Stochastic Processes and their Applications 33 (1989) 123-129
North-Holland

A LCFS FINITE BUFFER MODEL WITH BATCH INPUT AND NON-EXPONENTIAL SERVICES

Nico M. van DIJK

Free University, 1007 MC Amsterdam, The Netherlands

Received 16 March 1988
Revised 23 February 1989

A finite last-come first-served queueing system is studied with batch input and non-exponential services. A closed form expression is obtained for the steady-state queue length distribution and shown to be insensitive to service distributional forms (i.e. to depend only on mean service times). This result is of both practical and theoretical interest as an extension of the standard exponential case.

batch input * LCFS * insensitivity * batch balance

1. Introduction

Finite queueing systems are extensively studied for applications in telecommunication, computer performance evaluation and manufacturing. Generally the assumption is made that jobs arrive one at a time. In various present-day applications, however, jobs will arrive in batches of more than one job. For instance in satellite communication or packet switching networks a message may consist of various signals to be transmitted, in computer programming a program may initiate a number of modules to be run and in manufacturing parts to be processed upon may be transported grouped on pallets.

For non-exponential batch arrival systems with infinite capacity both exact results in terms of generating functions (cf. Burke, 1975; Cohen, 1976) and asymptotic expansions (cf. Van Ommeren, 1987) have been reported. For the special Poisson arrival case and Erlangian services also efficient computational methods have been developed (cf. Chaudhry and Templeton, 1983; Tijms, 1988). For systems with finite capacity, however, explicit results have been obtained for the case of Poisson input and exponential services only (cf. Kabak, 1970; Mansfield and Tran-Gia, 1982; Chaudhry and Templeton, 1983; Takahashi and Katayama, 1985; Nobel, 1987). Recently, the Poisson input has been relaxed to non-exponential finite source input (cf. Van Dijk, 1987) under the assumption of last-come first-served servicing. For the case of non-exponential services, however, no closed form expression has been reported in the literature. Clearly, such a result would be of practical interest as exponentiality assumptions are often far from realistic. Particularly insensitivity results (that is, independently of service distributional forms) are of interest, as these require knowledge of only mean services.

This paper deals with the non-exponential service case under the special last-come first-served queueing discipline. It is shown that the steady state queue length distribution has a closed form expression of a scaled geometric form and is insensitive to service distributional forms.

Though last-come first-served queueing disciplines are not the most common disciplines in practice, they do appear practical in some applications. For instance, stocks are often refilled but also worked off at the top. The result is also of theoretical interest for a twofold reason. (i) No insensitivity results at all have been reported for systems with batch input. (ii) The result does not fit in any of standard partial balance frameworks (see Remark 3.1), but as in Van Dijk (1987) is based upon a new notion of balance per batch.

Though this notion and the technique of this paper are much related to the latter reference, the inclusion of non-exponential services brings in essentially different complications in particular as batches are involved. For instance, while insensitivity is normally associated with a notion of balance per job (or lifetime), in the case of batches this balance fails as all jobs, except for one, have to wait for service (cf. Remark 3.1). The results of this paper therefore deserve special attention.

The organization is as follows. First, in Section 2, the model is described and the restriction to phase-type distributions is argued. Next, in Section 3 the steady-state expressions are derived and some remarks are made as to computation and finite source input.

2. Model

Consider a single-server facility with a storage constraint (buffer) for no more than N jobs, the one in service included. Batches of jobs arrive according to a Poisson process with parameter λ. A batch is of size k with probability $b(k)$, $k=1,2,\dots$. A batch of size k is accepted only when there are still k or more vacancies within the buffer. Otherwise the batch is completely rejected.

A last-come first-served preemptive resume service discipline is in order. That is, upon acceptance of a new batch the service of the batch presently in service is interrupted and the unit service speed is instantaneously allocated to this new batch. When all jobs of a batch are completed, the service of the last interrupted batch is resumed. Within a batch, jobs are served one at a time in some arbitrary but preassigned order.

Phase type restriction. The service requirement of a job is in principle allowed to be generally distributed with mean τ. For convenience of analysis, however, we will restrict the presentation to service distributions of the form

$$G = \sum_{k=1}^{\infty} a(k) E(k, \alpha) \tag{2.1}$$

where $E(k, \alpha)$ denotes an Erlang-k distribution with exponential parameter α and where $a(k)$ is the probability that the distribution consists of k successive exponential phases with parameter α. Hence,

$$\tau = \sum_{k=1}^{\infty} a(k)[k/\alpha], \tag{2.2}$$

while

$$u(r) = [\alpha\tau]^{-1} \sum_{k=r}^{\infty} a(k) \tag{2.3}$$

is known from renewal theory (cf. Kohlas, 1982, p. 47) as the "excess" probability of "r" residual exponential phases up to a next renewal in a renewal process with renewal distribution G.

