Stationary workload for some non-work-conserving M/G/1 Preemptive LIFO queues

Karl Sigman [*] Jacob Bergquist[†]

February 9, 2021

Abstract

We analyze two non-work-conserving variations of the M/G/1 preemptive last-in first-out (LIFO) queue with emphasis on deriving explicit expressions for the limiting (stationary) distribution of workload. In the first model, we assume that preempted customers are returned to the front of the queue with a new independent and identically distributed service time. In the second, we assume they are returned to the front of the queue with their original service time. Our analysis is based on queueing theory methods such as the Rate Conservation Law, PASTA, regenerative process theory and Little's Law ($l = \lambda w$) as well as some classic results about the standard (work-conserving) M/G/1 preemptive LIFO model. Along the way, we obtain stability results as well as explicit expressions for the limiting distribution of the service time found in service by an arrival. For the second model we even derive the joint distribution of age and excess (remaining service time) of such a service time, and they are quite different from what is found in standard work-conserving models; in fact, we prove that the excess is stochastically larger than the age. For example, in the M/M/1 case, they are independent exponentials but with different rates. We also give heavy-traffic limits and tail asymptotics for stationary workload.

Keywords LIFO QUEUES, PREEMPTIVE QUEUES, WORKLOAD, RATE CONSERVATION LAW, STOCHASTIC ORDERING

AMS 2020 Subject Classification: Primary: ;
Secondary: ;

[*] Department of IEOR, Columbia University, Mudd Building, 500 West 120th Street, New York, NY 10027

[†] Department of IEOR, Columbia University, Mudd Building, 500 West 120th Street, New York, NY 10027

1 Introduction

In this paper we consider two *non-work-conserving* variations of the M/G/1 preemptive LIFO (PL) queue. As with the classic/standard (work-conserving) M/G/1 PL model, when a new customer arrives they immediately bump out any customer in service and start service themself with their own independent and identically distributed (i.i.d) service time, while the preempted customer returns to the front of the queue. But in the classic model, the preempted customer retains its *remaining service time* and thus the model is work-conserving; in particular, the workload process is identical sample-path by sample-path to the standard first-in-first-out (FIFO) M/G/1 model. In the two models we analyze here, the preempted customer either receives a new i.i.d. service time (Model I), or retains its original service time (Model II), hence losing any progress that had been made (workload thus makes a jump upward larger than just the new arriving customer's service time).

These two models were presented and analyzed in the recent paper of Asmussen and Glynn [1] in which the focus was on determining the moments of sojourn time and establishing the stability conditions. They refer to Model I as 'preemptive-repeat different' (PRD) and to Model II as 'preemptive-repeat identical' (PRI), and their methods of analysis involve branching processes, Galton-Watson family trees and stochastic fixed point equations.

Our focus in the present paper, however, is on deriving the entire limiting (stationary) *distribution of workload*. We do so by giving an explicit random variable representation for the workload very much in the spirit of the classic Pollaczek-Khinichine formula for the standard M/G/1 queue in which the limiting distribution of the workload is expressed as a geometric sum of i.i.d. random variables endowed with the *equilibrium* (stationary excess) distribution of service. (See for example, Pages 386-387 in [6]) As part of this we show that for both Models I and II the limiting distribution of the number of customers in the system is geometric, and do so by using a classic queueing method going back to [3].

But as we discover by explicitly computing them, the parameters of these geometric distributions involves the entire distribution of service, not just its mean, and the distribution of the i.i.d. random variables is not the equilibrium distribution, nor the stationary *spread* distribution as found in the inspection paradox.

Along the way we also obtain (in two different ways) the stability results found in [1]– but with more of a 'queueing' interpretation–as well as explicit expressions for the limiting distribution of the service time found in service by an arrival. For Model II we even derive the joint distribution of age and excess (remaining service time) of such a service time, and discover that they are quite different from what is found in standard work-conserving models. Our analysis is based on queueing theory methods such as Rate Conservation Law, PASTA, regenerative process theory and Little's Law ($l = \lambda w$). We also present both heavy-traffic limits and tail asymptotics for the stationary workload.

1.1 Basic M/G/1 model notation and set up

The M/G/1 queue has a Poisson point process of customer arrival times $\{t_n : n \geq 1\}$ at rate λ, and (independently) i.i.d. service times $\{S_n : n \geq 1\}$ brought by each customer distributed as a general distribution $G(x) = P(S \leq x),\ x \geq 0$, where S denotes a generic such service time. We assume that $0 < E(S) = 1/\mu < \infty$. T denotes a generic interarrival time distributed as exponential at rate λ, and then we define $\rho \overset{\text{def}}{=} \lambda/\mu$. (A priori all we can say about ρ is that $0 < \rho < \infty$).

The workload $V(t)$ at time t is the sum of all whole or remaining service times in the system at time t: the sum of any service times of customers in the queue plus the remaining service time of the customer in service (if any). $\{V(t) : t \geq 0\}$ then denotes the workload stochastic process; its sample paths are continuous from the right with left hand limits. $V(t_n-)$ denotes the amount of work *found in the system* by the n^{th} arrival, and $V(t_n+)$ is the amount of work right after they arrive. For example, for work-conserving disciplines $V(t_n+) = V(t_n-) + S_n$.

The times at which an arrival finds the system empty, $V(t_n-) = 0$, serve as regeneration points with i.i.d. cycles. In the case when the cycle length distribution is proper and has finite first moment–the positive recurrent case–we are ensured the existence of a (proper) limiting (stationary) distribution of workload, and that is what we mean by *stability* in the present paper. We let V denote a random variable with this distribution and can define the distribution via w.p. 1 limits:

$$P(V \leq x) = \lim_{t\to\infty} \frac{1}{t} \int_0^t I\{V(s) \leq x\},\ x \geq 0. \tag{1}$$

Because we are assuming Poisson arrivals we can use *Poisson Arrivals See Time Averages (PASTA)* (see for example Theorem 6, Page 294 in [6]) to also express this distribution as a w.p. 1 customer average:

$$P(V \leq x) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I\{V(t_j-) \leq x\}, x \geq 0. \tag{2}$$

Of crucial importance in the present paper is computing

$$p_0 \overset{\text{def}}{=} P(V = 0), \tag{3}$$

which, because of PASTA, can be viewed as both the long-run proportion of time that the system is empty and as the long-run proportion of arriving customers who find the system empty. We can use the above limits to *determine* stability in advance: it is precisely the case when p_0 exists and satisfies $0 < p_0 < 1$ (we avoid the trivial case of $p_0 = 1$).

For work-conserving M/G/1 models it is well known that $0 < p_0 < 1$ if and only if $0 < \rho < 1$ in which case $p_0 = 1 - \rho$, but in the present paper with non-work-conservation in play, we will see that this does not hold in general. Moreover, in the work-conserving case workload $\{V(t) : t \geq 0\}$ forms a Markov process, but it does not in the non-work-conserving models here.

2 PRD: Computing p_0 and the distribution of V

We begin with the M/G/1 LIFO preemptive *repeat-different* queue (PRD). We recall this is the preemptive LIFO model in which each time a customer in service is preempted, they go to the front of the queue with a new i.i.d. service time S distributed as G.

Let S_r denote time-stationary remaining service time (of the customer, if any, in service); we then observe that $P(S_r = 0) = P(V = 0) = p_0$. Let $\hat{S}_r$ denote a random variable distributed as S_r *conditional on* $S_r > 0$; $(S_r \mid S_r > 0)$. From PASTA and the memoryless property of the exponential distribution, the following lemma follows immediately.

