Age of information using Markov-renewal methods

George Kesidis[*] Takis Konstantopoulos[†] Michael A. Zazanis[‡]

August 14, 2022

Abstract

When designing a message transmission system, from the point of view of making sure that the information transmitted is as fresh as possible, two rules of thumb seem reasonable: use small buffers and adopt a last-in-first-out policy. In this paper, the freshness of information is interpreted as the recently studied "age of information" performance measure. Considering it as a stochastic process operating in a stationary regime, we compute the whole marginal distribution of the age of information for some well-performing systems. We assume that the arrival process is Poisson and that the messages have independent service times with common distribution, i.e., the M/GI model. We demonstrate the usefulness of Palm and Markov-renewal theory to derive results for Laplace transforms. Our numerical studies address some aspects of open questions regarding the optimality of previously proposed scheduling policies, and a policy newly considered herein, for AoI management.

Keywords and phrases. Age of information, Markov renewal, Palm probability, Laplace transform, queueing

1 Introduction

The so-called "age of information" (AoI) performance criterion, defined as the time elapsed since the information possessed by a monitor was generated and time stamped at the source, has recently received a lot of attention, e.g., [16, 17]. The reason is simple: in several applications it is the freshness of information that is important rather than the correct transmission of all messages. Examples include virtual reality, online gaming, weather reports, semi or fully autonomous vehicles, stock market trading, power systems and other "cyber physical" systems. Often in these applications, bounds on the tail of the AoI *distribution* (not just its mean) need to be met.

We start by precisely defining the concept of AoI in general. Consider a message processing facility with one input stream of arriving messages. The facility can be a single queue or a complex network system. An arriving message has a certain positive "size" (expressed in time units and interpreted as processing or service time) and three things can happen: (i) the message is immediately rejected upon arrival; (ii) the message is accepted but rejected while in the system; (iii) the message is successfully transmitted as soon as it is processed in its entirety. We are interested in the time that the latter will happen in comparison to the time that the message arrives in the system. If messages are labeled by integers in a way that the message with label $n \in \mathbb{Z}$ arrives at time $T_n \in \mathbb{R}$ and $T_m < T_n$ when $m < n$, if $T_n + \Delta_n$ denotes the time at which message labeled n leaves the system either by being rejected or successfully transmitted, and if ψ_n is a binary variable indicating the latter ($\psi_n = 1$ if message n is successful or 0 if not), then we let

$$D(t) := \sup\{T_n + \Delta_n : n \in \mathbb{Z}, T_n + \Delta_n \le t, \psi_n = 1\}, \tag{1}$$

$$A(t) := \sup\{T_n : n \in \mathbb{Z}, T_n + \Delta_n \le t, \psi_n = 1\}, \tag{2}$$

[*]gik2@psu.edu; Pennsylvania State University, USA

[†]takiskonst@gmail.com; University of Liverpool, UK; research supported by Cast. Co. IIS-75

[‡]zazanis@aueb.gr; Athens University of Economics and Business, Greece

and define the AoI at time t by

$$\alpha(t) := t - A(t). \tag{3}$$

Note that $A(t) = A(D(t))$. This definition is quite general, that is, it does not depend on the details of the system design.

Typically, systems that adopt freshness of information as performance measure should be designed so that its AoI "be as small as possible". The last phrase can mean several things. For example, it can mean that the quantity $\alpha(t)$ is least for all t under identical traffic conditions. Or it could mean least in terms of an expectation or another functional of the process, e.g., $\mathbb{P}(\alpha > \gamma) < \varepsilon$ for some quality-of-service parameters γ, ε where here "α" has the stationary AoI distribution. Adopting AoI as a performance criterion immediately poses some simplifications over traditional queueing theory performance criteria but also presents some new challenges.

It is reasonable to conjecture that every time a message arrives we should start to process it immediately if possible (after all, we are not interested in obsolete information). That is, even if the server is busy at the moment of arrival, the currently served message is immediately discarded and the new one starts being processed. Systems working in this manner are service-preemptive. It also seems reasonable to serve messages in reverse order of arrival: the most recent message must be served first (LIFO).

One may thus conjecture that LIFO-preemptive (i.e., service preemptive) is "best". But numerical examples and simulations show that this may be false depending on the model assumptions. In particular, it has been shown, that a single buffer system with no service preemption (called $\mathcal{B}_1$ below) has smaller AoI both in expectation and stochastically under particular assumptions on the message size distribution, e.g., [7, 12].

The simplest systems with small-size buffer and no service preemption are defined next. One of them, denoted as $\mathcal{B}_2$, is a single-server FIFO queue with capacity for two messages and blocking. (Thus, an incoming message finding the buffer full is immediately discarded.) The other, denoted as $\mathcal{P}_2$, also works without service preemption. An arriving message under $\mathcal{P}_2$ finding the buffer full displaces or "pushes out" the stored message. See Figure 1 for a typical scenario in $\mathcal{P}_2$.

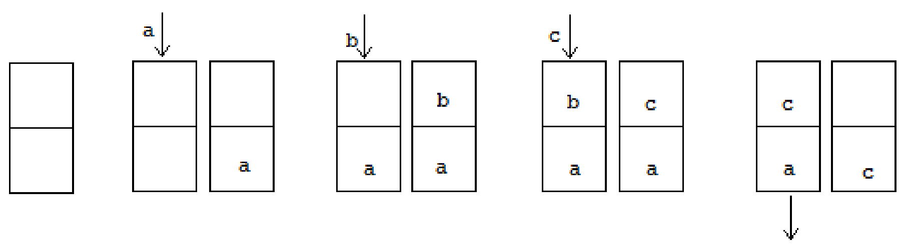

Figure 1: *The $\mathcal{P}_2$ system. The lower cell contains the message being processed (if any). Message a arrives at an empty system and is immediately processed. Message b arrives and is stored in the upper cell. While the buffer still contains a and b, a third message c arrives and immediately kicks b out. When a completes being processed it departs and c moves to the lower cell.*

More generally, we define, for $n \geq 2$, $\mathcal{B}_n$ and $\mathcal{P}_n$ as follows. The $\mathcal{B}_n$ system is a single-server queue with buffer of size n, blocking of any incoming message finding the buffer full, and operating under FIFO (First In First Out) dequeuing policy[1] for $n > 2$. The $\mathcal{P}_n$ system, for $n \geq 2$, works as follows: messages are stored in an order that is reverse to their order of arrival (i.e., LIFO dequeuing policy when $n > 2$)[2]. So if a message is being processed in cell 1 at time t, the message in cell 2 arrived last before t while the

[1]"Dequeuing" policy denotes how the server selects which to serve from among stored messages when it becomes free. FIFO or LIFO dequeuing policies are relevant when more than one message can be stored.

[2]In the $\mathcal{P}_2$ and $\mathcal{B}_2$ systems, there is obviously only at most one message to choose from when the server is free so these systems are both LIFO and FIFO.

message in cell n is the oldest. A new message arriving at a full buffer is always stored in cell 2, displacing the other messages upwards and pushing out the one sitting in cell n (oldest one). For $n \geq 2$, $\mathcal{P}_n$ has no service preemption.

For $n = 1$, the system $\mathcal{B}_1$ is simply a single-buffer blocking queue: an arriving message is immediately rejected if there is a message in the system. The $\mathcal{P}_1$ system is the single-buffer push-out system: an arriving message pushes out the currently processed message and takes its place. Of all systems $\mathcal{P}_n$ and $\mathcal{B}_n$, $n \geq 1$, the $\mathcal{P}_1$ system is special because it is the only one with service preemption.

In some practical scenarios, a LIFO queueing policy is infeasible[3]. Also, in some virtualized computational environments, whether or not the message transmission server is busy may not be observable. That is, an executing AoI-sensitive application may only be able to sense the occupancy of their transmission buffer memory, and send a recently generated message if the buffer memory is empty, otherwise drop the message. That is, buffered messages cannot be replaced, and so the $\mathcal{B}_2$ policy is implemented. Obviously, if the buffered message can be replaced, the $\mathcal{P}_2$ policy can be implemented.

The expectation of AoI for $\mathcal{B}_\infty$ system for the M/M model (i.e., the M/M/1 queue), a system with generally large AoI, was analyzed in [9]. Further work on the expectation of various systems was done in [14]. Other work for LIFO queues, including comparison of different systems, has been done in [10, 6, 3, 11, 7]. In [7], the stationary AoI distribution of the $\mathcal{P}_2$ policy for the M/GI model (among several others) is derived by a sample-path approach focusing on the (locally maximal) AoI at departure times and the system delay of successfully served messages; see their Theorem 10. The stationary AoI distribution of $\mathcal{P}_1$ [7, 12] and $\mathcal{B}_1$ [12] for the GI/GI model has been derived. [16, 17] are broader surveys of articles which address AoI-related issues, e.g., shared queueing systems, not considered herein.

In this paper, we analyze the AoI processes and derive the stationary AoI distribution for $\mathcal{B}_2$ (in Section 3) and $\mathcal{P}_2$ (in Appendix A) for the M/GI model. Though the stationary AoI distribution for $\mathcal{P}_2$ for the M/GI model was already derived in [7], we herein use a different approach (which also works for $\mathcal{P}_n$ LIFO systems when $n > 2$). In the following, we use the classical embedding technique, valid for queueing systems assuming Poisson arrivals, given that the system sampled at certain epochs (the departures of successful messages in our case) has a Markovian property, see, e.g., [4]. Some numerical comparisons for the $\mathcal{P}_n$ and $\mathcal{B}_n$ systems for the M/GI model, for $n = 1, 2$, are given in Section 5.

When the message sizes are i.i.d. exponential, having (for notational convenience) rate $\mu = 1$, we shall show (as a corollary) that, in steady-state, the value of AoI at some (and hence any) point of time, has density

$$f_{\mathcal{B}_2}(t) = c\,(q(t)e^{-t} + e^{-\lambda t}), \tag{4}$$

$$f_{\mathcal{P}_2}(t) = q_1(t)e^{-t} + q_2(t)e^{-(\lambda+1)t} - \frac{\lambda}{\lambda - 1}e^{-\lambda t}, \tag{5}$$

in the $\mathcal{P}_2$ and $\mathcal{B}_2$ cases, respectively, where

$$c = \frac{\lambda}{(\lambda^2+\lambda+1)(\lambda-1)^2}, \quad q(t) = \tfrac{1}{2}\lambda(\lambda-1)^2t^2 + \lambda(\lambda-1)t - 1,$$

$$q_1(t) = \frac{(\lambda^3+\lambda^2-2\,\lambda)t - \lambda^2 + \lambda + 3}{(\lambda^2+\lambda+1)(\lambda-1)}, \quad q_2(t) = \frac{(\lambda^2+\lambda)t + \lambda^2 + 3\,\lambda + 3}{\lambda^2+\lambda+1},$$

when $\lambda \neq 1$; while, for $\lambda = 1$, the densities become

$$f_{\mathcal{B}_2}(t) = \tfrac{1}{3}(t^2+t)e^{-t}, \quad f_{\mathcal{P}_2}(t) = \tfrac{1}{3}(7+2t)e^{-2t} + \tfrac{1}{3}(6t-7)e^{-t}.$$

(For general μ, simply replace λ by λ/μ and t by μt in the foregoing expressions.)

[3] At the "network layer", the Internet performs "in order" (FIFO) delivery of packets by rule.

In Section 6, *time-reversibility* is employed for the $\mathcal{P}_3$ system[4]. Again note that $\mathcal{P}_n$ systems for $n \geq 3$ employ a LIFO policy when selecting (successful) messages to enter the server and begin transmission. The same approach can be applied to $\mathcal{P}_n$ (and $\mathcal{B}_n$) for any positive integer $n > 2$ to support the conjecture that the stationary AoI of $\mathcal{P}_n$ is stochastically increasing in n for $n \geq 2$.

In Section 7, we show how the Markov embedding can be used to compute the stationary AoI for a more complex causal push-out policy, $\mathcal{P}_{2,\theta}$ with parameter $\theta \geq 0$, which can achieve lower mean AoI than $\mathcal{P}_1$ or $\mathcal{P}_2$ for some cases of service-time distribution.

In Section 8, we conclude with a short summary discussion including future work.

2 Basic framework

Under our assumptions, and because we consider finite buffers, it holds that there is a unique stationary version of the stochastic process α in all systems considered. We will not offer any reasons for this technical result, but only point out that even existence may not hold if the arrival and message size processes are neither independent nor renewal, and point out the difficulties by referring to [2]. We shall always be considering the stationary version. Hence $\alpha(t)$ has the same distribution for all t which we are interested in describing. We note that computing the expectation is, in general, not that much easier than deriving the whole distribution. We also note that deriving the distribution is essential in case that we are interested not just in maintaining a low AoI on the average but also in maintaining the tail of the probability distribution small.

Throughout the paper, we let λ be the rate of the (Poisson) arrival process and G the distribution of a typical message size σ, a random variable that is positive with probability 1 and has finite expectation denoted by $1/\mu$. We thus only assume that $\lambda > 0$ and $\mu > 0$ (but σ may have infinite variance). It is assumed that the message sizes are i.i.d. copies of σ and independent of the arrival process. The ratio $\rho = \lambda/\mu$ is referred to as traffic intensity.

We discuss the technique used in the analysis for all systems described in the introduction from the point of view of the distribution of the age of information. By this phrase, we will always mean that the age of information process $\alpha(t)$, $t \in \mathbb{R}$, is stationary and that we shall be interested in the distribution of $\alpha(t)$ for some, and hence all, t which will be taken to be the point $t = 0$. The goal is to derive a "fixed point equation" for $\alpha(0)$, see equation (11) below. The arrival process is always taken to be a Poisson process on $\mathbb{R}$ (=time) with rate λ. As mentioned above, T_n is the arrival time of message labeled n. Its size is σ_n. The collection of message sizes are i.i.d. and independent of the arrival process. Let

$$G(x) = \mathbb{P}(\sigma_1 \leq x)$$

be the distribution function of the typical size and let $1/\mu$ be its expectation, assumed to be finite. Also assume that $G(0) = 0$. Abusing notation, we shall let G denote the probability measure defined by the function $G(x)$ and by $\hat{G}(s)$ its Laplace-Stieltjes transform:

$$\hat{G}(s) = \int_0^\infty e^{-sx} dG(x).$$

Recall that $T_n + \Delta_n$ is defined as the time at which message n departs either because it was pushed out or rejected or because it was successfully processed ($\psi_n = 1$ in the latter case). See discussion around (1) and (2) where these symbols were introduced. Then the number of messages in the system at time $t \in \mathbb{R}$ is given by

$$\xi(t) = \sum_{n \in \mathbb{Z}} \mathbf{1}_{T_n \leq t < T_n + \Delta_n}.$$

[4] Time reversibility is also employed when using Markov embeddings to similarly study GI/M models which are not addressed herein.

