{b+d\leqslant C_n} e^{-s\frac{b-n}{\sqrt{n}}} e^{-t\gamma_n d} p_{b,d}^{(n)}$$
$$= p_0^{(n)} e^{s\sqrt{n}} \sum_{b=0}^{C_n} \frac{\left(ne^{-\frac{s}{\sqrt{n}}}\right)^b}{b!} \sum_{d=0}^{C_n-b} \left((1-\gamma_n)\, e^{-t\gamma_n}\right)^d.$$

We recognize sums similar to those in (2), but where C is replaced by C_n, ρ_1 is replaced by $n\exp(-s/\sqrt{n})$ and ρ_2 is replaced by $(n + n^{\alpha})\exp(-s/\sqrt{n})e^{t\gamma_n}$. Since $\gamma_n = \frac{n^{\alpha}}{n+n^{\alpha}}$, it follows that $\sigma = \rho_1/\rho_2$ is replaced by $(1-\gamma_n)\, e^{-t\gamma_n}$. Analogous to (4) we find, with $\delta_n(t) := (1-\gamma_n)e^{-t\gamma_n}$,

$$P_n(s,t) = p_0^{(n)} e^{s\sqrt{n}} \frac{1}{1-\delta_n(t)} \times \left(e^{ne^{-\frac{s}{\sqrt{n}}}} Q\left(C_n, ne^{-\frac{s}{\sqrt{n}}}\right) - \delta_n(t)^{C_n+1} e^{(n+n^{\alpha})e^{-\frac{s}{\sqrt{n}}} e^{t\gamma_n}} Q\left(C_n, (n+n^{\alpha})e^{-\frac{s}{\sqrt{n}}} e^{t\gamma_n}\right)\right)$$
$$= p_0^{(n)} e^{s\sqrt{n}+ne^{-\frac{s}{\sqrt{n}}}} \frac{f_n(s) - g_n(s,t)h_n(s,t)}{1-\delta_n(t)}. \tag{9}$$

To examine the behavior of $p_0^{(n)}$ as $n \to \infty$, we take $s = t = 0$ in (9) and realize that $P_n(0,0) = 1$ to find

$$p_0^{(n)} = e^{-n}\gamma_n\left(f_n(0) - g_n(0,0)h_n(0,0)\right)^{-1}.$$

Substituting this back in (9) then gives the desired result. □

As mentioned at the start of this section, the expression for $P_n(s,t)$ in Lemma 1 is a product of three terms, each playing an intuitively appealing role in relation to Corollary 1. More specifically, our analysis reveals that the first two terms correspond to the normal and exponential distribution, respectively, as $n \to \infty$. In addition, we show that the third term, which is significantly more subtle to analyze in the limiting regime $n \to \infty$, immediately relates to the truncation at β.

Let us start with the leftmost term of (8). Applying a standard Taylor expansion to $\exp(-s/\sqrt{n})$ around $s = 0$, we obtain

$$e^{-n+s\sqrt{n}+ne^{-\frac{s}{\sqrt{n}}}} = e^{-n+s\sqrt{n}+n\left(1-\frac{s}{\sqrt{n}}+\frac{s^2}{2n}+O\left(n^{-\frac{3}{2}}\right)\right)}. \tag{10}$$

As $n \to \infty$, this converges to $\exp(\frac{1}{2}s^2)$, which can be recognized as the LST of a standard-normal random variable. For the middle term, we can apply the same strategy: for $\alpha < 1$, as $n \to \infty$, we have $\gamma_n \to 0$ so that we can apply a Taylor expansion to $\exp(-t\gamma_n)$. Therefore,

$$\frac{\gamma_n}{1-(1-\gamma_n)e^{-t\gamma_n}} = \frac{\gamma_n}{1-(1-\gamma_n)\left(1-t\gamma_n+O\left(\gamma_n^2\right)\right)} = \frac{\gamma_n}{\gamma_n(1+t)+O\left(\gamma_n^2\right)}. \tag{11}$$

As $n \to \infty$, this converges to $\frac{1}{1+t}$, which is the LST of an exponentially distributed random variable with rate 1. Hence, to prove the theorem it remains to analyze the second fraction of (8) in various regimes of α. To this end, we inspect the behavior of the functions $f_n(s)$, $g_n(s,t)$ and $h_n(s,t)$ in the limiting regime separately for different values of α. In doing this, we rely on the following lemma.

