Two Extensions of Kingman's GI/G/1 Bound

Florin Ciucu
University of Warwick

Felix Poloczek

ABSTRACT

A simple bound in GI/G/1 queues was obtained by Kingman using a discrete martingale transform [30]. We extend this technique to 1) multiclass ΣGI/G/1 queues and 2) Markov Additive Processes (MAPs) whose background processes can be time-inhomogeneous or have an uncountable state-space. Both extensions are facilitated by a necessary and sufficient ordinary differential equation (ODE) condition for MAPs to admit continuous martingale transforms. Simulations show that the bounds on waiting time distributions are almost exact in heavy-traffic, including the cases of 1) heterogeneous input, e.g., mixing Weibull and Erlang-k classes and 2) Generalized Markovian Arrival Processes, a new class extending the Batch Markovian Arrival Processes to continuous batch sizes.

KEYWORDS

Queueing; Markov and Non-Renewal Processes; Stochastic Bounds

ACM Reference Format:
Florin Ciucu and Felix Poloczek. 2018. Two Extensions of Kingman's GI/G/1 Bound. In *Proceedings of ACM conference*. ACM, New York, NY, USA, Article 4, 18 pages.

1 INTRODUCTION

A milestone in queueing theory was relaxing the often implicit assumption that interarrival times in GI/G/1 queues are statistically independent. One such extension, applicable in manufacturing and production systems, is the multiclass ΣGI/G/1 queue in which multiple classes of jobs, each with its own arrival (renewal) process, are merged. Due to the general lack of closure of renewal processes, let alone the general lack of stationarity of the merged process, the analysis of the ΣGI/G/1 queue is challenging. Several studies in heavy-traffic regimes addressed functional central limits (e.g., of the waiting times) [27], approximations (e.g., of the workload) with a one-dimensional reflecting Brownian motion [17], or Laplace transforms (e.g., of the waiting times) [6].

Another extension also emerging in the 1970s was driven by the non-renewal traffic characteristics in packet switches [2, 32]. Two widely studied models accounting for 'bursty' traffic are Markov Modulated Fluid (MMF) and Markov Modulated Poisson Process (MMPP). The former was proposed in the seminal paper [2] by representing traffic as (continuous) 'fluid' evolving at some constant rate, depending on a modulating Markov process; queues with MMF input can be exactly analyzed using ODEs and matrix analysis; related methods include spectral decomposition [1] or Wiener-Hopf factorization [45]. MMPP is a more accurate 'packetized' version of MMF, i.e., traffic evolves as a Poisson process with state dependent rates according to a modulating Markov process; the typical queueing analysis rests on matrix analytical techniques [25] or spectral decompositions [1, 20]. A common challenge of analyzing MMF and MMPP is the underlying numerical complexity, which can become prohibitive when a large number of sources are multiplexed [48]. For related discussions and more comprehensive reference lists see [33] and [23].

A popular method to analyze queues with MMF and MMPP input is effective bandwidth [19]. Advantages include the availability of exact (asymptotic) results, negligible computational cost when multiplexing many sources, and simplicity in the sense that many arrival processes can be analyzed in a unified manner. However, this method can yield inaccurate (non-asymptotic) results unless the input is Poisson [13, 48]. A related technique with similar features is the probabilistic network calculus [34].

In this paper we develop a unified analysis of queues with two broad classes of non-renewal arrivals: 1) the multiclass ΣGI/G/1 queue and 2) queues with Markov Additive Processes (MAPs). Our framework provides (non-asymptotic) stochastic bounds (e.g., on waiting time distributions) by extending an approach of Kingman [30] who obtained such bounds in GI/G/1 queues by first constructing martingale transforms and then using martingale properties. While this approach has often been used [4, 18, 40, 41, 46], our novelty is a link between MAP martingales and a necessary and sufficient ODE condition. This applies to general MAPs, whereby the background process can be inhomogeneous or have an uncountable state-space; moreover, the martingales are constructed in *continuous-time*. These three features altogether are instrumental to the analysis of the ΣGI/G/1 model.

Besides generality, the proposed method can be applied in a rather straightforward manner. The ODE condition is elementary, and in particular it immediately lends itself to a MMF martingale which was obtained in [21] using an involved argument. We investigate several other scenarios, e.g., ΣWeibull/G/1, ΣErlang-k/G/1, ΣWeibull + ΣErlang-k/G/1 (a mix of Weibull and Erlang-k classes), and queues with MMF, MMPP, Markovian Arrival Processes (MArPs),[1] and Generalized Markovian Arrival Processes (GMArP)[2]. Remarkably, the method retains the key advantage of effective bandwidth, i.e., a straightforward analysis with negligible numerical complexity in multiplexing scenarios. Additionally, the bounds are shown through simulations to be almost exact in heavy-traffic. The method can be easily extended to account for non-stationary services and scheduling.

The highlights of this paper are:

ACM conference, 2018

ACM ISBN 123-4567-24-567/08/06.

[1] We adopt the acronyms MAP and MArP for Markov Additive and Arrival, respectively, Processes; see [4], p. 302.
[2] GMArP is our own generalization of Batch Markovian Arrival Processes (BMArPs), whereby batch sizes can be real numbers.

- A key result enabling *continuous* martingale constructions from general MAPs by solving ODEs (Lemmas 5 and 6).
- Providing (almost) explicit and closed-form bounds on waiting time distributions in multiclass ΣGI/G/1 queues, including heterogeneous scenarios (Examples 1-3 in § 4).
- Several simulations illustrating almost exact bounds in heavy-traffic.
- Linear time computational complexity in analyzing queues with a superposition of GMArPs (§ 5.3). Effective bandwidth achieves the same complexity but with very poor numerical accuracy, whereas exact results are typically subject to an exponential complexity.
- The overall method extends to random and possibly non-stationary service, using roughly the same underlying results.

An important auxiliary result for future studies is

- Isolating *a single* source for numerical inaccuracies in Kingman's technique (Lemma 2).

In the rest of the paper we first summarize Kingman's technique and give new insight into the bounds' (in)accuracy. In § 3 we provide the main technical result of the paper. Several applications to multiclass ΣGI/G/1 and Markov Additive Processes (MAPs) queues are considered in § 4 and § 5. In § 6 we provide a more comprehensive discussion on related work, and also comment on possible extensions of the proposed technique. We conclude the paper in § 7. Appendices § A and § B provide detailed proofs and additional numerical results.

2 KINGMAN'S BOUND IN SPACE AND TIME DOMAIN QUEUEING MODELS

In this section we summarize Kingman's [30] martingale-based technique in two queueing models:

- Queueing models in the space domain, i.e., GI/G/1 queues (the model originally solved in [30]) and discuss their extension to multiclass ΣGI/G/1 queues (whose input is not GI due to the lack of closure of renewal processes under multiplexing, unless Poisson);
- Queueing models in the time domain, i.e., queues with general Markov Additive Processes (MAPs) comprising many arrival models subject to correlation such as Markov Fluids (MFs), Markov Modulated Poisson Processes (MMPPs), or Markovian Arrival Processes (MArPs).

The purpose of this summary is to illustrate the key ideas and similarities in the two models, relative to Kingman's technique, and to thus justify the development of a "unified" analysis.

2.1 Space Domain

The classical queueing model consists of two sequences of identically distributed interarrival times $(T_i)_{i\in\mathbb{N}}$ (when do jobs arrive at some queueing server/station?) and service times $(S_i)_{i\in\mathbb{N}}$ (how long does each job take to being served?). A typical assumption is that $(T_i)_i$ and $(S_i)_i$ are mutually independent. This is the $GI/G/1$ queue.

2.1.1 Kingman's Bound. While an exact and computationally tractable analysis of queues with general distributions is hard, an approximate solution (in terms of stochastic bounds) can be quickly given. Focusing on the waiting time W_n (how long does the n^{th} job wait in the queue prior to being served?), its distribution converges to that of

$$W := \sup_{n\geq 0}\{U_1 + U_2 + \cdots + U_n\} \ , \tag{1}$$

where $U_n := S_n - T_n$ for $n \geq 1$ and subject to the stability condition $\mathbb{E}[U_n] < 0$ (by convention, when $n = 0$, the corresponding element in the 'sup' is 0) (see, e.g., Proposition 2.1 in [44]).

The key idea to approximate W's distribution is a duality between stationary distributions and first passage probabilities for random walks, i.e.,

$$\mathbb{P}(W \geq \sigma) = \mathbb{P}(T < \infty) \ , \tag{2}$$

where $T := \inf\{n : U_1 + \cdots + U_n \geq \sigma\}$ is the first passage time (also a stopping time)[3]. Let the exponential martingale

$$X_n := e^{\theta(U_1+U_2+\cdots+U_n)} \ ,$$

where $\theta > 0$ satisfies $\mathbb{E}\left[e^{\theta U_n}\right] = 1$ (its existence is guaranteed by stability). Then, according to the optional sampling theorem for some finite n

$$\begin{aligned} 1 = \mathbb{E}[X_0] =& \mathbb{E}[X_{T\wedge n}] = \mathbb{E}[X_{T\wedge n}1_{T\leq n}] + \mathbb{E}[X_{T\wedge n}1_{T>n}] \\ \geq& \mathbb{E}[X_{T\wedge n}1_{T\leq n}] = \mathbb{E}[X_T 1_{T\leq n}] \\ =& \mathbb{E}\left[e^{\theta(U_1+U_2+\cdots+U_T)}1_{T\leq n}\right] \\ \geq& e^{\theta\sigma}\mathbb{E}[1_{T\leq n}] = e^{\theta\sigma}\mathbb{P}(T \leq n) \ . \end{aligned} \tag{3}$$

The need for the parameter n stems from a technicality of the optional sampling theorem. By taking $n \to \infty$ the final result is

Theorem 1. (Kingman's Bound) *In the model above*

$$\mathbb{P}(W \geq \sigma) \leq e^{-\theta\sigma} \ . \tag{4}$$

The result is quite general in terms of the distributions of T_i and S_i; service times must however have a moment generating function, otherwise, θ could not be constructed as above. Note also that the result is (almost) explicit, except for the construction of θ which generally requires a numerical procedure.

2.1.2 On the Bound's Accuracy. There are two inequalities in the derivations of Kingman's bound from (3). We next show that the first one holds in the limit as an equality:

Lemma 2. *In the model above*

$$\lim_{n\to\infty}\mathbb{E}[X_{T\wedge n}1_{T>n}] = 0 \ .$$

Proof. Construct the stopped martingale

$$Y_n := X_{T\wedge n}$$

which satisfies $X_n 1_{T>n} = Y_n 1_{T>n}$. We show next that Y_n is uniformly integrable.

Fixing $\varepsilon > 0$ and $n \geq 0$ we need to find $K < \infty$, independent of n, such that

$$E\left[Y_n 1_{Y_n>K}\right] < \varepsilon \ .$$

[3]The same idea was also used in risk analysis, whereby the right-hand side in (2) has the interpretation of 'ruin probability' [3].

Let us rewrite

$$\begin{aligned} &E\left[Y_n 1_{Y_n>K}\right] \\ &\quad = E\left[X_{T\wedge n} 1_{T>n} 1_{Y_n>K}\right] + E\left[X_{T\wedge n} 1_{T\le n} 1_{Y_n>K}\right] \\ &\quad = E\left[X_n 1_{T>n} 1_{X_n>K}\right] + E\left[X_T 1_{T\le n} 1_{X_T>K}\right] \ . \end{aligned} \tag{5}$$

From the definition of T, the first term in the sum is 0 when $K > e^{\theta\sigma}$. Rewrite the second term as $E\left[X_T 1_{T\le n} 1_{X_T 1_{T\le n}>K}\right]$. From the second line of (3), with $n \to \infty$, we obtain that $X_T 1_{T<\infty}$ is integrable, and therefore (see, e.g., [50], p. 127) there exists a $K < \infty$ such that

$$E\left[X_T 1_{T<\infty} 1_{X_T 1_{T<\infty}>K}\right] < \varepsilon \ .$$

Since $X_T 1_{T\le n} 1_{X_T 1_{T\le n}>K} \le X_T 1_{T<\infty} 1_{X_T 1_{T<\infty}>K}$ it then follows that the second term in (5) can be made arbitrarily small. Hence, Y_n is uniformly integrable.

According to the martingale convergence theorem (see, e.g., [50], p. 134), $Y := \lim_n Y_n$ exists *a.s.* (and also in $\mathcal{L}^1$).

We finally obtain that

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n} 1_{T>n}\right] &= \lim_{n\to\infty} \mathbb{E}\left[X_n 1_{T>n}\right] = \mathbb{E}\left[\lim_n X_n 1_{T>n}\right] \\ &= \mathbb{E}\left[\lim_n Y_n 1_{T>n}\right] = \mathbb{E}\left[\lim_n Y_n \lim_n 1_{T>n}\right] \\ &= \mathbb{E}\left[\lim_n Y_n 1_{T=\infty}\right] = \mathbb{E}\left[\lim_n X_n 1_{T=\infty}\right] \le \mathbb{E}\left[\lim_n X_n\right] = 0 \ . \end{aligned}$$

In the first line we could exchange the limit with the expectation from the bounded convergence theorem (the definition of T implies that $X_{T\wedge n} 1_{T>n} \le e^{\theta\sigma}$). In the second line we could split the limit of a product in the product of limits due to the *a.s.* convergence of Y_n. In the last line we used the fact that $U_1 + U_2 + \cdots + U_n$ is a divergent random walk with negative drift. □

The previous result indicates that the accuracy of Kingman's bound reduces to that of the straightforward bound

$$\mathbb{E}\left[e^{\theta(U_1+U_2+\cdots+U_T)} 1_{T\le n}\right] \ge e^{\theta\sigma}\mathbb{P}(T \le n)$$

from the last inequality in (3). A refinement was provided by Ross [46], i.e.,

$$\begin{aligned} \sup_{y\ge 0} K(y) e^{\theta\sigma}\mathbb{P}(T \le n) &\ge \mathbb{E}\left[e^{\theta(U_1+U_2+\cdots+U_T)} 1_{T\le n}\right] \\ &\ge \inf_{y\ge 0} K(y) e^{\theta\sigma}\mathbb{P}(T \le n) \ , \end{aligned} \tag{6}$$

where

$$K(y) = \mathbb{E}\left[e^{\theta(U_1-y)} \mid U_1 \ge y\right] \ .$$

These bounds immediately lend themselves to bounds on the waiting time distribution:

Lemma 3. (Ross' Bounds) *In the model above*

$$\frac{1}{\sup_{y\ge 0} K(y)} e^{-\theta\sigma} \le \mathbb{P}(W \ge \sigma) \le \frac{1}{\inf_{y\ge 0} K(y)} e^{-\theta\sigma} \ . \tag{7}$$

Remarkably, these bounds are *exact* for the GI/M/1 queue (see [46]). As a side remark, the proof for the lower bound in (6) uses an ingenious argument involving an additional stopping time. Using Lemma 2, however, the lower bound can be derived exactly as the upper bound, except for replacing the 'inf' with 'sup'.

We give an alternative proof of Lemma 3 in Appendix § A which can be immediately extended to generalize Ross bound from (6) to the case when $(U_n)_n$ is a homogeneous Markov chain.

2.1.3 Open Question: ΣGI/G/1. Consider the multiclass ΣGI/G/1 queue, whereby the arrivals are driven by multiple renewal sequences $(T_i^k)_i$ with $k = 1, 2, \ldots$ Unless the individual sequences are exponentially distributed, the aggregate interarrival process (essentially the spacings of order statistics) is not a renewal process. Consequently, the corresponding process X_n is no longer a martingale and the above method fails. An additional complication is that, in general, the aggregate interarrival process is not even stationary, and hence the existence of a steady-state for W_n is not guaranteed by Loynes' condition for G/G/1 queues (which requires the stationarity of the sequence $(T_i, S_i)_i$ and $E[S_i] < E[T_i]$).

