On a Continuous-Time Martingale and Two Applications

Sima Mehri
University of Warwick

Florin Ciucu
University of Warwick

ABSTRACT

We construct a continuous-time martingale to analyze two queueing systems: One is the $GI^{X}/M/1$ queue with light-tailed batch arrivals for which we obtain an exact result in closed-form when X has a Geometric distribution, and stochastic bounds otherwise. The second application concerns queues with Semi-Markovian (SM) inter-arrival times; of particular interest is the scenario with multiplexed SM sources whose aggregate loses the SM property. Unlike existing exact but implicit solutions which rely on numerical transform methods, even in the $M^{r}/M/1$ case, our stochastic bounds are in closed-form and shown to be (mostly) numerically accurate.

CCS CONCEPTS

• **Mathematics of computing-Markov processes**; • **Networks-Network performance modeling**;

KEYWORDS

Queueing; Semi-Markov Processes; Stochastic Bounds

ACM Reference Format:
Sima Mehri and Florin Ciucu. 2023. On a Continuous-Time Martingale and Two Applications. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Martingale-based methods can address many challenging queues in terms of closed-form stochastic bounds; these are numerically accurate especially in heavy-traffic and significantly improve upon alternative bounds obtained using large-deviations or network calculus approaches. Starting with the $GI/G/1$ analysis by Kingman [16], in terms of exponential bounds, recent studies addressed diverse scenarios with Markovian arrival processes [4, 7, 9, 13, 14, 21], various scheduling algorithms [23], or multiple-access protocols [24].

In this paper we construct a continuous-time martingale which is structurally similar to existing martingales but is formulated in terms of the counting processes corresponding to arrivals and services, as opposed to using compound arrival processes as in [9]. This simple twist enables our main result in queues with batch arrivals, which are a natural abstraction in a variety of scenarios including sensor networks, cloud computing, or even cognitive radio networks [25]. Concretely, we provide an exact result in closed-form for the $GI^{X}/M/1$ queue, whereby the arrivals follow a renewal process but occur in batches/bulks of a (Geometric) random size X[1]. In the more general case, when X has a light-tailed distribution, our results are expressed as stochastic bounds. We note that the classical $GI^{X}/M/1$ analysis rests on transform methods; these are numerically challenging even in the $M^{r}/M/1$ case with deterministic $X = r$ (see § 3).

The second application is in queues with Semi-Markovian arrival processes (SMPs), which jointly generalize renewal and Markov processes: the inter-arrival times not only can have general distributions but can also be correlated. Studying these models is motivated both by measurements (e.g., [5, 10, 19, 30, 33]) and analytical studies demonstrating a wide variability in the queue size as a function of the autocorrelation. For a specific queue with Markov-dependent interarrivals and exponential service times, the queue size can grow by as much as a factor of 9 at high utilizations depending on the lag-1 autocorrelation r (i.e., the correlation coefficient between two consecutive inter-arrivals) [31]; the baseline (renewal) case is $r = 0$. Significantly larger differences have been reported in [22], in the case of a queue with Markov renewals inter-arrivals (the modulating Markov chain has two states only) and exponential inter-arrivals. In the case of high utilizations (around 0.9), the queue size can grow by a factor of 100 by increasing the lag-1 autocorrelation from $r = 0.1$ to 0.7. Other extraordinary queueing effects when injecting autocorrelation in the interarrival times were reported in [18].

There are several computational/algorithmic methods to exactly analyze SM queues: One of the earliest works used transform methods for $SM/M/1$ [6]. Wiener-Hopf factorization was considered in [11] for $SM/SM/1$; explicit factorizations are available in special cases, e.g., the $SM/M/1$ queue, subject to some technical conditions (see also [12]). An iterative procedure was proposed in [29] for the $SM/PH/1$ queue. Transform methods are used in [1, 3] for the analysis of the $SM/SM/1$ queue, additionally subject to correlation between arrivals and service (there is a single Markov chain modulating both arrivals and service times); a related more general model is analyzed in [2] using matrix-analytical methods. Earlier applications of matrix-analytical methods were considered for the $SM/PH/1$ queue in the classical text [20], p. 159. As pointed out in [11], some of these methods are insufficiently understood from a numerical/computational point of view.

Using the newly constructed martingale, the $SM/M/1$ analysis proceeds similarly as for the $GI^{X}/M/1$ queue. While an identical analysis could be obtained using a (discrete-time) martingale, for suitable random embedding points (see § 4.1), the key advantage of the proposed (continuous-time) martingale is that it also applies to the ΣSM/M/1 queue involving a superposition (the 'Σ') of multiple SM arrivals. The key challenge arising in these systems is that the superposed arrival process is not stationary, let alone SM. The same idea has recently been used for $\Sigma GI/G/1$ [9] but using a different

[1] This model corresponds to a $GI/M/1$ queue given that a geometric sum of iid exponential random variables yields an exponential random variable.

Conference'17, July 2017, Washington, DC, USA

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

(continuous-time) martingale, which does however not lend itself to the $GI^X/M/1$ exact solutions herein.

In the following, we first introduce the model and present two key Lemmas for the construction of continuous-time martingales. Next we analyze the $GI^X/M/1$ and then the $SM/M/1$ and $\Sigma SM/M/1$ queues, along with some numerical comparisons of the bounds against simulations. Some conclusions and an Appendix including some helpful auxiliary results conclude the paper.

2 MODEL AND TWO LEMMAS

Jobs arrive in the order $n+k-1, n+k-2, \ldots, 0$ with the convention that job 0 is the last before time 0. For $i \geq 1$, the inter-arrival time between jobs i and $i-1$ is denoted by T_i; job 0 arrives T_0 time units before time 0. The service time of job i is denoted by S_i. The sequences $(T_n)_{n\in\mathbb{N}}$ and $(S_n)_{n\in\mathbb{N}\cup\{0\}}$ are stationary and independent of each other. Let $E[T_n] = \frac{1}{\lambda}, n \geq 1, E[S_n] = \frac{1}{\mu}, n \geq 0$, and assume for stability that the intensity $\rho := \frac{\lambda}{\mu} < 1$.

We are interested in the number of jobs Q in the system (queue + service) at time 0, whose steady-state distribution can be written as

$$\mathbb{P}(Q \geq k) = \lim_{n\to\infty} \mathbb{P}\left(\max_{k\leq j\leq n+k}\left(\sum_{l=j}^{n+k-1} T_l + \sum_{l=k-1}^{j-1} S_l\right) > \sum_{l=0}^{n+k-1} T_l\right) .$$

By changing the subscript in S_l from l to $l-k+2$ this can be rewritten as

$$\begin{aligned}\mathbb{P}(Q \geq k) &= \mathbb{P}(\exists m \geq 1 : S_1 + \cdots + S_m > T_0 + \cdots + T_{m+k-2})\\ &= \mathbb{P}\left(\sup_{t\geq 0}\{A(t) - S(t)\} \geq k\right) ,\end{aligned}$$

where $A(t)$ is the number of arrival points within the interval $[-t, 0)$ and

$$S(t) := \inf\{n \geq 0 : S_1 + \cdots + S_{n+1} > t\}$$

is the other counting process corresponding to the service times' process. By abuse of notation, let

$$Q := \sup_{t\geq 0}\{A(t) - S(t)\} . \tag{1}$$

Besides $\mathbb{P}(Q \geq k)$, we also consider the corresponding Palm distribution conditional of an arrival at time 0, i.e.,

$$\mathbb{P}_a(Q \geq k) := \lim_{\delta\downarrow 0} \mathbb{P}(Q \geq k \mid A(\delta) > 0) ,$$

where $A(\delta)$ denotes, with abuse of notation, the number of arrival points in $(0, \delta]$; the limit is the right-limit. The Palm probability $\mathbb{P}_a(\cdot)$ is to be understood as the probability of an event just before an arrival; denote also by $\mathbb{E}_a$ the corresponding expectation.

2.1 Two Lemmas

Our technique relies on the construction of (continuous-time) martingales from the counting processes $A(t)$ and $S(t)$. The next two lemmas offer the technical support.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a common probability space and $\mathcal{L}^1$ the class of integrable random variables (i.e, $\mathbb{E}[|X|] < \infty$). Let $\mathbf{X} = (X_t)_{t\geq 0}$ be a (continuous-time) stochastic process in $\mathcal{L}^1$ and $\mathcal{F}_t := \sigma(X_s,\ 0 \leq s \leq t)$ be the corresponding natural filtration. By definition, X_t is a (continuous-time) martingale if $X_t \in \mathcal{L}^1$ and

$$\mathbb{E}[X_t \mid \mathcal{F}_s] = X_s\ \forall 0 \leq s \leq t .$$

Define the auxiliary (random) functions $\forall t \geq 0$

$$\begin{aligned} f_t(s) &:= \mathbb{E}\left[X_{t+s} \mid \mathcal{F}_t\right]\ \forall s \geq 0 \text{ and}\\ g_t(s) &:= \frac{f_t(s) - f_t(0)}{s} = \frac{\mathbb{E}\left[X_{t+s} - X_t \mid \mathcal{F}_t\right]}{s}\ \forall s > 0 \text{ a.s.}\end{aligned}$$

Lemma 1. *The process $X_t \in \mathcal{L}^1$ is a martingale if and only if it satisfies the following conditions for all $t \geq 0$:*

(i) $f_t(s)$ is continuous in s;
(ii) $f_t(s)$ is (right-)differentiable at $s = 0$ and $f_t'(0) = 0$;
(iii) $\exists\delta_t > 0\ \exists Y_t \in \mathcal{L}^1$ such that $|g_t(s)| \leq Y_t\ \forall s \in (0, \delta_t)$.

