J. Service Science & Management, **2010, 3, 512-519**
doi:10.4236/jssm.2010.34058 Published Online December 2010 (http://www.SciRP.org/journal/jssm)

Transient Little's Law for the First and Second Moments of G/M/1/N Queue Measures

Avi Herbon[1,2], Eugene Khmelnitsky[3]

[1]Dept. of Management, Bar-Ilan University, Ramat-Gan, Israel; [2]Dept. of Industrial Engineering and Management, Ariel University Center of Samaria, Ariel, Israel; [3]Dept. of Industrial Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel.
Email: xmel@eng.tau.ac.il

Received July 22nd, 2010; revised September 10th, 2010, accepted October 18th, 2010.

ABSTRACT

A customer in a service system and an outside observer (*manager or designer of the system*) *estimate the system performance differently. Unlike the outside observer, the customer can never find himself in an empty system. Therefore, the sets of scenarios, relevant for the two at a given time, differ. So differ the meanings and values of the performance measures of the queue: expected queue length and expected remaining waiting time* (*workload*). *The difference between the two viewpoints can be even more significant when steady-state values of the queue measures are reached slowly, or even are never reached. In this paper, we obtain the relations between the means and variances of the measures in transient time and in steady state for a capacitated FCFS queue with exponentially distributed service time. In particular, a formula similar to Little's law is derived for the means of the queue measures. Several examples support the validity and significance of the results.*

Keywords: *Queuing systems, Service operations, Transient time analysis, Customer-oriented measures*

1. Introduction

When dealing with queuing systems, the measures that are often of interest are mean queue length, L, and mean workload (remaining waiting time), W. The L and W measures are associated with customers that wait for service in a queue; however, its reduction is important for both the customers and the system manager. The meaning and the value of the L and W measures are different when the queue is observed from the viewpoint of either the customers, or the manager. The difference is in the fact that the scenarios at which the queue is empty, are not observable by the customers, and are not accounted for by the customer-oriented measures. The manager-oriented measures account for the entire set of scenarios. The manager has to be concerned not only about his/her measures, but also about the customer measures, since it is the customer measures that yield actual feedback to the manager decision support system.

With that in mind, we define the workload and the queue length stochastic processes as follows,

W_t - is the workload, *i.e.*, the remaining waiting time, observed by the last customer in the system at t.

L_t - is the number of customers in the system at t, observed by the last customer in the system.

$\hat{W}_t$ - is the workload, *i.e.*, the remaining waiting time of customers at t.

$\hat{L}_t$ - is the number of customers in the system at t.

The $\hat{L}_t$ and $\hat{W}_t$ processes are defined in terms of the entire system, rather than correspond to a specific customer within the system. If, for example, no customer is in the system at t, then $\hat{L}_t = 0$ and $\hat{W}_t = 0$. Both L_t and W_t are customer-oriented; they are not defined if no customer is in the system at t. In the next section, we will define the above stochastic processes more precisely. In the next section we also show that the formula $E\left(\hat{L}\right) = \lambda E\left(W\right)$ (λ is the arrival rate of the customers in a GI/M/1 queue) stated in our terms is similar to the Little's law (Little [1]).

The relations between the expected waiting time and the expected queue length were commented on and analyzed in the works of Eilon [2], Maxwell [3], Stidham [4], Keilson and Servi [5]. Other papers, such as Brumelle [6], Brumelle [7], and Heyman and Stidham [8] showed that similar relations exist between more general measures: customer-averages, H, and time-averages, G, which is represented by the formula $H = \lambda G$. Other extensions include the distributional version of the

Little's law due to Haji and Newell [9] and Keilson and Servi [5], and the continuous system version due to Rolski and Stidham [10] and Glynn and Whitt [11]. A review of $L = \lambda W$ and its extensions is given in Whitt [12].

Since real-life queuing systems operate in a dynamic environment characterized by different initial conditions, limited space for waiting customers, and unstable arrival rate, the steady state of the queue might be attained only after a long time, or even cannot be reached at all. Therefore, some works stress the importance of transient analysis of queuing models. The papers of Bertsimas and Mourtzinou [13] and Riaño *et al.* [14] formulate the transient Little's law in an integral form by relating the expected number of customers in the system at *t*, to the waiting time of customers joining the system in the interval (0, *t*]. Diverse applications of the transient analysis additionally motivate the research in this field. In the work of Zhang [15] transient behavior of time-dependent M/M/1 queues was studied. This work numerically calculated transient probability distributions of *L* and *W* from a set of integral equations. The paper by Perry *et al.* [16] studied the dynamics of workload in the M/G/1 queue and presented various results on its distribution. The paper by Garcia *et al.* [17] derived a closed-form solution for the time-dependent probability distribution of the number of customers of the M/D/1/N queues initialized in an arbitrary deterministic state. That paper gives a number of examples where transient oscillating, rather than steady state, of the expected number of customers, is the major phenomenon of queue dynamics.