The restriction to phase type distributions will justify a discrete Markovian analysis in the next section. It is well known, however, that any non-negative probability distribution can be approximated arbitrarily closely (in the sense of weak convergence) by distributions of the form (2.1) (cf. Hordijk and Schassberger, 1982). Based upon weak convergence limit theorems for the probability measures of the sample paths defined on so-called D-spaces (cf. Barbour, 1976; Whitt, 1980; Hordijk and Schassberger, 1982), the main result of this paper (Theorem 3.2) can therefore be extended to general service distributions G. The well-worn but technical details of these limiting steps are omitted. For details readers are referred to, among others, Whitt (1980) or Hordijk and Schassberger (1982).

3. Steady state distribution

For $n \leq N$, $k_1 + \cdots + k_n \leq N$ and $r_i > 0$, $i = 1, \ldots, n$, let the state vector

$$[\bar{k}_n, \bar{r}_n] = ((k_1; r_1), \ldots, (k_n; r_n))$$

denote that jobs are present from n different batches, with k_i jobs from batch i, the ith in order of arrival from the batches still present (thus the nth is the last entered batch), and where the job first to be completed within the ith batch still requires r_i exponential phases to be completed. By virtue of the exponential structure, the corresponding queueing process constitutes a continuous-time irreducible and aperiodic Markov chain with uniformly bounded jump rates. The existence of a unique steady state distribution is thus guaranteed (cf. Kohlas, 1982, p. 93). Throughout, a steady state distribution will be denoted by $\pi(\cdot)$ and is assumed to be zero for non-admissible states (i.e. with $k_1+\cdots+k_n>N$). The following theorem is the key result. To this end, for $k+l\leq N$ define

$$V(k\,|\,l)=\sum_{i=k}^{N-l} b(i). \tag{3.1}$$

Theorem 3.1. *With normalizing constant* $\pi(0)$, *we have*

$$\pi([\bar{k}_n, \bar{r}_n])=\pi(0)[\lambda\tau]^n \prod_{i=1}^{n} u(r_i)\, V(k_i\,|\,k_1+\cdots+k_{i-1}). \tag{3.2}$$

Proof. By virtue of the Markovian structure it suffices to verify the global balance equations (cf. Kohlas, 1982, p. 93) for any state $[\bar{k}_n, \bar{r}_n]$ while assuming (3.2). For notational convenience, for a vector $[\bar{k}_t, \bar{r}_t]=((k_1, r_1),\ldots,(k_t, r_t))$ let

$$[[\bar{k}_t, \bar{r}_t], (k, r)]=((k_1, s_1),\ldots,(k_t, s_t),(k, r)).$$

First, let $n>0$. Consider a state $[\bar{k}_n, \bar{r}_n]$ and assume $(k_n, r_n)=(k, r)$ to specify the last entered batch. The rate out of this state $[\bar{k}_n, \bar{r}_n]$, (where transitions from a state into itself due to a blocked arrival are included) is equal to

$$\pi([\bar{k}_n, \bar{r}_n])\{\alpha+\lambda\}. \tag{3.3}$$

The rate into state $[\bar{k}_n, \bar{r}_n]$ is given by

$$\begin{aligned}
&\pi([[\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r+1)])\alpha+\pi([\bar{k}_{n-1}, \bar{r}_{n-1}])\lambda b(k)a(r)\\
&\quad+\pi([[\bar{k}_{n-1}, \bar{r}_{n-1}], (k+1, 1)])\alpha a(r)\\
&\quad+\pi([[\bar{k}_n, \bar{r}_n], (1, 1)])\alpha+\pi([\bar{k}_n, \bar{r}_n])\lambda\left[\sum_{t=N-(k_1+\cdots+k_n)+1}^{\infty} b(t)\right]
\end{aligned} \tag{3.4}$$

where it is to be noted that the third term is 0 when $k_1+\cdots+k_{n-1}+k=N$. By assuming (3.2) one easily derives

$$\pi([[\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r+1)])=\pi([\bar{k}_n, \bar{r}_n])u(r+1)/u(r),$$

$$\pi([\bar{k}_{n-1}, \bar{r}_{n-1}])=\pi([\bar{k}_n, \bar{r}_n])/[\lambda\tau u(r)\, V(k\,|\,k_1+\cdots+k_{n-1})],$$

$$\begin{aligned}
&\pi([[\bar{k}_{n-1}, \bar{r}_{n-1}], (k+1, 1)])\\
&\quad=\pi([\bar{k}_n, \bar{r}_n])\, V(k+1\,|\,k_1+\cdots+k_{n-1})/ V(k\,|\,k_1+\cdots+k_{n-1}).
\end{aligned}$$

