UNPUBLISHED (HAS SOME ERRORS) The Variance Curve of M/G/1 Outputs – a Renewal Reward Perspective

Yoav Kerner* and Yoni Nazarathy†

July 2, 2009

WARNING: THIS PAPER HAS A MAJOR FLAW (in the proof of the Lemma 1) AND SOME OF THE RESULTS ARE WRONG. NEVERTHELESS SOME OF THE RESULTS ARE CORRECT AND ARE WORTH LOOKING AT.

Abstract

We present a method for the analysis of queueing output processes that is based on an embedding of the output process in a renewal-reward process in which the renewal periods are busy cycles and the rewards represent the numbers of customers served during busy cycles. Our method yields an approximation for the variance curve that is asymptotically exact as time grows to infinity.

For illustration, we consider the stable $M/G/1$ queue with an arbitrary, finite variance, distribution for the initial state. We show that under the assumption of a finite third moment for the service time distribution, this curve converges to a linear asymptote with a slope equal to the arrival rate and further show that the y-intercept of this asymptote depends only on the variance of the initial queue level, σ_0^2, and the traffic intensity, ρ, and is equal to $\sigma_0^2 - \rho/(1-\rho)^2$.

Our result immediately yields an expression for the y-intercept of the stationary case that depends on the first 3 moments of the service time distribution. This stimulates consideration of a 35 year old conjecture of D. J. Daley: Among the stationary $M/G/1$ queues, the $M/M/1$ is characterized by having a variance curve of λt, where λ is the arrival rate. While we leave this conjecture unproved, we narrow down the class of candidate service time distributions to those whose y-intercept of the linear asymptote is 0.

Keywords: Renewal Reward, M/G/1 Queue, Output Processes, Variance Curve.

1 Introduction

The theoretical study of queueing output processes has received considerable attention during the 50's, 60's and 70's. Burke's theorem, [3], stating that the output of a stationary M/M/1 queue is a Poisson process is probably the first major result of this era. It was followed by several dozens of studies which analyzed properties of inter-departure times, output counting processes and characterizations of queueing systems based on their input and output relations. A classic survey is in [5], other useful resources are [7] and [8].

*EURANDOM, Eindhoven University of Technology, Eindhoven The Netherlands.

†EURANDOM and the Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven and CWI, Amsterdam, The Netherlands.

One measure of performance that has received considerable attention is the variance of the number of outputs over time, see [18] for a survey. Evaluation of this variance plays a central role in the analysis of manufacturing and supply chain settings. Some recent studies have investigated computational procedures that aim to approximate this quantity for complex queueing systems (cf. [11, 12, 13, 14, 19, 20]). In this paper we propose an alternative approach for this analysis that is based on an embedding of the output process in a renewal reward process. Our approach may be applied to any regenerative queueing system and yields an approximation that is asymptotically exact as time grows to infinity. For illustration, we concentrate on the M/G/1 queue and complement a classic result of Daley [4]. As a consequence we shed some light on a 35 year old unproved conjecture.

We consider the M/G/1 queue operating under an arbitrary work-conserving non-preemptive service discipline with arrival rate λ and service time distribution $G(\cdot)$. Denote the Laplace-Stieltjes transform (LST) associated with $G(\cdot)$ by $G^*(\cdot)$ and the k'th moment of $G(\cdot)$ by g_k. We assume that $g_3 < \infty$ and $\rho = \lambda g_1 < 1$. Further assume that the number of customers in the system at time 0 and the residual service time of the customer in service, follow an arbitrary distribution with a finite variance, σ_0^2, for the number of customers.

Our main interest is the output counting process, $\{D(t), t \geq 0\}$ and in particular its *variance curve*: $\mathrm{Var}(D(t))$. It is well known that for the special case of the stationary M/M/1 queue, $D(t)$ is a Poisson process with rate λ and thus $\mathrm{Var}(D(t)) = \lambda t$, but in general this need not be the case. It can easily be shown that,

$$\mathrm{Var}(D(t)) = \lambda t + o(t), \tag{1}$$

where $o(t)$ is a function that increases at a slower than linear rate. Thus the *asymptotic variance rate* of outputs is λ and an immediate approximation for the variance curve is

$$\mathrm{Var}(D(t)) \approx \lambda t. \tag{2}$$

This simple approximation is attractive since it does not depend on the distribution of the initial state and the relative error, $|\mathrm{Var}(D(t)) - \lambda t|/t$, diminishes to 0 as $t \to \infty$. Nevertheless, the absolute error, $|\mathrm{Var}(D(t)) - \lambda t|$, does not necessarily vanish and in general depends on the initial state. The requirement that $g_3 < \infty$ implies that the asymptotic queue length has finite variance and this implies (1), the details are in Section 3. Some examples with $g_3 = \infty$ in which (1) does not hold are given in [6].

