The age of the arrival process in the G/M/1 and M/G/1 queues

Moshe Haviv* and Yoav Kerner†

Abstract

This paper shows that in the G/M/1 queueing model, conditioning on a busy server, the age of the inter-arrival time and the number of customers in the queue are independent. Explicit expressions for the density functions of this age conditioning on a busy server and conditioning on an idle server are given. Moreover, we show that this independence property, which we prove by elementary arguments, also leads to an alternative proof for the fact that given a busy server, the number of customers in the queue follows a geometric distribution. Also, we show that the residual inter-arrival time and the number of customers in the system given that the server is busy are independent. Moreover, we develops an explicit form for the density function of the conditional residual inter-arrival time given a busy and given an idle server. This is also repeated for the inter-arrival time itself. We conclude with a derivation for the Laplace Stieltjes Transform (LST) of the age of the inter-arrival time in the M/G/1 queue.

1 Introduction

Consider the G/M/1 queueing model. Let $G(x)$ be the cumulative distribution function of the inter-arrival time and let μ be the rate of service. Assume

*Department of Statistics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel. E-mail: haviv@huji.ac.il.

†EURANDOM, Eindhoven University of Technology, Eindhoven, The Netherlands. E-mail: kerner@eurandom.tue.nl.

$1/\mu < \overline{x}$ where the latter is the mean inter-arrival time. Denote $(\mu\overline{x})^{-1}$ by ρ and note that ρ is the limit probability that the server is busy. Let $G^*(s)$ be the LST of the inter-arrival time, namely $G^*(s) = \int_{x=0}^{\infty} e^{-sx} g(x)\, dx$ where $g(x)$ is the density function of the inter-arrival time. Also, let σ be the unique value between zero and one obeying $\sigma = G^*(\mu(1-\sigma))$. For future shortening, let $\eta = \mu(1-\sigma)$.
Assume steady-state conditions. Let π_i be the steady-state probability that an arrival finds i customers at the system upon arrival. It is well-known, see e.g. [4], p. 208, that $\pi_i = (1-\sigma)\sigma^i$, $i \geq 0$. Also, let e_i be the steady-state probability for having i customers in the system at a random time. Then, $e_0 = 1-\rho$ and $e_i = \rho(1-\sigma)\sigma^{i-1}$, $i \geq 1$. Note that given a busy server, the probability at random times that the number of customers in the queue equals i equals the probability that an arrival finds i in the system upon his arrival, $i \geq 0$. Recall that in the case of a first-come first-served (FCFS) regime, the sojourn time in a G/M/1 queue follows an exponential distribution with parameter η (see e.g., [4], p.229).

Consider the age of the inter-arrival time. By that we mean the time elapsed since the previous arrival which of course coincides with the past sojourn time (under FCFS) of the last one to arrive (in case there is at least one there). Denote this random variable by A. It is standard by renewal theory to observe that the corresponding density $f_A(a)$ equals $\overline{G}(a)/\overline{x}$ where $\overline{G}(a) = 1 - G(a)$. Also, denoting by $A^*(s)$ its LST, it is well-known that $A^*(s) = (1 - G^*(s))/(\overline{x}s)$. In particular, by trivial algebra,

$$A^*(\eta) = \mathrm{E}(e^{-\eta A}) = \rho = (\mu\overline{x})^{-1}. \tag{1}$$

Finally, note that the residual of the inter-arrival time, namely the time until the next arrival, follows the same distribution as its age.