Lemma 2. *Suppose $x_n \to \infty$ as $n \to \infty$. If $(C_n - x_n)/\sqrt{x_n} \to M$, then $Q(C_n, x_n) \to \Phi(M)$ as $n \to \infty$.*

Proof. Observe that a Poisson random variable with mean $x_n \geq 1$ can be written as a sum of $\lfloor x_n \rfloor$ Poisson random variables with mean $x_n/\lfloor x_n \rfloor$. Therefore, with $(Y_k)_{k\in\mathbb{N}}$ i.i.d. copies of $\mathscr{P}(x_n/\lfloor x_n \rfloor)$,

$$Q(C_n, x_n) = \mathbb{P}\left(\sum_{k=1}^{\lfloor x_n \rfloor} Y_k \leqslant C_n\right) = \mathbb{P}\left(\frac{1}{\sqrt{x_n}} \sum_{k=1}^{\lfloor x_n \rfloor}\left(Y_k - \frac{x_n}{\lfloor x_n \rfloor}\right) \leqslant \frac{C_n - x_n}{\sqrt{x_n}}\right).$$

The lemma now follows from the central limit theorem. □

We now explicitly analyze the asymptotic behavior of the functions $f_n(s)$, $g_n(s,t)$ and $h_n(s,t)$ for the cases $\alpha < \frac{1}{2}$, $\alpha > \frac{1}{2}$, and $\alpha = \frac{1}{2}$. Lemma 3 covers most of these results, while Lemma 4 provides the remaining analysis for $g_n(s,t)$ and $h_n(s,t)$ in case $\alpha > \frac{1}{2}$.

Lemma 3. *Let $n \to \infty$. Then,*

- *for $\alpha < \frac{1}{2}$, $f_n(s), g_n(s,t) \to 1$ and $h_n(s,t) \to e^{-\beta(t+1)}$;*
- *for $\alpha > \frac{1}{2}$, $f_n(s) \to \Phi(\beta + s)$;*
- *for $\alpha = \frac{1}{2}$, $f_n(s) \to \Phi(\beta + s)$, $g_n(s,t) \to \Phi(\beta + s - 1 - t)$ and $h_n(s,t) \to e^{-\frac{1}{2}(\beta+s-1-t)^2 - \frac{1}{2}(\beta+s)^2}$.*

Proof. We start with the case $\alpha < \frac{1}{2}$. Note that in this case $C_n = \lfloor n + \beta/\gamma_n \rfloor$. To prove $f_n(s) \to 1$, we apply Lemma 2 with $x_n = n\exp(-s/\sqrt{n}) = n - s\sqrt{n} + o\left(\sqrt{n}\right)$ (so that $M = \infty$). Along the same lines, it can be shown that $g_n(s,t) \to 1$.

For the limit of $f_n(s)$ in case $\alpha \geqslant \frac{1}{2}$, an application of Lemma 2 with $x_n = n\exp(-s/\sqrt{n})$ and $C_n = \lfloor n + \beta\sqrt{n} \rfloor$ leads to

$$M = \lim_{n\to\infty} \frac{C_n - x_n}{\sqrt{x_n}} = \lim_{n\to\infty} \frac{(\beta + s)\sqrt{n} + o\left(\sqrt{n}\right)}{\sqrt{n - s\sqrt{n} + o\left(\sqrt{n}\right)}} = \beta + s.$$

Hence, if $\alpha \geqslant \frac{1}{2}$, then $f_n(s) \to \Phi(\beta + s)$.