In this work we deal with the G/M/1/N queue, *i.e.*, with a capacitated, single server queue governed by the FCFS discipline for the entering customers, by arbitrary (nonstationary) rate of arrivals, and by exponentially distributed service time. We develop dynamic, transient time relations between the means $E(W_t)$, $E(\hat{W}_t)$, $E(L_t)$ and $E(\hat{L}_t)$, as well as between the variances. The practical and theoretical significance of G/M/1/N and many specific results on such queue can be found in Cohen [18], Asmussen [19], and Adan *et al.* [20]. We extend those results by considering customer-oriented and observer-oriented measures of the queue and by studying their dynamics. Section 2 presents the mathematical analysis of the queue dynamics. Section 3 considers several special cases, providing insight into the general properties. Section 4 concludes the paper.

2. Mathematical Analysis

A scenario of the G/M/1/N queue is described by $\hat{L}_0$, $\hat{L}_0 \le N$, queue length at t = 0, a sequence of arrival times a_j and a sequence of service times Δs_j, $j = 1, 2, \cdots, \infty$. In case when $\hat{L}_0 > 0$, the arrival times of the customers present in the system at t = 0 are assumed zero, $a_1 = a_2 = \cdots = a_{\hat{L}_0} = 0$, without loss of generality. The arrivals that encounter a blocked system $\left(\hat{L}_t = N\right)$ are discarded, and therefore not indexed. We also denote the departure times by s_j, $j = 1, 2, \cdots, \infty$. The service times are assumed to be exponentially distributed with the mean $1/\mu$.

In a scenario in which $\hat{L}_t > 0$, let j_t^{last} be the index of the last customer in the system, and j_t be the index of the customer in service. Then,

$$\hat{L}_t = j_t^{last} - j_t + 1, \tag{1}$$

$$L_t = \begin{cases} \hat{L}_t, & \text{if } \hat{L}_t > 0 \\ \text{undefined}, & \text{if } \hat{L}_t = 0 \end{cases} \tag{2}$$

$$\hat{W}_t = \begin{cases} s_{j_t^{last}} - t, & \text{if } \hat{L}_t > 0 \\ 0, & \text{if } \hat{L}_t = 0 \end{cases} \tag{3}$$

$$W_t = \begin{cases} \hat{W}_t, & \text{if } \hat{L}_t > 0 \\ \text{undefined}, & \text{if } \hat{L}_t = 0 \end{cases} \tag{4}$$

Following the above notation and definitions, we prove the following lemma, which will be used later in proving the main results of this section.

Lemma 1: The departure time of the last customer is:

$$s_{j_t^{last}} = \sum_{i=1}^{j_t^{last}} \Delta s_i + \int_0^{a_{j_t}} \left(1 - \theta\left(\hat{L}_t\right)\right) dt, \tag{5}$$

where $\theta(L)$ is the step-function, $\theta(L) = 0$ if $L = 0$, and $\theta(L) = 1$ if $L > 0$.

Proof: For an arbitrary *j*, the total time before the *j*-th departure includes the service times of the first *j* customers and the total idle time of the system up to a_j, *i.e.*,

$$s_j = \sum_{i=1}^{j} \Delta s_i + \int_0^{a_j} \left(1 - \theta\left(\hat{L}_t\right)\right) dt. \tag{6}$$

For $j = j_t^{last}$, the system is not idle within $t \in \left(a_{j_t}, a_{j_t^{last}}\right)$. Therefore, the second term of (6) is re-written as $\int_0^{a_{j_t}} \left(1 - \theta\left(\hat{L}_t\right)\right) dt$. The proof of the lemma is completed.

The next theorem states the basic relation between the expectations of customer-oriented workload and queue length.

Theorem 1. For all *t* and $\mu > 0$,

$$E(W_t)=\frac{1}{\mu}E(L_t). \tag{7}$$

Proof: By combining (3-5), we obtain,

$$E(W_t)=E\left(\sum_{i=1}^{j_t^{last}}\Delta s_i+\int_0^{a_{j_t}}\left(1-\theta(\hat{L}_\tau)\right)d\tau \mid \hat{L}_t>0\right)-t,$$

or, equivalently,

$$\begin{aligned}E(W_t)=&E\left(\sum_{i=j_t+1}^{j_t^{last}}\Delta s_i \mid \hat{L}_t>0\right)+\\&+E\left(\sum_{i=1}^{j_t}\Delta s_i+\int_0^{a_{j_t}}\left(1-\theta(\hat{L}_\tau)\right)d\tau \mid \hat{L}_t>0\right)-t\end{aligned} \tag{8}$$

From (6) it follows that the second term in (8) is

$$E\left(\sum_{i=1}^{j_t}\Delta s_i+\int_0^{a_{j_t}}\left(1-\theta(\hat{L}_\tau)\right)d\tau \mid \hat{L}_t>0\right)=E\left(s_{j_t} \mid \hat{L}_t>0\right).$$