Of course, any information given on the system itself changes this prior distribution of the age. This is, for example, the case if the number of customers in the system is given. Indeed, below we derive the conditional density of A given this number. As it turns out, for any positive number of customers in the system, the conditional density is the same. Put differently, given that the server is busy, the age of the inter-arrival time and the number of customers in the queue are independent. This result, which is proved in the next section, looks quite counter intuitive, as one might think that the larger the number of customers is, the smaller is the time elapsed since the last

arrival. Our proof is coupled with an explicit expression for the distribution of the age of the inter-arrival time under these two cases. It is then shown that this independence property is sufficient for the number of customers in the queue, given a busy server, to follow a geometric distribution. As we have established this independence property by primitive arguments, this can serve as an alternative proof for the geometric distribution phenomenon in the G/M/1 queue. We conclude Section 2 with deriving the corresponding results for the conditional residual and total length of the inter-arrival time, both enjoying the same above-mentioned independence property. Note that although the age and the residual of the inter-arrival time are identically distributed, this by no means is carried over to the conditional distributions. Finally, in Section 3 we derive the LST of the age of the inter-arrival time given the number of customers in the system for the M/G/1 queue. The conditional distribution of the past service time of the one currently in service, given the number of customers in the system in the M/G/1 queue, and the joint transform of these two random variables, were analyzed in Adan and Haviv [2].

2 The G/M/1 case

2.1 The conditional age of inter-arrival time

This section deals with the conditional age, residual and total length of the inter-arrival time, given the number of customers in the system for the G/M/1 model. Note that we do not need to specify the queueing regime and it can be any regime as long as it is work-conserving (namely, the total work in the system is as under the FCFS regime) and it is not-anticipating (namely, the decision who gets service and when (preemption included), does not depend on actual service requirements).

Theorem 2.1 *Let $f_{A|L=n}(a)$ be the density function of the age of the inter-arrival time, given that the number of customers in the system (a random variable denoted by L) equals n, where $a, n \geq 0$. Then, in the G/M/1 queue,*

$$f_{A|L=n}(a) = \overline{G}(a)\mu e^{-\eta a}, \quad a \geq 0, \ n \geq 1. \tag{2}$$

In particular, given that the server is busy, the age of the inter-arrival time

and the number of customers in the queue are independent. Also,

$$f_{A|L=0}(a) = \frac{\overline{G}(a)}{\overline{x}} \frac{1-e^{-\eta a}}{1-\rho}, \quad a \geq 0. \tag{3}$$

Proof. Recall that the probability to find m customers in the queue upon arrival equals $(1-\sigma)\sigma^m$, $m \geq 0$, and the probability of such an event at random times is $\rho(1-\sigma)\sigma^{m-1}$ for $m \geq 1$, and $1-\rho$ for $m = 0$. Thus, for the case where $n \geq 1$,

$$f_{A|L=n}(a) = \frac{f_A(a)}{\mathrm{P}(L=n)} \mathrm{P}(L=n|A=a)$$

$$= \frac{\frac{\overline{G}(a)}{\overline{x}}}{\rho(1-\sigma)\sigma^{n-1}} \sum_{m=n-1}^{\infty} (1-\sigma)\sigma^m e^{-\mu a} \frac{(\mu a)^{m-n+1}}{(m-n+1)!}.$$

Note that the summation is based on the fact that in order to see n, $n \geq 1$, customers given that the age of the inter-arrival time equals a and that the previous arrival saw $m \geq n-1 \geq 0$ upon arrival, one needs to have exactly $m-n+1$ service completions during the past a units of time. The probability of this event is $e^{-\mu a}(\mu a)^{m-n+1}/(m-n+1)!$. Some trivial algebra now concludes the proof of (2). As for proving (3), note that

$$f_A(a) = \frac{\overline{G}(a)}{\overline{x}} = (1-\rho) f_{A|L=0}(a) + \rho f_{A|L\geq 1}(a).$$

Since $f_{A|L\geq 1}(a)$ appears in (2), the rest is just simple algebra. □

Remark 2.1 Let Y be a random variable which follows the same distribution as that of an inter-arrival time and let S follow an exponential distribution with parameter η (as the sojourn time in G/M/1 in case of FCFS). Assume that Y and S are independent. Then, it is possible to see that the density function given in (2) is as that of $S|S \leq Y$. Thus, $A|L \geq 1$ is distributed as the sojourn time of the last customer in a busy period.

Remark 2.2 As there is no distribution function $G(x)$, for which the expressions in (2) and in (3) coincide for all values of a, $a \geq 0$, we conclude that in the G/M/1 model, the age of the inter-arrival time and the number of customers in the system are never independent.

