Flow time distribution in queues with customer batching and setup times
Qi-Ming He; E Jewkes
INFOR; Feb 1997; 35, 1; ABI/INFORM Global
pg. 76

FLOW TIME DISTRIBUTIONS IN QUEUES WITH CUSTOMER BATCHING AND SETUP TIMES[1]

QI-MING HE AND, E. JEWKES

Department of Management Sciences, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ABSTRACT

This paper examines the relationship between batch size and flow time in a single server queue in which individual customers are grouped before processing. A setup time precedes the processing time of each batch. The model has applications in manufacturing systems where flow and job shops follow one another or in assembly type manufacturing. In the first part of the paper, the Laplace-Stieltjes Transforms (LSTs) and the first and second moments of several flow time measures are derived. The second part of the paper presents numerical results which relate the batch size and the mean and variance of the flow time. It is shown computationally that the mean and variance of the flow time are generally minimized at different batch sizes, and hence tradeoffs exist in selecting an optimal batch size. The numerical results also show that the variability of the setup time and processing time increases the mean and variance of the flow time as well as the corresponding optimal batch sizes.

Key words: $E_N/G/1$ queue, flow time, customer batching, setup time, matrix analytic methods.

RÉSUMÉ

Cet article présente la relation entre la taille des lots et le temps de cycle dans une file d'attente à guichet simple dans laquelle les clients sont groupés avant d'être traités. Un temps de mise en course précède le traitement de chacun des lots. Le modèle proposé a des répercussions pratiques en ce qui concerne les systèmes manufacturiers où les lignes et les ateliers de production alternent, ainsi que pour les systèmes d'assemblage. Nous commençons par dériver les Transformées de Laplace-Stieltjes ainsi que les premiers et deuxièmes moments de plusieurs mesures de temps de cycle. Ensuite, nous présentons des résultats numériques reliant la taille des lots à la moyenne et à la variance du temps de cycle. Nous démontrons numériquement que la moyenne et la variance du temps de cycle sont généralement minimisées en fonction de la taille des lots, et que des compromis doivent être faits lors de la sélection d'une taille de lot optimale. Les résultats numériques démontrent aussi que la variabilité des temps de mise en course et de traitement augmente la moyenne et la variance du temps de cycle ainsi que la taille de lot optimale associée.

1. INTRODUCTION

In this paper, we examine a queueing system in which individual arrivals are gathered into batches of size N before they are processed. The need for batching is created by setup time which is carried out each time the server begins to process a group of customers. The service time of a batch consists of the setup time followed by the individual service times of all the customers in the batch.

The motivation for studying this problem stems from manufacturing systems where a job shop manufacturing stage follows a flow shop stage. Items processed one-by-one in the flow shop are gathered into batches to be processed in the job shop stage. An example in the electronics industry is where printed circuit boards have a variety of standard operations performed in a flow shop setting (e.g., attaching bar codes). Boards are processed one-by-one without significant setup time between even if the end products are different. Then the boards go to a component insertion stage (the job shop) where machine setup times are required between different types of boards. We are concerned with modelling the batch formation process between the flow shop and the job shop and determining the optimal batch size for the job shop stage.

Models of manufacturing systems which include batching generally examine mean performance measures as a function of the batch size. The variance of the flow time is also important in manufacturing settings as it is a measure of predictability of leadtimes. Generally, the lower

[1]Recd. Nov. 1995; Revd. Apl. and June 1996

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the variance of the flow time, the better a manufacturer is able to quote accurate delivery times to customers. Though more difficult to compute than mean flow times, the variance of flow times provides additional insight into the consequences of a particular batch sizing decision. If several batch sizes produce similar mean flow times but very different flow time variances, the nature of the tradeoffs made in selecting a particular batch size need to be understood.

Queues with batching of arrivals have been studied by several authors. Fabens [4] studied the $E_N/G/1$ queue without setup times. Fukuta [5] investigated a queueing model where an idle server will delay the start of service until $l > 1$ customers are waiting. However, batching of customers was not explicitly considered in this paper. More recently, Karmarkar [6] examined a system where individual *Poisson* arrivals form batches of size N before service and where service of a batch includes a setup time. $M/M/1$ and $M/G/1$ models were used as an approximation for the process and to explore the tradeoff between batch size and mean flow times and work in process levels. Sumita and Kijima [10] followed up on Karmarkar's work by numerically evaluating, with Laguerre transforms, the effect of batching on a number special cases of $GI/G/1$ queues. They obtained numerical results for the first moment of several related batch flow time measures for the m^{th} batch processed in a busy period for $m = 1, 10, 20, 30, 40$ and 50. They did not, however, provide analytical expressions for any of the performance measures. In addition, Sumita and Kijima [10] noted that the approximations made in Karmarkar [6] could be quite crude in estimating the performance measures.

The contribution of our paper is that it provides *exact* analytical expressions for the LST of various flow time measures when the arrival process of individual items is Poisson. From these analytical expressions, we are able to provide formulas and computational procedures for determining both the first and second moments of the flow times. We therefore give exact results for the approximations made by Karmarkar [10], give analytical expressions for the numerical work of Sumita and Kijima [10] and develop an understanding of how batch sizing affects flow time variance. Comparisons are made between the batch size which minimizes the mean of flow time and the batch size which minimizes the variance of the flow time. Numerical results show that the mean and variance are not generally minimized at the same batch size.

In Section 2, we define the model to be studied and several time measurements. In Section 3, the Laplace Stieltjes transform (LST) of the actual waiting time in an $E_N/G/1$ queue is given. The LSTs of the flow times are given in Section 4. In Section 5, the means and variances of the flow times are derived. In Section 6, an algorithm for computing the variances of flow times and their minimization is presented. Numerical results are given in Section 7, where a comparison is made between the batch size that minimizes the mean flow time and that which minimizes the flow time variance. Finally, in Section 8, we present proofs of some theorems in this paper.

2. PROBLEM DEFINITION AND NOTATION

In this section, we give a detailed description of the system and flow times of interest. We consider a single server queueing system where customers arrive according to a Poisson process with parameter λ. Let $\{U_n\}$ denote the interarrival times. The individual arrivals are gathered into batches of size N before service. The service time of a batch consists of a setup time, V_s (a general nonnegative r.v.), followed by N individual customer processing times. $F_s(x)$ and $f_s^*(s)$ are the distribution function and the LST of the setup time, respectively. Let $\{V_{pn}, n = 1, ..., N\}$ (*iid* general nonnegative r.v.s) be the processing times of the individual customers in a batch. $F_p(x)$ and $f_p^*(s)$ denote the distribution function and the LST of the processing time, respectively. The batches are served on a first formed first served basis and customers within a batch are served FCFS. Let

$$\rho(N) = \lambda \left(\mathbf{E}V_{p1} + \frac{\mathbf{E}V_s}{N} \right) \tag{1}$$

be the traffic intensity of the queueing system. We assume that $\rho(N) < 1$ so that the queueing system is stable. (Note: $\mathbf{E}X$ denotes the mathematical expectation of the random variable X.)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