Lemma 2.1 *For the M/G/1 PRD model*

$$\hat{S}_r \overset{\mathrm{d}}{=} (S - T \mid S > T). \tag{4}$$

We next compute $E(\hat{S}_r) = E(S - T \mid S > T)$:

Lemma 2.2 *For the M/G/1 PRD model*

$$E(\hat{S}_r) = \frac{\frac{1}{\mu}}{1 - E(e^{-\lambda S})} - \frac{1}{\lambda}. \tag{5}$$

Proof : We re-write

$$E(S - T \mid S > T) = \frac{E(S - T; S > T)}{P(S > T)},$$

where the denominator is computed as $1 - E(e^{-\lambda S})$. To compute the numerator, we condition on $T = t$, use the "integrate the tail" method for computing the expected value of a non-negative random variable, and switch the order of integration via Fubini's (Tonelli's) Theorem:

$$\begin{aligned}
\int_0^\infty \int_t^\infty P(S > x)\lambda e^{-\lambda t} dx dt &= \int_0^\infty \int_0^x P(S > x)\lambda e^{-\lambda t} dt dx \\
&= \int_0^\infty P(S > x)(1 - e^{-\lambda x}) dx \\
&= E \int_0^S (1 - e^{-\lambda x}) dx \\
&= E(S) - E \int_0^S e^{-\lambda x} dx \\
&= \frac{1}{\mu} - \frac{1}{\lambda}(1 - E(e^{-\lambda S}))
\end{aligned}$$

Dividing the numerator by the denominator then yields the result. ■

We now derive a formula for p_0. Using the Rate Conservation Law (RCL) (see for example Theorems 5.5 and 5.6, Page 116 in [5]) with $\{X(t)\}$ as the stochastic process $\{V(t)\}$, we obtain the following proposition.

Proposition 2.1 *When the M/G/1 PRD model is stable,*

$$p_0 = \frac{1 - 2\rho + \lambda E(\hat{S}_r)}{1 - \rho + \lambda E(\hat{S}_r)}. \tag{6}$$

Proof : For $X(t) = V(t)$, we have $X'(t) = -I\{V(t) > 0\}$ (right derivative) w.p. 1. Meanwhile, jumps occur at customer arrival times $\{t_j\}$ at rate λ, but we break them up into two disjoint streams (type 0 and 1, respectively): customers who find the system empty which arrive at rate $\lambda_0 \overset{\text{def}}{=} \lambda p_0$, and those that find the system busy who arrive at rate $\lambda_1 \overset{\text{def}}{=} \lambda(1 - p_0)$. This leads to

$$P(V > 0) = \lambda_0 E^0(-J(0)) + \lambda_1 E^1(-J(1)), \tag{7}$$

where $-J(0)$ denotes a jump amount $(V(t_j+) - V(t_j-))$ of type 0, and $-J(1)$ denotes a jump amount of type 1. We obtain $E^0(-J(0)) = E(S)$ and $E^1(-J(1)) = E(S_1 + S_2 - \hat{S}_r)$ (where S, S_1 and S_2 are i.i.d. distributed as G and independent of $\hat{S}_r$).

Then Equation 7 becomes

$$1 - p_0 = \rho p_0 + 2\rho(1 - p_0) - \lambda(1 - p_0)E(\hat{S}_r),$$

which solving for p_0 yields Equation 6. ■

As a sanity check, using Equation 6 on the special case when G is exponential at rate μ (the M/M/1 case), we know that $\hat{S}_r \sim exp(\mu)$ and hence $\lambda E(\hat{S}_r) = \rho$ and Equation 6 yields $p_0 = 1 - \rho$ as it should. (And of course we must have that $\rho < 1$ in this special case.)

In fact, a little thought reveals that the M/M/1 PRD model has the property that although the workload stochastic process $\{V(t) : t \geq 0\}$ does not have the same distribution as the workload process of the standard M/M/1 (because the sample paths are not work-conserving for PRD and can even take a jump downward), it *does* hold from the memoryless property of the exponential distribution of service that for each *fixed* t, $V(t)$ has the same distribution as the standard M/M/1, and hence has the same limiting distribution; V is the same for both, hence p_0 is the same for both.

Remark 2.1 Equation 6 still remains valid for general renewal arrivals at rate λ except then we no longer have a closed form expression for $\hat{S}_r$: Equation 4 no longer holds in general. In our use of RCL, $\hat{S}_r$ must have the stationary remaining service time distribution with respect to customer arrivals finding the system busy.

Remark 2.2 Our use of RCL requires verifying a priori that all three of the following are finite which they certainly are in our application: $E|X'|$ (since $|X'| \leq 1$), $E(S)$ and $E(|S_1 + S_2 - \hat{S}_r|)$ (because G is assumed to have finite first moment in the model).

2.1 Relating p_0 to the stability condition of PRD

Clearly, Equation 6 makes sense (proportion of time) only if the p_0 it yields satisfies

$$0 \leq p_0 \leq 1.$$

Putting aside *null recurrence*, we would require that $0 < p_0 \leq 1$ to ensure stability/positive recurrence of workload, and to avoid trivialities we thus would require that

$$0 < p_0 < 1.$$

A little thought reveals (since $\rho > 0$) that this happens if and only if the numerator in Equation 6 is strictly positive:

$$1 - 2\rho + \lambda E(\hat{S}_r) > 0, \tag{8}$$

which from Equation 4 is precisely

$$1 - 2\rho + \lambda E(S - T \mid S > T) > 0. \tag{9}$$

From Equation 5 we have

$$\lambda E(S-T \mid S>T) = -1 + \frac{\rho}{1-E(e^{-\lambda S})}, \tag{10}$$

and hence the stability condition 9 becomes

$$E(e^{-\lambda S}) > \frac{1}{2}; \tag{11}$$

we obtain the same condition as in Theorem 5 on Page 15 in [1], which asserts that the PRD queueing model empties infinitely often if and only if

$$P(S<T) = E(e^{-\lambda S}) \geq \frac{1}{2}. \tag{12}$$

(The case when equality holds corresponds to null recurrence.)

Furthermore, plugging Equation 10 back into the p_0 formula in Equation 6 furnishes a closed form solution for p_0:

Proposition 2.2 *For a stable M/G/1 PRD model*

$$p_0 = \frac{2E(e^{-\lambda S})-1}{E(e^{-\lambda S})}. \tag{13}$$

Remark 2.3 An alternative way of proving stability and deriving p_0 is by creating a regular (equivalent) FIFO M/G/1 queue (with a new service-time distribution different from the original G) and showing stability. Observe that

$$\tau = \min\{n \geq 1 : S_n < T_n\},$$

is (in distribution) the total number of times that a customer enters service before departing. It has a geometric distribution with success probability $P(S<T) = E(e^{-\lambda S})$, and hence expected value

$$E(\tau) = \frac{1}{E(e^{-\lambda S})}.$$

$\tau - 1 \geq 0$ then is the total number of times that a customer gets preempted. We thus can define an *effective* service time denoted by $\mathbb{S}$ as the total amount of work required by a customer in the system:

$$\mathbb{S} \overset{\mathrm{d}}{=} (S \mid S<T) + \sum_{j=1}^{\tau-1}(T_j \mid T_j < S_j),$$

where the S and T are independent of the i.i.d. $\{T_j\}$ and $\{S_j\}$ in the sum; each preemption adds in another independent T_j conditional on $T_j < S_j$. The first piece is the final service time, the one that did not get preempted. The second piece is the sum of all the preempted work that was already processed. These effective service times are i.i.d. among customers and can be viewed as "service times" to a standard FIFO M/G/1 queue with mean

$$E(\mathbb{S}) = E(S \mid S<T) + (E(\tau)-1)E(T \mid T<S),$$

and stability holds if and only if

$$\lambda E(\mathbb{S}) < 1.$$

Carrying out the computation yields that the condition is exactly as in Equation 11. Moreover, when stable, $p_0 = 1 - \lambda E(\mathbb{S})$ for standard M/G/1 queues, and the computation yields exactly the p_0 as in Equation 13.

2.2 An interesting PRD example

Consider the case when service times are a mixture of a point mass at 0 and an exponential at rate 0.01:

$$S \overset{\mathrm{d}}{=} (0.99)\delta_0 + (0.01)exp(0.01).$$

Thus $E(S) = 1 = \mu$, and $\rho = \lambda$.

Clearly, $\hat{S}_r \overset{\mathrm{d}}{=} exp(0.01)$ since the 0 ones are never found in service, and so $E(\hat{S}_r) = 100$. Thus Equation 6 becomes

$$p_0 = \frac{1 - 2\lambda + \lambda(100)}{1 - \lambda + \lambda(100)} = \frac{1 + 98\lambda}{1 + 99\lambda},$$

and we see that the system is stable for all values of λ and as $\lambda \to \infty$, p_0 decreases monotonically to 98/99. This is quite intuitive: any non-zero remaining service time that gets preempted is replaced by a 0 service time 99% of the time; the faster the arrival rate, the more likely any positive service time will be found by an arrival and get bumped out.