Remark 2.3 Next we give a stand-alone proof for the independence property mentioned in Theorem 2.1, namely a proof which does use the fact that conditioning on a busy server, the queue length follows a geometric distribution. Consider the last-come first-served G/G/1 queue with the preemption-resume service policy (LCFS-PR). We next argue that the independence property mentioned in Theorem 2.1 applies to this model too. Specifically, assume the server is busy and tag the customer who currently receives service. It is clear that any event or random variable, say Y, which is defined on the period which lasts from the arrival of the tagged customer until his/her departure, and the number of customers he/she has found upon arrival, are independent. Note that the latter random variable coincides with the current number of customers in the queue. An example for Y is the age of the inter-arrival time. Hence, this random variable and the queue length, given a busy server, are independent. This property also holds in the G/M/1 queue under any work-conserving and non-anticipating queue regime since at all of them the number of customers in the system follows the same distribution. For more on the G/G/1 LCFS-PR and such intuitive arguments see [9].

Example: M/M/1. In the case where the arrival process is Poisson, i.e., the inter-arrival times follow an exponential distribution, $\overline{G}(x) = e^{-\lambda x}$ for some constant $\lambda < \mu$. Moreover, $\overline{x} = \lambda^{-1}$, $\rho = \lambda/\mu$, $G^*(s) = \lambda/(\lambda + s)$ and $\sigma = \rho$. In this case, (2) and (3) become

$$f_{A|L=n}(a) = \mu e^{-\mu a}, \quad a \geq 0, \ n \geq 1, \tag{4}$$

and

$$f_{A|L=0}(a) = \frac{\lambda\mu}{\mu - \lambda}(e^{-\lambda a} - e^{-\mu a}), \quad a \geq 0, \tag{5}$$

respectively. Note that (4) is an exponential density function with parameter μ, where (5) is the density function of the sum of two independent and exponentially distributed random variables, one with parameter λ and one with parameter μ. These special cases for the M/M/1 queue are not a surprise given the fact that a M/M/1 queue length process is time-reversible. Specifically, since in the time-reversed process an arrival corresponds to a departure in the original process, and vice versa, the age of the inter-arrival time in the original process corresponds to its residual time until departure in the time-reversed process. Thus, by the fact that the process is time-reversible, the

time until the next departure in the time-reversed process follows an exponential distribution with parameter μ in case of a busy server. In case of an idle server, one needs to add an independent arrival time which follows an exponential distribution with parameter λ.

Remark 2.4 It is interesting to observe that in the M/M/1 case, given that the server is busy, the distribution of the age of the inter-arrival time is a function of the arrival rate λ only through the requirement that it should be smaller than the service rate μ (and not through its exact value).

Application: A two servers system where the first has a finite buffer. Consider the following queueing model. Two servers give service to a common arrival process, which is assumed to be Poisson. Service times follow an exponential distribution with server dependent rates. The arrivers join the queue in front of the first server as long as it is shorter than some agreed threshold (can be size of the waiting room). Otherwise, they join the other line. No regrets take place afterwards and each customer receives service from the server whose line he/she joins upon arrival. Note that arrival epochs to the second queue (which are arrivals epochs to the system when the first queue is full) are independent, as the number of customers in the first queue is a Markov process, and hence given that the buffer in front of the first server is full, the past and the next time that an arrival finds it full, are independent. Moreover, the times between such consequent epochs are identically distributed (again, due to the Markovian property). Thus, the second line process follows a G/M/1 model. In [7] (see also [3]) it was stated and proved that given that the second server is busy, the two queue lengths are independent. The proof there was based on deriving explicitly the limit probabilities of the corresponding two dimensional Markov process, from which this independence property is easily observed. Theorem 2.1 leads to an alternative and much simpler proof. Specifically, let $L1$ and $L2$ denote the number of customers in front of server 1 and server 2, respectively. Also, let A be the age of the inter-arrival time to $L2$. It is easy to see (for example, by conditioning on the number of customers in front of server 2 at the instant of the last arrival to this queue) that given A, $L1$ and $L2$ are independent. Thus, given $\{A, L2 \geq 1\}$, $L1$ and $L2$ are independent too. But, by Theorem 2.1, given $L2 \geq 1$, $L2$ and A are independent. Hence, given $L2 \geq 1$, $L1$ and $L2$ are independent too. We write the latter chain in terms

