The Markov-modulated Erlang loss system

M Mandjes, P G Taylor, Koen de Turck

To cite this version:

M Mandjes, P G Taylor, Koen de Turck. The Markov-modulated Erlang loss system. Performance Evaluation, 2017, 116 (nov), pp.53-69. 10.1016/j.peva.2017.08.005 . hal-01580374

HAL Id: hal-01580374
https://hal.science/hal-01580374

Submitted on 1 Sep 2017

The Markov-modulated Erlang Loss System

M. Mandjes [•] [⋆], P.G. Taylor [†], K. De Turck [○]

July 4, 2017

Abstract

This paper focuses on a loss system in which both the arrival rate and the per-customer service rate vary according to the state of an underlying finite-state, continuous-time Markov chain.
Our first contribution consists of a closed-form expression for the stationary distribution of this Markov-modulated Erlang loss queue. This, in particular, provides us with an explicit formula for the probability that the queue is full, which can be regarded as the Markov-modulated counterpart of the well-known Erlang loss formula. It facilitates the computation of the probability that an arbitrary arriving customer is blocked.

Furthermore, we consider a regime where, in a way that is common for this type of loss system, we scale the arrival rate and the number of servers, while also scaling the transition rates of the modulating Markov process. We establish convergence of the stationary distribution to a truncated Normal distribution, which leads to an approximation for the blocking probability. In this 'fast regime', the parameters of the limiting distribution critically depend on the precise scaling imposed. We also derive scaling results for a 'slow regime', in which the modulating Markov process is slow relative to the arrival process.

Numerical experiments show that the resulting approximations are highly-accurate.

KEYWORDS. Markov-modulation ○ loss systems ○ Erlang loss formula ○ scaling limits.

[•] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands. *Email*: `m.r.h.mandjes@uva.nl`

[⋆] CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands.

[†] School of Mathematics and Statistics, University of Melbourne, Victoria, 3010, Australia. *Email*: `p.taylor@ms.unimelb.edu.au`

[○] Laboratoire Signaux et Systèmes (L2S, CNRS UMR8506), École CentraleSupélec, Université Paris Saclay, 3 Rue Joliot Curie, Plateau de Moulon, 91190 Gif-sur-Yvette, France. *Email*: `koen.deturck@l2s.centralesupelec.fr`

M. Mandjes is also with EURANDOM, Eindhoven University of Technology, Eindhoven, the Netherlands, and Amsterdam Business School, Faculty of Economics and Business, University of Amsterdam, Amsterdam, the Netherlands.

M. Mandjes' research is partly funded by the NWO Gravitation project NETWORKS, grant number 024.002.003.

P.G. Taylor's research is supported by the Australian Research Council (ARC) Laureate Fellowship FL130100039 and the ARC Centre of Excellence for the Mathematical and Statistical Frontiers (ACEMS).

This research was partly performed when K. De Turck was a Postdoctoral Fellow of *Fonds Wetenschappelijk Onderzoek / Research Foundation – Flanders*.

1 Introduction

The *Erlang loss model*, or M/M/C/C queue, dates back to Erlang's original paper [9] in 1917. It models a setting in which calls arrive at a telephone exchange according to a Poisson process with rate λ, call holding times are exponentially distributed with mean μ, and C lines are available. The blocking probability turns out to have the clean, explicit form embodied in the expression

$$p_{\rm bl} = \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C \Bigg/ \sum_{\ell=0}^{C} \frac{1}{\ell!}\left(\frac{\lambda}{\mu}\right)^\ell. \tag{1}$$

Equation (1) is often referred to as the *Erlang-B formula*. Since Erlang's time, the model has been broadly applied for dimensioning circuit-switched telephone networks. Arguably, (1) is the most frequently used formula in communications engineering.

The model has been extended in many directions. In the first place, Erlang himself analyzed the model in which customers who find all servers busy are not lost but rather join a queue, leading to the *Erlang-C formula*. Later, it was found that $p_{\rm bl}$ depends on the service time distribution only through its mean; for nice accounts of this remarkable insensitivity property, see [15, 20].

A further generalization is the loss network [12], which was defined about 25 years ago to study the probabilistic properties of a network with R user classes (characterized by their Poisson arrival processes and exponential service times) and K links, with C_k circuits being available at link k. On arrival, a user of class r seizes a specified number A_{rk} circuits from each link k if this number of circuits is available, or is rejected and 'lost' from the system if there is a link ℓ at which the required $A_{r\ell}$ circuits are not free. Loss networks also have the insensitivity property with respect to the distribution of the holding times of each of the user classes [12].

Even with the advent of modern packet-switched data networks, the loss model (together with its many variants) has many uses in communications engineering via concepts such as *effective bandwidth* [8], and also in a broad range of other resource allocation problems.

One obvious variant of the Erlang loss model is the *Markov-modulated* M/M/C/C queue. The point of difference between this model and the standard Erlang loss model is that the arrival and service rates vary over time. More specifically, the system has Poisson arrivals and exponential service time distributions, but with the special feature that the arrival rates and service rates are determined by an external, independently-evolving, continuous-time Markov chain. This queue considered is still a *loss system*, with customers arriving to find all servers busy being rejected from the system and lost. However, despite the fact that M. Neuts analyzed the related Ph/M/C/$K+C$ model in his celebrated monograph [16, pp. 92–94], the Markov-modulated loss queue does not appear to have received much attention in the literature.

The evolution of the Markov-modulated M/M/C/C queue is uniquely described by the bivariate continuous-time Markov chain $(M(t), J(t))_{t\geqslant 0}$, where $M(t)$ denotes the number of customers in the system, and $J(t)$ the state of the modulating Markov chain (also referred to as the *background process*), at time t. In the literature on matrix-analytic methods, the process $(M(t))_{t\geqslant 0}$ is usually referred to as the *level process*, whereas $(J(t))_{t\geqslant 0}$ is called the *phase process.* The process $(J(t))_{t\geqslant 0}$

is assumed to be irreducible, taking values in some finite state space $S := \{1, \ldots, d\}$. We use $Q = (q_{ij})_{i,j\in S}$ to denote the generator of the modulating Markov chain. As usual, $q_{ij} \geqslant 0$ for $i \neq j$ and $Q\boldsymbol{e} = \boldsymbol{0}$, with $\boldsymbol{e}$ denoting a $(d \times 1)$ column vector of ones. The $(1 \times d)$ vector $\boldsymbol{\alpha}$ denotes the stationary distribution of the Markov chain $(J(t))_{t\geqslant 0}$: it is the unique non-negative vector satisfying $\boldsymbol{\alpha}Q = \boldsymbol{0}$ and $\boldsymbol{\alpha}\boldsymbol{e} = 1$.

The vectors $\boldsymbol{\lambda} = (\lambda_j)_{j\in S}$ and $\boldsymbol{\mu} = (\mu_j)_{j\in S}$ contain the arrival rates and per-customer service rates. At time t the arrival rate is $\lambda_{J(t)}$, whereas the total service rate at time t is $\mu_{J(t)}M(t)$, proportional to the number of customers currently present. The number of arrivals between 0 and t is Poisson with the *random* parameter

$$\int_0^t \lambda_{J(s)}\mathrm{d}s. \tag{2}$$

Under the assumption that at least one λ_j and one μ_j are positive, the irreducibility of Q leads to the fact that the whole queue is modelled by an irreducible finite state Markov chain. As a consequence, the stationary distribution

$$\pi_{k,j} := \lim_{t\to\infty} \mathrm{P}(M(t) = k, J(t) = j) \tag{3}$$

exists for $k \in \{0, \ldots, C\}$ and $j \in S$. We shall use (M, J) to denote random variables distributed according to this stationary distribution.

In this paper we provide an explicit analysis of the stationary distribution (3) of the Markov-modulated M/M/C/C queue, which we encode by its parameters $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$. In addition, we provide remarkably simple, yet accurate approximations in particular scaling regimes. More specifically, the main contributions of our work are the following.

- We derive an explicit expression for the probabilities $\boldsymbol{\pi}_k \equiv (\pi_{k,j})_{j=1,\ldots,d}$. In particular, we identify the blocking probability in closed form, which can be considered the true Markov-modulated counterpart of the classical formula (1).

- We consider the scaling $(N\boldsymbol{\lambda}, \boldsymbol{\mu}, N^f Q, C_N)$ for N large; here $C_N := N\varrho + \beta N^\gamma$, with ϱ the system's offered load $(\boldsymbol{\alpha}\boldsymbol{\lambda})/(\boldsymbol{\alpha}\boldsymbol{\mu})$, $\beta > 0$ a given constant (typically referred to as the 'hedge'), and $\gamma := \max\{1 - f/2, 1/2\}$. In this central-limit type of scaling, we show that an appropriately centered and scaled version of the stationary number of customers converges to a truncated Normal random variable. This finding also provides us with a compact, closed-form approximation of the blocking probability in this 'fast' regime. Importantly, the parameters of the truncated Normal random variable critically rely on whether $f < 1$, $f = 1$, or $f > 1$. The proof relies on a variant of the methodology developed in [4].

- In addition, we study a 'slow' regime of the type $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon Q, C)$, with $\varepsilon \downarrow 0$. Relying on the theory of singularly perturbed Markov chains, we propose explicit approximations in this case.

- The paper concludes with a series of representative experiments. In both the fast and slow regime, the proposed approximations perform remarkably well, despite their simple form.

The organization of our paper is as follows. Section 2 presents the explicit results for the non-scaled model $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$. Results for the fast and slow regimes are covered in Section 3. The numerical experiments can be found in Section 4.

2 Exact analysis

The objective of this section is to identify the blocking probability in the Markov-modulated counterpart of the M/M/C/C queue characterized by the parameters $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$. We do so by providing an expression for the vector of probabilities $\boldsymbol{\pi}_C$ that $M = C$ while J runs over the various values in S. In addition, we identify the full distribution $\boldsymbol{\pi}_k \equiv (\pi_{k,j})$ for $k \in \{0, \dots, C\}$ and $j \in S$.

In the sequel we frequently work with probability generating functions. For $k \in \{0, \dots, C\}$, we write $\boldsymbol{\pi}_k$ for the $1 \times d$ vector whose entries are $\pi_{k,j}$, with $j \in S$. Then

$$\boldsymbol{\Pi}(z) := \sum_{k=0}^{C} \boldsymbol{\pi}_k z^k$$

is the vector-valued probability generating function of the level process, defined for all $z \in \mathbb{C}$.

We first derive the equations that determine the stationary distribution of $(M(t), J(t))$. Denoting by 1_A the indicator of the event A, it is immediate that the vectors $\boldsymbol{\pi}_k$, with $k = 0, \dots, C$, satisfy the equations

$$\boldsymbol{\pi}_k \left[Q - 1_{\{k<C\}}\text{diag}\{\boldsymbol{\lambda}\} - k\text{diag}\{\boldsymbol{\mu}\}\right] + \boldsymbol{\pi}_{k-1} 1_{\{k>0\}}\text{diag}\{\boldsymbol{\lambda}\} + \boldsymbol{\pi}_{k+1} 1_{\{k<C\}}(k+1)\text{diag}\{\boldsymbol{\mu}\} = 0, \quad (4)$$

with diag$\{\boldsymbol{\eta}\}$ the diagonal matrix whose entries are given by the vector $\boldsymbol{\eta}$. Multiplying Equation (4) by z^k and summing over $k = 0, \dots, C$, we derive the fact that $\boldsymbol{\Pi}(z)$ satisfies

$$\boldsymbol{\Pi}(z)Q + (z-1)\left[\left(\boldsymbol{\Pi}(z) - \boldsymbol{\pi}_C z^C\right)\text{diag}\{\boldsymbol{\lambda}\} - \frac{\mathrm{d}\boldsymbol{\Pi}(z)}{\mathrm{d}z}\text{diag}\{\boldsymbol{\mu}\}\right] = \mathbf{0}. \quad (5)$$

By substituting $z = 1$ in Equation (5), we obtain the equation

$$\boldsymbol{\Pi}(1)Q = \mathbf{0}, \quad (6)$$

which makes sense, because $\boldsymbol{\Pi}(1) = \sum_{k=0}^{C} \boldsymbol{\pi}_k$ is the marginal stationary distribution $\boldsymbol{\alpha}$ of the phase, which must satisfy $\boldsymbol{\alpha}Q = \mathbf{0}$. Combining this with the obvious identity $\boldsymbol{\Pi}(1)\boldsymbol{e} = 1$, Equation (6) uniquely defines $\boldsymbol{\Pi}(1) = \boldsymbol{\alpha}$.