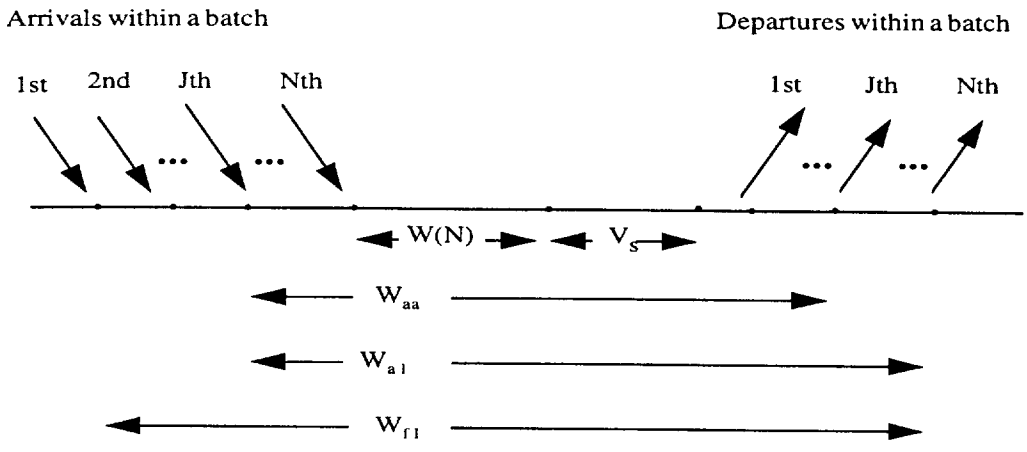

Figure 1: The Customer Arrival and Service Process

In a manufacturing system where individual items are batched for processing, there may be several flow times of interest. In particular, if the batch remains intact after processing, we may be interested in the time between the arrival of a random item to the time its batch is formed. If, on the other hand, the next stage in the manufacturing process is flow-shop in nature, items, once processed, may leave their batch as soon as they are processed and continue downstream without waiting for the rest of the batch to complete. The flow time of interest would be the time between the arrival of a random customer and its service completion. Finally, we may be interested in how long a reusable container (kanban) used to hold items in a batch is used at the manufacturing stage. In this case we would want to model the time from the first arrival in a batch to the time the batch completes service.

Based on this motivation, three flow times in our model are of interest. They are:

1. $W_{fl}(N)$: The duration between the first arrival in a batch and the batch completion time.
2. $W_{al}(N)$: The duration between the arrival of a random customer and its batch completion time.
3. $W_{aa}(N)$: The flow time of a random customer.

Each can be decomposed into the sum of three consecutive time periods. First, the time between the arrival epoch of an individual customer and the time its batch forms. Second, the time from when the batch is formed until the batch begins service. With exponential interarrival times, the inter-batch formation time has an *Erlang* Stage-N (E_N) distribution. If the batch formation times are considered as individual 'arrivals', existing flow time results from the $E_N/G/1$ queue can be used for this portion of the overall customer flow time. The waiting time for a batch in the $E_N/G/1$ queue will be denoted by $W(N)$. Third, the duration between when the batch in consideration begins service and when the customer or batch in consideration completes service.

Having defined $W(N)$, the three flow times can be specified (see Figure 1):

$$W_{fl}(N) = U_2 + \ldots + U_N + W(N) + V_s + V_{p1} + \ldots + V_{pN}; \quad (2)$$

$$W_{al}(N) = U_{J+1} + \ldots + U_N + W(N) + V_s + V_{p1} + \ldots + V_{pN}, \quad (3)$$

where J is a random variable with $P\{J = j\} = 1/N,\ 1 \leq j \leq N$; and

$$W_{aa}(N) = U_{J+1} + \ldots + U_N + W(N) + V_s + V_{p1} + \ldots + V_{pJ}, \quad (4)$$

given that the random customer is the J^{th} $(1 \leq J \leq N)$ arrival in its batch.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

It is worth noting that though results in the literature are available for computing $W(N)$, deriving the LST of the distribution of $W_{fl}(N), W_{al}(N)$ or $W_{aa}(N)$ is not straightforward because the interarrival times are not independent of $W(N)$.

3. THE ACTUAL WAITING TIME $W(N)$ IN AN $E_N/G/1$ QUEUE

In this section, we consider the actual waiting time $W(N)$ in the $E_N/G/1$ queue. Some literature pertaining to this system is Chaudhry and Templeton [1], Chaudhry *et al.* [2], Cohen [3], Neuts [8] and Prabhu [9]. Although the LST of $W(N)$, $w^*(N,s)$, has been obtained elsewhere (see Prabhu [9]), we give an alternate derivation which results in a form convenient for obtaining the LSTs of the three flow time distributions.

We denote by $V = V_s + V_{p1} + \ldots + V_{pN}$ the service time of a customer (a batch) in an $E_N/G/1$ queue. Let $F(x)$ be the distribution function of V with LST $f^*(s)$. Then $K = W(N) + V$ represents the sojourn time of a batch in the $E_N/G/1$ queue. The LST of K is $k^*(N,s) = w^*(N,s)f^*(s)$, since $W(N)$ and V are independent.

It is easy to see that $W(N)$ has the following recursive form:

$$W(N) = \max\{0, \hat{W}(N) + \hat{V} - (U_1 + \ldots + U_N)\}, \tag{5}$$

where $\hat{W}(N)$ (the waiting time of the previous batch) and $W(N)$ have the same probabilistic distribution, as do $\hat{V}$ (the service time of the previous batch) and V. Since $\hat{W}(N), \hat{V}$ and $\{U_i, 1 \leq i \leq N\}$ are independent, the following theorem holds:

Theorem 3.1.

The LST of the actual waiting time $W(N)$ of a batch in the $E_N/G/1$ queue satisfies equation (6) for $s \geq 0$ and $s \neq \lambda$,

$$w^*(N,s) = \frac{\lambda^N}{(\lambda - s)^N} w^*(N,s) f^*(s) + \sum_{i=0}^{N-1} \left(1 - \frac{\lambda^{N-i}}{(\lambda - s)^{N-i}}\right) k_i(N). \tag{6}$$

where

$$k_i(N) = \frac{(-\lambda)^i}{i!} \left.\frac{d^i k^*(N,s)}{ds^i}\right|_{s=\lambda} = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^i}{i!}\, dP\{W(N) + V < x\}, \quad i \geq 0. \tag{7}$$

Proof.

By equation (5), we have, for $s > 0$ and $s \neq \lambda$,

$$\begin{aligned}
\mathbf{E}\exp\{-sW(N)\} &= \mathbf{E}\exp\{-s\max\{0, \hat{W}(N) + \hat{V} - (U_1 + \ldots + U_N)\}\} \\
&= \int_0^\infty \mathbf{E}\exp\{-s\max\{0, x - (U_1 + \ldots + U_N)\}\}\, dP\{\hat{W}(N) + \hat{V} < x\} \\
&= \int_0^\infty \left[\int_0^x e^{-s(x-y)}\, dP\{U_1 + \ldots + U_N < y\}\right. \\
&\qquad \left. + P\{U_1 + \ldots + U_N \geq x\}\right] dP\{\hat{W}(N) + \hat{V} < x\}.
\end{aligned} \tag{8}$$

Since $U_1, \ldots, U_N$ are *iid* exponential random variables with parameter λ, we have

$$\begin{aligned}
P\{U_1 + \ldots + U_N \geq x\} &= \sum_{i=0}^{N-1} \frac{e^{-\lambda x}(\lambda x)^i}{i!} \\
\int_0^x e^{-s(x-y)}\, dP\{U_1 + \ldots + U_N < y\} &= \frac{\lambda^N}{(\lambda - s)^N} \left[e^{-sx} - \sum_{i=0}^{N-1} \frac{e^{-\lambda x}(\lambda - s)^i x^i}{i!}\right].
\end{aligned} \tag{9}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Substituting the equations in (9) into (8), we obtain

$$w^*(N,s) = \int_0^\infty \left[\frac{\lambda^N}{(\lambda-s)^N} e^{-sx} - \sum_{i=0}^{N-1} \frac{\lambda^N x^i}{i!(\lambda-s)^{N-i}} e^{-\lambda x} + \sum_{i=0}^{N-1} \frac{e^{-\lambda x}(\lambda x)^i}{i!} \right] dP\{\hat{W}(N) + \hat{V} < x\}. \quad (10)$$

By the definitions of $k^*(N,s), w^*(N,s), f^*(s)$ and $k_i(N)$, we obtain (6). •

An algorithm for computing $\{k_i(N)\}$ will be given in Section 6.