of probabilities (although in a non formal way) to clarify the idea. First,

$$\mathrm{P}(L1, L2|A) = P(L1|A)\mathrm{P}(L2|A). \tag{6}$$

Then,

$$\mathrm{P}(L1, L2|A, L2 \geq 1) = \mathrm{P}(L1|A, L2 \geq 1)P(L2|A, L2 \geq 1)$$

$$\mathrm{P}(L1|A)P(L2|A, L2 \geq 1) = \mathrm{P}(L1|A)\mathrm{P}(L2|L2 \geq 1)$$

where the last equality is due to Theorem 2.1.

In the following theorem we compare stochastically the marginal distribution of the age of the inter-arrival time and its two conditional distributions, first given that the server is busy and second, when it is not.

Theorem 2.2 *In a G/M/1 queue, the random variable of the age of the inter-arrival time given that the server is busy, is stochastically smaller than the unconditional one, which in turn is stochastically smaller than this age conditioning on an idle server.*

Proof. For any function ψ we get by (2) and (1) that

$$\mathrm{E}(\psi(A)|L \geq 1) = \int_{a=0}^{\infty} \psi(a) f_{A|L\geq 1}(a)da = \int_{a=0}^{\infty} \psi(a)\overline{G}(a)\mu e^{-\eta a}\, da$$

$$= \mu\overline{x} \int_{a=0}^{\infty} \psi(a)\frac{\overline{G}(a)}{\overline{x}}e^{-\eta a}\, da = \frac{\int_{a=0}^{\infty} \psi(a) f_A(a)e^{-\eta a}\, da}{\mathrm{E}(e^{-\eta A})} = \frac{\mathrm{E}(\psi(A)e^{-\eta A})}{\mathrm{E}(e^{-\eta A})}. \tag{7}$$

Since $e^{-\eta x}$ is a decreasing function, then for any function $\psi(x)$ which is non-decreasing, $\psi(A)$ and $e^{-\eta A}$ are negatively correlated, namely,

$$\mathrm{E}(\psi(A)e^{-\eta A}) \leq \mathrm{E}(\psi(A))\mathrm{E}(e^{-\eta A}).$$

Thus, the expression in (7) is smaller than $\mathrm{E}(\psi(A))$ for any non-decreasing function $\psi(x)$. In summary, for any non-decreasing function $\psi(x)$, $\mathrm{E}(\psi(A)|L \geq 1) \leq \mathrm{E}(\psi(A))$. This is equivalent to saying that $A|L \geq 1$ is stochastically smaller than A. Finally, the fact that $A|L = 0$ is stochastically larger than A is now immediate. □

The following theorem is well-known. It appears in virtually any queueing text and it comes with a commonly repeated proof. An original proof is given in [9]. It fact this proof also holds for the case of the G/G/1 queue under the LCFS-PR regime. We give a new proof which is based on the fact, established independently in Remark 2.3, that in a G/M/1 queue, given a busy server, the age of the inter-arrival time and the number of customers in the queue are independent.

Theorem 2.3 *In a G/M/1 queue, given that the server is busy, the number of customers in the queue follows a geometric distribution.*