Differentiating Equation (5) with respect to z and putting $z = 1$ we see that

$$\boldsymbol{\Pi}'(1)\left[Q - \text{diag}\{\boldsymbol{\mu}\}\right] + \left(\boldsymbol{\Pi}(1) - \boldsymbol{\pi}_C\right)\text{diag}\{\boldsymbol{\lambda}\} = \mathbf{0}. \quad (7)$$

Because $Q -$ diag$\{\boldsymbol{\mu}\}$ is an irreducible generator with at least one row sum strictly less than zero, it must be non-singular (see, for example, [17, pp. 8 and 31]). It follows that we can rewrite Equation (7) as

$$\boldsymbol{\Pi}'(1) = \left[\boldsymbol{\Pi}(1) - \boldsymbol{\pi}_C\right]\text{diag}\{\boldsymbol{\lambda}\}\left[\text{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}. \quad (8)$$

We can also derive a similar relationship for the higher derivatives of $\mathbf{\Pi}(z)$, again evaluated at $z = 1$. Repeated differentiation of (5) shows us that, for $k = 1, \ldots, C - 1$,

$$\begin{aligned}\mathbf{\Pi}^{(k)}(z)Q + k&\left[\left(\mathbf{\Pi}^{(k-1)}(z) - \frac{C!}{(C-k+1)!}z^{C-k+1}\boldsymbol{\pi}_C\right)\mathrm{diag}\{\boldsymbol{\lambda}\} - \mathbf{\Pi}^{(k)}(z)\mathrm{diag}\{\boldsymbol{\mu}\}\right] \\ + \quad (z-1)&\left[\left(\mathbf{\Pi}^{(k)}(z) - \frac{C!}{(C-k)!}z^{C-k}\boldsymbol{\pi}_C\right)\mathrm{diag}\{\boldsymbol{\lambda}\} - \mathbf{\Pi}^{(k+1)}(z)\mathrm{diag}\{\boldsymbol{\mu}\}\right] = \mathbf{0},\end{aligned}$$

where we have written $\mathbf{\Pi}^{(k)}(z)$ for the k-th derivative of $\mathbf{\Pi}(z)$. Inserting $z = 1$ we see that

$$\mathbf{\Pi}^{(k)}(1)\left[Q - k\cdot\mathrm{diag}\{\boldsymbol{\mu}\}\right] + k\left(\mathbf{\Pi}^{(k-1)}(1) - \frac{C!}{(C-k+1)!}\boldsymbol{\pi}_C\right)\mathrm{diag}\{\boldsymbol{\lambda}\} = 0. \tag{9}$$

The matrix $Q - k \cdot \mathrm{diag}\{\boldsymbol{\mu}\}$ is nonsingular for the same reason that $Q - \mathrm{diag}\{\boldsymbol{\mu}\}$ is nonsingular, and so we can write

$$\mathbf{\Pi}^{(k)}(1) = k\left(\mathbf{\Pi}^{(k-1)}(1) - \frac{C!}{(C-k+1)!}\boldsymbol{\pi}_C\right)\mathrm{diag}\{\boldsymbol{\lambda}\}\left[k\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}. \tag{10}$$

We are now in a position to determine the vector $\boldsymbol{\pi}_C$. It follows from Equations (8) and (10) that

$$\begin{aligned}\mathbf{\Pi}^{(k)}(1) &= k!\,\mathbf{\Pi}(1)\prod_{i=1}^{C}\left(\mathrm{diag}\{\boldsymbol{\lambda}\}\left[i\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}\right) \\ &- k!\,\boldsymbol{\pi}_C\sum_{\ell=1}^{k}\binom{C}{\ell-1}\prod_{i=\ell}^{k}\left(\mathrm{diag}\{\boldsymbol{\lambda}\}\left[i\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}\right)\end{aligned} \tag{11}$$

where the matrix products multiply *on the right* as i runs through successive values.

As an immediate consequence of the fact that the level process attains values in $\{0, \ldots, C\}$, observe that, for any value of z,

$$\mathbf{\Pi}^{(C)}(z) = \frac{\mathrm{d}^C}{\mathrm{d}z^C}\sum_{k=0}^{C}\boldsymbol{\pi}_k z^k = C!\,\boldsymbol{\pi}_C.$$

Substituting this into Equation (11), we see that

$$\boldsymbol{\pi}_C\left[\sum_{\ell=0}^{C}\binom{C}{\ell}\prod_{i=\ell+1}^{C}\left(\mathrm{diag}\{\boldsymbol{\lambda}\}\left[i\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}\right)\right] = \boldsymbol{\alpha}\prod_{i=1}^{C}\left(\mathrm{diag}\{\boldsymbol{\lambda}\}\left[i\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}\right), \tag{12}$$

with the empty product taken to be the identity matrix. We thus arrive at the following theorem.

Theorem 2.1. *In the Markov-modulated* M/M/C/C *queue with parameters* $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$*, the vector* $\boldsymbol{\pi}_C$ *satisfies equation (12).*
If the matrix

$$A(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C) := \sum_{\ell=0}^{C}\binom{C}{\ell}\prod_{i=\ell+1}^{C}\left(\mathrm{diag}\{\boldsymbol{\lambda}\}\left[i\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}\right) \tag{13}$$

is nonsingular, then $\boldsymbol{\pi}_C$ *is given by the expression*

$$\boldsymbol{\pi}_C = \boldsymbol{\alpha}\prod_{i=1}^{C}\left(\mathrm{diag}\{\boldsymbol{\lambda}\}\left[i\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}\right)\left[\sum_{\ell=0}^{C}\binom{C}{\ell}\prod_{i=\ell+1}^{C}\left(\mathrm{diag}\{\boldsymbol{\lambda}\}\left[i\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}\right)\right]^{-1}. \tag{14}$$

Remark 2.1. The expression (14) is the Markov-modulated analogue of the celebrated *Erlang loss formula*. The fraction of time that all servers are busy and the background process is in phase j is given by $(\boldsymbol{\pi}_C)_j$ and the fraction of time that all servers are busy irrespective of the state of the background process is $\boldsymbol{\pi}_C\,\boldsymbol{e}$. In all instances we considered, the matrix $A(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$ turned out to be nonsingular, and we conjecture that this is true for all $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$. However, a proof of this conjecture has eluded us to date.

Expression (14) also provides us with an explicit formula for the customer blocking probability $p_{\rm bl}$, defined as the probability that an arbitrary arriving customer finds all servers busy. It is obtained by weighting the probabilities $\pi_{C,j}$ by the fractions $\lambda_j/\boldsymbol{\alpha}\,\boldsymbol{\lambda}$.

Corollary 2.1. *The blocking probability $p_{\rm bl}$ is given by*

$$p_{\rm bl} = \frac{\boldsymbol{\pi}_C\,\boldsymbol{\lambda}}{\boldsymbol{\alpha}\,\boldsymbol{\lambda}} = \sum_{j=1}^{d} \pi_{C,j}\lambda_j \Bigg/ \sum_{j=1}^{d} \alpha_j\lambda_j\ .$$

In fact, the above analysis not only yields the vector $\boldsymbol{\pi}_C$, but actually all $\boldsymbol{\pi}_k$ for $k \in \{0, \dots, C\}$. To this end, realize that

$$\bar{\boldsymbol{\pi}}_k := \boldsymbol{\Pi}^{(k)}(1) = \frac{{\rm d}^k}{{\rm d}z^k} \sum_{\ell=0}^{C} \boldsymbol{\pi}_\ell z^\ell \Bigg|_{z\uparrow 1} = \sum_{\ell=k}^{C} \frac{\ell!}{(\ell-k)!}\,\boldsymbol{\pi}_\ell;$$

having identified the vectors $\bar{\boldsymbol{\pi}}_k$, it is trivial to obtain $\boldsymbol{\pi}_k$. Directly from Equation (11) and Theorem 2.1,

$$\begin{aligned} \boldsymbol{\Pi}^{(k)}(1) \;&=\; k!\,\boldsymbol{\pi}_C \left[\sum_{\ell=0}^{C} \binom{C}{\ell} \prod_{i=\ell+1}^{C} {\rm diag}\{\boldsymbol{\lambda}\}\left[i\cdot{\rm diag}\{\boldsymbol{\mu}\} - Q\right]^{-1} \right] - \\ &\; k!\,\boldsymbol{\pi}_C \left[\sum_{\ell=0}^{k-1} \binom{C}{\ell} \prod_{i=\ell+1}^{k} {\rm diag}\{\boldsymbol{\lambda}\}\left[i\cdot{\rm diag}\{\boldsymbol{\mu}\} - Q\right]^{-1} \right]. \end{aligned}$$

In the special case where $\lambda_i \equiv \lambda$ and $\mu_i \equiv \mu$, for all $i \in S$, we can verify that

$$\bar{\boldsymbol{\pi}}_k = \frac{1}{B(C)} \sum_{\ell=k}^{C} \frac{1}{(\ell-k)!}\left(\frac{\lambda}{\mu}\right)^{\ell},\quad B(C) := \sum_{\ell=0}^{C} \frac{1}{\ell!}\left(\frac{\lambda}{\mu}\right)^{\ell}.$$

In particular, as desired, $\bar{\boldsymbol{\pi}}_C = C!\,\boldsymbol{\pi}_C = (\lambda/\mu)^C/B(C)$, demonstrating that the classical Erlang loss formula (1) is indeed a special case of our model.

We can find the moments of the stationary number of clients in the system M, jointly with the state J of the background process. It takes an elementary computation to verify that

$$\begin{aligned} \left(\mathbb{E}M1_{\{J=1\}}, \dots, \mathbb{E}M1_{\{J=d\}}\right) \;&=\; (\boldsymbol{\alpha} - \boldsymbol{\pi}_C)\,{\rm diag}\{\boldsymbol{\lambda}\}\left[{\rm diag}\{\boldsymbol{\mu}\} - Q\right]^{-1} \\ &=\; \left(\sum_{k=0}^{C-1} \boldsymbol{\pi}_k\right) {\rm diag}\{\boldsymbol{\lambda}\}\left[{\rm diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}, \end{aligned}$$

and likewise

$$
\begin{aligned}
\big(\mathbb{E}M(M-1)1_{\{J=1\}},\dots,\mathbb{E}M(M-1)1_{\{J=d\}}\big) &= 2\left(\mathbf{\Pi}^{(1)}(1) - C\boldsymbol{\pi}_C\right)\mathrm{diag}\{\boldsymbol{\lambda}\}\left[2\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1} \\
&= 2\left(\sum_{k=0}^{C-1} k\,\boldsymbol{\pi}_k\right)\mathrm{diag}\{\boldsymbol{\lambda}\}\left[2\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - Q\right]^{-1}.
\end{aligned}
$$

These expressions facilitate the computation of $\mathbb{V}\mathrm{ar}\,M$.

Remark 2.2. The naïve way to numerically compute the process's equilibrium distribution amounts to solving the $(C+1)d$-dimensional system of balance equations, which (in its basic form) takes $O(C^3d^3)$ operations. This can already be substantially be reduced by exploiting the system's band structure; from the results in e.g. [18] it immediately follows that in our model the computational cost drops to $(C+1)d(d+1)^2/4 = O(Cd^3)$. Careful counting reveals that evaluation of (14) has complexity $O(Cd^3)$ as well. This means that such methods may be prohibitively slow when C or d is large (where the value of d has more impact than the value of C).