This result generalizes a particular result from [9] for the construction of martingales from MAPs.

Proof. The necessity of the three properties is trivial: The first two follow from $f_t(s) = X_t\ \forall t \geq 0$; the last follows from $g_t(s) = 0\ \forall s > 0$.

For the other direction, fix $t \geq 0$ and $s > 0$. We can write

$$\begin{aligned} f_t'(s) &= \lim_{\delta\downarrow 0}\frac{f_t(s+\delta) - f_t(s)}{\delta} = \lim_{\delta\downarrow 0}\frac{\mathbb{E}\left[X_{t+s+\delta} - X_{t+s} \mid \mathcal{F}_t\right]}{\delta}\\ &= \lim_{\delta\downarrow 0}\frac{\mathbb{E}\left[\mathbb{E}\left[X_{t+s+\delta} - X_{t+s} \mid \mathcal{F}_{t+s}\right] \mid \mathcal{F}_t\right]}{\delta}\\ &= \lim_{\delta\downarrow 0}\mathbb{E}\left[g_{t+s}(\delta) \mid \mathcal{F}_t\right] = \mathbb{E}\left[\lim_{\delta\downarrow 0} g_{t+s}(\delta) \mid \mathcal{F}_t\right]\\ &= \mathbb{E}\left[\lim_{\delta\downarrow 0}\frac{f_{t+s}(\delta) - f_{t+s}(0)}{\delta} \mid \mathcal{F}_t\right]\\ &= \mathbb{E}\left[f_{t+s}'(0) \mid \mathcal{F}_t\right] = \mathbb{E}\left[0 \mid \mathcal{F}_t\right] = 0 .\end{aligned}$$

In the first line we used the Tower Property for conditional expectations. In the next we applied (iii) and the Dominated Convergence Theorem, and lastly we applied (ii). This shows that $f_t'(s)$ is right-differentiable and $f_t'(s) = 0\ \forall s \geq 0$. Using the continuity of $f_t(s)$ from (i) it then follows that $f_t(s)$ is a constant, i.e.,

$$\mathbb{E}\left[X_{t+s} \mid \mathcal{F}_t\right] = f_t(s) = f_t(0) = \mathbb{E}\left[X_t \mid \mathcal{F}_t\right] = X_t\ \forall s \geq 0 .$$

□

Note that the continuity condition (i) is needed for the last argument, whereas (iii) is needed for the interchange of the limit with the expectation.

We shall use Lemma 1, and more precisely the differentiability condition (ii), for martingale constructions[2]. Because validating the condition (iii) is exceptionally difficult/tedious, we will prove the martingale property using the next result.

Lemma 2. *Let Y_t be a Markov process on a state space E with generator $(A, D(A))$. Assume that there exist a sequence of stopping times $\tau_K \uparrow \infty, K \in \mathbb{N}$, such that $\{Y_{t\wedge\tau_K}, t \geq 0\}$ is a stopped Markov process with state space $E_K \subseteq E$, for each $K \in \mathbb{N}$. Let $f : E \longrightarrow \mathbb{R}$ be a measurable function bounded on E_K, for each $K \in \mathbb{N}$. If $Af(Y_t) = 0$*

[2] An example of a stochastic process which does not satisfy (ii) is $X_t = |B_t|$, where B_t is the standard Brownian motion. Indeed $f_0(s) = \mathbb{E}\left[|B_s| \mid \mathcal{F}_0\right] = \mathbb{E}\left[|B_s|\right] = \int_0^\infty 2x\frac{1}{\sqrt{2\pi s}}e^{-\frac{x^2}{2s}}\,dx = \sqrt{\frac{2s}{\pi}}$, which is not right-differentiable at 0.

for all $t < \tau_K$ *then* $\{f(Y_{t\wedge\tau_K}), t \geq 0\}$ *is a martingale, and hence* $\{f(Y_t), t \geq 0\}$ *is a local martingale.*

This is an application of Dynkin's Formula (see Appendix § A). This result along with a careful technical argument (see the proofs of Corollaries 3 and 5) are particularly needed because Dynkin's Formula cannot be directly applied to show that the constructed processes $X(t)$ are martingales, because they are not bounded.

Proof. Fix $K \in \mathbb{N}$ and let $f|_{E_K} : E_K \longrightarrow \mathbb{R}$ be a bounded measurable function. Applying Dynkin's Formula for $\{Y_{t\wedge\tau_K}, t \geq 0\}$ yields the martingale

$$M_t := f(Y_{t\wedge\tau_K}) - \int_0^t Af(Y_{s\wedge\tau_K})ds \,,$$

and hence

$$M_{t\wedge\tau_K} := f(Y_{t\wedge\tau_K}) - \int_0^{t\wedge\tau_K} Af(Y_s)ds = f(Y_{t\wedge\tau_K})$$

by using $Af(Y_s) = 0$ for $s < \tau_K$. The rest is clear. □

3 GI^X/M/1

We can now proceed with the general analysis of the $GI^X/M/1$ queue and then consider several special cases.

The arrivals are driven by a renewal process; denote by $f(t)$ and $F(t)$ the density and distribution, respectively, of the inter-arrivals. The arrivals occur in bulks/batches following the distribution of a (non-negative) discrete random variable $X > 0$; the arrival rate is thus $\lambda\mathbb{E}[X]$ and the service rate is scaled to $\mu := \frac{\lambda\mathbb{E}[X]}{\rho}$. Let the moment generating function (MGF) $M_X(\theta) := \sum_{k=1}^{\infty} \mathbb{P}(X = k)e^{\theta k}$, for some $\theta > 0$. Let also $\gamma(t) := \frac{f(t)}{1-F(t)}$ be the hazard rate of T_1 and $L(\zeta) := E\left[e^{-\zeta T_1}\right]$ be the corresponding Laplace transform, where $\zeta := \mu(1 - e^{-\theta})$. Denote also by $R(t)$ the remaining lifetime of the arrival process at time $-t$; recall that we are interested in the steady-state queue size $Q = \sup_{t\geq 0}\{A(t) - S(t)\}$ at time 0.

We assume that the batch size is *light-tailed*, i.e., it admits finite moment generating functions. Concretely, we assume that

$$\sup\{\theta > 0 : \mathbb{E}[e^{\theta X}] < \infty\} \in (0, \infty]$$

and also that there exists $\theta > 0$ such that

$$g(\theta) := \mathbb{E}[e^{\theta X}]L(\zeta) \in [1, \infty) \,. \qquad (2)$$

This latter condition prevents X from having a 'thin' tail; such a case could be easily treated by our framework by constructing super-martingales rather than martingales, as done next. A certainly interesting case, but not covered by our framework, is when X has a heavy-tailed distribution.

Corollary 3. (The martingale) *The process*

$$X_t := h(R(t))e^{\theta(A(t)-S(t))} \,, t \geq 0$$

is a (continuous-time) martingale, where

$$h(t) = \frac{M_X(\theta)\int_t^{\infty} e^{-\zeta s}f(s)ds}{(1 - F(t))e^{-\zeta t}}$$

and $\theta > 0$ *is the largest solution of*

$$M_X(\theta)^{-1} = L(\zeta) \,. \qquad (3)$$

Moreover, $h(t)$ *is bounded.*

A key element in the structure of X_t is the random function $h()$ which takes as parameter the 'memory' $R(t)$ at time $-t$. This is needed to compute the probability of an arrival within the interval $[-(t + \delta), -t)$ in terms of the corresponding hazard rate, i.e., $\gamma(R(t))\delta + o(\delta)$ for some small $\delta > 0$, where $\lim_{\delta\to 0} \frac{o(\delta)}{\delta} = 0$. Note that the hazard rate replaces the rate λ, had the arrival process been a Poisson process.

Proof. Fix $t \geq 0$. The expressions for θ and $h(t)$ follow by invoking condition (ii) from Lemma 1, i.e.,

$$f_t'(0) = 0 \Leftrightarrow \lim_{\delta\downarrow 0} \frac{\mathbb{E}\left[X_{t+\delta} - X_t \mid \mathcal{F}_t\right]}{\delta} = 0$$

$$\Leftrightarrow \lim_{\delta\downarrow 0} \frac{\mathbb{E}\left[h(R(t+\delta))e^{\theta(A(t,t+\delta)-S(t,t+\delta))} - h(R(t)) \mid \mathcal{F}_t\right]}{\delta} = 0 \,.$$

Expanding the expectation yields

$$\begin{aligned}\lim_{\delta\downarrow 0}\frac{1}{\delta}\Big(&\gamma(R(t))\delta(1-\mu\delta)h(0)M_X(\theta)\\ &+(1-\gamma(R(t))\delta)(1-\mu\delta)h(R(t)+\delta)\\ &+(1-\gamma(R(t))\delta)\mu\delta h(R(t)+\delta)e^{-\theta} - h(R(t)) + o(\delta)\Big) = 0 \qquad (4)\end{aligned}$$

We note that on an arrival $R(t + \delta) = 0$, i.e., the residual lifetime resets itself, whence the term $h(0)$ in the first line; we also used the conditional independence of A and S. Denoting for convenience $R(t) = t$ and rearranging terms we obtain

$$h'(t) = (\gamma(t) + \zeta)h(t) - \gamma(t)h(0)M_X(\theta) \,.$$

Setting the initial value problem with $h(0) = 1$ (note that $h(t)$ can be arbitrarily scaled) yields

$$h(t) = \frac{1 - M_X(\theta)\int_0^t e^{-\zeta s}f(s)ds}{(1 - F(t))e^{-\zeta t}} \,.$$

Imposing that $h(t)$ is to be bounded, $M_X(\theta)\int_0^{\infty} e^{-\zeta s}f(s)ds = 1$ must necessarily hold because $(1 - F(t))e^{-\zeta t} \xrightarrow[t\to\infty]{} 0$. This yields the expression of $h(t)$ and the condition on θ from the Corollary.