Proof. Consider the Markov process whose typical state is (n, a), where n refers to the number of customers in the system and a to the age of the inter-arrival time, $n, a \geq 0$. The set of states sharing the same value for n, will be referred to as the *macro-state* n, $n \geq 0$. Note that the process among macro-states is *not* a Markov process. Let $\pi(n, a)$ be the limit probability-density of state (n, a) and denote $\int_{a=0}^{\infty} \pi(n, a)\, da$ by $e(n)$, $n \geq 0$. Clearly, $e(n)$ is the limit probability of the macro-state n, $n \geq 0$. Also, $\pi(n, a)/e(n)$, $a \geq 0$, is the conditional density function of the age given that n customers are in the system, $n \geq 0$. Denote by $h(x)$ the hazard function of the arrival process, namely $h(x) = g(x)/\overline{G}(x)$, $x \geq 0$. Then,

$$\lambda_n \equiv \int_{a=0}^{\infty} \frac{\pi(n, a)}{e(n)} h(a)\, da, \quad n \geq 0, \tag{8}$$

is the transition rate from macro-state n into macro-state $n + 1$, $n \geq 0$. Due to service times being exponentially distributed, the corresponding transition rate from n into $n - 1$, $n \geq 1$, equals μ. Although the process among macro-states is not Markovian, its limit probabilities, namely $e(n)$, $n \geq 0$, coincide with those of the Markov process having these transition rates. This fact is argued for example in [6], p.814 and in [5], p. 1184, for a discrete state space. This auxiliary Markov process is in fact a birth and death process with (8) being the rates of birth, and μ being the common rate of death.

Now inspect (8). Remark 2.4 says that $\pi(n, a)/e(n)$ is homogeneous with n, as long as $n \geq 1$. Thus, the same is the case regarding the transition rates (or birth rates in the auxiliary process) given in (8). The death rates are of course homogeneous as they equal μ for any $n \geq 1$. Thus, the auxiliary birth and death process has homogeneous transition rates as long as $n \geq 1$. Hence,

$e(n+1) = Ce(n)$, or, equivalenlty, $e(n) = C^{n-1}e(1)$, for some constant C, $n \geq 1$, as required. $\square$

Remark 2.5 Note that C as defined in the above proof equals σ as this is the common multiplier in the geometric probabilities. Hence, λ_n, as defined in (8), equals $\mu\sigma$, when $n \geq 1$. The reason behind this is that $\lambda_n \pi_n = \mu\pi_{n+1}$ and $\pi_{n+1}/\pi_n = \sigma$, $n \geq 1$. Finally, in the M/M/1 case $\lambda_n = \lambda$, $n \geq 0$, where λ is the (unconditional) arrival rate.

Remark 2.6 In Theorem 2.3, the age of the inter-arrival time can be replaced by the residual inter-arrival time. This is the case since the process whose typical state is (n, r) where n is as above and where r is the residual inter-arrival time, is a Markov process with the same limit probabilities defined in the proof of Theorem 2.3.

2.2 The conditional residual and total inter-arrival time in the G/M/1

The following two corollaries are immediate from Theorem 2.1.

Corollary 2.1 *Let $f_{R|L=n}(r)$ be the density function of the age of the inter-arrival time, given that the number in the system equals n, where $r, n \geq 0$. Then, in the G/M/1 queue,*

$$f_{R|L=n}(r) = \mu \int_{a=0}^{\infty} e^{-\eta a} g(a+r)\, da, \quad r \geq 0,\ n \geq 1. \tag{9}$$

In particular, given that the server is busy, the residual of the inter-arrival time and the number in the queue are independent. Also,

$$f_{R|L=0}(r) = \frac{1}{\overline{x}(1-\rho)} \int_{a=0}^{\infty} (1 - e^{-\eta a}) g(a+r)\, da, \quad r \geq 0.$$

Remark 2.7 Observing the density function given in (9) we see that this is the same density as that of the random variable $Y - S | Y \geq S$ where Y has the same distribution as that of an inter-arrival time, where S follows an exponential distribution with parameter η (as the sojourn time in G/M/1 in case of FCFS), and where Y and S are independent. By a similar argument to the one made in Remark 2.1, one can conclude that this is also the density function of an idle period in the $G/M/1$ queue. This observation was first pointed out in [1].

Example: M/M/1. For the sake of completeness we state the trivial fact that the residual inter-arrival time in case of an M/M/1 queue follows an exponential distribution with parameter λ and this is the case conditional on any number of customers at the system.