3 Asymptotic analysis

As argued in Remark 2.2, the evaluation of the quantities derived in the previous section may be time consuming in specific regimes, in particular when C and/or d is large. This motivates interest in considering scalings under which more explicit results can be derived. In this section we consider two scalings, which we refer to as the *fast regime* and the *slow regime*. The fast scaling, which directly relates to the scalings proposed in [1, 4, 5], involves blowing up the arrival rates and the transition rates of the background process, but, importantly, *at different rates.* In the slow regime, the transitions of the background process are made increasingly rare. For convenience, we assume henceforth that $\lambda_j > 0$ and $\mu_j > 0$ for all $j \in S$.

3.1 The fast regime

We proceed by introducing the scaling of the fast regime in more detail; it involves adaptation of (i) the arrival rates $\boldsymbol{\lambda}$, (ii) the transition rate matrix Q, and (iii) the capacity C. We first define by ϱ the load imposed on the system: with $\lambda_\infty := \boldsymbol{\alpha}\,\boldsymbol{\lambda} = \sum_{i\in S}\alpha_i\lambda_i$ and $\mu_\infty := \boldsymbol{\alpha}\,\boldsymbol{\mu} = \sum_{i\in S}\alpha_i\mu_i$, this offered load is defined as

$$\varrho := \frac{\lambda_\infty}{\mu_\infty}.$$

The scaling regime is now defined as follows:

- We linearly scale the arrival rates by a factor N: $\boldsymbol{\lambda} \mapsto N\boldsymbol{\lambda}$ (and leave the service rates unaltered).
- The jumps of the modulating process are sped up by a factor N^f, for some $f > 0$, which amounts to $Q \mapsto N^f Q$.
- In, for example, [4] it was pointed out that in the corresponding *infinite-server* queue the variance of the stationary number of customers essentially grows like $N^{2\gamma}$, with the parameter

γ defined as $\frac{1}{2}\max\{2-f,1\} \in [\frac{1}{2},1]$. This motivates us to consider a scaling of the capacity $C \mapsto C_N$, with

$$C_N := \lfloor N\varrho + \beta N^{\gamma} \rfloor,$$

for $\beta > 0$. The parameter β is usually referred to as the *hedge*: the higher the hedge, the lower the blocking probability.

As both the arrival process and the jump process of the modulating Markov chain $(J(t))_{t\geqslant 0}$ are accelerated (albeit at different rates), this regime can be thought of as being *fast*.

Denote the stationary number of clients in the resulting scaled system by the random variable $M^{(N)} \in \{0,\ldots,C_N\}$. In self-evident notation, the objective of this section is to study the stationary number of clients in the scaled version $(N\boldsymbol{\lambda}, \boldsymbol{\mu}, N^f Q, C_N)$ of our original system $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$, when N grows large.

We now further motivate the relevance of this scaling regime, by pointing at a link with the phenomenon of *overdispersion.* As was mentioned in the introduction, the number of arrivals between times 0 and t has a mixed Poisson distribution, where the random parameter is given by (2). In the asymptotic regime introduced above, more explicit statements can be made: as shown in [14], the number of arrivals up to time t in the scaled system, denoted by $A^{(N)}(t)$, obeys a (functional) central limit theorem with scale function N^{γ}. More specifically, as N grows large,

$$\frac{A^{(N)}(t) - N\lambda_\infty t}{N^{\gamma}}$$

converges to a zero-mean Normally distributed random variable. It entails the fact that the mean of the number of arrivals grows essentially linearly in N, whereas the variance behaves as $N^{2\gamma}$. We observe that the case where $f > 1$ (and therefore $\gamma = \frac{1}{2}$) corresponds to the 'traditional Poisson regime', in which the mean and the variance of the number of arrivals roughly match. The situation where $f < 1$ (in which $\gamma > \frac{1}{2}$), however, exhibits overdispersion [3]: the variance grows faster than the mean. The phenomenon of overdispersion has been observed in various operational contexts such as call center data [6] and hospital arrivals [13].

The boundary case when $f = 1$ is more subtle and has to be dealt with separately: the variance behaves linearly, as is the case for $f > 1$, but the system is more variable than in the Poisson regime. This is expressed by the fact that the mean and variance are both essentially proportional to N, but the proportionality constant corresponding to the variance is larger than that corresponding to the mean.

Before we can state our main scaling result, we have to introduce new notation. Define by D the *deviation matrix* corresponding to the background process $J(\cdot)$; its (i,j)-th entry is given by

$$D_{ij} = \int_0^\infty \left((e^{Qt})_{ij} - \alpha_j\right)\,\mathrm{d}t,$$

for $i,j = 1,\ldots,d$. This allows us to introduce $\sigma^2 := \tau^2\, 1_{\{f\leqslant 1\}} + \varrho\, 1_{\{f \geqslant 1\}}$, with

$$\tau^2 := \frac{U}{\mu_\infty}, \quad \text{where } \; U := \boldsymbol{\alpha}\,(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\,D\,(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\,\boldsymbol{e}; \tag{15}$$

below, in Corollary 3.1, we show that U plays a central role in an asymptotically exact approximation of $p_{\rm bl}^{(N)}$.

Our main result in this section shows that an appropriately centered and normalized version of $M^{(N)}$ converges in distribution to a truncated Normal random variable; the variance of this Normal distribution critically depends on whether f is smaller or larger than 1. Here and in the sequel, let $\mathcal{N}(\mu, \sigma^2)$ denote a random variable that has a Normal distribution with mean μ and variance σ^2.

Proposition 3.1. *Consider the scaling $(N\boldsymbol{\lambda}, \boldsymbol{\mu}, N^f Q, C_N)$. The random variable*

$$\bar{M}^{(N)} := \frac{M^{(N)} - N\varrho}{N^\gamma}$$

converges, as $N \to \infty$, to $(\mathcal{N}(0,\sigma^2) \,|\, \mathcal{N}(0,\sigma^2) \leqslant \beta)$.

With $\boldsymbol{\pi}_k^{(N)}$ denoting the counterpart of $\boldsymbol{\pi}_k$ in the N-scaled model, define

$$\boldsymbol{P}^{(N)}(s) := \left(\mathbb{E}\, e^{-s\bar{M}^{(N)}} 1_{\{J=1\}}, \ldots, \mathbb{E}\, e^{-s\bar{M}^{(N)}} 1_{\{J=d\}}\right) = \sum_{k=0}^{C_N} \exp\left(-s\left(\frac{k - N\varrho}{N^\gamma}\right)\right) \boldsymbol{\pi}_k^{(N)}.$$

In order to prove Proposition 3.1, by Lévy's convergence theorem [21, Thm. 18.1] it suffices to establish the convergence

$$\boldsymbol{P}^{(N)}(s)\boldsymbol{e} \to \mathbb{E}(\exp(-s\mathcal{N}(0,\sigma^2)) \,|\, \mathcal{N}(0,\sigma^2) \leqslant \beta), \tag{16}$$

as $N \to \infty$. From (16) it then follows after a routine calculation that we should prove

$$\boldsymbol{P}^{(N)}(s)\boldsymbol{e} \to e^{\sigma^2 s^2/2} \frac{\mathbb{P}(\mathcal{N}(0,1) \leqslant \sigma s + \beta/\sigma)}{\mathbb{P}(\mathcal{N}(0,1) \leqslant \beta/\sigma)}. \tag{17}$$

Several techniques can be used to prove this convergence. Perhaps the most straightforward among these is a variation of that presented in [4], of which we have included a sketch in Appendix A.

The following corollary of the derivation in Appendix A provides us with the asymptotics of the blocking probability, with this probability in the N-scaled model denoted by

$$p_{\rm bl}^{(N)} := \frac{\boldsymbol{\pi}_C^{(N)} \boldsymbol{\lambda}}{\boldsymbol{\alpha} \boldsymbol{\lambda}} = \sum_{j=1}^d \pi_{C_N,j}^{(N)} \lambda_j \Bigg/ \sum_{j=1}^d \alpha_j \lambda_j \ .$$

Let $\phi_{\mathcal{N}}(\cdot)$ and $\Phi_{\mathcal{N}}(\cdot)$ denote the density and cumulative distribution function of a standard Normal random variable, respectively.

Corollary 3.1. *Consider the scaling $(N\boldsymbol{\lambda}, \boldsymbol{\mu}, N^f Q, C_N)$. As $N \to \infty$,*

$$N^{1-\gamma} p_{\rm bl}^{(N)} \to \frac{\mu_\infty}{\lambda_\infty} \left(\int_{-\infty}^0 e^{-r^2\sigma^2/2 - r\beta}\, {\rm d}r\right)^{-1} = \frac{\sigma}{\varrho} \frac{\phi_{\mathcal{N}}(\beta/\sigma)}{\Phi_{\mathcal{N}}(\beta/\sigma)} =: b. \tag{18}$$

The above corollary is the Markov-modulated counterpart of the Halfin-Whitt-type result for the non-modulated loss system, as has been presented in e.g. [19, Thm. 10.4.3]. The non-modulated result [19, Thm. 10.4.3] states that, under essentially the same scaling as the one that we consider, the blocking probability is inversely proportional to $\sqrt{N}$ (where the proportionality constant was explicitly given). Corollary 3.1 entails that in the Markov-modulated case the blocking probability is inversely proportional to $N^{1-\gamma}$, rather than $\sqrt{N}$, where it is noted that $N^{1-\gamma} = \sqrt{N}$ for $f \geqslant 1$.

Now we study the asymptotics of the blocking probability more closely, To this end, consider the $f > 1$, $f < 1$, and $f = 1$ cases separately. We denote by $g_N \sim h_N$ that $g_N/h_N \to 1$ as $N \to \infty$.

- For $f > 1$, Corollary 3.1 yields (realizing that $\gamma = \frac{1}{2}$),

$$p_{\mathrm{bl}}^{(N)} \sim \frac{b}{\sqrt{N}} = \frac{1}{\sqrt{N}} \frac{1}{\sqrt{\varrho}} \frac{\phi_{\mathcal{N}}(\beta/\sqrt{\varrho})}{\Phi_{\mathcal{N}}(\beta/\sqrt{\varrho})},$$

  in line with the (classical) asymptotics for the non-modulated M/M/C/C queue under a Halfin-Whitt-type scaling [11] and [19, Thm. 10.4.3]. For the non-scaled model $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$, this leads to the approximation

$$p_{\mathrm{bl}} \approx \frac{1}{\sqrt{\varrho}} \frac{\phi_{\mathcal{N}}((C-\varrho)/\sqrt{\varrho})}{\Phi_{\mathcal{N}}((C-\varrho)/\sqrt{\varrho})}. \tag{19}$$

- For $f < 1$, we have $\gamma = 1 - f/2$, and therefore Corollary 3.1 yields

$$p_{\mathrm{bl}}^{(N)} \sim \frac{b}{\sqrt{N^f}} = \frac{1}{\sqrt{N^f}} \frac{\sqrt{U\mu_\infty}}{\lambda_\infty} \frac{\phi_{\mathcal{N}}(\beta\sqrt{\mu_\infty/U})}{\Phi_{\mathcal{N}}(\beta\sqrt{\mu_\infty/U})},$$

  with U as in (15). For the non-scaled model $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$, we thus obtain the approximation

$$p_{\mathrm{bl}} \approx \frac{\sqrt{U\mu_\infty}}{\lambda_\infty} \frac{\phi_{\mathcal{N}}((C-\varrho)\sqrt{\mu_\infty/U})}{\Phi_{\mathcal{N}}((C-\varrho)\sqrt{\mu_\infty/U})}. \tag{20}$$

- For $f = 1$, again due to Corollary 3.1, we see that

$$p_{\mathrm{bl}}^{(N)} \sim \frac{b}{\sqrt{N}} = \frac{1}{\sqrt{N}} \frac{1}{\sqrt{\varrho}} \sqrt{\frac{U}{\lambda_\infty} - 1}\, \frac{\phi_{\mathcal{N}}(\beta/\sqrt{\tau^2+\varrho})}{\Phi_{\mathcal{N}}(\beta/\sqrt{\tau^2+\varrho})}.$$

  We observe the $\sqrt{N}$ scaling as in the case $f > 1$, but with a different proportionality constant. This yields for the non-scaled model

$$p_{\mathrm{bl}} \approx \frac{\sqrt{U/\mu_\infty + \varrho}}{\varrho} \cdot \left[\phi_{\mathcal{N}}\left(\frac{C-\varrho}{\sqrt{U/\mu_\infty+\varrho}}\right) \Big/ \Phi_{\mathcal{N}}\left(\frac{C-\varrho}{\sqrt{U/\mu_\infty+\varrho}}\right)\right]. \tag{21}$$

In Section 4 we get back to these approximations, also proposing a single approximation that works for all three regimes simultaneously, across all values of $f > 0$.