To show the existence of θ we need to show that there exists a solution to $g(\theta) = 1$. Using $g(0) = 1$, the assumption from (2), and the stability condition $g'(0) = \mathbb{E}[X] - \mu\mathbb{E}[T_{n+1}] < 0$ the existence of θ is proved. We take $\theta := \max\{r : g(r) = 1\}$.

Next, to show that $h(t)$ is bounded we integrate by parts:

$$\begin{aligned} h(t) &= M_X(\theta)\left(1 + \frac{\zeta\int_t^{\infty} e^{-\zeta s}F(s)ds - e^{-\zeta t}}{(1 - F(t))e^{-\zeta t}}\right)\\ &\leq M_X(\theta)\left(1 + \frac{\zeta\int_t^{\infty} e^{-\zeta s}ds - e^{-\zeta t}}{(1 - F(t))e^{-\zeta t}}\right)\\ &= M_X(\theta) \,.\end{aligned}$$

Finally, to prove that X_t is a martingale, let the Markov process $Y_t := (A(t) - S(t), R(t)), t \geq 0$ and define

$$\tau_K := \inf\{t \geq 0 : A(t) - S(t) \geq K\}$$

for all $K \in \mathbb{N}$, with the convention that $\inf\emptyset = \infty$, and

$$f(N, t) := h(t)e^{\theta N} \,,$$

for all $N \in \mathbb{N}$ and $t \geq 0$. Consider the stopped Markov process $Y_{t \wedge \tau_K}$. We have for $N \leq K - 1$

$$\begin{aligned} Af(N,t) &= \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{E}[f(Y_\delta) - f(N,t) \mid Y_0 = (N,t)] \\ &= \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{E}\Bigg[h(t+\delta) e^{\theta N} (1 - \gamma(t)\delta - \mu\delta) \\ &\qquad + h(0) e^{\theta X} e^{\theta N} \gamma(t)\delta + h(t+\delta) e^{\theta(N-1)} \mu\delta + o(\delta) \Bigg] \\ &= e^{\theta N} \left[h'(t) - (\zeta + \gamma(t)) h(t) + \gamma(t) h(0) \mathbb{E}[e^{\theta X}] \right] . \end{aligned}$$

Therefore $Af(N,t) = 0$ if and only if

$$h'(t) - (\zeta + \gamma(t)) h(t) + \gamma(t) \mathbb{E}[e^{\theta X}] h(0) = 0 ,$$

which holds by construction. Hence, by Lemma 2, $f(Y_t) = X_t$ is a local martingale. According to Theorem 9, the local martingale $f(Y_t)$ is a martingale if and only if it is of class DL, i.e, for every $t \geq 0$ the set $\{f(Y_\tau),\ \text{stopping time } \tau \leq t \text{ a.s.}\}$ is uniformly integrable. We have for $\tau \leq t$

$$|f(Y_\tau)| \leq |h|_\infty e^{\theta A(t)}$$

and

$$\mathbb{E}[e^{\theta A(t)}] = \sum_{k=1}^{\infty} \mathbb{E}[e^{\theta X}]^k \mathbb{P}(A_1(t) = k)$$

where $A_1(t) := \inf\{n \geq 0 : T_0 + T_1 + \cdots + T_n > t\}$. By applying Lemma 12 for A_1 we get $\mathbb{E}[e^{\theta A(t)}] < \infty$. Therefore, $f(Y_t) = X_t, t \geq 0$ is a martingale. □

Theorem 4. (Queue Distribution) *Let*

$$K_X(i) := \mathbb{E}[e^{\theta(X-i)} \mid X \geq i], i \in \mathbb{N} .$$

Then for all $k \geq 1$

$$\frac{e^{-\theta k}}{\sup_{i \in \mathbb{N}} K_X(i)} \leq \mathbb{P}_a(Q \geq k) \leq \frac{e^{-\theta k}}{\inf_{i \in \mathbb{N}} K_X(i)}$$

and

$$\begin{aligned} \frac{\rho}{\mathbb{E}[X]} \frac{\mathbb{E}[e^{\theta X}] - 1}{\sup_{i \in \mathbb{N}} K_X(i)(1 - e^{-\theta})} e^{-\theta k} &\leq \mathbb{P}(Q \geq k) \\ &\leq \frac{\rho}{\mathbb{E}[X]} \frac{\mathbb{E}[e^{\theta X}] - 1}{\inf_{i \in \mathbb{N}} K_X(i)(1 - e^{-\theta})} e^{-\theta k} , \end{aligned}$$

with θ *and* $h(t)$ *from Corollary 3.*

Proof. Fix $k \geq 0$ and let the stopping time

$$T := \inf\{t : A(t) - S(t) \geq k\} .$$

For some $t \geq 0$, applying the Optional Stopping Theorem to the martingale X_t from Corollary 3 yields:

$$\begin{aligned} \mathbb{E}_a[X_0] &= \mathbb{E}_a[X_{T \wedge t}] = \mathbb{E}_a[X_T \mathbb{1}_{T \leq t}] + \mathbb{E}_a[X_t \mathbb{1}_{T > t}] \\ &= \mathbb{E}_a\left[h(R(T)) e^{\theta(A(T) - S(T))} \mathbb{1}_{T \leq t}\right] + \mathbb{E}_a[X_t \mathbb{1}_{T > t}] , \end{aligned}$$

where $\mathbb{1}_\cdot$ denotes the indicator function. A key observation is that $h(R(T)) = h(0) = 1$, because T must happen at an arrival point (of a batch). Moreover, $\mathbb{E}_a[X_0] = \mathbb{E}_a[h(R(0))] = 1$ because $R(0) = 0$ (there is an arrival right after 0). Taking $t \to \infty$ the second term in the derivation above vanishes according to Lemma 11 (see Appendix § A) and hence

$$1 = \mathbb{E}_a\left[e^{\theta(A(T) - S(T))} \mathbb{1}_{T < \infty}\right] . \tag{5}$$

For small $\delta > 0$ we can write

$$\begin{aligned} &\mathbb{E}_a\left[e^{\theta(A(T) - S(T))} \mathbb{1}_{t < T \leq t+\delta}\right] = \mathbb{E}_a\Big[e^{\theta(A((t,t+\delta]) - S((t,t+\delta]))} \\ &e^{\theta(A(t) - S(t))} \mathbb{1}_{t < T} \mathbb{1}_{A((t,t+\delta]) - S((t,t+\delta]) \geq k - A(t) + S(t)}\Big] + o(\delta) \\ &= \mathbb{E}_a\Big[\mathbb{E}_a\big[e^{\theta(A((t,t+\delta]) + A(t) - S(t))} \mathbb{1}_{A((t,t+\delta]) \geq k - A(t) + S(t)} \\ &\qquad \mid \mathcal{F}_t, A((t,t+\delta]) > 0\big] \mathbb{1}_{t < T, A((t,t+\delta]) > 0}\Big] + o(\delta) \\ &= \mathbb{E}_a\Big[\mathbb{E}_a\big[e^{\theta(X + A(t) - S(t))} \mathbb{1}_{X \geq k - A(t) + S(t)} \mid \mathcal{F}_t, A((t,t+\delta]) > 0\big] \\ &\qquad \mathbb{1}_{t < T, A((t,t+\delta]) > 0}\Big] + o(\delta) \\ &= e^{\theta k} \mathbb{E}_a\left[K_X(k - A(t) + S(t)) \mathbb{1}_{X \geq k - A(t) + S(t)} \mathbb{1}_{T > t, A((t,t+\delta]) > 0}\right] \\ &\qquad + o(\delta) \\ &= e^{\theta k} \mathbb{E}_a\left[K_X(k - A(t) + S(t)) \mathbb{1}_{t < T \leq t+\delta}\right] + o(\delta) , \end{aligned}$$

immediately implying that

$$\begin{aligned} \inf_{i \in \mathbb{N}} K_X(i) \mathbb{P}_a(t < T \leq t+\delta) e^{k\theta} + o(\delta) &\leq \mathbb{E}_a\left[e^{\theta(A(T) - S(T))} \mathbb{1}_{t < T \leq t+\delta}\right] \\ &\leq \sup_{i \in \mathbb{N}} K_X(i) e^{k\theta} \mathbb{P}_a(t < T \leq t+\delta) + o(\delta) . \end{aligned}$$

Further integrating over $[0, \infty)$ yields

$$\begin{aligned} \inf_{i \in \mathbb{N}} K_X(i) \mathbb{P}_a(T < \infty) e^{k\theta} &\leq \mathbb{E}_a\left[e^{\theta(A(T) - S(T))} \mathbb{1}_{T < \infty}\right] \\ &\leq \sup_{i \in \mathbb{N}} K_X(i) e^{k\theta} \mathbb{P}_a(T < \infty) , \end{aligned}$$

and finally using (5) we obtain

$$\frac{e^{-\theta k}}{\sup_{i \in \mathbb{N}} K_X(i)} \leq \mathbb{P}_a(Q \geq k) = \mathbb{P}_a(T < \infty) \leq \frac{e^{-\theta k}}{\inf_{i \in \mathbb{N}} K_X(i)} .$$