Now (8) can be re-written as

$$E(W_t)=E\left(\sum_{i=j_t+1}^{j_t^{last}}\Delta s_i \mid \hat{L}_t>0\right)+E\left(s_{j_t}-t \mid \hat{L}_t>0\right). \tag{9}$$

Since the service times, Δs_i, are mutually independent, and independent of $\hat{L}_t$, the first term in (9) is,

$$\begin{aligned}E\left(\sum_{i=j_t+1}^{j_t^{last}}\Delta s_i \mid \hat{L}_t>0\right)&=\frac{1}{\mu}E\left(j_t^{last}-j_t \mid \hat{L}_t>0\right)\\&=\frac{1}{\mu}E\left(\hat{L}_t-1 \mid \hat{L}_t>0\right)=\frac{1}{\mu}E(L_t)-\frac{1}{\mu}\end{aligned}$$

The second term in (9) is the expected remaining service time of the customer in service; it is equal to $1/\mu$. This completes the proof of the theorem.

The proven relationship agrees with the intuition of a customer who relates the expected remaining waiting time to the number of customers in the system and to the service rate.

Theorem 2 below derives the relationship between the variances of workload and queue length. In the theorem p_t^0 denotes the probability of observing an empty system at *t*, *i.e.*, $\Pr\left(\hat{L}_t=0\right)=p_t^0$.

Theorem 2. For $\mu>0$ and all *t* when $p_t^0<1$,

$$Var(W_t)=\frac{1}{\mu^2}\left(Var(L_t)+E(L_t)\right). \tag{10}$$

Proof: The workload variance is given as,

$$Var(W_t)=E\left(W_t^2 \mid \hat{L}_t>0\right)-\left(E(W_t \mid \hat{L}_t>0)\right)^2. \tag{11}$$

From (9) we notice that W_t is composed of L_t terms each distributed exponentially. That is, for the set of scenarios for which $L_t=k$ for every $k=1,2,\cdots,N$, W_t is distributed Erlang with rate μ and shape parameter *k*. The first term in (11) is,

$$\begin{aligned}E\left(W_t^2 \mid \hat{L}_t>0\right)&=\frac{1}{1-p_t^0}\sum_{k=1}^{N}\Pr\left(\hat{L}_t=k\right)E\left(W_t^2 \mid \hat{L}_t=k\right)\\&=\frac{1}{1-p_t^0}\sum_{k=1}^{N}\Pr\left(\hat{L}_t=k\right)\left(\frac{k}{\mu^2}+\frac{k^2}{\mu^2}\right)=\\&=\frac{1}{\mu^2}\frac{1}{1-p_t^0}\left(E\left(\hat{L}_t\right)+E\left(\hat{L}_t^2\right)\right)\end{aligned} \tag{12}$$

In order to substitute $\hat{L}_t$ with L_t in (12), we take the expectation of both sides in (2) and obtain,

$$E(L_t)=E\left(\hat{L}_t \mid \hat{L}_t>0\right)=\frac{E\left(\hat{L}_t\right)}{1-p_t^0},$$

or, equivalently,

$$E\left(\hat{L}_t\right)=\left(1-p_t^0\right)E(L_t). \tag{13}$$

Similarly,

$$E\left(\hat{L}_t^2\right)=\left(1-p_t^0\right)E\left(L_t^2\right). \tag{14}$$

Now, by substituting (13,14) into (12), we have

$$\begin{aligned}E\left(W_t^2 \mid \hat{L}_t>0\right)&=\frac{1}{\mu^2}\left(E(L_t)+E\left(L_t^2\right)\right)\\&=\frac{1}{\mu^2}\left(E(L_t)+\left(E(L_t)\right)^2+Var(L_t)\right)\end{aligned}. \tag{15}$$

From Theorem 1, the second term in (11) is,

$$\left(E\left(W_t \mid \hat{L}_t>0\right)\right)^2=\frac{1}{\mu^2}\left(E(L_t)\right)^2. \tag{16}$$

By substituting (15,16) into (11), we obtain the theorem.

Theorem 2 shows that the variance of workload is affected by the first two moments of queue length. The next two theorems prove the transient relations between the customer- and observer-oriented measures. They also further strengthen the analysis of the transient behavior of the measures.

Theorem 3. For $\mu>0$ and all *t* when $p_t^0<1$,

$$E\left(\hat{L}_t\right)=\left(1-p_t^0\right)\mu E(W_t) \tag{17}$$

Proof: The proof follows immediately from (7) and

(13).

Note that if $p_t^0 = 1$, the left-hand side of (17) equals zero, while the right-hand side is not defined, since $E(W_t)$ is not defined. If $\mu = 0$ and $\hat{L}_0 > 0$, $E(W_t)$ is infinite, and $E(\hat{L}_t)$ is finite; $E(\hat{L}_t) = N$ after the arrival of the *N*-th customer. Expressions (13,17), as well as,

$$E(\hat{W}_t) = (1 - p_t^0) E(W_t)$$

that follows from (4), determine the relationships of the expected queue length and expected workload between the two viewpoints.