3.2 The slow regime

A crucial characteristic of the regime discussed in Section 3.1 is that the modulating Markov chain jumps relatively fast. In this section we consider the situation in which the opposite applies: we study the scaling $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon Q, C)$ for $\varepsilon \downarrow 0$. Note that for $\varepsilon = 0$, the system decomposes into d different ergodic classes, each constituting an (unmodulated) Erlang loss model, while for any positive ε, the system is ergodic. As a consequence, the regime $\varepsilon \downarrow 0$ can be analyzed relying on the theory of *singularly perturbed Markov chains*.

To this end, let $\mathcal{Q}(\varepsilon)$ denote the family of generator matrices of the process $(M(t), J(t))_{t\geqslant 0}$, indexed by the 'rarity parameter' ε. For notational convenience, we swap levels and phases compared with the rest of the paper: the levels represent the background state and the phases represent the number of customers present. Thus, $\mathcal{Q}(\varepsilon)$ can be decomposed as follows: with $\otimes$ as usual denoting the Kronecker product,

$$\mathcal{Q}(\varepsilon) = \begin{pmatrix} \tilde{\mathcal{Q}}_1 & & \\ & \ddots & \\ & & \tilde{\mathcal{Q}}_d \end{pmatrix} + \varepsilon \cdot (Q \otimes I) =: \mathcal{Q}^{(0)} + \varepsilon \cdot \mathcal{Q}^{(1)},$$

where $\tilde{\mathcal{Q}}_i$ represents the $(C+1) \times (C+1)$-dimensional generator matrix of a (non-modulated) Erlang loss model with parameters (λ_i, μ_i, C). We let $\boldsymbol{\nu}_i$ denote the corresponding the stationary distribution vector: for $j = 0, \dots, C$,

$$\nu_{i,j} = \frac{1}{j!}\left(\frac{\lambda_i}{\mu_i}\right)^j \Bigg/ \sum_{\ell=0}^{C} \frac{1}{\ell!}\left(\frac{\lambda_i}{\mu_i}\right)^\ell .$$

The idea now is to determine the stationary distribution $\boldsymbol{\pi}(\varepsilon)$, making use of the results of [2, Section 3]. As established there, $\boldsymbol{\pi}(\varepsilon)$ is analytic around $\varepsilon = 0$, and as such admits a Taylor-series expansion:

$$\boldsymbol{\pi}(\varepsilon) = \sum_{k=0}^{\infty} \varepsilon^k \boldsymbol{\pi}^{(k)}, \tag{22}$$

for appropriate (vector-valued) coefficients $\boldsymbol{\pi}^{(k)}$. Before pointing out the main result of this subsection, we first introduce a number of matrices; D is the $(d \times d)$ deviation matrix corresponding to Q, as before:

- V and W are $d \times (C+1)d$ and $(C+1)d \times d$ matrices, respectively, defined by

$$V := \begin{pmatrix} \boldsymbol{\nu}_1^{\mathrm{T}} & & \\ & \ddots & \\ & & \boldsymbol{\nu}_d^{\mathrm{T}} \end{pmatrix}; \qquad W := \begin{pmatrix} \boldsymbol{e} & & \\ & \ddots & \\ & & \boldsymbol{e} \end{pmatrix} = I \otimes \boldsymbol{e},$$

- and $\mathcal{D}$ represents the deviation matrix for $\mathcal{Q}(0)$, with $\tilde{\mathcal{D}}_i$ the deviation matrix of $\tilde{\mathcal{Q}}_i$,

$$\mathcal{D} = \begin{pmatrix} \tilde{\mathcal{D}}_1 & & \\ & \ddots & \\ & & \tilde{\mathcal{D}}_d \end{pmatrix}.$$

Then the results of [2] imply that the vectors $\boldsymbol{\pi}^{(k)}$ have the geometric structure given by

$$\boldsymbol{\pi}^{(k)} = \boldsymbol{\pi}^{(0)} \left(\mathcal{Q}^{(1)} \mathcal{D} (I + \mathcal{Q}^{(1)} W D V) \right)^k .$$

This expression can be simplified since, after some manipulations, it turns out that $\mathcal{D}\mathcal{Q}^{(1)}WDV$ vanishes in our specific case. Indeed, as $\mathcal{Q}^{(1)} = (Q \otimes I)$ and $W = I \otimes \boldsymbol{e}$, we have

$$\mathcal{D}\mathcal{Q}^{(1)}W = \begin{pmatrix} \tilde{\mathcal{D}}_1 & & \\ & \ddots & \\ & & \tilde{\mathcal{D}}_d \end{pmatrix} \cdot (Q \otimes \boldsymbol{e}) = 0,$$

where the last step is due to the fact that $\tilde{\mathcal{D}}_i \boldsymbol{e} = 0$. We obtain the following result.

Proposition 3.2. *Consider the scaling $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon Q, C)$. As $\varepsilon \downarrow 0$, $\boldsymbol{\pi}(\varepsilon)$ admits the Taylor-series expansion (22), where*

$$\boldsymbol{\pi}^{(k)} = \boldsymbol{\pi}^{(0)} \left(\mathcal{Q}^{(1)} \mathcal{D} \right)^k .$$

We now analyze the complexity of computing $\boldsymbol{\pi}^{(k)}$ relying on Prop. 3.2. Observe that (in self-evident notation) $\boldsymbol{\pi}_i^{(0)} = \alpha_i \boldsymbol{\nu}_i$, which is in line with the intuition that when the switching between different background states is very slow, the stationary distribution is effectively reached within each regime. Note that, due to the fact that the matrices $\tilde{\mathcal{Q}}_i$ are tridiagonal, computing products of the form $\boldsymbol{v}\tilde{\mathcal{D}}_i$ is an operation that has a computational complexity of $O(C)$, and hence computing an additional term $\boldsymbol{\pi}^{(k)}$ takes $O(Cd)$ operations. The multiplication with $\mathcal{Q}^{(1)}$ takes $O(Cd^2)$ operations, and is therefore the bottleneck. It means that when K terms of the expansion (22) are to be evaluated, the complexity is $O(d^3 + KCd^2)$. Our experiments (reported in the next section) indicate that the evaluation of $\boldsymbol{\pi}^{(k)}$ through Prop. 3.2 typically compares favorably with the complexity of the approaches discussed in Remark 2.2 (which were $O(Cd^3)$ or higher).

The above results lead to the following approximation in the slow regime.

Corollary 3.2. *Consider the scaling $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon Q, C)$. As $\varepsilon \downarrow 0$, with $\boldsymbol{\pi}(\varepsilon)$ admitting the Taylor-series expansion (22), and $\boldsymbol{\pi}^{(k)}$ as in Prop. 3.2, the following Taylor expansion applies to $p_{\rm bl}$:*

$$p_{\rm bl} = \frac{1}{\boldsymbol{\alpha}\boldsymbol{\lambda}} \sum_{j=1}^{d} \pi_{C,j}(\varepsilon) \lambda_j .$$

4 Numerical experiments

In this section we report on the output of a set of numerical experiments. Section 4.1 studies the impact of the model's parameters. In Section 4.2 we perform extensive accuracy tests for a broad range of parameter settings.

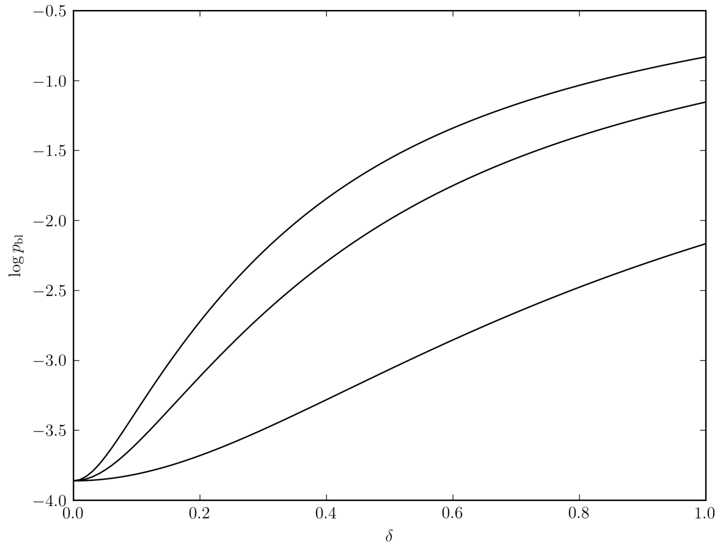

Figure 1: $\log p_{\rm bl}$ as a function $\delta \in [0, 1]$, for three values of q. The top curve corresponds to $q = \frac{1}{10}$, the middle one to $q = 1$, and the bottom one to $q = 10$.

4.1 Impact of model parameters

In this subsection we consider the following experiments:

- We first demonstrate by means of a concrete example that modulation can result in an increase as well as a decrease of the blocking probability.
- Next, we present an example that is indicative of the potential errors that would result from replacing a Markov-modulated loss system by its non-modulated counterpart, relying on the expressions obtained in Section 2. This example underscores the relevance of the model studied in this paper; naïvely using the 'classical' Erlang loss model typically leads to substantial errors.
- For the fast regime, we then study the numerical performance of approximations based on the asymptotic results derived in Section 3. Importantly, we derive a new approximation of the blocking probability that is valid across all $f > 0$, with the situations $f < 1$, $f > 1$, and $f = 1$ not needing to be distinguished, which outperforms (19)–(20).
- We conclude the section with experiments for the slow regime, showing a high degree of accuracy for small ε.

A. Modulation can affect blocking probability both ways. Following the intuition that modulation increases the variance and more variance means worse performance, one might be led to think that adding modulation (while keeping ϱ constant) will always increase the blocking probability. We show that this is not the case and that the effect can go both directions. To this end, consider a system with $C = 5$ and a three-state background chain with generator Q defined as follows:

$$Q = \begin{pmatrix} -10 & 5 & 5 \\ 1 & -2 & 1 \\ 1 & 8 & -9 \end{pmatrix}.$$

- If we take the arrival and service rates constant (thus effectively ending up in an unmodulated system) at $\lambda_1 = \lambda_2 = \lambda_3 = 2.0727$ and $\mu_1 = \mu_2 = \mu_3 = 1$, we find an offered load of $\varrho = 2.0727$, and a blocking probability of 0.0409.
- With the same C and Q but with $\lambda_1 = 12$, $\lambda_2 = 4$, $\lambda_3 = 8$, $\mu_1 = 3$, $\mu_2 = 5$, and $\mu_3 = 1$, we find that ϱ is still 2.0727, but the blocking probability drops to 0.031.
- Alternatively, with the same C and Q, but with $\lambda_1 = 0.6820$, $\lambda_2 = 2.0727$, $\lambda_3 = 3$, and $\mu_1 = \mu_2 = \mu_3 = 1$, we find again the same ϱ, but the blocking probability is now 0.0422, which is higher than in the unmodulated case.