Obtaining queueing bounds in multiclass ΣGI/G/1 queues, alike (4), is open. The related literature include exact results in terms of Laplace transforms (see Theorem 4 in [6]) and approximations on the expected waiting time W in heavy-traffic (see Proposition 1 in [6]). Our contribution is the derivation of closed-form stochastic bounds on the *distribution* of W, alike in the GI/G/1 case.

2.2 Time Domain

The other common queueing model consists of a compound arrival process $A(t)$ (how many jobs arrived by time t?) and a server processing the arrivals at some rate (either constant or random). The index t represents 'time', whereas the index n in the previous model represents 'space' (i.e., job number).

Assume a continuous-time model, a constant rate $C > 0$ for the server, and a stability condition $\limsup_t \frac{A(t)}{t} < C$. Focusing on the backlog process $Q(t)$ (how many jobs are in the queue at time t), under certain stationarity and ergodicity conditions, a limiting distribution of $Q(t)$ exists, and that is equal to that of

$$Q := \sup_{t\ge 0} \{A(t) - Ct\} \ . \tag{8}$$

(we assume that $A(t)$ is a reversible process to simplify notation).

To compute stochastic bounds on the distribution of Q, Kingman's technique can be extended from the space to the time domain. One has to first construct an appropriate martingale, e.g.,

$$X_t := e^{\theta(A(t)-Ct)} \ ,$$

in the case when $A(t)$ has independent increments, under an appropriate condition on θ. Following the same steps as before, the same elegant approximation can be obtained

$$\mathbb{P}(Q \ge \sigma) \le e^{-\theta\sigma} \ .$$

(For a complete proof in the general case with not necessarily independent increments see Theorem 7.)

An important observation about the technique is that it does not require the existence of a steady-state (non-ergodic Markovian arrival processes can be addressed). The explanation is that the produced backlog bounds are transient, i.e., they hold for $\mathbb{P}(Q(t) \ge \sigma)$ for any time t; the same observation holds in the space domain.

An advantage of the time domain model is its suitability to encode the correlation structure in the arrivals (e.g., driven by some Markov process). Moreover, analyzing queues with multiplexed arrivals $A_i(t)$ is very convenient. Indeed, by assuming the statistical independence of $A_i(t)$ and a constant rate server, one can let $A(t) := \sum_i A_i(t)$ in the representation of Q from (8) and apply the same steps as above to obtain a bound on Q's distribution.

Based on this last observation, we will analyze the multiclass ΣGI/G/1 queue by framing the model in the time domain where multiplexing is seemingly 'easy' (see § 4). What is noteworthy is that the martingale construction in the transformed domain is driven by the same general/unified result which provides conditions for the martingale construction from pure time-domain based arrivals.

3 A MARTINGALE TRANSFORM VIA ODE

Here we present the main result of this paper, i.e., a necessary and sufficient condition for *Markov Additive Processes* (MAPs) to admit martingale representations. In a continuous-time model, we adopt a simplified definition of a MAP by Pacheco and Prabhu [39] (for a more general version see [14]):

Definition 4. *A bivariate process* $(A(t), M_t)_t$ *is a* Markov Additive Process *if and only if*

(1) *the pair* $(A(t), M_t)$ *is a Markov process in* $\mathbb{R}^2$,
(2) $A(0) = 0$ *and* $A(t)$ *is nondecreasing,*
(3) *the (joint and conditional) distribution of*

$$(A(s,t), M_t \mid A(s), M_s)$$

depends only on M_s.

M_t is a *background process* and $A(t)$ is an additive processes counting arrivals up to time t; we write $A(s,t) := A(t) - A(s)$. Note that M_t is a Markov process and $A(t)$ has conditionally independent increments (conditioning on the states of M_t).

Next we give the main result, first in the (time) homogenous case, i.e., the law $\mathbb{P}(A(s+\tau, t+\tau) \le x, M_{t+\tau} = y \mid M_{s+\tau} = z)$ is invariant under the time shift τ. First, denote by 'Im' the *image* of a function, e.g., $\mathrm{Im}(M_t)$ is the set of states of M_t.

Lemma 5. (Time-Homogeneous Case) *Consider a time-homogenous Markov Additive Process* $(A(t), M_t)$*, a random function* $h : \mathrm{Im}(M) \to \mathbb{R}^+$*, the parameters* $y \in \mathrm{Im}(M)$*,* $C, \theta > 0$*, and define for* $s \ge 0$

$$\varphi_y(s) := \mathbb{E}\left[h(M_s) e^{\theta(A(s)-Cs)} \;\middle|\; M_0 = y \right] .$$

Then $\frac{d}{ds}\varphi_y(s)\Big|_{s=0} = 0$ *for all* $y \in \mathrm{Im}(M)$ *if and only if the process*

$$h(M_t) e^{\theta(A(t)-Ct)} \tag{9}$$

is a martingale relative to the natural filtration.

An explicit exponential martingale for MAPs is given in Asmussen [4] (see Proposition 2.4, p. 312) by solving for an eigenvalue/vector problem. In connection to this result, Lemma 5 is more general in that the state-space of M_t can be uncountable (e.g., $\mathbb{R}$); moreover, the lemma can be immediately extended to the time-inhomogeneous case (see Lemma 6). These two features are instrumental for the later applications. An additional advantage of Lemma 5 is that the necessity of the differentiability condition ensures the uniqueness of exponential martingales of the form from Eq. (9) for several MAP examples treated in § 5.

We remark that the sufficiency of the differentiability condition is trivial. Indeed, let a time-continuous martingale X_t and $\varphi_{X_0}(s) := \mathbb{E}[X_s \mid X_0]$. Then $\frac{d}{ds}\varphi_{X_0}(s) = 0$ because $\varphi_{X_0}(s) = X_0$, i.e., a constant, by definition. The key result in Lemma 5 is thus the necessary condition, which critically relies on the underlying Markov structure.

Proof. Let $(\mathcal{F}_t)_t$ be the natural filtration generated by $(A(t), M_t)$. Note first that, by homogeneity, for any $t \ge 0$:

$$\mathbb{E}\left[h(M_{t+s}) e^{\theta(A(t,t+s)-Cs)} \;\middle|\; M_t = y \right] = \varphi_y(s) .$$

The martingale property is equivalent to

$$\mathbb{E}\left[h(M_{t+s}) e^{\theta(A(t,t+s)-Cs)} \;\middle|\; \mathcal{F}_t \right] = h(M_t) ,$$

for any $s, t \ge 0$. However, it suffices to show that for any $s \ge 0$

$$\varphi_{M_0}(s) = \mathbb{E}\left[h(M_s) e^{\theta(A(s)-Cs)} \;\middle|\; M_0 \right] = h(M_0) ,$$

due to the time-homogeneity and the Markov property. By assumption, the derivative of $\varphi_{M_0}(s)$ vanishes at $s = 0$. Next, we show that the derivative also vanishes for arbitrary $s > 0$, i.e., $\frac{d}{ds}\varphi_{M_0}(s) \equiv 0$:

$$\begin{aligned}
\frac{d}{ds}\varphi_{M_0}(s) &= \lim_{\Delta s \to 0} \frac{1}{\Delta s} \mathbb{E}\Big[h(M_{s+\Delta s}) e^{\theta(A(s+\Delta s)-C(s+\Delta s))} \\
&\qquad - h(M_s) e^{\theta(A(s)-Cs)} \Big| M_0 \Big] \\
&= \lim_{\Delta s \to 0} \frac{1}{\Delta s} \mathbb{E}\Big[\mathbb{E}\Big[h(M_{s+\Delta s}) e^{\theta(A(s+\Delta s)-C(s+\Delta s))} \\
&\qquad - h(M_s) e^{\theta(A(s)-Cs)} \Big| \mathcal{F}_s \Big] \Big| M_0 \Big] \\
&= \lim_{\Delta s \to 0} \frac{1}{\Delta s} \mathbb{E}\Big[e^{\theta(A(s)-Cs)} \mathbb{E}\Big[h(M_{s+\Delta s}) e^{\theta(A(s,s+\Delta s)-C\Delta s)} \\
&\qquad - h(M_s) \Big| \mathcal{F}_s \Big] \Big| M_0 \Big] \\
&= \lim_{\Delta s \to 0} \frac{1}{\Delta s} \mathbb{E}\Big[e^{\theta(A(s)-Cs)} \mathbb{E}\Big[h(M_{s+\Delta s}) e^{\theta(A(s,s+\Delta s)-C\Delta s)} \\
&\qquad - h(M_s) \Big| M_s \Big] \Big| M_0 \Big] \\
&= \lim_{\Delta s \to 0} \mathbb{E}\left[e^{\theta(A(s)-Cs)} \frac{1}{\Delta s} \left(\varphi_{M_s}(\Delta s) - \varphi_{M_s}(0) \right) \;\middle|\; M_0 \right] \\
&= \mathbb{E}\left[e^{\theta(A(s)-Cs)} \lim_{\Delta s \to 0} \frac{1}{\Delta s} \left(\varphi_{M_s}(\Delta s) - \varphi_{M_s}(0) \right) \;\middle|\; M_0 \right] \\
&= \mathbb{E}\left[e^{\theta(A(s)-Cs)} \frac{d}{ds}\varphi_{M_s}(0) \;\middle|\; M_0 \right] = 0 .
\end{aligned}$$

In the sixth equation we applied the dominated convergence theorem, along with the definition of differentiability (the function $\frac{1}{\Delta s}\left(\varphi_{M_s}(\Delta s) - \varphi_{M_s}(0)\right)$ is bounded within a vicinity of 0), to interchange the limit and the expectation. The proof completes by the observation:

$$\varphi_{M_0}(s) = \varphi_{M_0}(0) + \int_0^s \frac{d}{du}\varphi_{M_0}(u) du = h(M_0) + 0 .$$

□

Next we present the extension to the time-inhomogeneous case.

Lemma 6. (Time-Inhomogeneous Case) *Under the same conditions from Lemma 5, except for allowing the MAP to be inhomogeneous, define*

$$\varphi_{t,y}(s) := \mathbb{E}\left[h(M_{t+s}) e^{\theta(A(t,t+s)-Cs)} \;\middle|\; M_t = y \right] .$$

Then $\frac{d}{ds}\varphi_{t,y}(s)\Big|_{s=0} = 0$ *for all* $y \in \mathrm{Im}(M)$ *and* $t \ge 0$ *if and only if the process*

$$h(M_t) e^{\theta(A(t)-Ct)}$$

is a martingale.

We note that Lemmas 5 and 6, as well as their proofs, are almost identical, with the difference of specifically accounting for the starting time t in the latter.

In the analysis of the ΣGI/G/1 queue we shall consider M_t as the remaining lifetime of a renewal process, in which case the associated MAP is inhomogeneous; in all other examples from § 5 we shall consider homogeneous MAPs.

3.1 Queueing Metrics

Recalling our goal of developing a unified framework for multiclass ΣGI/G/1 and MAPs queues, we present such a unified result next.

Theorem 7. *Consider an arrival process $A(t)$ being served at rate C, and suppose that there exists the martingale process*

$$X_t := h(M_t)e^{\theta(A(t)-Ct)}$$

for some parameter $\theta > 0$, random process M_t, and non-negative function $h()$. Then the stationary backlog process Q satisfies

$$\mathbb{P}(Q \geq \sigma) \leq \frac{\mathbb{E}[h(M_0)]}{\inf_{m\in\mathrm{Im}(M)} h(m)} e^{-\theta\sigma} .$$

Moreover, if the sizes of the arrivals' data units are bounded by ξ, then the following lower bound holds:

$$\mathbb{P}(Q \geq \sigma) \geq \frac{\mathbb{E}[h(M_0)]}{\sup_{m\in\mathrm{Im}(M)} h(m)} e^{-\theta(\sigma+\xi)} .$$

We denoted with abuse of notation

$$\mathrm{Im}(M) = \{m \mid \exists t : M_t = m \wedge a(t) \geq C\} ,$$

where $a(t)$ is the instantaneous arrival process of $A(t)$, i.e., $A(t) = \int_0^t a(s)ds$. The clause '$a(t) \geq C$' becomes clear in the proof and it can tighten the bounds significantly. We note that waiting time bounds are similar.

The parameter θ is exactly the asymptotic decay rate of the backlog process from the large-deviation limit $\sigma^{-1}\log\mathbb{P}(Q \geq \sigma) \to -\theta$, as $\sigma \to \infty$, which is at the basis of the effective bandwidth approximation $\mathbb{P}(Q \geq \sigma) \approx e^{-\theta\sigma}$ [13]; note the exact match between the decay rates in the upper and lower bounds from the theorem. Compared to this approximation, the crucial difference in the upper bound is the prefactor in front of the exponential. For some multiplexed arrivals the prefactor is exponential in the number of multiplexed sources (see, e.g., (13)), as conjectured in [13], which can make a substantial numerical difference to the effective approximation (see [13, 15] for numerical results).

The random process M_t depends on the structure of $A(t)$; in the case of the GI/G/1 queue, M_t is the remaining lifetime of the arrivals' renewal process (see § 4); in the case of MAP, M_t is the background process itself (see § 5). The random function $h()$ captures the correlation structure of the arrivals. In the case of renewal processes, $h()$ is a constant for discrete-time martingales (see the Kingman's martingale from § 2.1); a more general form holds for continuous-time martingales (see the construction from Corollary 8) to capture the construction in continuous time. In the MAP case, h() is constant for processes with independent increments, and non-constant otherwise; see the constructions from § 5.

The proof for the upper bound (see Appendix § A) is a straightforward adaptation of the proof of Kingman's bound from (3) to the given martingale; similar results, and proofs, are available in the literature (e.g., [9, 15, 40]). The proof for the lower bound is an immediate extension of the proof for the upper bound by leveraging Lemma 2; an alternative yet more compounded proof follows by defining an additional stopping time as in [46] (this ingenious idea was employed in [9], p. 342, and [16]). For a follow-up discussion see the Related-Work section § 6.1.

3.2 Multiplexing

An important benefit of the martingale characterization from Lemma 5 is that analyzing queues with multiplexed MAPs is convenient. Let two independent MAPs $(A_1(t), M_{1,t})$ and $(A_2(t), M_{2,t})$ being served at rate C. One needs a split $C_1 + C_2 = C$ to construct the martingales $h_1(M_{1,t})e^{\theta(A_1(t)-C_1t)}$ and $h_2(M_{2,t})e^{\theta(A_2(t)-C_2t)}$, respectively, subject to the conditions from Lemma 5, and with the *same* 'θ'. Then the closure property of independent martingales under multiplication yields the martingale

$$h_1(M_{1,t})h_2(M_{2,t})e^{\theta(A_1(t)+A_2(t)-t(C_1+C_2))} .$$

In this way the result from Theorem 7 applies directly. We shall provide several examples in § 4 and § 5.

We also note that the alternative approach of constructing an aggregate MAP from $(A_1(t), M_{1,t})$ and $(A_2(t), M_{2,t})$ can be computationally very expensive (e.g., exponential explosion in the number of states) due to Kronecker sums (see [39] and § 5.3.1 for a concrete example); moreover, constructing martingales with different θ's and then normalizing (e.g., using Jensen's inequality as in [41]) can lend itself to numerical accuracy issues.