Next, for $g_n(s,t)$ as $\alpha = \frac{1}{2}$, an application of Lemma 2 to $x_n = (n + \sqrt{n})\, \exp(-s/\sqrt{n})\, e^{t\gamma_n}$ and $C_n = \lfloor n + \beta\sqrt{n} \rfloor$ leads to

$$M = \lim_{n\to\infty} \frac{C_n - x_n}{\sqrt{x_n}} = \lim_{n\to\infty} \frac{(\beta + s - 1 - t)\sqrt{n} + o\left(\sqrt{n}\right)}{\sqrt{n - (s-1-t)\sqrt{n} + o\left(\sqrt{n}\right)}} = \beta + s - 1 - t.$$

We conclude that if $\alpha = \frac{1}{2}$, then $g_n(s,t) \to \Phi\left(\beta + s - 1 - t\right)$.

For $h_n(s,t)$ when $\alpha \leqslant \frac{1}{2}$, we explicitly consider the two components of $h_n(s,t)$ separately. In the proceeding calculations, terms vanishing as $n \to \infty$ will be integrated into $o(1)$. From the first component of $h_n(s,t)$, we extract the leading terms by applying Taylor expansions. With $s_n := 1 - s/\sqrt{n}$, this component equals

$$\begin{aligned}
&\exp\left((n + n^\alpha)\, e^{-\frac{s}{\sqrt{n}}} e^{t\gamma_n} - ne^{-\frac{s}{\sqrt{n}}}\right) \\
&= \exp\left(n^\alpha e^{-\frac{s}{\sqrt{n}}} e^{t\gamma_n} + ne^{-\frac{s}{\sqrt{n}}}\left(e^{t\gamma_n} - 1\right)\right) \\
&= \exp\left(n^\alpha s_n\,(1 + t\gamma_n) + ns_n\left(t\gamma_n + \frac{1}{2}t^2\gamma_n^2\right) + o(1)\right) \\
&= \exp\left((n + n^\alpha)\, t\gamma_n + n^\alpha - s\left(n^{\alpha-\frac{1}{2}} + \sqrt{n}t\gamma_n\right) + \frac{1}{2}nt^2\gamma_n^2 + o(1)\right) \\
&= \exp\left((t+1)n^\alpha - s\,(1+t)\, n^{\alpha-\frac{1}{2}} + \frac{1}{2}nt^2\gamma_n^2 + o(1)\right). \qquad (12)
\end{aligned}$$

Likewise, defining $\eta_n(t) := 1 - e^{-t\gamma_n} + \gamma_n e^{-t\gamma_n}$, we have for the second component of $h_n(s,t)$ that

$$\begin{aligned}
&\left((1-\gamma_n)e^{-t\gamma_n}\right)^{C_n+1} \\
&\quad= \left((1-\gamma_n)\, e^{-t\gamma_n}\right)\exp\left(C_n \ln\left(1 - \eta_n(t)\right)\right) \\
&\quad= \exp\left(-C_n\eta_n(t) - \frac{1}{2}C_n\eta_n(t)^2 + o(1)\right) \\
&\quad= \exp\left(-C_n\left(t\gamma_n - \frac{1}{2}t^2\gamma_n^2 + \gamma_n - t\gamma_n^2 + \frac{1}{2}(t+1)^2\gamma_n^2\right) + o(1)\right) \\
&\quad= \exp\left(-C_n\left((t+1)\gamma_n + \frac{1}{2}\gamma_n^2\right) + o(1)\right). \qquad (13)
\end{aligned}$$

Multiplying the two components (12) and (13), we conclude that $h_n(s,t) = \exp(H_n(s,t) + o(1))$ as $n \to \infty$, with

$$H_n(s,t) = (t+1)\left(n^\alpha - C_n\gamma_n - sn^{\alpha-\frac{1}{2}}\right) + \frac{1}{2}\gamma_n^2\left(nt^2 - C_n\right). \quad (14)$$