B. Impact of burstiness on blocking probability. In this example, with $d = 2$, we assume $q_{12} = q_{21} =: q$, so that $\alpha_1 = \alpha_2 = \frac{1}{2}$, and $\mu_1 = \mu_2 = 1$. The number of servers C is set to 100. We create burstiness by assuming heterogeneity in the arrival rates: $\lambda_1 = 70 + 50\,\delta$ and $\lambda_2 = 70 - 50\,\delta$, with $\delta \in [0, 1]$. Observe that (i) the system is underloaded on average, in the sense that $\ell := (\alpha_1\varrho_1 + \alpha_2\varrho_2)/C = 0.7 < 1$ (with $\varrho_i := \lambda_i/\mu_i$), irrespective the value of δ, (ii) for $\delta = 0$ the system is an ordinary (non-modulated) Erlang-loss system operating in underload (in that $\varrho/C < 1$), (iii) for $\delta = 1$ the system alternates between 'overloaded periods' (as $\varrho_1/C = 120/100 > 1$) and 'underloaded periods' (as $\varrho_2/C = 20/100 < 1$). We conclude that the burstiness in the arrival process increases when δ goes from 0 to 1; δ can therefore be interpreted as a 'burstiness parameter'.

When evaluating the performance of this system, one could naïvely use the classical (i.e., non-modulated) Erlang-loss formula, neglecting the burstiness in the arrival process. This approach evidently gives the correct result for $\delta = 0$, but becomes increasingly questionable when δ grows. Fig. 1 shows $\log p_{\rm bl}$ as a function of $\delta \in [0, 1]$, for three values of q. We observe that the blocking probability significantly grows with δ. This effect is not that strong for large q, due to the fact that the modulating process is relatively fast: the periods with overload are so short that hardly any blocking occurs before the modulating process jumps back to the underloaded state. For small q, however, the increase of the blocking probability is rather pronounced.

C. Fast regime, convergence to truncated Normal distribution. In the second series of experiments we study the convergence of the random variable $\bar{M}^{(N)}$ to $(\mathcal{N}(0, \sigma^2) \,|\, \mathcal{N}(0, \sigma^2) \leqslant \beta)$ as N tends to ∞. We choose $\beta = 0.5$, and

$$\boldsymbol{\lambda} = \begin{pmatrix} 1.2 \\ 0.6 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} 0.6 \\ 1.8 \end{pmatrix}, \ Q = \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix}.$$

The QQ-plots in Fig. 2 illustrate the convergence as N becomes large, which is particularly fast for larger values of f.

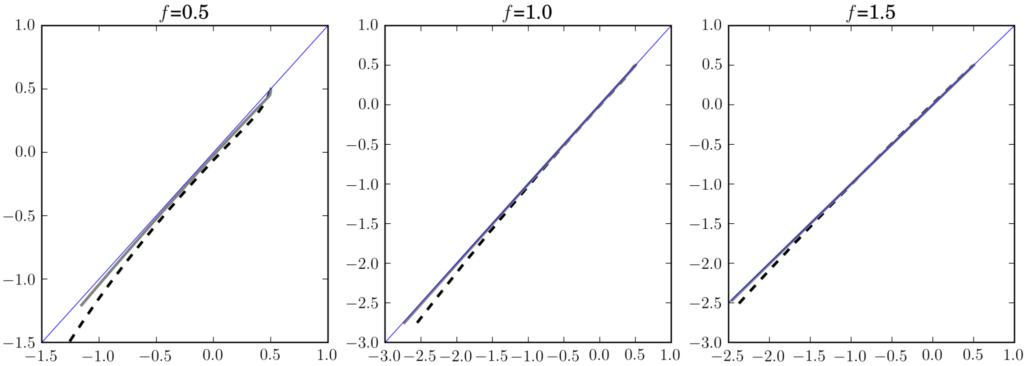

Figure 2: QQ-plots of $\bar{M}^{(N)}$ versus $(\mathcal{N}(0,\sigma^2)\,|\,\mathcal{N}(0,\sigma^2)\leqslant\beta)$. The dashed curve corresponds to $N=100$, and the solid one to $N=500$; to assess the convergence, the diagonal (in blue) has also been depicted.

D. Fast regime, convergence to the limiting variance. In the next experiments we numerically study the speed of convergence of $N^{-2\gamma}\,\mathbb{V}\mathrm{ar}\bar{M}^{(N)}$ to its limiting value. It is readily verified that, with $\eta:=\beta/\sigma$,

$$\begin{aligned} m &:= \mathbb{E}[\mathcal{N}(0,\sigma^2)\,|\,\mathcal{N}(0,\sigma^2)\leqslant\beta] = -\sigma\,\frac{\phi_{\mathcal{N}}(\eta)}{\Phi_{\mathcal{N}}(\eta)}, \\ v &:= \mathbb{V}\mathrm{ar}[\mathcal{N}(0,\sigma^2)\,|\,\mathcal{N}(0,\sigma^2)\leqslant\beta] = \sigma^2\left(1-\frac{\phi_{\mathcal{N}}(\eta)}{\Phi_{\mathcal{N}}(\eta)}\left(\eta+\frac{\phi_{\mathcal{N}}(\eta)}{\Phi_{\mathcal{N}}(\eta)}\right)\right). \end{aligned}$$

We use the same values for Q, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ as in the previous example. The numerical experiments show that $v_N:=\mathbb{V}\mathrm{ar}\,\bar{M}^{(N)}$ indeed converges to v, but that this convergence is rather slow: compare $N^{-2\gamma}\,\mathbb{V}\mathrm{ar}\,M^{(N)}$ in the top panels of Fig. 3 (the solid curves) with its theoretical limit (the blue step function, with the isolated point corresponding to $f=1$).

As we concluded above, the approximation $\mathbb{V}\mathrm{ar}\,M^{(N)}\approx N^{2\gamma}\,v$ typically performs poorly. Interestingly, however, there is a highly-accurate alternative. The idea is to replace σ^2 by its 'prelimit counterpart'

$$\sigma_N^2 := \frac{1}{\mu_\infty}\left(N^{2-f-2\gamma}U+N^{1-2\gamma}\lambda_\infty\right) = N^{2-f-2\gamma}\frac{U}{\mu_\infty}+N^{1-2\gamma}\varrho, \tag{23}$$

as in Equation (27) in Appendix A; likewise, we also replace η by $\eta_N:=\beta/\sigma_N$. This leads to the approximation:

$$\mathbb{V}\mathrm{ar}\,M^{(N)}\approx N^{2\gamma}\,v_N,\ \ \text{with}\ \ v_N:=\sigma^2\left(1-\frac{\phi_{\mathcal{N}}(\eta_N)}{\Phi_{\mathcal{N}}(\eta_N)}\left(\eta_N+\frac{\phi_{\mathcal{N}}(\eta_N)}{\Phi_{\mathcal{N}}(\eta_N)}\right)\right).$$

The plots in the bottom panels of Fig. 3 show that the approximation is remarkably accurate: the solid and dashed curves are close. Important from an application perspective is that the approximations are conservative.

E. Fast regime, approximation of the blocking probability. Here we see behavior that is very similar to that observed for the variance of $M^{(N)}$, in the sense that (i) the convergence (18) turns out to be

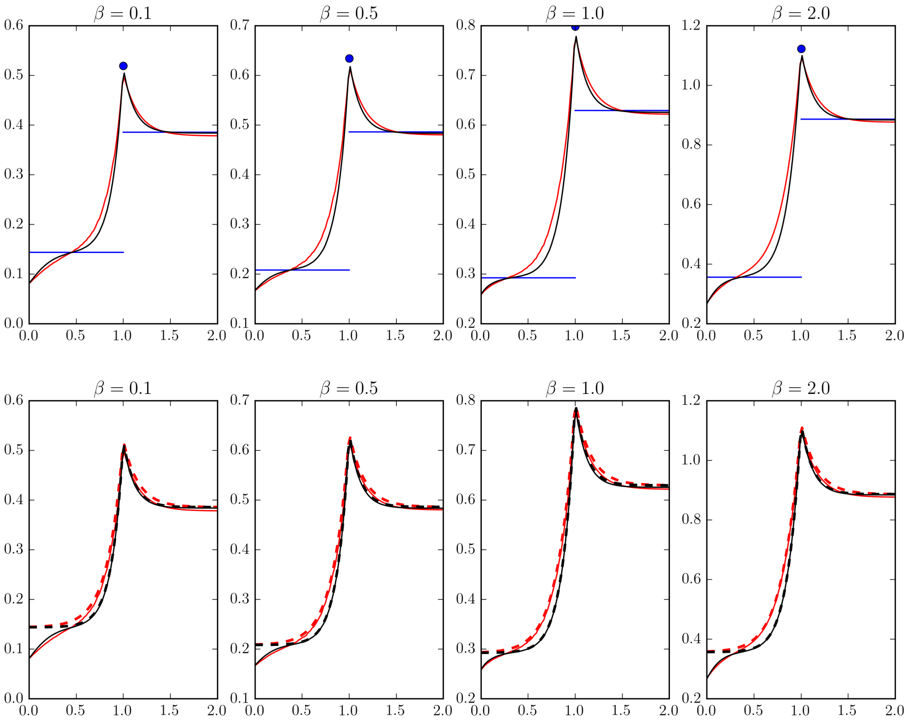

Figure 3: The solid curves are $N^{-2\gamma}\,\mathbb{V}\mathrm{ar}\,M^{(N)}$ as a function of f, for four values of β (the same in the top graphs and the bottom graphs). In the top graph $N^{-2\gamma}\,\mathbb{V}\mathrm{ar}\,M^{(N)}$ can be compared with v (in blue), and in the bottom graphs with v_N (dashed curves). The red curves correspond to $N = 500$ and the black curves to $N = 5000$.

slow, and (ii) replacing σ^2 by σ_N^2 (as given by Equation (23)) drastically improves the performance. As a result, we have the accurate approximation

$$p_{\rm bl}^{(N)} \approx N^{\gamma-1} b_N, \;\; \text{with} \;\; b_N := \frac{\sigma_N}{\varrho}\frac{\phi_{\mathbb{N}}(\eta_N)}{\Phi_{\mathbb{N}}(\eta_N)}.$$

The results displayed in Fig. 4 again show a nearly perfect fit for $f > 0.5$, and still reasonably accurate performance for $f \in (0.1, 0.5)$; again, we have used the same values for Q, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$.

Based on the above observations, we therefore propose the replacement of the approximations (19) and (20) for the blocking probability in the non-scaled model $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$ for $f > 1$ and $f < 1$ by the approximation (21) that corresponds to $f = 1$ *across all values of* $f > 0$. Throughout our experiments, we have observed that this approximation is typically highly-accurate, and when there is a bias it tends to be conservative.