For GI/M/1 if $E(\hat{L}_t)$ and $E(W_t)$ converge when $t \to \infty$, and p_t^0 converges to $1 - \lambda/\mu$, then (17) takes a form similar to Little's law,

$$E(\hat{L}) = \lambda E(W). \tag{18}$$

In the other cases, *i.e.*, either in a transient state, or when one of the variables in (17) does not converge at all, expression (17) generalizes (18).

Theorem 4. For $\mu > 0$ and all *t* when $p_t^0 < 1$,

$$Var(\hat{L}_t) = \mu\left((1 - p_t^0)\right)\left(\mu Var(W_t) + \mu p_t^0 \left(E(W_t)\right)^2 - E(W_t)\right) \tag{19}$$

Proof: When calculating the variance of $\hat{L}_t$ in terms of the first two moments of W_t, we make use of Theorems 1-3, as follows,

$$\begin{aligned} Var(\hat{L}_t) &= E(\hat{L}_t^2) - \left(E(\hat{L}_t)\right)^2 = (1 - p_t^0) E(L_t^2) - \left(E(\hat{L}_t)\right)^2 \\ &= (1 - p_t^0)\left(Var(L_t) + \left(E(L_t)\right)^2\right) - \left(E(\hat{L}_t)\right)^2 \\ &= (1 - p_t^0)\left(\mu^2 Var(W_t) - E(L_t) + \left(E(L_t)\right)^2\right) - \left(E(\hat{L}_t)\right)^2 \\ &= (1 - p_t^0)\mu\left(\mu Var(W_t) - E(W_t) + \mu p_t^0 \left(E(W_t)\right)^2\right) \end{aligned}$$

The proof of the theorem is completed.

For GI/M/1 if the distributions of $\hat{L}_t$ and W_t converge, when $t \to \infty$, and p_t^0 converges to $1 - \lambda/\mu$, then (19) can be considered as the Little's law for second moments,

$$Var(\hat{L}) = \lambda\left((\mu - \lambda)\left(E(W)\right)^2 + \mu Var(W) - E(W)\right) \tag{20}$$

Similar to Theorems 1 and 2, the relationships within the manager-oriented frame are,

$$E(\hat{W}_t) = \frac{1}{\mu} E(\hat{L}_t) \text{ and } Var(\hat{W}_t) = \frac{1}{\mu^2}\left(Var(\hat{L}_t) + E(\hat{L}_t)\right) \tag{21}$$

The next theorem proves that when the coefficient of variation of L_t is not too large, the variance of L_t is not greater than the variance of $\hat{L}_t$.

Theorem 5. For $\mu > 0$ and all *t* when $p_t^0 < 1$, if $\frac{Var(L_t)}{\left(E(L_t)\right)^2} \le 1 - p_t^0$, then

$$Var(\hat{L}_t) \ge Var(L_t). \tag{22}$$

Proof: From (13) and (14) we have

$$\begin{aligned} Var(\hat{L}_t) &= E(\hat{L}_t^2) - \left(E(\hat{L}_t)\right)^2 \\ &= (1 - p_t^0) E(L_t^2) - (1 - p_t^0)^2 \left(E(L_t)\right)^2 \\ &= (1 - p_t^0)\left(Var(L_t) + \left(E(L_t)\right)^2 - (1 - p_t^0)\left(E(L_t)\right)^2\right) \\ &= Var(L_t) - p_t^0\left(Var(L_t) - (1 - p_t^0)\left(E(L_t)\right)^2\right) \end{aligned}$$

Now, the theorem immediately follows.

3. Examples

This section illustrates the results obtained in the previous section and highlights the differences in the queue measures as viewed by the customers and by the manager, as well as the transient behavior of the measures. The examples below also show how (17) can be used in determining the expected dynamic workload.

Example 1: Pure death process

The pure death process is a special case of G/M/1, where no new customers arrive after *t* = 0 and $\hat{L}_0 > 0$.

The customers are served with the rate μ until all of them leave the system. The probability, p_t^n, of *n* customers at time *t* is known to be

$$p_t^n = \begin{cases} e^{-\mu t} \dfrac{(\mu t)^{\hat{L}_0 - n}}{(\hat{L}_0 - n)!} & n = 1, \cdots, \hat{L}_0 \\ 1 - \displaystyle\sum_{k=0}^{\hat{L}_0 - 1} \frac{e^{-\mu t}(\mu t)^k}{k!} & n = 0 \end{cases} \tag{23}$$

The transient Little's law (17) for this queue takes the following form,

$$E(\hat{L}_t) = \mu \sum_{k=0}^{\hat{L}_0 - 1} \frac{e^{-\mu t}(\mu t)^k}{k!} E(W_t). \tag{24}$$