Our contribution is a refinement to the approximation (2) for which the absolute error vanishes as $t \to \infty$. We show that there exists a constant $\overline{B}$ such that the $o(t)$ term in (1) equals $\overline{B} + o(1)$ where $o(1)$ is a function that vanishes as $t \to \infty$. We refer to $\overline{B}$ as the *y-intercept* and denote the *linear asymptote* by, $\overline{v}(t) = \lambda t + \overline{B}$. We thus have:

$$\mathrm{Var}(D(t)) = \overline{v}(t) + o(1). \tag{3}$$

Our key result is the following simple expression for the y-intercept:

$$\overline{B} = \sigma_0^2 - \frac{\rho}{(1-\rho)^2}. \tag{4}$$

Considering the stationary system as a special case, set $\sigma_0^2 = \sigma_\pi^2$, where σ_π^2 denotes the steady state variance of the number of customers in the system. This yields an expression for $\overline{B}$ that depends on ρ and the first 3 moments of $G(\cdot)$. Note that for the stationary M/M/1 queue, $\sigma_\pi^2 = \rho/(1-\rho)^2$ and thus starting with stationary initial queue level yields $\overline{B} = 0$ as expected. It is further evident that all stationary systems in which $\sigma_\pi^2 = \rho/(1-\rho)^2$ yield $\overline{v}(t) = \lambda t$. One such class of systems are those having a service time distribution whose first moments agree with the exponential distribution.

Daley [4], conjectured the following: *The M/M/1 queue is the only stationary M/G/1 queue having* $Var(D(t)) = \lambda t$. While it is known ([10], see also [9]), that an output Poisson process characterizes the

M/M/1 within the class of stationary M/G/1 queues, it is not known if other stationary M/G/1 queues (other than M/M/1) yield a "Poisson like" variance curve. Our result implies that if there are any other such queues (service distributions), then they must have $\sigma_\pi^2 = \rho/(1-\rho)^2$ since they must have $\overline{B} = 0$.

Daley posed the above conjecture after finding the following form for the LST of the variance curve for the stationary system operating under a non-preemptive service policy:

$$V^*(s) = \int_0^\infty e^{-st} d\mathrm{Var}(D(t)) = \frac{\lambda}{s} + \frac{2\lambda}{s}\left(\frac{G^*(s)}{1-G^*(s)}\left(1 - \frac{s\Pi(\Gamma(s,1))}{s+\lambda(1-\Gamma(s,1))}\right) - \frac{\lambda}{s}\right). \tag{5}$$

Here $\Pi(\cdot)$ is the probability generating function of the stationary number of customers in the system and $\Gamma(s,z) = \mathbb{E}\left[e^{-sX'}z^Y\right]$, where the random variables X' and Y are respectively distributed as the duration of an M/G/1 busy period and the number of customers served during a busy period. $\Gamma(s,z)$ may be found as the root of smallest modulus of the well known busy period functional equation (cf. [17], pp. 50):

$$\Gamma(s,z) = zG^*(s+\lambda(1-\Gamma(s,z))). \tag{6}$$

One method for calculating the y-intercept of the stationary case is to apply the so-called final value theorem to the right hand side term of (5). One may then apply a simple coupling argument, stated in Section 4 to obtain the y-intercept for an arbitrary initial distribution. In this paper we choose a different path in order to illustrate a broader method: *Embedding the output process in a renewal reward process*: To this end, we analyze a renewal-reward process in which the renewal periods are busy cycles and the rewards represent the numbers of customers served during busy cycles. In [2], the authors specify the linear asymptote of a renewal reward process based on some moments and joint moments of the renewal intervals and the rewards. We utilize their result to obtain the linear asymptote of $\mathrm{Var}(D(t))$. Note that in principle, our method can be used to find the linear asymptote of the variance curve of output processes that result from any regenerative queueing process.

The remainder of the paper is organized as follows: In Section 2 we present our method of embedding $D(t)$ in a renewal reward process. The results are stated in a general context with out specific reference to the M/G/1 queue. In Section 3 we show the existence of a linear asymptote for the M/G/1 queue. In Section 4 we derive the y-intercept of the M/G/1 queue. Our proof utilizes our renewal reward approach along with a simple coupling argument that gives some insight to the result and simplifies the computations. We also present a corollary: an expression for the limiting value of the covariance between the number of arrivals and the queue size. In Section 5 we consider the special case of a stationary $M/G/1$ queue and discuss Daley's conjecture.

2 Embedding Departure Counts in a Renewal Reward Process

We now present a framework for obtaining the linear asymptote. The results of this section are stated in a general setting and do not make specific reference to the M/G/1 queue.