If $\alpha < \frac{1}{2}$, many of these terms vanish; noticing $n^\alpha - n\gamma_n \to 0$,

$$\begin{aligned}
h_n(s,t) &= \exp\left((t+1)\,(n^\alpha - C_n\gamma_n) + o(1)\right) \\
&= \exp\left((t+1)\,(n^\alpha - n\gamma_n - \beta) + o(1)\right) \to \exp\left(-\beta(t+1)\right).
\end{aligned}$$

If on the other hand $\alpha = \frac{1}{2}$, then with (14), $h_n(s,t)$ equals

$$\begin{aligned}
&\exp\left((t+1)\left(\sqrt{n} - \frac{n\sqrt{n} + \beta n}{n + \sqrt{n}} - s\right) + \frac{1}{2}\gamma_n^2\left(nt^2 - n\right) + o(1)\right) \\
&\to \exp\left(-(t+1)\,(\beta + s - 1) + \frac{1}{2}\left(t^2 - 1\right)\right) \\
&= \exp\left(\frac{1}{2}\,(\beta + s - t - 1)^2 - \frac{1}{2}(\beta + s)^2\right).
\end{aligned}$$

This completes the proof of Lemma 3. □

The analysis of the remaining functions, i.e. $g_n(s,t)$ and $h_n(s,t)$ for $\alpha > \frac{1}{2}$, requires a more subtle reasoning. Since $g_n(s,t) \to 0$ and $h_n(s,t) \to \infty$ as $n \to \infty$, we must analyze the product of the two functions before taking the limit, which can be done by a change-of-measure argument.

Lemma 4. *Let $n \to \infty$. If $\alpha > \frac{1}{2}$, then $g_n(s,t)\, h_n(s,t) \to 0$.*

Proof. Let $x_n(s,t) = (n + n^\alpha)e^{-\frac{s}{\sqrt{n}}} e^{t\gamma_n}$; in the sequel we write just x_n for brevity. In this proof, our first objective is to identify the asymptotics of $g_n(s,t) = \mathbb{P}(\mathscr{P}(x_n) \leqslant C_n)$. To this end, let $\mathbb{Q}$ be an alternative measure under which the Poisson random variable has mean C_n, such that

$$g_n(s,t) = \mathbb{E}_\mathbb{P}\left(\mathbb{1}\{P \le C_n\}\right) = \mathbb{E}_\mathbb{Q}\left(L\,\mathbb{1}\{P \le C_n\}\right),$$

with L denoting the likelihood ratio or Radon-Nikodym derivative

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

We thus arrive at

$$g_n(s,t) = e^{C_n - x_n}\mathbb{E}_\mathbb{Q}\left(\left(x_n/C_n\right)^P\ \mathbb{1}\{P \le C_n\}\right).$$

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

$$g_n(s,t) = e^{C_n - x_n}\left(\frac{x_n}{C_n}\right)^{C_n} q_n, \qquad (15)$$

where

$$q_n = \mathbb{E}_\mathbb{Q}\left(\left(\left(x_n/C_n\right)^{\sqrt{C_n}}\right)^{\bar{P}}\ \mathbb{1}\{\bar{P} \le 0\}\right).$$

We now replace $\bar{P}$ for n large by a standard-normal random variable, which can be formally justified by applying the Berry-Esseen bound, precisely following the lines of the proof of [6, Thm. 3.7.4]. Hence we obtain, with $d_n := \sqrt{C_n}\ln(x_n/C_n)$, that

$$\begin{aligned}
q_n &\sim \int_{-\infty}^0 \left(\left(\frac{x_n}{C_n}\right)^{\sqrt{C_n}}\right)^y \frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,\mathrm{d}y \\
&= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}}\exp\left(yd_n - \frac{y^2}{2}\right)\mathrm{d}y.
\end{aligned}$$