2.3 Deriving the distribution of V

With p_0 in hand via Proposition 2.2, we now obtain an explicit random variable expression for V. To prepare, let $\hat{Q}$ denote a rv with a geometric distribution (with mass at 0) with success probability p_0:

$$P(\hat{Q} = n) = (1 - p_0)^n p_0,\ n \geq 0.$$

The rv $\hat{Q}$ and the i.i.d. service time sequence $\{S_j\}$ distributed as G, and the random variable $\hat{S}_r$ (distributed as in Lemma 4), are taken as independent in what follows.

Proposition 2.3 *For the stable M/G/1 PRD model,*

$$(V \mid V > 0) \overset{\mathrm{d}}{=} \hat{S}_r + \sum_{j=1}^{\hat{Q}} S_j. \tag{14}$$

Thus the distribution of V, F_V, is a mixture

$$F_V = p_0\delta_0 + (1 - p_0)F_{\hat{V}}$$

where $F_{\hat{V}}$ denotes the distribution of $(V \mid V > 0)$ given in Equation 14, and δ_0 denotes the point mass at 0.

Proof : The key to the proof is in recalling that for a regular M/G/1 PL queue, the stationary number of customers N_{PL} in system is geometric with success probability $1 - \rho$:

$$P(N_{PL} = n) = \rho^n(1 - \rho),\ n \geq 0,$$

and we can still use the simple kind of proof going back to Fakinos in 1981([3], [4]) (In the textbook [6], this is given nicely (guided along) as exercise 9-14 on Page 434.) We use the same method of proof here, the only difference being that ρ is replaced by $1 - p_0$. Before doing so, let us see how this yields a proof to our Proposition. We let N here denote stationary number in system for the PRD model. Assume that we have proved that

$$P(N=n)=(1-p_0)^n p_0,\ n\geq 0. \tag{15}$$

Given $V>0$, the workload is equal to the work in queue (line) plus the (>0) work in service $\hat{S}_r$. But the event $\{V>0\}$ is the same as $\{N\geq 1\}$, and the distribution of the number in queue is simply that of $(N-1\mid N\geq 1)$ which is exactly that of $\hat{Q}$ given in our statement of the proposition, the very same geometric distribution as N. But unlike the regular PL model, the service times in queue are, under PRD, i.i.d. distributed as G. (And they are independent of $\hat{S}_r$.) Thus the work in queue is the second piece in our expression; Equation 14 is derived.

We now prove Equation 15 holds. We recall from PASTA that we can assume N is constructed as a customer-stationary distribution; the distribution found by an arriving customer (denoted by C_0) at arrival time $t_0=0$ given that the system started at time $-\infty$. We already know that $P(N\geq 1)=P(V>0)=1-p_0$. Now let us consider the event $\{N\geq 2\}$. This means that the arriving customer C_0 found a customer C_{-k} in service (who arrived in the past at time t_{-k} for some $k\geq 1$) *and* C_{-k} found a customer in service when they arrived. The customers in the queue found by C_0 upon arrival are precisely the same customers that C_{-k} found upon arrival at time t_{-k}. But these events are independent because of the PL discipline so we get $P(N\geq 2)=P(N\geq 1)^2=(1-p_0)^2$. The proof continues similarly for any $n\geq 0$, with each customer independently having found the system busy thus obtaining our result that $P(N\geq n)=(1-p_0)^n,\ n\geq 1$; Equation 15 is derived. ■

2.4 Heavy-traffic limits for stationary workload V

We now give a simple argument to characterize the asymptotic behavior of stationary workload in the heavy-traffic regime. $\Longrightarrow$ denotes convergence in distribution, while $\overset{p}{\Longrightarrow}$ denotes convergence in probability in what follows. Here we consider a heavy-traffic regime analogous to how it is considered for a regular (work-conserving) M/G/1 queue with G fixed (with tail denoted by $\overline{G}(x),\ x\geq 0$) in which by letting $\lambda\uparrow\mu$ (equivalently $\rho\uparrow 1$) it holds that $(1-\rho)V\Longrightarrow exp(\alpha)$ (the exponential distribution at rate α). In that classic case $\alpha^{-1}=E(S_e)=E(S^2)/2E(S)$, where S_e has density $g_e(x)=\mu\overline{G}(x)$, the *equilibrium* (or stationary excess) distribution of G. For the PRD model we replace $1-\rho$ by $p_0=p_0(\lambda)$, from Equation 13 and consider what happens to p_0V as $p_0\to 0$. Stability for PRD is that $E(e^{-\lambda S})>1/2$, and thus by increasing λ to the value λ_2 such that $E(e^{-\lambda_2 S})=1/2$ results in $p_0\downarrow 0$; *heavy-traffic.*

Theorem 2.1 *Let $\lambda_2>\lambda$ be the solution to $E[e^{-\lambda_2 S}]=1/2$ (which exists by the monotone convergence theorem). Then as $\lambda\uparrow\lambda_2$,*

$$p_0V\Longrightarrow exp(\mu),$$

where $\mu^{-1}=E(S)$.

Proof : Let $N(\lambda)$ denote the stationary number-in-system. By Proposition 2.3, $N(\lambda)$ is a geometric random variable with success probability $p_0(\lambda)=\frac{2E(e^{-\lambda S})-1}{E(e^{-\lambda S})}$ from Equation 13. Since $p_0(\lambda)\downarrow 0$ as $\lambda\uparrow\lambda_2$, it is easily seen that

$$p_0(\lambda)N(\lambda)\Longrightarrow exp(1), \tag{16}$$

as $\lambda \uparrow \lambda_2$. (Based on the fact that $(1-1/a)^a \to e^{-1}$ as $a \to \infty$.) Because $p_0 \to 0$, it follows from Proposition 2.3 that we only need to consider $\hat{V} \overset{\mathrm{d}}{=} (V \mid V > 0)$ and we have

$$p_0(\lambda)\hat{V}(\lambda) \overset{\mathrm{d}}{=} p_0(\lambda)S_r(\lambda) + p_0(\lambda)\sum_{i=1}^{N(\lambda)} S_i \tag{17}$$

But $p_0(\lambda)S_r(\lambda) \overset{p}{\Longrightarrow} 0$ hence can be ignored. Indeed, take $x > 0$ arbitrary. Applying Markov's Inequality and using Lemma 5 gives

$$\begin{aligned}
P(p_0(\lambda)S_r(\lambda) > x) &\le p_0(\lambda)E[S_r(\lambda)]\frac{1}{x} \\
&= p_0(\lambda)\left(\frac{\frac{1}{\mu}}{1-E(e^{-\lambda S})} - \frac{1}{\lambda}\right)\frac{1}{x} \\
&\le p_0(\lambda)\left(\frac{\frac{1}{\mu}}{1-E(e^{-\lambda S})}\right)\frac{1}{x} \\
&\to 0 \times 2/\mu = 0,
\end{aligned}$$

as $\lambda \uparrow \lambda_2$.
Now to handle the second term, we write

$$p_0(\lambda)\sum_{i=1}^{N(\lambda)} S_i = \frac{\sum_{i=1}^{N(\lambda)} S_i}{N(\lambda)} p_0(\lambda)N(\lambda), \tag{18}$$

and note that because $\frac{1}{n}\sum_{i=1}^{n} S_i \overset{a.s.}{\to} E[S]$ as $n \to \infty$ and $N(\lambda) \overset{p}{\Longrightarrow} \infty$ as $\lambda \uparrow \lambda_2$, we have

$$\frac{\sum_{i=1}^{N(\lambda)} S_i}{N(\lambda)} \overset{p}{\Longrightarrow} E[S] = \frac{1}{\mu}, \tag{19}$$

as $\lambda \uparrow \lambda_2$. Thus it follows from Equations 16 and 18 that

$$p_0(\lambda)\sum_{i=1}^{N(\lambda)} S_i \Longrightarrow exp(\mu)$$

■

2.5 Average sojourn time

Since N has a geometric distribution as explained in the proof of Proposition 2.3, we obtain the time average number in system as

$$l = E(N) = \frac{1-p_0}{p_0}.$$

Using our solution to p_0 from Equation 13 then yields

$$l = \frac{\rho}{1 - 2\rho + \lambda E(\hat{S}_r)} = \frac{1 - E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1} = -1 + \frac{E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1}.$$

From Little's law ($l = \lambda w$) we thus can also solve for average sojourn time $w = l/\lambda$:

$$w = \frac{1}{\lambda}\Big[-1 + \frac{E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1}\Big]. \tag{20}$$

Remark 2.4 In [1], the formula given for w in Proposition 6, Page 17 is incorrect as has been confirmed by the authors (private communication). The error is only an algebraic one; carrying out the computation as they suggest, one indeed obtains our Equation 20.