Consider the renewal reward process $C(t)$ in which the renewal periods are busy cycles and the rewards are the numbers of customers served during busy cycles. This process may also be viewed as an output process of a modified system in which customers do not leave immediately after service but rather all customers served during a busy period leave in a single batch once the busy period is complete. Our assumption is that (as in the M/G/1 queue) $Q(t) = 0$ implies that $C(t) = D(t)$ and as a result the processes $D(t)$ and $C(t)$ are equal at regeneration epochs and their difference during busy periods is never more than the number of customers served in a busy period. This relationship is the key to obtain the linear asymptote $\overline{v}(t)$ from the linear asymptote of $\mathrm{Var}(C(t))$. The latter is obtained from the following:

Theorem 1 *(Brown and Solomon 1975, [2]): Consider a sequence of random vectors* $\{(X_i, Y_i), i = 0, 1, \ldots\}$ *where* (X_i, Y_i), $i \geq 1$ *are identically distributed. Further assume that* X_i *are spread-out (cf. [1], pp. 186). Denote the moments* $x_j = \mathbb{E}\,[X_1^j]$, $y_j = \mathbb{E}\,[Y_1^j]$, $n_{jk} = \mathbb{E}\,[X_1^j Y_1^k]$, $\tilde{x}_j = \mathbb{E}\,[X_0^j]$, $\tilde{y}_j = \mathbb{E}\,[Y_0^j]$ *and* $\tilde{n}_{jk} = \mathbb{E}\,[X_0^j Y_0^k]$. *Further assume that* $x_3, y_2, \tilde{x}_2, \tilde{y}_2 < \infty$. *Define the renewal process,* $N(t)$ *and the renewal reward process,* $C(t)$ *as follows:*

$$N(t) = \inf\{n : \sum_{k=1}^{n} X_k > t\}, \qquad C(t) = \sum_{i=0}^{N(t)-1} Y_i.$$

Then $Var(C(t)) = \overline{V}_C t + \overline{B}_C + o(1)$, *where,*

$$\begin{aligned}
\overline{V}_C &= \frac{x_2 y_1^2}{x_1^3} - 2\frac{n_{11} y_1}{x_1^2} + \frac{y_2}{x_1}, \\
\overline{B}_C^0 &= \frac{5}{4}\frac{x_2^2 y_1^2}{x_1^4} - \frac{2}{3}\frac{x_3 y_1^2}{x_1^3} + 2\frac{n_{21} y_1}{x_1^2} - 3\frac{x_2 y_1 n_{11}}{x_1^3} + \frac{n_{11}^2}{x_1^2} + \frac{1}{2}\frac{x_2 y_2}{x_1^2} - \frac{n_{12}}{x_1}, \\
\overline{B}_C &= \overline{B}_C^0 - \frac{\tilde{x}_1 x_2 y_1^2}{x_1^3} + \frac{y_1^2(\tilde{x}_2 - \tilde{x}_1^2) + 2\tilde{x}_1 n_{11} y_1}{x_1^2} - \frac{y_2 \tilde{x}_1 + 2y_1(\tilde{n}_{11} + \tilde{x}_1 \tilde{y}_1)}{x_1} + \tilde{y}_2 - \tilde{y}_1^2.
\end{aligned}$$

Note that if (X_0, Y_0) is distributed as (X_1, Y_1) then $\overline{B}_C = \overline{B}_C^0$. Using the next lemma we relate the output process $D(t)$ to the renewal reward process $C(t)$.

Lemma 1 *Consider a regenerative process* $\{\mathcal{Q}(t), t \geq 0\}$ *with regeneration epochs* $\{\tau_1, \tau_2, \ldots\}$ *and a process* $\{D(t)\}$ *defined on the same probability space such that the sequence* $\{D(\tau_{n+1}) - D(\tau_n), n \geq 1\}$ *is i.i.d. Denote* $\tau(t) = \sup\{\tau_1, \tau_2, \ldots | \tau_n \leq t\}$ *and assume that* $C(t) = D(\tau(t))$ *w.p. 1. Denote the limiting life random variable by* $\tau = \lim_{t\to\infty}(t - \tau(t))$ *and further denote* $x_i = \mathbb{E}\,[(\tau_2 - \tau_1)^i]$ *and* $y_1 = \mathbb{E}\,[D(\tau_2) - D(\tau_1)]$ *and assume that* $x_2 < \infty$ *and* $y_1 < \infty$.

If there exist constants $\overline{V}_C, \overline{B}_C, \overline{V}_D, \overline{B}_D$ *such that*

$$Var(D(t)) = \overline{V}_D t + \overline{B}_D + o(1) \quad and \quad Var(C(t)) = \overline{V}_C t + \overline{B}_C + o(1), \tag{7}$$

then,

(i) $\overline{V}_D = \overline{V}_C$.

(ii) $\overline{B}_D = \overline{B}_C + \overline{V}_D \mathbb{E}\,[\tau] - \frac{y_1^2}{x_1^2} Var(\tau)$.