The second result proceeds similarly. Note that $h(R(T)) = 1$ for the same reason as above;the rest follows by immediate integration

$$\mathbb{E}[X_0] = \mathbb{E}[h(R(0))] = \int_0^\infty h(x) \frac{1 - F(x)}{\mathbb{E}[T_1]} dx = \frac{\rho}{\mathbb{E}[X]} \frac{\mathbb{E}[e^{\theta X}] - 1}{1 - e^{-\theta}} .$$

□

In the literature, the distribution of Q is available implicitly in terms of the probability generating function (PGF) of Q, i.e., $G(z) := \sum_{k=0}^{\infty} q_k z^k$, where $q_k := \mathbb{P}(Q = k)$. The PGF $G(z)$ is implicit itself in terms of the roots inside the unit circle $|z| = 1$ of the equation $L(\mu(1 - z)) = z^r$, which are guaranteed by Rouché's Theorem ([8], p. 118). In the $\mathrm{M^r}/M/1$ case, $G(z)$ is explicit ([8], p. 121), i.e.,

$$G(z) = \frac{(1 - z)(1 - \rho)}{(1 - z) - \frac{\rho}{r}(1 - z^r)} .$$

However, the derivation of the q_k's and consequently of $\mathbb{P}(Q \geq k)$ requires numerical inversions of $G(z)$, which is a "reasonable task when r is small" ([15], p. 120).

Therefore, the contribution of Theorem 4 are closed-form (and transform-free) bounds on Q's distribution for $\mathrm{GI^X}/M/1$; see also below for some cases with exact results. We note that while the parameter θ is implicit, it can be obtained using a simple (logarithmic-time) binary search. For instance, in the $\mathrm{M^r}/M/1$ case, θ satisfies

$$\lambda e^{\theta r} + \mu e^{-\theta} - (\lambda + \mu) = 0 \ .$$

Denoting $x = e^{\theta}$, this can be rewritten as

$$x + \cdots + x^r = \frac{\mathbb{E}[X]}{\rho} \ , \tag{6}$$

which has a unique (real) solution when $\rho < 1$. Some numerical results will be shown in § 4.2.

3.1 Special Cases: $M/M/1$, $GI/M/1$, $GI^{\mathbf{Geo}}/M/1$

When the arrival process is Poisson there is no need for tracking the 'memory' $R(t)$, due to the memoryless property, and hence $h = 1$. The martingale process becomes

$$X_t := e^{\theta(A(t)-S(t))} \ , \ t \geq 0 \ ,$$

where $\theta := -\ln \rho$. Remarkably, the bound on $\mathbb{P}(Q \geq k)$ becomes exact, i.e., $\mathbb{P}(Q \geq k) = \rho^k$; the reason is that $A(T) - S(T) = k$ when the batch size is 1, as a by-product of using *counting processes* in the representation of X_t. The exact result is also captured using Kingman's (discrete-time) martingale

$$X_n = e^{\theta \sum_{i=1}^{n}(S_i - T_i)} \ \forall n \in \mathbb{N} \ .$$

The $GI/M/1$ analysis follows by simply letting $X = 1$. For the same reason as in the $M/M/1$ case, i.e., $A(T) - S(T) = k$, the exact result (shown in (7) below) is recovered. We point out that both Kingman's martingale as well as an alternative recent (continuous-time) martingale for $GI/G/1$ from [9]

$$X_t = h(R(t)) e^{\theta(\sum_{i=1}^{N(t)} S_i - t)} \ \forall t \geq 0 \ ,$$

only yield upper bounds, due to the existence of an overshoot at time T; here, the martingale is built in terms of a compound process drained at a rate 1 corresponding to the '$-t$' term in the exponent. An exact result was obtained however earlier by Ross [27] using an additional stopping time, a neat conditioning argument, and the memoryless property of the service times.

Lastly, the $GI^X/M/1$ analysis when X has a Geometric distribution with parameter p, i.e., $\mathbb{P}(X = i) = p(1-p)^{i-1}, i \in \mathbb{N}$, follows by noting that $K_X(i) = \frac{p}{1-e^{\theta}(1-p)}$ is invariant to i (as a by-product of the memoryless property of X). Applying Theorem 4 we obtain the exact result for $k \geq 1$

$$\mathbb{P}(Q \geq k) = \rho e^{\theta(1-k)} . \tag{7}$$

This is the same result as for $GI/M/1$ except for the value of θ driven by (3) from Corollary 3. We also note that our solution drastically simplifies the standard $GI/M/1$ solution (see, e.g., [15], p.259), especially concerning the existence of the unique solution which relies on a much more involved argument based on Rouché's Theorem from complex analysis. While only the existence of θ was proven in Corollary 3, we note that in both $GI/M/1$ and $GI^{\mathrm{Geo}}/M/1$ cases unicity is an immediate by-product of the exact result since $\lim_{k\to\infty} \mathbb{P}(Q \geq k)^{1/k} = e^{-\theta}$.

4 SM/M/1

A Semi-Markovian Process (SMP), also called Markov Renewal Process or (discrete) Markov Additive Process [4], is defined by the kernel

$$\begin{aligned} F_{i,j}(x) &:= \mathbb{P}\left(M_{n+1} = j, T_{n+1} \leq x \mid M_n = i\right) \\ &= \mathbb{P}\left(M_{n+1} = j, T_{n+1} \leq x \mid M_n = i, M_{n-1}, M_{n-2} \ldots, T_n, T_{n-1} \ldots\right) \ , \end{aligned}$$

for all $n \in \mathbb{Z}, \ i, j \in \mathcal{S}, \ x \geq 0$, where the stationary process M_n denotes the state after n transitions and T_n denotes the sojourn time in the $n - 1^{\text{th}}$ state; the state space $\mathcal{S}$ is finite and denote the number of states by $|\mathcal{S}|$. The stochastic process $(M_n, T_n)_n$ itself is called a Markov Renewal Process.

At one extreme, if there is a single state (i.e., $|\mathcal{S}| = 1$), then the point process T_n (i.e., the collection of points generated by the intervals T_n) is a renewal process and hence the $SM/M/1$ queue instantiates to the $GI/M/1$ queue. At the other extreme, if the sojourn times in all of the states are exponentially distributed (i.e., $F_i(x) = 1 - e^{-\lambda_i x} \ \forall i$), then the SMP instantiates to a Markov process.

In an SMP, the sojourn times in the states not only follow general distributions but are not necessarily independent; they are however conditionally independent on the states, e.g.,

$$\begin{aligned} &\mathbb{P}(T_{n+1} \leq x, T_n \leq y \mid M_{n+1}, M_n, M_{n-1}) \\ &\quad = \mathbb{P}(T_{n+1} \leq x \mid M_{n+1}, M_n)\mathbb{P}(T_n \leq y \mid M_n, M_{n-1}) \ . \end{aligned}$$

The transition probabilities of the embedded (and assumed ergodic) Markov chain are

$$q_{i,j} = \mathbb{P}\left(M_{n+1} = j \mid M_n = i\right) \ \forall i, j = 1, \ldots, |\mathcal{S}| \ ,$$

and the stationary distribution is

$$q_i = \mathbb{P}(M_n = i) \ \forall n \in \mathbb{Z}, \ i \in \mathcal{S} \ ,$$

i.e., $qQ = q$, where $Q := (q_{i,j})_{i,j=1,\ldots,|\mathcal{S}|}$. Therefore,

$$\frac{1}{\lambda} = \sum_i q_i \frac{1}{\lambda_i} \ ,$$

where $\frac{1}{\lambda_i} = \mathbb{E}\left[T_{n+1} \mid M_n = i\right]$ is the expected sojourn time in state i. Due to the representation of Q from (1), we assume without loss of generality that the Markov chain M_n is reversible. Otherwise, one would need to construct and work with the corresponding reversed process, and more exactly to replace Q and $F_{i,j}$ by

$$\begin{cases} Q^* = D^{-1} Q^T D \\ F^*_{i,j} = \frac{q_j}{q_i} F_{j,i} \ , \end{cases}$$

where D is a diagonal matrix with the q_i's on its main diagonal.

We also assume that at time 0 the state is M_0 such that the associated SMP is $M(t) = M_{A(t)}$, i.e., the state at time $-t$. Its stationary distribution is ([17], p. 411)

$$\pi_i = \frac{q_i \frac{1}{\lambda_i}}{\sum_j q_j \frac{1}{\lambda_j}} = q_i \frac{\lambda}{\lambda_i} \ . \tag{8}$$

The distribution of the sojourn time in state i is denoted by

$$F_i(x) = \mathbb{P}\left(T_{n+1} \leq x \mid M_n = i\right) \ .$$

If this distribution is (absolutely) continuous[3], the corresponding density is denoted by $f_i(x)$ and the hazard rate by

$$\gamma_i(x) := \lim_{\delta \downarrow 0} \frac{\mathbb{P}\left(x < T_{n+1} \le x + \delta \mid x < T_{n+1}, M_n = i\right)}{\delta} = \frac{f_i(x)}{1 - F_i(x)} \ .$$

Denoting $f_{i,j}(x) = F'_{i,j}(x)$ we also consider the hazard rate

$$\gamma_{i,j}(x) := \lim_{\delta \downarrow 0} \frac{\mathbb{P}\left(x < T_{n+1} \le x + \delta, M_{n+1} = j \mid x < T_{n+1}, M_n = i\right)}{\delta}$$
$$= \frac{f_{i,j}(x)}{1 - F_i(x)} \ .$$

We can now proceed to the construction of an $SM/M/1$ martingale, similarly as in the $GI^X/M/1$ case. First, let $L_{i,j}(\zeta)$ be the Laplace transform of the sojourn time in state i with a transition to state j, i.e., $L_{i,j}(\zeta) := \int_0^\infty e^{-\zeta x} dF_{i,j}(x)$, where $\zeta := \mu(1 - e^{-\theta})$ for some $\theta > 0$.

