

ET

49

05348.-

1987

SERIE RESEARCH MEMORANDA

A LCFS FINITE BUFFER MODEL WITH
FINITE SOURCE BATCH INPUT

Nico M. van Dijk

Researchmemorandum 1987-49 Dec. '87



VRIJE UNIVERSITEIT
FACULTEIT DER ECONOMISCHE WETENSCHAPPEN
EN ECONOMETRIE
AMSTERDAM

A LCFS finite buffer model with finite source batch input

Nico M. van Dijk

Free University, Amsterdam, The Netherlands

Abstract

Queueing systems are studied with a last-come first-served queueing discipline and batch arrivals generated by a finite number of non-exponential sources. A closed form expression for the steady-state distribution of jobs at the queue is derived. This expression has a scaled geometric form and is insensitive to the input distribution. A recursive computation of the normalizing constant and busy source distribution is included. The results are of both practical and theoretical interest as an extension of the standard Poisson batch input case.

Queueing * finite source input * batches * last-come first-served *
finite buffer * closed form expression * insensitivity * normalizing
constant.

1. Introduction

Motivated by Erlang's classical loss model, queueing systems with finite capacity or storage constraints have been extensively investigated over about fifty years. Most notably at present, applications can be found in telecommunication, computer performance evaluation and manufacturing.

Generally the assumption is made that jobs arrive one at a time. However, in various applications it appears practical that jobs arrive in batches of more than one job. For instance, packets of messages such as in voice data traffic may simultaneously arrive at a link for transmission, a computer program may initiate a number of modules at the same time, or parts to be worked on in manufacturing are often offered in groups such as at pallets. In such applications finite capacity constraints naturally arise such as from storage limitations, store and forward buffers, restricted pooled output, or limits on busy periods. In various applications a batch is either fully rejected or accepted (known in the literature as the "whole batch acceptance strategy" in contrast with the "partial batch acceptance strategy"). For instance, a truck is to be unloaded completely, a program is to be read from begin to end, or a manufacturing requires all parts to be finished.

For unrestricted delay systems with non-exponential batch input and non-exponential services both exact results in terms of generating forms (cf. Burke 1975, Cohen 1976) and asymptotic expansions (cf. Van Ommeren 1987) for the waiting time distribution have been obtained. For the special Poisson arrival case and Erlangian services also efficient computational methods have been developed (cf. Chaundry and Templeton 1983, Eikeboom and Tijms 1987). For restricted batch arrival systems, however, explicit results are reported for only the Poisson arrival case. More precisely, also assuming exponential services, one can then recursively determine the steady state queue length distribution by a "flow in equals flow out" principle for sets of the form $\{j, j+1, \dots, j+N\}$ (cf. Kabak 1970, Manfield and Tran-Gia 1982, Chaundry and Templeton 1983, Takahashi and Katayman 1985 and Nobel 1987).

Particularly the generation of batch inputs, however, seems often more natural to be non-exponential as it may involve a number of underlying random stages. For example, a packet of messages might not be sent out before an acknowledgement is received for each message from its addresser, a program execution cannot be started before all subroutines are read in, while interarrival times of trucks to be unloaded involve loading and travel times. In addition, many input systems are by nature of a finite source type such as by a fixed number of local transmission stations, disk drives, or plants.

This paper attempts to attack the case of finite source non-exponential batch input for a particular discipline. Unfortunately, the recursive flow in is flow out principle for Poisson input does not generally extend to finite source input as one has to keep track of the number of idle sources and thus also of how many jobs of which batches (sources) are present and which jobs are worked upon. The non-exponentiality of inputs, moreover, encounters an extra complication as the residual times or number of exponential phases (under Erlang type assumptions) are non-reversible by nature while closed form expressions for systems with capacity constraints generally require reversibility. It seems that no explicit expressions or recursion relations have been reported for restricted systems with finite source or non-exponential batch input.