Approximation 4.1. *Consider the non-scaled model* $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Q, C)$*. Then, with* $C > \varrho$*,*

$$p_{\rm bl} \approx \frac{\sqrt{U/\mu_\infty + \varrho}}{\varrho} \cdot \left[\phi_{\mathbb{N}}\left(\frac{C-\varrho}{\sqrt{U/\mu_\infty + \varrho}} \right) \Big/ \Phi_{\mathbb{N}}\left(\frac{C-\varrho}{\sqrt{U/\mu_\infty + \varrho}} \right) \right]. \tag{24}$$

F. Slow regime, convergence to the limiting variance and approximation of the blocking probability. This last example considers the approximation proposed for the slow regime. We conduct an experiment

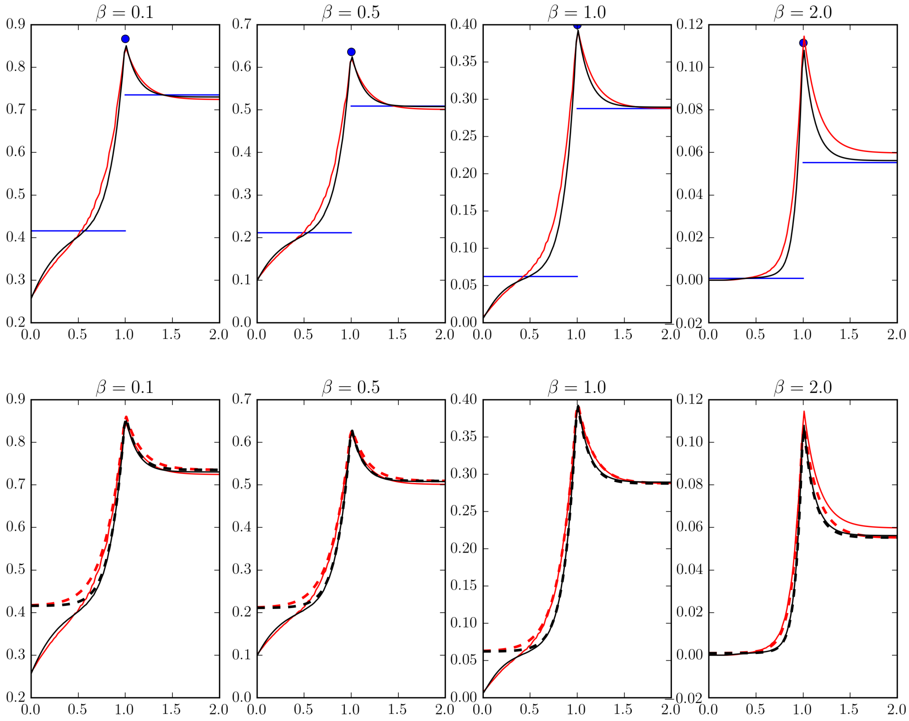

Figure 4: The solid curves are $N^{1-\gamma} p_{\mathrm{bl}}^{(N)}$ as a function of f, for four values of β (the same in the top graphs and the bottom graphs). In the top graph $N^{1-\gamma} p_{\mathrm{bl}}^{(N)}$ can be compared with b (in blue), and in the bottom graphs with b_N (dashed curves). The red curves correspond to $N = 500$ and the black curves to $N = 5000$.

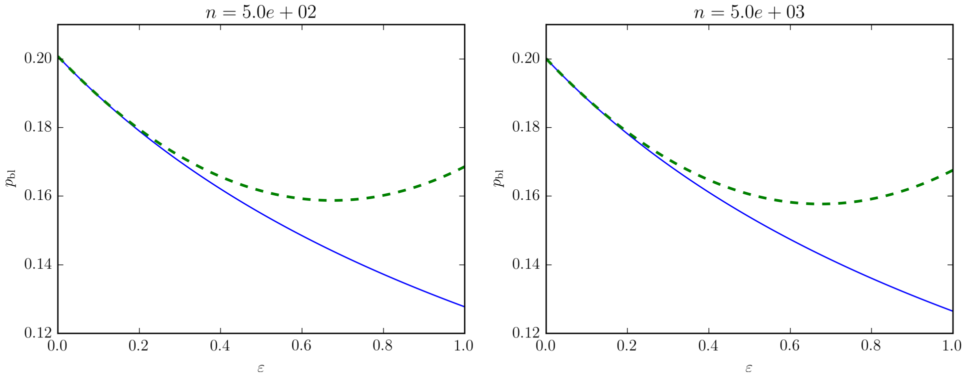

Figure 5: p_{bl} as a function of $\varepsilon \in (0, 1)$, with the quadratic quasi-stationary approximation indicated in green.

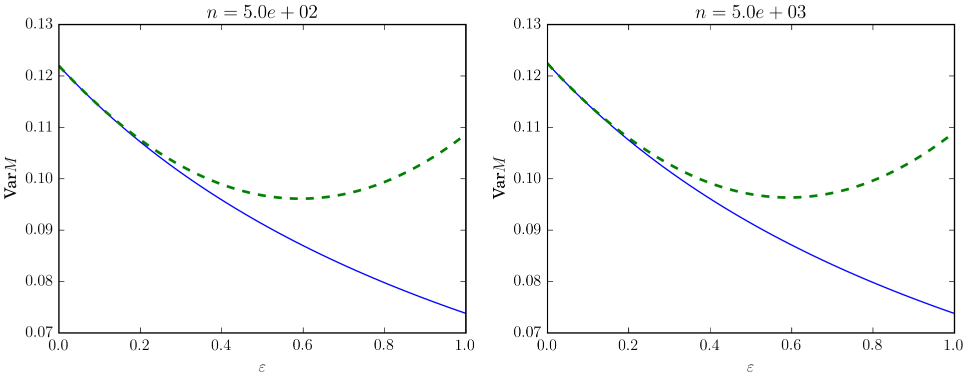

Figure 6: The variance $\mathbb{V}\mathrm{ar}\, M$ as a function of $\varepsilon \in (0, 1)$, with the quadratic quasi-stationary approximation indicated in green.

showing the good fit even a lower-order approximation can provide. With $d = 2$, we assume $q_{12} = q_{21} = 1$, so that $\alpha_1 = \alpha_2 = \frac{1}{2}$. Also, we take $\boldsymbol{\mu} = (1,\ 2)$, and $\boldsymbol{\lambda} = (2,\ 1)$, and as a consequence the offered load ϱ is equal to 1. We consider the scaling $(n\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon Q, 1.2\, n)$ and let ε be small; we take $n = 500$ and $n = 5\,000$. In Figs. 5-6 we consider the blocking probability $p_{\rm bl}$ and the variance of the stationary number of customers $\mathbb{V}\mathrm{ar}\, M$, by truncating the series (22) after the quadratic term.

First observe that the blocking probability is decreasing in ε, which can be understood as follows. The parameters are chosen such that the system is a temporarily underloaded (overloaded) when in state 1 (state 2) respectively. When ε increases, the background process jumps more frequently between the states, thus reducing the burstiness, and hence the blocking probability. A similar reasoning applies to the variance.

Figs. 5-6 show that for relatively small ε the approximation is accurate, even though just three terms, a constant, a linear, and aquadratic term have beeen included. The fit can be made more accurate for larger ε by taking into account a quadratic term.

4.2 Validation of approximation

In this subsection we discuss the performance, under a broad range of possible parameter values, of the approximation (23) of the variance of $M^{(N)}$, and the approximation (24) of $p_{\rm bl}$. Recall that these approximations rely on our result that the $\bar{M}^{(N)}$ converges to a truncated Normal random variable. As this convergence can be made arbitrarily slow (just as this can be done in a conventional central limit theorem setting), it is clear that it is possible to construct cases in which the approximations lose accuracy. More specifically, when (i) the transition rates of the modulating Markov chain are low, or when (ii) the arrival rates (and/or service rates) during different states are highly heterogeneous, we anticipate that the accuracy of our approximation is likely to degrade. We shall see that this intuition is confirmed by the experiments below.

In view of the above observations, we performed the following extensive set of numerical experiments. We chose thoroughly to examine the case where the modulating Markov chain has two states, fixing the hedge β to some values that we will detail below, and fixing f to 1 (as any other value of f can be obtained by rescaling the parameters). This brings us to six free parameters, namely

$$\lambda_1,\ \lambda_2,\ \mu_1,\ \mu_2,\ q_1 := q_{12},\ \text{ and }\ q_2 := q_{21}.$$

We bring this down to four by imposing the constraint that the average arrival rate $\alpha_1\lambda_1 + \alpha_2\lambda_2$, as well as the average service rate $\alpha_1\mu_1 + \alpha_2\mu_2$, must be equal to 1. As we scale up time and space afterwards, these constraints in fact do not restrict the parameter space, it just makes it so that for the same scaling parameter N, and identical β and f, we have the same service capacity (because the loads are the same), and the same 'relaxation rate' μ_∞ (which governs the rate of convergence to equilibrium in the model's infinite-server counterpart; see for example [4]).

This leaves us with four (positive) parameters: q_1, q_2, and the ratios $r_1 := \lambda_1/\lambda_2$ and $r_2 := \mu_1/\mu_2$. In order to explore this four dimensional parameter space, we randomly generated these parameters in the following way. Let $\nu_1, \ldots, \nu_4$ denote four standard normal random variables and $c_1, \ldots, c_4$ four constants. Then the random parameters Q_1, Q_2, R_1, R_2 were given by

$$Q_1 = 10^{c_1+\nu_1};\quad Q_2 = 10^{c_2+\nu_2};\quad R_1 = 10^{c_3+\nu_3};\quad R_2 = 10^{c_4+\nu_4}.$$

Hence, the parameters essentially adhere to a log-normal law. Choosing the constants c_i appropriately allow us to explore more extreme areas of the parameter space.

For each row of the two tables in Appendix B, we generated $50\,000$ points in parameter space. We list the constants c_i, β, and four characterizations of the relative error of (A) the standard deviation of the system content and (B) the blocking probability (indicated in the following by a superscript '(s)' and '(b)', respectively): with $\star \in \{(\mathrm{s}), (\mathrm{b})\}$,

- $p_5^\star$ and $p_{10}^\star$ denote the fraction of samples that have a relative error of more than 5% resp. 10%, and

- $E^\star$ and $\sigma^\star$ denote the average relative error and its standard deviation.

The first table shows the results for scaling parameter $N = 500$, the second for $N = 5\,000$. The general conclusion is that the approximations perform well with our procedure being set up to cover a broad range of parameter values (some of which rather extreme), typically a relatively low fraction of scenarios leads to substantial inaccuracies. In addition, the tables show that even if the low value of N brings already quite satisfactory results, they improve significantly when N is increased to 5 000. We also observe an asymmetric effect when the mean of the Q_2 samples is shifted: we see an improvement when the values are on average higher, and a marked deterioration when the values are lower. This is as could have been expected, as the convergence of the background chain will be respectively higher and lower in these cases. For the other two parameters, shifts in both directions lead to loss of accuracy.

The quality of the approximation of the standard deviation generally improves when β increases, whereas the opposite is true when the blocking probability is concerned. This can be explained from the fact that for higher β the blocking probability requires convergence of the tail of the distribution.

5 Concluding remarks

In this paper, we have presented a full analysis of the Markov-modulated counterpart of the classical Erlang loss system. Our first main result is the derivation of an explicit formula describing the stationary behavior, which also leads to a closed-form expression for the blocking probability. As a second contribution we provide simple, yet highly-accurate approximations which are provably correct under specific scalings.

The approximations relate to a fast regime (in which arrivals as well as the background process' transitions are sped up) and a slow regime (in which the background process jumps slowly). In the fast regime, the steady-state distribution tends to that of a truncated Normal random variable. Potential topics for future research are the following:

- Large deviations of the blocking probability under the scaling $(N\boldsymbol{\lambda}, \boldsymbol{\mu}, N^f Q, NC)$, where the capacity is blown up proportionally to the arrival rate, whereas the transitions of the background process can be relatively slow ($f < 1$) or fast ($f > 1$). We anticipate that the blocking probability will decay essentially exponentially. Recent work on similar models [10] has shown that the case $f = 1$ is delicate and should be handled separately.
- The development of sound staffing rules for the Markov-modulated loss system. A similar issue could be studied for the model in which customers finding all servers busy join a waiting line.

Acknowledgements

The authors thank S. Asmussen and the anonymous referees for useful comments.

References

[1] D. Anderson, J. Blom, M. Mandjes, H. Thorsdottir, and K. de Turck (2016). A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodology and Computing in Applied Probability*, Vol. 18, pp. 153-168.

[2] K. Avrachenkov and M. Haviv (2004). The first Laurent series coefficients for singularly perturbed stochastic matrices. *Linear Algebra and its Applications*, Vol. 386, pp. 243-259.

[3] A. Bassamboo, S. Ramandeep, and A. Zeevi (2010). Capacity sizing under parameter uncertainty: safety staffing principles revisited. *Management Science*, Vol. 56, pp. 1668-1686.

[4] J. Blom, K. de Turck, and M. Mandjes (2015). Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probability in the Engineering and Informational Sciences*, Vol. 29, pp. 433-459.

[5] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir (2014). Markov-modulated infinite-server queues with general service times. *Queueing Systems*, Vol. 76, pp. 403-424.

[6] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, Vol. 100, pp. 36-50.

[7] B. D'Auria (2007). Stochastic decomposition of the M/G/∞ queue in a random environment. *Operations Research Letters*, Vol. 35, pp. 805-812.