The expected queue length is found from (23),

$$E(\hat{L}_t) = e^{-\mu t} \sum_{k=1}^{\hat{L}_0} k \frac{(\mu t)^{\hat{L}_0 - k}}{(\hat{L}_0 - k)!}. \tag{25}$$

Now, the expected customer-oriented workload is found by substituting (25) into (24),

$$E(W_t) = \frac{1}{\mu} \frac{\displaystyle\sum_{k=1}^{\hat{L}_0} k \frac{(\mu t)^{\hat{L}_0 - k}}{(\hat{L}_0 - k)!}}{\displaystyle\sum_{k=0}^{\hat{L}_0 - 1} \frac{(\mu t)^k}{k!}}. \tag{26}$$

When $t \to \infty$, $E(L_t) \to 1$, $E(\hat{L}_t) \to 0$, $E(W_t) \to 1/\mu$ and $E(W_t) \to 0$, which supports the intuition regarding the time-limit behavior of the pure death process.

Similarly, the customer-oriented workload variance is calculated from (10),

$$Var(W_t) = \frac{1}{\mu^2} \frac{\sum_{k=1}^{\hat{L}_0} k^2 \frac{(\mu t)^{\hat{L}_0 - k}}{(\hat{L}_0 - k)!}}{\sum_{k=0}^{\hat{L}_0 - 1} \frac{(\mu t)^k}{k!}} - \frac{1}{\mu^2} \left(\frac{\sum_{k=1}^{\hat{L}_0} k \frac{(\mu t)^{\hat{L}_0 - k}}{(\hat{L}_0 - k)!}}{\sum_{k=0}^{\hat{L}_0 - 1} \frac{(\mu t)^k}{k!}} \right)^2 + \frac{1}{\mu^2} \frac{\sum_{k=1}^{\hat{L}_0} k \frac{(\mu t)^{\hat{L}_0 - k}}{(\hat{L}_0 - k)!}}{\sum_{k=0}^{\hat{L}_0 - 1} \frac{(\mu t)^k}{k!}} \tag{27}$$

When $t \to \infty$, $Var(L_t) \to 0$ and $Var(W_t) \to 1/\mu^2$.

Example 2: Numerical simulation of D/M/1/N

The customers arrive with fixed interarrival times equal to $1/\lambda = 4$ and served with exponentially distributed service times with the mean $1/\mu = 4/5$. The initial queue length is 4, and the capacity of the queue is infinite. In running the simulation code, we have estimated the queue length measures. The remaining waiting time measures have been calculated from (7), (10) and (21).

Figure 1 presents the plots of the measures' means, $E(L_t)$, $E(\hat{L}_t)$, $E(W_t)$ and $E(\hat{W}_t)$. **Figure 2** presents the plots of the measures' variances, $Var(L_t)$, $Var(\hat{L}_t)$, $Var(W_t)$ and $Var(\hat{W}_t)$. This experiment shows that the perception of the dynamic behavior of a queue can differ significantly whether the manager or the customers are required to assess it. The simulated queue does not converge in time (from the manager's viewpoint) in spite of low utilization $(\lambda < \mu)$, while the customers observe the convergent behavior with low variability of the queue measures.

The next experiment illustrates queue behavior when the arrival rate is higher than the service rate, $\lambda = 1.9$, $\mu = 1.3$, $\hat{L}_0 = 8$, $N = 14$. Since the probability of observing an empty queue is always close to zero, both the manager and customers measures almost coincide. Therefore, **Figures 3** and **4** present only the customers measures. **Figure 5** compares $E(W_t)$ with the expected waiting time, which would follow from Little's law, *i.e.*, $E(\hat{L}_t)/\lambda(1-p_t^{14})$. A significant deviation between the last two expected waiting time measures can be noticed. For example, at t = 39, $E(L_t) \approx E(\hat{L}_t) \approx 13.185$, the expected waiting time of the last customer is estimated by $E(W_t)$ as 10.142, while the estimation made by the $E(\hat{L}_t)/\lambda(1-p_t^{14})$ measure is 16.662. Therefore, the $E(\hat{L}_t)/\lambda(1-p_t^{14})$ measure significantly over-estimates the expected waiting time in this case.