Remark 2.5 In all the PL models, the distribution of sojourn time W is identical to that of a busy period. So Equation 20 is also the expected value of a busy period in the M/G/1 PRD case. Moreover, from PASTA, p_0 is equal to the long-run proportion of arrivals who find the system empty (hence begin a busy period, hence begin a regenerative cycle). Thus by regenerative process theory,

$$p_0 = \frac{1}{E(N_B)},$$

where N_B denotes the number of customers served during a busy period. Using Proposition 2.2, we thus can solve for $E(N_B)$ yielding

$$E(N_B) = \frac{1}{p_0} = \frac{E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1}. \tag{21}$$

3 PRI Model

We now consider the LIFO *repeat-identical* M/G/1 queue (PRI) as introduced in [1]. In this model, whenever a customer is preempted, it retains its identical whole original service time S, as opposed to a new i.i.d. one when it gets sent to the front of the queue. This model is more complicated to analyze than PRD as we shall see, but we first give some preliminary results.

Once again we let V denote stationary workload, $p_0 = P(V = 0)$, and N denotes stationary number in system. (We will solve for p_0 later.)

One important distinction concerning assumptions is that:

> **We can and will assume without loss of generality that $P(S > 0) = 1$.** Why? Because any arriving customer with a 0 service time will pass through without notice; they have no effect on anything, they will never get preempted. A customer can be preempted only if $S > 0$ in which case, what gets placed back in queue (first time of preemption) is an $(S \mid S > 0)$ which remains the identical positive service time forever after. In the PRD model, if $S > 0$, the customer if preempted gets placed back in queue with a new i.i.d. copy distributed as G and hence will still have the possibility of entering service later with a 0. Thus, for PRI: If $P(S > 0) < 1$, then one can replace λ with $\lambda P(S > 0)$ and G with the conditional distribution of $(S \mid S > 0)$ reducing the model to assuming that $P(S > 0) = 1$.

Unlike $\hat{S}_r$ for PRD in Proposition 6, deriving the distribution of such things for PRI appears to be much more challenging, thus solving for p_0 is less direct than for PRD. The difficulty is that although $\hat{S}_r$ is of the form $(S - T \mid T < S)$, the S here has a biased distribution (different from G) since it may have been preempted in the past. (The more times it has already been preempted, the more biased it is.) We also have the stationary *age* of the service found when preempted; we denote it by $\hat{B}$, and we have the sum of the two which is the whole service time of the customer found in service when preempted (in stationarity),

$$\hat{S} = \hat{B} + \hat{S}_r. \tag{22}$$

In stationarity, it is $\hat{S}$ that gets placed back in queue when preempted; it is biased and not (in general) distributed as G. For the standard M/G/1 PL model, what gets placed back is an i.i.d. copy of $\hat{S}_r$ distributed as the *equilibrium distribution* of service time, G_e, with density $g_e(x) = \mu P(S > x),\ x \geq 0$. So, it might be tempting to "guess" that for PRI the distributions of both $\hat{B}$ and $\hat{S}_r$ are that of G_e and the distribution of $\hat{S}$ is that of the stationary *spread* distribution of G found in the inspection paradox (if G has a density $g(x)$, then the spread has a density $\mu x g(x)$). This is not true. For example in the case when $P(S = 1) = 1$, the PRI model is identical to the PRD model and hence $\hat{S}_r$ is distributed as $(1 - T \mid T < 1)$ and $\hat{B}$ is that of $(T \mid T < 1)$, while G_e is the uniform distribution over $(0, 1)$; all three are different. ($\hat{S} = S = 1$ is the spread but that is a triviality because the spread of a constant is itself).

3.1 Preliminary results

3.1.1 Solving for p_0 via computing $E(\hat{B})$

Using RCL on workload $V(t)$ similar to the derivation of Proposition 6 yields the following equation

$$1 - p_0 = \rho + \lambda(1 - p_0)E(\hat{B}), \tag{23}$$

where $\hat{B}$ denotes the stationary *age* of the service time of the customer found in service by an arrival. Age here refers to the amount of the service time that was already processed before getting preempted by the arrival.

Solving for p_0 yields

$$p_0 = \frac{1 - \rho - \lambda E(\hat{B})}{1 - \lambda E(\hat{B})}. \tag{24}$$

We now proceed to compute $E(\hat{B})$ by first determining the distribution of the number of times that an arrival enters service before completing service and departing.

Proposition 3.1 *For a fixed service time S distributed as G, of an arriving customer, let $\tau = \tau(S) \geq 1$ denote the total number of times that it enters service before completion, hence $\tau - 1 \geq 0$ denotes the total number of times it was preempted. Then conditional on S, the distribution of τ is geometric with success probability $e^{-\lambda S}$ and hence the expected total number of times a job enters service is given by*

$$E(\tau) = E(e^{\lambda S}). \tag{25}$$

Proof : Let $\{T_n : n \geq 1\}$ denote i.i.d. exponential random variables at rate λ. By the memoryless property of the exponential distribution in the Poisson arrival process, we have that each time the service time enters service it will be preempted if, after an independent exponential (at rate λ) amount of time, it is still in service, and thus

$$\begin{aligned} P(\tau = 1) &= P(S \leq T_1) \\ P(\tau = n) &= P(S > T_1, \ldots, S > T_{n-1},\ S \leq T_n),\ n \geq 2. \end{aligned}$$

Thus, conditional on S,

$$P(\tau = n \mid S) = (1 - e^{-\lambda S})^{n-1} e^{-\lambda S},\ n \geq 1,$$

the geometric distribution with success probability $e^{-\lambda S}$ and hence

$$E(\tau \mid S) = \frac{1}{e^{-\lambda S}} = e^{\lambda S}.$$

Finally

$$E(\tau) = E(E(\tau \mid S)) = E(e^{\lambda S}).$$

■

3.1.2 Computing the stationary mean and distribution of $\hat{B}$ and determining stability

Recalling Equation 22 and the various definitions of $\hat{B}$, $\hat{S}$ and $\hat{S}_r$ which are defined the same here for PRI, let $K = \tau - 1 \geq 0$ denote the number of times that a service time S is preempted as defined in Proposition 3.1. Then $E(K) = E(e^{\lambda S}) - 1$ from Proposition 3.1.

Note that if a service time S is preempted then, because of the Poisson arrival process and the memoryless property of the exponential distribution, the *age* of the service time at that point (how much service has been completed) is an i.i.d. length T_j distributed as exponential at rate λ conditional on $T_j > S$. Conditional on S, the T_j up to time τ are i.i.d.

From PASTA we know that time averages are the same as customer averages; for example

the time-stationary distribution of age in service at time t, $B(t)$, is the same as the average found by arriving customers; in particular w.p. 1 it holds that (for all non-negative measurable functions f)

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t f(B(s)) ds = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} f(B(t_j-)).$$

We let B denote stationary age, and thus $\hat{B} = (B \mid B > 0)$.

To compute $E(\hat{B})$, we focus on the customer average and let $N^1(n)$ denote the number of arrivals out of the first n for which $B(t_j-) > 0$, and consider the average

$$E(\hat{B}) = E(B \mid B > 0) = \lim_{n \to \infty} \frac{1}{N^1(n)} \sum_{j=1}^{N^1(n)} B(t_j-) I\{B(t_j-) > 0\}. \tag{26}$$

When the system is stable (actually in this case we only need assume that the long-run departure rate equals the arrival rate λ thus ensuring that each customer completes service), each arriving customer will be included in the limit in Equation 26 until they finally depart.

As explained in the proof of Proposition 3.1, we can thus take each customer alone with their independent service time S and their own i.i.d. sequence of interarrival times $\{T_n\}$ and sum up their ages when preempted as if they form one regenerative cycle which ends at interarrival time $K = \tau - 1$. Doing this sequentially for the i.i.d. service times $\{S_n : n \geq 1\}$, the cycles are i.i.d. and thus can be used to form a regenerative process in discrete time (equivalently a renewal reward process).