Note that if the message is immediately rejected then $\Delta_n = 0$ and so this message does not contribute to ξ. We let

$$\{S_m : m \in \mathbb{Z}\} := \{T_n + \Delta_n : n \in \mathbb{Z}, \psi_n = 1\},$$

and, thinking of the two sets as sequences, $\{S_m\}$ is a subsequence of $\{T_n + \Delta_n\}$ and is enumerated so that $S_{m_1} < S_{m_2}$ if $m_1 < m_2$. We note that ξ is right-continuous for all t. Recalling the notions of Markov renewal and semi-Markov processes, see, e.g., [1, VII.4], our first observation is:

Lemma 1. *For both $\mathcal{B}_2$ and $\mathcal{P}_2$ cases, the process $\xi(t)$, $t \in \mathbb{R}$, is a semi-Markov process [4, Ch.10] with respect to the points S_n, $n \in \mathbb{Z}$. Moreover, the distribution of ξ is the same in both $\mathcal{B}_2$ and $\mathcal{P}_2$ cases.*

This follows easily by standard arguments in queueing theory, for instance in the analysis of a queue with Poisson arrivals; see, e.g., [4, Ch. 6, Sec. 5]. Thus, ξ does not "see" the difference between $\mathcal{B}_2$ and $\mathcal{P}_2$. The distinction between these two will become important in the next section when we discuss the details about α in each case.

We further assume that the arrival process together with the process ξ are stationary under a probability measure $\mathbb{P}$. (This assumption is non-vacuous; we shall not elaborate on this further but refer the reader to [2] for an exposition of techniques used to establish it.)

We refer to the intervals $[S_n, S_{n+1})$ as *segments* and split the paths of ξ into a union of paths over segments. (See Figure 2.) By convention, we assume that the segment labelled 0 contains the point $t = 0$. Denote by $\mathbb{P}^0$ the Palm probability of $\mathbb{P}$ with respect to the point process $\{S_n\}$. We refer to [2] for this

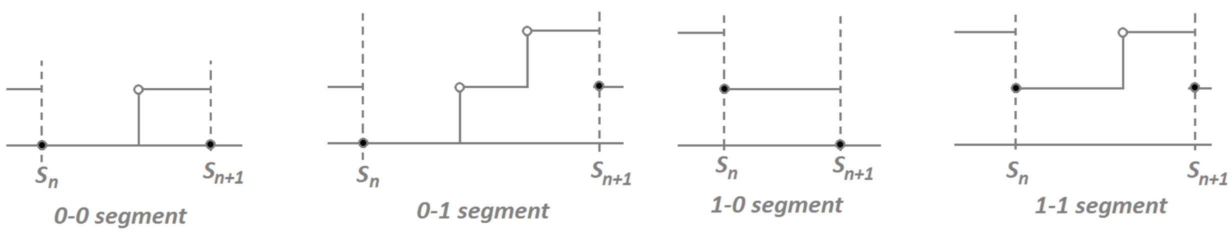

Figure 2: *What $\xi(t)$ looks like when $S_n \le t < S_{n+1}$, regardless of the policy used, where S_n denotes the departure time of message with label n provided that it is successful.*

concept. Intuitively, $\mathbb{P}^0$ is $\mathbb{P}$ conditional on the event that $0 \in \{S_n, n \in \mathbb{Z}\}$. Hence $\mathbb{P}^0(S_0 = 0) = 1$. Let

$$K_n := \xi(S_n), \quad n \in \mathbb{Z}.$$

The sequence $\{K_n\}$ is a Markov chain with state space $\{0, 1\}$ while $\{(S_n, K_n)\}$ is the Markov renewal sequence [1, 4] associated with the semi-Markov process ξ. The latter has transition kernel

$$Q_{ij}(x) := \mathbb{P}^0(S_{n+1} - S_n \le x, K_{n+1} = j \,|\, K_n = i), \quad i, j \in \{0, 1\},$$

explicitly given by

$$\begin{bmatrix} Q_{00}(x) & Q_{01}(x) \\ Q_{10}(x) & Q_{11}(x) \end{bmatrix} = \begin{bmatrix} \int_0^x \left(1 - e^{-\lambda(x-u)}\right) e^{-\lambda u} dG(u) & \int_0^x \left(1 - e^{-\lambda(x-u)}\right)\left(1 - e^{-\lambda u}\right) dG(u) \\ \int_0^x e^{-\lambda u} dG(u) & \int_0^x \left(1 - e^{-\lambda u}\right) dG(u) \end{bmatrix}, \tag{6}$$

as follows easily by considering the cases of Figure 2. Letting $x \to \infty$ in (6) we obtain the transition matrix for the Markov chain $\{K_n\}$,

$$\begin{bmatrix} Q_{00}(\infty) & Q_{01}(\infty) \\ Q_{10}(\infty) & Q_{11}(\infty) \end{bmatrix} = \begin{bmatrix} \hat{G}(\lambda) & 1 - \hat{G}(\lambda) \\ \hat{G}(\lambda) & 1 - \hat{G}(\lambda) \end{bmatrix},$$

from which it is evident that $K_n, n \geq 1$, is not just Markovian but also a sequence of independent Bernoulli random variables with

$$\mathbb{P}^0(K_n = 0) = \hat{G}(\lambda) = 1 - \mathbb{P}^0(K_n = 1). \tag{7}$$

Figure 2 shows the four different types of segments depending on the values of K_n and K_{n+1}. We next define

$$\Phi_i(s) := \mathbb{E}^0[e^{-s(S_1-S_0)}|K_0 = i],$$

and, using the kernel (6), we obtain

$$\Phi_0(s) = \int_0^\infty e^{-sx} dQ_{00}(x) + \int_0^\infty e^{-sx} dQ_{01}(x) = \frac{\lambda}{\lambda+s}\hat{G}(s), \tag{8}$$

$$\Phi_1(s) = \int_0^\infty e^{-sx} dQ_{10}(x) + \int_0^\infty e^{-sx} dQ_{11}(x) = \hat{G}(s). \tag{9}$$

From (8), (9), and (7) we obtain the Laplace transform of the segment length:

$$\Phi(s) \ := \ \mathbb{E}^0[e^{-s(S_1-S_0)}] \ = \ \left(1 - \hat{G}(\lambda) + \hat{G}(\lambda)\frac{\lambda}{\lambda+s}\right)\hat{G}(s).$$

From this, we obtain the mean length of a segment as

$$\mathbb{E}^0[S_1 - S_0] \ = \ \frac{1}{\mu} + \frac{\hat{G}(\lambda)}{\lambda}. \tag{10}$$

We shall henceforth use the abbreviation $\mathbb{E}(X; A)$ for the expectation of a random variable X on the event A, that is, the quantity $\mathbb{E}(X\mathbf{1}_A)$. The following result depends entirely on the semi-Markov property of ξ.

Proposition 1. *For both $\mathcal{B}_2$ and $\mathcal{P}_2$ cases, the random variable $\alpha(0)$ satisfies*

$$\mathbb{E}[e^{-s\alpha(0)}] \ = \ \frac{\lambda}{s} \cdot \frac{\mathbb{E}^0[e^{-s\alpha(0)}; K_0 = 0]\left(1 - \frac{\lambda}{\lambda+s}\hat{G}(s)\right) + \mathbb{E}^0[e^{-s\alpha(0)}; K_0 = 1]\left(1 - \hat{G}(s)\right)}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}. \tag{11}$$

Proof. The Palm inversion formula [1, 2] applied to the $\mathbb{P}$-stationary process α gives

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\mathbb{E}^0[\int_{S_0}^{S_1} e^{-s\alpha(t)}dt]}{\mathbb{E}^0[S_1 - S_0]}. \tag{12}$$

Take a look at (3) and notice that the process α is right-continuous. Its set of discontinuities is $\{S_n\}$. Moreover, it increases at unit rate on each segment:

$$\alpha(t) = \alpha(S_n) + t - S_n, \quad \text{for } t \in [S_n, S_{n+1}). \tag{13}$$

To see this, notice that, for $S_n \leq t < S_{n+1}$, we have $D(t) = D(S_n) = S_n$, by the definition of D in (1), and so $A(D(t)) = A(D(S_n)) = A(S_n)$. Since, from the definition (3), $\alpha(t) = t - A(D(t))$ for all t, we have

$$\alpha(t) = t - A(S_n),$$

whenever $S_n \leq t < S_{n+1}$. Writing this for $t = S_n$, we have

$$\alpha(S_n) = S_n - A(S_n),$$

and so (13) is obtained by subtracting the last two displays. In particular, $S_0 = 0$ and $\alpha(t) = \alpha(0) + t$ for $t \in [S_0, S_1)$, $\mathbb{P}^0$-a.s. Hence, for $i = 0, 1$,

$$\mathbb{E}^0\left[\int_{S_0}^{S_1} e^{-s\alpha(t)}dt; K_0 = i\right] = \mathbb{E}^0\left[e^{-s\alpha(0)}\,\frac{1-e^{-sS_1}}{s}; K_0 = i\right] = \frac{1-\Phi_i(s)}{s}\mathbb{E}^0[e^{-s\alpha(0)}; K_0 = i], \tag{14}$$

where the last equality follows from the fact that $\alpha(0)$ and $S_1 - S_0$ are conditionally independent given $\{K_0 = i\}$, a consequence of the semi-Markov structure of the process $\{\xi(t)\}$, see Lemma 1. Using expressions (8) and (9) and adding the terms in (14) we obtain the numerator of (12). The denominator is given by (10). This shows the validity of (11). □

Remark 1. See the related proof of Proposition 1 of [8]. It should be clear that (11) holds for a much larger class of systems with one (or several independent) Poisson arrival process(es). Also, though specialized for the $n = 2$ case, (11) is based on an expression given by Palm's theorem that works for any $n \geq 2$, as more clearly indicated by the statements of Propositions 2 and 3 below. For example, we may define $\mathcal{B}_n$ to be an extension of $\mathcal{B}_2$ when the buffer has n cells where messages are stored according to the order of their arrivals and a message arriving to a full buffer is immediately rejected (the so-called M/GI/1/n queue). On the other hand, we may define $\mathcal{P}_n$ to be an extension of $\mathcal{P}_2$: messages are stored in an order that is reverse to their order of arrival; so if there is a message being processed in cell 1 at time t, the message in cell 2 arrived last before t while the message in cell n is the oldest; a new message arriving at a full buffer is always stored in cell 2, displacing the other messages upwards and expels the one sitting in cell 1 (oldest one). In both $\mathcal{B}_n$ and $\mathcal{P}_n$, the process ξ is semi-Markov and thus Proposition 1, depending only on this semi-Markov property, applies and formula (11) holds. In fact, one can assert that the proposition holds for networks with i.i.d. message sojourn times, e.g., due to a single bottleneck server. We shall not attempt to formalize this further in this paper.

3 The $\mathcal{B}_2$ system

Recall that the $\mathcal{B}_2$ system is the same as a single server queue with buffer size 2. Under our Poisson assumption for the arrival process and i.i.d. assumptions for message sizes, this is further denoted by M/GI/1/2 in standard queueing terminology. We are, however, interested not in the number of messages in the system neither on message delays but, rather, on the age of information process α. Assuming that α is stationary, we compute the Laplace transform of $\alpha(0)$ under $\mathbb{P}$ by using (11) which requires knowledge of $\mathbb{E}^0[e^{-s\alpha(0)}; K_0 = j]$, $j = 0, 1$. To obtain the latter, we consider the segment $[S_{-1}, S_0)$, further condition on K_{-1}, and summarize the results in Lemma 2 below. In what follows, we let τ, σ, be two independent random variables, where τ is exponential with rate λ and σ has distribution G.

Lemma 2. *For* $\mathcal{B}_2$,

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 0] = \hat{G}(\lambda)\hat{G}(s+\lambda), \tag{15}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 1] = \hat{G}(\lambda)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right), \tag{16}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 0] = \frac{\lambda}{\lambda - s}\left(\hat{G}(s) - \hat{G}(\lambda)\right)\hat{G}(s+\lambda), \tag{17}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 1] = \frac{\lambda}{\lambda - s}\left(\hat{G}(s) - \hat{G}(\lambda)\right)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right). \tag{18}$$

Proof. Recall that the K_n are i.i.d. with distribution (7): $\mathbb{P}^0(K_n = 0) = \hat{G}(\lambda)$. We shall consider the four cases separately and, in each case, we shall be referring to the definition (3) to figure out what $\alpha(0)$ is.

Case 1. $K_{-1}=0, K_0=0$. Observe $\alpha(0)=S_0-T_0$, see Figure 3. But

$$\mathbb{P}^0(S_0-T_0\in dx \mid K_0=0,K_{-1}=0) \;=\; \mathbb{P}(\sigma\in dx\mid \sigma<\tau) \;=\; \frac{e^{-\lambda x}dG(x)}{\hat G(\lambda)},$$

and so

$$\mathbb{E}^0[e^{-s\alpha(S_0)};K_{-1}=0,K_0=0] \;=\; \int_0^\infty e^{-sx}\frac{e^{-\lambda x}}{\hat G(\lambda)}dG(x)\left(\hat G(\lambda)\right)^2 \;=\; \hat G(\lambda)\hat G(s+\lambda).$$

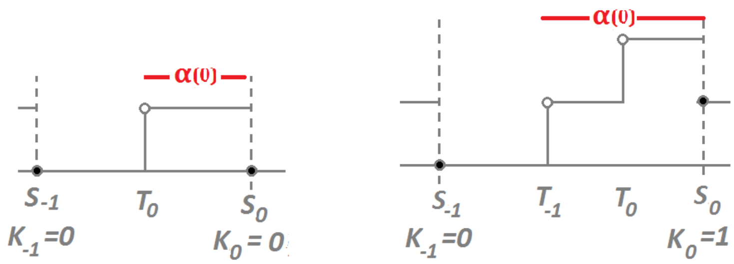

Figure 3: *The segment $[S_{-1},S_0)$ when $K_{-1}=K_0=0$ (left) and $K_{-1}=0$, $K_0=1$ (right), with $S_0=0$.*

Case 2. $K_{-1}=0, K_0=1$. We have $\alpha(0)=S_0-T_{-1}$, see Figure 3. Since

$$\mathbb{P}^0(S_0-T_{-1}\in dx \mid K_0=1,K_{-1}=0) \;=\; \mathbb{P}(\sigma\in dx\mid \tau<\sigma) \;=\; \frac{\left(1-e^{-\lambda x}\right)dG(x)}{1-\hat G(\lambda)}$$

we obtain

$$\begin{aligned}\mathbb{E}^0[e^{-s\alpha(S_0)};K_{-1}=0,K_0=1] \;&=\; \mathbb{P}^0(K_{-1}=0,K_0=1)\mathbb{E}^0[e^{-s\alpha(S_0)}\mid K_{-1}=0,K_0=1]\\ &=\; \hat G(\lambda)\int_0^\infty e^{-sx}\left(1-e^{-\lambda x}\right)dG(x) \;=\; \hat G(\lambda)\left(\hat G(s)-\hat G(s+\lambda)\right).\end{aligned}$$