4. LSTS OF THE FLOW TIMES

In this section, we present the LSTs of the flow times W_{fl}, W_{al} and W_{aa} as defined in Section 2. Recall that the batch service time is defined by $V = V_s + V_{p1} + \ldots + V_{pN}$ so that $f^*(s) = f_s^*(s)[f_p^*(s)]^N$. Let us begin with W_{fl} (we have suppressed N if it is understood).

Theorem 4.1.

The LST of $W_{fl} = U_2 + \ldots + U_N + W(N) + V_s + V_{p1} + \ldots + V_{pN}$ *is given by*

$$\mathbf{E}\exp\{-sW_{fl}\} = f^*(s)\left\{ \left(\frac{\lambda^{N-1}}{(\lambda+s)^{N-1}} - \frac{\lambda}{\lambda-s} \right) k_0(N) + \frac{\lambda}{(\lambda-s)} k^*(N,s) + \sum_{i=0}^{N-2} \frac{1}{s^{i+1}} \left(\frac{\lambda^N}{(\lambda+s)^{N-1-i}} - \lambda^{i+1} \right) \left[k_0(N) - \sum_{j=0}^{i} \frac{(-s)^j}{j!} \left(\frac{d^j k^*(N,\hat{s})}{d\hat{s}^j}\Big|_{\hat{s}=s+\lambda} \right) \right] \right\}. \quad (11)$$

Proofs for this, and all subsequent theorems, appear in Section 8.

Theorem 4.2.

The LST of $W_{al} = U_{J+1} + \ldots + U_N + W(N) + V_s + V_{p1} + \ldots + V_{pN}$ *is:*

$$\mathbf{E}\exp\{-sW_{al}\} = \frac{f^*(s)}{N}\left\{ \sum_{i=0}^{N-1} \left[\sum_{j=i+1}^{N} \left(\frac{\lambda^{N-j+i}}{(\lambda+s)^{N-j}} - \frac{\lambda^j}{(\lambda-s)^{j-i}} \right) \right] k_i(N) + \left(\sum_{j=1}^{N} \frac{\lambda^j}{(\lambda-s)^j} \right) k^*(N,s) + \frac{1}{s} \sum_{i=0}^{N-2} \left(\frac{\lambda^N}{(\lambda+s)^{N-i-1}} - \lambda^{i+1} \right) \left[k_i(N) - \frac{(-1)^i}{i!} \frac{d^i k^*(N,\hat{s})}{d\hat{s}^i}\Big|_{\hat{s}=s+\lambda} \right] \right\}. \quad (12)$$

Theorem 4.3.

The LST of $W_{aa} = U_{J+1} + \ldots + U_N + W(N) + V_s + V_{p1} + \ldots + V_{pJ}$ *is:*

$$\mathbf{E}\exp\{-sW_{aa}\} = \frac{f_s^*(s)}{N}\left\{ \sum_{j=1}^{N} \frac{(\lambda f_p^*(s))^j}{(\lambda-s)^j} k^*(N,s) + \sum_{i=0}^{N-1} k_i(N) \sum_{j=i+1}^{N} (f_p^*(s))^j \left(\frac{\lambda^{N-j+i}}{(\lambda+s)^{N-j}} - \frac{\lambda^j}{(\lambda-s)^{j-i}} \right) \right.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$+\sum_{i=0}^{N-2} k_i(N) \sum_{j=i}^{N-2} \frac{[1-f_p^*(s)]^{j-i}[f_p^*(s)]^{i+1}}{s^{j+1-i}} \left(\frac{\lambda^N}{(\lambda+s)^{N-j-1}} - \lambda^{j+1} \right)$$

$$-f_p^*(s) \sum_{i=0}^{N-2} \frac{(-1)^i}{i!} \left[\sum_{j=i}^{N-2} \frac{[1-f_p^*(s)]^{j-i}}{s^{j+1-i}} \left(\frac{\lambda^N}{(\lambda+s)^{N-j-1}} - \lambda^{j+1} \right) \right]$$

$$\left. \frac{d^i k^*(N,\hat{s})}{d\hat{s}^i} \right|_{\hat{s}=s+\lambda} \Bigg\}. \tag{13}$$

The LSTs (11), (12), and (13), while complex, lead to computationally tractable methods of obtaining moments of the flow time measures. In Section 5, we use these LSTs to obtain the first two moments of the flow time distributions.

5. FIRST AND SECOND MOMENTS OF FLOW TIMES

The mean of the three flow times presented earlier can be obtained in terms of the mean interarrival time, batch size, mean batch waiting time, mean setup time and mean processing time. By definitions (2), (3), and (4), the mean flow times can be obtained in a straightforward fashion:

$$\begin{aligned} \mathbf{E}W_{fl}(N) &= \frac{N-1}{\lambda} + \mathbf{E}W(N) + \mathbf{E}V_s + N\mathbf{E}V_{p1}; \\ \mathbf{E}W_{al}(N) &= \frac{N-1}{2\lambda} + \mathbf{E}W(N) + \mathbf{E}V_s + N\mathbf{E}V_{p1}; \\ \mathbf{E}W_{aa}(N) &= \frac{N-1}{2\lambda} + \mathbf{E}W(N) + \mathbf{E}V_s + \frac{(N+1)}{2}\mathbf{E}V_{p1}. \end{aligned} \tag{14}$$

By equation (5.2.35) and (5.2.36), Theorem 5.2.5, in Neuts [8], the following formula for the mean waiting time is obtained:

$$\mathbf{E}W(N) = \frac{1}{\lambda[1-\rho(N)]} \sum_{j=1}^{N} (j-1)y_{0j} + \frac{1}{2[1-\rho(N)]} \left[\rho(N) \frac{(\sigma^2+\mu^2)}{\mu} - \frac{N-1}{\lambda} \right]. \tag{15}$$

where y_{0j} is the probability that at an arbitrary time there is no formed batch in the system and the number of customers in the system is $j-1$, $1 \le j \le N$. An algorithm for computing $y_0 = (y_{01}, \ldots, y_{0N})$ is given in Section 6.

We note that the computation of the mean flow times does not actually require the LSTs in equations (11), (12) and (13) as the results can be obtained with (15) and some straightforward manipulations. In contrast, the LSTs *are* needed to obtain the second moments (and hence the variance) of flow times $W_{fl}(N), W_{al}(N)$ and $W_{aa}(N)$.

The second moments of flow times will be presented in terms of $\{k_i(N)\}$, $\mathbf{E}V$, $\mathbf{E}V^2$, $\mathbf{E}V^3$, $\mathbf{E}W(N)$, and $\mathbf{E}(W(N))^2$. We assume that $\mathbf{E}V^3$ is finite (implying that $\mathbf{E}V$ and $\mathbf{E}V^2$ are finite). We begin by finding $\mathbf{E}W(N))^2$ from (6) by differentiating three times, taking limits as $s \to 0$ and simplifying the resulting equation:

$$\begin{aligned} \mathbf{E}(W(N))^2 &= \frac{1}{[1-\rho(N)]} \Bigg\{ \sum_{j=1}^{N} \frac{(N-j+1)(N-j+2)}{\lambda^2} y_{0j} - \frac{(N+1)(N+2)}{3\lambda^2} + \frac{\lambda}{3} \frac{\mathbf{E}V^3}{N} \\ &\quad + \mathbf{E}W(N) \left[\frac{\lambda}{N}\mathbf{E}V^2 - 2\mathbf{E}V + \frac{N+1}{\lambda} \right] - \mathbf{E}V^2 + \frac{(N+1)}{\lambda}\mathbf{E}V \Bigg\}. \end{aligned} \tag{16}$$

where we used the relationship between $k_i(N)$ and y_{0i} given by (22).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1. The second moment of W_{fl} :

Since the service time $V = V_s + V_{p1} + ...V_{pN}$ of a batch is independent of the interarrival times $U_i, 1 \leq i \leq N$ and $W(N)$ of that batch, we have

$$\mathrm{Var}(W_{fl}) = \mathrm{Var}(U_2 + \ldots + U_N + W(N)) + \mathrm{Var}(V). \tag{17}$$

Therefore, we only need to find the second moment of $U_2 + ... + U_N + W(N)$.