Corollary 2.2 *Let $f_{X|L=n}(x)$ be the density function of the current inter-arrival time (i.e., its age plus its residual), given that the number of customers in the system equals n, where $x, n \geq 0$. Note that $X = A + R$. Then, in the G/M/1 queue,*

$$f_{X|L=n}(x) = \mu \int_{a=0}^{x} e^{-\eta a} g(x-a)\, da, \quad x \geq 0,\ n \geq 1. \tag{10}$$

In particular, given that the server is busy, the length of the current inter-arrival time and the number of customers in the queue are independent. Also,

$$f_{X|L=0}(x) = \frac{1}{\overline{x}(1-\rho)} \int_{a=0}^{x} (1 - e^{-\eta a}) g(x-a)\, da, \quad x \geq 0.$$

Example: M/M/1 (continued). From the fact that the age and the residual time are independent in the case of exponentially distributed inter-arrival times, it is possible to conclude that $X|L = n$ for $n \geq 1$, is the sum of two independent and exponentially distributed random variables, one, $A|L \geq 1$, with parameter μ (see (2)) and the other, R, with parameter λ. Finally, $X|L = 0$ is the sum of three such random variables, two with parameter λ and one with parameter μ.

In Theorem 2.2 we showed the stochastic order between the conditional age and unconditional one. Investigating the same order regarding the residual yields, unlike the case of the age in which the order is distribution free, that an order does not always exist. Moreover, when it exists, its direction (as we show shortly) depends on the inter-arrival distribution. We were, however, able to give a definite answer in the case of increasing failure rate (IFR) service distributions. This is done in the theorem below. In the case of service distributions having a decreasing failure rate (DFR), all orders are reversed.

Theorem 2.4 *In a G/M/1 queue, if the inter-arrival time distribution is with IFR, then the residual of the inter-arrival time, given that the server is busy, is stochastically larger than the unconditional one, which in turn is stochastically larger than this age, conditioning on an idle server.*

Proof. Let ϕ be a non-decreasing function. Denote $\psi(A) = \mathrm{E}(\phi(R|A))$. Since given A, the random variables R and L are independent, we have

$$\mathrm{E}(\phi(R)|L \geq 1) = \mathrm{E}[\mathrm{E}(\phi(R)|A)|L \geq 1] = \mathrm{E}[\psi(A)|L \geq 1]. \tag{11}$$

By definition, if the inter-arrival time distribution is IFR, $\psi(A)$ is non-increasing with A. Thus, in the IFR case by Theorem 2.2, $\mathrm{E}[\psi(A)|L \geq 1] \geq \mathrm{E}[\psi(A)]$ and $R|L \geq 1$ is stochastically larger than R. The order between $R|L = 0$ and R is now straightforward. $\square$.

Remark 2.8 Theorem 2.4, coupled with the observation in Remark 2.7 give conditions under which the idle period in $G/M/1$ and the residual inter-arrival time are stochastically ordered.

Remark 2.9 For the G/M/c model, where c identical servers all serve at the common rate μ, all of the above properties hold as long as one conditions on all servers being busy. For example, given all servers are busy, the age of the inter-arrival time and the number in the queue are independent. Also, Equations (2),(9) and (10) hold for $n \geq c$ where μ and η are replaced by $c\mu$ and $c\mu(1-\sigma)$, respectively, where now σ obeys $\sigma = G^*(c\mu(1-\sigma))$.

3 The age of inter-arrival time in the $M/G/1$

In this section we deal with the same question dealt with in the previous section, but for the M/G/1 model. Here on top of requiring the service regime to be work-conserving and non-anticipating, we need to assume that no preemptions take place. Among such regimes, one can find the FCFS, LCFS and random order. It is clear that when conditioning on the number of customers in the system, the age of the inter-arrival time follows the same distribution under all these regimes. For simplicity, during the rest of this section we assume FCFS. Below we use the standard notation for the M/G/1 queue. In particular, λ denotes the arrival rate and $G^*(s)$ the LST of service

times. Also, π_i is the steady-state probability of having i customers in the system, $i \geq 0$.