Corollary 5. (The martingale - Construction) *Let $e^{-\theta}$, with $\theta > 0$, and $h(0) = (h_i(0))_{i \in \mathcal{S}}$ be respectively the (unique) Perron eigenvalue (i.e., positive and maximal) and eigenvector of the matrix $[L_{ij}(\zeta)]_{i,j \in \mathcal{S}}$. The process*

$$X_t := h_{M(t)}(R(t)) e^{\theta(A(t) - S(t))} \ , t \ge 0$$

is a martingale, where[4]

$$h_i(t) = \frac{e^\theta \sum_j h_j(0) \int_t^\infty e^{-\zeta s} f_{i,j}(s) ds}{(1 - F_i(t)) e^{-\zeta t}} \ \forall i$$

are bounded.

Proof. As in the proof of Corollary 3, we determine θ and $h_i(t)$ by applying condition (ii) from Lemma 1. Writing for simplicity $R(t) = t$ the differentiability condition becomes for all $i \in \mathcal{S}$

$$\lim_{\delta \downarrow 0} \frac{1}{\delta} \Big(\sum_j \gamma_{i,j}(t) \delta (1 - \mu\delta) h_j(0) e^\theta + (1 - \gamma_i(t)\delta)(1 - \mu\delta) h_i(t + \delta)$$
$$+ (1 - \gamma_i(t)\delta) \mu\delta h_i(t + \delta) e^{-\theta} - h_i(t) \Big) = 0 \ ,$$

which leads to the system of ODEs:

$$h'_i(t) = (\gamma_i(t) + \zeta) h_i(t) - e^\theta \sum_j \gamma_{i,j}(t) h_j(0) \ .$$

Setting the initial value problem with some values $h_i(0)$ yields

$$h_i(t) = \frac{h_i(0) - \sum_j h_j(0) \int_0^t e^{-\zeta s} f_{i,j}(s) e^\theta ds}{(1 - F_i(t)) e^{-\zeta t}} \ .$$

Because $(1 - F(t)) e^{-\zeta t} \xrightarrow[t \to \infty]{} 0$, in order for $h_i(t)$ to be bounded $M_X(\theta) \int_0^\infty e^{-\zeta s} f(s) ds = 1$ necessarily holds, whence the expressions of $h_i(t)$ and the condition on θ from Corollary 5.

Next we show that the solution to the eigenvalue problem exists and is unique. By Perron-Frobenius Theorem, for some arbitrary $\zeta = \mu(1 - e^{-\theta}) > 0$, there exists a unique positive and maximal eigenvalue $e^{-\kappa(\zeta)}$, and a positive vector $h^{(\zeta)}$ for the positive matrix $[L_{ij}(\zeta)]$, i.e.,

$$[L_{ij}(\zeta)] h^{(\zeta)} = e^{-\kappa(\zeta)} h^{(\zeta)} \ .$$

We have to show that there exists a unique solution $\theta > 0$ for the fixed point equation

$$\kappa(\mu(1 - e^{-\theta})) = \theta \ . \quad (9)$$

Define the function $f(\theta) := \kappa(\mu(1 - e^{-\theta})) - \theta$. Then $f(0) = 0$, $f(\infty) = -\infty$, and $f'(0) = \mu\kappa'(0) - 1$. By Corollary XI.2.9 from [4],

$$\kappa'(0) = \sum_{i,j \in \mathcal{S}} q_i q_{i,j} \mathbb{E}[T_{n+1} \mid M_n = i, M_{n+1} = j]$$
$$= \sum_{i \in \mathcal{S}} q_i \mathbb{E}[T_{n+1} \mid M_n = i] = \frac{1}{\lambda} \ ,$$

i.e., the inverse of the arrivals' (stationary) rate. From the stability condition $\frac{\mu}{\lambda} > 1$ it follows that $f'(0) > 0$ and hence there exists a zero root for the function f. This proves the existence of θ and $h(0)$. Let us now briefly refer to the follow-up Theorem 6 and note that $\lim_{k \to \infty} \mathbb{P}(Q \ge k)^{1/k} = e^{-\theta}$, which proves that θ is unique. By Perron-Frobenius theorem, there is no other positive eigenvector of $[L_{ij}(\zeta)]$ except for positive multiples of $h(0)$.

Showing that $h_i(t)$'s are bounded follows as in the proof of Corollary 3, i.e., for all $i \in \mathcal{S}$

$$h_i(t) \le e^\theta \sum_i h_i(0) \ .$$

Finally, to prove that X_t is indeed a martingale, let the Markov process $Y_t := (A(t) - S(t), R(t), M(t)), t \ge 0$ and define

$$\tau_K := \inf\{t \ge 0 : A(t) - S(t) = K\} \ ,$$

for all $K \in \mathbb{N}$, with the convention that $\inf \emptyset = \infty$, and

$$f(N, t, i) := h_i(t) e^{\theta N} \ ,$$

for all $N \in \mathbb{N}$, $t \ge 0$, and $i \in \mathcal{S}$. For the stopped Markov process $Y_{t \wedge \tau_K}$ we can write for $N \le K - 1$

$$Af(N, t, i) = \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{E}[f(Y_\delta) - f(N, t, i) \mid Y_0 = (N, t, i)]$$
$$= \lim_{\delta \to 0} \frac{1}{\delta} \Bigg[h_i(t + \delta) e^{\theta N} (1 - \gamma(t)\delta - \mu\delta)$$
$$+ \left(\sum_j p_{i,j} h_j(0) \right) e^{\theta(N+1)} \gamma(t) \delta + h_i(t + \delta) e^{\theta(N-1)} \mu\delta + o(\delta) \Bigg]$$
$$= e^{\theta N} \left[h'_i(t) - (\mu(1 - e^{-\theta}) + \gamma(t)) h_i(t) + \gamma(t) e^\theta \sum_j p_{i,j} h_j(0) \right] \ .$$

Note that $Af(N, t, i) = 0$ if and only if

$$h'_i(r) - (\mu(1 - e^{-\theta}) + \gamma(r)) h_i(r) + \gamma(r) e^\theta \sum_j p_{i,j} h_j(0) = 0 \ ,$$

which holds by construction. Hence, by Lemma 2, $f(Y_t) = X_t$ is a local martingale; according to Theorem 9, this is further a martingale if and only if it is of class DL, i.e, for every $t \ge 0$ the set $\{f(Y_\tau), \text{ stopping time } \tau \le t \text{ a.s.}\}$ is uniformly integrable. We have for $\tau \le t$

$$|f(Y_\tau)| \le |h|_\infty e^{\theta A(t)} \ .$$

Because $\mathbb{E}[e^{\theta A(t)}] < \infty$ from Lemma 12, it finally follows that $f(Y_t), t \ge 0$ is a martingale. □

[3] While the construction of martingales require (absolute) continuity, the produce martingales hold for general distributions.

[4] All sums in this (sub-)section are taken over the state-space $\mathcal{S}$ of the chain M_n.

Theorem 6. (Queue Distribution) *Denote by $\mathbb{P}_{a,i}$ the Palm probability of an event just before an arrival with state i. Then for all $k \geq 0$*

$$\frac{h_i(0)}{\max_i h_i(0)}e^{-\theta k} \leq \mathbb{P}_{a,i}(Q \geq k) \leq \frac{h_i(0)}{\min_i h_i(0)}e^{-\theta k}$$

and

$$\rho \frac{\sum_i q_i h_i(0)}{\max_i h_i(0)}e^{-\theta(k-1)} \leq \mathbb{P}(Q \geq k) \leq \rho \frac{\sum_i q_i h_i(0)}{\min_i h_i(0)}e^{-\theta(k-1)} ,$$

with θ and $h_i(t)$ from Corollary 5.

Proof. Fix $k \geq 0$. The Palm distribution proceeds exactly as in the proof of Theorem 4. For the other we need to compute

$$\begin{aligned} E[X_0] &= E[h_{M(0)}(R(0))] = \sum_i \pi_i \mathbb{E}[h_i(R(0)) \mid M(0) = i] \\ &= \sum_i \lambda_i \pi_i \int_0^\infty h_i(x)(1-F_i(x))dx \\ &= \sum_i \lambda_i \pi_i \int_0^\infty \frac{e^\theta \sum_j h_j \int_x^\infty e^{-\zeta y} f_{i,j}(y)dy}{(1-F_i(x))e^{-\zeta x}}(1-F_i(x))dx \\ &= \sum_i \lambda_i \pi_i e^\theta \sum_j h_j(0) \int_0^\infty e^{-\zeta y} f_{i,j}(y) \int_0^y e^{\zeta x} dx dy \\ &= \sum_i \frac{\lambda_i \pi_i e^\theta}{\zeta} \sum_j h_j(0)(q_{i,j} - L_{ij}(\zeta)) , \end{aligned}$$

and the rest follows from (8) and the construction of $e^{-\theta}$ and $h_j(0)$ from Corollary 5. □

4.1 ΣSM/M/1

The input is a superposition of two[5] SMP processes. We use the same notation as earlier except for subscripting the parameters accordingly (e.g., λ_i for the arrival rates of the SMPs); the overall service rate is $\mu > \lambda_1 + \lambda_2$.