Theorem 5.1.

$$\begin{aligned}\mathbf{E}(U_2 + \ldots + U_N + W(N))^2 &= \mathbf{E}(W(N))^2 + 2\mathbf{E}W(N)\mathbf{E}V + \mathbf{E}V^2 - \frac{2}{\lambda}(\mathbf{E}W(N) + \mathbf{E}V) \\ &\quad + \frac{2}{\lambda^2} + \frac{(N^2 - N - 2)}{\lambda^2}k_0(N) + \sum_{i=0}^{N-2}(N - i - 1)\lambda^i \\ &\quad \left[2(i+1)k_{i+2}(N) + \frac{(N-i)k_{i+1}(N)}{\lambda}\right].\end{aligned} \tag{18}$$

2. The second moment of W_{al} :

We have

$$\mathrm{Var}(W_{al}) = \mathrm{Var}(U_{J+1} + \ldots + U_N + W(N)) + \mathrm{Var}(V). \tag{19}$$

Therefore we only need to find the second moment of $U_{J+1} + ... + U_N + W(N)$.

Theorem 5.2.

$$\begin{aligned}\mathbf{E}(U_{J+1} + \ldots + U_N + W(N))^2 &= \mathbf{E}(W(N))^2 + 2\mathbf{E}W(N)\mathbf{E}V + \mathbf{E}V^2 \\ &\quad - \frac{(N+1)}{\lambda}(\mathbf{E}W(N) + \mathbf{E}V) + \frac{(N+1)(N+2)}{3\lambda^2} \\ &\quad - \frac{1}{N}\sum_{i=0}^{N-1}\lambda^{i-2}(N-i)(N-i+1)k_i(N) \\ &\quad + \frac{1}{N}\sum_{i=0}^{N-2}\lambda^i(N-i-1)(i+1) \\ &\quad \left[\frac{(N-i)k_{i+1}(N)}{\lambda} + (i+2)k_{i+2}(N)\right].\end{aligned} \tag{20}$$

3. The second moment of W_{aa} :

For this flow time, the length of service, $V_s + V_{p1} + ... + V_{pJ}$, is not independent of $U_{J+1}, \ldots, U_N$, $W(N)$. Therefore, the result obtained is somewhat more complicated than the previous second moments:

Theorem 5.3.

$$\begin{aligned}\mathbf{E}(W_{aa})^2 &= \mathbf{E}(W(N))^2 + 2\mathbf{E}W(N)\mathbf{E}V + \mathbf{E}V^2 + (N+1)\left(\mathbf{E}V_p - \frac{1}{\lambda}\right)(\mathbf{E}W(N) + \mathbf{E}V) \\ &\quad + \frac{(N+1)(N+2)}{3\lambda^2} - \frac{(N+1)(2N+1)}{3\lambda}\mathbf{E}V_p + \frac{(N+1)}{2}\mathbf{E}V_p^2 + \frac{(N^2-1)}{3}(\mathbf{E}V_p)^2 \\ &\quad + \frac{1}{N}\sum_{i=0}^{N-1}\lambda^{i-1}(N-i)\left[(N+i+1)(N-i)\mathbf{E}V_p - \frac{N-i+1}{\lambda}\right]k_i(N) \\ &\quad + \frac{1}{N}\sum_{i=0}^{N-2}\lambda^i(N-i-1)(i+1) \\ &\quad \left[\left(\frac{N-i}{\lambda} + (i+2)\mathbf{E}V_p\right)k_{i+1}(N) + (i+2)k_{i+2}(N)\right] \\ &\quad + 2(\mathbf{E}W_{aa} - \mathbf{E}V_s)\mathbf{E}V_s + \mathbf{E}V_s^2.\end{aligned} \tag{21}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

6. COMPUTATIONAL PROCEDURES FOR DETERMINING THE OPTIMAL BATCH SIZE

In this section, we outline the computational procedure for determining the mean and variance of flow time as a function of the batch size. We also explore the batch size that minimizes either the flow time mean or variance. Much of the computational workload lies in the computation of the constants $\{k_i(N),\ 1 \le i \le N\}$, or equivalently the vector $\mathbf{y}_0$. The constants $\{k_i(N),\ 1 \le i \le N\}$ and $\mathbf{y}_0$ are the focus of this section. An efficient algorithm for computing $\mathbf{y}_0$ is proposed based on the matrix analytic method.

By (7), we see that $k_i(N)$ is the probability that $i-1$ customers arrived during the waiting and service period of an arbitrary batch in the $E_N/G/1$ queue. This means that $\mathbf{k} = (k_0(N), \ldots, k_{N-1}(N))$ is the probability vector that there is no batch in the queueing system. With this observation, there is a simple relationship between $\mathbf{k}$ and $\mathbf{y}_0$ (see Lucantoni [7]) : $\mathbf{k} = -N\mathbf{y}_0 D_0/\lambda$ (D_0 is defined in (25)), *i.e.*,

$$k_i(N) = N(y_{0(i+1)} - y_{0i}), \quad 0 \le i \le N-1. \tag{22}$$

As for $k_N(N)$, we can show (see Section 8) that:

$$k_N(N) = N\left(\frac{y_{01}}{f^*(\lambda)} - y_{0N}\right). \tag{23}$$

An algorithm to compute $\mathbf{y}_0$ for the $BMAP/G/1$ queue can be found in Lucantoni [7]. However, that algorithm involves matrices of dimension N, which makes implementation difficult for large batch sizes. In what follows, we simplify the algorithm in [7] for the $E_N/G/1$ queue so that it involves only vectors of dimension N. A major advantage of this modified algorithm is that it is possible for the program to run efficiently for large N.

Let G (an $N \times N$ matrix) be the minimal nonnegative solution to the equation

$$G = \int_1^\infty \exp\{x(D_0 + D_1 G)\}\, dF(x). \tag{24}$$

where

$$D_0 = \begin{pmatrix} -\lambda & \lambda & & \\ & \ddots & \ddots & \\ & & \ddots & \lambda \\ & & & -\lambda \end{pmatrix} \quad \text{and} \quad D_1 = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & \ddots & \\ \lambda & & & 0 \end{pmatrix} \tag{25}$$

are $N \times N$ matrices, the matrix representation of the input Erlang process of batches. Let g_{ij} be the $(i,j)^{\text{th}}$ element of matrix G and vector $\mathbf{g}_1 = (g_{11}, \ldots, g_{1N})$ (see Neuts [8] and Lucantoni [7] for more on G and equation (24)).

Property 1.

For the $E_N/G/1$ queue:

$$y_{0j} = [1 - \rho(N)]\frac{\sum_{i=1}^{j} g_{1i}}{\sum_{i=1}^{N}(N-i+1)g_{1i}}, \quad 1 \le j \le N. \tag{26}$$

Based on (24), and the special structure of the input process of the $E_N/G/1$ queue, we propose the following modified algorithm. Let $\theta = \lambda + 1$ and vector $\mathbf{g}_1$ be the first row of matrix G. Let $\mathbf{g}_1[0] = 0, \mathbf{h}_{0,k} = (1, 0, \ldots 0)$ and

$$\begin{aligned} \mathbf{h}_{n+1,k} &= \mathbf{h}_{n,k} + \frac{\lambda}{\theta}\left[-\mathbf{h}_{n,k} + \phi(\mathbf{h}_{n,k}) + (\mathbf{h}_{n,k})_N \mathbf{g}_1[k]\right]; \\ \mathbf{g}_1[k+1] &= \sum_{n=0}^{\infty} \gamma_n \mathbf{h}_{n,k}, \end{aligned} \tag{27}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where $\phi(\mathbf{h}_{n,k}) = (0, (\mathbf{h}_{n,k})_1, \ldots, (\mathbf{h}_{n,k})_{N-1})$ and

$$\gamma_n = \int_0^\infty \frac{e^{-\theta x}(\theta x)^n}{n!}\, dF(x), \quad 0 \le n \le \infty. \tag{28}$$

Then $\mathbf{g}_1 = \lim \mathbf{g}_1[k]$ and $\mathbf{y}_0$ can be calculated with (26). We note that this algorithm for our problem is much more efficient than its original form given in Lucantoni [6].