4 APPLICATION 1: THE ΣGI/G/1 QUEUE

We start with a single (stable) GI/G/1 queue. To focus on the stationary waiting time distribution, it is convenient to represent the interarrivals as $(T_i)_{i\in\mathbb{Z}^*}$ such that $T_i \geq 0$ and

$$\cdots < -T_{-1} - T_0 < -T_0 \leq 0 < -T_0 + T_1 < -T_0 + T_1 + T_2$$

(note that T_0 is used for centering). Let $\mathbb{P}^0(\cdot) = \mathbb{P}(\cdot \mid T_0 = 0)$ be the Palm (conditional) probability that one job arrives at time 0. In other words, in the conditional space, the arrival points are

$$\cdots < -T_{-2} - T_{-1} < -T_{-1} < 0 < T_1 < T_1 + T_2 < \ldots .$$

For brevity, we shall drop the superscript in $\mathbb{P}^0$ in this section; also, the expectation $\mathbb{E}[\cdot]$ is relative to the same Palm measure.

Denote the service times by $(S_j)_{j\in\mathbb{Z}}$. As mentioned in § 2.2, we will analyze the GI/G/1 queue by framing it in a time domain model: Define the compound arrival process up to time 0 as

$$A(t) := \sum_{j=1}^{N(t)} S_{-j}$$

for $t > 0$ and $A(0) := 0$, where $N(t)$ is the counting process

$$N(t) := \max\left\{ n \in \mathbb{N} \mid \sum_{j=1}^{n} T_{-j} \leq t \right\} .$$

(again, for brevity, we prefer to write $A(t)$ instead of $A(-t)$, and similarly for $N(t)$).

The stationary waiting time distribution is

$$\mathbb{P}(W \geq \sigma) = \mathbb{P}\left(\sup_{t\geq 0}\{A(t)-t\} \geq \sigma\right) . \tag{10}$$

Recall that $\mathbb{P}$ is the Palm measure under having an arrival at time 0. The event in the right-hand side (Palm) probability corresponds to the waiting time of the arrival at 0; while slightly cumbersome for a single queue, the Palm representation will be helpful in the multiclass case.

Let us remark that unless $N(t)$ is Poisson then neither the exponential process

$$X_t := e^{\theta(A(t)-t)} ,$$

nor a re-weighed one with $A(t)$ replaced by $N(t)$ can be martingales, for non-trivial values of θ. To enable martingale constructions suitable for Theorem 7, we shall regard $N(t)$ as an inhomogeneous Poisson process with a random rate $\lambda(R(t))$ where

$$R(t) := t - \sum_{j=1}^{N(t)} T_{-j} ,$$

i.e., the time elapsed from some time $-t$ to the first arrival time (also called the remaining lifetime in the language of renewal processes), whereas $\lambda(s)$ is the hazard rate

$$\lambda(s) := \lim_{\Delta s\to 0} \frac{\mathbb{P}(s < T_1 \leq s+\Delta s \mid s < T_1)}{\Delta s} = \frac{f(s)}{1-F(s)} ,$$

and $f()$ and $F()$ are the density and distribution functions of T_1 (under the original probability measure); note that the hazard rate resets itself at the arrival times $\sum_j T_{-j}$.

We can now apply Lemma 6 to construct a martingale for the GI/G/1 queue:

Corollary 8. GI/G/1 Martingale (Time Domain) *In the scenario above, let θ satisfying $E\left[e^{-\theta T_1}\right] E\left[e^{\theta S_1}\right] = 1$ and*

$$h(t) := \frac{1 - E\left[e^{\theta S_1}\right] \int_0^t e^{-\theta s} f(s) ds}{e^{-\theta t}\left(1-F(t)\right)} .$$

Then the process

$$h(R(t)) e^{\theta(A(t)-t)}$$

is a martingale.

The condition on θ ensures the non-negativity of $h()$.

Proof. Let a time t. Since T_i's are independent, the probability that a job arrives during $(t, t+\Delta t]$ is $\lambda(R(t))\Delta t + o(\Delta t)$ where $\lim_{t\to 0} \frac{o(\Delta t)}{\Delta t} = 0$. Note that the hazard rate replaces the constant rate λ in the case of the Poisson process, and that we are in the context of Lemma 6 with $M_t = R(t)$.

Due to the underlying renewal property, we can assume without loss of generality that $t \in [0, T_1)$, i.e., $R(t) = t$. The martingale condition from Lemma 6 becomes

$$\lim_{\Delta t\to 0} \frac{1}{\Delta t}\Big[\lambda(t)\Delta t h(0)\mathbb{E}[e^{\theta S_1}] e^{-\theta\Delta t} + (1-\lambda(t)\Delta t) h(t+\Delta t) e^{-\theta\Delta t} - h(t)\Big] = 0 .$$

Note that in the first term we do have $h(0)$, and not $h(t+\Delta t)$, because a job arrival "refreshes" the counter $R(t)$. Taking the limit and applying Taylor's expansion (i.e., $e^{x\Delta t} = 1 + x\Delta t + o(\Delta t)$) leads to the ODE

$$h'(t) = h(t)\left(\lambda(t)+\theta\right) - \lambda(t) h(0) \mathbb{E}[e^{\theta S_1}] . \tag{11}$$

By setting the initial value problem with $h(0) = 1$ the proof is complete. □

Next we give three applications of Corollary 8 to ΣGI/G/1 queues.

4.1 Example 1: ΣWeibull/G/1

There are N mutually independent homogeneous classes (indexed by i) having Weibull distributed interarrivals $T_{i,j}$ with scale parameter 1 and shape parameter 2, i.e., $\mathbb{P}(T_{1,1} \leq t) = 1 - e^{-t^2}$ for which $E[T_{1,1}] = \frac{\sqrt{\pi}}{2}$. To have a utilization factor $\rho < 1$, the service times of the jobs $S_{i,j}$ satisfy $E[S_{1,1}] = \frac{\sqrt{\pi}}{2N}\rho$.

Corollary 9. *A bound on the waiting time for each class is*

$$\mathbb{P}(W \geq \sigma) \leq K(\theta)^{N-1} e^{-\theta N \sigma} ,$$

where

$$K(\theta) := E\left[e^{\theta N S_{1,1}}\right] e^{\frac{\theta^2}{4}} erfc\left(\frac{\theta}{2}\right)$$

and θ satisfies $E\left[e^{-\theta T_1}\right] E\left[e^{\theta N S_{1,1}}\right] = 1$.

We use the standard notation $erf(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2} ds$ and $erfc(x) := 1 - erf(x)$; $E\left[e^{-\theta T_1}\right]$ is given in (21).

Recalling that we work with a Palm measure, the (Palm) bound holds for the arrivals of a particular class. It is important to remark that in the case of a single class ($N = 1$), the bound (relying on a continuous-time martingale) recovers Kingman's bound from Theorem 1 (relying on a discrete-time martingale); that is because $R(0) = 0$ and thus $h()$ is a constant. In the case of $N-1$ additional classes, we need to keep track of the remaining lifetimes of these at time 0—when an arrival from the first class happens—which essentially lend themselves to the prefactor $K(\theta)^{N-1}$ (for more details see the proof).

4.2 Example 2: ΣErlang-k/G/1

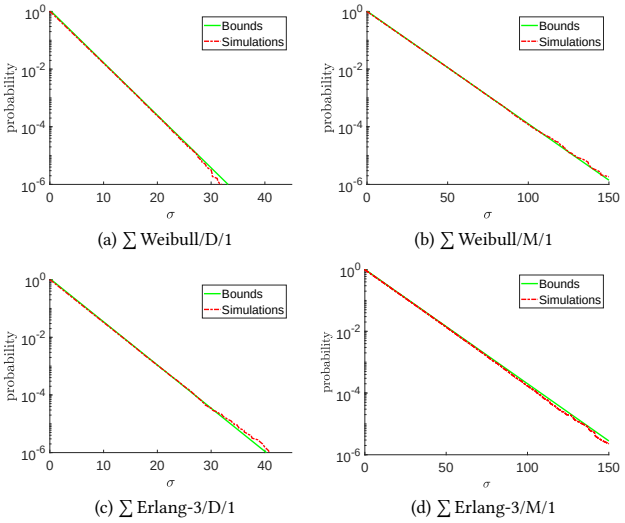
Here $T_{i,j}$ are Erlang-k distributed with parameter λ, i.e., $E[T_{1,1}] = \frac{k}{\lambda}$. The service times satisfy $E[S_{1,1}] = \frac{k}{\lambda N}\rho$.

Corollary 10. *A bound on the waiting time is the same as in Corollary 9 except for*

$$K(\theta) := \frac{\lambda}{k} \frac{E\left[e^{\theta N S_{1,1}}\right] - 1}{\theta} ,$$

and θ satisfying $\left(1+\frac{\theta}{\lambda}\right)^{-k} E\left[e^{\theta N S_{1,1}}\right] = 1$.

Figs. 1.(a-d) illustrate upper bounds vs. simulations for the CCDF of the waiting time in heavy-traffic ($\rho = 0.99$). In the Erlang-k case, $\lambda := \frac{2k}{\sqrt{\pi}}$ such that $E\left[T_{1,1}\right]$ is the same as in the Weibull case. The simulations are obtained from 10^7 samples, each representing the waiting time of the 10^5th job starting from an empty system. The tail instability is due to the simulation length; note that $\Theta\left(10^{12}\right)$ simulation runtime is insufficient to render stable tails in the shown



(a) ∑ Weibull/D/1 (b) ∑ Weibull/M/1

(c) ∑ Erlang-3/D/1 (d) ∑ Erlang-3/M/1

Figure 1: Waiting-time CCDF (upper bounds vs. simulations); ($N = 5$, $\rho = 0.99$)

intervals. Besides the accuracy of the bounds, an interesting observation is that in the case of constant service times, the inter-arrival distribution makes a substantial difference on waiting times; this effect disappears however in the case of exponential service times. Appendix § B provides additional simulations (Fig. 9) illustrating that the bounds degrade at lower utilizations, and especially for constant service times.

The issue of the bounds' tightness is closely related to the estimation of the *overshoot*. Having a (Markov) random walk with increments $(U_i)_i$, and a value $\sigma \geq 0$, the overshoot is defined as

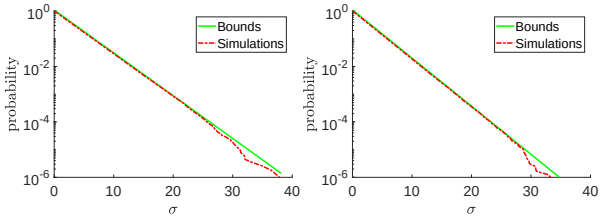
$$R_\sigma = \inf\{U_1 + U_2 + \cdots + U_n - \sigma \mid U_1 + U_2 + \cdots + U_n \geq \sigma\} .$$

In the proof of Theorem 7, the derivation of the bounds mainly relies on the crude estimation $R_\sigma \geq 0$; see also the discussion around Lemma 2. Without resorting on a rigorous argument, we believe that in heavy-traffic the last increment behaves as a typical increment, whereas in lower-traffic the last increment gets larger; ignoring this information is a possible cause for the bounds degradation. For potential improvements of the crude overshoot estimation see Chang [11].

4.3 Example 3: ΣWeibull + ΣErlang-k/G/1

Let us now consider a heterogeneous mix of N_1 Weibull and N_2 Erlang-k classes, mutually independent. We use the same parameters as before, including $\lambda := \frac{2k}{\sqrt{\pi}}$ in the Erlang-k case, to normalize the arrival rates of the two classes. The service times satisfy $E[S_{1,1}] = \frac{k}{\lambda N}\rho$ where $N := N_1 + N_2$ and ρ is the overall utilization. Denote the Weibull and Erlang-k compound processes as $A_i(t)$ for $i = 1 \ldots N_1$ and $i = N_1 + 1, \ldots, N$, respectively.

We next illustrate the algorithm for computing a waiting time bound in the case of heterogeneous input. Recall the key idea from § 3.2 of obtaining martingales with the same 'θ' for both classes (in this case N_1 Weibull and N_2 Erlang-k), and also the



(a) $N_1 = 1$ and $N_2 = 4$ (b) $N_1 = 4$ and $N_2 = 1$

Figure 2: Waiting-time CCDF for a Weibull job; N_1 Weibull and N_2 Erlang-k classes; constant (D) service times; ($N = 5$, $k = 3$, $\rho = 0.99$)

proofs of Corollaries 9 and 10. We thus look for a split

$$w_1 N_1 + w_2 N_2 = N$$

which yields the martingales

$$h_W(R_1(t))e^{\theta_1 \frac{N}{w_1}(A_1(t) - \frac{w_1}{N}t)}$$

for a single Weibull compound process $A_1(t)$ and

$$h_E(R_{N_1+1}(t))e^{\theta_2 \frac{N}{w_2}(A_{N_1+1}(t) - \frac{w_2}{N}t)}$$

for a single Erlang-k compound process $A_2(t)$; the 'W' and 'E' subscripts correspond to the two classes.

The same 'θ' constraint reduces to

$$\theta := \frac{\theta_1 N}{w_1} = \frac{\theta_2 N}{w_2} .$$

We also note the additional constraints on w_1 and w_2 to guarantee the existence of the two martingales above

$$\rho < w_1 < \frac{N - N_2\rho}{N_1} ,$$

which are merely stability conditions (e.g., the rate of $A_1(t)$ is less than $\frac{w_1}{N}$). The existence of w_1 satisfying the same 'θ' constraint is guaranteed by the continuity of $f_1(w_1) := \frac{\theta_1 N}{w_1}$ and $f_2(w_1) := \frac{\theta_2 N}{w_2}$, and the extreme points $f_1(\rho) = 0$ (because the corresponding θ_1 is zero) and $f_2(\frac{N-N_2\rho}{N_1}) = 0$.

Multiplexing N_1 Weibull classes and N_2 Erlang-k classes yields the martingale

$$\prod_{i=1}^{N_1} h_W(R_i(t)) \prod_{i=N_1+1}^{N} h_E(R_i(t))e^{\theta(A(t)-t)}$$

where $A(t) := \sum_{i=1}^{N} A_i(t)$ is the overall compound process. Therefore, a bound on the waiting-time of a Weibull class is

$$\mathbb{P}(W \geq \sigma) \leq K_W(\theta)^{N_1-1} K_E(\theta)^{N_2} e^{-\theta\sigma} ,$$

where $K_W(\theta)$ and $K_E(\theta)$ are the $K(\theta)$'s from Corollaries 9 and 10, respectively. In turn, the waiting time of an Erlang-k class is the same except for the prefactor $K_W(\theta)^{N_1} K_E(\theta)^{N_2-1}$.

We illustrate the accuracy of these bounds for a ΣWeibull + ΣErlang-k/D/1 queue in Fig. 2; both cases of disproportionate Weibull and Erlang-k classes relative to the other are addressed in (a) and (b). The numerical settings are the same as in Fig. 1. Results with similar accuracy were obtained for exponential service jobs (not

shown here), whereas the accuracy of the bounds degrade at lower utilization (similar as in Fig. 9 from Appendix § B).

5 APPLICATION 2: QUEUES WITH MARKOVIAN ARRIVALS

We now apply Lemma 5 to several subclasses of MAPs from teletraffic theory: Markov Modulated Fluid (MMF, § 5.1), Markov Modulated Poisson Process (MMPP, § 5.2), and (Generalized) Markovian Arrival Processes ((G)MArP, § 5.3).

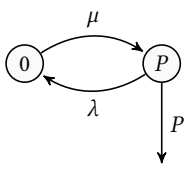


Figure 3: MMOO process

5.1 Fluid Scenario. MMF

The MMF model assumes that data is infinitely divisible (i.e., a continuous 'fluid'), whereas a background process M_t determines the rate at which the fluid arrives at the server:

$$A(t) = \int_0^t M_s ds \ . \tag{12}$$

In the basic Markov-Modulated On-Off (MMOO) model [2], M_t has two states (denoted for convenience 0 and P) with transition rates λ and μ (see Fig. 3). While in state 0 (also referred to as 'off') the process does not generate any fluid; while in state P (also referred to as 'on') the process generates 'fluid' at some constant rate P.