Completing the square in the exponent now leads to

$$\begin{aligned}
q_n &= \exp\left(\frac{d_n^2}{2}\right)\int_{-\infty}^0 \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\,(y - d_n)^2\right)\mathrm{d}y \\
&= \exp\left(\frac{d_n^2}{2}\right)\Phi(-d_n) = \exp\left(\frac{d_n^2}{2}\right)(1 - \Phi(d_n))
\end{aligned} \qquad (16)$$

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

$$e^{x^2/2}\,(1 - \Phi(x)) \sim \frac{1}{x\sqrt{2\pi}} \qquad (17)$$

as $x \to \infty$ (cf. [9, p. 175]). To use this property it is necessary to verify that d_n goes to ∞ as $n \to \infty$. This can be seen by relying on a Taylor expansion, and recalling that $\alpha > \frac{1}{2}$:

$$\sqrt{C_n}\ln\frac{x_n}{C_n} = \sqrt{\lfloor n + \beta\sqrt{n}\rfloor}\cdot\ln\left(\frac{(n + n^\alpha)e^{-\frac{s}{\sqrt{n}}}e^{t\gamma_n}}{\lfloor n + \beta\sqrt{n}\rfloor}\right)$$

$$= \sqrt{n+o(n)} \cdot \ln\left(1+\omega\left(n^{-\frac{1}{2}}\right)\right) = \omega(1) \to \infty$$

as $n \to \infty$, where $\omega(w_n)$ denotes a sequence v_n for which it holds that $\lim_{n\to\infty} v_n/w_n = \infty$. Using property (17) in (16), and substituting the result in (15), we thus obtain that, as $n \to \infty$,

$$g_n(s,t) \sim e^{C_n - x_n} \left(\frac{x_n}{C_n}\right)^{C_n} \frac{1}{d_n \sqrt{2\pi}}.$$

Multiplying with $h_n(s,t)$, and using $(1-\gamma_n)e^{-t\gamma_n} x_n = ne^{-s/\sqrt{n}}$, it holds that

$$\begin{aligned} & g_n(s,t)\, h_n(s,t) \\ & \sim e^{C_n - x_n} \left(\frac{x_n}{C_n}\right)^{C_n} \frac{1}{d_n \sqrt{2\pi}} \cdot e^{x_n - ne^{-\frac{s}{\sqrt{n}}}} \left((1-\gamma_n)e^{-t\gamma_n}\right)^{C_n+1} \\ & \leqslant \frac{1}{d_n} e^{C_n - ne^{-\frac{s}{\sqrt{n}}}} \left(\frac{ne^{-\frac{s}{\sqrt{n}}}}{C_n}\right)^{C_n}. \end{aligned}$$

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

$$\begin{aligned} & g_n(s,t)\, h_n(s,t) \\ & \leqslant \frac{1}{d_n} \exp\left(C_n - ne^{-\frac{s}{\sqrt{n}}} + C_n \ln\left(1 + \left(\frac{ne^{-\frac{s}{\sqrt{n}}} - C_n}{C_n}\right)\right)\right) \\ & = \frac{1}{d_n} \exp\left(C_n - ne^{-\frac{s}{\sqrt{n}}} + C_n \frac{ne^{-\frac{s}{\sqrt{n}}} - C_n}{C_n} + O(1)\right) \to 0 \end{aligned}$$

as $n \to \infty$. □

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

Proof of Theorem 1. The result is a consequence of Lemma 1 when substituting equations (10), (11), in combination with the functions that we asymptotically evaluated in Lemmas 3 and 4 (both for general s,t and for $s = t = 0$, that is). □