The sum over a cycle is the sum of the K ages $T_1 + \cdots + T_K$ (defined to be 0 if $K = 0$; the service was not preempted.) The cycle length is K.

Using regenerative process theory, we thus can express, for any non-negative measurable function f,

$$E(f(\hat{B})) = \frac{E\Big[\sum_{j=1}^{K} f(T_j)\Big]}{E(K)}. \tag{27}$$

K is not a stopping but $\tau = K + 1$ is;

$$\tau = \min\{n \geq 1 : T_n > S\},$$

the first time that S completes service. Thus we can compute the numerator of $E(f(\hat{B}))$ (expected sum over a cycle) using Wald's equation and subtracting the last piece:

$$E\Big[\sum_{j=1}^{K} f(T_j)\Big] = E(\tau)E(f(T)) - E(f(T_\tau)). \tag{28}$$

For $f(b) = b$ this then yields

$$E(\hat{B}) = \frac{\frac{1}{\lambda}E(e^{\lambda S}) - E(T_\tau)}{E(e^{\lambda S}) - 1}. \tag{29}$$

We next compute $E(T_\tau)$ and obtain (on its own) a very interesting result:

Lemma 3.1

$$E(T_\tau) = \frac{1}{\mu} + \frac{1}{\lambda} = E(S + T). \tag{30}$$

Proof : We re-write

$$E(T_\tau) = \sum_{n=1}^{\infty} E(T_\tau \ ; \ \tau = n), \tag{31}$$

where

$$E(T_\tau \ ; \ \tau = 1) = E(T_1; T_1 > S),$$

$$E(T_\tau \ ; \ \tau = n) = E(T_n \ ; \ T_n > S, T_{n-1} < S, \ldots, T_1 < S), \ n \geq 2.$$

But conditional on S, the T_j are i.i.d. and so conditional on S we have

$$E(T_\tau \ ; \ \tau = n \mid S) = E(T; T > S \mid S)(1 - e^{-\lambda S})^{n-1}, \ n \geq 1.$$

Thus conditional on S, using Equation 31 yields (via a geometric series)

$$E(T_\tau \mid S) = E(T\ ;\ T > S \mid S)\sum_{n=1}^{\infty}(1-e^{-\lambda S})^{n-1} = e^{\lambda S}E(T; T > S \mid S). \tag{32}$$

A quick calculation yields

$$E(T; T > S \mid S) = Se^{-\lambda S} + \frac{1}{\lambda}e^{-\lambda S},$$

and hence

$$E(T_\tau \mid S) = S + \frac{1}{\lambda},$$

and the result then follows by taking expected values,

$$E(T_\tau) = E(E(T_\tau \mid S)) = \frac{1}{\mu} + \frac{1}{\lambda}.$$

■

Inserting Equation 30 into Equation 29 then yields

Proposition 3.2 *For the M/G/1 PRI model*

$$E(\hat{B}) = \frac{1}{\lambda} - \frac{\frac{1}{\mu}}{E(e^{\lambda S}) - 1} \tag{33}$$

From Equation 33 for $E(\hat{B})$ we plug it into to obtain

Proposition 3.3 *For the M/G/1 PRI model*

$$p_0 = 2 - E(e^{\lambda S}).$$

Stability $0 < p_0 < 1$ *is thus*

$$E(e^{\lambda S}) < 2.$$

(In particular, S must have a finite moment generating function in a neighborhood of 0*.)*

3.1.3 A FIFO M/G/1 proof of stability

Here, we proceed along the same lines of what is explained in Remark 2.3.

Proposition 3.4 *The PRI model is stable if and only if* $E(e^{\lambda S}) < 2$.

Proof : Let $\mathbb{S}$ denote the total amount of work required by a customer (arriving with i.i.d. service time $S \sim G$) during their sojourn time. Then

$$\mathbb{S} \stackrel{\mathrm{d}}{=} S + \sum_{j=1}^{K} T_j,$$

that is, S itself plus all the ages added on due to preemption. This is the *effective* service time of a customer and they are i.i.d. among customers. If we operate the system under FIFO sequentially using these effective i.i.d. service times $\{\mathbb{S}_n : n \geq 1\}$ for each Poisson arrival, we have an ordinary FIFO M/G/1 queue with G distributed as $\mathbb{S}$. The two pieces (S and the sum) are not independent but we can still take the expected value:

$$E(\mathbb{S}) = E(S) + E\Big[\sum_{j=1}^{K} T_j\Big].$$

We already computed the second piece (added on ages) when proving Proposition 3.2:

$$E\Big[\sum_{j=1}^{K} T_j\Big] = \frac{1}{\lambda}E(e^{\lambda S}) - (\frac{1}{\mu} + \frac{1}{\lambda}),$$

so by adding on $E(S) = 1/\mu$ yields

$$E(\mathbb{S}) = \frac{1}{\lambda}E(e^{\lambda S}) - \frac{1}{\lambda},$$

and hence

$$\overline{\rho} \overset{\text{def}}{=} \lambda E(\mathbb{S}) = E(e^{\lambda S}) - 1.$$

Recalling that a standard M/G/1 queue is stable if and only if $\overline{\rho} < 1$, we get the result; $E(e^{\lambda S}) - 1 < 1$ if and only if $E(e^{\lambda S}) < 2$.

■

As a final observation, we note that using Proposition 3.1 reveals a nice intuitive interpretation of stability:

> *The PRI model is stable if and only if the expected number of times a customer is preempted is strictly less than* 1.

3.1.4 Computing the distribution of $\hat{B}$

Using the same Wald's equation method from Equation 28 we can determine the probability distribution of $\hat{B}$. For a fixed $x \geq 0$ we use $f(b) = I\{b > x\}$ to first compute the tail $P(\hat{B} > x)$. Then obtain the cdf via $1 - P(\hat{B} > x)$. We also obtain the density, denoted by $f_{\hat{B}}(x)$.

Proposition 3.5 *For the M/G/1 PRI model, the cumulative distribution function (cdf), $F_{\hat{B}}(x)$ of $\hat{B}$ is given by*

$$F_{\hat{B}}(x) = \frac{E(e^{\lambda S}) - e^{-\lambda x}E(e^{\lambda S}; S > x) - G(x)}{E(e^{\lambda S}) - 1}, \ x \geq 0.$$

$\hat{B}$ always has a density (it is always a continuous r.v.) given by

$$f_{\hat{B}}(x) = \frac{\lambda e^{-\lambda x}E(e^{\lambda S}; S > x)}{E(e^{\lambda S}) - 1}, \ x \geq 0.$$

Proof : From Equation 28, we have

$$\begin{aligned} P(\hat{B} > x) &= \frac{E\Big[\sum_{j=1}^{K+1} I\{T_j > x\}\Big] - P(T_\tau > x)}{E(K)} && (34) \\ &= \frac{E(e^{\lambda S})e^{-\lambda x} - P(T_\tau > x)}{E(e^{\lambda S}) - 1}. && (35) \end{aligned}$$

Similar to Equation 32 we obtain

$$\begin{aligned} P(T_\tau > x) &= E(e^{\lambda S} P(T > S, T > x \mid S)) && (36) \\ &= E(e^{\lambda S}(e^{-\lambda S} I\{S > x\} + e^{-\lambda x} I\{S \le x\})) && (37) \\ &= P(S > x) + e^{-\lambda x} E(e^{\lambda S}; S \le x). && (38) \end{aligned}$$

Since $E(e^{\lambda S}) - E(e^{\lambda S}; S \le x) = E(e^{\lambda S}; S > x)$, when we subtract $P(T_\tau > x)$ in Equation 35 we obtain

$$P(\hat{B} > x) = \frac{e^{-\lambda x} E(e^{\lambda S}; S > x) - P(S > x)}{E(e^{\lambda S}) - 1}.$$