Before applying Lemma 5, we remark that the parameter C has the meaning of the rate of a hypothetical queueing server for the process $A(t)$. To avoid trivial situations we assume that $P > C$ (i.e., the peak rate is greater than the capacity) and that the utilization factor $\rho = \frac{\frac{\mu}{\lambda+\mu}P}{C}$ satisfies the stability condition $\rho < 1$.

Corollary 11. (Single MMOO) *In the scenario above, let*

$$\theta := \frac{\lambda}{P-C} - \frac{\mu}{C} \ , \quad h(P) := \frac{\theta C + \mu}{\mu} \ , \quad \textit{and} \quad h(0) := 1 \ .$$

Then the process

$$h(M_t)e^{\theta(A(t)-Ct)}$$

is a martingale.

Proof. We distinguish two cases. First, if $M_0 = 0$, then in a small interval $[0, \Delta s]$ the process M_s jumps to the 'on'-state with probability $\mathbb{P} \approx \mu\Delta s$ (more precisely $\mathbb{P} = \mu\Delta s + o(\Delta s)$). We have

$$\begin{aligned}
&\frac{d}{ds}\varphi_0(s)\Big|_{s=0} \\
&= \lim_{\Delta s\to 0} \frac{1}{\Delta s}\mathbb{E}\Big[h(M_{\Delta s})e^{\theta(A(\Delta s)-C\Delta s)} - h(0)\Big|M_0 = 0\Big] \\
&= \lim_{\Delta s\to 0} \frac{1}{\Delta s}\left(\mu\Delta s h(P)e^{\theta\Delta s(P-C)} + (1-\mu\Delta s)e^{-\theta C\Delta s} - 1\right) \\
&= \mu h(P) - \mu - \theta C = 0 \ ,
\end{aligned}$$

after applying Taylor's expansion $e^{x\Delta s} = 1 + x\Delta s + o(\Delta s)$.

Similarly, if $M_0 = P$ then the process jumps in $[0, \Delta s]$ with probability $\mathbb{P} \approx \lambda\Delta s$ so that

$$\begin{aligned}
&\frac{d}{ds}\varphi_P(s)\Big|_{s=0} \\
&= \lim_{\Delta s\to 0} \frac{1}{\Delta s}\mathbb{E}\Big[h(M_{\Delta s})e^{\theta(A(\Delta s)-C\Delta s)} - h(P)\Big|M_0 = P\Big] \\
&= \lim_{\Delta s\to 0} \frac{1}{\Delta s}\left(\lambda\Delta s e^{-\theta C\Delta s} + (1-\lambda\Delta s)h(P)e^{\theta\Delta s(P-C)} - h(P)\right) \\
&= \lambda - \lambda h(P) + h(P)\theta(P-C) \\
&= h(P)\left(\lambda\frac{\mu}{\theta C+\mu} - \lambda + \theta(P-C)\right) \\
&= h(P)\left(\lambda\frac{\mu(P-C)}{C\lambda} - \lambda + \lambda - \frac{\mu(P-C)}{C}\right) = 0 \ .
\end{aligned}$$

□

The MMOO martingale appeared in a general form for Markov fluids in Ethier and Kurtz [21] (see Lemma 3.2 therein), which was instantiated in the MMOO case by Palmowski and Rolski [40]. Note that Corollary 11 not only provides an elementary proof, but it also guarantees the unicity of exponential martingales of the form from Eq. (9) for the MMOO process (subject to a fixed C).

Next we consider an aggregate of N MMOO processes represented in Fig. 4. The corresponding aggregate process is $A(t)$ and the background process with $N+1$ states is M_t; the utilization factor $\rho = \frac{\frac{\mu}{\lambda+\mu}PN}{C}$ satisfies $\rho < 1$.

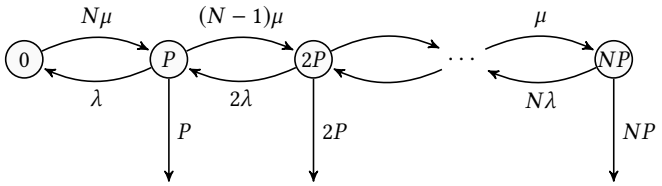


Figure 4: An aggregate of N MMOO processes

Corollary 12. (Multiplexed MMOO) *In the scenario above, let*

$$\theta = \frac{N}{C}\left(\frac{\lambda C}{NP-C} - \mu\right), \ h(iP) = \left(1 + \frac{C\theta}{N\mu}\right)^i \ i = 0, \ldots, N \ .$$

Then the process

$$h(M_t)e^{\theta(A(t)-Ct)}$$

is a martingale.

Bounds on the waiting time distribution follow directly from Theorem 7. Denoting for convenience $c := \frac{C}{N}$ and $b := 1 + \frac{c\theta}{\mu}$ we have

$$\mathbb{P}(W \geq \sigma) \leq \frac{\sum_{i=0}^N \pi_i b^i}{b^{\frac{c}{P}}} e^{-\theta\sigma} \ ,$$

where $\pi_i = \binom{N}{i}\left(\frac{\mu}{\lambda+\mu}\right)^i\left(\frac{\lambda}{\lambda+\mu}\right)^{N-i}$ are the stationary probabilities of M_t. We deliberately used the weaker bound with $b^{\frac{c}{P}}$, instead of $b^{\lceil\frac{c}{P}\rceil}$, which lends itself to the 'expressive' bound from [15]

$$\mathbb{P}(W \geq \sigma) \leq K^N e^{-\theta\sigma}\ , \tag{13}$$

where $K := \rho\left(\frac{\rho-p_{on}}{1-p_{on}}\right)^{\frac{p_{on}}{\rho}-1} < 1$ and $p_{on} := \frac{\mu}{\lambda+\mu}$; the same bound appeared in [40] yet without the explicit exponential representation of the prefactor. We also note that in the application of Theorem 7 we have $\mathrm{Im}(M_t) = \{\lceil\frac{c}{P}\rceil, \ldots, N\}$ because at least $\lceil\frac{c}{P}\rceil$ individual sources must be 'on' to guarantee $a(T) \geq C$ at the stopping time T; the rest follows from the monotonicity of $h(iP)$. The bounds from (13) are accurate, at both high ($\rho = .9$) and moderate ($\rho = .75$) utilizations, as illustrated through simulations in [15]. The fundamental reason is that the bound from (13) captures the right scaling in N, as conjectured by Choudhury *et al.* [13].

5.2 Packet Scenario. MMPP

Here we analyze the 'packetized' version of the MMF model; we consider both constant and random packet sizes.

5.2.1 Constant Packet Size. Data consists of indivisible units (i.e., 'packets') of size 1. The instantaneous probability of a packet arrival is determined by a background process M_t, whereas the cumulative arrivals process $A(t)$ evolves according to

$$\mathbb{P}(A(t+\Delta t) - A(t) = 1) = r(M_t)\Delta t + o(\Delta t)\ , \tag{14}$$

where $r(\cdot)$ is a rate function. For instance, we let M_t be the Markov process from Fig. 5a, i.e., state space $\{1, 2\}$ and transition rates μ_1 and μ_2, in which case $r(1) = \lambda_1$ and $r(2) = \lambda_2$.

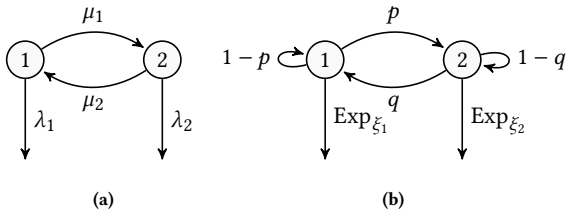


Figure 5: MMPP (a) and packet size modulator (b)

To construct a martingale from $A(t)$ using Lemma 5 we need the following matrix transform: For $\theta > 0$, let

$$T_\theta := \begin{pmatrix} \lambda_1 e^\theta - \mu_1 - \lambda_1 & \mu_1 \\ \mu_2 & \lambda_2 e^\theta - \mu_2 - \lambda_2 \end{pmatrix}$$

and denote by $\lambda(\theta)$ its spectral radius.

Corollary 13. *In the scenario above, pick $\theta > 0$ such that $\lambda(\theta) = \theta C$, and let $h = (h_1, h_2)$ be an eigenvector corresponding to T_θ and $\lambda(\theta)$. Then the process*

$$h(M_t)e^{\theta(A(t)-Ct)}$$

is a martingale; for notation's convenience $h(i) \equiv h_i$.

We next apply Theorem 7 in the case of N multiplexed (homogeneous) MMPPs $A_i(t)$, with background processes $M_{i,t}$, served at rate C, and utilization $\rho < 1$. Letting the individual martingales

$$h(M_{i,t})e^{\theta\left(A_i(t)-\frac{C}{N}t\right)}$$

with $h(\cdot)$ and θ as in Corollary 13 (with C replaced by $\frac{C}{N}$), the aggregate martingale is

$$\prod_i h(M_{i,t})e^{\theta(\sum_i A_i(t)-Ct)}\ .$$

We then obtain the following upper bound on the waiting time

$$\mathbb{P}(W \geq \sigma) \leq \frac{E\left[h(M_{1,0})\right]^N}{\min\{h_1, h_2\}^N}e^{-\theta C\sigma}\ . \tag{15}$$

Assuming that the system is initially stationary, $E\left[h(M_{1,0})\right] = h_1\frac{\mu_2}{\mu_1+\mu_2} + h_2\frac{\mu_1}{\mu_1+\mu_2}$. The lower bound is similar except for replacing the 'min' by 'max', and σ by $\sigma + 1$ (as packets have size 1).

5.2.2 Random Packet Size. We extend the previous model from constant to random packet sizes. We assume that a Markov chain L_n determines the size of the n-th packet. The chain L_n alternates between two states with transition probabilities p and q as in Fig. 5b. The packets are exponentially distributed with rates ξ_1 and ξ_2 depending on the chain's state; other types of distributions can be considered. Note that in the case $\xi_1 = \xi_2$ we have the scenario with i.i.d. packet sizes.

If $A(t)$ is the cumulative arrival process with constant packet sizes (as in Subsection § 5.2.1), the arrival process with random packets $A^{\mathrm{rnd}}(t)$ has the representation

$$A^{\mathrm{rnd}}(t) := \sum_{k=1}^{A(t)} S_{L_k,k}\ ,$$

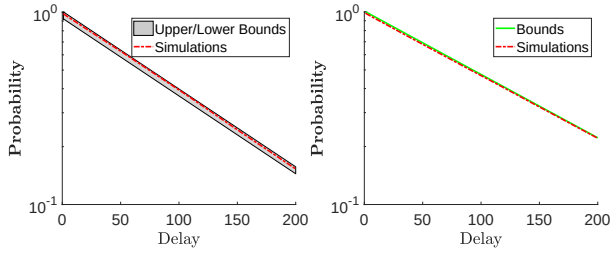
where $(S_{1,k})_{k\in\mathbb{N}}$ and $(S_{2,k})_{k\in\mathbb{N}}$ are i.i.d. sequences of exponential random variables with rates ξ_1 and ξ_2, respectively. Note that the process

$$\left(A^{\mathrm{rnd}}(t), \left(M_t, L_{A(t)}\right)\right)$$

is a MAP in the sense of Definition 4.

In order to apply Lemma 5 to this example, we need the following matrix transform T_θ for $\theta > 0$

$$T_\theta := \begin{pmatrix} (1-p)\,\lambda_1\mathbb{E}e^{\theta S_{1,1}} - \mu_1 - \lambda_1 & p\lambda_1\mathbb{E}e^{\theta S_{2,1}} & \mu_1 & 0 \\ q\lambda_1\mathbb{E}e^{\theta S_{1,1}} & (1-q)\,\lambda_1\mathbb{E}e^{\theta S_{2,1}} - \mu_1 - \lambda_1 & 0 & \mu_1 \\ \mu_2 & 0 & (1-p)\,\lambda_2\mathbb{E}e^{\theta S_{1,1}} - \mu_2 - \lambda_2 & p\lambda_2\mathbb{E}e^{\theta S_{2,1}} \\ 0 & \mu_2 & q\lambda_2\mathbb{E}e^{\theta S_{1,1}} & (1-q)\,\lambda_2\mathbb{E}e^{\theta S_{2,1}} - \mu_2 - \lambda_2 \end{pmatrix}\ .$$



(a) constant (b) random

Figure 6: Waiting-time CCDF for N MMPPs; constant and random packet sizes; ($N = 5$, $\mu_1 = 0.1$, $\mu_2 = 0.5$, $\lambda_1 = 1$, $\lambda_2 = 25$, $p = 0.1$, $q = 0.9$, $E[\xi_1] = 0.2$, $\rho = 0.99$)

Let $\lambda(\theta)$ be its spectral radius.

Corollary 14. *In the scenario above, pick $\theta > 0$ such that $\lambda(\theta) = \theta C$, and let $h = (h_{1,1}, h_{1,2}, h_{2,1}, h_{2,2})$ be an eigenvector corresponding to T_θ and $\lambda(\theta)$. Then the process*

$$h(M_t)e^{\theta\left(A^{rnd}(t)-Ct\right)}$$

is a martingale.

An upper bound on the waiting time is the same as in Eq. (15) except for the denominator in the prefactor, which is replaced by $\min\{h_{1,1}, h_{1,2}, h_{2,1}, h_{2,2}\}^N$ according to Corollary 14. In turn, a lower bound cannot be obtained with Theorem 7 because packet sizes are unbounded.

Figure 6 illustrates the accuracy of the bounds in the case of an aggregate of MMPP flows in heavy-traffic ($\rho = 0.99$). Both cases of constant and random-size packets are considered; in both cases the upper bound and simulation lines almost overlap, the former being slightly above the other. Simulations are obtained from a run of 10^{10} packets of which the first 10% were discarded. Additional simulations for smaller utilization $\rho = 0.75$ are shown in Figure 10 in Appendix § B.

5.3 Packet Scenario. MArP and GMArP

As in the MMPP case we address both constant and random packet sizes.

5.3.1 Constant Packet Size. First we consider Markovian Arrival Processes (MArPs) that generalize the Markov Modulated Poisson processes from § 5.2.1.

Definition 15. *A Markovian Arrival Process is defined via a pair (D_0, D_1) of $n \times n$-matrices such that:*

$$d_{i,j} := D_0(i,j) \geq 0\,, i \neq j\,, \quad d'_{i,j} := D_1(i,j) \geq 0\,,$$
$$d_{i,i} := D_0(i,i) = -\sum_{i\neq j} d_{i,j} - \sum_j d'_{i,j}\,.$$

*The background process M_t is a Markov process with generator D_0+D_1 and steady-state distribution π. If a transition of M_t is triggered by an element of D_1, a packet is generated and $A(t)$ increases by 1 (*active transitions*); transitions triggered by D_0 do not increase $A(t)$ (*hidden transitions*):*

$$\mathbb{P}\left(A(t, t+\Delta t) = 0, M_{t+\Delta t} = j \mid M_t = i\right) = D_0(i,j)\Delta t + o(\Delta t)\,,$$

and

$$\mathbb{P}\left(A(t, t+\Delta t) = 1, M_{t+\Delta t} = j \mid M_t = i\right) = D_1(i,j)\Delta t + o(\Delta t)\,.$$

Corollary 16. *In the scenario above, for $\theta > 0$, let $\lambda(\theta)$ be the spectral radius of the matrix*

$$D_0 + e^\theta D_1\,.$$

If $\lambda(\theta) = \theta C$ and h is a corresponding eigenvector then the process

$$h(M_t)e^{\theta(A(t)-Ct)} \tag{16}$$

is a martingale. Moreover, if h^r is an eigenvector corresponding to the spectral radius of the transform matrix

$$\Pi^{-1}\left(D_0 + e^\theta D_1\right)^T \Pi\,,$$

where Π is the matrix with the steady state distribution π on its diagonal, then the process

$$h^r(M_t^r)e^{\theta(A^r(t)-Ct)}$$

is a martingale as well.