It is shown that for the special case of a last-come first-served pre-emptive (LCPS-pre) discipline an explicit expression for the steady state queue size distribution can be obtained. This expression has a scaled geometric form and can be computed recursively. Moreover, it is insensitive (robust) to the distributional form of the input (i.e. it depends upon only the mean interarrival time),

Though in practice LCFS-pre disciplines are not commonly in order, they do seem practical in some applications. For instance, stocks are frequently refilled but also worked off at the top, or jobs may deteriorate in value while waiting for service so that it can be profitable to assign higher priority to fresher jobs. Moreover, the closed form expression is

of theoretical interest in itself, because of its geometric type form and its insensitivity as based upon a notion of local balance per batch rather than per job. For systems without batches the latter notion is responsible for product forms (cf. Hordijk and Van Dijk (1983a)). In contrast, while standard product forms are valid also for exponential first-come first-served or non-exponential infinite server disciplines, such extensions do not seem to be possible with batch arrivals (see remark 3.3). Finally, as a finite source input can also be seen as a cyclic closed queueing network, the result of this paper can be regarded as a first step towards queueing networks with batch jobs. Such networks are currently of significant interest to voice-data transmission analysis, packet switching and flexible manufacturing.

The organization is as follows. First, in section 2 the model is described. Next, in section 3 we derive the steady-state job distribution. In section 4, finally, it is shown how this distribution can be recursively computed.

2. Model

Consider a single-server facility which provides service at a unit rate and which has a restricted storage capacity (buffer) for no more than K jobs, the one in service included. Attached to this facility are a fixed finite number of sources that at regular times generate jobs to be served in the following manner. After a source has been idle for some random amount of time, called its think time, with probability $b(k)$ it generates a batch of k jobs that are to be served, $k=1,2,\dots$. When there are still k vacancies within the buffer this whole batch is accepted for service and the source becomes busy. When all jobs of this batch are completed the source becomes idle again and restarts a new think time. When, however, there are only less than k vacancies within the buffer, the whole batch is lost and the source instantaneously restarts a new think time.

The think time distribution is allowed to be generally distributed, say with distribution function A and mean σ . The service requirement of a job

is assumed to be exponential with mean $1/\mu$. All think times, service requirements and batch generations are assumed to be independent.

The facility provides service in a last-come first-served preemptive manner. More precisely, upon acceptance of a batch the service of the batch presently in service is interrupted and the unit service speed is instantly assigned to this new batch. When a batch finishes, the service of the last interrupted batch is resumed. Within a batch the unit service speed can be arbitrarily assigned to the remaining jobs such as in a processor sharing manner.

→ Phase type restriction. For convenience of analysis, we will restrict the presentation to think time distributions of the form:

$$(2.1) \quad A = \sum_{k=1}^{\infty} a(k)E(k, \alpha).$$

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

$$(2.2) \quad \sigma = \sum_{k=1}^{\infty} a(k) [k/\alpha].$$

and

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

is known for renewal theory (cf. Kohlas 1982, p. 47) as the stationary excess probability of " r " residual exponential phases up to the next renewal in a renewal process with renewal distribution A . This restriction will enable us to present a Markovian description so as to verify stationarity by means of global balance equations. It is well-known, however, that any non-negative probability distribution can be arbitrarily closely approximated (in the sense of weak convergence) by the above mixtures of Erlang distributions (cf. Hordijk and Schassberger 1982). Based upon standard weak convergence limit theorems for the prob-

is assumed to be exponential with mean $1/\mu$. All think times, service requirements and batch generations are assumed to be independent.

The facility provides service in a last-come first-served preemptive manner. More precisely, upon acceptance of a batch the service of the batch presently in service is interrupted and the unit service speed is instantly assigned to this new batch. When a batch finishes, the service of the last interrupted batch is resumed. Within a batch the unit service speed can be arbitrarily assigned to the remaining jobs such as in a processor sharing manner.