Next we derive the conditional age of the inter-arrival time given the number in the system. Towards this goal, we first need to look into the conditional residual service time. Specifically, denote by R such a random variable and by R_n the same residual but given that there are n customers in the system, $n \geq 1$. Let $r_n(x)$ and $R_n^*(s)$ be the corresponding density function and LST, respectively, $n \geq 1$. Recursive expressions for $r_n(x)$ and for $R_n^*(s)$, $n \geq 1$, can be found in [8]. Here we quote only the recursion for $R_n^*(s)$:

$$R_1^*(s) = \frac{\lambda}{s-\lambda} \frac{G^*(\lambda) - G^*(s)}{1 - G^*(\lambda)}$$

and

$$R_{n+1}^*(s) = \frac{\lambda}{s-\lambda} \left(G^*(\lambda) \frac{1 - R_n^*(s)}{1 - R_n^*(\lambda)} - G^*(s) \right), \quad n \geq 1$$

Recall by the PASTA property that the distributions of R at random times and at epochs of arrival, coincide. The same is the case regarding R_n, $n \geq 1$.

Theorem 3.1 *Let $A_n^*(s)$ be the LST of the age of the inter-arrival time given that there are n customers in the system, $n \geq 0$. Then, for the M/G/1 queue*

$$A_0^*(s) = \frac{\lambda G^*(\lambda + s)}{s + \lambda G^*(\lambda + s)} \tag{12}$$

and for $n \geq 1$,

$$A_n^*(s) = \frac{\lambda}{(\lambda + s)\pi_n} \times \tag{13}$$

$$\left(\pi_{n-1}(1 - R_{n-1}^*(\lambda + s)) + \sum_{m=n}^{\infty} \pi_m (1 - G^*(\lambda + s))(G^*(\lambda + s))^{m-n} R_m^*(\lambda + s) \right)$$

with $R_0^*(s) = G^*(s)$.

Proof. We deal first with the case where $n = 0$. As before, we write

$$f_{A|L=0}(a) = \frac{\mathrm{P}(L = 0|A = a) f_A(a)}{\mathrm{P}(L = 0)} = \frac{\mathrm{P}(L = 0|A = a) \lambda e^{-\lambda a}}{1 - \rho}. \tag{14}$$

Notice that $\mathrm{P}(L = 0|A = a)$ is the probability, according to the FCFS discipline, that the sojourn time of the last customer to arrive is smaller

than or equal to a. Indeed, had someone arrived a units of time ago, he/she would have cleared the system by now. Denote the LST of this sojourn time by $W^*(s)$ and recall that

$$W^*(s) = (1-\rho)\frac{sG^*(s)}{\lambda G^*(s) + s - \lambda}$$

(see, e.g., [4], p.433). Also, recall that if $F^*(s)$ is the LST of a nonnegative random variable and if $F(x)$ is its cumulative distribution function, then

$$\int_{x=0}^{\infty} F(x)e^{-sx}\,dx = \frac{F^*(s) - F(0)}{s}.$$

For simplicity, assume that $G(0) = 0$. Equivalent results can be proved in the same way for the case where $G(0) > 0$. Using all of that, multiply both hand sides of (14) by e^{-sa} and integrate with respect to a, $0 \le a < \infty$. By some algebra one then gets (12).

For $n \ge 1$ we have,

$$f_{A|L=n}(a) = \frac{\mathrm{P}(L=n|A=a)f_A(a)}{\pi_n} = \frac{\sum_{m=n-1}^{\infty} \mathrm{P}(L=n|A=a, L_a=m)\lambda e^{-\lambda a}\pi_m}{\pi_n} \tag{15}$$

where L_a stands for the number in the system at the point of the previous arrival. We now develop individually each addend in this sum. For $n \ge 1$ and $m = n-1$, the addend in the sum given in (15) equals $\mathrm{P}(R_{n-1} > a)e^{-\lambda a}\pi_{n-1}$. For $m \ge n$ the addend equals

$$\pi_m \lambda e^{-\lambda a} \int_{r=0}^{a} r_m(r) \int_{x=0}^{a-r} g^{(m-n)}(x)\bar{G}(a-r-x)dxdr \tag{16}$$

where $g^{(\ell)}(x)$ is the density function of the sum of ℓ independent random variables all having the same density function $g(x)$, $\ell \ge 1$. Multiplying both hand sides of (15) by e^{-sa}, integrating with respect to a, $0 \le a < \infty$, and some straightforward algebra, complete the proof. $\Box$