Proof. We compute the variance of $D(\tau(t))$ conditioned on $\tau(t)$:

$$\begin{aligned}
\mathrm{Var}(D(\tau(t)) &= \mathbb{E}\,[\mathrm{Var}(D(\tau(t)))|\tau(t))] + \mathrm{Var}(\mathbb{E}\,[D(\tau(t))|\tau(t)]) \\
&= \mathbb{E}\,[\overline{V}_D \tau(t) + \overline{B}_D + o(1)] + \mathrm{Var}(\frac{y_1}{x_1}\tau(t) + \overline{b}' + o(1)),
\end{aligned}$$

where we used $\mathbb{E}\,[D(\tau(t))] = \mathbb{E}\,[C(\tau(t))] = \frac{y_1}{x_1} t + \overline{b}' + o(t)$ for some finite $\overline{b}'$ (see lemma 1 of [2]). Now using (7) and the fact that $\mathrm{Var}(C(t)) = \mathrm{Var}(D(\tau(t)))$ we have:

$$\overline{V}_C t + \overline{B}_C = \overline{V}_D \mathbb{E}\,[\tau(t)] + \overline{B}_D + \frac{y_1^2}{x_1^2}\mathrm{Var}(\tau(t)) + o(t). \tag{8}$$

Observe that $\lim_{t\to\infty} \frac{\mathbb{E}\,[\tau(t)]}{t} = 1$ and $\lim_{t\to\infty} \frac{\mathrm{Var}(\tau(t))}{t} = 0$, thus dividing (8) by t and taking $t \to \infty$ of (8), we obtain (i).

To obtain (ii), apply (i) into (8):

$$\overline{B}_D = \overline{B}_C + \overline{V}_D \mathbb{E}\left[t - \tau(t)\right] - \frac{y_1^2}{x_1^2}\mathrm{Var}(t - \tau(t)) + o(1).$$

Taking $t \to \infty$ we obtain (ii). ■

Using Theorem 1 and Lemma 1 we are able to get an expression for the linear asymptote of $\overline{v}(t)$ based on moments and joint moments of the busy cycle and the number of customers served. We summarize this in the following theorem (stated within a general setting of regenerative processes):

Theorem 2 *Consider a regenerative process $\{\mathcal{Q}(t), t \geq 0\}$ with regeneration epochs $\{\tau_1, \tau_2, \ldots\}$ and a process $\{D(t)\}$ defined on the same probability space such that the sequence $\{D(\tau_{n+1}) - D(\tau_n), n \geq 1\}$ is i.i.d. Further denote,*

$$\begin{array}{lll} x_i = \mathbb{E}\left[(\tau_2 - \tau_1)^i\right], & y_i = \mathbb{E}\left[\big(D(\tau_2) - D(\tau_1)\big)^i\right], & n_{ij} = \mathbb{E}\left[(\tau_2 - \tau_1)^i \big(D(\tau_2) - D(\tau_1)\big)^j\right], \\ \tilde{x}_i = \mathbb{E}\left[\tau_1^i\right], & \tilde{y}_i = \mathbb{E}\left[D(\tau_1)^i\right], & \tilde{n}_{ij} = \mathbb{E}\left[\tau_1^i D(\tau_1)^j\right]. \end{array}$$

Assume $x_3, y_2, \tilde{x}_2, \tilde{y}_2 < \infty$ and assume that the distribution of $\tau_2 - \tau_1$ is spread-out. If there exist constants $\overline{V}_D, \overline{B}_D$ such that $Var(D(t)) = \overline{V}_D t + \overline{B}_D + o(1)$ then,

$$\begin{aligned} \overline{V}_D &= \frac{x_2 y_1^2}{x_1^3} - 2\frac{n_{11} y_1}{x_1^2} + \frac{y_2}{x_1}, \\ \overline{B}_D &= \frac{1}{x_1^4}\Big(2x_2^2 y_1^2 - x_1 y_1 (4x_2 n_{11} + x_3 y_1 + x_2 \tilde{x}_1 y_1) - x_1^3 (n_{12} + 2\tilde{n}_{11} y_1 + \tilde{x}_1 y_2 - 2\tilde{x}_1 y_1 \tilde{y}_1) \\ &\quad + x_1^4 (\tilde{y}_2 - \tilde{y}_1^2) + x_1^2 (n_{11}^2 + 2\tilde{x}_1 n_{11} y_1 + 2n_{21} y_1 - \tilde{x}_1^2 y_1^2 + \tilde{x}_2 + y_1^2 + x_2 y_2)\Big). \end{aligned} \tag{9}$$

Proof. $\overline{V}_D$ equals $\overline{V}_C$ of Theorem 1 due to (i) of Lemma 1. To obtain $\overline{B}_D$ use the formula for $\overline{B}_C$ in Theorem 1 combined with (ii) of Lemma 1. The moments of the limiting life random variable τ are known to be $\mathbb{E}[\tau] = x_2/2x_1$ and $\mathbb{E}[\tau^2] = x_3/3x_1$. Simplifying, we obtain the result. ■

Theorem 2 reduces the problem of calculating $\overline{V}_D$ and $\overline{B}_D$ to that of calculating moment values. The asymptotic variance rate requires 5 quantities that are independent of the initial state: $(x_1, x_2, y_1, y_2, n_{11})$. Finding the y-intercept requires an additional set of quantities, first (x_3, n_{12}, n_{21}), and further $(\tilde{x}_1, \tilde{x}_2, \tilde{y}_1, \tilde{y}_2, \tilde{n}_{11})$ which depend on the distribution of the initial state.