An immediate consequence of the second part of the Corollary is that in the general case of not necessarily reversible processes, an upper bound on the waiting time is the same as in (15), except for accounting for the "reversed" eigenvector h^r.

A key property of MArPs is their stability under superposition: Given two MArPs $(A(t), M_t)$ and $(A'(t), M'_t)$ with corresponding matrices (D_0, D_1) and (D'_0, D'_1), respectively, the aggregate arrival process $A(t) + A'(t)$ is a MArP with matrices

$$(D_0 \oplus D'_0, D_1 \oplus D'_1)\,,$$

where '$\oplus$' stands for the Kronecker sum. The next result gives the resulting martingale:

Corollary 17. *In the situation with two MArPs as above, for $\theta > 0$, let $\lambda(\theta)$ and $\lambda'(\theta)$ denote the spectral radii of the matrices*

$$D_0 + e^\theta D_1 \text{ and } D'_0 + e^\theta D'_1\,,$$

respectively; let also h and h' be the corresponding eigenvectors. If $\lambda(\theta) + \lambda'(\theta) = \theta C$ then the process

$$h(M_t)h'(M'_t)e^{\theta(A(t)+A'(t)-Ct)}$$

is a martingale.

The result generalizes immediately to any number of MArPs.

5.3.2 Random Packet Size. We finally consider Generalized Markovian Arrival Processes (GMArPs) that generalize the MArPs from § 5.3.1 by allowing for random packet sizes.

Definition 18. *A Generalized Markovian Arrival Process (GMArP) is defined via a sequence $(\mathcal{L}_k)_{1\leq k<\infty}$ of strictly positive distributions and a sequence $(D_k)_{0\leq k<\infty}$ of $n \times n$-matrices such that*

$$D_k(i,j) \geq 0\,, i \neq j\,, \quad \text{for all } k \geq 0\,, \quad \text{and}$$
$$D_0(i,i) = -\sum_{i\neq j} D_0(i,j) - \sum_{k=1}^{\infty}\sum_j D_k(i,j)\,.$$

The background process M_t is a Markov process with generator $\sum_{k=0}^\infty D_k$, and π denotes its steady-state distribution. If a transition of M_t is triggered by an element of D_k, a packet is generated with size given

by $\mathcal{L}_k$. Accordingly, $A(t)$ increases by X_k, i.e., a random variable independently drawn from the distribution $\mathcal{L}_k$.

If in the above definition we let $D_k := 0$ for all $k \geq 2$, and $\mathcal{L}_1 := \delta_1$, i.e., the deterministic distribution on 1, we recover the MArP scenario from the previous section. Moreover, if only $\mathcal{L}_k := \delta_k$, i.e., the deterministic distribution on k, GMArP instantiates to the Batch Markovian Arrival Process (BMArP) [37].

Corollary 19. *In the scenario above, for $\theta > 0$, let $\lambda(\theta)$ denote the spectral radius of the matrix*

$$\sum_{k=0}^{\infty} \mathbb{E}[e^{\theta X_k}] D_k \ .$$

If $\lambda(\theta) = \theta C$, and h is a corresponding eigenvector, then the process

$$h(M_t)e^{\theta(A(t)-Ct)}$$

is a martingale. Moreover, if h^r is an eigenvector corresponding to the spectral radius of the transposed matrix

$$\Pi^{-1}\left(\sum_{k=0}^{\infty} \mathbb{E}[e^{\theta X_k}] D_k\right)^T \Pi \ ,$$

where Π denotes the matrix with the steady state distribution π on its diagonal, then the process

$$h^r(M_t^r)e^{\theta(A^r(t)-Ct)}$$

is a martingale as well.

Proof. Analogously to the proof of Corollary 16. □

We also note that multiplexing GMArPs can be treated in the same manner as in Corollary 17, whereas a bound on the waiting time follows exactly as in the MArP case.

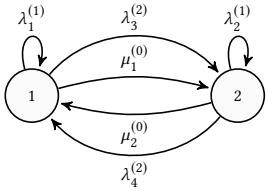


Figure 7: Example of GMArP

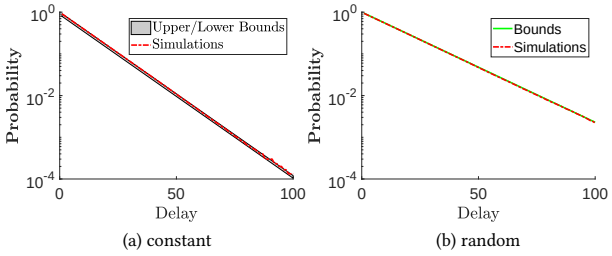
To provide numerical results we consider the GMarP process from Fig. 7. By convention, the superscript in each transition corresponds to the 'k' from Def. 18. More precisely

$$D_0 = \begin{bmatrix} -\lambda_1 - \lambda_3 - \mu_1 & \mu_1 \\ \mu_2 & -\lambda_2 - \lambda_4 - \mu_2 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \ D_2 = \begin{bmatrix} 0 & \lambda_3 \\ \lambda_4 & 0 \end{bmatrix} \ .$$

Note that unlike λ_1 and λ_2, the transitions λ_3 and λ_4 involve a change of state, in addition to drawing a packet size from a different distribution.

In Fig. 8 we consider an aggregate of $N = 5$ homogeneous GMArPs, and both constant and exponential packet sizes. The numerical settings normalize the average rate as in the MMPP case



(a) constant (b) random

Figure 8: Waiting-time CCDF for N GMArPs; constant and random packet sizes; ($N = 5$, $\mu_1 = 0.1$, $\mu_2 = 0.5$, $\lambda_1 = 0.3$, $\lambda_2 = 10$, $\lambda_3 = 0.7$, $\lambda_4 = 15$, $E[X_1] = 1$, $E[X_2] = 3.01$, $\rho = 0.99$)

(Fig. 6); however, we now consider much burstier processes. Simulations are run as in the MMPP case; similarly, the upper bound and simulation lines almost overlap.

Let us now comment on the numerical complexity in analyzing queues with a superposition of N BMArP. The standard approach consists in computing the generator matrix of the superposed process, which has an exponential number of states (in N) as a consequence of the Kronecker product. Exact results (e.g., on the waiting time distribution) can be obtained by applying a mix of matrix-analytic techniques and inversion algorithms of Laplace transforms (for an overview see [37]). A computationally more effective approach in the case of MArPs consists in building a n-dimensional Markov process, where n is the number of states for each (i.i.d.) MArP; the overall number of states is $\binom{N+n-1}{n-1}$ which is generally much smaller than the exponential. This approach has its roots in the analysis of GI/PH/N queues [43]; for a discussion of the applications of this approach, including queues with superposed MArPs, see [24]. In turn, bounding approaches as in this paper or the literature (e.g., [9, 36]) are subject to a linear complexity.

6 DISCUSSION

Here we discuss some related work in more detail and comment on possible extensions of our results.

6.1 Related Work

Kingman's GI/G/1 bound from (4) was extended to the case of *discrete-time* MAPs in Chang and Cheng [10]. Using a different martingale transform, Duffield [18] improved the bounds by essentially capturing the positiveness of the instantaneous drift at the underlying stopping time (this fact holds by default in the renewal case and does not have to be properly accounted for). This improvement can be substantial because in some cases, e.g., bursty On-Off processes whereby the sum of the transition probabilities between the two states is less than 1, the prefactor in the exponential bound is also less than 1; in turn, the prefactor from [10] is always greater or equal than 1. Another martingale transform was constructed by Fang *et al.* [22] using a fixed point argument in the case of the G/GI/1 queue, allowing for Markovian inter-arrivals; while there is similarity to Duffield's approach (which essentially relies on the eigenvalue/eigenvector problem – a fixed point problem itself), a qualitative comparison is challenging due to the different bounds' structures.

In a more recent work, Jiang and Misra [29] obtained bounds in $\Sigma GI/G/1$ queues. In the $\Sigma D/D/1$ case, tight worst-case bounds are obtained by relying on network calculus models and techniques. The general case is treated by discretizing time and then directly applying Kingman's technique, as outlined in § 2. A proof for the claimed discrete-time martingale is however not given, and we believe that it may be challenging due to the loss of the renewal property in the general case. For Poisson arrivals, the renewal property is preserved under superposition and the martingale construction holds; the obtained bounds—which are essentially the same as in this work, as well as in [30] by properly instantiating the general results—are shown to be numerically accurate.

Kingman also provided a more powerful GI/G/1 bound in [31]. In the notation from § 2.1

$$\mathbb{P}(W \geq \sigma) \leq \gamma(\sigma)\,,$$

where $\gamma(\sigma)$ is a non-increasing function with $0 \leq \gamma(\sigma) \leq 1$ such that for all $\sigma > 0$

$$\int_{-\infty}^{\sigma} \gamma(\sigma - y) dF(y) + 1 - F(\sigma) \leq \gamma(\sigma)\,, \tag{17}$$

where $F(y)$ is the distribution of U_1. The bound facilitates the discovery of tighter bounds than the original bound from (4), which is recovered with $\gamma(\sigma) := e^{-\theta\sigma}$.

This idea was exploited by Liu, Nain, and Towsley [35, 36] in the case of general discrete-time MAPs, whereby the background Markov chain can have a general state space. The method extends immediately to continuous-time MAPs by embedding a Markov chain to account for the the (discrete-time) structure of the integral inequality from (17). Notably, the obtained bounds are *exact* for the GI/M/1 queue, which also holds for Ross' bounds from [46] (see (7)); based on this match, it is of interest to qualitatively compare the bounds from [36, 46] (see the proof of Lemma 3 for the extension of Ross bounds to the non-renewal case).

Such a qualitative comparison is provided in [35, 36] for the bounds therein and those from [18], and also from Asmussen and Rolski [5]; the latter are derived in the context of risk theory (for the analogy between ruin probabilities and tail bounds on waiting time see [3]). A deep comparison is however very challenging due to the different structures of the bounds. Numerical comparison between the three bounds (and also some corresponding lower bounds) are given in [36]; we reproduce some tables in Appendix § B (see Figs. (12) and (13), and include our bounds from § 5.2.1 for the MMPP/D/1 queue (see (15)) and § 5.2.2 for the MMPP/M/1 queue; we refer to our bounds as CP (the authors' initials), and to the other three similarly (LNT-Liu/Nain/Towsley, D-Duffield, and AR-Asmussen/Rolski). In the MMPP/D/1 case the CP-bounds are essentially identical to the AR-bounds. In the MMPP/M/1 case the CP-bounds are only slightly better than the D-bounds, which were identified in [36] as the loosest for the numerical settings therein.

From a qualitative point of view, the CP-bounds are most 'similar' to the D-bounds. The fundamental difference is that the CP-bounds are derived exclusively in continuous-time, using a continuous-martingale, whereas the D-bounds are derived in discrete-time but using the same technique from Theorem 7 extending Kingman's original idea to the non-renewal case. A slight difference is that the CP-bounds hold for the virtual delay process whereas the D-bounds hold for the packet delay; a normalization between the two measures can be obtained using a Palm argument (see Shakkottai and Srikant [47]). There is also a deeper difference in that continuous and discrete-time models (e.g., Markov On-Off processes/chains) can lend themselves to qualitatively different bounds (see the exponential decay with prefactor less than 1 from (13); the same holds in the case of an On-Off chain but under a specific burstiness condition on the transition probabilities, see Buffet and Duffield [8], which is the same as the embeddability condition of Markov chains in Markov processes, see Poloczek and Ciucu [41]).

The CP-bounds (reproduced from [15]) are almost identical to those from Palmowski and Rolski [40] in the case of the continuous-time Markovian fluid; only the MMOO model was considered in § 5.1 due to its expressiveness. As in Theorem 7, [40] exclusively works in continuous-time using a continuous-time martingale from Ethier and Kurtz [21]. Unlike the MMOO case, the general case from [40] appears to miss the fundamental improvement of the bounds related to the property of the instantaneous increment at the stopping time; this likely overlook was rectified by Ciucu *et al.* [16].

6.2 Extensions

The results in this paper assume a constant-rate service rate; even the GI/G/1 queue was treated by constructing a compound arrival process to be served at rate one. The underlying principle behind this approach is to encode *all* the information about arrivals, including the service times of packets in the GI/G/1 case, in a single model, i.e., the martingale representation; this model is referred to in Poloczek and Ciucu [42] as an *arrival-martingale*.

A fundamental motivation of this approach, which essentially follows from the network calculus principles (see Chang [9], Le Boudec and Thiran [7], and Jiang and Liu [28]), is to decouple arrivals from service. One key benefit is the straightforward extension to random service rates, by encoding *all* the information about service in a *service-martingale* [42] (defined therein for some (discrete-time) Markov-modulated processes modelling specific wireless channels). In our context, we can represent service in terms of a MAP $(S(t), L_t)_t$ and slightly change Lemmas 5, 6 to construct service-martingales in the homogeneous or inhomogeneous cases. The main difference is a sign-change in the exponential of the martingale, i.e.,

$$h(L_t)e^{-\theta(S(t)-Ct)}\,.$$

(a service-martingale essentially extends an arrival-martingale in the same way that effective-capacity (Wu and Negi [51]) extends effective bandwidth).

Given an arrival-martingale $h_a(M_t)e^{\theta_a(A(t)-C_a t)}$ and a service-martingale $h_s(L_t)e^{-\theta_s(A(t)-C_s t)}$, the bounds from Theorem 7 extend easily. C_a and C_b should be selected such that $\theta_a = \theta_s =: \theta$, using the algorithm from § 4.3; existence is again guaranteed from stability. A backlog upper bound is then

$$\mathbb{P}(Q \geq \sigma) \leq \frac{\mathbb{E}[h_a(M_0)]\mathbb{E}[h_s(L_0)]}{\max_{(m,l)\in D} h_a(m)h_s(l)} e^{-\theta\sigma}\,, \tag{18}$$

where $D = \{(m,l) \mid \exists t : M_t = m \wedge L_t = l \wedge a(t) \geq s(t)\}$ ($s(t)$ is the instantaneous service, i.e., $S(t) = \int_0^t s(u)du$).

Another key benefit of the decoupling principle is that scheduling can be encoded in the service-martingale itself, and the bound from (18) would still hold; such service-martingales have been *implicitly* used in Ciucu *et al.* [15] for several scheduling algorithms. The aggregate models in this paper are implicitly restricted to FIFO scheduling.

7 CONCLUSIONS

We have proposed a novel method to construct martingale representations from MAPs by solving for ODEs. Besides its elegance, the key benefit of the proposed method is covering the case when the background Markov process has an uncountable state-space and can be inhomogeneous. The obtained MAP martingales, in continuous time, enabled the analysis of the multiclass ΣGI/G/1 queues in terms of closed-form and almost explicit bounds, alike the classical Kingman's bounds for GI/G/1 queues. The key idea is that fully working in continuous-time circumvents the non-renewal/non-stationary technical issue characteristic to ΣGI/G/1. Using the same method, we have also also derived bounds in queueing systems with a broad range of Markovian arrival processes, including a novel Batch Markovian Arrival Process with continuous batch sizes. What it noteworthy is that the computational complexity is linear (in the number of multiplexed arrivals), whereas all the derived bounds are almost exact in heavy-traffic according to simulations.