The fundamental difficulty of such a system is that the superposed input is not stationary, even when the individual inputs are renewals (unless Poisson). To see why our continuous-time martingale is particularly suitable to capture this system, consider the alternative (discrete-time) $SM/M/1$ $\mathcal{F}_{\tau_n}$-martingale

$$X_n := h_{M(\tau_n)} e^{\theta(A(\tau_n)-S(\tau_n))} , n \geq 1,$$

where $\tau_n := \inf\{t \geq 0 : A(t) = n\}$ are the arrival points and $h_i \equiv h_i(0)$ for brevity. The proof follows from ([4], Proposition XI.2.4) by noting that $J_n := M(\tau_n)$ is an $\mathcal{F}_{\tau_n}$-MAP; it can be shown that this martingale yields the same results as Theorem 6. The key reason why this martingale is not suitable for the superposed process is that the embedding points corresponding to the individual arrivals are *different* (i.e., τ_n^1 and τ_n^2).

In continuous time, however, multiplexing martingales proceeds by first building two martingales for two fictitious $SM/M/1$ queues with service rates μ_1 and μ_2, respectively, such that $\mu_1 + \mu_2 = \mu$; also, $\mu_i > \lambda_i$ for the stability of the two queues. Applying Corollary 5, the two martingales are

$$X_{i,t} := h_{i,M_i(t)}(R_i(t))e^{\theta_i(A_i(t)-S_i(t))} , i = 1, 2, \ t \geq 0,$$

where $S_i(t)$ is a Poisson process with rate μ_i. Recall that θ_i depends on μ_i, and we write this dependency in terms of the functions $\theta_1(\mu_1)$ and $\theta_2(\mu_2)$, which are positive (see the argument about the existence and uniqueness of 'θ' from Corollary 5).

Next we show how to find μ_1 to obtain a suitable martingale for the original ΣSM/M/1 queue from $X_{1,t}$ and $X_{2,t}$. First, the stability conditions yield the following margins for the parameter μ_1:

$$\lambda_1 < \mu_1 < \mu - \lambda_2 .$$

From the last part of the proof of Corollary 5, involving the existence of $\theta > 0$, we obtain that

$$\lim_{\mu_1 \downarrow \lambda_1} \theta_1(\mu_1) = \theta_1(\lambda_1) = 0 .$$

Note that when $\mu_1 = \lambda_1$ the first $SM/M/1$ queue becomes unstable and $\theta = 0$ would be the only solution of the fixed point equation (9). Similarly,

$$\lim_{\mu_1 \uparrow \mu-\lambda_2} \theta_2(\mu_2) = \lim_{\mu_1 \uparrow \mu-\lambda_2} \theta_2(\mu - \mu_1) = \theta_2(\lambda_2) = 0 .$$

From continuity (of the eigenvalue solution), there exists a value $\mu_1 \in (\lambda_1, \mu - \lambda_2)$ such that

$$\theta_1(\mu_1) = \theta_2(\mu - \mu_1) = \theta_2(\mu_2) =: \theta .$$

Using this value of μ_1, the product of the two (independent) martingales $X_{1,t}$ and $X_{2,t}$ is itself a martingale

$$X_t = h_{1,M_1(t)}(R_1(t))h_{2,M_2(t)}(R_2(t))e^{\theta(A_1(t)+A_2(t)-S(t))} ,$$

where $S(t) = S_1(t) + S_2(t)$ is a Poisson process with rate μ, by the superposition property of Poisson processes.

The key reason for finding μ_1 to guarantee the same θ for the individual martingales $X_{1,t}$ and $X_{2,t}$ is that we could express the martingale X_t in terms of $A_1(t) + A_2(t) - S(t)$; this term drives the queueing process in the original ΣSM/M/1 queue (recall (1)). Therefore, stochastic bounds on Q follow immediately as in Theorem 6:

$$\mathbb{P}(Q \geq k) \leq \rho_1\rho_2 \frac{\sum_i q_{1,i}h_{1,i}(0)}{\inf_t \min_i h_{1,i}(t)} \frac{\sum_i q_{2,i}h_{2,i}(0)}{\inf_t \min_i h_{2,i}(t)} e^{-\theta(k-2)} ,$$

and similarly for the lower bound, where $\rho_i = \frac{\lambda_i}{\mu_i}$; note the exponent $(k-2)$, whereby the '-2' stems from the calculations of $\mathbb{E}[X_{i,0}]$ (see the proof of Theorem 6). Moreover, the additional '$\inf_t$' (to be replaced by '$\sup_t$' for the lower bound) are needed because in the case of two arrival streams the stopping time T can happen on an arrival from a single stream only, i.e., either $R_1(T) = 0$ or $R_2(T) = 0$.

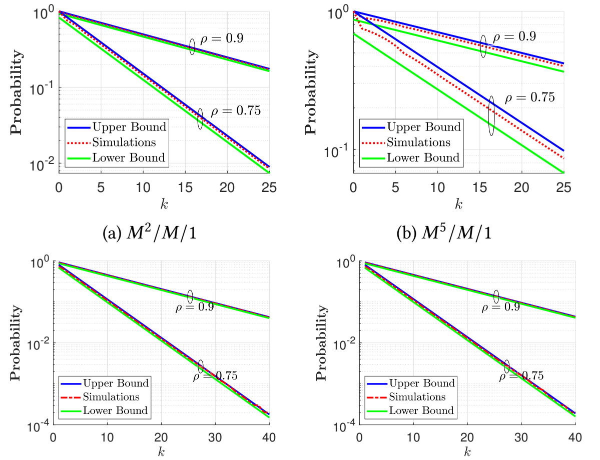
4.2 Numerical Results and Open Problem(s)

Figs. 1.(a-b) illustrate the bounds' accuracy for $M^r/M/1$, with the decay rate θ obtained from (6). While reasonably accurate, the exponential upper bounds from Theorem 4 do not capture the concave-like behavior (on a log-scale) for $k \in \{1, \ldots, r+1\}$, as clearly seen in (b) at 75% utilization[6].

Moving to $SM/M/1$, an intuitive SM but non-renewal process is an alternating renewals (AR) process with states $M(t) \in \{1, 2\}$,

[5] The general case of multiple processes follow similarly.

[6] We observed the same behavior at all the values of r we simulated.



(a) $M^2/M/1$

(b) $M^5/M/1$

(c) AR: Exponential(λ_1), Exponential(λ_2) (d) AR: Weibull(2,1), Exponential(λ_2)

Figure 1: $M^r/M/1$ and $AR/M/1$: bounds on the CCDF $\mathbb{P}(Q \geq k)$ in (a-b) and $\mathbb{P}_{a,2}(Q \geq k)$ in (c-d); $\lambda = 1$ in (a-b) and $\lambda_1 = 1$, $\lambda_2 = 0.1$ in (c-d); $\rho = 0.75,\ 0.9$

alternating distributions $F_1(t)$ and $F_2(t)$, and means $\frac{1}{\lambda_1}$ and $\frac{1}{\lambda_2}$, respectively. The corresponding kernel is

$$\begin{bmatrix} 0 & F_1(x) \\ F_2(x) & 0 \end{bmatrix}$$

and the stationary distributions of the chain M_n and SMP $M(t)$ are

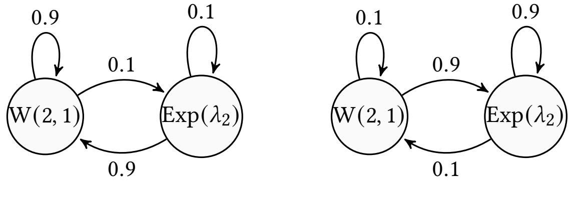
$$q_i = \frac{1}{2} \text{ and } \pi_i = \frac{\lambda_{3-i}}{\lambda_1 + \lambda_2}, i = 1, 2 .$$

The correlation structure can be explained in an extreme scenario with small λ_1 and large λ_2: if an arbitrary interarrival is large then the next is likely small, and vice-versa.

Fig. 1.(c-d) illustrate the bounds' accuracy for $AR/M/1$ scenarios, the latter with a Weibull(2,1) distribution with shape 2 and scale 1 (i.e., $\mathbb{P}(T_1 \leq x) = 1 - e^{-x^2}$) alternating with an Exponential with rate λ_2. The service rate μ depends on the intensity ρ. We consider the bounds $\mathbb{P}_{a,2}(Q \geq k)$, i.e., concerning the queue size just before triggering the Exponential interarrival with rate λ_2 (the bounds are immediate applications of Theorem 6).[7]

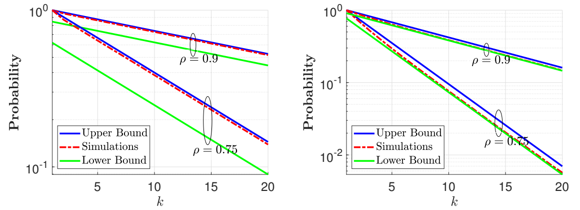
Fig. 3 illustrates the accuracy of the $SM/M/1$ bounds from Theorem 6. The underlying Markov chain has two states (see Fig. 2) and the corresponding kernel is detailed in the caption (the distributions are as in the $AR/M/1$ case). We consider two scenarios depending on which distribution (Weibull or Exponential) is triggered much more frequently than the other: the former in (c), given that $q_{1,2} < q_{2,1}$, and the latter in (d).