We note that (i) of the above theorem validates the asymptotic variance rate in (1). To do so, use (6) to evaluate $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathbb{E}[Y]$, $\mathbb{E}[Y^2]$ and $\mathbb{E}[XY]$ where X is distributed as the M/G/1 busy cycle (busy + idle period) and Y is distributed as the number of customers served in a busy period. Inserting these moments in (i) yields $\overline{V}_D = \lambda$. In fact, (i) also provides an elementary way to obtain the asymptotic variance rate of outputs of the M/M/1/K queue which was obtained in [15] using more involved methods. Another application of (i) to a simple queueing network is in Section 6.5 of [16].

We also comment that in regenerative queueing systems where some of the moments may not be calculated, they may be estimated by the use of regenerative simulation and the result can be substituted in (i) to obtain estimates of $\overline{V}_D$. This method of estimating the asymptotic variance rate is superior to the naive method of estimating $\overline{V}_D$: simulate an i.i.d. sequence $\{D_1(T), \ldots, D_n(T)\}$ for a large time T and take the sample variance of the sequence divided by T. Observe that the naive method contains an inherit bias equal to,

$$\frac{\overline{B}_D}{T} + o(1).$$

3 Asymptotic Form of the M/G/1 Variance Curve

In this section we show that the variance curve has a linear asymptote with slope λ and a finite y-intercept. Denote the arrival Poisson process by $\{A(t), t \geq 0\}$ and denote the number of customers in the system at time t by $Q(t)$. First note that showing $\mathrm{Var}(D(t)) = \lambda t + o(t)$ is almost immediate. To do so, observe that the variance of

$$D(t) = A(t) + Q(0) - Q(t), \tag{10}$$

is,

$$\mathrm{Var}(D(t)) = \lambda t + \sigma_0^2 + \mathrm{Var}(Q(t)) - 2\mathrm{Cov}(Q(0), Q(t)) - 2\mathrm{Cov}(A(t), Q(t)). \tag{11}$$

Now since $\rho < 1$ and $g_3 < \infty$, $\mathrm{Var}(Q(t))$ converges to σ_π^2 which is finite. Further, the covariance terms satisfy:

$$\mathrm{Cov}(Q(0), Q(t)) \leq \sqrt{\sigma_0^2 \,\mathrm{Var}(Q(t))}, \qquad \mathrm{Cov}(A(t), Q(t)) \leq \sqrt{\lambda t \,\mathrm{Var}(Q(t))}.$$

Thus we have that $\mathrm{Var}(D(t)) = \lambda t + O(\sqrt{t})$, where $O(\sqrt{t})$ denotes some function that increases at a rate bounded by a multiple of $\sqrt{t}$.

We now show that the $O(\sqrt{t})$ term is in fact $\overline{B} + o(1)$.

Theorem 3 *Consider the output counting process, $\{D(t), t \geq 0\}$ of the M/G/1 queue operating under an arbitrary work-conserving service discipline with $\rho < 1$, $g_3 < \infty$ and $\sigma_0^2 < \infty$. Then there exists a finite $\overline{B}$, such that*

$$Var(D(t)) = \lambda t + \overline{B} + o(1).$$

Proof. Our goal is to show that the two covariances in (11) converge to a constant: First observe that

$$\lim_{t\to\infty} \mathrm{Cov}(Q(0), Q(t)) = 0.$$

This follows from the regenerative structure of the queueing process and the fact that $g_3 < \infty$.
To show that $\mathrm{Cov}(A(t), Q(t))$ converges to a constant define $\tau(t) = \sup\{u \leq t | Q(u) = 0\}$. Now,

$$\begin{aligned} \mathbb{E}\left[A(t)Q(t)\right] &= \mathbb{E}\left[(A(\tau(t)) + A(t) - A(\tau(t)))Q(t)\right] \\ &= \mathbb{E}\left[A(\tau(t))\right]\mathbb{E}\left[Q(t)\right] + \mathbb{E}\left[A(t-\tau(t))\bar{Q}(t-\tau(t))\right], \end{aligned}$$

where $\bar{Q}(t)$ is distributed as the queue length of an M/G/1 queue at time t starting empty. The first term of the second equality follows from independence due to regeneration at $\tau(t)$. The second term follows since the pair $(A(t) - A(\tau(t)), Q(t))$ is distributed as the pair $\big(A(t-\tau(t)), \bar{Q}(t-\tau(t))\big)$. Combining the above we have:

$$\begin{aligned} \mathrm{Cov}(A(t), Q(t)) &= \mathbb{E}\left[A(t)Q(t)\right] - \mathbb{E}\left[A(t)\right]\mathbb{E}\left[Q(t)\right] \\ &= \mathbb{E}\left[A(t-\tau(t))\bar{Q}(t-\tau(t))\right] - \mathbb{E}\left[A(t) - A(\tau(t))\right]\mathbb{E}\left[Q(t)\right] \\ &= \mathbb{E}\left[A(t-\tau(t))\bar{Q}(t-\tau(t))\right] - \mathbb{E}\left[A(t-\tau(t))\right]\mathbb{E}\left[Q(t))\right]. \end{aligned} \tag{12}$$