Once $\mathbf{y}_0$ is determined, the first moments, $\mathbf{E}W(N)$, $\mathbf{E}W_{fl}$, $\mathbf{E}W_{al}$ and $\mathbf{E}W_{aa}$, can be calculated with simple formulas derived from (2), (3), (4) and (15). For the second moments or variances, we first use (22) and (23) to calculate $\{k_i(N),\ 0 \le i \le N\}$. Then (16) and Theorems 5.1, 5.2 and 5.3 are used to compute $\mathbf{E}(W(N))^2$, $\mathbf{E}W_{fl}^2$, $\mathbf{E}W_{al}^2$ and $\mathbf{E}W_{aa}^2$. The corresponding variances can be easily found with the formula $\mathrm{Var}(X) = \mathbf{E}X^2 - (\mathbf{E}X)^2$.

To illustrate the algorithm for computing the variances, we use $\mathrm{Var}(W_{aa})$:

1. Compute $\mathbf{y}_0$.
2. Compute $\{k_i(N),\ 0 \le i \le N\}$ by (22)and (23).
3. Compute $\mathbf{E}W(N)$, $\mathbf{E}W_{aa}(N)$ and $\mathbf{E}(W(N))^2$ with (16).
4. Compute $\mathbf{E}(W_{aa}(N))^2$ with (21) and $\mathrm{Var}(W_{aa}(N)) = \mathbf{E}(W_{aa}(N))^2 - (\mathbf{E}W_{aa}(N))^2$

The above procedure is used to do computations for a single batch size N. To find the batch size which minimizes the mean or the variance of $W_{aa}(N)$, we begin with the smallest batch size which yields an ergodic system. The procedure is repeated for larger and larger N until the optimal batch size is determined. The following properties are used to stop searching.

Lemma 6.1.

$$\begin{aligned}
\mathbf{E}W_{aa}(N) &\ge \frac{N-1}{2\lambda} + \mathbf{E}V_s + \frac{(N+1)}{2}\mathbf{E}V_p; \\
\mathrm{Var}(W_{aa}(N)) &= \mathrm{Var}(V_s) + \mathrm{Var}(W(N)) + \mathrm{Var}(V_p) + (N-1)\left[\frac{1}{2}\left(\frac{1}{\lambda} + \mathbf{E}V_p\right) - \mathbf{E}W(N)\right] \\
&\quad + \frac{(N^2-1)}{12}\left(\frac{1}{\lambda} - \mathbf{E}V_p\right)^2 + \frac{2}{N}\sum_{j=1}^{N}\mathbf{E}[W(N)(U_{j+1} + \ldots + U_N)].
\end{aligned} \tag{29}$$

The right hand side of the first equation in (29) is an increasing function of N. Therefore, the search should stop when the right hand side of (29) exceeds the mean flow time already calculated for a smaller batch size.

As for the variance, it can be proved that $\mathbf{E}W(N) \to 0$ when $N \to \infty$. In addition, numerical results show that $\mathbf{E}W(N)$ decreases monotonically in N. Then we have

$$\mathrm{Var}(W_{aa}(N)) \ge \frac{(N^2-1)}{12}\left(\frac{1}{\lambda} - \mathbf{E}V_p\right)^2, \tag{30}$$

when $\mathbf{E}W(N) < (1/\lambda + \mathbf{E}V_p)/2$. The search is stopped at a batch size where the right hand side of (30) is larger than the minimal variance computed so far.

In summary, this section presented formulas for the mean and variance of each flow time and an algorithm for computing the optimal batch size.

7. NUMERICAL RESULTS

In this section, we report the numerical results for four examples. We shall mainly concentrate on the comparison of the batch size that minimizes the mean flow time as compared to that which minimizes the variance of flow time. For brevity, we only present the results for $W_{aa}(N)$. The results for $W_{fl}(N)$ and $W_{al}(N)$ are similar in nature.

In the examples, the setup time and processing time are either deterministic (C) or exponential (E). We shall use (C, C), (C, E), (E, C) and (E, E) to represent the type of setup time and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

	(C,C)	(C,E)	(E,C)	(E,E)
N for $\mathbf{E}W_{aa}$	2	2	2	2
N for $\mathrm{Var}(W_{aa})$	3	4	3	6

Table 1: The Optimal Batch Sizes for Example 1

	(C,C)	(C,E)	(E,C)	(E,E)
N for $\mathbf{E}W_{aa}$	6	7	7	7
N for $\mathrm{Var}(W_{aa})$	10	12	11	16

Table 2: The Optimal Batch Sizes for Example 2

processing time. For instance, (C,E) means that the setup time is constant and the processing time is exponential.

Example 1. Let $\lambda = 0.7$, $\mu_s = 0.2$ and $\mu_p = 1$. The setup time is short but batching still makes sense to reduce the mean and variance of flow time. The batch sizes which minimize each of the mean and variance (respectively) of $W_{aa}(N)$ are shown in Table 1.

Table 1 shows that the mean and variance of $W_{aa}(N)$ are minimized at different batch sizes. This raises a question of how to choose the best batch size. We will discuss this more after showing all four examples.

Example 2. Let $\lambda = 0.7$, $\mu_s = 1.5$ and $\mu_p = 1$. The setup time is comparable to the processing time. The batch sizes which minimize the mean and variance of $W_{aa}(N)$ are shown in Table 2.

Example 3. Let $\lambda = 0.7$, $\mu_s = 20$ and $\mu_p = 1$. The setup time is much longer than the processing time. In this case, we see a big difference brought in by the variability of the setup time. The variance is minimized at $N = 71$ with a deterministic setup time and exponential processing time but it is minimized at $N = 113$ with an exponential setup time and deterministic processing time. Intuitively, the setup time has great influence on the flow time, since it is much longer than processing time. Then the variance of setup time dominates the selection of the optimal batch size. Therefore, when an exponential setup time replaces a deterministic setup time, there should be a big change in the optimal batch size. In fact, this change turns out to be a sharp increase in the optimal batch size.

Example 4. Let $\lambda = 0.3$, $\mu_s = 20$ and $\mu_p = 1$. In comparison to Example 3, the traffic intensity is less for a given N. The mean, second moment and variance are all minimized at smaller batch sizes. However, we still see the effect of batching and the influence of the variability of the setup time on the optimal solution.

Generally, increasing the setup time increases the variability of the flow times (see Figures 2 and 3). Also, increasing the setup time increases the difference between the batch sizes which minimize the mean and variance of flow time, respectively.