5 MODERATE REPAIR ($\alpha = \frac{1}{2}$)

Now that Theorem 1 has been proved, Corollary 1 provides the limiting marginal distributions of B_n and D_n when $\alpha < \frac{1}{2}$ or $\frac{1}{2} < \alpha < 1$. In particular, the corollary shows that for these values of α, B_n and D_n are asymptotically independent. In this section, we discuss the implications of Theorem 1 for $\alpha = \frac{1}{2}$, which turn out to be more delicate. In this case, the repair queue size fluctuates precisely as much as the product queue size, as we pointed out in Section 2.3. This causes the constraint $B_n + D_n \leqslant C_n$ to affect the limiting distributions of both $\bar{B}_n$ and $\bar{D}_n$ (instead of just either of them as in Corollary 1). According to Theorem 1, the LST satisfies

$$P_n(s,t) \to e^{-s\beta} \frac{1}{1+t} \frac{\Psi(\beta+s) - \Psi(\beta+s-t-1)}{\Psi(\beta) - \Psi(\beta-1)}, \qquad (18)$$

as $n \to \infty$, where we recall that $\Psi(x)$ denotes the Mills ratio of the standard-normal distribution. It is unclear how to invert this LST, but we can extract meaningful information from it. A crucial observation is that the limit of $P_n(s,t)$ no longer factorizes into components involving s and t separately, as was the case in Corollary 1. As a result, we conclude that $\bar{B}_n$ and $\bar{D}_n$ are now asymptotically *dependent*. This finding is especially intriguing because for all other values of α we found asymptotic independence of $\bar{B}_n$ and $\bar{D}_n$.

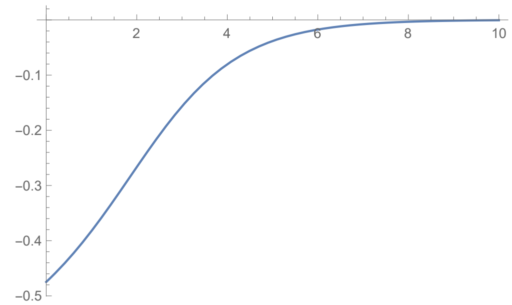

Figure 3: Cor($\bar{B}, \bar{D}$) for $\alpha = \frac{1}{2}$, as a function of β.

To characterize the dependence that arises, we compute (joint) moments of $\bar{B}_n$ and $\bar{D}_n$ exploiting the moment-generating properties of $P_n(s,t)$. Let $\bar{B} := \lim_{n\to\infty} \bar{B}_n$ and $\bar{D} := \lim_{n\to\infty} \bar{D}_n$. Then by differentiation of (18) and inserting $s = t = 0$, we have

$$\mathbb{E}\left(\bar{B}\right) = -\frac{\Psi(\beta-1)}{\Psi(\beta) - \Psi(\beta-1)},$$

$$\mathbb{E}\left(\bar{D}\right) = 1 - \frac{1 + (\beta-1)\Psi(\beta-1)}{\Psi(\beta) - \Psi(\beta-1)}.$$

By standard properties of $\Psi(\cdot)$, one can show that $\mathbb{E}(\bar{B}) < 0$ and $\mathbb{E}(\bar{D}) < 1$. This can be explained by the fact that 0 and 1 are the expectations of the untruncated distributions as presented in Corollary 1. However, as $\alpha = \frac{1}{2}$ is the critical point between the two cases presented in Corollary 1, it makes sense that the constraint $B_n + D_n \leqslant C_n$ has a decreasing effect on both $\mathbb{E}(\bar{B})$ and $\mathbb{E}(\bar{D})$, rather than just one of them. Furthermore, by computing $\mathbb{E}(\bar{B}\bar{D})$ using (18), we obtain

$$\mathrm{Cov}\left(\bar{B}, \bar{D}\right) = -\frac{(\Psi(\beta-1))^2 + \Psi(\beta) + (\beta-2)\Psi(\beta)\Psi(\beta-1)}{(\Psi(\beta) - \Psi(\beta-1))^2}.$$

Similarly, the Pearson correlation coefficient between $\bar{B}$ and $\bar{D}$ can be derived; see Figure 3. The figure shows that the correlation is both negative and significant for small β, but that $\bar{B}$ and $\bar{D}$ become independent as $\beta \to \infty$. This behavior also follows from the constraint $B_n + D_n \leqslant C_n$: it yields negative correlation, but it vanishes as $\beta \to \infty$. In that case, C_n becomes large with respect to B_n and D_n, nullifying the constraint and removing the source of dependence.