Denote $U(t) = A(t-\tau(t))\bar{Q}(t-\tau(t))$ and observe that it is a regenerative process. Further observe that $U(t) \leq A(t-\tau(t))^2 \leq Y^2$ w.p. 1 where Y denotes a generic random variable of the number of customers served in an M/G/1 busy cycle. Further, denoting X as the duration of the busy cycle, observe that,

$$\lim_{t\to\infty} \mathbb{E}\left[U(t)\right] = \frac{\mathbb{E}\left[\int_0^X U(s)ds\right]}{\mathbb{E}\left[X\right]} \leq \frac{\mathbb{E}\left[XY^2\right]}{\mathbb{E}\left[X\right]} < \infty.$$

The last inequality follows from the fact that $g_3 < \infty$. Similar reasoning along with the fact that $\mathbb{E}\left[Q(t)\right]$ converges, shows that the right hand side term of (12) also converges to a constant and thus (12) converges to a constant as $t \to \infty$. ■

Note that with minor modifications the above result also applies to the GI/G/1 queue.

4 The M/G/1 Linear Asymptote

We now prove our main result regarding the linear asymptote of the M/G/1. Our proof is in two steps. We first apply Theorem 2 for the case where $Q(0) = 0$ w.p. 1 by using the moments $(x_1, x_2, x_3, y_1, y_2, n_{11}, n_{12}, n_{21})$. We then use a coupling argument to obtain the general result.

Lemma 2 *Let $\overline{B}_0$ denote the y-intercept in the case where the system starts empty. Then,*

$$\overline{B} = \sigma_0^2 + \overline{B}_0.$$

Proof. We use a coupling argument. Consider the following two systems under the same sample path of the arrival process and service times. System 0 starts with $Q(0) = 0$ and system 1 starts with $Q(0) = Q_0$. Operate system 1 by giving low priority to the initial Q_0 customers, that is, these customers are only served with preemption when there are no other customers in system. This implies that the first Q_0 customers of system 1 are being served only during idle periods of System 0. Thus after a finite time T, the trajectories of the two systems coincide. Thus for $t \geq T$, $D_1(t) = D_0(t) + Q_0$, where $D_i(\cdot)$ denote the output counting processes. Taking variances yields the result. A similar argument hold for the non-preemptive case. ■

The above proof explains the effect of the initial queue level distribution on the y-intercept. We now prove our main result:

Theorem 4 *Consider the output counting process, $\{D(t), t \geq 0\}$, of the M/G/1 queue operating under an arbitrary work-conserving non-preemptive service discipline with $\rho < 1$ and $g_3 < \infty$. If $\sigma_0^2 < \infty$ then,*

$$Var(D(t)) = \lambda t + \sigma_0^2 - \frac{\rho}{(1-\rho)^2} + o(1).$$

Proof. [1] First we calculate $\overline{B}_0$ (as defined in Lemma 2). The moments $(x_1, x_2, x_3, y_1, y_2, n_{11}, n_{12}, n_{21})$ are evaluated in the standard manner by using (6) and taking the idle period into account. For completeness, we list the expressions, denoting $\rho_k = \lambda^k g_k$ for $k = 2, 3$:

$$\begin{aligned}
x_1 &= \lambda^{-1} \cdot \frac{1}{1-\rho}, \\
x_2 &= \lambda^{-2} \cdot \frac{\rho_2 + 2(1-\rho)^2}{(1-\rho)^3}, \\
x_3 &= \lambda^{-3} \cdot \frac{6 - 24\rho^3 + 6\rho^4 + 3\rho_2 + 3\rho_2^2 + 3\rho^2(12+\rho_2) + \rho_3 - \rho(24 + 6\rho_2 + \rho_3)}{(1-\rho)^5}, \\
y_1 &= \frac{1}{1-\rho}, \\
y_2 &= \frac{\rho_2 - \rho^2 + 1}{(1-\rho)^3}, \\
n_{11} &= \lambda^{-1} \cdot \frac{\rho_2 + 1 - \rho}{(1-\rho)^3}, \\
n_{12} &= \lambda^{-2} \cdot \frac{3\rho_2^2 + 3\rho_2(1-\rho) + (1-\rho)(\rho_3 + 2(1-\rho)^2)}{(1-\rho)^5},
\end{aligned}$$