Using $F_{\hat{B}}(x) = 1 - P(\hat{B} > x)$ then yields the cdf. Because of the presence of $G(x) = P(S \le x)$ in the cdf formula, it appears at first sight that we need G to have a density $g(x)$ (so that we can differentiate G) in order to obtain a density for $\hat{B}$, so we will first assume that G has such a density, just so as to quickly get the formula for $f_{\hat{B}}(x)$ given in the Proposition, but then we will show that the existence of $g(x)$ is not required. Writing out

$$E(e^{\lambda S}; S \le x) = \int_0^x e^{\lambda s} g(s) ds,$$

we observe its derivative with respect to x is $e^{\lambda x} g(x)$ so $f_{\hat{B}}(x) = F'_{\hat{B}}(x)$ yields the density in our Proposition. But this formula for the density does not contain $g(x)$ in it. So we will integrate our density formula for $f_{\hat{B}}(x)$ and see that we always get back out our cdf formula thus proving that the existence of a density for G is not required. That is, we will show that

$$\int_0^x f_{\hat{B}}(y) dy = F_{\hat{B}}(x);$$

equivalently that for $c = E(e^{\lambda S}) - 1$, that

$$c \int_0^x f_{\hat{B}}(y) dy = c F_{\hat{B}}(x).$$

To this end due to non-negativity, we use Fubini's (Tonelli's) Theorem via

$$\begin{aligned}
\int_0^x cf_{\hat{B}}(y)dy &= \int_0^x \lambda e^{-\lambda y} E(e^{\lambda S}) I\{S > y\} dy \\
&= E(e^{\lambda S} \int_0^x \lambda e^{-\lambda y} I\{S > y\})dy \\
&= E(e^{\lambda S} \int_0^{\min\{x,S\}} \lambda e^{-\lambda y})dy \\
&= E(e^{\lambda S}(1 - e^{-\lambda \min\{x,S\}})) \\
&= E(e^{\lambda S}) - E(e^{\lambda S} e^{-\lambda \min\{x,S\}}) \\
&= E(e^{\lambda S}) - E(e^{\lambda S}; S > x) - P(S \leq x) \\
&= cF_{\hat{B}}(x).
\end{aligned}$$

■

Interestingly, when G is exponential at rate μ (M/M/1 case), we get

$$f_{\hat{B}}(x) = \mu e^{-\mu x}; \tag{39}$$

$\hat{B}$ is exponential at rate μ, the same as G. Hence $E(\hat{B}) = \frac{1}{\mu}$.

As a check in general, it is easy to verify that Equation 33 can be derived using the density in Proposition 3.5;

$$\int_0^\infty x f_{\hat{B}}(x) dx = \frac{1}{\lambda} - \frac{\frac{1}{\mu}}{E(e^{\lambda S}) - 1}.$$

The following holds using exactly the same proof given in Proposition 2.3 for the PRD model:

Proposition 3.6 *When stable, N has a geometric distribution with success probability p_0:*

$$P(N = n) = (1 - p_0)^n p_0, \ n \geq 0.$$

3.1.5 Computing $E(\hat{S})$ and $E(\hat{S}_r)$ of a service time preempted from service

We also can also compute $E(\hat{S})$ which represents the mean (in stationarity) of the total service time found in service by a customer who preempts them.

Proposition 3.7 *For the M/G/1 PRI model*

$$E(\hat{S}) = \frac{E(Se^{\lambda S}) - \frac{1}{\mu}}{E(e^{\lambda S}) - 1} \tag{40}$$

Proof : We directly compute it as (similar to how we computed $E(\hat{B})$, but no need to use Wald's equation):

$$E(\hat{S}) = \frac{E\Big[\sum_{j=1}^{K} S\Big]}{E(K)}.$$

Conditioning the numerator on S yields it as

$$E(K \mid S)S = (e^{\lambda S} - 1)S = Se^{\lambda S} - S.$$

Taking expected values then dividing by $E(K) = E(e^{\lambda S}) - 1$ yields the result. ■

Finally since $E(\hat{S}) = E(\hat{B}) + E(\hat{S}_r)$ we can compute the mean remaining service time $E(\hat{S}_r) = E(\hat{S}) - E(\hat{B})$ using Equations 40 and 33:

Proposition 3.8 *For the M/G/1 PRI model*

$$E(\hat{S}_r) = \frac{E(Se^{\lambda S})}{E(e^{\lambda S}) - 1} - \frac{1}{\lambda}. \tag{41}$$

When we apply Proposition 3.7 to the M/M/1 case, recalling Equation 39, we get

$$E(\hat{S}) = \frac{1}{\mu} + \frac{1}{\mu - \lambda},$$

which implies that

$$E(\hat{S}_r) = \frac{1}{\mu - \lambda};$$

in particular, this shows that $\hat{B}$ and $\hat{S}_r$ do not have the same distribution. We will determine the distribution of $\hat{S}_r$ and even the joint distribution of $(\hat{B}, \hat{S}_r)$ next.

3.2 Distributions of $\hat{S}$, $\hat{S}_r$ and the joint distribution of $(\hat{B}, \hat{S}_r)$

Proposition 3.9 *The cumulative distribution function (cdf) of $\hat{S}$, $F_{\hat{S}}(x) = P(\hat{S} \leq x)$ is given by*

$$F_{\hat{S}}(x) = \frac{E(e^{\lambda S}; S \leq x) - G(x)}{E(e^{\lambda S}) - 1}, x \geq 0.$$

In particular, if G has a density $g(x)$ then so does $\hat{S}$ and it is given by (via differentiation)

$$f_{\hat{S}}(x) = \frac{g(x)(e^{\lambda x} - 1)}{E(e^{\lambda S}) - 1}, x \geq 0.$$

Proof :

Following the proof of Proposition 3.7, we can express

$$F_{\hat{S}}(x) = \frac{E(E(K \mid S)I\{S \leq x\})}{E(e^{\lambda S}) - 1}, x \geq 0.$$

The numerator becomes

$$E(e^{\lambda S}; S \leq x) - G(x);$$

the cdf follows. With density $g(x)$,

$$E(e^{\lambda S}; S \leq x) = \int_0^x g(s)e^{\lambda s} ds,$$

thus differentiating the numerator yields

$$e^{\lambda x}g(x) - g(x) = g(x)(e^{\lambda x} - 1).$$

■

We now proceed to obtain the distribution of $\hat{S}_r$ and the joint distribution of $(\hat{B}, \hat{S}_r)$. $\overline{G}(x) = P(S > x)$, denotes the tail of G.

Proposition 3.10 *The cumulative distribution function (cdf) of $\hat{S}_r$, $F_{\hat{S}_r}(x) = P(\hat{S}_r \leq x)$ is given by*

$$F_{\hat{S}_r}(x) = \frac{e^{\lambda x}\overline{G}(x) + E(e^{\lambda S}; S \leq x) - 1}{E(e^{\lambda S}) - 1}, x \geq 0.$$

$\hat{S}_r$ always has a density (it is alway a continuous r.v.) and it is given by

$$f_{\hat{S}_r}(x) = \frac{\lambda e^{\lambda x}\overline{G}(x)}{E(e^{\lambda S}) - 1}, x \geq 0.$$

Proof : (Sketch) Here we follow the Wald's equation method used in Proposition 3.2. We have

$$E(f(\hat{S}_r)) = \frac{E\Big[\sum_{j=1}^{K} f(S - T_j)\Big]}{E(K)}.$$

Using $f(b) = I\{b \leq x\}$ we have

$$P(\hat{S}_r \leq x) = \frac{E\Big[\sum_{j=1}^{K} I\{S - T_j \leq x\}\Big]}{E(K)}.$$

We note that *conditional on* S, the sum

$$\sum_{n}^{K+1} f(S - T_n)$$

is a stopping time sum of i.i.d. random variables and hence has conditional (given S) expected value

$$E(K + 1 \mid S)Ef(S - T_j \mid S) = e^{\lambda S}Ef(S - T_j \mid S).$$

Taking expected values then subtracting $E(f(S - T_\tau))$ then yields then numerator which we then divide by $E(K) = E(e^{\lambda S}) - 1$. To obtain the density, we use the same trick that we used in the proof of Proposition 3.5: We first assume that G has a density $g(x)$ so as to obtain our formula for the density $f_{\hat{S}_r}(x)$ by differentiation, and then by integration, show that $g(x)$ is not required.

■

In the M/M/1 case, the above yields that $\hat{S}_r$ is exponential at rate $\mu - \lambda$. Recalling that in this case we also have that $\hat{B}$ is exponential at rate μ, it begs the question of whether in this M/M/1 case, $\hat{S}_r$ and $\hat{B}$ are independent. We will answer that question in the affirmative next by computing, in general, the joint distribution of $(\hat{B}, \hat{S}_r)$.