Remark 3.1 Next we give a probabilistic interpretation to (13). First, recall that the LST of a random variable can be interpreted as follows. Let $F^*(\cdot)$ be the LST of a nonnegative random variable Y. Also, let S be a random

variable which follows an exponential distribution with parameter s. Assume also that Y and S are independent. Then, $F^*(s) = \mathrm{P}(Y \leq S)$. Let X be a random variable which is distributed as an inter-arrival time. Also, let $N_m(\cdot)$ be the delayed renewal process with the first renewal time distributed as R_m and the rest as the service times. Using the fact that $\min\{X, S\}$ and $I_{\{X \leq S\}}$ are independent, we have, by (13), that

$$A_n^*(s)\pi_n = \sum_{m=n-1}^{\infty} \mathrm{P}\left(N_m(X) = m - n + 1, L_a = m, X \leq S\right). \qquad (17)$$

Example: M/M/1. Much is being simplified in the M/M/1 case due to the fact that the distribution of $R|L = i$ is the same for all $i \geq 1$. In particular, it follows an exponential distribution with parameter μ. Thus, inserting $R_i^*(s) = \mu/(\mu + s)$, $i \geq 1$, and $\pi_i = (1 - \rho)\rho^i$, $i \geq 0$, in (12) and (13) leads after some trivial algebra to $A_0^*(s) = \mu\lambda/[(\mu + s)(\lambda + s)]$ and $A_n^*(s) = \mu/(\mu + s)$, $n \geq 1$. These results were already derived in the previous section where the M/M/1 model was treated as a special case of the G/M/1 model.

Remark 3.2 In the previous section we showed that having exponential service times is a sufficient condition for the independence between the age of inter-arrival time and the number of customers in the system, given that the latter is positive. Here we see that in the $M/G/1$ case this independence property does not hold. This leads to the conclusion that having exponential service times is also a necessary condition for this independence property to hold.

Acknowledgement

We like to acknowledge many insightful comments made by Onno Boxma. In particular, he referred us to [9]. We also like to thank Ivo Adan and David Perry for stimulating discussions.

References

[1] Adan, I., Boxma, O. and D. Perry (2005), "The $G/M/1$ queue revisited," *Mathematical Methods of Operations Research* Vol. 62, pp. 437-452.

[2] Adan, I. and M. Haviv (2008), "Conditional ages and residual service times in the M/G/1 queue," *Stochastic Models*, (to appear).

[3] Altman, E., Jimenez, T., Nunez-Queija, R., and U. Yechiali (2004), "Optimal routing among $\cdot$/M/1 queues with partial information," *Stochastic Models*, Vol. 20, pp. 149-172.

[4] Cohen, J.W. (1982), *The Single Server Queue, 2nd Edition*, North-Holland, Amsterdam.

[5] Haviv, M. and G.-J. van Houtum (1998), "The critical offered load in variants of the symmetric shortest and longest queue systems," *Stochastic Models*, Vol. 14, pp. 1179–1195.

[6] Haviv, M. and L. van der Heyden (1984), "Perturbation bounds for the stationary probabilities of a finite Markov chain," *Advances in Applied Probability*, Vol. 16, pp. 804–818.

[7] Haviv, M. and R. Zlotnikov (2008), "Computational schemes for two exponential servers where the first has a finite buffer," (Submitted for publication).

[8] Kerner, Y. (2008), "The conditional distribution of the residual service time in the $M_n/G/1$ queue," *Stochastic Models* Vol. 24, pp. 364-375.

[9] Nunez-Queija, R. (2001), "Note on the GI/GI/1 queue with LCFS-PR observed at arbitrary times," *Probability in the Engineering and Informational Sciences*, Vol. 15, pp. 179–187.