[1] Note that an alternative (less elegant) proof follows by a direct application of Theorem 2 along with the additional moments $(\tilde{x}_1, \tilde{x}_2, \tilde{y}_1, \tilde{y}_2, \tilde{n}_{11})$ of the initial busy cycle. In fact, this method is slightly less general since it requires the initial residual service time to be of finite variance. This fact appears since the moments of the initial busy cycle depend on the joint moments of the initial queue level and residual service time. Once these moments are plugged in (9), the residual service time moments and cross moments are canceled out.

$$n_{21} \quad = \quad \lambda^{-1} \cdot \frac{3\rho_2^2 + 3\rho_2(1-\rho) + (1-\rho)(\rho_3 + 2(1-\rho)^2)}{(1-\rho)^5}.$$

Inserting the above into (9) and using $(\tilde{x}_1, \tilde{x}_2, \tilde{y}_1, \tilde{y}_2, \tilde{n}_{11}) = (\lambda^{-1}, 2\lambda^{-2}, 0, 0, 0)$ we obtain

$$\overline{B}_0 = -\frac{\rho}{(1-\rho)^2}.$$

The result now follows from Lemma 2. ■

Some immediate properties of $\overline{B}$ should be mentioned: First observe that $\overline{B}_0$ is a lower bound for the y-intercept that is obtained when $Q(0)$ is constant w.p. 1. Second, as $\rho \to 1$ from below, $\overline{B} \to -\infty$. Finally, observe that the sign of $\overline{B}$ is determined by the variability of $Q(0)$ compared to the variability of the queue length of a stationary M/M/1 queue.

Covariance of Arrivals and Queue Size

The proof of Theorem 3 in Section 3 showed that $\mathrm{Cov}(A(t), Q(t))$ converges to a constant. Having obtained $\overline{B}$, this constant is actually immediately available:

Corollary 3 *Consider the M/G/1 queue with arrival counting process $A(t)$ and queue level process $Q(t)$. Assume $\rho < 1$, $g_3 < \infty$ and $\sigma_0^2 < \infty$. Then,*

$$\lim_{t\to\infty} Cov(A(t), Q(t)) = \frac{\sigma_\pi^2 + \frac{\rho}{(1-\rho)^2}}{2}.$$

Proof. Applying Theorem 4 to (11) we obtain:

$$\lambda t + \sigma_0^2 - \frac{\rho}{(1-\rho)^2} + o(1) = \lambda t + \sigma_0^2 + \mathrm{Var}(Q(t)) - 2\mathrm{Cov}(Q(0), Q(t)) - 2\mathrm{Cov}(A(t), Q(t)).$$

Cancelling terms, taking t to infinity and rearranging, we obtain the result. ■

Note that Corollary 3 implies that for the M/M/1 queue with an arbitrary distribution of the number of customers having finite variance,

$$\lim_{t\to\infty} \mathrm{Cov}(A(t), Q(t)) = \frac{\rho}{(1-\rho)^2}.$$

5 The Stationary M/G/1

We now consider the linear asymptote of the stationary M/G/1. In this case,

$$\overline{B} = \sigma_\pi^2 - \frac{\rho}{(1-\rho)^2},$$

and σ_π^2 can be easily evaluated by using the well known Pollaczek-Khinchine formula. For our purposes it is fruitful to parameterize σ_π^2 by the offered load ρ, the squared coefficient of variation c^2 and the skewness coefficient γ of $G(\cdot)$. To avoid ambiguity, we specify these parameters in terms of the moments:

$$c^2 = \frac{g_2}{g_1^2} - 1, \qquad \gamma = \frac{2g_1^3 - 3g_1g_2 + g_3}{(g_2 - g_1^2)^{3/2}}.$$

A useful representation of σ_π^2 for our purposes is:

$$\sigma_\pi^2 = (L+1)\frac{\rho}{(1-\rho)^2},$$

where,

$$L = \Big(\frac{1}{4}c^4 - \frac{1}{3}\gamma c^3 + \frac{1}{2}c^2 - \frac{1}{12}\Big)\rho^3 + \Big(\frac{1}{3}\gamma c^3 - \frac{3}{2}c^2 + \frac{5}{6}\Big)\rho^2 + \Big(\frac{3}{2}c^2 - \frac{3}{2}\Big)\rho.$$

It is immediately evident that the signs of L and $\overline{B}$ are equal: A positive sign of L implies higher variability than that of an M/M/1 queue and in turn $\overline{B}$ is positive. A negative sign implies the opposite. Having $L = 0$ implies that $\overline{B} = 0$. Note that for the exponential distribution, $c = 1$, $\gamma = 2$ and indeed (as expected) $L = 0$. This also shows that all service time distributions that agree with the exponential distribution on the first 3 moments yield $\overline{v}(t) = \lambda t$. Further it is easy to check that when $\gamma = 2$ the sign of L equals the sign of $c - 1$ and it is monotone in c. Moreover, when $c = 1$, the sign of L equals the sign of $\gamma - 2$ and it is monotone in γ.