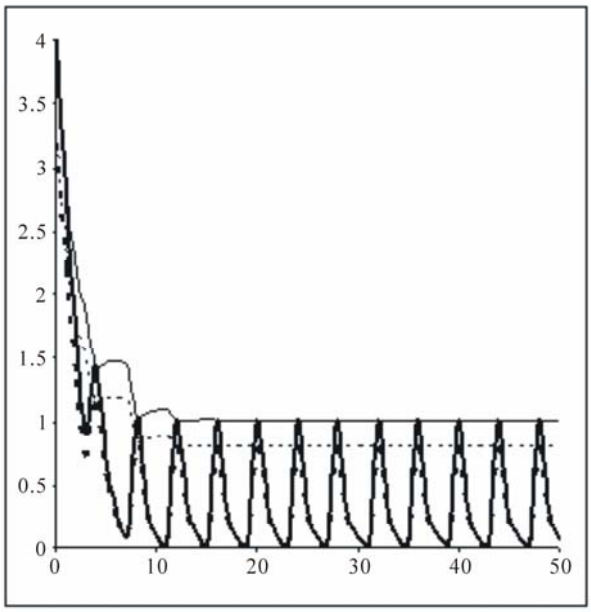

Figure 1. $E(L_t)$ **- thin line,** $E(\hat{L}_t)$ **- bold line,** $E(W_t)$ **- thin dashed line , and** $E(\hat{W}_t)$ **- bold dashed line.**

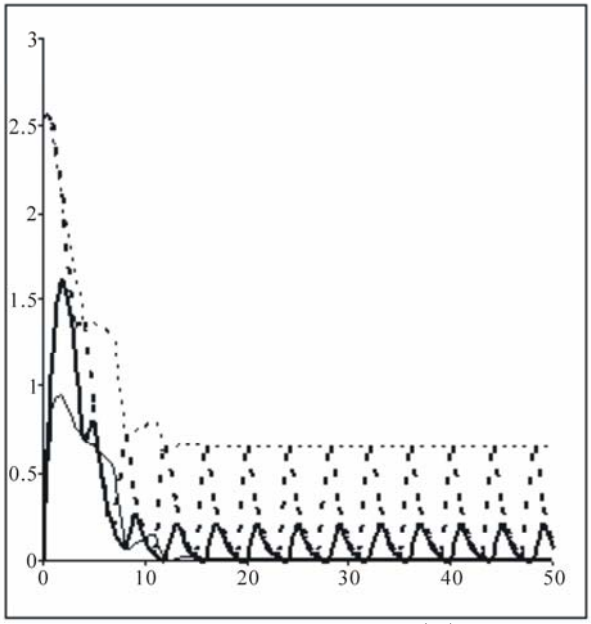

Figure 2. $Var(L_t)$ **- thin line,** $Var(\hat{L}_t)$ **- bold line,** $Var(W_t)$ **- thin dashed line, and** $Var(\hat{W}_t)$ **- bold dashed line.**