Case 3. $K_{-1}=1, K_0=0$. To figure out $\alpha(0)$ we are here forced to consider two consecutive segments. We then have

$$\alpha(0)=(S_0-S_{-1})+(S_{-1}-T_0)\,,$$

see Figure 4. Note that S_0-S_{-1} and $S_{-1}-T_0$ are conditionally independent given $K_{-1}=1$ and $\{S_{-1}-T_0\in dx; K_{-1}=1\}$ is independent of K_{-2} with $\mathbb{P}^0(S_0-S_{-1}\in dx;K_0=0\mid K_{-1}=1)=\mathbb{P}(\sigma\in dx;\sigma<\tau)$ and $\mathbb{P}^0(S_{-1}-T_0\in dx;K_{-1}=1\mid K_{-2}=0)=\mathbb{P}(\sigma-\tau\in dx;\sigma>\tau)$, respectively. Thus

$$\begin{aligned}\mathbb{E}^0[e^{-s\alpha(0)};K_{-1}=1,K_0=0] \;&=\; \mathbb{E}[e^{-s\sigma};\sigma<\tau]\,\mathbb{E}[e^{-s(\sigma-\tau)};\sigma>\tau]\\ &=\; \mathbb{E}[e^{-s\sigma}e^{-\lambda\sigma}]\,\mathbb{E}[e^{-s\sigma}\int_0^\sigma \lambda e^{-(\lambda-s)t}dt] \;=\; \hat G(s+\lambda)\frac{\lambda}{\lambda-s}\left(\hat G(s)-\hat G(\lambda)\right).\end{aligned}$$

Case 4. Again, we have to consider two consecutive segments to realize that

$$\alpha(0)=S_0-T_{-1}=(S_{-1}-T_{-1})+(S_0-S_{-1})\,,$$

see Figure 5. The two random variables

$$(S_{-1}-T_{-1}) \text{ and } (S_0-S_{-1}) \text{ are conditionally independent given that } K_{-1}=1$$

and thus

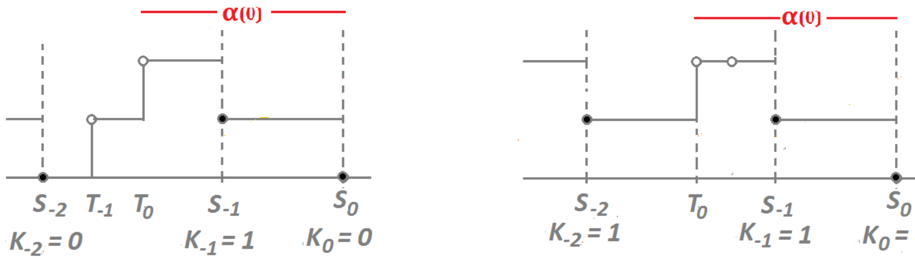

Figure 4: *The segments* $[S_{-2}, S_{-1})$, $[S_{-1}, S_0)$ *when* $K_{-1} = 1, K_0 = 0$ *in the case where* $K_{-2} = 0$ *(left) and* $K_{-2} = 1$ *(right).*

Figure 5: *The segments* $[S_{-2}, S_{-1})$, $[S_{-1}, S_0)$ *when* $K_{-1} = 1, K_0 = 1$ *in the case where* $K_{-2} = 0$ *(left) and* $K_{-2} = 1$ *(right).*

$$\begin{aligned}
\mathbb{E}^0[e^{-s\alpha(0)}; K_{-1} = 1, K_0 = 1] \;&=\; \mathbb{E}^0[e^{-s(S_{-1}-T_{-1})-s(S_0-S_{-1})}; K_{-1} = 1, K_0 = 1] \\
= \mathbb{E}[e^{-s(\sigma-\tau)}; \sigma > \tau]\, \mathbb{E}[e^{-s\sigma}; \sigma > \tau] \;&=\; \mathbb{E}\left[\int_0^\sigma e^{-s(\sigma-t)} \lambda e^{-\lambda t} dt\right] \mathbb{E}\left[e^{-s\sigma}\left(1 - e^{-\lambda\sigma}\right)\right] \\
= \frac{\lambda}{\lambda - s}\left(\hat{G}(s) - \hat{G}(\lambda)\right)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right).&
\end{aligned}$$

This completes the proof. □

Define

$$\hat{G}_I(s) = \frac{1 - \hat{G}(s)}{s}\mu. \tag{19}$$

This is the Laplace transform of a probability measure G_I that is well-known in renewal theory: If we consider a renewal process with points, say, Z_n, $n \in \mathbb{Z}$, such that $Z_0 = 0$ and $Z_{n+1} - Z_n$ having distribution G, then it has a stationary version (with no point at 0) and in such a way that Z_1 has distribution G_I.

Theorem 1. *For* $\mathcal{B}_2$*, the Laplace transform of the stationary Age of Information is given by*

$$\mathbb{E}[e^{-s\alpha(0)}] \;=\; \hat{G}(s)\left(\hat{G}(\lambda) + \lambda\frac{\hat{G}(s) - \hat{G}(\lambda)}{\lambda - s}\right)\left(\frac{\hat{G}(\lambda)}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}\frac{\lambda}{\lambda + s}\frac{\hat{G}(s+\lambda)}{\hat{G}(\lambda)} + \frac{\frac{\lambda}{\mu}}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}\hat{G}_I(s)\right). \tag{20}$$

Proof. Summing (15) and (17) and summing (16) and (18) we obtain

$$\begin{aligned}
\mathbb{E}^0[e^{-s\alpha(S_0)}; K_0 = 0] \;&=\; \hat{G}(s+\lambda)\left[\frac{\lambda}{\lambda - s}\hat{G}(s) - \frac{s}{\lambda - s}\hat{G}(\lambda)\right] \\
\mathbb{E}^0[e^{-s\alpha(S_0)}; K_0 = 1] \;&=\; \left(\hat{G}(s) - \hat{G}(s+\lambda)\right)\left[\frac{\lambda}{\lambda - s}\hat{G}(s) - \frac{s}{\lambda - s}\hat{G}(\lambda)\right].
\end{aligned}$$

Substituting the last two lines into the right hand side of (11) we obtain

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\hat{G}(s)\left[\frac{\lambda}{\lambda-s}\hat{G}(s) - \frac{s}{\lambda-s}\hat{G}(\lambda)\right]\left(\frac{\hat{G}(s+\lambda)}{s+\lambda} + \frac{1-\hat{G}(s)}{s}\right)}{\frac{1}{\lambda}\left(\frac{\lambda}{\mu} + \hat{G}(\lambda)\right)}, \tag{21}$$

which gives (20) if we take into account the definition of $\hat{G}_I$. □

Corollary 1. *Expression (20) gives the stationary AoI as a sum of three independent random variables. In particular, the middle term in the right hand side of (20) corresponds to the Laplace transform of the random variable $(\sigma - \tau)^+$. Moreover, the expectation of $\alpha(0)$ is given by*

$$\mathbb{E}[\alpha(0)] = \frac{1}{\mu} + \left(\frac{1}{\mu} - \frac{1-\hat{G}(\lambda)}{\lambda}\right) + \frac{\hat{G}(\lambda) - \lambda\hat{G}'(\lambda) + \frac{1}{2}\lambda^2\hat{G}''(0)}{\lambda\left(\frac{\lambda}{\mu} + \hat{G}(\lambda)\right)}. \tag{22}$$

We obtained this corollary directly from the Laplace transform (20) where we recognize that $\frac{1}{2}\mu\int_0^\infty x^2 dG(x) = \int_0^\infty x dG_I(x)$. Notice that if the message size has high variance then so does $\mathbb{E}\alpha(0)$. In particular, $\mathbb{E}\alpha(0) = \infty$ if $\int x^2 dG(x) = \infty$. Rather than seeing this as a problem, one could change the point of view and adopt another function of α as a performance measure, for instance, $\mathbb{E}\alpha(0)^p$ for some $p < 1$.

Corollary 2. *For $\mathcal{B}_2$, with G being exponential with mean $1/\mu$ we have*

$$\mathbb{E}[e^{-s\alpha(0)}] = \left(\frac{\mu}{s+\mu}\right)^3 \frac{\lambda}{s+\lambda} \frac{s^2 + 2s(\lambda+\mu) + \lambda^2 + \lambda\mu + \mu^2}{(\lambda^2 + \lambda\mu + \mu^2)}, \tag{23}$$

$$\mathbb{E}[\alpha(0)] = \frac{3\lambda^3 + 2\lambda^2\mu + 2\lambda\mu^2 + \mu^3}{\lambda\mu\,(\lambda^2 + \lambda\mu + \mu^2)}. \tag{24}$$

Inverting the Laplace transform (23) gives density (4), i.e., when $\mu = 1$: if $\lambda \neq 1$,

$$f_{\mathcal{B}_2}(t) = c\,(q(t)e^{-t} + e^{-\lambda t}),$$

where

$$c = \frac{\lambda}{(\lambda^2 + \lambda + 1)(\lambda - 1)^2}, \quad q(t) = \tfrac{1}{2}\lambda(\lambda-1)^2 t^2 + \lambda(\lambda-1)t - 1,$$

else if $\lambda = 1$,

$$f_{\mathcal{B}_2}(t) = \tfrac{1}{3}(t^2 + t)e^{-t}$$

Interestingly, as $\lambda \to \infty$ we immediately see from (23) that $\mathbb{E}[e^{-s\alpha(0)}] \to (\mu/(s+\mu))^3$, the Laplace transform of the sum of three i.i.d. exponentials. See Appendix C for an explanation.

4 The $\mathcal{P}_2$ system

We remind the reader that $\mathcal{P}_2$ differs from $\mathcal{B}_2$ in that the arriving message is always admitted by replacing the message (if any) sitting in the second cell of the buffer, see Figure 1. Again, $\mathcal{P}_2$ is not service-preemptive: once a message starts being processed it will not be interrupted. The strategy for obtaining the Laplace transform of $\alpha(0)$ is the same as before. The following theorem, whose proof is given in Appendix A, was derived in [7] by different means.

Theorem 2. *For $\mathcal{P}_2$, the Laplace transform of the stationary Age of Information is given by*

$$\mathbb{E}[e^{-s\alpha(0)}] = \hat{G}(s)\left(\hat{G}(\lambda) + \lambda\frac{1-\hat{G}(s+\lambda)}{s+\lambda}\right)\left(\frac{\hat{G}(\lambda)}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}\frac{\lambda}{\lambda+s}\frac{\hat{G}(s+\lambda)}{\hat{G}(\lambda)} + \frac{\frac{\lambda}{\mu}}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}\hat{G}_I(s)\right) \tag{25}$$

Corollary statements for deterministic or exponential service-time distributions are also given in the Appendix. The mean AoI derived from Theorem 2 is also consistent with [14].

Corollary 3. *In expression (25) we recognize that $\alpha(0)$ is equal in distribution to the sum of three independent random variables. In particular, the middle term corresponds to the Laplace transform of the random variable $\tau\mathbf{1}_{\tau\leq\sigma}$. Moreover,*

$$\mathbb{E}[\alpha(0)] \; = \; \frac{1}{\mu} + \frac{1}{\lambda}\left(1 - \hat{G}(\lambda) + \lambda\hat{G}'(\lambda)\right) + \frac{\hat{G}(\lambda) - \lambda\hat{G}'(\lambda) + \frac{1}{2}\lambda^2\hat{G}''(0)}{\lambda\left(\frac{\lambda}{\mu} + \hat{G}(\lambda)\right)} \tag{26}$$

Recall that a real random variable X is stochastically smaller than Y, and write $X \leq_{\mathrm{st}} Y$, if

$$\mathbb{P}(X > u) \leq \mathbb{P}(Y > u) \text{ for all } u \in \mathbb{R}.$$

Note that stochastic ordering is a partial order in the space of probability measures on the real line so two random variables may not be comparable at all.

The following simple lemma compares the middle terms of $\mathcal{B}_2$ and $\mathcal{P}_2$ to show that AoI is stochastically larger under $\mathcal{B}_2$ (consistent with the elementary coupling argument).

Lemma 3. *If τ is exponentially distributed and independent of σ, then*

$$\tau\mathbf{1}_{\tau\leq\sigma} \leq_{st} (\sigma - \tau)^+.$$

Proof. For all $a > 0$,

$$\mathbb{P}((\sigma-\tau)^+ > a) = \int_a^\infty \mathbb{P}(\tau < x - a)dG(x) = \int_a^\infty (1 - \mathrm{e}^{-\lambda(x-a)})dG(x)$$
$$\mathbb{P}(\tau\mathbf{1}_{\tau\leq\sigma} > a) = \int_a^\infty (\mathrm{e}^{-\lambda a} - \mathrm{e}^{-\lambda x})dG(x)$$

Since for all $a > 0$ and $x > a$, $\mathrm{e}^{-\lambda a} - \mathrm{e}^{-\lambda x} = \mathrm{e}^{-\lambda a}(1 - \mathrm{e}^{-\lambda(x-a)}) \leq 1 - \mathrm{e}^{-\lambda(x-a)}$, the lemma is proved. □

5 Numerical comparisons

Slightly abusing notation, we write $\alpha_{\mathcal{P}_n}$ instead of $\alpha_{\mathcal{P}_n}(0)$; to further simplify life, we shall now use normalized units, assuming $\mu = 1$. Following [14, 7, 12] and the above calculatons, we now summarize observations regarding $\mathcal{P}_n, \mathcal{B}_n$, $n = 1, 2$.

5.1 Recalling formulas for $\mathcal{P}_1$ and $\mathcal{B}_1$

Concerning the $\mathcal{P}_1$ system we have, from [12, Corollary 4] (see also [7]),

$$\mathbb{E}[e^{-s\alpha_{\mathcal{P}_1}}] = \frac{\rho\hat{G}(s+\rho)}{s+\rho\hat{G}(s+\rho)}, \quad \mathbb{E}[\alpha_{\mathcal{P}_1}] = \frac{1}{\rho\hat{G}(\rho)}. \tag{27}$$

On the other hand, for $\mathcal{B}_1$, [12, Corollary 9(i)], gives

$$\mathbb{E}[e^{-s\alpha_{\mathcal{B}_1}}] = \frac{\rho}{1+\rho}\cdot\frac{(s+\rho-\rho\hat{G}(s))\hat{G}(s)}{s(s+\rho)}, \quad \mathbb{E}[\alpha_{\mathcal{B}_1}] = 1 + \frac{1}{\rho} + \frac{\rho}{2}\cdot\frac{\mathbb{E}\sigma^2}{1+\rho}. \tag{28}$$

We now have information about all systems that we now compare. The comparisons depend on the message size distributions. We choose to consider two "extremes". First, exponentially distributed size, second, deterministic, representing maximal and minimal randomness. The observations are summarized in plots rather than formulas because the latter, albeit explicit in almost all cases are not succinctly presentable.