	(C,C)	(C,E)	(E,C)	(E,E)
N for $\mathbf{E}W_{aa}$	56	58	68	69
N for $\mathrm{Var}(W_{aa})$	65	72	113	119

Table 3: The Optimal Batch Sizes for Example 3

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

	(C,C)	(C,E)	(E,C)	(E,E)
N for $\mathbf{E}W_{aa}$	11	12	15	15
N for $\mathrm{Var}(W_{aa})$	12	12	21	21

Table 4: The Optimal Batch Sizes for Example 4

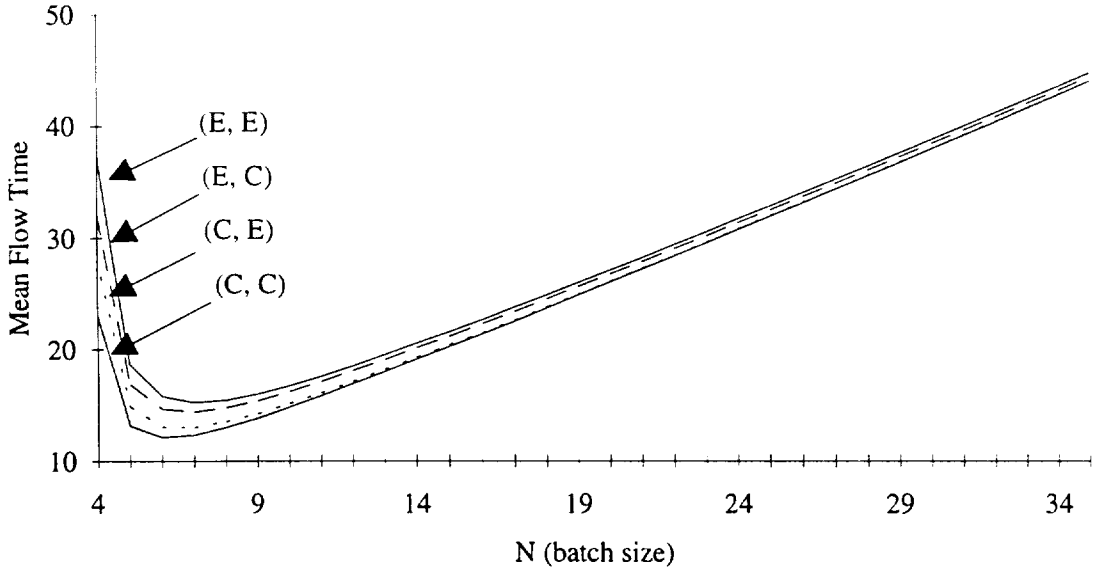

Figure 2: The Mean Flow Time $\mathbf{E}W_{aa}$ (when $\lambda = 0.7$, $\mu_s = 1.5$ and $\mu_p = 1$)

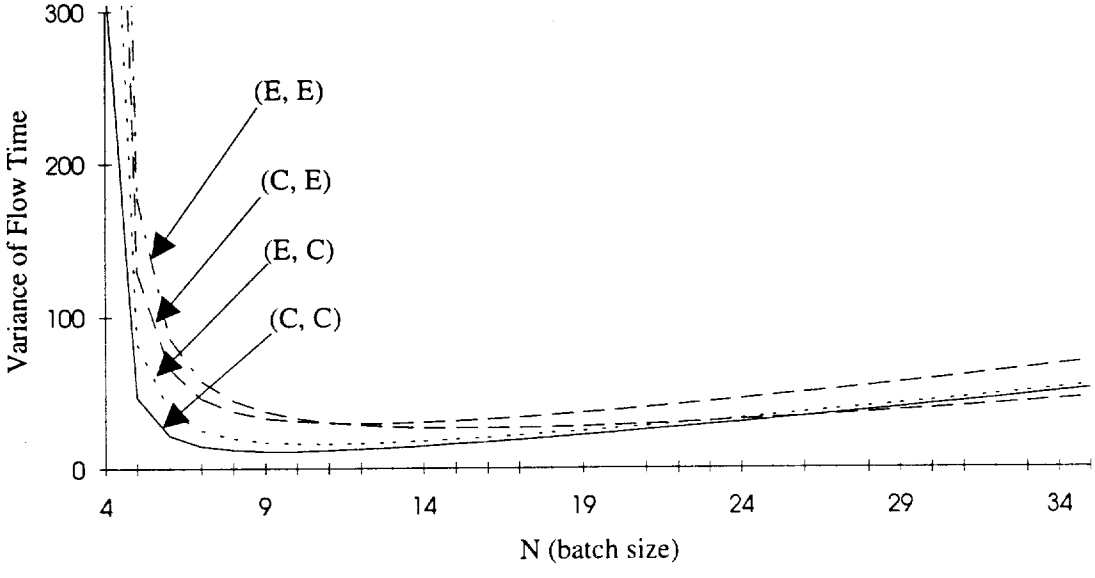

Figure 3: The Variance of the Flow Time $\mathrm{Var}W_{aa}$ (when $\lambda = 0.7$, $\mu_s = 1.5$ and $\mu_p = 1$)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Our computations also show that $\mathbf{E}W_{aa}(N)$ and $\mathrm{Var}(W_{aa}(N))$ generally are not minimized at the same batch size. Usually, the variance is minimized with a larger batch size (refer to Tables 1, 2, 3 and 4). As illustrated in Table 3, sometimes there is a large difference between the two optimal batch sizes. Therefore, tradeoffs must be made between mean and variance in selecting the batch size. In practice, the best batch size depends on the preferences of the decision maker. This may be formalized through the use of some objective function which includes both the mean and variance of the flow times.

This section concludes our analysis of the queueing system with setup times and batching. The four examples illustrate how the setup time and the processing time affect both the mean and variance of the flow time. Because the batch size that minimizes the mean flow time differs from that which minimizes the variance, a tradeoff may need to be made in selecting a batch size that somehow balances the two performance measures.

8. PROOFS

In order to prove Theorems 4.1, 4.2 and 4.3, we need the following result.

Lemma 8.1.

For $1 \le j \le N$,

$$\begin{aligned}
&\mathbf{E}\exp\{-s(U_{j+1}+\ldots+U_N+\max\{0,x-(U_1+\ldots+U_N)\})\} \\
&= \frac{\lambda^j}{(\lambda-s)^j}e^{-sx} + \sum_{i=0}^{j-1}\frac{\left[\frac{\lambda^{N-j+i}}{(\lambda+s)^{N-j}} - \frac{\lambda^j}{(\lambda-s)^{j-i}}\right]}{i!}x^i e^{-\lambda x} \\
&\quad + \sum_{i=0}^{N-j-1}\frac{1}{i!}\left(\frac{\lambda^{N-j}}{(\lambda+s)^{N-j-i}} - \lambda^i\right)e^{-\lambda x}\int_0^x \frac{e^{-s(x-y)}\lambda^j y^{j-1}(x-y)^i}{(j-1)!}\,dy. \qquad (31)
\end{aligned}$$

Proof.

First, since $U_1+\ldots+U_j$ and $U_{j+1}+\ldots+U_N$ are independent, we have

$$\begin{aligned}
&\mathbf{E}\exp\{-s(U_{j+1}+\ldots+U_N+\max\{0,x-(U_1+\ldots+U_N)\})\} \\
&= \int_0^x \mathbf{E}\exp\{-s(U_{j+1}+\ldots+U_N+\max\{0,x-y-U_{j+1}-\ldots-U_N\})\} \\
&\quad dP\{U_1+\ldots+U_j<y\} \\
&\quad + \int_x^\infty \mathbf{E}\exp\{-s(U_{j+1}+\ldots+U_N)\}\,dP\{U_1+\ldots+U_j<y\}. \qquad (32)
\end{aligned}$$

Since $U_1+\ldots+U_j$ has an Erlang Stage-j distribution, and $U_{j+1}+\ldots+U_N$ has an Erlang Stage-$(N-j)$ distribution, the last line in (32) is given by

$$\frac{\lambda^{N-j}}{(\lambda+s)^{N-j}}\sum_{i=0}^{j-1}\frac{e^{-\lambda x}(\lambda x)^i}{i!}. \qquad (33)$$

The first item on the right hand side of (32) can be further decomposed into:

$$\begin{aligned}
&\int_0^x\left[\int_0^{x-y} e^{-s(x-y)}\,dP\{U_{j+1}+\ldots+U_N<z\} + \int_{x-y}^\infty e^{-sz}\,dP\{U_{j+1}+\ldots+U_N<z\}\right] \\
&\quad dP\{U_1+\ldots+U_j<y\} \\
&= \int_0^x\left[e^{-s(x-y)}\left(1-\sum_{i=0}^{N-j-1}\frac{e^{-\lambda(x-y)}\lambda^i(x-y)^i}{i!}\right)\right.
\end{aligned}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$+\frac{\lambda^{N-j}}{(\lambda+s)^{N-j}}\left(\sum_{i=0}^{N-j-1}\frac{e^{-(s+\lambda)(x-y)}(\lambda+s)^i(x-y)^i}{i!}\right)\Bigg]\frac{e^{-\lambda y}\lambda^j y^{j-1}}{(j-1)!}\,dy. \tag{34}$$

Substituting (33) and (34) into (32), we obtain (31). •

Proof of Theorem 4.1.