REFERENCES

[1] Nail Akar and Khosrow Sohraby. 2004. Infinite- and Finite-Buffer Markov Fluid Queues: A Unified Analysis. *Journal of Applied Probability* 41, 2 (2004), 557–569.

[2] David Anick, Debasis Mitra, and Man M. Sondhi. 1982. Stochastic Theory of a Data-Handling System with Multiple Sources. *Bell Systems Technical Journal* 61, 8 (Oct. 1982), 1871–1894.

[3] Søren Asmussen. 1995. Stationary Distributions via First Passage Times. In *Advances in Queueing Theory: Theory, Methods and Open Problems (Editor J. H. Dshalalow)*. CRC Press, Inc., 79–102.

[4] Søren Asmussen. 2003. *Applied Probability and Queues*. Springer.

[5] Søren Asmussen and Tomasz Rolski. 1994. Risk Theory in a Periodic Environment: The Cramér-Lundberg Approximation and Lundberg's Inequality. *Mathematics of Operations Research* 19, 2 (May 1994), 410–433.

[6] Demitris Bertsimas and Georgia Mourtzinou. 1997. Multiclass Queueing Systems in Heavy Traffic: An Asymptotic Approach Based on Distributional and Conservation Laws. *Operations Research* 45, 3 (June 1997), 470–487.

[7] Jean-Yves Le Boudec and Patrick Thiran. 2001. *Network Calculus*. Springer Verlag, Lecture Notes in Computer Science, LNCS 2050.

[8] Emmanuel Buffet and Nick G. Duffield. 1994. Exponential Upper Bounds via Martingales for Multiplexers with Markovian Arrivals. *Journal of Applied Probability* 31, 4 (Dec. 1994), 1049–1060.

[9] Cheng-Shang Chang. 2000. *Performance Guarantees in Communication Networks*. Springer Verlag.

[10] Cheng-Shang Chang and Jay Cheng. 1995. Computable Exponential Bounds for Intree Networks with Routing. In *Proceedings of IEEE Infocom*. 197–204.

[11] Joseph T. Chang. 1994. Inequalities for the overshoot. *The Annals of Applied Probability* 4, 4 (Nov. 1994), 1223–1233.

[12] Julian Cheng, Chintha Tellambura, and Norman C. Beaulieu. 2004. Performance of Digital Linear Modulations on Weibull Slow-Fading Channels. *IEEE Transactions on Communications* 52, 8 (Aug 2004), 1265–1268.

[13] Gagan L. Choudhury, David M. Lucantoni, and Ward Whitt. 1996. Squeezing the Most out of ATM. *IEEE Transactions on Communications* 44, 2 (Feb. 1996), 203–217.

[14] Erhan Çinlar. 1972. Markov Additive Processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 24, 2 (1972), 85–93.

[15] Florin Ciucu, Felix Poloczek, and Jens Schmitt. 2014. Sharp Per-Flow Delay Bounds for Bursty Arrivals: The Case of FIFO, SP, and EDF Scheduling. In *IEEE Infocom*. 1896–1904.

[16] Florin Ciucu, Felix Poloczek, and Jens Schmitt. 2016. Stochastic Upper and Lower Bounds for General Markov Fluids. In *International Teletraffic Congress (ITC)*.

[17] Jim G. Dai and Thomas G. Kurtz. 1995. A multiclass Station with Markovian Feedback in Heavy Traffic. *Mathematics of Operations Research* 20, 3 (Aug. 1995), 721–742.

[18] Nick G. Duffield. 1994. Exponential Bounds for Queues with Markovian Arrivals. *Queueing Systems* 17, 3-4 (Sept. 1994), 413–430.

[19] Anwar I. Elwalid and Debasis Mitra. 1993. Effective bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks. *IEEE/ACM Transactions on Networking* 1, 3 (June 1993), 329–343.

[20] Anwar I. Elwalid, Debasis Mitra, and Thomas E. Stern. 1991. Statistical Multiplexing of Markov-Modulated Sources: Theory and Computational Algorithms. In *International Teletraffic Congress (ITC)*.

[21] Stewart N. Ethier and Thomas G. Kurtz. 1986. *Markov processes – Characterization and Convergence*. John Wiley & Sons Inc.

[22] Youjian Fang, Michael Devetsikiotis, Ioannis Lambadaris, and A. Roger Kaye. 1995. Exponential Bounds for the Waiting Time Distribution in Markovian Queues, with Applications to TES/GI/1 Systems. In *Proceedings of the 1995 ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '95/PERFORMANCE '95)*. 108–115.

[23] Wolfgang Fischer and Kathleen Meier-Hellstern. 1993. The Markov-Modulated Poisson Process (MMPP) Cookbook. *Performance Evaluation* 18, 2 (1993), 149 – 171.

[24] Qi-Ming He and Attahiru A. Alfa. 2018. Space Reduction for a Class of Multidimensional Markov Chains: A Summary and Some Applications. *INFORMS Journal on Computing* 30, 1 (2018), 1–10.

[25] Harry Heffes and David M. Lucantoni. 1986. A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. *IEEE Journal on Selected Areas in Communications* 4, 6 (Sept. 1986), 856–867.

[26] Roger Horn and Charles R. Johnson. 1991. *Topics in Matrix Analysis*. Cambridge University Press.

[27] Donald L. Iglehart and Ward Whitt. 1970. Multiple Channel Queues in Heavy Traffic. I. *Advances in Applied Probability* 2, 1 (Spring 1970), 150–177.

[28] Yuming Jiang and Yong Liu. 2008. *Stochastic Network Calculus*. Springer.

[29] Yuming Jiang and Vishal Misra. 2017. Delay Bounds for Multiclass FIFO. *CoRR* abs/1605.05753v2 (2017). https://arxiv.org/abs/1605.05753v2

[30] John F. C. Kingman. 1964. A Martingale Inequality in the Theory of Queues. *Cambridge Philosophical Society* 60, 2 (April 1964), 359–361.

[31] John F. C. Kingman. 1970. Inequalities in the Theory of Queues. *Journal of the Royal Statistical Society, Series B* 32, 1 (1970), 102–110.
[32] Leendert Kosten. 1974. *Stochastic Theory of a Multi-Entry Buffer I.* Delft Progress Report, Series F. University of Delft. 10–18 pages.
[33] Vidyadhar G. Kulkarni. 1997. Fluid Models for Single Buffer Systems. In *Frontiers in Queueing: Models and Applications in Science and Engineering. (Editor J. H. Dshalalow).* CRC Press, Inc., 321–338.
[34] Chengzhi Li, Almut Burchard, and Jörg Liebeherr. 2007. A Network Calculus with Effective Bandwidth. *IEEE/ACM Transactions on Networking* 15, 6 (Dec. 2007), 1442–1453.
[35] Zhen Liu, Philippe Nain, and Don Towsley. 1994. On a Generalization of Kingman's Bounds. Technical Report No. 2423, INRIA.
[36] Zhen Liu, Philippe Nain, and Don Towsley. 1997. Exponential Bounds with Applications to Call Admission. *J. ACM* 44, 3 (May 1997), 366–394.
[37] David M. Lucantoni. 1993. The BMAP/G/1 Queue: A Tutorial. In *Performance Evaluation of Computer and Communication Systems*, Lorenzo Donatiello and Randolph Nelson (Eds.). Springer, 330–358.
[38] Edward W. Ng and Murray Geller. 1969. A Table of Integrals of the Error Functions. *J. Res. Natl. Bur. Stand., Sec. B: Math. Sci.* 73B, 1 (Jan.-Mar. 1969), 1–20.
[39] António Pacheco and Narahari U. Prabhu. 1995. Markov-Additive Processes of Arrivals. In *Advances in Queueing Theory: Theory, Methods and Open Problems (Editor J. H. Dshalalow).* CRC Press, Inc., 167–194.
[40] Zbigniew Palmowski and Tomasz Rolski. 1996. A Note on Martingale Inequalities for Fluid Models. *Statistics & Probability Letters* 31, 1 (Dec. 1996), 13–21.
[41] Felix Poloczek and Florin Ciucu. 2014. Scheduling Analysis with Martingales. *Performance Evaluation (Special Issue: IFIP Performance 2014)* 79 (Sept. 2014), 56 – 72.
[42] Felix Poloczek and Florin Ciucu. 2015. Service-Martingales: Theory and Applications to the Delay Analysis of Random Access Protocols. In *IEEE Infocom.* 945–953.
[43] Vaidyanathan Ramaswami. 1985. Independent Markov Processes in Parallel. *Communications in Statistics. Stochastic Models* 1, 3 (1985), 419–432.
[44] Philippe Robert. 2003. *Stochastic Networks and Queues.* Springer.
[45] Leonard C. G. Rogers. 1994. Fluid Models in Queueing Theory and Wiener-Hopf Factorization of Markov Chains. *The Annals of Applied Probability* 4, 2 (May 1994), 390–413.
[46] Sheldon M. Ross. 1974. Bounds on the Delay Distribution in GI/G/1 queues. *Journal of Applied Probability* 11, 2 (June 1974), 417–421.
[47] Sanjay Shakkottai and Rayadurgam Srikant. 2000. Delay Asymptotics for a Priority Queueing System. In *ACM Sigmetrics.* 188–195.
[48] Ness B. Shroff and Mischa Schwartz. 1998. Improved Loss Calculations at an ATM Multiplexer. *IEEE/ACM Transactions on Networking* 6, 4 (Aug. 1998), 411–421.
[49] Stanislaw J. Szarek and Elisabeth Werner. 1999. A Nonsymmetric Correlation Inequality for Gaussian Measure. *Journal of Multivariate Analysis* 68, 2 (Feb 1999), 193–211.
[50] David Williams. 1991. *Probability with Martingales.* Cambridge University Press.
[51] Dapeng Wu and Rohit Negi. 2003. Effective Capacity: A Wireless Link Model for Support of Quality of Service. *IEEE Transactions on Wireless Communication* 2, 4 (July 2003), 630–643.

A PROOFS

Proof of Lemma 3. We only give the proof for the upper bound; the other is almost identical. Let us expand

$$\mathbb{E}\left[e^{\theta(U_1+U_2+\cdots+U_T)}1_{T\le n}\right]=\sum_{k=1}^{n}\mathbb{E}\left[e^{\theta(U_1+U_2+\cdots+U_k)}1_{T=k}\right]$$

and denote by $f(x)$ the density of U_1. We can write for each term

$$\begin{aligned}
&\mathbb{E}\left[e^{\theta(U_1+U_2+\cdots+U_k)}1_{T=k}\right]\\
&=\int_{-\infty}^{\sigma}e^{\theta x_1}f(x_1)\int_{-\infty}^{\sigma-x_1}e^{\theta x_2}f(x_2)\cdots\int_{-\infty}^{\sigma-L_{k-2}}e^{\theta x_{k-1}}f(x_{k-1})\\
&\qquad E\left[e^{\theta U_k}1_{U_k\ge\sigma-L_{k-1}}\right]dx_{k-1}\ldots dx_1\\
&=\int_{-\infty}^{\sigma}e^{\theta x_1}f(x_1)\int_{-\infty}^{\sigma-x_1}e^{\theta x_2}f(x_2)\cdots\int_{-\infty}^{\sigma-L_{k-2}}e^{\theta x_{k-1}}f(x_{k-1})\\
&\qquad K(\sigma-L_{k-1})e^{\theta(\sigma-L_{k-1})}\mathbb{P}\left(U_k\ge\sigma-L_{k-1}\right)dx_{k-1}\ldots dx_1\\
&\ge\inf_{y\ge0}K(y)e^{\theta\sigma}\mathbb{P}(T=k)\,.
\end{aligned}$$

Here we denoted $L_{k-1}:=x_1+\cdots+x_{k-1}$. Therefore,

$$\mathbb{E}\left[e^{\theta(U_1+U_2+\cdots+U_T)}1_{T\le n}\right]\ge\inf_{y\ge0}K(y)e^{\theta\sigma}\sum_{k}\mathbb{P}(T=k)$$

and the rest is identical as in the proof of the Kingman's bound.

As a side remark, the Ross bound from (6) can be immediately generalized to the case when $(U_n)_n$ is a homogeneous Markov chain. Indeed, the same bound from (6) would hold but with $K(y)$ replaced by

$$K(y,z)=\mathbb{E}\left[e^{\theta(U_2-y)}\mid U_2\ge y,U_1=z\right]\,.$$

(additionally, the 'inf' and 'sup' must be also taken after z, i.e, the state-space of U_n). This bound can be leveraged to improve existing bounds in queues with Markov modulated arrivals in discrete-time models (e.g., [18]); such bounds would have an additional factor in (7), due to the use of a different martingale for Markov modulated arrivals. □

Proof of Theorem 7. The proof is similar to the one for Kingman's bound from § 2.1.1; what is different is the continuous-time model and also the additional prefactor in the exponential martingale.

The stationary backlog distribution Q has the representation

$$Q:=\sup_{t\ge0}\{A(t)-Ct\}\,.$$

Define the stopping time T by

$$T:=\inf\{t\ge0\mid A(t)-Ct\ge\sigma\}\,,$$

and note that $\{Q\ge\sigma\}=\{T<\infty\}$. Now for $n\in\mathbb{N}$, by the optional stopping theorem:

$$\begin{aligned}
\mathbb{E}[h(M_0)]&=\mathbb{E}[X_0]=\mathbb{E}[X_{T\wedge n}]\\
&\ge\mathbb{E}[h(M_T)e^{\theta(A(T)-CT)}1_{T\le n}]\\
&\ge e^{\theta\sigma}\inf_{m\in\mathrm{Im}(M)}h(m)\mathbb{P}(T\le n)\,.
\end{aligned}\tag{19}$$

Recalling the definition of 'Im(M)', we remark that $a(T)\ge C$ from the definition of T. The upper bound on $\mathbb{P}(Q\ge\sigma)$ follows immediately by taking the limit $n\to\infty$.

For the lower bound, the standard proof from [46] is to first define an additional stopping time and then invoke a more elaborate application of the optional stopping theorem. We next give a more direct proof using Lemma 2. In the limit $n\to\infty$ the first inequality in (19) holds as an equality, and thus

$$\begin{aligned}
\mathbb{E}[h(M_0)]&=\mathbb{E}[X_0]=\mathbb{E}[X_{T\wedge n}]\\
&=\lim_n\mathbb{E}[h(M_T)e^{\theta(A(T)-CT)}1_{T\le n}]\\
&\le\lim_n e^{\theta(\sigma+\xi)}\sup_{m\in\mathrm{Im}(M)}h(m)\mathbb{P}(T\le n)\,,
\end{aligned}$$

which completes the proof. (Note that just before T it holds $A(T-)-(T-)C<\sigma$ and hence $A(T)-CT<\sigma+\xi$.) □

Proof of Corollary 9. We focus on class 1 and consider the Palm conditional space that one of its jobs arrives at time 0. For $t\ge0$ let $R_i(t)$ be the remaining lifetimes for each class i, i.e., the time it takes from $-t$ to the next arrival; note that, in particular, $R_1(0)=0$.