5.2 Exponential message sizes

We obtain explicit formulas from Corollary 2 for $\alpha_{\mathcal{B}_2}$, Corollary 4 for $\alpha_{\mathcal{P}_2}$ and (28), (27), for $\alpha_{\mathcal{B}_1}$, $\alpha_{\mathcal{P}_1}$, respectively. We use the notation $M_{\mathcal{P}_1}(\rho)$ for $\mathbb{E}[\alpha_{\mathcal{P}_1}]$, where $\rho = \lambda/\mu = \lambda$ in normalized units. Similarly for other systems. We summarize the comparisons in a plot:

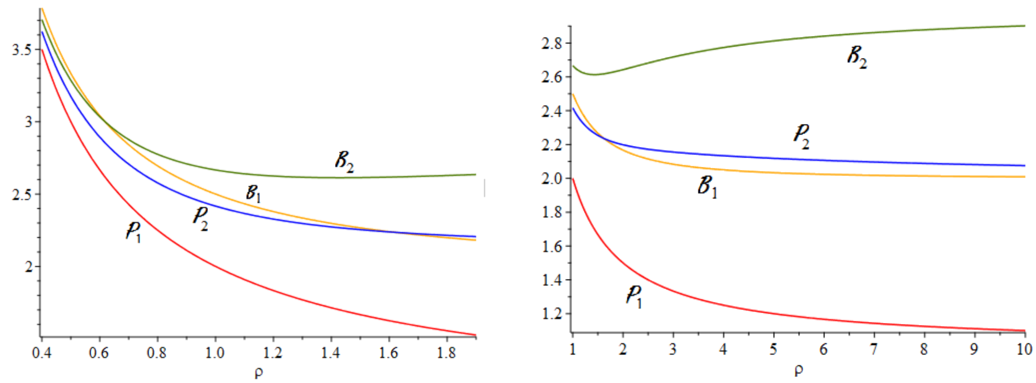

Figure 6: *Mean AoI as a function of ρ; the left plot is for $0.4 \leq \rho \leq 1.9$; the right is for $\rho \geq 1$.*

We see that

$$M_{\mathcal{P}_1}(\rho) < M_{\mathcal{P}_2}(\rho) < M_{\mathcal{B}_2}(\rho) \text{ for all } \rho.$$

The odd system is $\mathcal{B}_1$. For small ρ, $M_{\mathcal{B}_1}(\rho)$ is worst (highest). For large ρ, $M_{\mathcal{B}_1}(\rho)$ is between $M_{\mathcal{P}_1}(\rho)$ and $M_{\mathcal{P}_2}(\rho)$. There is also an intermediate zone, where $M_{\mathcal{B}_1}(\rho)$ is between $M_{\mathcal{P}_2}(\rho)$ and $M_{\mathcal{B}_2}(\rho)$.

We can also ask whether the comparisons above remain true in the sense of stochastic ordering. The information is obtained by inverting the Laplace transforms (23) and (48) which give the densities (4) and (5), respectively. It is also easy to invert the Laplace transforms (27) and (28). Integrating the densities from t to ∞, we obtain the complementary distribution functions, better summarized in a couple of plots:

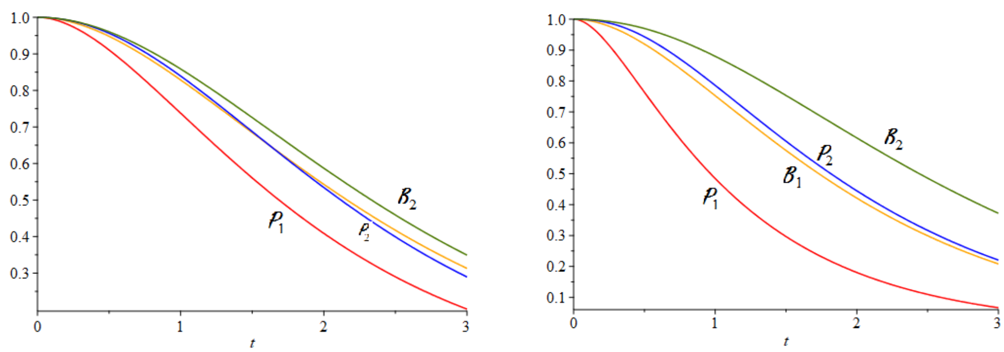

Figure 7: $\mathbb{P}(\alpha > t)$ *plotted against t for small ρ on the left and high ρ on the right.*

We obtain that

$$\alpha_{\mathcal{P}_1} \leq_{\text{st}} \alpha_{\mathcal{P}_2} \leq_{\text{st}} \alpha_{\mathcal{B}_2} \text{ for all } \rho.$$

Moreover,

$$\alpha_{\mathcal{P}_1} \leq_{\text{st}} \alpha_{\mathcal{B}_1} \leq_{\text{st}} \alpha_{\mathcal{P}_2} \text{ for all sufficiently high } \rho.$$

The following figure gives plots of variances as functions of ρ.

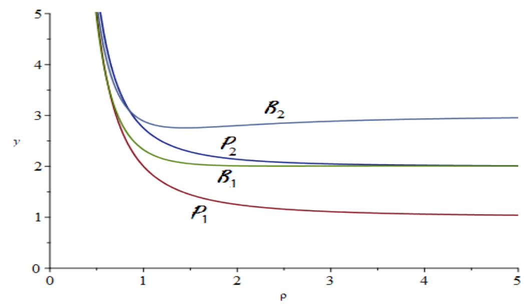

Figure 8: *Variances as a functions of ρ*

Note that they all converge to integers. To see why, see Appendix C and then recall that we assume in this section that $\mu = 1$, so the service-time σ and σ_I distributed as $\hat{G}_I$ of (19) both have variance one, where $\hat{G}_I = \hat{G}$ since σ is assumed to be exponential in this subsection.

5.3 Deterministic message sizes

We now assume that $\mathbb{P}(\sigma = 1) = 1$: message sizes are all equal to 1 with probability 1. We can thus easily obtain $M(\rho)$ in all cases by setting $\sigma = 1$ in the formulas of Corollaries 1 and 3 and in (28) and (27). They are summarized in Figure 9. We observe that

$$M_{\mathcal{P}_2}(\rho) < M_{\mathcal{B}_2}(\rho) < M_{\mathcal{P}_1}(\rho) \text{ for all } \rho.$$

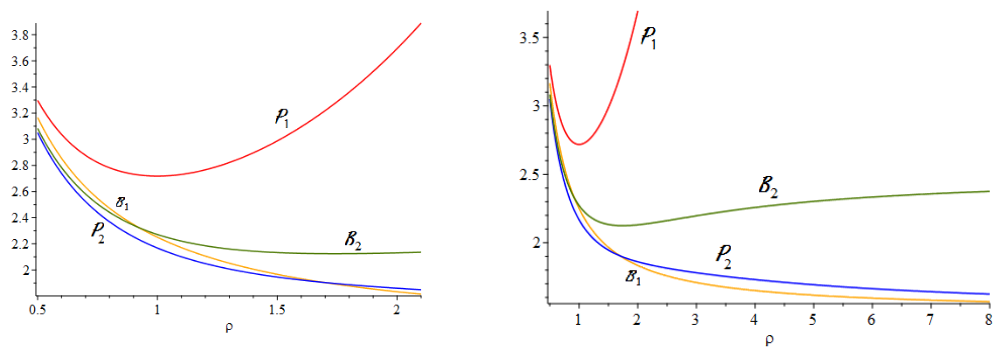

Figure 9: *Mean AoI as a function of ρ; right plot extends to high values of ρ*

Whereas $\mathcal{P}_1$ was best in the exponential case, it is now worst. In fact,

$$\lim_{\rho\to\infty} M_{\mathcal{P}_1}(\rho) = \infty.$$

The worst system, from the point of view of expectation, is thus $\mathcal{P}_1$. However, as in the exponential case, $\mathcal{B}_1$ is the odd system in that it is between $\mathcal{B}_2$ and $\mathcal{P}_1$ for small ρ, but $M_{\mathcal{B}_1}(\rho) < M_{\mathcal{P}_2}(\rho)$ for all large enough ρ. However, the difference between the two goes to 0 as $\rho \to \infty$. We can easily see that $\lim_{\rho\to\infty} M_{\mathcal{B}_1}(\rho) = \lim_{\rho\to\infty} M_{\mathcal{P}_2}(\rho) = 3/2$, while $\lim_{\rho\to\infty} M_{\mathcal{B}_2}(\rho) = 5/2$.

We again ask whether the comparisons in the mean translate to stochastic comparisons. We observe that

$$\alpha_{\mathcal{P}_1} \xrightarrow[\rho\to\infty]{\text{d}} \infty.$$

The reason for this is clear: when ρ is high, the message being processed is constantly interrupted. Since the message size is always 1 no message has a chance to ever be completed. To obtain information about $\mathbb{P}(\alpha_{\mathcal{P}_1} > x)$ for all x, we resort to numerics as the Laplace transform (27) with $\hat{G}(s) = e^{-s}$ is not invertible.

Luckily, the Laplace transforms for all other variables, $\alpha_{\mathcal{B}_1}, \alpha_{\mathcal{B}_2}, \alpha_{\mathcal{P}_2}$ are all invertible and correspond to random variables with densities that can all be analytically computed. We summarize the comparisons of the distributions in the plot below.

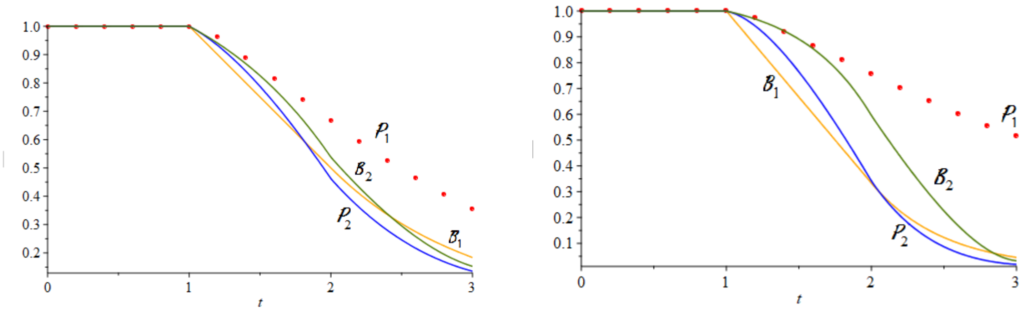

Figure 10: $\mathbb{P}(\alpha > t)$ *plotted against* t *for small* ρ *on the left and high* ρ *on the right.*

Our observation is then that

$$\alpha_{\mathcal{P}_2} \leq_{\rm st} \alpha_{\mathcal{B}_2} \leq_{\rm st} \alpha_{\mathcal{P}_1}, \;\; \alpha_{\mathcal{B}_1} \leq_{\rm st} \alpha_{\mathcal{P}_1} \text{ for all } \rho,$$

whereas $\alpha_{\mathcal{B}_1}$ is not comparable to any of the other three random variables. Figure 11 shows the densities for $\alpha_{\mathcal{P}_2}$ and $\alpha_{\mathcal{B}_2}$ for various values of ρ.

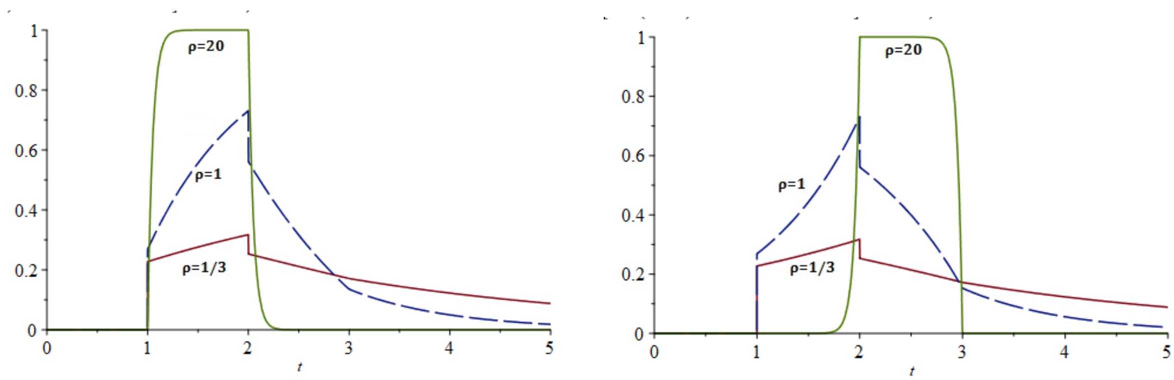

Figure 11: *Densities of* $\alpha_{\mathcal{P}_2}$ *(left) and* $\alpha_{\mathcal{B}_2}$ *(right) for various traffic intensities.*

Variance plots are in Figure 12.

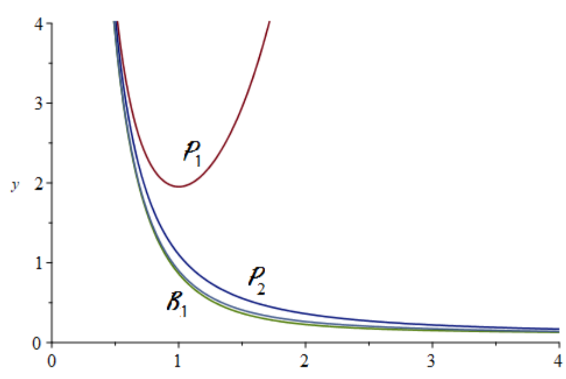

Figure 12: *variances of all systems as a function of* ρ

We have $\lim_{\rho\to\infty} \mathrm{var}_{\mathbb{P}}(\alpha_{\mathcal{B}_1}) = \lim_{\rho\to\infty} \mathrm{var}_{\mathbb{P}}(\alpha_{\mathcal{P}_2}) = \lim_{\rho\to\infty} \mathrm{var}_{\mathbb{P}}(\alpha_{\mathcal{B}_2}) = 1/12$.

6 The $\mathcal{P}_3$ LIFO system

In this section, we consider larger finite buffers, i.e., the $\mathcal{P}_3$ case for the M/GI model. Again note that $\mathcal{P}_n$ systems for $n \geq 3$ employ a LIFO dequeuing policy when selecting (successful) messages to enter the server and begin transmission. We show that $\mathcal{P}_3$ has higher mean AoI than $\mathcal{P}_2$ for all traffic loads, and higher mean AoI than $\mathcal{P}_1$ for exponential service times, in support of the "small buffers" conjecture for minimal AoI. The details for stationary mean AoI involve an interesting use of time-reversibility.

6.1 Why simple pathwise arguments won't work to compare $\mathcal{P}_m$ policies for $m > 1$

Recall that saying that a message that is "$\mathcal{P}_m$ successful" means that it captures the server and so completes service under $\mathcal{P}_m$ for $m > 1$ (which does not preempt service).

Obviously, some messages are $\mathcal{P}_3$ successful but not $\mathcal{P}_2$ successful. Note that service of $\mathcal{P}_3$ successful messages that were formerly pushed back in the queue do not reduce $\mathcal{P}_3$'s AoI.