Daley's Conjecture

It is an open problem to find the class of service time distributions for which (5) yields $V^*(s) = \lambda/s$ (equivalent to $\mathrm{Var}(D(t)) = \lambda t$). Daley conjectured, [4], that this is true only for the exponential service time distribution. Note that in [4] he also obtains the LST of the variance curve of the stationary G/M/1 queue and proves that within this class of queueing systems, $\mathrm{Var}(D(t)) = \lambda t$ only when the arrival process is Poisson.

It is obvious that for a steady state system to have a variance curve of λt we must have $\overline{B} = 0$. Thus having $L = 0$ is a necessary condition for the variance curve to be λt. Surprisingly it is not a sufficient condition. For example, consider a service time distribution with the following LST:

$$G^*(s) = \frac{1}{384}\big(192\frac{1}{1+s} + 147e^{-\frac{1}{2}s} + 8e^{-\frac{3}{2}s} + 30e^{-\frac{5}{2}s} + 7e^{-\frac{9}{2}s}\big).$$

It is a mixture of an exponential distribution and masses at 4 points. It is straightforward to check that it agrees with a mean 1 exponential random variable on the first 3 moments[2] and thus it yields $L = 0$. Yet the variance curve is not λt. We observed this by calculating the LST (5) for a grid of points of s, where for each point we solved the fixed point equation (6) numerically and the resulting LST was not equal to λ/s.

Acknowledgment

We thank the following persons for useful discussions and advice: Ahmad Al Hanbali, Onno Boxma and David Perry.

References

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

[2] M. Brown and H. Solomon. A second order approximation for the variance of a renewal reward process. *Stochastic Processes and their Applications*, 3:301–314, 1974.

[3] P.J. Burke. The output of a queuing system. *Operations Research*, 4(6):699–704, 1956.

[4] D.J. Daley. Further second-order properties of certain single-server queueing systems. *Stochastic Processes and their Applications*, 3:185–191, 1975.

[2] A comment of independent interest: We tried finding such a distribution within the class of Matrix-Exponential distributions (rational LST) and have so far been unsuccessful.

[5] D.J. Daley. Queueing output processes. *Adv. Appl. Prob.*, 8:395–415, 1976.

[6] D.J Daley and R. Vesilo. Long range dependence of point processes, with queueing examples. *Stochastic Processes and Their Applications*, 70(2):265–282, 1997.

[7] R. L. Disney and P. C. Kiessler. *Traffic Processes in Queueing Networks – A Markov Renewal Approach*. The Johns Hopkins University Press, 1987.

[8] R. L. Disney and D. Konig. Queueing networks: A survey of their random processes. *SIAM Review*, 27(3):335–403, 1985.

[9] R.L. Disney, R.L. Farrell, and P.R. de Morais. A characterization of M/G/1 queues with renewal departure processes. *Management Science*, 19(11):1222–1228, 1973.

[10] P. D. Finch. The output process of the queueing system M/G/1. *Journal of the Royal Statistical Society, Series B (Methodological)*, 21(2):375–380, 1959.

[11] S. B. Gershwin. Variance of output of a tandem production system. *in: Queueing Networks with Finite Capacity,* eds R. Onvural and I. Akyildiz, Proceedings of the Second International Conference on Queueing Networks with Finite Capacity (Elsevier, Amsterdam)., 1993.

[12] K. B. Hendricks. The output processes of serial production lines of exponential machines with finite buffers. *Operations Research*, 40(6):1139–1147, 1992.

[13] K. B. Hendricks and J. O. McClain. The output processes of serial production lines of general machines with finite buffers. *Management Science*, 39(10):1194–1201, 1993.

[14] G. J. Miltenburg. Variance of the number of units produced on a transfer line with buffer inventories during a period of length T. *Naval Research Logistics*, 34:811–822, 1987.

[15] Y. Nazarathy and G. Weiss. The asymptotic variance rate of finite capacity birth-death queues. *Queueing Systems*, 59(2):135–156, 2008.

[16] Y. Nazarathy and G. Weiss. Positive Harris recurrence and diffusion scale analysis of a push pull queueing network. *Performance Evaluation - To appear*, 2008.

[17] N.U. Prabhu. *Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication*. Springer, 1998.

[18] J. F. Reynolds. The covariance structure of queues and related processes - a survey of recent work. *Adv. Appl. Prob.*, 7:383–415, 1975.

[19] B. Tan. Variance of the output as a function of time: Production line dynamics. *European Journal of Operational Research*, 177(3):470–484, 1999.

[20] B. Tan. Asymptotic variance rate of the output in production lines with finite buffers. *Annals of Operations Research*, 93:385–403, 2000.