Let the compound process

$$A_i(t) := \sum_{j=1}^{N_i(t)} S_{i,-j} \,,$$

and note that the waiting time W of the job of class 1 arriving at time 0 is *bounded*, in distribution, by[4]

$$\begin{aligned}\mathbb{P}(W \geq \sigma) &\leq \mathbb{P}\left(\sup_{t\geq 0}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N_i(t)} S_{i,-j} - t\right\} \geq \sigma\right) \\ &= \mathbb{P}\left(\sup_{t\geq 0}\left\{\sum_{i=1}^{N}\left(A_i(t) - \frac{t}{N}\right)\right\} \geq \sigma\right) . \qquad (20)\end{aligned}$$

To use the multiplexing property from § 3.2, we consider a single class system but keep the utilization ρ (e.g., the service times of $A_1(t)$ are scaled by N). Let $L := E\left[e^{\theta N S_{1,1}}\right]$. Since $\lambda(t) = 2t$ for the Weibull distribution, the ODE from Lemma 8 becomes

$$h'(t) - h(t)(2t+\theta) = -2th(0)L \,.$$

Choosing the initial condition $h(0) = 1$ yields the unique solution

$$h(t) = \frac{1 - 2L\int_0^t se^{-(s^2+\theta s)}ds}{e^{-(t^2+\theta t)}}$$

and consequently the martingale process

$$h(R_1(t))e^{\theta(NA_1(t)-t)} \,.$$

Repeating the argument for all classes $A_i(t)$ we obtain the product martingale

$$\prod_{i=1}^{N} h(R_i(t))e^{\theta N\sum_{i=1}^{N}\left(A_i(t)-\frac{t}{N}\right)} \,.$$

is a martingale. Recalling the expression from (20) and applying Theorem 7 yields

$$\mathbb{P}(W \geq \sigma) \leq \frac{\prod_i E\left[h(R_i(0))\right]}{\prod_i \inf_{t\geq 0} h(t)} e^{-\theta N\sigma} \,.$$

To complete the proof we will first prove that $E\left[h(R_i(0))\right] = K(\theta)$ for $i \geq 2$ (note that $E\left[h(R_1(0))\right] = E\left[h(0)\right] = 1$) and second that $\inf_{t\geq 0} h(t) = 1$.

Fix $t \geq 0$. Given that the density of $R(0)$ (we drop the index i) is $\frac{2}{\sqrt{\pi}}e^{-t^2}$ we have

$$E\left[h(R(0))\right] = \int_0^\infty \frac{1 - 2E\left[e^{\theta y}\right]\int_0^t se^{-(s^2+\theta s)}ds}{\frac{\sqrt{\pi}}{2}e^{-\theta t}}dt$$

The inner integral can be rewritten as

$$\int_0^t se^{-(s^2+\theta s)}ds = \int_0^t e^{\frac{\theta^2}{4}}se^{-\left(s+\frac{\theta}{2}\right)^2}ds$$

and by the change of variable $s + \frac{\theta}{2} = u$ it becomes

$$\begin{aligned}&\int_{\frac{\theta}{2}}^{t+\frac{\theta}{2}} e^{\frac{\theta^2}{4}}ue^{-u^2}du - \frac{\theta}{2}e^{\frac{\theta^2}{4}}\int_{\frac{\theta}{2}}^{t+\frac{\theta}{2}} e^{-u^2}du \\ &= \frac{1}{2}\left(1 - e^{-(t^2+\theta t)}\right) - \frac{\theta}{2}\frac{\sqrt{\pi}}{2}e^{\frac{\theta^2}{4}}\left(erf\left(t+\frac{\theta}{2}\right) - erf\left(\frac{\theta}{2}\right)\right) .\end{aligned}$$

[4] W is generally not a *stationary* waiting time, alike in the GI/G/1 case (see (10)), due to the general lack of stationarity; W should be regarded as *transient* delay.

By rearranging terms $E\left[h(R(0))\right]$ is

$$\int_0^\infty \frac{1 - L\left(1 - e^{-(t^2+\theta t)} + \theta\frac{\sqrt{\pi}}{2}e^{\frac{\theta^2}{4}}\left(1 - erf\left(t+\frac{\theta}{2}\right)\right)\right.}{\left.-\theta\frac{\sqrt{\pi}}{2}e^{\frac{\theta^2}{4}}\left(1 - erf\left(\frac{\theta}{2}\right)\right)\right)}{\frac{\sqrt{\pi}}{2}e^{-\theta t}}dt$$

Using the identity [12]

$$\mathbb{E}\left[e^{-\theta T_1}\right] = 1 - \theta e^{\frac{\theta^2}{4}}\frac{\sqrt{\pi}}{2}\left(1 - erf\left(\frac{\theta}{2}\right)\right) \qquad (21)$$

and the definition of θ the integral simplifies to

$$\begin{aligned}&\frac{2}{\sqrt{\pi}}L\int_0^\infty \frac{e^{-(t^2+\theta t)} - \theta\frac{\sqrt{\pi}}{2}e^{\frac{\theta^2}{4}}\left(1 - erf\left(t+\frac{\theta}{2}\right)\right)}{e^{-\theta t}}dt \\ &= \frac{2}{\sqrt{\pi}}L\left(\int_0^\infty e^{-t^2}dt - \int_0^\infty \theta\frac{\sqrt{\pi}}{2}e^{\frac{\theta^2}{4}}e^{\theta t}erfc\left(t+\frac{\theta}{2}\right)\right)dt \\ &= L - \theta Le^{\frac{\theta^2}{4}}\int_0^\infty e^{\theta t}erfc\left(t+\frac{\theta}{2}\right)dt. \qquad (22)\end{aligned}$$

By a change of variable $t + \frac{\theta}{2} = s$ the integral becomes

$$\begin{aligned}e^{-\frac{\theta^2}{2}}\int_{\frac{\theta}{2}}^\infty e^{\theta s}erfc(s)ds &= e^{-\frac{\theta^2}{2}}\left(-\frac{1}{\theta}e^{\frac{\theta^2}{2}}erfc\left(\frac{\theta}{2}\right) + \frac{1}{\theta}e^{\frac{\theta^2}{4}}\right) \\ &= -\frac{1}{\theta}erfc\left(\frac{\theta}{2}\right) + \frac{1}{\theta}e^{-\frac{\theta^2}{4}} ,\end{aligned}$$

after using in the first line the identity [38]

$$\int e^{\theta s}erfc(s)ds = \frac{1}{\theta}e^{\theta s}erfc(s) + \frac{1}{\theta}e^{\frac{\theta^2}{4}}erf\left(s - \frac{\theta}{2}\right) .$$

We can now complete the derivation of Eq. (22) as

$$L - \theta Le^{\frac{\theta^2}{4}}\left(-\frac{1}{\theta}erfc\left(\frac{\theta}{2}\right) + \frac{1}{\theta}e^{-\frac{\theta^2}{4}}\right) = Le^{\frac{\theta^2}{4}}erfc\left(\frac{\theta}{2}\right) = K(\theta) .$$

Lastly, to prove $\inf_{t\geq 0} h(t) = 1$, we follow the equations above and rewrite

$$h(t) = L\left(1 - \frac{\theta\frac{\sqrt{\pi}}{2}\left(1 - erf\left(t+\frac{\theta}{2}\right)\right)}{e^{-\left(t+\frac{\theta}{2}\right)^2}}\right) .$$

The proof is complete from $h(0) = 1$ and the monotonicity of $\frac{1-erf(x)}{e^{-x^2}}$ [49]. □

Proof of Corollary 10. The proof is similar to that for the Weibull case. Differently, we compute the numerator in the expression of $h(t)$ from Corollary 8

$$1 - E\left[e^{\theta S_1}\right]\int_0^t e^{-\theta s}f(s)ds = e^{-(\lambda+\theta)t}\sum_{l=0}^{k-1}\frac{(t(\lambda+\theta))^l}{l!} ,$$

after elementary integrations involving the Erlang-k density $f(t) = \frac{\lambda^k t^{k-1}e^{-\lambda t}}{(k-1)!}$. Since the density of $R(0)$ (the remaining lifetime) is $\frac{1-F(t)}{E\left[T_{1,1}\right]}$ we obtain that

$$E[h(R(0))] = \frac{\lambda}{k}\int_0^\infty e^{-\lambda t}\left(\sum_{l=0}^{k-1}\frac{(t(\lambda+\theta))^l}{l!}\right)dt = \frac{1}{k}\sum_{l=0}^{k}\left(1 + \frac{\theta}{\lambda}\right)^l .$$

The proof is complete after rearranging terms and noting that $\inf_{t\geq 0} h(t) = 1$ ($h(0) = 1$ and $h(t)$ is non-decreasing). □

Proof of Corollary 12. A direct proof follows from Lemma 5. We present however a much more concise proof by using the multiplexing property from § 3.2. Indeed, let $A_i(t)$ and $M_{i,t}$ be the arrival and background processes, respectively, of the individual MMOO processes. According to Corollary 11 the processes

$$h_i(M_{i,t})e^{\theta\left(A_i(t)-\frac{C}{N}t\right)}$$

are martingales, where $h_i = h$ for $i = 1, 2, \ldots, N$ and θ is obtained similarly but with C replaced by $\frac{C}{N}$. The proof is complete by letting

$$h(M_t) := h(\sum_i M_{i,t}) := \prod_i h(M_{i,t}) .$$

As a side remark, the 'split' mentioned in § 3.2 is uniform (i.e., the capacity C is equally split) since $A_i(t)$'s are themselves uniform. Should that not be the case, then one would have to search for a split guaranteeing the same 'θ' as in § 4.3; recall the remark that constructing martingales with different θ's and then normalizing them as in [41] can be prone to numerical inaccuracies (due to the use of Jensen's inequality). □

Proof of Corollary 13. We again apply Lemma 5.

Assume $M_0 = 1$. In a small interval $[0, \Delta s]$, three 'events' can happen:

(1) M stays at state 1 and A transmits:

$$\mathbb{P} \approx (1 - \mu_1\Delta s)\lambda_1\Delta s ;$$

(2) M stays at 1 and A does not transmit:

$$\mathbb{P} \approx (1 - \mu_1\Delta s)(1 - \lambda_1\Delta s) ;$$

(3) M jumps to state 2 and A does not transmit:

$$\mathbb{P} \approx \mu_1\Delta s(1 - \lambda_1\Delta s) .$$

Note that, due to the independence assumption, the probability of the fourth event, i.e., *both* a jump $1 \to 2$ of M *and* a transmission of A, is of order $o(\Delta s)$, and can be ignored.

Therefore

$$\begin{aligned}\varphi_1(\Delta s) =&\mathbb{E}\left[h(M_{\Delta s})e^{\theta(A(\Delta s)-C\Delta s)} \mid M_0 = 1\right]\\ =&(1 - \mu_1\Delta s)\,\lambda_1\Delta s\,h_1\,e^{\theta(1-C\Delta s)}\\ &+ (1 - \mu_1\Delta s)\,(1 - \lambda_1\Delta s)\,h_1\,e^{-\theta C\Delta s}\\ &+ \mu_1\Delta s\,(1 - \lambda_1\Delta s)\,h_2\,e^{-\theta C\Delta s} + o(\Delta s) ,\end{aligned}$$

which simplifies to

$$\begin{aligned}&h_1\,e^{-\theta C\Delta s} + \Delta s\,h_1(\lambda_1\,e^{\theta} - \mu_1 - \lambda_1)e^{-\theta C\Delta s}\\ &\quad + \Delta s\,h_2\,\mu_1 e^{-\theta C\Delta s} + o(\Delta s) .\end{aligned}$$

Accounting for $\varphi_1(0) = h_1$ we have

$$\lim_{\Delta s\to 0}\frac{1}{\Delta s}\left(h_1\,e^{-\theta C\Delta s} - h_1\right) = -h_1\theta C = -\lambda(\theta)h_1 ,$$

so that finally

$$\left.\frac{d}{ds}\varphi_1(s)\right|_{s=0} = h_1(\lambda_1\,e^{\theta} - \mu_1 - \lambda_1) + h_2\,\mu_1 - \lambda(\theta)h_1 . \tag{23}$$

Analogously, one obtains

$$\left.\frac{d}{ds}\varphi_2(s)\right|_{s=0} = h_2(\lambda_2\,e^{\theta} - \mu_2 - \lambda_2) + h_1\,\mu_2 - \lambda(\theta)h_2 . \tag{24}$$

Both final terms in (23) and (24) vanish if and only if

$$T_\theta\begin{pmatrix}h_1\\ h_2\end{pmatrix} = \lambda(\theta)\begin{pmatrix}h_1\\ h_2\end{pmatrix} ,$$

which is true by assumption. □

Proof of Corollary 14. We again apply Lemma 5. Assume $M_0 = \lambda_1$ and $L_0 = 1$. In a small interval $(0, \Delta s)$, four 'events' can happen:

(1) M stays at state 1, A^{rnd} transmits, S stays:

$$\mathbb{P} = (1 - \mu_1\Delta s)\lambda_1(1 - p)\Delta s ;$$

(2) M stays at 1, A^{rnd} transmits, S jumps:

$$\mathbb{P} = (1 - \mu_1\Delta s)\lambda_1 p\Delta s ;$$

(3) M stays at 1, and A^{rnd} does not transmit:

$$\mathbb{P} = (1 - \mu_1\Delta s)(1 - \lambda_1\Delta s) ;$$

(4) M jumps to state 2, and A^{rnd} does not transmit:

$$\mathbb{P} = \mu_1\Delta s(1 - \lambda_1\Delta s) .$$

Then,

$$\begin{aligned}\varphi_{(1,1)}(\Delta s) =&\mathbb{E}\left[h(M_{\Delta s})e^{\theta(A^{\text{rnd}}(\Delta s)-C\Delta s)} \mid M_0 = 1,\ L_0 = 1\right]\\ =&(1 - \mu_1\Delta s)\,\lambda_1\,(1 - p)\,\Delta s\,h_{1,1}\,e^{\theta(S_{1,1}-C\Delta s)}\\ &+ (1 - \mu_1\Delta s)\,\lambda_1\,p\,\Delta s\,h_{1,2}\,e^{\theta(S_{2,1}-C\Delta s)}\\ &+ (1 - \mu_1\Delta s)\,(1 - \lambda_1\Delta s)\,h_{1,1}\,e^{-\theta C\Delta s}\\ &+ \mu_1\Delta s\,(1 - \lambda_1\Delta s)\,h_{2,1}\,e^{-\theta C\Delta s} + o(\Delta s) ,\end{aligned}$$

Similarly as in the proof of Example 13 one obtains:

$$\begin{aligned}\left.\frac{d}{ds}\varphi_{(1,1)}(s)\right|_{s=0} =&\left((1 - p)\lambda_1\mathbb{E}[e^{\theta S_{1,1}}] - \mu_1 - \lambda_1 - \theta C\right)h_{1,1}\\ &+ p\lambda_1\mathbb{E}[e^{\theta S_{2,1}}]h_{1,2} + \mu_1 h_{2,1} .\end{aligned}$$

Analogously, one obtains

$$\begin{aligned}\left.\frac{d}{ds}\varphi_{(1,2)}(s)\right|_{t\,s=0} =&q\lambda_1\mathbb{E}[e^{\theta S_{1,1}}]h_{1,1} + \mu_1 h_{2,2}\\ &+ \left((1 - q)\,\lambda_1\mathbb{E}[e^{\theta S_{2,1}}] - \mu_1 - \lambda_1 - \theta C\right)h_{1,2} ,\end{aligned}$$

$$\begin{aligned}\left.\frac{d}{ds}\varphi_{(2,1)}(s)\right|_{s=0} =&\left((1 - p)\lambda_2\mathbb{E}[e^{\theta S_{1,1}}] - \mu_2 - \lambda_2 - \theta C\right)h_{2,1}\\ &+ p\lambda_2\mathbb{E}[e^{\theta S_{2,1}}]h_{2,2} + \mu_2 h_{1,1} ,\end{aligned}$$

and

$$\begin{aligned}\left.\frac{d}{ds}\varphi_{(2,2)}(s)\right|_{s=0} =&q\lambda_2\mathbb{E}[e^{\theta S_{1,1}}]h_{2,1} + \mu_2 h_{1,2}\\ &+ \left((1 - q)\,\lambda_2\mathbb{E}[e^{\theta S_{2,1}}] - \mu_2 - \lambda_2 - \theta C\right)h_{2,2} .\end{aligned}$$

By the choice of θ, all four terms vanish. □

Proof of Corollary 16. Apply Lemma 5. For an arbitrary state i it holds:

$$\begin{aligned}\varphi_i(\Delta s) :=& \mathbb{E}\left[h(M_{\Delta s})e^{\theta(A(\Delta s)-C\Delta s)} \,\middle|\, M_0 = i\right] \\ =& \sum_{j\neq i} d_{i,j}\Delta s h(j) e^{-\theta C\Delta s} + \sum_j d'_{i,j}\Delta s h(j) e^{\theta(1-C\Delta s)} \\ &+ \left(1 + d_{i,i}\Delta s\right) h(i) e^{-\theta C\Delta s} + o(\Delta s)\,,\end{aligned}$$

such that

$$\begin{aligned}\left.\frac{d}{dt}\varphi_i(s)\right|_{s=0} &= \lim_{\Delta s\to 0} \left(\varphi_i(\Delta s) - h(i)\right)/\Delta s \\ &= \sum_j \left(d_{i,j} + e^{\theta} d'_{i,j}\right) h(j) - \theta C h(i) \\ &= \left(\left(D_0 + e^{\theta} D_1\right) h\right)_i - (\lambda(\theta) h)_i \ .\end{aligned}$$

By assumption, the last term vanishes, which completes the first part of the proof.