Conversely, it's possible some $\mathcal{P}_2$ successful messages are not $\mathcal{P}_3$ successful. To see why, note that a message that is $\mathcal{P}_3$ successful but not $\mathcal{P}_2$ successful may have a long service time during which one or more $\mathcal{P}_2$ successful messages (arriving in a burst) are pushed out under $\mathcal{P}_3$.

Consider messages that are **both** $\mathcal{P}_2$ and $\mathcal{P}_3$ successful. It's possible that these messages are not served in the same order in these systems. To see why, suppose consecutive arriving messages i and $i+1$ are successful in both. If they both arrive when $\mathcal{P}_3$'s server is busy with an earlier message having very short remaining service time, but i arrives when $\mathcal{P}_2$'s server is idle, then $i+1$ is served before i in $\mathcal{P}_3$, unlike $\mathcal{P}_2$. In this case, service of i under $\mathcal{P}_3$ does not change $\mathcal{P}_3$'s AoI, and $\mathcal{P}_3$ enjoys a lower AoI than $\mathcal{P}_2$ for an interval of time after $i+1$ departs under $\mathcal{P}_3$. On the other hand, if i arrives when $\mathcal{P}_3$'s server is busy and $\mathcal{P}_2$'s server is idle, but $i+1$ arrives after i departs both systems, then i experiences queueing delay only in $\mathcal{P}_3$. Thus, $\mathcal{P}_2$ enjoys a lower AoI than $\mathcal{P}_3$ for an interval of time after i departs under $\mathcal{P}_2$.

6.2 Mean AoI for $\mathcal{P}_3$ under Poisson arrivals

Let τ, τ_k represent i.i.d. message interarrival times with $\mathbb{P}(\tau > t) = e^{-\lambda t}$ for $t > 0$ and $\lambda^{-1} := \mathbb{E}\tau < \infty$. Again, let σ, σ_k represent i.i.d. message service times which are independent of interarrival times with $\mathbb{P}(\sigma \leq t) =: G(t)$ for $t > 0$, $\mathbb{P}(\sigma > 0) = 1$, $\mu^{-1} := \mathbb{E}\sigma < \infty$, and $\hat{G}(s) = \int_0^\infty e^{-sx} dG(x)$. Let S_n denote the moment at which the n^{th} successful message completes service and departs the system. Let $K_n \in \{0, 1, 2\}$ be the number of messages in the system immediately after S_n. So $\{K_n, S_n\}$ is a Markov-renewal process. In particular, under $\mathbb{P}^0$, $\alpha(0)$ is independent of $S_1 - S_0$ given K_0.

Let $P_{ij} = \mathbb{P}(K_n = j | K_{n-1} = i)$ so that

$$\mathbf{P} = \begin{bmatrix} \mathbb{P}(\tau_1 > \sigma) & \mathbb{P}(\tau_1 \leq \sigma, \tau_1 + \tau_2 > \sigma) & \mathbb{P}(\tau_1 + \tau_2 \leq \sigma) \\ \mathbb{P}(\tau_1 > \sigma) & \mathbb{P}(\tau_1 \leq \sigma, \tau_1 + \tau_2 > \sigma) & \mathbb{P}(\tau_1 + \tau_2 \leq \sigma) \\ 0 & \mathbb{P}(\tau_1 > \sigma) & \mathbb{P}(\tau_1 \leq \sigma) \end{bmatrix}$$
$$= \begin{bmatrix} \hat{G}(\lambda) & -\lambda \hat{G}'(\lambda) & 1 - \hat{G}(\lambda) + \lambda \hat{G}'(\lambda) \\ \hat{G}(\lambda) & -\lambda \hat{G}'(\lambda) & 1 - \hat{G}(\lambda) + \lambda \hat{G}'(\lambda) \\ 0 & \hat{G}(\lambda) & 1 - \hat{G}(\lambda) \end{bmatrix} \tag{29}$$

If $u = \hat{G}(\lambda)$ and $v = -\lambda \hat{G}'(\lambda)$ then the stationary distribution of K, $\pi_i := \mathbb{P}^0(K = i)$, is

$$\pi_0 = \frac{u^2}{1-v}, \ \pi_1 = \frac{u(1-u)}{1-v}, \ \pi_2 = \frac{1-u-v}{1-v}$$

As Proposition 1, we have:

Proposition 2. *The stationary mean AoI is*

$$\mathbb{E}\alpha(0) = \frac{\sum_{i=0}^{2} \mathbb{E}^0(\alpha(0)|K_0 = i)\mathbb{E}^0(S_1 - S_0|K_0 = i)\pi_i + \frac{1}{2}\mathbb{E}^0(S_1 - S_0)^2}{\mathbb{E}^0(S_1 - S_0)}, \tag{30}$$

where $\mathbb{P}^0$ *and* $\mathbb{E}^0$ *are respectively probability and expectation given* $S_0 = 0$.

Proof. For $t \in [S_0, S_1)$, $\alpha(t) = \alpha(S_0) + t - S_0$. So, by the Palm inversion formula [2],

$$\begin{aligned}\mathbb{E}\alpha(0) &= \frac{\mathbb{E}^0 \int_{S_0}^{S_1} \alpha(t)dt}{\mathbb{E}^0(S_1 - S_0)} \\ &= \frac{\mathbb{E}^0(\alpha(S_0)(S_1 - S_0)) + \frac{1}{2}\mathbb{E}^0(S_1 - S_0)^2}{\mathbb{E}^0(S_1 - S_0)} \end{aligned} \tag{31}$$

□

The terms in (31) are computed as per the following lemmas.

The following three lemmas are straightforward.

Lemma 4. *It holds that*

$$\mathbb{E}^0(S_1 - S_0|K_0) = \mu^{-1} + \lambda^{-1}\mathbf{1}\{K_0 = 0\}, \tag{32}$$

$$\mathbb{E}^0(S_1 - S_0) = \pi_0\lambda^{-1} + \mu^{-1}, \tag{33}$$

$$\mathbb{E}^0(S_1 - S_0)^2 = \mathbb{E}\sigma^2 + \pi_0 2(\mu^{-1} + \lambda^{-1})\lambda^{-1}. \tag{34}$$

Proof. The proof is straightforward after noting that $S_1 - S_0 \overset{d}{=} \sigma + \tau\mathbf{1}\{K_0 = 0\}$. □

Lemma 5. *It* τ, τ_1, τ_2 *are exponential random variables with rate* λ *and* σ *has distribution* G *on* $\mathbb{R}^+$ *with Laplace transform* $\hat{G}$*, and are all independent, then the following hold:*

$$\begin{aligned}
\mathbb{E}(\sigma|\tau > \sigma) &= \frac{-\hat{G}'(\lambda)}{\hat{G}(\lambda)} \\
\mathbb{E}(\sigma|\tau \le \sigma) &= \frac{\mu^{-1} + \hat{G}'(\lambda)}{1 - \hat{G}(\lambda)} \\
\mathbb{E}(\sigma|\tau_1 + \tau_2 \le \sigma) &= \frac{\mu^{-1} + \hat{G}'(\lambda) - \lambda\hat{G}''(\lambda)}{1 - \hat{G}(\lambda) + \lambda\hat{G}'(\lambda)} \\
\mathbb{E}(\sigma|\tau_1 \le \sigma, \tau_1 + \tau_2 > \sigma) &= \frac{\hat{G}''(\lambda)}{-\hat{G}'(\lambda)} \\
\mathbb{E}(\tau|\tau \le \sigma) &= \frac{\lambda^{-1} - \lambda^{-1}\hat{G}(\lambda) + \hat{G}'(\lambda)}{1 - \hat{G}(\lambda)} \\
\mathbb{E}(\tau_1|\tau_1 + \tau_2 \le \sigma) &= \frac{\lambda^{-1} - \lambda^{-1}\hat{G}(\lambda) + \hat{G}'(\lambda) - \frac{1}{2}\lambda\hat{G}''(\lambda)}{1 - \hat{G}(\lambda) + \lambda\hat{G}'(\lambda)} \\
\mathbb{E}(\tau_1|\tau_1 \le \sigma, \tau_1 + \tau_2 > \sigma) &= \frac{\hat{G}''(\lambda)}{-2\hat{G}'(\lambda)} \\
q := \mathbb{P}(\tau_1 + \tau_2 > \sigma|\tau_1 \le \sigma) &= \frac{-\lambda\hat{G}'(\lambda)}{1 - \hat{G}(\lambda)}
\end{aligned}$$

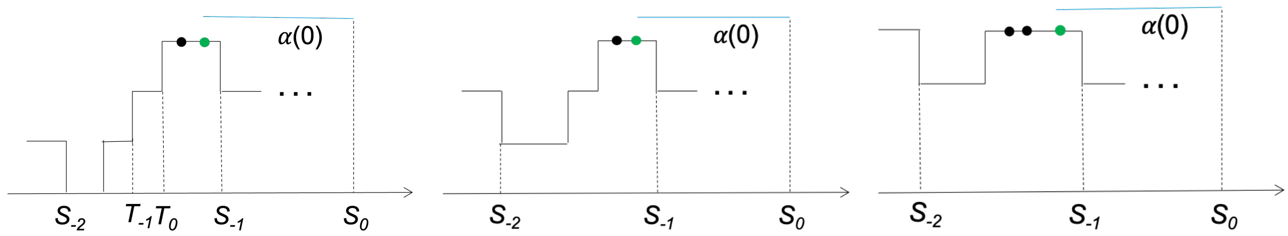

Figure 13: *Cases for* $K_{-2} \in \{0,1,2\}$ *when* $K_{-1} = 2$

Lemma 6. *For* $\ell \in \{0,1,2\}$*,*

$$\mathbb{E}^0(\alpha(0)|K_0 = \ell) = \sum_{i=0}^{2}\sum_{j=0}^{2} \mathbb{E}^0(\alpha(0)|K_{-2} = i, K_{-1} = j, K_0 = \ell)P^r_{\ell j}P^r_{ji} \tag{35}$$

where the transition probabilities of the time-reversed Markov chain K are $P^r_{ji} = \pi_i P_{ij}/\pi_j$*.*

The conditional expectations

$$\mathbb{E}^0(\alpha(0)|K_{-2} = i, K_{-1} = j, K_0 = \ell) \tag{36}$$

are evaluated in the following three lemmas.

Lemma 7. *For* $i \in \{0,1\}$*,*

$$\mathbb{E}^0(\alpha(0)|K_{-2} = i, K_{-1} = 0, K_0 = \ell) = \begin{cases} \mathbb{E}(\sigma|\tau > \sigma) & \text{if } \ell = 0 \\ \mathbb{E}(\sigma|\tau_1 \le \sigma, \tau_1 + \tau_2 > \sigma) & \text{if } \ell = 1 \\ \mathbb{E}(\sigma|\tau_1 + \tau_2 \le \sigma) & \text{if } \ell = 2 \end{cases} \tag{37}$$

Proof. If $K_{-1} = 0$ then $\alpha(0)$ is just the service time of the first message arriving in $[S_{-1}, S_0)$, irrespective of the value i of K_{-2} which, however, cannot be equal to 2 (since $P_{20} = 0$ as is seen in (29)). □

Lemma 8. *For* $i \in \{0,1\}$*,*

$$\mathbb{E}^0(\alpha(0)|K_{-2} = i, K_{-1} = 2, K_0 = \ell) = \mathbb{E}(\tau_1|\tau_1 + \tau_2 \le \sigma) + \begin{cases} \mathbb{E}(\sigma|\tau > \sigma) & \text{if } \ell = 1 \\ \mathbb{E}(\sigma|\tau \le \sigma) & \text{if } \ell = 2 \end{cases} \tag{38}$$

$$\mathbb{E}^0(\alpha(0)|K_{-2} = 2, K_{-1} = 2, K_0 = \ell) = \mathbb{E}(\tau|\tau \le \sigma) + \begin{cases} \mathbb{E}(\sigma|\tau > \sigma) & \text{if } \ell = 1 \\ \mathbb{E}(\sigma|\tau \le \sigma) & \text{if } \ell = 2 \end{cases} \tag{39}$$

Proof. Given $K_{-1} = 2$, see Figure 13. If $K_{-2} = i$ for $i \in \{0,1\}$, there must be at least two message arrivals during the service interval within $[S_{-2}, S_{-1})$, i.e., at T_{-1}, T_0. If there are only two arrivals, then $\alpha(0) = S_0 - T_0$, but there may be more than two (black and green dots, where the green arrival is the one served in $[S_{-1}, S_0)$), in which case $\alpha(0)$ is as indicated in the two left sub-figures. So (38), where the first term is by time-reversibility of the Poisson arrival process. Similarly, if $K_{-2} = 2, K_{-1} = 2$ then there is at least one arrival in $[S_{-2}, S_{-1})$ and so (39). □

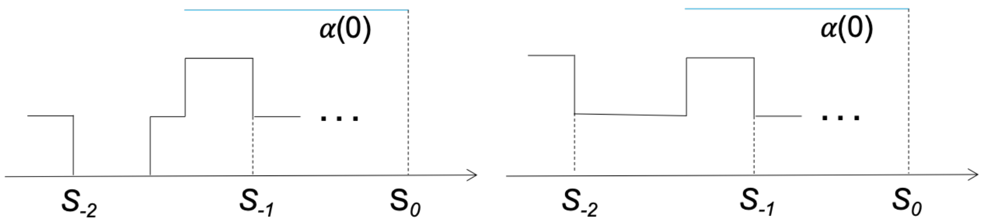

Figure 14: *Cases for* $K_{-2} \in \{0, 1\}$ *when* $K_{-1} = 1$

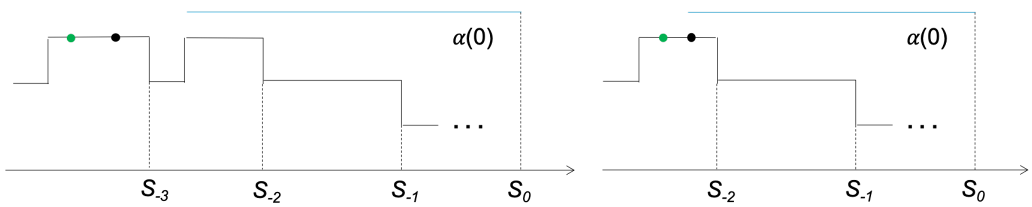

Figure 15: *Cases for* $K_{-2} = 2$ *when* $K_{-1} = 1$

Lemma 9. *For* $i \in \{0, 1\}$,

$$\mathbb{E}^0(\alpha(0)|K_{-2} = i, K_{-1} = 1, K_0 = \ell) = \mathbb{E}(\tau_1|\tau_1 \leq \sigma, \tau_1 + \tau_2 > \sigma) + \begin{cases} \mathbb{E}(\sigma|\tau > \sigma) & \text{if } \ell = 0 \\ \mathbb{E}(\sigma|\tau_1 \leq \sigma, \tau_1 + \tau_2 > \sigma) & \text{if } \ell = 1 \\ \mathbb{E}(\sigma|\tau_1 + \tau_2 \leq \sigma) & \text{if } \ell = 2 \end{cases} \tag{40}$$

$$\begin{aligned} \mathbb{E}^0(\alpha(0)|K_{-2} = 2, K_{-1} = 1, K_0 = \ell) &= q\mathbb{E}(\tau_1|\tau_1 \leq \sigma, \tau_1 + \tau_2 > \sigma) + (1 - q)\mathbb{E}(\tau_1|\tau_1 + \tau_2 \leq \sigma) \\ &+ \mathbb{E}(\sigma|\tau > \sigma) + \begin{cases} \mathbb{E}(\sigma|\tau > \sigma) & \text{if } \ell = 0 \\ \mathbb{E}(\sigma|\tau_1 \leq \sigma, \tau_1 + \tau_2 > \sigma) & \text{if } \ell = 1 \\ \mathbb{E}(\sigma|\tau_1 + \tau_2 \leq \sigma) & \text{if } \ell = 2 \end{cases} \end{aligned} \tag{41}$$

Proof. First see Figure 14. Given $K_{-1} = 1$ again we condition on K_{-2}. If $K_{-2} \in \{0, 1\}$ then the message that departs at S_0 is the last one that arrives in the service interval within $[S_{-2}, S_{-1})$. So (40), where the first term is by time-reversibility of the Poisson arrivals.