By substituting these relations, noting that $u(1)=1/[\alpha\tau]$ and that $V(k+1\,|\,l)=0$ for $k+l=N$, the first three terms of (3.4) can be written as

$$\begin{aligned}
&\pi([\bar{k}_n, \bar{r}_n])\alpha\{u(r+1)+[\alpha\tau]^{-1}a(r)b(k)/ V(k\,|\,k_1+\cdots+k_{n-1})\\
&\qquad+[\alpha\tau]^{-1}a(r)\, V(k+1\,|\,k_1+\cdots+k_{n-1})/ V(k\,|\,k_1+\cdots+k_{n-1})\}/u(r)\\
&\quad=\pi([\bar{k}_n, \bar{r}_n])\alpha,
\end{aligned} \tag{3.5}$$

where the latter equality follows from $b(k)+V(k+1|l)=V(k|l)$ for $k+l<N$ and $b(k)=V(k|l)$ for $k+l=N$ as by (3.1), and the equality $u(r+1)+[\alpha\tau]^{-1}a(r)=u(r)$ as by (2.3). From (3.2) again, we also derive

$$\pi([\bar{k}_n,\bar{r}_n],(1,1)])=\pi([\bar{k}_n,\bar{r}_n])\lambda\tau u(1)V(1|k_1+\cdots+k_n) \quad (3.6)$$

with

$$\sum_{t=l+1}^{\infty} b(t)=[1-V(1|l)] \quad (3.7)$$

following from (3.1) and $u(1)=1/[\alpha\tau]$, we can thus rewrite the last two terms of (3.4) by

$$\pi([\bar{k}_n,\bar{r}_n])\lambda\{V(1|k_1+\cdots+k_n)+[1-V(1|k_1+\cdots+k_n)]\}=\pi([\bar{k}_n,\bar{r}_n])\lambda. \quad (3.8)$$

By combining (3.5) and (3.8) we have thus shown equality of (3.3) and (3.4), assuming (3.2) and $n>0$. For $n=0$, finally, the global balance equations or equivalently equality of (3.3) and (3.4) leads to the boundary condition

$$\pi(0)\lambda=\pi((1,1))\alpha+\pi(0)\lambda\left[\sum_{t=N+1}^{\infty} b(t)\right]. \quad (3.9)$$

Assuming (3.2) this condition is directly verified similarly to (3.6)–(3.8). The global balance equations are thus guaranteed by (3.2) for any state, so that the proof is completed. □

By noting that $\sum_{r=1}^{\infty} u(r)=1$ and summing over all possible numbers of residual phases for all batches, the following main result is immediate from expression (3.2). This result shows that the steady state batch size distribution has a scaled geometric form and depends upon the services only through their means. It is thus insensitive to service distributional forms.

Theorem 3.2. *For $n>0$ and with $(k_1,\ldots,k_n)$ denoting that jobs are present from n different batches with k_i jobs of batch i, the ith in order of arrival, the steady state distribution is given by*

$$\pi(k_1,\ldots,k_n)=\pi(0)[\lambda\tau]^n\prod_{i=1}^{n}V(k_i|k_1+\cdots+k_{i-1}). \qquad \square \quad (3.10)$$

Remark 3.1 (batch balance). The proof of Theorem 3.1 is actually based upon verifying the global balance equations in the detailed manner of (3.5) and (3.6). Relation (3.5) in particular can be interpreted as equality of the rate into and out of a state as due to a specified batch. This notion of "balance per batch" seems to be a useful extension of earlier notions as partial-, local- or job-local-balance, which are known to be responsible for insensitive closed form results (cf. Kelly, 1976, Schassberger, 1978; Cohen, 1979, Hordijk and Van Dijk, 1983a, b). These latter notions are easily shown to fail for the present system. (For instance, except for one job all jobs arriving in a batch have to wait for service so that their inrate is positive whereas their outrate zero. This also conflicts with the notion of instantaneous attention as usually required for insensitivity (e.g. Schassberger, 1978).) Also, while for non-batch input systems insensitivity results for LCFS disciplines extend for instance to processor-sharing disciplines, such extensions seem to be impossible for systems with batch input. The notion of "balance per batch" therefore is of interest in itself.

Remark 3.2 (recursive computation). As in Van Dijk (1987), a recursive scheme can be derived for computing the normalizing constant $p(0)$ and the steady state probabilities (3.10). We refer to this reference for the essential steps.