For the reversed process, note first that by Bayes' theorem

$$\begin{aligned}&\mathbb{P}\left(A^r(t, t+\Delta t) = 0, M^r_{t+\Delta t} = j \,\middle|\, M^r_t = i\right) \\ &\quad = D_0(j,i)\frac{\pi_j}{\pi_i}\Delta t + o(\Delta t)\,, \text{ and} \\ &\mathbb{P}\left(A^r(t, t+\Delta t) = 1, M^r_{t+\Delta t} = j \,\middle|\, M^r_t = i\right) \\ &\quad = D_1(j,i)\frac{\pi_j}{\pi_i}\Delta t + o(\Delta t)\,,\end{aligned}$$

such that the reversed MArP process is characterized by the pair (D_0^r, D_1^r):

$$D_0^r = \Pi^{-1} D_0^T \Pi \quad \text{and} \quad D_1^r = \Pi^{-1} D_1^T \Pi\,.$$

Since

$$\begin{aligned}D_0^r + e^{\theta} D_1^r &= \Pi^{-1} D_0^T \Pi + e^{\theta} \Pi^{-1} D_1^T \Pi \\ &= \Pi^{-1}\left(D_0^T + e^{\theta} D_1^T\right)\Pi\,,\end{aligned}$$

the proof follows as in the first part. Note that eigenvalues are preserved under transposition and similarity transformations, i.e., $\lambda(\theta)$ is also the spectral radius of $\Pi^{-1}\left(D_0 + e^{\theta} D_1\right)^T \Pi$. □

Proof of Corollary 17. With '$\otimes$' denoting the Kronecker product and I_n denoting the $n \times n$-unit matrix, we have

$$\begin{aligned}&D_0 \oplus D'_0 + e^{\theta}(D_1 \oplus D'_1) \\ &\quad = \left(D_0 \otimes I_n + I_n \otimes D'_0\right) + e^{\theta}\left(D_1 \otimes I_n + I_n \otimes D'_1\right) \\ &\quad = \left(D_0 + e^{\theta} D_1\right) \otimes I_n + I_n \otimes \left(D'_0 + e^{\theta} D'_1\right) \\ &\quad = \left(D_0 + e^{\theta} D_1\right) \oplus \left(D'_0 + e^{\theta} D'_1\right)\,,\end{aligned}$$

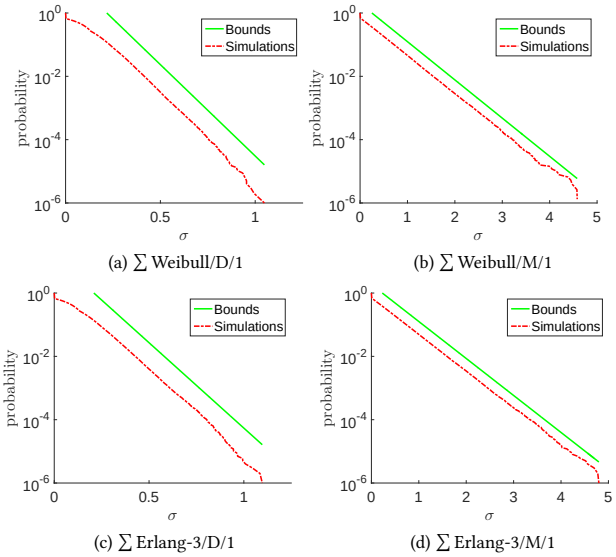
whose spectral radius is $\lambda(\theta) + \lambda'(\theta)$; the corresponding eigenvector is $h' \otimes h$ (see Theorem 4.4.5 in [26]).[5] Denote $M''(t)$ the background Markov process of $A(t)+A'(t)$ (i.e., with generator $D_0 \oplus D'_0 + D_1 \oplus D'_1$) and observe that

$$h' \otimes h(M''_t) = h(M_t) h'(M'_t)\,.$$

The proof is complete by applying Corollary 16. □

[5] We use the definition of the Kronecker sum from [26]; other definitions are available in the literature.

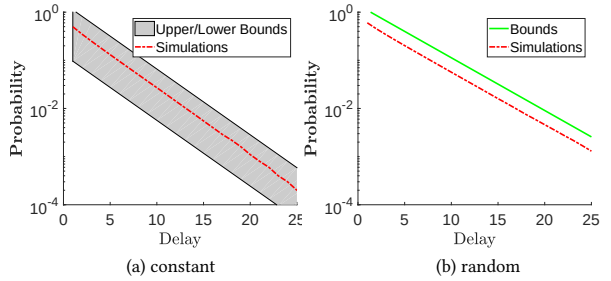
B ADDITIONAL NUMERICAL RESULTS



(a) $\sum$ Weibull/D/1 (b) $\sum$ Weibull/M/1

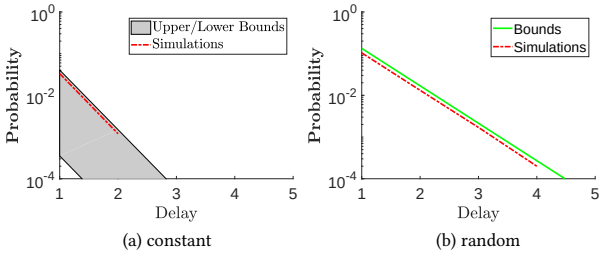
(c) $\sum$ Erlang-3/D/1 (d) $\sum$ Erlang-3/M/1

Figure 9: Waiting-time CCDF (upper bounds vs. simulations); ($N = 5$, $\rho = 0.75$)



(a) constant (b) random

Figure 10: Waiting-time CCDF for N MMPPs; constant and random packet sizes; ($N = 5$, $\mu_1 = 0.1$, $\mu_2 = 0.5$, $\lambda_1 = 1$, $\lambda_2 = 25$, $p = 0.1$, $q = 0.9$, $E[\xi_1] = 0.2$, $\rho = 0.75$)



(a) constant (b) random

Figure 11: Waiting-time CCDF for N GMArPs; constant and random packet sizes; ($N = 5$, $\mu_1 = 0.1$, $\mu_2 = 0.5$, $\lambda_1 = 0.3$, $\lambda_2 = 10$, $\lambda_3 = 0.7$, $\lambda_4 = 15$, $E[X_1] = 1$, $E[X_2] = 3.01$, $\rho = 0.75$)

σ	0	50	100	150	200
LNT u.b.	1.017	$1.472\ 10^{-2}$	$2.131\ 10^{-4}$	$3.086\ 10^{-6}$	$4.468\ 10^{-8}$
LNT l.b.	0.939	$1.360\ 10^{-2}$	$1.969\ 10^{-4}$	$2.851\ 10^{-6}$	$4.128\ 10^{-8}$
AR u.b.	1.008	$1.459\ 10^{-2}$	$2.113\ 10^{-4}$	$3.059\ 10^{-6}$	$4.429\ 10^{-8}$
AR l.b.	0.898	$1.300\ 10^{-2}$	$1.882\ 10^{-4}$	$2.724\ 10^{-6}$	$3.945\ 10^{-8}$
D u.b.	1.009	$1.058\ 10^{-2}$	$1.110\ 10^{-4}$	$1.164\ 10^{-6}$	$1.220\ 10^{-8}$
CP u.b.	1.008	$1.459\ 10^{-2}$	$2.113\ 10^{-4}$	$3.059\ 10^{-6}$	$4.429\ 10^{-8}$
CP l.b.	0.898	$1.300\ 10^{-2}$	$1.882\ 10^{-4}$	$2.724\ 10^{-6}$	$3.944\ 10^{-8}$

(a) $\rho = 0.95$ ($\lambda_1 = 0.6,\ \lambda_2 = 2,\ \mu_1 = 1,\ \mu_2 = 3$)

σ	0	8	16	24	32
LNT u.b.	1.044	$1.620\ 10^{-2}$	$2.514\ 10^{-4}$	$3.901\ 10^{-6}$	$6.052\ 10^{-8}$
LNT l.b.	0.702	$1.089\ 10^{-2}$	$1.690\ 10^{-4}$	$2.623\ 10^{-6}$	$4.069\ 10^{-8}$
AR u.b.	1.028	$1.594\ 10^{-2}$	$2.474\ 10^{-4}$	$3.838\ 10^{-6}$	$5.956\ 10^{-8}$
AR l.b.	0.550	$0.853\ 10^{-2}$	$1.323\ 10^{-4}$	$2.053\ 10^{-6}$	$3.185\ 10^{-8}$
CP u.b.	1.028	$1.594\ 10^{-2}$	$2.474\ 10^{-4}$	$3.838\ 10^{-6}$	$5.956\ 10^{-8}$
CP l.b.	0.550	$0.853\ 10^{-2}$	$1.323\ 10^{-4}$	$2.053\ 10^{-6}$	$3.185\ 10^{-8}$

(b) $\rho = 0.75$ ($\lambda_1 = 0.6,\ \lambda_2 = 1.2,\ \mu_1 = 1,\ \mu_2 = 3$)

σ	0	3	6	9	12
LNT u.b.	1.184	$1.357\ 10^{-2}$	$1.555\ 10^{-4}$	$1.783\ 10^{-6}$	$2.044\ 10^{-8}$
LNT l.b.	0.341	$0.390\ 10^{-2}$	$0.445\ 10^{-4}$	$0.551\ 10^{-6}$	$0.589\ 10^{-8}$
AR u.b.	1.092	$1.252\ 10^{-2}$	$1.435\ 10^{-4}$	$1.645\ 10^{-6}$	$1.886\ 10^{-8}$
AR l.b.	0.169	$0.193\ 10^{-2}$	$0.222\ 10^{-4}$	$0.254\ 10^{-6}$	$0.291\ 10^{-8}$
D u.b.	1.064	$18.26\ 10^{-2}$	$9.220\ 10^{-4}$	$0.799\ 10^{-6}$	$2.352\ 10^{-8}$
CP u.b.	1.092	$1.252\ 10^{-2}$	$1.435\ 10^{-4}$	$1.645\ 10^{-6}$	$1.886\ 10^{-8}$
CP l.b.	0.169	$0.193\ 10^{-2}$	$0.222\ 10^{-4}$	$0.254\ 10^{-6}$	$0.291\ 10^{-8}$

(c) $\rho = 0.4$ ($\lambda_1 = 0.3,\ \lambda_2 = 0.8,\ \mu_1 = 1,\ \mu_2 = 4$)

Figure 12: Bounds on the waiting-time distribution $\mathbb{P}(W \geq \sigma)$ for the MMPP/D/1 queue (notations from § 5.2.1; average service time is 1)

σ	0	100	200	300	400
LNT u.b.	0.956	$1.003\ 10^{-2}$	$1.052\ 10^{-4}$	$1.103\ 10^{-6}$	$1.157\ 10^{-8}$
LNT l.b.	0.952	$0.999\ 10^{-2}$	$1.047\ 10^{-4}$	$1.099\ 10^{-6}$	$1.152\ 10^{-8}$
AR u.b.	0.958	$1.005\ 10^{-2}$	$1.054\ 10^{-4}$	$1.105\ 10^{-6}$	$1.159\ 10^{-8}$
AR l.b.	0.942	$0.988\ 10^{-2}$	$1.036\ 10^{-4}$	$1.087\ 10^{-6}$	$1.140\ 10^{-8}$
D u.b.	1.009	$1.058\ 10^{-2}$	$1.110\ 10^{-4}$	$1.164\ 10^{-6}$	$1.220\ 10^{-8}$
CP u.b.	1.004	$1.053\ 10^{-2}$	$1.104\ 10^{-4}$	$1.157\ 10^{-6}$	$1.214\ 10^{-8}$

(a) $\rho = 0.95$ ($\lambda_1 = 0.6,\ \lambda_2 = 2,\ \mu_1 = 1,\ \mu_2 = 3$)

σ	0	3	12	48	72
LNT u.b.	0.759	$4.040\ 10^{-2}$	$1.145\ 10^{-4}$	$6.099\ 10^{-6}$	$1.729\ 10^{-8}$
LNT l.b.	0.749	$3.993\ 10^{-2}$	$1.132\ 10^{-4}$	$6.027\ 10^{-6}$	$1.709\ 10^{-8}$
AR u.b.	0.765	$4.073\ 10^{-2}$	$1.155\ 10^{-4}$	$6.148\ 10^{-6}$	$1.743\ 10^{-8}$
AR l.b.	0.728	$3.878\ 10^{-2}$	$1.099\ 10^{-4}$	$5.853\ 10^{-6}$	$1.659\ 10^{-8}$
D u.b.	1.020	$5.431\ 10^{-2}$	$1.540\ 10^{-4}$	$8.197\ 10^{-6}$	$2.323\ 10^{-8}$
CP u.b.	1.012	$5.391\ 10^{-2}$	$1.528\ 10^{-4}$	$8.136\ 10^{-6}$	$2.306\ 10^{-8}$

(b) $\rho = 0.75$ ($\lambda_1 = 0.6,\ \lambda_2 = 1.2,\ \mu_1 = 1,\ \mu_2 = 3$)

σ	0	3	12	24	30
LNT u.b.	0.417	$7.150\ 10^{-2}$	$3.611\ 10^{-4}$	$0.313\ 10^{-6}$	$0.921\ 10^{-8}$
LNT l.b.	0.403	$6.912\ 10^{-2}$	$3.491\ 10^{-4}$	$0.302\ 10^{-6}$	$0.891\ 10^{-8}$
AR u.b.	0.426	$7.302\ 10^{-2}$	$3.688\ 10^{-4}$	$0.319\ 10^{-6}$	$0.941\ 10^{-8}$
AR l.b.	0.367	$6.294\ 10^{-2}$	$3.179\ 10^{-4}$	$0.275\ 10^{-6}$	$0.811\ 10^{-8}$
D u.b.	1.064	$18.26\ 10^{-2}$	$9.220\ 10^{-4}$	$0.799\ 10^{-6}$	$2.352\ 10^{-8}$
CP u.b.	1.032	$17.706\ 10^{-2}$	$8.942\ 10^{-4}$	$0.774\ 10^{-6}$	$2.281\ 10^{-8}$

(c) $\rho = 0.4$ ($\lambda_1 = 0.3,\ \lambda_2 = 0.8,\ \mu_1 = 1,\ \mu_2 = 4$)

Figure 13: Bounds on the waiting-time distribution $\mathbb{P}(W \geq \sigma)$ for the MMPP/M/1 queue (notations from § 5.2.2; average service time is 1)