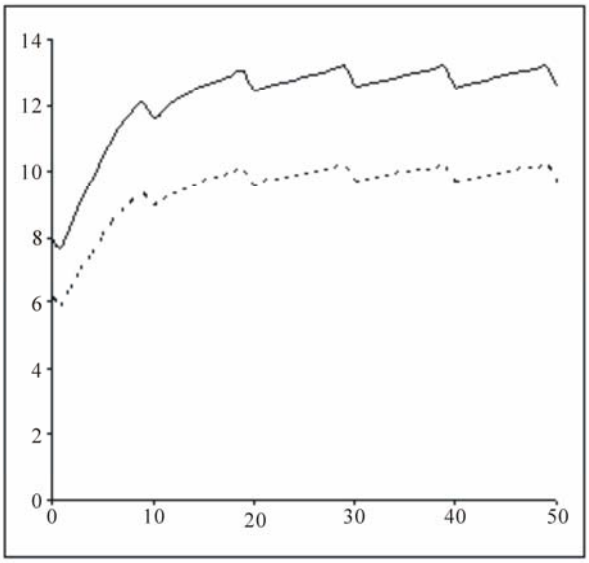

Figure 3. $E(L_t)$ **- thin line,** $E(W_t)$ **- thin dashed line**

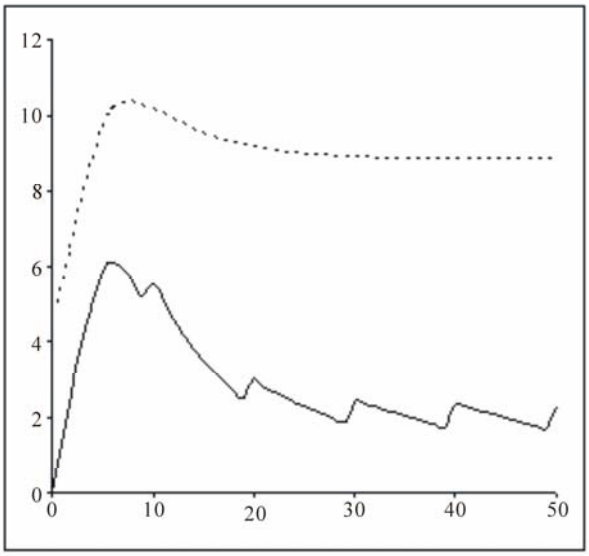

Figure 4. $Var(L_t)$ **- thin line,** $Var(W_t)$ **- thin dashed line.**

Example 3: Numerical simulation of G/M/1

Contrary to the previous example where the variance of the interarrival times was zero, here the variance of interarrival times is infinite. The interarrival times are distributed with the density $f(x) = 2/(1+x)^3$, whose mean is $1/\lambda = 1$. The service time parameter $\mu = 1.3$, $\hat{L}_0 = 8$, $N = \infty$. **Figure 6** presents the dynamics of $E(L_t)$ and $E(\hat{L}_t)$ simulated over the time interval of 300 time units, as well as the dynamics of $E(W_t)$ and $E(\hat{W}_t)$ calculated from (17,21). The transient time is very long, which makes the use of the Little's formula impractical in this case. **Figure 7** plots the dynamics of the ratio $\mu(1-p_t^0)/\lambda$, which shows the deviation of the expected waiting time obtained from the transient relation (17), from the expected waiting time which would follow from the steady-state Little's formula. As *t* goes to infinity, the ratio converges to one, and the relevance of the Little's law for the calculation of the expected waiting time increases.

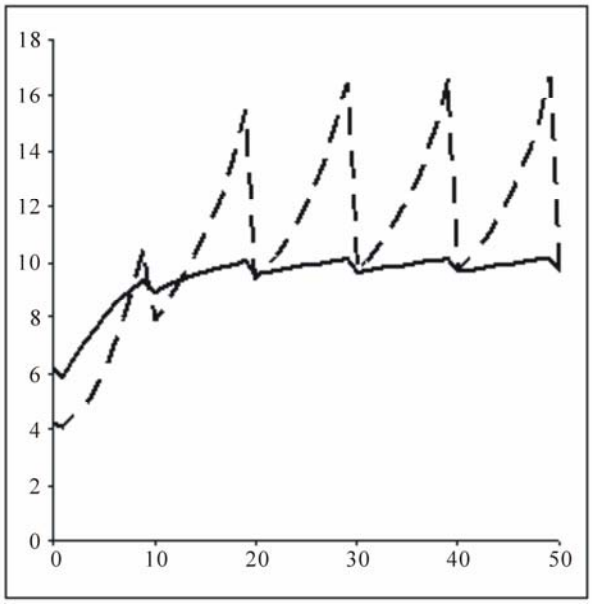

Figure 5. $E(W_t)$ **- bold line, and** $E(\hat{L}_t)/\lambda(1-p_t^{14})$ **- dashed line.**

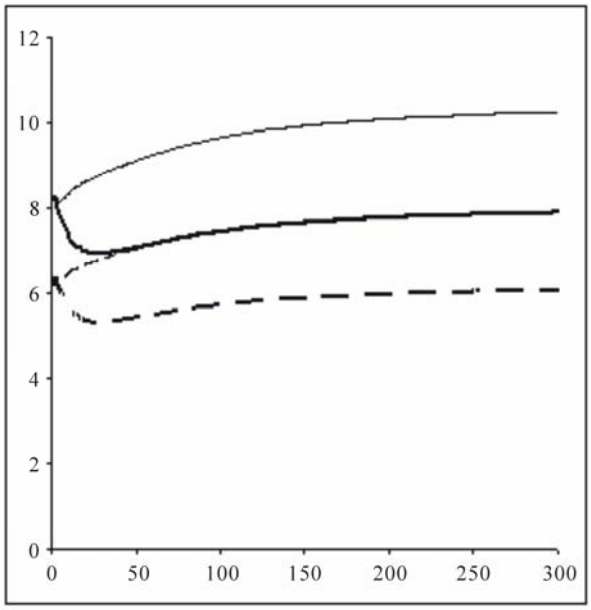

Figure 6. $E(L_t)$ **- thin line,** $E(\hat{L}_t)$ **- bold line,** $E(W)_t$ **- thin dashed line , and** $E(\hat{W}_t)$ **- bold dashed line.**

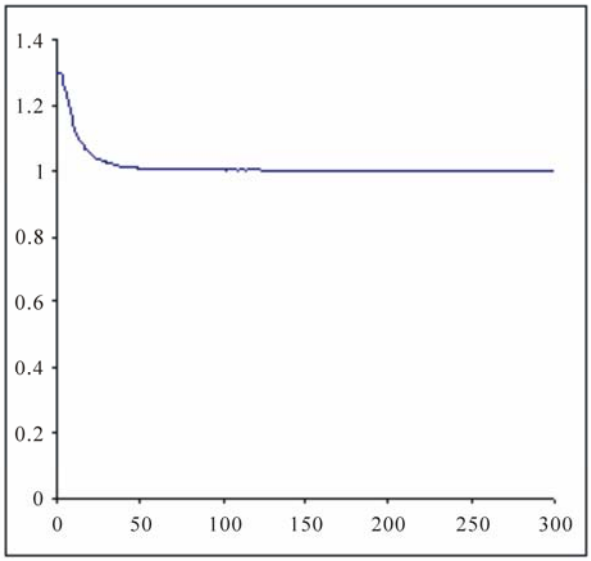

Figure 7. The dynamics of the ratio $\mu\left(1-p_t^0\right)/\lambda$ **.**

Note the difference between $E\left(L_t\right)$ and $E\left(\hat{L}_t\right)$ in the transient period as well as in steady-state. The existence of the difference indicates that the assessment of the queue length made by the manager significantly differs from the assessment made by the customers. We believe it is reasonable for the manager to take into account both observer-and customer-oriented measures in decision making. In particular, the manager can decide about how much space should be available for the waiting customers on the basis of either $E\left(\hat{L}_t\right)$, or $E\left(L_t\right)$.