Remark 3.3 (finite source batch input). As in Van Dijk (1987), the results here can be extended to non-exponential finite source rather than Poisson batch input. More precisely, with M sources and mean think times λ^{-1}, expression (3.10) remains valid with an additional factor $M!/(M-n)!$

Remark 3.4 (single arrival case). In the single arrival case, geometric steady state expressions under LCFS-disciplines have been widely reported, both with non-exponential services and interarrival times (cf. Cohen, 1982; Cooper and Niu, 1986; Fakinos, 1981; Yamazaki, 1984; Shanthikumar and Sumita 1986). The present paper, however, cannot be seen as a direct application of any of these references as batch arrivals and blocking are of an essentially different nature.

References

A. Barbour, Networks of queues and the method of stages, Adv. in Appl. Probab. 8 (1976) 584–591.

P.J. Burke, Delays in single-server queues with batch-input, Oper. Res. 23 (1975) 830–833.

M.L. Chaudhry and J.G.C. Templeton, A First Course in Bulk Queues (Wiley, New York, 1983).

J.W. Cohen, On a single server queue with group arrivals, J. Appl. Probab. 13 (1976) 619–622.

J.W. Cohen, The multiple phase service network with generalized processor sharing, Acta Inform. 2 (1979) 245–284.

J.W. Cohen, The Single-Server Queue (North-Holland, Amsterdam, 1982).

R.B. Cooper and S.C. Niu, Benes's formula for M/G/1-FIFO "Explained" by preemptive-resume LIFO, J. Appl. Probab. 23 (1986) 550–554.

N.M. van Dijk, A LCFS finite buffer model with finite source batch input, Res. Rept., Free University, Amsterdam, 1987.

A.M. Eikeboom and H.C. Tijms, Waiting time percentiles in the multiserver $Mx/G/c$ queue with batch arrivals, Probab. Eng. Inf. Sci. 1 (1987) 163–174.

D. Fakinos, The $G/G/1$ queueing system with a particular queue discipline, J. Roy. Statist. Soc. 43 (1980) 190–196.

A. Hordijk and N.M. van Dijk, Networks of queues. Part I: Job-local-balance and the adjoint process. Part II: General routing and service characteristics, in: Lecture Notes in Control and Information Sci., Vol. 60 (Springer, Berlin, 1983a) pp. 158–205.

A. Hordijk and N.M. van Dijk, Adjoint processes, job-local-balance and insensitivity of stochastic networks, Bull. 44th Session Int. Statist. Inst. 50 (1983b) 776–788.

A. Hordijk and R. Schassberger, Weak convergence of generalized semi-Markov processes, Stochastic Process. Appl. 12 (1982) 271–291.

I.W. Kabak, Blocking and delays in $M^{(x)}|M|c$-bulk arriving queueing systems, Management Sci. 17 (1970) 112-115.

F.P. Kelly, Reversibility and Stochastic Networks (Wiley, London, 1979).

J. Kohlas, Stochastic Methods of Operations Research (Cambridge Univ. Press, Cambridge, 1982).

D.R. Mansfield and P. Tran-Gia, Analysis of a finite storage system with batch input arising out of message packetization, IEEE Trans. Comm. 30 (1982) 456-463.

R. Nobel, Practical approximations for finite buffer queueing models with batch arrivals, Res. Rept., Free University, Amsterdam, 1982.

J.C.W. van Ommeren, Exponential expansion for the tail of the waiting time probability in the single server queue with batch arrivals, Appl. Probab. (1987), to appear.

R. Schassberger, The insensitivity of stationary probabilities in networks of queues, Adv. in Appl. Probab. 10 (1978) 906-912.

J.G. Shanthikumar and U. Sumita, On $G/G/1$ queues with LIFO-P service discipline, J. Oper. Res. Soc. Japan 29 (1986) 220-230.

Y. Takahashi and T. Katayama, Multi-server system with batch arrivals of queueing and non-queueing calls, in: ITC 11, Vol. 3.2A (Elsevier Science Publishers (North-Holland), Amsterdam, 1985) pp. 41-47.

H.C. Tijms, Algorithms and approximations for batch-arrival queues, in: ITC 1988, Torino, Italy, 1988, to appear.

W. Whitt, Continuity of generalized semi-Markov processes, Math. Oper. Res. 5 (1980) 494-501.

G. Yamazaki, Invariance relations of $GI/G/1$ queueing systems with preemptive-resume last-come-first-served queue discipline, J. Oper. Res. Soc. Japan 27 (1984) 338-246.