The case given $K_{-2} = 2, K_{-1} = 1$ is interesting. See Figure 15 and recall q from Lemma 5. $K_{-2} = 2$ implies at least one arrival in $[S_{-3}, S_{-2})$. Generally, there is a geometric number $N \sim \mathsf{geom}(q)$ of consecutive intervals like $[S_{-3}, S_{-2})$ of the left sub-figure with exactly one arrival.[5] Just before these N intervals is an interval with more than one arrival, where the second-last arrival (green dot) therein is serviced in $[S_{-1}, S_0)$. Recall that the definition of AoI is the time since the arrival of the most recently arrived message which has been completely served, which at time $t = S_0 = 0$ is *not* the (green) message served in $[S_{-1}, S_0)$ in this case. Note that $N = 1$ in the left sub-figure, $N = 0$ in the right one, and $\mathbb{P}(N > 0) = q$. Also, $S_{-1} - S_{-2} \stackrel{\mathrm{d}}{=} (\sigma|\tau > \sigma)$. So (41). □

Note that in the case of Figure 15 at left, the message being served in $[S_{-1}, S_0)$ could be very stale - a geometric number of fresher messages have been served before it. This can also occur in $\mathcal{P}_m$ for $m > 3$ but not in $\mathcal{P}_2$.

[5] That is, $\mathbb{P}(N = k) = q^k(1 - q)$ for $k = 0, 1, 2, \ldots$

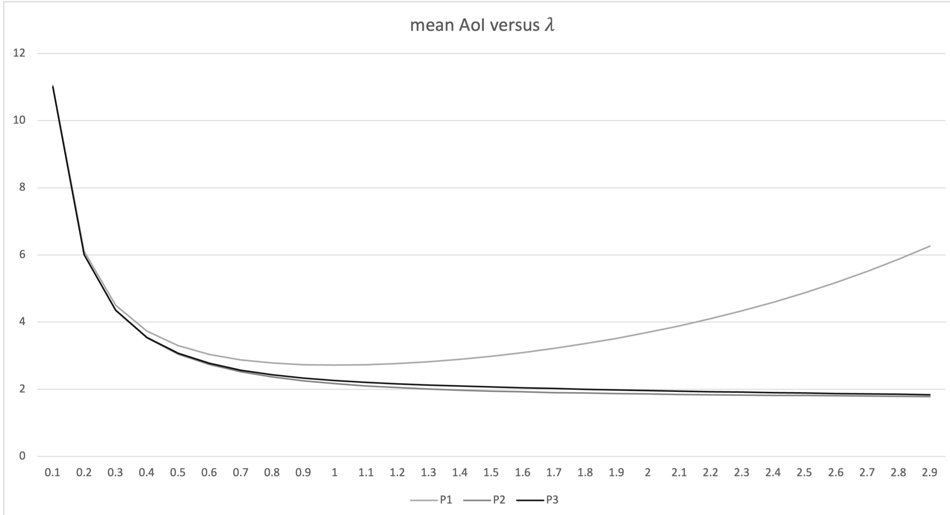

Figure 16: *For Poisson arrivals and deterministic service times with* $\mu = 1$

6.3 Comparing $\mathbb{E}\alpha(0)$ for different $\mathcal{P}_m$ policies

Here let α_k be the AoI process for $\mathcal{P}_k$.

Recall from Section 5 that for deterministic service times ($\hat{G}(\lambda) = e^{-\lambda/\mu}$), $\mathcal{P}_2$ was shown to have lower mean AoI than $\mathcal{P}_1$ (and $\mathcal{B}_1$ for sufficiently small traffic loads). See Figure 16 which shows that $\mathcal{P}_2$ has smaller mean AoI than $\mathcal{P}_1$ or $\mathcal{P}_3$.

For exponential service times ($\hat{G}(\lambda) = \mu/(\lambda + \mu)$), recall that $\mathcal{P}_1$ has lowest mean AoI, consistent with Figure 17. Also, $\mathcal{P}_2$ has lower mean AoI than $\mathcal{P}_3$.

The preceding analysis is easily extended to compute the Laplace transform of the stationary AoI distribution $\mathcal{P}_3$, as above for $\mathcal{B}_2$, and can be adapted in a straightforward way for $\mathcal{P}_m$ LIFO systems for $m > 3$.

7 The $\mathcal{P}_{2,\theta}$ system - a hybrid $\mathcal{P}_1/\mathcal{P}_2$ policy

In this section, for the two-buffer case, we consider a causal service policy that is not "pure" pushout, rather a hybrid of $\mathcal{P}_1$ and $\mathcal{P}_2$ policies. We numerically show that it can produce lower mean AoI than $\mathcal{P}_1$ or $\mathcal{P}_2$ for some service-time distributions.

That is, again suppose there is a buffer consisting of two cells. Cell 1 is reserved for the message receiving service and cell 2 for the message waiting. If there is a message in cell 1 at time t we let $\mathfrak{u}(t)$ be the amount of service received by this message up to t; if the system is empty, we set $\mathfrak{u}(t) = 0$. Fix $\theta \in [0, \infty]$. If a message arrives at time t and $\mathfrak{u}(t) \leq \theta$ then the arriving message pushes-out the message in cell 1 and takes its place. Otherwise, if $\mathfrak{u}(t) > \theta$ then the arriving message occupies cell 2 (pushing out the message sitting there, if any). We call this system $\mathcal{P}_{2,\theta}$. Note that $\mathcal{P}_{2,0}$ and $\mathcal{P}_{2,\infty}$ make sense too and that the collection $\mathcal{P}_{2,\theta}$, $0 \leq \theta \leq \infty$, is a "homotopy" between these two systems. In fact, $\mathcal{P}_{2,0} = \mathcal{P}_2$ and $\mathcal{P}_{2,\infty} = \mathcal{P}_1$. (In the latter system, cell 2 will never be occupied, so, effectively, it has buffer of size 1.)

Thus, a contiguous service interval that ends with a message departure is a sequence of preempted message-service periods followed by a completed/successful message-service period. During the successful message-service period, any arriving messages obviously fail to preempt and, under queue pushout, the *last* such arriving message is queued and begins service once the successful message-service period ends.

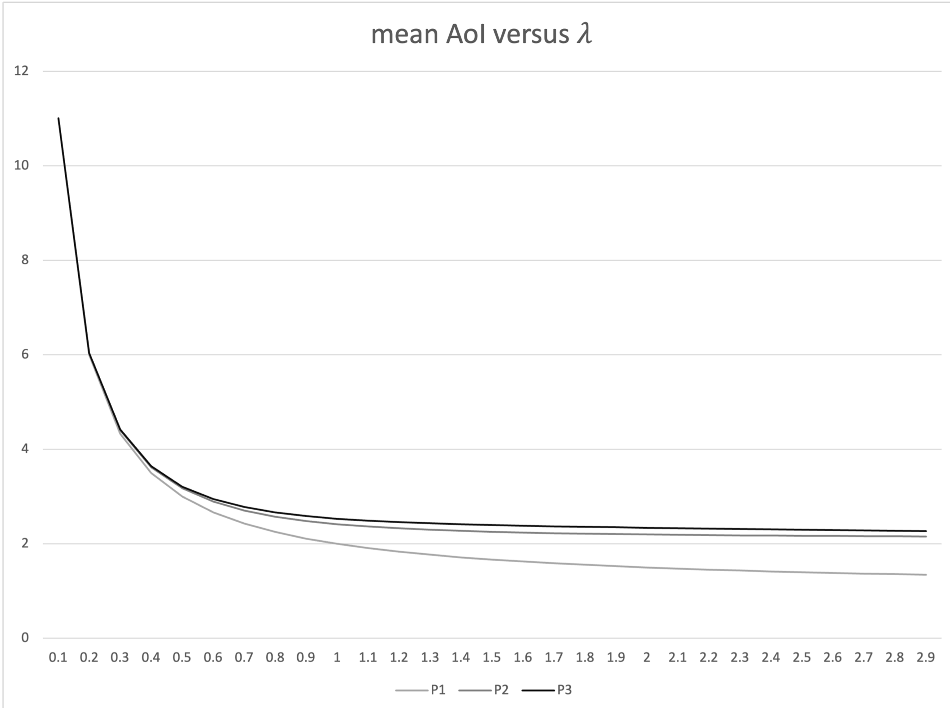

Figure 17: *For Poisson arrivals and exponential service times with* $\mu = 1$

Again assume that the arrival process is Poisson with rate $\lambda > 0$ and that messages have i.i.d. service times (independent of arrivals) distributed like a random variable σ such that $\sigma > 0$ a.s. with expectation $1/\mu < \infty$. Again let G be the distribution function of σ and set $\hat{G}(s) = \mathbb{E}e^{-s\sigma}$. Also assume that the system $\mathcal{P}_{2,\theta}$ is in steady-state (again taking into account that there is a unique such steady-state, the reasons for which are classical and will not be discussed here).

The stationary mean AoI can be obtained from the analysis of (50) in Appendix B again using a Markov embedding approach. Note that it can be numerically minimized over θ for a given set of M/GI model parameters, i.e., for an arbitrary service-time distribution G. The possibility of selecting an optimal θ using perturbation analysis, e.g., [13], is left to future work. Again recall from Section 5 that for exponential service times ($dG(x) = \mu e^{-\mu x} dx$), $\theta = \infty$ ($\mathcal{P}_1$) achieves minimal mean AoI. Also, for constant service time ($dG(x) = \delta_{1/\mu}(dx)$), $\theta = 0$ ($\mathcal{P}_2$) achieves smaller mean AoI than $\mathcal{P}_1$ (and also smaller than $\mathcal{B}_1$ for sufficiently small traffic loads).

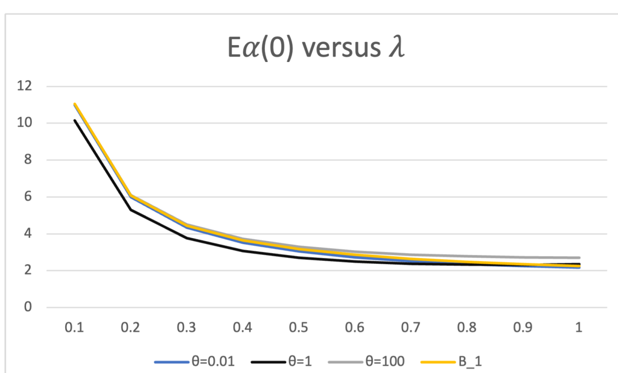

Figure 18: $\mathbb{E}\alpha(0)$ *for service-time distribution which is a mixture of deterministic and exponential with mean* $\mu^{-1} = 1$.

Now consider the mixture service-time distribution defined by $dG(x) = \frac{1}{2}\delta_1(x)dx + \frac{1}{2}e^{-x}dx$ for $x > 0$, and so $\mu = 1$. From Figure 18, which was obtained numerically using (50), we see that in some cases $\mathbb{E}\alpha(0)$ is *not* minimized at either $\theta = 0$ or $\theta = \infty$. That is, in some cases (specifically, traffic loads ≤ 0.8), the $\mathcal{P}_{2,\theta}$ policy for a finite $\theta > 0$ is better than $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{B}_1$.

There are variations of the $\mathcal{P}_{2,\theta}$ policy one could consider. For example, consider the policy where a message arriving at time t does not preempt the in-service message (but joins the queue in cell 2) if $0 < \mathfrak{u}(t) < \theta$, otherwise the message captures the server. The stationary AoI distribution of this policy is similarly derived using the Markov embedding.

8 Summary and future work

In summary, we have newly derived the stationary AoI mean and distribution for the $\mathcal{B}_2$, $\mathcal{P}_3$ and $\mathcal{P}_{2,\theta}$ queueing systems by method of Markov embedding for M/GI models. The same method can be used for the $\mathcal{P}_n$ and $\mathcal{B}_n$ systems for any integer $n > 1$. In particular, the stationary AoI distribution was also thus derived for $\mathcal{P}_2$ having been formerly derived in [7] by different means. Though we've shown how to numerically establish whether stationary AoI is stochastically larger under $\mathcal{P}_{n+1}$ than $\mathcal{P}_n$ for particular service time distributions and $n > 1$, to our knowledge this has not yet been generally proved. (This can be generally shown for the $\mathcal{B}_n$ cases by a simple pathwise argument.)

Based on the results herein and prior work (particularly [6, 3, 14, 7, 12]), rules of thumb for minimal AoI of the M/GI model can roughly be summarized as follows: If the message sizes are deterministic or nearly so then it is best to use $\mathcal{B}_1$ or $\mathcal{P}_2$, and $\mathcal{P}_1$ gives large AoI. On the other hand, if message sizes are "very random" (i.e., close to exponentially distributed), we expect the opposite: $\mathcal{P}_1$ is stochastically minimal (and this is so even for non-Poisson arrivals). Also, we showed that for some cases of service-time distributions, the $\mathcal{P}_{2,\theta}$ policy for some finite $\theta > 0$ gives smaller AoI than $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{B}_1$.

Granted, the study in this paper has been done only for Poisson arrivals. For the GI/M cases, it's not surprising that the same method of Markov embedding shown above can be used to derive stationary AoI distributions [4, 7]. Note that bounds are obtained in [5] the $\mathcal{P}_2$ system for the GI/GI model were based on the approach of [7]. Analysis of the AoI distribution of, e.g., variations of the $\mathcal{P}_{2,\theta}$ policy, LIFO blocking systems, processor sharing policies, or multiserver systems are left to future work.