Note that $\{U_1,\ldots,U_N,\ W(N)\}$ and $\{V_s, V_{p1},\ldots,V_{pN}\}$ are independent. By (2) and (5), applying Lemma 8.1 with $j=1$, we have

$$\begin{aligned}
\mathbf{E}\exp\{-sW_{fl}\} &= \mathbf{E}\exp\{-s[U_2+\ldots+U_N+W(N)]\}f^*(s)\\
&= f^*(s)\int_0^\infty \exp\{-s(U_2+\ldots+U_N+\max\{0,x-U_1-\ldots-U_N\})\}\\
&\quad dP\{\hat{W}(N)+\hat{V}<x\}\\
&= f^*(s)\Bigg[\frac{\lambda}{(\lambda-s)}k^*(N,s)+\left(\left(\frac{\lambda}{\lambda+s}\right)^{N-1}-\frac{\lambda}{\lambda-s}\right)k_0(N)\\
&\quad+\sum_{i=0}^{N-2}\left(\frac{\lambda^{N-1}}{(\lambda+s)^{N-i-1}}-\lambda^i\right)\int_0^\infty e^{-\lambda x}\\
&\quad\int_0^x e^{-s(x-y)}\frac{\lambda(x-y)^i}{i!}\,dy\,dP\{\hat{W}(N)+\hat{V}<x\}\Bigg].
\end{aligned}\tag{35}$$

This leads to (11). •

Proof of Theorem 4.2.

Since J is uniformly distributed on $\{1,\ldots,N\}$, we have

$$\mathbf{E}\exp\{-sW_{al}\}=\sum_{j=1}^{N}\frac{1}{N}\mathbf{E}\exp\{-s[U_{j+1}+\ldots+U_N+W(N)]\}f^*(s). \tag{36}$$

By Lemma 8.1 and the following simplification

$$\begin{aligned}
&\sum_{j=1}^{N}\sum_{i=0}^{N-j-1}\frac{1}{i!}\left(\frac{\lambda^{N-j}}{(\lambda+s)^{N-j-i}}-\lambda^i\right)e^{-\lambda x}\int_0^x\frac{e^{-s(x-y)}\lambda^j y^{j-1}(x-y)^i}{(j-1)!}\,dy\\
&=\sum_{m=0}^{N-2}e^{-\lambda x}\int_0^x\left[\sum_{i=0}^{m}\left(\frac{\lambda^N}{(\lambda+s)^{N-m-1}}-\lambda^{m+1}\right)\frac{y^{m-i}(x-y)^i}{i!(m-i)!}\right]e^{-s(x-y)}\,dy\\
&=\sum_{m=0}^{N-2}e^{-\lambda x}\int_0^x\left(\frac{\lambda^N}{(\lambda+s)^{N-m-1}}-\lambda^{m+1}\right)\frac{x^m}{m!}e^{-s(x-y)}\,dy,
\end{aligned}\tag{37}$$

we obtain

$$\begin{aligned}
\mathbf{E}\exp\{-sW_{al}\} &= \frac{f^*(s)}{N}\Bigg[\sum_{j=1}^{N}\frac{\lambda^j}{(\lambda-s)^j}k^*(N,s)+\sum_{i=0}^{N-1}\sum_{j=i+1}^{N}\left(\frac{\lambda^{N-j+i}}{(\lambda+s)^{N-j}}-\frac{\lambda^j}{(\lambda-s)^{j-i}}\right)k_i(N)\\
&\quad+\sum_{i=0}^{N-2}\frac{1}{i!}\left(\frac{\lambda^N}{(\lambda+s)^{N-i-1}}-\lambda^{i+1}\right)\\
&\quad\int_0^\infty x^i e^{-(\lambda+s)x}\int_0^x e^{sy}\,dy\,dP\{W(N)+V<x\}\Bigg],
\end{aligned}\tag{38}$$

which leads to (12). •

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Proof of Theorem 4.3.

By definition, we have

$$\mathbf{E}\exp\{-sW_{aa}\} = \sum_{j=1}^{N} \frac{1}{N} \mathbf{E}\exp\{-s[U_{j+1} + \ldots + U_N + W(N)]\} f_s^*(s)(f_p^*(s))^j. \tag{39}$$

By Lemma 8.1 and simplification, we have

$$\begin{aligned}
&\mathbf{E}\exp\{-sW_{aa}\} \\
&= \frac{f_s^*(s)}{N}\left\{\sum_{j=1}^{N} \frac{(\lambda f_p^*(s))^j}{(\lambda - s)^j} k^*(N,s) + \sum_{i=0}^{N-1}\left[\sum_{j=i+1}^{N} [f_p^*(s)]^j \left(\frac{\lambda^{N-j+i}}{(\lambda+s)^{N-j}} - \frac{\lambda^j}{(\lambda-s)^{j-i}}\right)\right] k_i(N)\right. \\
&\quad + \sum_{i=0}^{N-2} \frac{f_p^*(s)}{i!}\left(\frac{\lambda^N}{(\lambda+s)^{N-i-1}} - \lambda^{i+1}\right)\int_0^\infty e^{-\lambda x} \\
&\quad \left.\int_0^x e^{-s(x-y)}[x - y\{1 - f_p^*(s)\}]^i \, dy \, dP\{W(N) + V < x\}\right\}.
\end{aligned} \tag{40}$$

Here we use

$$\begin{aligned}
&\int_0^x e^{-s(x-y)}[x - y\{1 - f^*(s)\}]^i \, dy \\
&= \exp\left\{\frac{s f_p^*(s)}{1 - f_p^*(s)x}\right\}\int_0^x \exp\left\{\frac{-s}{1 - f_p^*(s)}\{x - y[1 - f_p^*(s)]\}\right\}[x - y(1 - f_p^*(s))]^i \, dy \\
&= \exp\left\{\frac{s f_p^*(s)}{1 - f_p^*(s)}s\right\}\int_{x f_p^*(s)}^x \exp\left\{-\frac{s}{1 - f_p^*(s)}y\right\} y^i \, dy \frac{1}{1 - f_p^*(s)} \\
&= \frac{i![1 - f_p^*(s)]^i}{s^{i+1}} \sum_{j=0}^{i} \frac{s^j x^j}{j![1 - f_p^*(s)]^j}[f_p^*(s)^j - e^{-sx}].
\end{aligned} \tag{41}$$

Equations (40) and (41) lead to (13). •

Proof of Theorem 5.1.

Differentiate the last expression in (35) (without $f^*(s)$) twice with respect to s and let $s \to 0$ in the resulting expression. (18) is obtained by simple algebra and

$$\begin{aligned}
\left.\frac{dk^*(N,s)}{ds}\right|_{s=0} &= -(\mathbf{E}W(N) + \mathbf{E}V); \\
\left.\frac{d^2k^*(N,s)}{ds^2}\right|_{s=0} &= \mathbf{E}(W(N))^2 + 2\mathbf{E}W(N)\mathbf{E}V + \mathbf{E}V^2. \quad \bullet
\end{aligned} \tag{42}$$

Proof of Theorem 5.2.