Proposition 3.11 *For the M/G/1 PRI model,*

$$P(\hat{B} > x,\ \hat{S}_r > y) = \frac{e^{-\lambda x}E(e^{\lambda S}; S > x+y) - e^{\lambda y}P(S > x+y)}{E(e^{\lambda S}) - 1},\ x \geq 0, y \geq 0.$$

Thus, if G has a density g, then the joint density of $(\hat{B}, \hat{S}_r)$ exists and is given by

$$f_{(\hat{B},\hat{S}_r)}(x,y) = \frac{\lambda e^{\lambda y} g(x+y)}{E(e^{\lambda S}) - 1},\ x \geq 0, y \geq 0.$$

Proof : We again use the regenerative method and conditional Wald's equation via the expression

$$P(\hat{B} > x,\ \hat{S}_r > y) = \frac{E\Big[\sum_{n=1}^{K+1} I\{T_n > x, S - T_n > y\}\Big] - P(T_\tau > x, S - T_\tau > y)}{E(K)}.$$

First off, it is immediate that $P(T_\tau > x, S - T_\tau > y) = 0$ since by definition of τ it must hold that $T_\tau > S$ which excludes $T_\tau < S - y$. Thus the numerator conditional on S becomes

$$e^{\lambda S} P(x < T < S - y \mid S).$$

It must hold that $S > x + y$ or else $P(x < T < S - y \mid S) = 0$; thus we end up with

$$e^{\lambda S}(e^{-\lambda x} - e^{-\lambda(S-y)}) I\{S > x+y\} = e^{-\lambda x} e^{\lambda S} I\{S > x+y\} - e^{\lambda y} I\{S > x+y\}.$$

Taking expected values and dividing by $E(K)$ then yields the joint tail result.

To obtain the density we compute it as $\frac{\partial}{\partial y}\frac{\partial}{\partial x} P(\hat{B} > x,\ \hat{S}_r > y)$: Letting $c = E(e^{\lambda S}) - 1$, we have (after a cancellation via $-e^{\lambda y} g(x+y) + e^{\lambda y} g(x+y) = 0$)

$$c\frac{\partial}{\partial x} P(\hat{B} > x,\ \hat{S}_r > y) = -\lambda e^{-\lambda x} E(e^{\lambda S}; S > x+y).$$

Then

$$-\frac{\partial}{\partial y} \lambda e^{-\lambda x} E(e^{\lambda S}; S > x+y) = \lambda e^{\lambda y} g(x+y).$$

■

As the reader can check: using Propositions 3.10 and 3.5 (using the tail), it is confirmed that $P(\hat{B} > x) = P(\hat{B} > x,\ \hat{S}_r > 0)$ and $P(\hat{S}_r > y) = P(\hat{B} > 0,\ \hat{S}_r > y)$ as should be. Moreover, as promised, in the M/M/1 case Proposition 3.11 yields

$$P(\hat{B} > x,\ \hat{S}_r > y) = e^{-\mu x} \times e^{-(\mu - \lambda)y};$$

$\hat{B}$ and $\hat{S}_r$ are independent exponentials. (The stability condition for the M/M/1 PRI is easily seen to be $\rho < 1/2$, i.e. $\lambda < \mu/2$)

We now present some relationships between the tails of $\hat{S}_r$ and $\hat{B}$.

Corollary 3.1 *For a stable M/G/1 PRI model it holds that*

$$P(\hat{S}_r > x) = e^{\lambda x} P(\hat{B} > x),\ x \geq 0, \tag{42}$$

and hence

1. *$\hat{S}_r$ is stochastically larger than $\hat{B}$; that is,*

$$P(\hat{S}_r > x) \geq P(\hat{B} > x),\ x \geq 0.$$

2. *If $P(S > x) > 0$ for all x, then*

$$\frac{P(\hat{S}_r > x)}{P(\hat{B} > x)} \to \infty \text{ as } x \to \infty;$$

the tail of $\hat{S}_r$ is heavier the the tail of $\hat{B}$.

Proof : Using Proposition 3.11 to compute the individual tails, we have

$$P(\hat{S}_r > x) = \frac{E[e^{\lambda S}; S > x] - e^{\lambda x} P(S > x)}{E(e^{\lambda S}) - 1},$$

while

$$P(\hat{B} > x) = \frac{e^{-\lambda x} E[e^{\lambda S}; S > x] - P(S > x)}{E(e^{\lambda S}) - 1},$$

immediately yielding Equation 42, and from which the next two statements directly follow. ■

3.3 $\hat{S}_r$ for PRI can be heavy-tailed even though S must be light-tailed

Corollary 3.1 begs the question as to just how heavy the tail of $\hat{S}_r$ can be. Recall from Propositions 3.3 and 3.4 the stability condition for PRI, $E(e^{\lambda S}) < 2$. In particular S must be light-tailed, by which we mean here that it has a finite moment generating function in a neighborhood of the origin; in this case $E(e^{sS}) < \infty,\ s \in (0, \lambda]$. What about $\hat{S}_r$? We give here an example where $E(e^{\lambda S}) < 2$ but

$$E(e^{s\hat{S}_r}) = \infty,\ s > 0; \tag{43}$$

$\hat{S}_r$ is heavy-tailed.

In what follows we let $M_X(s) = E(e^{sX}),\ s \geq 0$, denote the moment generating function of a non-negative r.v. X, and we note that it can be computed in terms of the tail $P(X > x)$ via

$$\int_0^\infty e^{sx} P(X > x) dx = \frac{1}{s}(M_X(s) - 1). \tag{44}$$

From Proposition 3.10, the density of $\hat{S}_r$ is given by

$$f_{\hat{S}_r}(x) = \frac{\lambda e^{\lambda x} P(S > x)}{E(e^{\lambda S}) - 1},\ x \geq 0.$$

Thus using Equation 44 yields

$$M_{\hat{S}_r}(s) = b(M_S(s + \lambda) - 1), \tag{45}$$

where

$$b = \frac{\lambda}{(s + \lambda)(E(e^{\lambda S}) - 1)}.$$

This implies that we need, for our example, an S such that $M_S(\lambda) < \infty$ but $M_S(\lambda + s) = \infty,\ s > 0$. We consider density functions for S of the form

$$g(x) = c(p)e^{-\lambda x}(1 + x)^{-p},\ p > 1,$$

where $c(p)$ is the normalizing constant, given explicitly by

$$c(p) = (p - 1)e^{\lambda^2/2}.$$

It is immediate that with this type of $g(x)$, it always holds that $M_S(\lambda) < \infty$ and $M_S(\lambda + s) = \infty,\ s > 0$. But we need to ensure stability, that is, find a value of λ such that

$$M_S(\lambda) = c(p)\int_0^\infty (1 + x)^{-p}dx = e^{\lambda^2/2} < 2.$$

We note that $e^{\lambda^2/2} < 2$ if and only if $\lambda < \sqrt{\ln(4)} \approx 1.177$.

Remark 3.1 By using the relation in Equation 45 and the fact that $P(\hat{B} > x) = e^{-\lambda x}P(\hat{S}_r > x)$ (recall Corollary 3.1), it is straightforward to prove that $M_{\hat{B}}(s) < \infty,\ s < \lambda$; stationary age is always light-tailed. But interestingly, $M_{\hat{B}}(\lambda) = 1 + \lambda E(\hat{S}_r)$, and hence involves $E(Se^{\lambda S})$, from Proposition 3.8, and hence could be infinite: $M_{\hat{B}}(\lambda) < \infty$ if and only if $E(\hat{S}_r) < \infty$.

3.3.1 Expected sojourn time

As with the PRD model, we can use Little's law ($l = \lambda w$) to obtain expected sojourn time.

$$l = E(N) = \frac{1 - p_0}{p_0} = \lambda w, \tag{46}$$

and hence that

$$w = \frac{1}{\lambda}\Big[\frac{1 - p_0}{p_0}\Big] = \Big[\frac{1}{2 - E(e^{\lambda S})} - 1\Big]. \tag{47}$$

We note that this formula (using different methods) was given as Theorem 2 on Page 6 in [1] (Equation 2.8).