Even under the assumptions of Poisson arrivals and a single dedicated server, given an arbitrary service-time distribution G, determining which causal service policy will result in somehow "minimal" AoI remains an open problem. This problem may be subject to technological constraints (e.g., whether non-FIFO dequeueing or preemptive service policies are feasible, and what aspects of the queueing system are observable). On the other hand, the choice may be simplified, e.g., $\mathcal{B}_2$ may be the best policy if LIFO dequeueing, queue pushout, and service preemption are infeasible.

Finally, we mention that the AoI α defined in (3) may not be the most appropriate measure of freshness as it incorporates information about the arrival process as well. In the notation of the processes introduced in (1) and (2), a different measure is $\beta(t) = A_0(t) - A(t)$ where $A_0(t)$ is the most recent arrival prior to time t (whether it is ultimately successful or not). The distribution of β may differ significantly from that of α [12]. If the model allows, the Markov embedding approach can also be used to find the stationary marginal distribution of β. Also, it may be interesting to study other performance criteria for AoI-sensitive applications including Cost of Update Delay [15, 7] or message blocking probability.

References

[1] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003.

[2] F. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences, 2nd Ed.* Springer-Verlag, Berlin, 2003.

[3] A.M. Bedewy, Y. Sun and N.B. Shroff. Minimizing the Age of Information Through Queues. *IEEE Trans. Info. Th.* **65**(8):5125-5232, Aug. 2019.

[4] E. Çinlar. *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, NJ, 1975.

[5] J. P. Champati, H. Al-Zubaidy, and J. Gross. On the Distribution of AoI for the GI/GI/1/1 and GI/GI/1/2* Systems: Exact Expressions and Bounds. In *Proc. IEEE INFOCOM*, 2019.

[6] M. Costa, M. Codreanu, and A. Ephremides. On the Age of Information in Status Update Systems With Packet Management. *IEEE Trans. Info. Th.* **62**(4):1897-1910, April 2016.

[7] Y. Inoue, H. Masuyama, T. Takine and T. Tanaka. A general formula for the stationary distribution of the age of information and its application to single-server queues. *IEEE Trans. Inf. Th.* **65**(12):8305-8324, Dec. 2019.

[8] Y. Jiang and N. Miyoshi. Joint performance analysis of ages of information in a multi-source pushout server. *IEEE Trans. Info. Th.* **68**(2), Feb 2022.

[9] S. Kaul, R.D. Yates, and M. Gruteser. Real-time status: How often should one update? In *Proc. IEEE INFOCOM*, 2012.

[10] S. Kaul, R.D. Yates, and M. Gruteser. Status updates through queues. In *Proc. Conf. on Information Sciences and Systems (CISS)*, March 2012.

[11] V. Kavitha, E. Altman, and I. Saha. Controlling Packet Drops to Improve Freshness of information. https://arxiv.org/pdf/1807.09325.pdf, July 18, 2018.

[12] G. Kesidis, T. Konstantopoulos, and M.A. Zazanis. The new age of information: a tool for evaluating the freshness of information in bufferless processing systems. *Queueing Systems* **95**(3):203–250, Aug. 2020.

[13] P. Konstantopoulos and M. Zazanis. Sensitivity analysis for stationary and ergodic queues. *Adv Appl Prob*, 24(3):738-750, Sept. 1992.

[14] A. Kosta, N. Pappas, and V. Angelakis. Age of information: A new concept, metric, and tool. *Foundations and Trends in Networking* **12**(3):162–259, 2017.

[15] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis. Age and value of information: Non-linear age case. In *Proc. IEEE ISIT*, Jun. 2017.

[16] Y. Sun, I. Kadota, R. Talak, and E. Modiano. *Age of Information: A New Metric for Information Freshness*. Synthesis Lectures on Communication Networks, Dec. 2019.

[17] R. Yates, Y. Sun, D. Brown, S. Kaul, E. Modiano and S. Ulukus. Age of information: an introduction and survey. *IEEE J. Sel. Areas Comm.* **39**(5):1183-1210, May 2021.

Appendix A: Proof of Theorem 2 by means of Markov embedding

We make use of (11) of Proposition 1 which needs computation of the quantities involving $\alpha(0)$ in its right-hand side. The analog of Lemma 2 is Lemma 10 below which looks conspicuously the same. In fact, the first two formulas are identical. The last two differ.

Lemma 10. *For $\mathcal{P}_2$,*

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 0] = \hat{G}(\lambda)\hat{G}(s + \lambda), \tag{42}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 1] = \hat{G}(\lambda)\left(\hat{G}(s) - \hat{G}(s + \lambda)\right), \tag{43}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 0] = \frac{\lambda}{\lambda + s}\left(1 - \hat{G}(s + \lambda)\right)\hat{G}(s + \lambda), \tag{44}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 1] = \frac{\lambda}{\lambda + s}\left(1 - \hat{G}(\lambda + s)\right)\left(\hat{G}(s) - \hat{G}(s + \lambda)\right) \tag{45}$$

Proof. 1) When $K_{-1} = 0, K_0 = 0$ or when $K_{-1} = 0, K_0 = 1$ the AoI is the same as in the $\mathcal{B}_2$ system, the reason being that the number of messages in the system is always at most 1, see Figure 3.

2) Suppose next that $K_{-1} = 1, K_0 = 0$. In Figure 19 we depict the two scenaria corresponding to the possible values of K_{-2}, namely $(K_{-2}, K_{-1}, K_0) = (0, 1, 0)$ or $(1, 1, 0)$. In both cases, (3) and the system dynamics imply that

$$\alpha(0) = S_0 - S_{-1} + V,$$

V is the time elapsed between the last arrival in the interval (S_{-2}, S_{-1}) and S_{-1}. If there is only one arrival in this interval then $V := S_{-1} - T_0$. In any case,

conditionally on $\{K_{-1} = 1, K_0 = 0\}$, the random variables V and $S_0 - S_{-1}$ are independent.

Therefore,

$$\mathbb{E}^0[e^{-s(S_0 - S_{-1} + V)} \mid K_{-1} = 0, K_0 = 1] = \mathbb{E}^0[e^{-s(S_0 - S_{-1})} \mid K_{-1} = 0, K_0 = 1]\,\mathbb{E}^0[e^{-sV} \mid K_{-1} = 0, K_0 = 1]$$

and the first factor on the right is easy:

$$\mathbb{E}^0[e^{-s(S_0 - S_{-1})} \mid K_{-1} = 0, K_0 = 1] \;=\; \int_0^\infty e^{-sx}\frac{e^{-\lambda x}dG(x)}{\hat{G}(\lambda)} \;=\; \frac{\hat{G}(s + \lambda)}{\hat{G}(\lambda)}.$$

To evaluate $\mathbb{E}^0[e^{-sV} \mid K_{-1} = 0, K_0 = 1]$ we note that, V is the distance of the last Poisson point inside the interval (T_{-1}, S_{-1}) (in the left scenario of Figure 19) or the interval (T_{-1}, S_{-1}) in the right scenario. (In both cases the length of the interval is that of a message size conditioned on containing at least one Poisson point.) To obtain the Laplace transform of V look backward in time starting from S_{-1} until the first Poisson point appears and condition on the event that this occurs between S_{-1} and S_{-2}. Thus, the density of V at $v > 0$ is

$$\frac{(1 - G(v))\lambda e^{-\lambda v}}{1 - \hat{G}(\lambda)},$$

which gives

$$\mathbb{E}^0[e^{-sV} \mid K_{-1} = 0, K_0 = 1] \;=\; \int_0^\infty e^{-sv}\frac{\lambda e^{-\lambda v}(1 - G(v))}{1 - \hat{G}(\lambda)}dv \;=\; \frac{\lambda}{\lambda + s}\,\frac{1 - \hat{G}(s + \lambda)}{1 - \hat{G}(\lambda)}. \tag{46}$$

Putting these together we obtain (44).

3) Finally, assume that $K_{-1} = 1, K_0 = 1$. This situation is similar to the previous one and thus will be treated succinctly. We are guided by Figure 20. Firstly, we have

$$\mathbb{E}^0[e^{-s(S_0 - S_{-1})} \mid K_{-1} = 1, K_0 = 1] \;=\; \int_0^\infty e^{-sx}\frac{1 - e^{-\lambda x}dG(x)}{1 - \hat{G}(\lambda)} \;=\; \frac{\hat{G}(s) - \hat{G}(s + \lambda)}{1 - \hat{G}(\lambda)}.$$

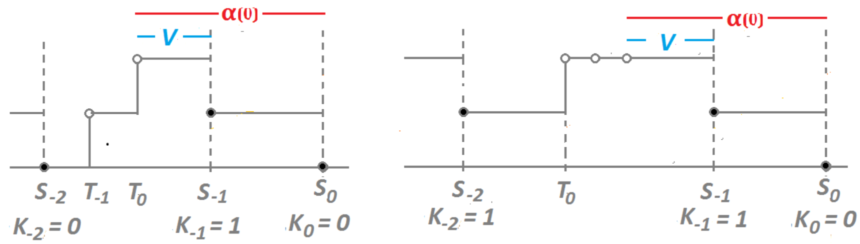

Figure 19: *Under $\mathcal{P}_2$, the segments $[S_{-2}, S_{-1})$, $[S_{-1}, S_0)$ when $K_{-1} = 1, K_0 = 0$ in the case where $K_{-2} = 0$ (left) and $K_{-2} = 1$ (right).*

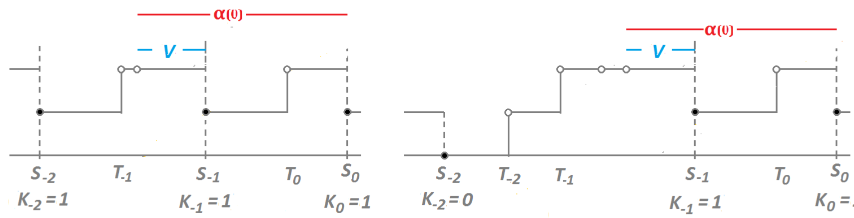

Figure 20: *Under $\mathcal{P}_2$, the segments $[S_{-2}, S_{-1})$, $[S_{-1}, S_0)$ when $K_{-1} = 1, K_0 = 1$ in the case where $K_{-2} = 0$ (left) and $K_{-2} = 1$ (right).*

Secondly, the argument used to derive (46) can be used here again with no changes to obtain

$$\mathbb{E}^0[e^{-sV} \mid K_{-1} = 1, K_0 = 1] = \int_0^\infty e^{-sv} \frac{\lambda e^{-\lambda v}(1 - G(v))}{1 - \hat{G}(\lambda)} dv = \frac{\lambda}{\lambda + s} \frac{1 - \hat{G}(s + \lambda)}{1 - \hat{G}(\lambda)}.$$

Putting these together we obtain (44) as well. □

The formula for $\mathbb{E}e^{-s\alpha(0)}$ in Theorem 2 is now clear:

Proof. Adding up (42) and (43) of Lemma 10 and similarly (44) and (45) we obtain

$$\mathbb{E}[e^{-s\alpha(S_0)}; K_0 = 0] = \hat{G}(s + \lambda)\left[\hat{G}(\lambda) + \frac{\lambda}{\lambda + s}\left(1 - \hat{G}(s + \lambda)\right)\right]$$

$$\mathbb{E}[e^{-s\alpha(S_0)}; K_0 = 1] = \left(\hat{G}(s) - \hat{G}(s + \lambda)\right)\left[\hat{G}(\lambda) + \frac{\lambda}{\lambda + s}\left(1 - \hat{G}(\lambda + s)\right)\right].$$

Substituting these expressions in the numerator of (11), and recalling the definition (19) of G_I, we obtain (25). □

An alternative expression for (25) is:

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\hat{G}(s)\left(\hat{G}(\lambda) + \frac{\lambda}{\mu}\hat{G}_I(s + \lambda)\right)\left(\frac{\lambda}{\lambda+s}\hat{G}(s + \lambda) + \frac{\lambda}{\mu}\hat{G}_I(s)\right)}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}. \tag{47}$$

Corollary 4. *For $\mathcal{P}_2$ with exponential message sizes,*

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\mu}{\mu + s}\left(\frac{\mu}{\mu + \lambda} + \frac{\lambda}{\lambda + \mu + s}\right)\left(\frac{\mu^2}{\lambda^2 + \lambda\mu + \mu^2}\frac{\lambda}{\lambda + s}\frac{\lambda + \mu}{\lambda + \mu + s} + \frac{\lambda^2 + \lambda\mu}{\lambda^2 + \lambda\mu + \mu^2}\frac{\mu}{\mu + s}\right), \tag{48}$$

$$\mathbb{E}[\alpha(0)] = \frac{2\lambda^5 + 7\lambda^4\mu + 8\lambda^3\mu^2 + 7\lambda^2\mu^3 + 4\lambda\mu^4 + \mu^5}{\lambda\mu(\lambda + \mu)^2(\lambda^2 + \lambda\mu + \mu^2)}, \tag{49}$$

and, with $\rho = \lambda/\mu$, the standard deviation of $\alpha(0)$ under $\mathbb{P}$ is

$$\mathrm{sd}_{\mathbb{P}}(\alpha(0)) = \frac{1}{\mu}\frac{\sqrt{2\rho^{10}+12\rho^9+35\rho^8+60\rho^7+66\rho^6+56\rho^5+45\rho^4+34\rho^3+18\rho^2+6\rho+1}}{\rho\,(\rho+1)^2\,(\rho^2+\rho+1)}.$$

The expectation and variance have been computed by summing up the expectations and variance of the three independent random variables comprising $\alpha(0)$. Inverting $\mathbb{E}e^{-s\alpha(0)}$ shows that $\alpha(0)$ has density (5), i.e., when $\mu = 1$:

$$f_{\mathcal{P}_2}(t) = q_1(t)e^{-t} + q_2(t)e^{-(\lambda+1)t} - \frac{\lambda}{\lambda-1}e^{-\lambda t},$$

if $\lambda \neq 1$ where

$$q_1(t) = \frac{(\lambda^3+\lambda^2-2\,\lambda)t - \lambda^2+\lambda+3}{(\lambda^2+\lambda+1)(\lambda-1)}, \quad q_2(t) = \frac{(\lambda^2+\lambda)t+\lambda^2+3\,\lambda+3}{\lambda^2+\lambda+1},$$

else if $\lambda = 1$,

$$f_{\mathcal{P}_2}(t) = \tfrac{1}{3}(7+2t)e^{-2t} + \tfrac{1}{3}(6t-7)e^{-t}.$$

It is easy to see from (48) that $\lim_{\lambda\to\infty}\mathbb{E}[e^{-s\alpha(0)}] = (\mu/(s+\mu))^2$, the sum of 2 i.i.d. exponentials.

Appendix B: Analysis of the $\mathcal{P}_{2,\theta}$ system

We now analyze the system described in Section 7.

Given $K_{-1} = 0$, consider Figure 21 at right.