Lastly, Fig. 4 illustrates the accuracy of the ΣSM/M/1 bounds; the caption details the precise numerical settings. We only consider the upper bounds to highlight their crucial challenge. In scenarios with burstiness, driven by the underlying Markov chains from Fig. 2.(a,b), the 'true' tail of the queue (on a log-scale) is clearly not a straight line. This behavior has already been apparent in the other scenarios (including the $M^r/M/1$ case) but it is more pronounced in the current ΣSM/M/1 case, subject to the two SMPs with different combinations of burstiness and sojourn times distributions.



(a) **(b)**

Figure 2: The SMP processes used for the numerical evaluations of the $SM/M/1$ and ΣSM/M/1 bounds; the state label denotes the distribution of the sojourn time in that particular state; $W(2,1)$ stands for Weibull distribution with shape 2 and scale 1; Exp(λ_2) stands for exponential distribution with rate λ_2; according to the transition probabilities, $W(2,1)$ is the dominating distribution (i.e., triggered more frequently) in (a), whereas Exp(λ_2) is the dominating one in (b).



(a) $q_{1,2} = 0.1,\ q_{2,1} = 0.9$ (b) $q_{1,2} = 0.9,\ q_{2,1} = 0.1$

Figure 3: SM/M/1: bounds on the CCDF $\mathbb{P}_{a,2}(Q \geq k)$; 2-state Markov chain with transition probabilities $(q_{i,j})$; $F_{1,j} = q_{1,j}\left(1 - e^{-x^2}\right)$, $F_{2,j} = q_{2,j}\left(1 - e^{-\lambda_2 x}\right)$, $\lambda_2 = 0.1$, $\rho = 0.75,\ 0.9$

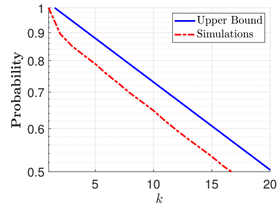


Figure 4: ΣSM/M/1: bounds on the CCDF $\mathbb{P}(Q \geq k)$; the superposed SMPs are those from Fig. 2.(a,b); $\rho = 0.9$

The upper bounds are tight at small k (i.e., $k = 1$); however, as they have an exact asymptotic rate (the 'θ') and their prefactor is independent of k, they follow a straight line which inevitably deviates from the real behavior. In some scenarios, depending on burstiness, the lower bounds which also follow straight lines become tight in the tail (see Fig. 3.(b)). In others, they are consistently loose as apparent from Fig. 3.(a), which is subject to the SMP from Fig. 2.(a) dominated by the Weibull distribution; the same holds in the ΣSM/M/1 case, as it accounts for the same SMP from Fig. 2.(a).

[7] All simulation results are obtained from 10^6 runs, which is sufficient to yield stable results; we omit confidence intervals to avoid clutter.

The crucial and arguably remaining challenge for the bounds' accuracy is to properly capture the initial 'bend' characteristic to the 'true' behavior of the tail. From a technical point of view, the open problem is to more precisely characterize

$$\mathbb{E}\left[h_{M(T)}\left(R(T)\right)1\!\!1_{T<\infty}\right] \ ,$$

where T is the recurring stopping time from all the proofs. In the $GI/M/1$ case, in which there is a single state for the underlying Markov chain, whereas T happens on an arrival (i.e., $R(T)=0$), exact results could be obtained; as a side remark, the assumption of exponential service times is also crucial for the exact results, as there is no need for an additional random function '$h^S()$' to capture the remaining lifetime in the service process due to the memoryless property. In the $AR/M/1$ or $SM/M/1$ cases, in which the Markov chains can have more than one state, or in the ΣSM/M/1 case, in which there is more than a single SMP, the current technique only uses 'rough' and deterministic bounds. In particular, the deterministic nature of these bounds is sufficient to break the above expectation and capture the key metric $\mathbb{E}\left[1\!\!1_{T<\infty}\right]=\mathbb{P}(Q\geq k)$.

5 CONCLUSIONS

We have developed a methodology to construct a wide class of continuous-time martingales, which were used to derive stochastic bounds in several practically motivated queueing systems. The overall methodology is not only an intuitive and much simpler alternative to rediscover the classical $GI/M/1$ exact result but it enables a modular treatment of significantly more complex queues by manipulating martingales from simpler queues. Several extensions to other even more complex queueing systems are possible, e.g., $SM^{\mathrm{X}}/SM/1$ or $\Sigma SM/SM/1$. Moreover, our queueing results retain the expressiveness of the exact $GI/M/1$ result, are asymptotically optimal, and are also computationally light – essentially involving an eigenvalue problem in the number of states $|\mathcal{S}|$ and binary searches. Further improving the bounds' accuracy, including capturing the non-exponential initial behavior (e.g., for $M^{\mathrm{r}}/M/1$ or $\Sigma SM/M/1$), remains an open problem.

REFERENCES

[1] Ivo J. B. F. Adan and Vidyadhar G. Kulkarni. 2006. Single-Server Queue with Markov-Dependent Inter-Arrival and Service Times. *Queueing Syst. Theory Appl.* 54, 1 (2006), 79. https://doi.org/10.1007/s11134-006-9860-1

[2] Nail Akar and Khosrow Sohraby. 2009. System-theoretical algorithmic solution to waiting times in semi-Markov queues. *Perform. Evaluation* 66, 11 (2009), 587–606. https://doi.org/10.1016/j.peva.2009.05.001

[3] E. Arjas. 1972. On the use of a fundamental identity in the theory of semi-Markov queues. *Advances in Applied Probability* 4, 2 (1972), 271–284. https://doi.org/10.2307/1425999

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

[5] Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. 2005. Statistical Analysis of a Telephone Call Center. *J. Amer. Statist. Assoc.* 100, 469 (2005), 36–50. https://doi.org/10.1198/016214504000001808 arXiv:https://doi.org/10.1198/016214504000001808

[6] Erhan Çinlar. 1967. Queues with Semi-Markovian Arrivals. *Journal of Applied Probability* 4, 2 (1967), 365–379. http://www.jstor.org/stable/3212030

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

[8] M. L. Chaudhry and J. G. C. Templeton. 1983. *A First Course in Bulk Queues.* John Wiley and Sons.

[9] Florin Ciucu and Felix Poloczek. 2018. Two Extensions of Kingman's GI/G/1 Bound. *Proc. of the ACM on Measurement and Analysis of Computing Systems - ACM Sigmetrics / IFIP Performance* 2, 3 (Dec. 2018), 43:1–43:33.

[10] Mark E. Crovella and Azer Bestavros. 1996. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. In *ACM SIGMETRICS Conf. Measurement & Modeling of Comput. Syst.* 160–169.

[11] Jos H. A. de Smit. 1995. Explicit Wiener-Hopf factorizations for the analysis of multidimensional queues. In *Advances in Queueing Theory: Theory, Methods and Open Problems (Editor J. H. Dshalalow).* CRC Press, Inc., 293–309.

[12] Jos H. A. de Smit and G. J. K. Regterschot. 1986. A semi-Markov queue with exponential service times. In *Semi-Markov Models: Theory and Applications (Editor J. Janssen).* Springer, 369–382.

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

[14] 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 *Proc. of the 1995 ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems.* 108–115.

[15] D. Gross and C.M. Harris. 1985. *Fundamentals of Queueing Theory (2nd Edition).* John Wiley and Sons.

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

[17] Vidyadhar G. Kulkarni. 2017. *Modeling and Analysis of Stochastic Systems* (3rd ed.). Chapman & Hall, Ltd.

[18] Miron Livny, Benjamin Melamed, and Athanassios K. Tsiolis. 1993. The Impact of Autocorrelation on Queuing Systems. *Management Science* 39, 3 (1993), 322–339. http://www.jstor.org/stable/2632647

[19] Ningfang Mi, Qi Zhang, Alma Riska, Evgenia Smirni, and Erik Riedel. 2007. Performance impacts of autocorrelated flows in multi-tiered systems. *Performance Evaluation* 64, 9 (2007), 1082–1101. https://doi.org/10.1016/j.peva.2007.06.016 Performance 2007.

[20] Marcel F. Neuts. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Dover.

[21] Zbigniew Palmowski and Tomasz Rolski. 1996. A Note on Martingale Inequalities for Fluid Models. *Statistics & Probability Letters* 31, 1 (Dec. 1996), 13–21.

[22] B. Eddy Patuwo, Ralph L. Disney, and Donald C. McNickle. 1993. The Effect of Correlated Arrivals on Queues. *IIE Transactions* 25, 3 (1993), 105–110. https://doi.org/10.1080/07408179308964296

[23] Felix Poloczek and Florin Ciucu. 2014. Scheduling Analysis with Martingales. *Performance Evaluation (Special Issue: IFIP Performance 2014)* 79 (Sept. 2014), 56 – 72.

[24] Felix Poloczek and Florin Ciucu. 2015. Service Martingales: Theory and Applications to the Delay Analysis of Random Access Protocols. In *IEEE Infocom.* 945–953.