4. Conclusions

Transient time relationships between customer- and manager-oriented measures of the FCFS G/M/1/N queue are derived and discussed. In particular, expression (17), which links the expected queue length as viewed by the manager to the customers' waiting time, generalizes Little's law for the considered queue. The relationships allow the first two moments of the expected remaining waiting time of the last customer (workload) to be easily calculated if the distribution of the queue length is known theoretically or by simulation. We have shown that the customer-and manager-oriented measures can significantly differ both qualitatively and quantitatively. The numerical study conducted in the paper stresses cases when the queue measures converge only in the long run, or do not converge at all. Further research is required in order to generalize the results of the paper to general queuing systems.

REFERENCES

[1] J. D. C. Little, "A Proof for The Queuing Formula," *Operations Research*, Vol. 9, No. 3, 1961, pp. 383-387.

[2] S. Eilon, "A Simpler Proof of $L=\lambda W$," *Operations Research*, Vol. 17, No. 5, 1969, pp. 915-917.

[3] W. Maxwell, "On the Generality of the Equation $L=\lambda W$ ", *Operations Research*, Vol. 18, No. 1, 1970, pp. 172-174.

[4] S. J. Stidham, "A Last Word on $L=\lambda W$," *Operations Research*, Vol. 22, No. 2, 1974, pp. 417-421.

[5] J. Keilson and L. D. Servi, "The Distributional Form of Little's Law", *Operations Research Letters*, Vol. 9, No. 4, 1990, pp. 239-247.

[6] S. L. Brumelle, "On the Relation between Customer and Time Averages in Queues," *Journal of Applied Probability*, Vol. 8, No. 3, 1971, pp. 508-520.

[7] S. L. Brumelle, "A Generalization of $L=\lambda W$ to Moments of Queue Length and Waiting Times," *Operations Research*, Vol. 20, No. 6, 1972, pp. 1127-1136.

[8] D. P. Heyman and S. Stidham, "The Relation between Customer and Time Averages in Queues," *Operations Research*, Vol. 28, No. 4, 1980, pp. 983-994.

[9] R. Haji and G. F. Newell, "A Relation between Stationary Queue and Waiting-Time Distributions," *Journal of Applied Probability*, Vol. 8, No. 3, 1971, pp. 617-620.

[10] T. Rolski and S. Stidham, "Continuous Versions of the Queuing Formulas $L=\lambda W$ and $H=\lambda G$," *Operations Research Letters*, Vol. 2**,** No. 5, 1983, pp. 211-215.

[11] P. W. Glynn and W. Whitt, "Extensions of the Queuing Relations $L=\lambda W$ and $H=\lambda G$," *Operations Research*, Vol. 37, No. 4, 1989, pp. 634-644.

[12] W. Whitt, "A Review of $L=\lambda W$ and Extensions," *Queueing Systems*, Vol. 9, No. 3, 1991, pp. 235-268.

[13] D. Bertsimas and G. Mourtzinou, "Transient Laws of Non-Stationary Queueing Systems and Their Applications," *Queueing Systems*, Vol. 25, No. 1-4, 1997, pp. 115-155.

[14] G. Riaño, R. Serfozo and S. Hackman, "A Transient Little's Law," Technical report, 2003, COPA Centro de Optimización y Probabilidad Aplicada, Universidad de los Andes and Georgia Institute of Technology.

[15] J. I. Zhang, "The Transient Solution of Time-Dependent *M*/*M*/1 Queues," *IEEE Transactions on Information Theory*, Vol. 37, No. 6, 1991, pp. 1690-1696.

[16] D. Perry, W. Stadje and S. Zacks, "The M/G/1 Queue with Finite Workload Capacity," *Queueing Systems*, Vol. 39, No. 1, 2001, pp. 7-22.

[17] J. -M. Garcia, O. Brun and D. Gauchard, "Transient Analytical Solution of M/D/1/N Queues," *Journal of Applied Probability*, Vol. 39, No. 4, 2002, pp. 853-

864.

[18] J. W. Cohen, "The Single Server Queue," North-Holland, Amsterdam, 1982.

[19] S. Asmussen, "Applied Probability and Queues," 2nd ed. Springer, Berlin, 2003.

[20] I. Adan, O. Boxma and D. Perry, "The $G/M/1$ Queue Revisited," *Mathematical Methods of Operations Research*, Vol. 62, No. 3, 2005, pp. 437-452.