The proof is similar to Theorem 5.1 but we differentiate the *right* hand side of (38) (without $f^*(s)$). •

Proof of Theorem 5.3.

Since V_p is independent from $U_J, \ldots, U_N$, $W(N)$, $V_1, \ldots, V_J$, we have

$$\mathbf{E}(W_{aa})^2 = \mathbf{E}(W_{aa} - V_s)^2 + 2\mathbf{E}(W_{aa} - V_s)\mathbf{E}V_s + \mathbf{E}V_s^2. \tag{43}$$

The LST of $W_{aa} - V_s$ is given by the right hand side of (13) without $f_s^*(s)$. Differentiating the second and third lines of (40) (without $f_s^*(s)$) twice with respect to s and noticing that

$$\frac{d}{ds}\left[\int_0^\infty f_p^*(s)\int_0^x e^{-s(x-y)}\frac{[x - y(1 - f_p^*(s))]^i}{i!}\, dy\, e^{-\lambda x}\, dP\{\hat{W}(N) + \hat{V} < x\}\right]\Bigg|_{s=0}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$= \int_0^\infty \left[-\mathbf{E}V_p \int_0^x \frac{x^i}{i!}\, dy - \int_0^x \frac{x^i(x-y)}{i!}\, dy - i\mathbf{E}V_p \int_0^x \frac{x^{i-1}y}{i!}\, dy \right] e^{-\lambda x}$$
$$dP\{\hat{W}(N) + \hat{V} < x\}$$
$$= -\left(\frac{i}{2} + 1\right)(i+1)\mathbf{E}V_p k_{i+1}(N) - \frac{(i+1)(i+2)}{2} k_{i+2}(N), \tag{44}$$

we obtain

$$\begin{aligned}
\mathbf{E}(W_{aa} - V_s)^2 &= \mathbf{E}(W(N))^2 + 2\mathbf{E}W(N)\mathbf{E}V + \mathbf{E}V^2 + \frac{2\sum_{j=1}^N j}{N}\left(\mathbf{E}V_p - \frac{1}{\lambda}\right)(\mathbf{E}W(N) + \mathbf{E}V) \\
&\quad + \sum_{j=1}^N \frac{j(j+1)}{N\lambda^2} - 2\sum_{j=1}^N \frac{j^2}{N\lambda}\mathbf{E}V_p + \left(\sum_{j=1}^N j\right)\mathbf{E}V_p^2 + \sum_{j=1}^N \frac{j(j-1)}{N}(\mathbf{E}V_p)^2 \\
&\quad + \frac{1}{N}\sum_{j=1}^N \left[\sum_{i=0}^{j-1} \lambda^{i-1} \left(\frac{(N-j)(N-j+1) - (j-i)(j-i+1)}{\lambda} \right.\right. \\
&\quad \left.\left. + 2j(N-i)\mathbf{E}V_p \right) k_i(N) \right] \\
&\quad + \frac{1}{N}\sum_{i=0}^{N-2} \lambda^{i-1}(N-i-1)(i+1)\{(N-i)k_{i+1}(N) + \lambda(i+2) \\
&\quad [\mathbf{E}V_p k_{i+1}(N) + k_{i+2}(N)]\}.
\end{aligned} \tag{45}$$

The rest of the proof consists of routine simplifications. •

Proof of equation (23).

By (6), we have

$$k^*(N,s) = \frac{\left[\lambda^N k^*(N,s) + \sum_{i=0}^{N-1} [\lambda^i(\lambda - s)^N - \lambda^N(\lambda - s)^i] k_i(N)\right]}{(\lambda - s)^N} f^*(s). \tag{46}$$

Letting $s \to \lambda$ on both sides of the equation and applying l'Hospitals rule N times (it can be verified that this is appropriate), we obtain

$$\begin{aligned}
k_0(N) = \lim_{s\to\lambda} k^*(N,s) &= \frac{\left[\lambda^N(-1)^N N! k_N(N) + \sum_{i=0}^{N-1} \lambda^i k_i(N)(-1)^N N!\right]}{(-1)^N N!} f^*(\lambda) \\
&= \left[\lambda^N k_N(N) + \sum_{i=0}^{N-1} \lambda^i k_i(N)\right] f^*(\lambda).
\end{aligned} \tag{47}$$

Rearranging this equation gives $k_N(N)$. •

Proof of Property 1.

Corollary 1 and (55) in Lucantoni [7] shows that $\mathbf{y}_0(D_0 + D_1 G) = 0$. Therefore, we have $\mathbf{y}_0 = -\mathbf{y}_0 D_1 G D_0^{-1}$. The inverse matrix of D_0 is given by

$$D_0^{-1} = -\frac{1}{\lambda}\begin{pmatrix} 1 & 1 & \dots & 1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & 1 \\ & & & 1 \end{pmatrix}. \tag{48}$$

Then (26) is obtained by simple calculations with (9), (10) and $\mathbf{y}_0\mathbf{e} = 1 - \rho(N)$. •

Proof of Lemma 6.1.

The proof is completed by definitions and simple calculation. •

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

REFERENCES

1. Chaudhry, M. L., and J. G. C. Templeton (1983), *A First Course in Bulk Queues*, Wiley, New York.
2. Chaudhry, M.L., Agarwal, M., and Templeton, J.G.C. (1992), Exact and approximate numerical solutions of steady-state distributions arising in the queue GI/G/1, *Queueing Systems,* **10**, 105-152.
3. Cohen, J. W. (1982), *The Single Server Queue*, North-Holland.
4. Fabens, A. (1961), The Solution of Queueing and Inventory Models by semi-Markov Process, *J. Roy. Statist. Soc.*, Series B, **23**, 113-117.
5. Fukuta, J. (1969), Concentrated service queue with limited unit source, *J. Oper. Res. Soc.* (Jpn.), **12**, 21-35.
6. Karmarkar, U. S. (1987), Lot sizes, lead times and in-process inventory, *Management Sciences*, **33**, 395-408.
7. Lucantoni, D. M. (1991), New results on the single server queue with a batch Markovian arrival process, *Stochastic Models*, **7**, 1-46.
8. Neuts, M. F. (1989), *Structured Stochastic Matrices of M/G/1 type and Their Applications.* New York: Marcel Dekker Inc.
9. Prabhu, N. U. (1980), *Stochastic Storage Processes, Queues, Insurance Risk, and Dams*, Springer-Verlag.
10. Sumita, U. and M. Kijima (1986), On optimal bulk size of single-server bulk-arrival queueing systems with set-up times: numerical exploration via the Laguerre transform, *Selected Statistica Canadiana*, **VII**, 79-108.

Qi-Ming HE is a Ph.D candidate in the Department of Management Sciences, University of Waterloo. His research interests are the design and application of operations research models in production management systems and telecommunication systems. Much of his research is in the areas of algorithmic methods in applied probability, queueing theory, inventory control and production management. His work has been published (or to appear) in a number of journals, including *Advances in Applied Probability, Stochastic Models, SIAM Journal of Matrix Analysis and Applications, Journal of Asian-Pacific Operations Research,* and *Microelectronics and Reliability.*

Beth Jewkes is an Associate Professor in the Department of Management Sciences at the University of Waterloo. Her interests are generally in the area of stochastic models of planning and control of manufacturing systems, more specifically in quality modeling, feedback and priority queues, and in simulation methods. She has published articles in *Queuing Systems: Theory and Practice, IIE Transactions, Stochastic Models, Interfaces* and several other journals. Her refereeing activities include journals such as *INFORM, Management Science, Operations Research and IIE Transactions.*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.