6 NEARLY-INSTANTANEOUS REPAIR ($\alpha \geqslant 1$)

Since Theorem 1 only covers $\alpha < 1$, it remains to analyze the opposite case in this section. Recall that in Section 2.3, we introduced normalized versions of B_n and D_n in order to preserve finite mean. When $\alpha \geqslant 1$ however, the repair speed is so large that the repair queue size D_n converges to a finite mean random variable (which was not the case before). In fact, if $\alpha > 1$, the distribution D_n converges to a degenerate distribution with value 0. Because of this, it is no longer necessary to normalize D_n. We will therefore consider the joint distribution of the random variables $\bar{B}_n$ and D_n.

In this regime, a statement similar to Theorem 1 holds, which is given in the following proposition. For convenience we write $U_n(s,t) := \mathbb{E}(e^{-s\bar{B}_n}e^{-tD_n})$, where D_n is unscaled.

Proposition 2. *For* $\alpha \geqslant 1$,

$$U_n(s,t) \sim e^{\frac{1}{2}s^2}\frac{\Phi(\beta+s)}{\Phi(\beta)}\frac{1-(1-\gamma_n)}{1-(1-\gamma_n)e^{-t}}$$

as $n \to \infty$.

Proof. The proof relies heavily on its counterpart for $\alpha < 1$. Observe that $U_n(s,t) = P_n(s, t/\gamma_n)$. When replacing t by t/γ_n, the result immediately follows from Lemma 1, Equation (10) and Lemmas 3 and 4. □

As in Corollary 1, we recognize the Laplace-Stieltjes transforms of a truncated standard-normal distribution, and of a geometric with parameter γ_n, so that $\bar{B}_n$ and D_n are asymptotically independent. In addition, as $n \to \infty$, $\bar{B}_n$ converges in distribution to $(\mathscr{N} \mid \mathscr{N} \leqslant \beta)$, whereas D_n behaves as $\mathscr{G}(\gamma_n)$. As expected, for $\alpha > 1$, $\gamma_n \to 0$, so that the distribution of D_n becomes degenerate.

7 NUMERICAL ILLUSTRATIONS

Corollary 1 describes the convergence in distribution for $\bar{B}_n$ and $\bar{D}_n$ when $\alpha \neq \frac{1}{2}$. However, there is no guarantee on the speed at which these random variables converge to their limits. In this section, we assess the pre-limit distributions by means of numerical experiments. In doing so, we depend on simulation to evaluate (1), since the exact computation of the normalization constant is challenging. Indeed, the sums in (3) contain many terms involving computationally demanding powers and factorials.

In Figure 4 the dots represent an estimated density for $\bar{B}_n$. The left plot results from taking $\alpha = 1$, and the standard-normal density is added for comparison (more precisely, the line is actually the standard-normal density times the constant $1/\Phi(\beta)$, so that it matches with the truncated normal density for values smaller than β). The simulation results almost coincide with the truncated normal density, entailing that the distribution of $\bar{B}_n$ is close to its limiting distribution. The right plot corresponds to $\alpha = \frac{1}{2}$, and shows a density estimation that is neither a standard-normal nor a truncated version. Its decline due to the constraint $B_n + D_n \leqslant C_n$ is well visible, yet not as sharp as for $\alpha > \frac{1}{2}$.