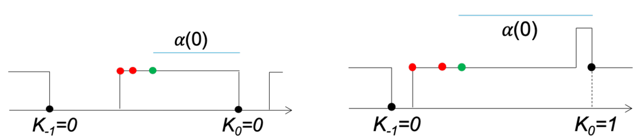

Figure 21: *Number of messages in the system (black line) when $K_{-1} = 0, K_0 = 0$ (left) and $K_{-1} = 0, K_0 = 1$ (right). A black dot indicates a (successful) message departure. A green dot indicates an arrival of a message that will successfully depart. A red dot indicates an arrival of an unsuccessful message. If the fate of an arrival is not indicated in the figure, then it has no indicating dot.*

A message service period is successful with probability

$$1 - q := \mathbb{P}(\tau > \theta \wedge \sigma) = (1-G(\theta))e^{-\lambda\theta} + \int_0^\theta e^{-\lambda s}dG(s).$$

So, considering the successful service period which concludes at S_0:

$$\mathbb{P}(K_0 = 0|K_{-1} = 0) = \mathbb{P}(\tau > \sigma \mid \tau > \theta \wedge \sigma) = (1-q)^{-1}\hat{G}(\lambda) = 1 - \mathbb{P}(K_1 = 1|K_{-1} = 0).$$

Given $K_{-1} = 1$, consider Figure 22 below.

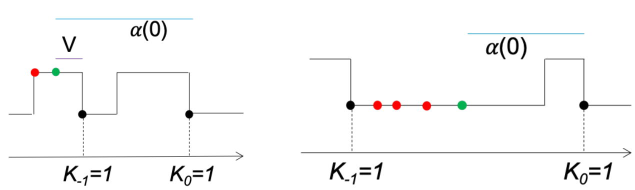

Figure 22: $K_{-1} = 1, K_0 = 1$ *cases.*

.

Use the memoryless property of interarivals to similarly obtain

$$\mathbb{P}(K_0 = 0|K_{-1} = 1) = (1-q)^{-1}\hat{G}(\lambda) = 1 - \mathbb{P}(K_0 = 1|K_{-1} = 1)$$

So, K_n is an i.i.d. Bernoulli sequence with

$$p_0 := \mathbb{P}(K_n = 0) = (1-q)^{-1}\hat{G}(\lambda) = 1 - \mathbb{P}(K_n = 1) =: 1 - p_1.$$

The queueing process over consecutive intervals $[S_{i-1}, S_i)$ and $[S_i, S_{i+1})$ are conditionally independent given K_i. Thus, $\{S_i, K_i\}_{i\in\mathbb{Z}}$ is Markov-renewal with renewal times S_i and the queueing process is semi-Markov [4]. In particular, $\alpha(S_i)$ and $S_{i+1} - S_i$ are conditionally independent given K_i.

As Proposition 1, we have:

Proposition 3. *The Laplace transform of the stationary AoI distribution is*

$$\mathbb{E}e^{-s\alpha(0)} = \frac{\sum_{i=0}^{1} \mathbb{E}^0(e^{-s\alpha(0)}|K_0 = i)\Big(1 - \mathbb{E}^0(e^{-s(S_0-S_{-1})}|K_{-1} = i)\Big)p_i}{s\mathbb{E}^0(S_1 - S_0)}, \tag{50}$$

where $\mathbb{P}^0$ and $\mathbb{E}^0$ are respectively probability and expectation given $S_0 = 0$.

Proof. By the Palm inversion formula [2],

$$\mathbb{E}e^{-s\alpha(0)} = \frac{\mathbb{E}^0 \int_{S_0}^{S_1} e^{-s\alpha(t)}dt}{\mathbb{E}^0(S_1 - S_0)},$$

where the numerator

$$\begin{aligned}
\mathbb{E}^0 \int_{S_0}^{S_1} e^{-s\alpha(t)}dt &= \mathbb{E}^0 \int_{S_0}^{S_1} e^{-s(\alpha(S_0)+t-S_0)}dt \\
&= \mathbb{E}^0 \int_{S_0}^{S_1} \sum_{i=0}^{1} (e^{-s\alpha(0)}|K_0 = i)\mathbb{E}^0(e^{-s(t-S_0)}|K_0 = i)p_i dt \\
&= \sum_{i=0}^{1} \mathbb{E}^0(e^{-s\alpha(0)}|K_0 = i)\mathbb{E}^0\left(\frac{1 - e^{-s(S_1-S_0)}}{s}\middle|K_0 = i\right)p_i
\end{aligned}$$

□

To calculate the terms in (50) we need to follow the steps outlined in the lemmas below. Let $\hat{F}_0(s) = \hat{G}(s+\lambda)/\hat{G}(\lambda)$ and $\hat{J}(s) = \int_0^\infty e^{-sy}dJ(y)$ where $J(y) = 1$ for $y \geq \theta$ and, for $0 \leq y < \theta$,

$$dJ(y) = q^{-1}\lambda e^{-\lambda y}(1 - G(y))dy$$

Lemma 11.

$$\mathbb{E}^0(e^{-s(S_0-S_{-1})} \mid K_{-1} = 0, K_0 = 0) = \frac{\lambda}{\lambda+s} \cdot \frac{1-q}{1-q\hat{J}(s)}\hat{F}_0(s) \tag{51}$$

$$\mathbb{E}^0(e^{-s\alpha(S_0)} \mid K_{-1} = 0, K_0 = 0) = \hat{F}_0(s) \tag{52}$$

Proof. See Figure 21 at left and consider the interval $[S_{-1}, S_0)$. Let τ_{-1} be first message arrival time in this interval minus S_{-1}, so that $\tau_{-1} \sim \exp(\lambda)$ by the memoryless property. Note that there is a geometric number N of interarrival times each of which is smaller than *both* θ and the associated service time; $N = 2$ in Figure 21. Again, the probability of such unsuccessful service is

$$\mathbb{P}(\tau < \theta \wedge \sigma) = q.$$

So, $\mathbb{P}(N = k) = (1 - q)q^k$ for $k = 0, 1, 2, \ldots$. Let $Y \overset{(d)}{=} (\tau | \tau < \theta \wedge \sigma)$ so that $\mathbb{P}(Y \le y) = J(y)$. Finally, let $\tau_0 \sim \exp(\lambda)$ be the duration between the arrival time (green dot) of the message that departs at S_0 and the next arrival time. The service time (from the green dot to S_0) σ_0 is independent of τ_0. Considering the prior N unsuccessful service completions, we are given that $\tau_0 > \sigma_0$ or $\tau_0 > \theta$. Given $K_0 = 0$, $\tau_0 > \sigma_0$. Let $X_0 \overset{(d)}{=} (\sigma_0 | \tau_0 > \sigma_0)$ which has distibution

$$dF_0(x) = \hat{G}(\lambda)^{-1} e^{-\lambda x} dG(x), \quad x > 0,$$

with $\mathbb{E}e^{-sX_0} = \hat{F}_0(s)$. So,

$$(S_0 - S_{-1} \mid K_{-1} = 0, K_0 = 0) \overset{(d)}{=} \tau_{-1} + \sum_{n=1}^{N} Y_n + X_0,$$

a sum of independent terms with $Y_n \overset{(d)}{=} Y$. Also $\alpha(S_0) \overset{(d)}{=} X_0$ in this case. □

Define

$$dF_1(x) = \frac{(e^{-\lambda\theta} - e^{-\lambda x})dG(x)}{\int_\theta^\infty (e^{-\lambda\theta} - e^{-\lambda z})dG(z)}, \quad x > \theta$$
$$\hat{F}_1(s) = \int_\theta^\infty e^{-sx} dF_1(x).$$

Lemma 12.

$$\mathbb{E}^0(e^{-s(S_0 - S_{-1})} \mid K_{-1} = 0, K_0 = 1) = \frac{\lambda}{\lambda + s} \cdot \frac{1 - q}{1 - q\hat{J}(s)} \hat{F}_1(s) \tag{53}$$
$$\mathbb{E}^0(e^{-s\alpha(S_0)} \mid K_{-1} = 0, K_0 = 1) = \hat{F}_1(s). \tag{54}$$

Proof. See Figure 21 at right. The difference between this and the previous case is that here $\sigma_0 > \tau_0 > \theta$. So, the distribution of $X_1 \overset{(d)}{=} (\sigma_0 \mid \sigma_0 > \tau_0 > \theta)$ is $dF_1(x)$. □

Define

$$dH(v) = \frac{\lambda e^{-\lambda v}(1 - G(v + \theta))dv}{\int_\theta^\infty (1 - e^{-\lambda(x - \theta)})dG(x)}, \quad v \ge 0,$$
$$\hat{H}(s) = \int_0^\infty e^{-sv} dH(v).$$

Lemma 13.

$$\mathbb{E}^0(e^{-s(S_0 - S_{-1})} \mid K_{-1} = 1, K_0 = 0) = \frac{1 - q}{1 - q\hat{J}(s)} \hat{F}_0(s) \tag{55}$$
$$\mathbb{E}^0(e^{-s\alpha(S_0)} \mid K_{-1} = 1, K_0 = 0) = q\hat{F}_0(s) + (1 - q)\hat{H}(s) \times (55) \tag{56}$$

Proof. See Figure 23 below.

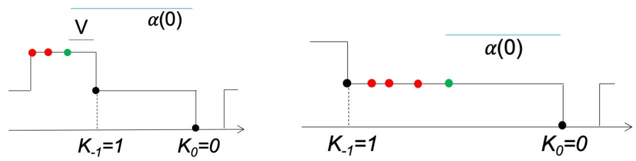

Figure 23: $K_{-1} = 1, K_0 = 0$ *cases.*

.

For $\alpha(S_0)$ there are two subcases depending of whether there are initial unsuccessful arrivals in the interval $[S_{-1}, S_0)$, i.e., whether $N > 0$. When $N > 0$ (Figure 23 right), this case is like when $K_{-1} = 0, K_0 = 0$. Otherwise (Figure 23 left),

$$\alpha(S_0) = V + S_0 - S_{-1},$$

where V is the duration between S_{-1} the last arrival before S_{-1} (green dot), and V and $S_0 - S_1$ are independent given $K_{-1} = 1$. Starting from S_{-1}, look backward in time until the first Poisson point appears (green dot) and condition on the event that this occurs at least θ units of time before the service time ends. Thus, $V \overset{(d)}{=} (\tau | \tau < \sigma - \theta, \sigma > \theta)$ with distribution H. Also, $S_0 - S_{-1}$ is distributed as for the case where $K_{-1} = 0, K_0 = 0$ except the first interarrival time is absent. □

Lemma 14.

$$\mathbb{E}^0(e^{-s(S_0 - S_{-1})} \mid K_{-1} = 1, K_0 = 1) = \frac{1-q}{1-q\hat{J}(s)}\hat{F}_1(s) \tag{57}$$

$$\mathbb{E}^0(e^{-s\alpha(S_0)} \mid K_{-1} = 1, K_0 = 1) = q\hat{F}_1(s) + (1-q)\hat{H}(s) \times (57) \tag{58}$$

Proof. See Figure 22. For $S_0 - S_{-1}$, this case is a combination of the previous two cases, and for $\alpha(S_0)$ follow the previous case except use (54) instead of (52). □

The final stage: The formulas obtained in the lemmas above must now be substituted into (50) as follows:

$$\mathbb{E}^0(e^{-s(S_0 - S_{-1})} | K_{-1} = 0) = (53) \times p_1 + (51) \times p_0 \tag{59}$$

$$\mathbb{E}^0(e^{-s(S_0 - S_{-1})} | K_{-1} = 1) = (57) \times p_1 + (55) \times p_0. \tag{60}$$

So,

$$\mathbb{E}^0(S_1 - S_0) = -\frac{d}{ds}\mathbb{E}^0 e^{-s(S_0 - S_{-1})}\Big|_{s=0} = -\frac{d}{ds}(p_1 \times (60) + p_0 \times (59))\Big|_{s=0}.$$

Moreover,

$$\mathbb{E}^0(e^{-s\alpha(0)} | K_0 = 0) = p_0 \times (52) + p_1 \times (56)$$

$$\mathbb{E}^0(e^{-s\alpha(0)} | K_0 = 1) = p_0 \times (54) + p_1 \times (58).$$

Appendix C: High traffic asymptotics

"High traffic asymptotics" refers to the regime $\rho = \lambda/\mu \to \infty$. Even though we have no explicit formulas for $\mathcal{P}_n$ or $\mathcal{B}_n$ when $n \geq 3$, we can easily obtain asymptotics from the system dynamics.

Proposition 4. *Let $\sigma, \sigma_1, \sigma_2, \ldots$ be i.i.d. copies of σ. Let σ_I be distributed as $\hat{G}_I$ as in (19). Then*

$$\alpha_{\mathcal{P}_n} \xrightarrow[\rho\to\infty]{\text{d}} \sigma + \sigma_I, \quad n \geq 2,$$

while

$$\alpha_{\mathcal{B}_n} \xrightarrow[\rho\to\infty]{\text{d}} \sigma_1 + \cdots + \sigma_n + \sigma_I, \quad n \geq 1.$$

Sketch of proof. In both systems, the buffer consists of n cells. The message being processed sits in cell 1. In $\mathcal{P}_n$, the freshest message is either in cell 1, in which case all other cells are empty, or in cell 2. When ρ is high there is always a message being processed in cell 1 and the freshest message is in cell 2. Hence, at any time t, the AoI $\alpha_{\mathcal{P}_n}(t)$ equals the service time of the message in cell 2 plus the remaining service time of the message in cell 1. These are two independent random variables. The first one is distributed as σ. The second is distributed as σ_I since the system is stationary. For $\mathcal{B}_1$, we can obtain the limit from the Laplace transform of (28). It is easy to to see that $\lim_{\rho\to\infty} \mathbb{E}[e^{-s\alpha_{\mathcal{B}_1}}] = \hat{G}(s)\,\hat{G}_I(s)$ and so $\alpha_{\mathcal{B}_1} \xrightarrow[\rho\to\infty]{\text{d}} \sigma + \sigma_I$. For general n, when ρ is high, the AoI $\alpha_{\mathcal{B}_n}(t)$ equals the remaining service time of the message in cell 1 (in distribution equal to σ_I) plus the time elapsed until the beginning of its service which is, in distribution, equal to the sum of n independent service times. □

Remark 2. When $\sigma = 1$ with probability 1, σ_I is a uniform random variable in the interval $[0, 1]$. Hence $\sigma_1 + \cdots + \sigma_n + \sigma_I = n + \sigma_I$ and the variance of this random variable is 1/12. Similarly, for $\mathcal{P}_2$, the asymptotic variance is again 1/12.

George Kesidis
School of EECS, The Pennsylvania State University, University Park, PA 16802, USA; gik2@psu.edu

Takis Konstantopoulos
Department of Mathematical Sciences, The University of Liverpool, Liverpool L69 7ZL, UK; takiskonst@gmail.com

Michael A. Zazanis
Department of Statistics, Athens University of Economics and Business, 76 Patission St., Athens 104 34, Greece; zazanis@aueb.gr