[8] A. ELWALID AND D. MITRA (1993). Effective bandwidth of general Markovian traffic sources and admission control of high speed networks *IEEE/ACM Transactions on Networking*, Vol 1(3), pp. 329-343.

[9] A.K. ERLANG (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Engineer's Journal*, Vol. 10, pp. 189-197.

[10] G. HUANG, M. MANDJES, and P. SPREIJ (2016). Large deviations for Markov-modulated diffusion processes with rapid switching. *Stochastic Processes and their Applications*, Vol. 126, pp. 1785-1818.

[11] D. JAGERMAN (1974). Some properties of the Erlang loss function. *Bell System Technical Journal*, Vol. 53, pp. 525-551.

[12] F. KELLY (1991). Loss networks. *Annals of Applied Probability*, Vol. 1, pp. 319-378.

[13] S. KIM and W. WHITT (2014). Are call center and hospital arrivals well modeled by non-homogeneous Poisson processes? *Manufacturing & Service Operations Management*, Vol. 16, pp. 464-480.

[14] H. LU, G. PANG, and M. MANDJES (2015). A functional central limit theorem for Markov additive arrival processes and its applications to queueing systems. *Submitted.*

[15] M. MIYAZAWA (1993). Insensitivity and product form decomposability of reallocatable GSMP. *Advances in Applied Probability*, Vol. 25, pp. 415-437.

[16] M. NEUTS (1981). *Matrix Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore, United States.

[17] E. SENETA (1981). *Non-negative matrices and Markov chains*. Springer, New York, USA.

[18] G. THORSON (2000). Gaussian elimination on a banded matrix. *Stanford Exploration Project*, 12 pages. See: `http://sep.stanford.edu/data/media/public/oldreports/sep20/20_11.pdf`.

[19] W. WHITT (2002). *Stochastic-process Limits*. Springer, New York, USA.

[20] P. WHITTLE (1986). *Systems in Stochastic Equilibrium*. Wiley, Chichester, UK.

[21] D. WILLIAMS (1991). *Probability with Martingales.* Cambridge University Press, Cambridge, UK.

A Sketch of the proof of Proposition 3.1

First observe that

$$(\boldsymbol{P}^{(N)})'(s) = -\frac{1}{N^\gamma}\sum_{k=0}^{C_N} k\exp\left(-s\left(\frac{k-N\varrho}{N^\gamma}\right)\right)\boldsymbol{\pi}_k^{(N)} + N^{1-\gamma}\varrho\,\boldsymbol{P}^{(N)}(s),$$

or, alternatively,

$$\sum_{k=0}^{C_N} k\exp\left(-s\left(\frac{k-N\varrho}{N^\gamma}\right)\right)\boldsymbol{\pi}_k^{(N)} = N\varrho\,\boldsymbol{P}^{(N)}(s) - N^\gamma(\boldsymbol{P}^{(N)})'(s). \tag{25}$$

Multiplying the balance equations of the scaled model by $\exp(-sN^{-\gamma}(k-N\varrho))$ and summing over $k\in\{0,\ldots,C_N\}$ yields the identity

$$\begin{aligned}\sum_{k=0}^{C_N}\exp\left(-s\left(\frac{k-N\varrho}{N^\gamma}\right)\right)\boldsymbol{\pi}_k^{(N)}\Big[N\,\mathrm{diag}\{\boldsymbol{\lambda}\}1_{\{k<C_N\}} + k\cdot\mathrm{diag}\{\boldsymbol{\mu}\} - N^f Q\Big]&\\ = \sum_{k=0}^{C_N}\exp\left(-s\left(\frac{k-N\varrho}{N^\gamma}\right)\right)\boldsymbol{\pi}_{k-1}^{(N)}\,N\,\mathrm{diag}\{\boldsymbol{\lambda}\}1_{\{k>0\}}&\\ +\sum_{k=0}^{C_N}\exp\left(-s\left(\frac{k-N\varrho}{N^\gamma}\right)\right)\boldsymbol{\pi}_{k+1}^{(N)}\,\mathrm{diag}\{\boldsymbol{\mu}\}1_{\{k<C_N\}}.&\end{aligned} \tag{26}$$

Now let us study (26) in greater detail. Writing $\boldsymbol{\kappa}_N := \boldsymbol{\pi}^{(N)}_{C_N}$, the first term on the left hand side can be rewritten as

$$N\left(\boldsymbol{P}^{(N)}(s) - \boldsymbol{\kappa}_N e^{-s\beta}\right)\,\mathrm{diag}\{\boldsymbol{\lambda}\};$$

the first term on the right hand side is precisely equal to this expression, but multiplied by a factor $\exp(-sN^{-\gamma})$. Using (25), the second term on the left hand side reads

$$N\varrho\,\boldsymbol{P}^{(N)}(s)\,\mathrm{diag}\{\boldsymbol{\mu}\} - N^{\gamma}(\boldsymbol{P}^{(N)})'(s)\,\mathrm{diag}\{\boldsymbol{\mu}\};$$

and the second term on the right hand side is equal to this expression multiplied by $\exp(sN^{-\gamma})$. We thus arrive at the system of differential equations

$$\begin{aligned}\boldsymbol{P}^{(N)}(s)Q &= N^{1-f}\left(1-e^{-s/N^{\gamma}}\right)\left(\boldsymbol{P}^{(N)}(s) - \boldsymbol{\kappa}_N e^{-s\beta}\right)\,\mathrm{diag}\{\boldsymbol{\lambda}\}\\ &\quad + N^{1-f}\left(1-e^{s/N^{\gamma}}\right)\varrho\,\boldsymbol{P}^{(N)}(s)\,\mathrm{diag}\{\boldsymbol{\mu}\} - N^{\gamma-f}\left(1-e^{s/N^{\gamma}}\right)(\boldsymbol{P}^{(N)})'(s)\,\mathrm{diag}\{\boldsymbol{\mu}\}.\end{aligned}$$

The next steps are now exactly as in [4], and we therefore omit details here. We postmultiply the equation by the fundamental matrix $F := D + A$ (where $A := \boldsymbol{e}\boldsymbol{\alpha}$), use $QF = I - A$, bring $\boldsymbol{P}^{(N)}(s)A$ to the right hand side, to obtain an expression for $\boldsymbol{P}^{(N)}(s)$ in terms of itself and its derivative. Then we use the Taylor-series expansion

$$1 - e^{\pm s/N^{\gamma}} = \mp sN^{-\gamma} - \frac{1}{2}s^2N^{-2\gamma} + O(N^{-3\gamma}),$$

to iterate the resulting identity, which we then postmultiply with $\boldsymbol{e}N^{f}$. Mimicking the reasoning in [4], one obtains, with $\psi^{(N)}(s) := \boldsymbol{P}^{(N)}(s)\boldsymbol{e}$,

$$\begin{aligned}0 &= -sN^{1-\gamma}\psi^{(N)}(s)\cdot\boldsymbol{\alpha}(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\boldsymbol{e}\\ &\quad + s^2N^{2-f-2\gamma}\psi^{(N)}(s)\cdot\boldsymbol{\alpha}(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})F(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\boldsymbol{e}\\ &\quad + \frac{s^2}{2}N^{1-2\gamma}\psi^{(N)}(s)\cdot\boldsymbol{\alpha}(\mathrm{diag}\{\boldsymbol{\lambda}\} + \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\boldsymbol{e}\\ &\quad - s\cdot(\psi^{(N)})'(s)\cdot\boldsymbol{\alpha}\,\mathrm{diag}\{\boldsymbol{\mu}\}\boldsymbol{e} + se^{-s\beta}\,N^{1-\gamma}\boldsymbol{\kappa}_N\,\mathrm{diag}\{\boldsymbol{\lambda}\}\boldsymbol{e} + o(1),\end{aligned}$$

where the first term on the right hand side cancels (as an immediate consequence of the definition of ϱ). After dividing by $\mu_\infty\,s$, we obtain the one-dimensional ordinary differential equation,

$$(\psi^{(N)})'(s) = s\psi^{(N)}(s)\cdot\frac{1}{\mu_\infty}\left(N^{2-f-2\gamma}U + N^{1-2\gamma}\lambda_\infty\right) + e^{-s\beta}N^{1-\gamma}\frac{\boldsymbol{\kappa}_N\,\mathrm{diag}\{\boldsymbol{\lambda}\}\boldsymbol{e}}{\mu_\infty}, \tag{27}$$

where U, defined in (15), is given by

$$\begin{aligned}U &= \boldsymbol{\alpha}\,(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\,F\,(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\boldsymbol{e}\\ &= \boldsymbol{\alpha}\,(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\,D\,(\mathrm{diag}\{\boldsymbol{\lambda}\} - \varrho\,\mathrm{diag}\{\boldsymbol{\mu}\})\boldsymbol{e}.\end{aligned}$$

Now we consider the limiting solution $\psi(\cdot)$ when $N \to \infty$, first looking at the homogeneous version of the differential equation (27) which is precisely the same as the equation in [4]). By distinguishing between the cases $f < 1$, $f > 1$, and $f = 1$, the reasoning presented in [4] yields the solution $g(s) := c_1\exp(s^2\sigma^2/2)$, with σ^2 as defined above, for some free constant c_1. In [4] it could be concluded that $c_1 = 1$ due to the lack of the non-homogeneous term; evidently, we cannot do this here.

The next (standard) step is to try $g(s)h(s)$ as solution to the non-homogeneous limiting differential equation, with the function $h(\cdot)$ to be identified. To this end, we first define

$$\kappa := \lim_{N\to\infty} N^{1-\gamma}(\boldsymbol{\kappa}_N \operatorname{diag}\{\boldsymbol{\lambda}\}\boldsymbol{e})/\mu_\infty,$$

whose value we determine below. We obtain the equation

$$g'(s)h(s) + g(s)h'(s) = sg(s)h(s)\sigma^2 + \kappa e^{-s\beta},$$

such that $h'(s) = \kappa e^{-s\beta}/g(s)$, and hence

$$h(s) = \int_{-\infty}^{s} \frac{\kappa}{c_1}\, e^{-r^2\sigma^2/2 - r\beta}\,\mathrm{d}r + c_2.$$

We thus have found the solution

$$\psi(s) = c_1 e^{s^2\sigma^2/2} \cdot \left(\frac{\kappa}{c_1}\int_{-\infty}^{s} e^{-r^2\sigma^2/2 - r\beta}\,\mathrm{d}r + c_2\right).$$

The next goal is to identify the unknown constants. To this end, notice that

$$\int_{-\infty}^{s} e^{-r^2\sigma^2/2 - r\beta}\,\mathrm{d}r = \int_{-\infty}^{s\sigma} \frac{1}{\sigma}\exp\left(-\frac{1}{2}\left(u + \frac{\beta}{\sigma}\right)^2\right) e^{\beta^2/(2\sigma^2)}\mathrm{d}u,$$

which can be interpreted as $\mathbb{P}\left(\mathcal{N}(0,1) < \sigma s + \beta/\sigma\right)\sigma^{-1} e^{\beta^2/(2\sigma^2)}$. Using the standard asymptotic relation $\mathbb{P}(\mathcal{N}(0,1) > x)\, x e^{x^2/2}\sqrt{2\pi} \to 1$, this quantity behaves, for $s \to -\infty$, as

$$\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma s + \beta/\sigma} e^{-(\sigma s + \beta/\sigma)^2/2}\frac{1}{\sigma} e^{\beta^2/(2\sigma^2)} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma^2 s + \beta} e^{-\beta s}\, e^{-s^2\sigma^2/2}.$$

We conclude that

$$\lim_{s\to-\infty}\frac{1}{s}\log\left(e^{s^2\sigma^2/2}\cdot\frac{1}{c_1}\int_{-\infty}^{s} e^{-r^2\sigma^2/2 - r\beta}\,\mathrm{d}r\right) = \beta.$$

Because $\bar{M}^{(N)} \leqslant \beta$, we also know that, for $s < 0$, $\psi(s) \leqslant e^{-s\beta}$, and hence

$$-\lim_{s\to-\infty}\frac{1}{s}\log\psi(s) \leqslant \beta,$$

and therefore $c_2 = 0$. Now from $\psi(0) = 1$ it follows that

$$\kappa = \left(\int_{-\infty}^{0} e^{-r^2\sigma^2/2 - r\beta}\,\mathrm{d}r\right)^{-1} = \sigma\,\frac{\phi_{\mathcal{N}}(\beta/\sigma)}{\Phi_{\mathcal{N}}(\beta/\sigma)}.$$

where we recall that $\phi_{\mathcal{N}}(\cdot)$ and $\Phi_{\mathcal{N}}(\cdot)$ denote the density and cumulative distribution function of a standard Normal random variable, respectively. It takes elementary calculus to verify that the resulting expression for $\psi(s)$ coincides with (16).