Figure 5 shows two similar plots for $\bar{D}_n$. Left, where $\alpha = -2$, the simulation results are close to the truncated exponential density. There is no truncation in the right plot ($\alpha = \frac{1}{2}$), but the estimated density of $\bar{D}_n$ is significantly steeper than an exponential density.

As one might expect, the convergence is slower when α is close to $\frac{1}{2}$. This holds in particular for $\bar{D}_n$, where we need to set α significantly below $\frac{1}{2}$.

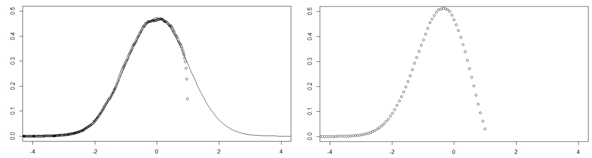

Figure 4: Simulated density of $\bar{B}_n$ (dots) with $\beta = 1$. On the left for $\alpha = 1$ and $n = 2000$, compared to the appropriate normal density form (line), on the right for $\alpha = \frac{1}{2}$ and $n = 500$.

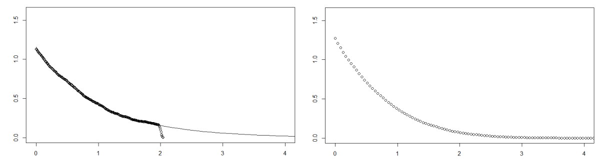

Figure 5: Simulated density of $\bar{D}_n$ (dots) with $\beta = 2$. On the left for $\alpha = -2$ and $n = 5$, compared to the appropriate exponential density form (line), on the right for $\alpha = \frac{1}{2}$ and $n = 1000$.

ACKNOWLEDGMENTS

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

REFERENCES

[1] A. Aveklouris, M. Vlasiou, and B. Zwart. 2017. Heavy-traffic approximations for a layered network with limited resources. *Prob. and Math. Stat.* 37 (2017), 497–532.
[2] A. Aveklouris, M. Vlasiou, and B. Zwart. 2019. Bounds and limit theorems for a layered queueing model in electric vehicle charging. *Queueing Syst.* 93 (2019), 83–137.
[3] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. 1975. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM* 22 (1975), 248–260.
[4] A. Birman and Y. Kogan. 1996. Error bounds for asymptotic approximations of the partition function. *Queueing Syst.* 23 (1996), 217–234.
[5] R. J. Boucherie and N. M. van Dijk. 2010. *Queueing Networks: A Fundamental Approach* (1st. ed.). Springer.
[6] A. Dembo and O. Zeitouni. 1998. *Large Deviations Techniques and Applications.* Springer.
[7] J.L. Dorsman. 2015. *Layered queueing networks: performance modelling, analysis and optimisation.* Ph.D. Dissertation. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR784295
[8] D. Fakinos. 1983. The $M/G/k$ loss system with servers subject to breakdowns. *J. Appl. Probab.* 20 (1983), 706–712.
[9] W. Feller. 1968. *An Introduction to Probability Theory and Its Applications* (3rd. ed.). Vol. 1. Wiley.
[10] S. Halfin and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29 (1981), 567–587.
[11] L. Haque and M.J. Armstrong. 2007. A survey of the machine interference problem. *Eur. J. Oper. Res.* 179 (2007), 469–482.
[12] A. Mandelbaum, W.A. Massey, and M. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Syst.* 30 (1998), 149–201.
[13] D. Mitra and J. McKenna. 1986. Asymptotic expansions for closed Markovian networks with statedependent service rates. *J. ACM* 33 (1986), 568–592.
[14] R. Nelson. 1995. *Probability, Stochastic Processes, and Queueing Theory.* Springer.
[15] L. Takács. 1962. *Introduction to the Theory of Queues.* Oxford University Press.
[16] D. Williams. 1991. *Probability with Martingales.* Cambridge University Press.