[25] Mohammad M. Rashid, Md. J. Hossain, Ekram Hossain, and Vijay K. Bhargava. 2009. Opportunistic spectrum scheduling for multiuser cognitive radio: a queueing analysis. *IEEE Transactions on Wireless Communications* 8, 10 (2009), 5259–5269. https://doi.org/10.1109/TWC.2009.081536

[26] Daniel Revuz and Marc Yor. 1999. *Continuous martingales and Brownian motion* (3rd ed.). Number 293 in Grundlehren der mathematischen Wissenschaften. Springer.

[27] Sheldon M. Ross. 1974. Bounds on the Delay Distribution in GI/G/1 queues. *Journal of Applied Probability* 11, 2 (June 1974), 417–421.

[28] Sheldon M. Ross. 1996. *Stochastic Processes* (2nd ed.). Wiley.

[29] Bhaskar Sengupta. 1990. The semi-markovian queue: theory and applications. *Communications in Statistics. Stochastic Models* 6, 3 (1990), 383–413. https://doi.org/10.1080/15326349908807154

[30] K. Sriram and W. Whitt. 1986. Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data. *IEEE JSAC* SAC-4, 6 (Sep 1986), 833–846.

[31] Pyke Tin. 1985. A Queueing System with Markov-Dependent Arrivals. *Journal of Applied Probability* 22, 3 (1985), 668–677. http://www.jstor.org/stable/3213869

[32] S. R. S. Varadhan. 2007. *Stochastic processes.* Number 16 in Courant Lecture Notes in Mathematics. Courant Institute of Mathematical Sciences, New York.

[33] Peter P. Ware, Thomas W. Page, and Barry L. Nelson. 1998. Automatic Modeling of File System Workloads Using Two-Level Arrival Processes. *ACM Trans. Model. Comput. Simul.* 8, 3 (July 1998), 305–330. https://doi.org/10.1145/290274.290317

A AUXILIARY RESULTS

First we include for completeness some known results needed for our main results.

Definition 7. *An adapted real-valued stochastic process $\{M_t, t\geq 0\}$ is called a local martingale if there exists a sequence of stopping times $\tau_K, K\in\mathbb{N}$, such that $\tau_K\uparrow\infty$ a.s. and $\{M_{t\wedge\tau_K}, t\geq 0\}$ is a martingale for all $K\in\mathbb{N}$.*

Definition 8. *A real value adapted process Y is of class DL if for every $t>0$ the family of random variables Y_τ, where τ ranges* *through all stopping times satisfying* $\mathbb{P}(\tau \le t) = 1$, *is uniformly integrable.*

Theorem 9 (e.g., [26], p. 124). *A local martingale is a martingale if and only if it is of class DL.*

Theorem 10 (Dynkin's Formula (e.g., [32], p. 110). *Let* X *be a homogeneous Markov process with state space* E *with cádlág paths for all* $\omega \in \Omega$ *and transition function* $\{P_t(x, A)\}$. *Let* $\{T(t); t \ge 0\}$ *denote its semigroup* $T(t)f(x) = \int_E f(y)P_t(x, dy)$, $f \in \mathcal{B}_E$ *and* $(A, D(A))$ *its generator. Then, for any bounded function* $g \in D(A)$, *the stochastic process* $\{M_t, t \ge 0\}$ *is an* $\{\mathcal{F}_t^X, t \ge 0\}$ *martingale, where*

$$M_t := g(X_t) - \int_0^t Ag(X_s)ds \, .$$

Next we provide some useful variations of known results concerning the process

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

which plays the role of a martingale for the $SM/M/1$ queue. $A(t)$ an $S(t)$ are counting processes associated to stationary interarrival and service times, respectively. The function $h(t) \equiv h_{M(t)}(R(t))$ was shown to be bounded from above, where $M(t)$ is the Semi-Markov Process and $R(t)$ is the remaining lifetime corresponding to the arrivals. Recall first the definition of $T := \inf\{t : A(t) - S(t) = k\}$.

Lemma 11. *The following limit holds*

$$\lim_{t \to \infty} \mathbb{E}\left[X_t \mathbb{1}_{t<T}\right] = 0 \, .$$

This result enables the parallel derivation of both upper and lower bounds. It first appeared in [9] in a less general setting. We present a (simpler) proof for completeness.

Proof. If the interarrivals T_n form a renewal process then as $t \longrightarrow \infty$ (see, e.g., [28], p. 102)

$$\frac{A(t)}{t} \longrightarrow \frac{1}{\mathbb{E}\left[T_1\right]} \, ,$$

as an immediate consequence of the Strong Law of Large Numbers. The result extends immediately to the (stationary) SMP case, using Birkhoff's Ergodic Theorem. Similarly,

$$\frac{S(t)}{t} \longrightarrow \frac{1}{\mathbb{E}\left[S_1\right]} \, ,$$

and using $\rho < 1$ we obtain that

$$\frac{A(t) - S(t)}{t} \longrightarrow \alpha \, ,$$

for some $\alpha < 0$. This convergence implies that for any $0 < \varepsilon < |\alpha|$ there exists T_ε such that $A(t) - S(t) < t(\alpha + \varepsilon)$ $\forall t > T_\varepsilon$ and hence

$$A(t) - S(t) \longrightarrow -\infty \text{ a.s.}$$

Because $\theta > 0$ it immediately follows that $X_t \mathbb{1}_{t<T} \longrightarrow 0$. Also, $|X_t \mathbb{1}_{t<T}| \le \|h\|_\infty e^{\theta k}$, where the infinite norm $\|h\|_\infty := \sup_t h(t)$ is finite. We can now apply the Dominated Convergence Theorem and conclude that $\lim_t \mathbb{E}\left[X_t \mathbb{1}_{t<T}\right] = \mathbb{E}\left[\lim_t X_t \mathbb{1}_{t<T}\right] = 0$. □

Lemma 12. *The moment generating function (MGF)* $E\left[e^{\theta A(t)}\right]$ *is bounded for all* $\theta > 0$ *and* $t \ge 0$.

Proof. Fix $\theta > 0$ and $t \ge 0$, and assume without loss of generality that there is an arrival at time 0. For clarity, we first give the proof in the renewal case. We can write

$$A(t) = \min\{n : T_1 + T_2 + \cdots + T_{n+1} > t\} \, .$$

Construct the renewal process

$$T_{\alpha,n} := \alpha \mathbb{1}_{T_n \ge \alpha} \; \forall n > 0,$$

for some $\alpha > 0$ such that $p := \mathbb{P}(T_1 \ge \alpha)$ satisfies $b := e^\theta(1-p) < 1$. The corresponding counting process is

$$A_\alpha(t) := \min\{n : T_{\alpha,1} + T_{\alpha,2} + \cdots + T_{\alpha,n+1} > t\} \, .$$

Denoting $k := \lfloor \frac{t}{\alpha} \rfloor$ and noting that

$$\mathbb{P}\left(A_\alpha(t) = n\right) = p^{k+1}(1-p)^{n-k}\binom{n}{n-k} \; \forall n \ge k \, ,$$

we obtain that

$$\mathbb{E}\left[e^{\theta A_\alpha(t)}\right] = e^{\theta k}p^{k+1}\sum_{n \ge 0}\binom{k+n}{n}b^n = e^{\theta k}\left(\frac{p}{1-b}\right)^{k+1} \, ,$$

which is bounded.

Finally, $T_n \ge T_{\alpha,n}$ implies that $A(t) \le A_\alpha(t)$ and therefore the MGF of $A(t)$ is also bounded.

In the SMP case, we ce can write

$$A(t) = \min\{n : U_1 + U_2 + \cdots + U_{n+1} > t\} \, ,$$

where $U_n := T^{M_n}_{v(M_n,n)}$ and $v(M_n, n)$ is the number of visits of state M_n during steps $\{1, 2, \ldots, n\}$.

For each $i \in \mathcal{S}$ construct the renewal processes

$$T^i_{\alpha,n} := \alpha \mathbb{1}_{T^i_n \ge \alpha} \; \forall n > 0,$$

for some $\alpha > 0$. Denote $p_i := \mathbb{P}(T^i_{\alpha,1} = \alpha)$ and assume without loss of generality that $p_1 = \min_{i \in \mathcal{S}} p_i$.

Let now the counting processes

$$A_\alpha(t) := \min\{n : U_{\alpha,1} + U_{\alpha,2} + \cdots + U_{\alpha,n+1} > t\} \, ,$$

where $U_{\alpha,n} := T^{M_n}_{\alpha,v(M_n,n)}$ and

$$A^1_\alpha(t) := \min\{n : T^1_1 + T^1_2 + \cdots + T^1_{n+1} > t\} \, .$$

We can now write

$$\mathbb{P}\left(A^1_\alpha(t) \ge n\right) = \mathbb{P}\left(T^1_{\alpha,1} + \cdots + T^1_{\alpha,n} \le t\right) \ge \mathbb{P}\left(U_{\alpha,1} + \cdots + U_{\alpha,n} \le t\right)$$
$$= \mathbb{P}\left(A_\alpha(t) \ge n\right) \ge \mathbb{P}\left(A(t) \ge n\right)$$

The first inequality follows from the choice of p_1: $T^1_{\alpha,n}$ is stochastically smaller than $U_{\alpha,n}$ for all n. The second follows from $T^i_{\alpha,n} \le T^i_n$.

Therefore, the MGF of $A(t)$ is bounded by the MGF of $A^1_\alpha(t)$, which is subject to a renewal structure. The rest follows as in the renewal case by properly choosing α. □