B Output of numerical experiments

$[c_1, c_2, c_3, c_4]$	β	$p_5^{(s)}$	$p_{10}^{(s)}$	$E^{(s)}$	$\sigma^{(s)}$	$p_5^{(b)}$	$p_{10}^{(b)}$	$E^{(b)}$	$\sigma^{(b)}$
$[0, 0, 0, 0]$	0.5	7.75%	3.23%	1.95 (-02)	6.61 (-02)	8.34%	3.10%	2.74 (-02)	4.49 (-02)
$[0, 0, 0, 0]$	1.0	6.67%	2.66%	1.74 (-02)	7.38 (-02)	6.92%	3.04%	1.33 (-02)	4.28 (-02)
$[0, 0, 0, 0]$	1.5	5.95%	2.31%	1.60 (-02)	4.26 (-02)	6.31%	2.78%	-1.03 (-02)	6.35 (-02)
$[0, 0, 0, 0]$	2.0	4.84%	2.03%	1.46 (-02)	4.69 (-02)	83.16%	3.89%	-4.98 (-02)	9.44 (-02)
$[0, 2, 0, 0]$	0.5	0.54%	0.14%	7.59 (-03)	8.70 (-03)	0.51%	0.10%	1.71 (-02)	6.32 (-03)
$[0, 2, 0, 0]$	1.0	0.33%	0.08%	7.38 (-03)	9.77 (-03)	0.61%	0.23%	1.73 (-03)	1.01 (-02)
$[0, 2, 0, 0]$	1.5	0.49%	0.15%	7.65 (-03)	9.46 (-03)	0.26%	0.07%	-2.67 (-02)	9.15 (-03)
$[0, 2, 0, 0]$	2.0	0.29%	0.08%	6.57 (-03)	5.70 (-03)	97.69%	0.29%	-7.23 (-02)	2.01 (-02)
$[0, 0, 2, 0]$	0.5	14.39%	6.77%	3.15 (-02)	1.02 (-01)	12.73%	4.67%	3.24 (-02)	4.65 (-02)
$[0, 0, 2, 0]$	1.0	12.28%	5.82%	2.76 (-02)	8.40 (-02)	10.65%	4.20%	1.90 (-02)	4.80 (-02)
$[0, 0, 2, 0]$	1.5	11.28%	5.38%	2.58 (-02)	7.93 (-02)	9.51%	4.22%	-2.60 (-03)	5.78 (-02)
$[0, 0, 2, 0]$	2.0	10.18%	4.88%	2.51 (-02)	9.95 (-02)	71.63%	4.54%	-3.75 (-02)	7.85 (-02)
$[0, 0, 0, 2]$	0.5	15.55%	7.62%	3.33 (-02)	8.02 (-02)	11.70%	4.67%	3.17 (-02)	4.81 (-02)
$[0, 0, 0, 2]$	1.0	14.11%	6.15%	2.92 (-02)	7.18 (-02)	9.38%	4.00%	1.71 (-02)	4.99 (-02)
$[0, 0, 0, 2]$	1.5	11.73%	5.42%	2.73 (-02)	7.75 (-02)	8.25%	4.03%	-5.36 (-03)	7.01 (-02)
$[0, 0, 0, 2]$	2.0	10.85%	4.60%	2.51 (-02)	6.94 (-02)	76.54%	5.25%	-4.26 (-02)	9.36 (-02)
$[0, -2, 0, 0]$	0.5	21.90%	13.95%	7.31 (-02)	3.89 (-01)	28.24%	18.46%	8.40 (-02)	2.15 (-01)
$[0, -2, 0, 0]$	1.0	20.78%	13.27%	6.93 (-02)	3.74 (-01)	26.20%	18.44%	7.93 (-02)	2.36 (-01)
$[0, -2, 0, 0]$	1.5	18.80%	12.13%	6.29 (-02)	4.20 (-01)	23.98%	16.92%	5.89 (-02)	2.96 (-01)
$[0, -2, 0, 0]$	2.0	17.73%	11.64%	6.40 (-02)	2.84 (-01)	86.37%	23.07%	5.68 (-02)	5.83 (-01)
$[0, 0, -2, 0]$	0.5	14.21%	6.49%	3.05 (-02)	1.24 (-01)	12.43%	4.31%	3.19 (-02)	4.82 (-02)
$[0, 0, -2, 0]$	1.0	12.50%	5.66%	2.65 (-02)	7.28 (-02)	10.94%	4.08%	1.89 (-02)	4.55 (-02)
$[0, 0, -2, 0]$	1.5	11.18%	5.18%	2.58 (-02)	9.08 (-02)	9.67%	4.65%	-1.47 (-03)	6.18 (-02)
$[0, 0, -2, 0]$	2.0	10.37%	4.97%	2.41 (-02)	8.16 (-02)	72.37%	5.23%	-3.51 (-02)	9.33 (-02)
$[0, 0, 0, -2]$	0.5	16.73%	7.47%	3.44 (-02)	1.16 (-01)	12.01%	4.68%	3.23 (-02)	5.80 (-02)
$[0, 0, 0, -2]$	1.0	14.46%	6.61%	3.00 (-02)	7.67 (-02)	9.23%	3.85%	1.72 (-02)	5.71 (-02)
$[0, 0, 0, -2]$	1.5	12.36%	5.67%	2.80 (-02)	7.25 (-02)	8.18%	4.10%	-4.64 (-03)	8.30 (-02)
$[0, 0, 0, -2]$	2.0	10.58%	4.67%	2.58 (-02)	1.12 (-01)	77.73%	5.41%	-4.19 (-02)	8.85 (-02)

Table 1: Results for $N = 500$.

$[c_1, c_2, c_3, c_4]$	β	$p_5^{(\mathrm{s})}$	$p_{10}^{(\mathrm{s})}$	$E^{(\mathrm{s})}$	$\sigma^{(\mathrm{s})}$	$p_5^{(\mathrm{b})}$	$p_{10}^{(\mathrm{b})}$	$E^{(\mathrm{b})}$	$\sigma^{(\mathrm{b})}$
$[0, 0, 0, 0]$	0.5	1.34%	0.32%	5.51 (-03)	1.52 (-02)	1.01%	0.13%	6.15 (-03)	1.12 (-02)
$[0, 0, 0, 0]$	1.0	1.09%	0.24%	5.76 (-03)	1.47 (-02)	1.17%	0.36%	-1.20 (-03)	1.46 (-02)
$[0, 0, 0, 0]$	1.5	1.08%	0.18%	3.04 (-03)	1.04 (-02)	1.69%	0.45%	5.07 (-03)	1.68 (-02)
$[0, 0, 0, 0]$	2.0	0.80%	0.16%	3.41 (-03)	9.92 (-03)	2.02%	0.60%	-1.04 (-02)	2.17 (-02)
$[0, 2, 0, 0]$	0.5	0.06%	0.01%	3.04 (-03)	4.42 (-03)	0.06%	0.01%	3.00 (-03)	2.71 (-03)
$[0, 2, 0, 0]$	1.0	0.03%	0.01%	3.43 (-03)	2.29 (-03)	0.08%	0.00%	-5.59 (-03)	3.28 (-03)
$[0, 2, 0, 0]$	1.5	0.02%	0.00%	1.15 (-03)	1.81 (-03)	0.03%	0.01%	1.76 (-03)	2.74 (-03)
$[0, 2, 0, 0]$	2.0	0.03%	0.00%	1.59 (-03)	2.00 (-03)	0.15%	0.02%	-1.60 (-02)	5.68 (-03)
$[0, 0, 2, 0]$	0.5	3.61%	1.07%	9.35 (-03)	4.02 (-02)	1.88%	0.32%	8.18 (-03)	1.63 (-02)
$[0, 0, 2, 0]$	1.0	2.97%	0.81%	8.39 (-03)	2.13 (-02)	1.64%	0.48%	1.12 (-03)	1.61 (-02)
$[0, 0, 2, 0]$	1.5	2.71%	0.69%	6.39 (-03)	2.73 (-02)	2.14%	0.53%	6.79 (-03)	1.75 (-02)
$[0, 0, 2, 0]$	2.0	2.20%	0.65%	6.12 (-03)	1.94 (-02)	2.62%	0.59%	-7.15 (-03)	2.23 (-02)
$[0, 0, 0, 2]$	0.5	3.04%	0.68%	9.12 (-03)	1.90 (-02)	1.68%	0.40%	7.65 (-03)	1.46 (-02)
$[0, 0, 0, 2]$	1.0	2.80%	0.68%	8.93 (-03)	1.79 (-02)	1.76%	0.51%	8.10 (-04)	1.84 (-02)
$[0, 0, 0, 2]$	1.5	2.55%	0.65%	6.68 (-03)	1.89 (-02)	2.28%	0.61%	5.74 (-03)	1.78 (-02)
$[0, 0, 0, 2]$	2.0	2.06%	0.58%	6.62 (-03)	2.20 (-02)	2.60%	0.87%	-8.57 (-03)	2.77 (-02)
$[0, -2, 0, 0]$	0.5	11.68%	5.32%	1.61 (-02)	8.93 (-02)	11.24%	4.96%	2.23 (-02)	5.79 (-02)
$[0, -2, 0, 0]$	1.0	11.10%	5.51%	1.62 (-02)	9.98 (-02)	11.92%	6.00%	1.95 (-02)	7.95 (-02)
$[0, -2, 0, 0]$	1.5	9.96%	4.34%	1.30 (-02)	8.28 (-02)	13.11%	6.82%	2.68 (-02)	1.00 (-01)
$[0, -2, 0, 0]$	2.0	9.21%	4.15%	1.07 (-02)	6.42 (-02)	16.68%	8.81%	1.90 (-02)	1.67 (-01)
$[0, 0, -2, 0]$	0.5	3.29%	1.05%	9.00 (-03)	2.58 (-02)	1.69%	0.28%	7.96 (-03)	1.36 (-02)
$[0, 0, -2, 0]$	1.0	3.32%	0.81%	9.07 (-03)	2.59 (-02)	1.58%	0.34%	1.25 (-03)	1.59 (-02)
$[0, 0, -2, 0]$	1.5	2.51%	0.88%	6.57 (-03)	2.28 (-02)	2.11%	0.46%	6.67 (-03)	2.13 (-02)
$[0, 0, -2, 0]$	2.0	2.21%	0.55%	5.83 (-03)	1.61 (-02)	2.52%	0.73%	-6.99 (-03)	2.28 (-02)
$[0, 0, 0, -2]$	0.5	3.40%	0.83%	9.68 (-03)	2.34 (-02)	1.89%	0.42%	7.99 (-03)	2.00 (-02)
$[0, 0, 0, -2]$	1.0	2.72%	0.52%	8.89 (-03)	1.99 (-02)	1.69%	0.46%	4.94 (-04)	1.73 (-02)
$[0, 0, 0, -2]$	1.5	2.23%	0.41%	6.30 (-03)	1.95 (-02)	2.19%	0.71%	5.61 (-03)	1.86 (-02)
$[0, 0, 0, -2]$	2.0	2.17%	0.48%	6.56 (-03)	2.32 (-02)	2.62%	0.86%	-8.74 (-03)	2.58 (-02)

Table 2: Results for $N = 5\,000$.