Remark 3.2 In all the PL models, the distribution of sojourn time W is identical to that of a busy period. So Equation 47 is also the expected value of a busy period in the M/G/1 PRI case. Moreover, from PASTA, p_0 is equal to the long-run proportion of arrivals who find the system empty (hence begin a busy period, hence begin a regenerative cycle). Thus by regenerative process theory,

$$p_0 = \frac{1}{E(N_B)},$$

where N_B denotes the number of customers served during a busy period. Using Equation 3.3, we thus can solve for $E(N_B)$ yielding

$$E(N_B) = \frac{1}{p_0} = \frac{1}{2 - E(e^{\lambda S})}. \tag{48}$$

3.4 Deriving the distribution of V for PRI

From Proposition 3.6, and Proposition 3.3 the stationary distribution of number in system N is geometric with success probability $p_0 = 2 - E(e^{\lambda S})$. A perusal of the standard M/G/1 PL queue proof that we used to determine the geometric distribution property also yields that conditional on $N = n$, the n service times (whole) are i.i.d. distributed as $\hat{S} = \hat{B} + \hat{S}_r$. So we can extract from this a representation of stationary workload. In what follows the rv $\hat{Q}$ and the i.i.d. service times $\{\hat{S}_j\}$, and the random variable $\hat{S}_r$, are taken as independent (with their distributions provided in Propositions 3.9 and 3.10) with $\hat{Q}$ having a geometric distribution (with mass at 0) and success probability $p_0 = 2 - E(e^{\lambda S})$.

Proposition 3.12 *For the M/G/1 PRI model,*

$$(V \mid V > 0) \overset{\mathrm{d}}{=} \hat{S}_r + \sum_{j=1}^{\hat{Q}} \hat{S}_j. \tag{49}$$

Thus the distribution of V, F_V, is a mixture

$$F_V = p_0 \delta_0 + (1 - p_0) F_{\hat{V}}$$

where $F_{\hat{V}}$ denotes the distribution of $(V \mid V > 0)$ given in Equation 49.

From the above, we can now compute $E(V)$, for example, by utilizing the solved value of p_0 and the various expected value results and the fact that $\hat{Q}$ is geometric:

$$E(V) = (1 - p_0)\Big[E(\hat{S}_r) + E(\hat{Q})E(\hat{S})\Big].$$

$$= (E(e^{\lambda S}) - 1)\Big[\frac{E(Se^{\lambda S})}{E(e^{\lambda S}) - 1} - \frac{1}{\lambda} + \big[\frac{1}{2 - E(e^{\lambda S})} - 1\big]\big[\frac{E(Se^{\lambda S}) - \frac{1}{\mu}}{E(e^{\lambda S}) - 1}\big]\Big].$$

For the M/M/1 case this becomes

$$E(V) = \frac{\rho}{1 - \rho}\Big[\frac{1}{\mu - \lambda} + \frac{\rho}{1 - 2\rho}(\frac{1}{\mu} + \frac{1}{\mu - \lambda})\Big].$$

3.5 Heavy-traffic limits for stationary workload V

Here we include a result for PRI in the same spirit as in Theorem 2.1 but requiring more care in its proof do to the additional complexity of stationary workload for PRI as compared to PRD. Of importance too is the fact that unlike PRI, there may not be a solution $\lambda_2 > \lambda$ to $p_0(\lambda) = 0$ (equivalently for PRI a solution to $E(e^{\lambda S}) = 2$); recall the counterexample given in Section 3.3.

Theorem 3.1 *Suppose that there exists a $\lambda_2 > \lambda$ such that as $\lambda \uparrow \lambda_2$, $E(e^{\lambda S}) \to E(e^{\lambda_2 S}) = 2$, and that $E(Se^{\lambda_2 S}) < \infty$. Then as $\lambda \uparrow \lambda_2$*

$$p_0 V \Longrightarrow exp(\alpha),$$

where $\alpha^{-1} = E(\hat{S}(\lambda_2)) = E(Se^{\lambda_2 S}) - 1/\mu$.

Proof : The proof follows the same steps as in Theorem 2.1, while now using the workload representation from Proposition 3.12, but there is additional complexity in establishing the convergence in probability as in Equation 19 due to the fact that $\hat{S} = \hat{S}(\lambda)$ depends on λ, and iid copies of it are included in the sum; we need to establish

$$\frac{\sum_{i=1}^{N(\lambda)} \hat{S}_i(\lambda)}{N(\lambda)} \overset{p}{\Longrightarrow} E(\hat{S}_i(\lambda_2)), \tag{50}$$

as $\lambda \uparrow \lambda_2$.

To get around this, we first construct the rvs by using the inverse transform method from simulation to couple them (see for example, Pages 37-38 in [2]): Let $U, U_1, U_2 \ldots$ denote iid copies of $Unif(0,1)$ rvs (one can take, for example, as the probability space for a uniform U, the interval $(0,1)$ under Lesbesgue measure and take $U(\omega) = \omega$, the identity map). We use the same U for constructing $N(\lambda)$ for all $\lambda < \lambda_2$, and we use U_i for constructing $\hat{S}_i(\lambda)$ for all $\lambda \leq \lambda_2$. Then we re-write

$$\frac{\sum_{i=1}^{N(\lambda)} \hat{S}_i(\lambda)}{N(\lambda)} = \frac{\sum_{i=1}^{N(\lambda)} \hat{S}_i(\lambda_2)}{N(\lambda)} + A, \tag{51}$$

where

$$A = \frac{1}{N(\lambda)} \sum_{i=1}^{N(\lambda)} (\hat{S}_i(\lambda_2) - \hat{S}_i(\lambda)),$$

the error term. If we can show that A converges to 0 in probability as $\lambda \uparrow \lambda_2$, hence can be ignored, then in the right-hand side of Equation 51 the iid summands only depend on λ_2, and we get the result exactly as was done for the PRD case of Theorem 2.1. To this end, observe that since $N(\lambda)$ is independent of the iid $\hat{S}_i(\lambda_2) - \hat{S}_i(\lambda)$, we have (via first conditioning on $N(\lambda)$) that

$$E(|A|) \leq E(|\hat{S}(\lambda_2) - \hat{S}(\lambda)|),$$

where $\hat{S}(\lambda_2) - \hat{S}(\lambda)$ denotes $\hat{S}_1(\lambda_2) - \hat{S}_1(\lambda)$.

It thus suffices to show that $E(|\hat{S}(\lambda_2) - \hat{S}(\lambda)|) \to 0$.

We now note that from the coupling using the inverse transform method (and the fact that from Proposition 3.9, it follows that $\hat{S}(\lambda)$ converges in distribution to $\hat{S}(\lambda_2)$ as $\lambda \uparrow \lambda_2$), that $\hat{S}(\lambda) \to \hat{S}(\lambda_2)$ wp1 as $\lambda \uparrow \lambda_2$, and thus $|\hat{S}(\lambda_2) - \hat{S}(\lambda)| \to 0$, wp1.

It thus suffices to show that the collection $\{|\hat{S}(\lambda_2) - \hat{S}(\lambda)|\}$ is uniformly integrable (in $\lambda \leq \lambda_2$). We do so by noting further that

$$|\hat{S}(\lambda_2) - \hat{S}(\lambda)| \leq X(\lambda) = \hat{S}(\lambda_2) + \hat{S}(\lambda),$$

and $\{X(\lambda)\}$ is uniformly integrable: It is a non-negative collection such that $X(\lambda) \to X(\lambda_2) = 2\hat{S}(\lambda_2)$ wp1, and $E(X(\lambda)) \to E(X(\lambda_2)) = 2E(\hat{S}(\lambda_2)) < \infty$ (finite by assumption). ■

References

[1] Asmussen, S. & Glynn, P.W. (2017). On preemptive-repeat LIFO queues. Springer.

[2] Asmussen, S. & Glynn, P.W. (2007). *Stochastic Simulation*, **87**, 1-22.

[3] Fakinos, D. (1981). The G/G/1 Queueing System with a Particular Queue Discipline. *J.R. Statist. Soc.*, **B43**, 190-196.

[4] Fakinos, D. (1986). On the Single-Server Queue with the Preemptive-Resume Last-Come- First-Served Queue Discipline. *Journal of Applied Probability*, **23**, 243-248.

[5] Sigman, K. (1995). *Stationary Marked Point Processes: An Intuitive Approach.* Chapman and Hall. New York.

[6] Wolff, R.W. (1989). *Stochastic Modeling And The Theory Of Queues.* Prentice Hall. New Jersey.

[7] Woods, F.S. (1926). *Advanced Calculus: A Course Arranged with Special Reference to the Needs of Students of Applied Mathematics.* Springer-Verlag. New York. 148-151.