Phase type restriction. For convenience of analysis, we will restrict the presentation to think time distributions of the form:

$$(2.1) \quad A = \sum_{k=1}^{\infty} a(k)E(k, \alpha).$$

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

$$(2.2) \quad \sigma = \sum_{k=1}^{\infty} a(k) [k/\alpha].$$

and

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

is known for renewal theory (cf. Kohlas 1982, p. 47) as the stationary excess probability of "r" residual exponential phases up to the next renewal in a renewal process with renewal distribution A . This restriction will enable us to present a Markovian description so as to verify stationarity by means of global balance equations. It is well-known, however, that any non-negative probability distribution can be arbitrarily closely approximated (in the sense of weak convergence) by the above mixtures of Erlang distributions (cf. Hordijk and Schassberger 1982). Based upon standard weak convergence limit theorems for the prob-

ability measures of the sample paths on appropriate so-called D-spaces (cf. Barbour 1976, Whitt 1980, Hordijk and Schassberger 1982), the result of theorem 3.2 can therefore be extended to general think times. The highly technical but well-worn details of this approach are referred to.

3. Steady-state distribution

Let the sources be numbered $1, \dots, M$ and denote by

$$\{(s_i, r_i), i=1, \dots, M-n; (s_j, r_j), j=1, \dots, n\}$$

the state in which $M-n$ sources are idle with for source s_i a residual number of r_i exponential phases up to completion of its current think time, $i=1, \dots, n$, and with n busy sources of which source s_j , the j -th in order of having become busy, still requires k_j jobs to be completed, $j=1, \dots, M-n$. Abbreviate this state by: $[\bar{s}, \bar{r}, \bar{k}, n]$. By virtue of the exponential structure, note that the corresponding queueing process constitutes a continuous-time irreducible aperiodic Markov chain with bounded jump rates at the set of admissible states, that is all states with $k_1, \dots, k_n \leq K$. The existence of a unique steady state distribution is thus guaranteed (cf. Kohlas 1982, p. 93). In what follows, steady state distributions are always denoted by $\pi(\cdot)$ and restricted without mentioning to admissible states only.

The following theorem is the key-result of this paper. To this end, for any $1 \leq n \leq M$ and vector $(k_1, \dots, k_n)$ with $k_1 + \dots + k_n \leq K$ define:

$$(3.1) \quad V(k_n | k_1, \dots, k_{n-1}) = \sum_{k=k_n}^{K-(k_1+\dots+k_{n-1})} b(k).$$

Theorem 3.1. With c a normalizing constant, we have

$$(3.2) \quad \pi([\bar{s}, \bar{r}, \bar{k}, n]) = c[\mu\sigma]^{-n} \prod_{i=1}^{M-n} u(r_i) \prod_{j=1}^n V(k_j | k_1, \dots, k_{j-1}).$$

Proof. By virtue of the Markov structure it suffices to verify the global balance equations (cf. Kohlas 1982, p. 93). These require that in any state the total rate (or probability flux) out of that state due to any of the sources $s=1, \dots, M$ is equal to the total rate into that state due to any of the sources $s=1, \dots, M$. Clearly, however, by summation over all sources $s=1, \dots, M$, this in turn is guaranteed if for each source s separately, $s=1, \dots, M$:

$$(3.3) \quad \text{"The rate out of any state due to source } s = \\ \text{the rate into that state due to source } s".$$

The proof is thus completed by verifying (3.3) assuming that (3.2) holds. To this end, consider a fixed source s and a fixed state:

$$[\bar{s}, k, \bar{r}, n].$$

Since only changes due to source s are to be considered while the source specification for all other sources remains fixed, for expository convenience let

$$[a, p, r]$$

denote the same state except for that source s is now in status "a" where $a=0$ means idle and $a=1$ means busy, with for $a=0$: r residual phases up to completion of its think time with dummy position value $p=0$, and for $a=1$: r jobs of its batch still to be completed while at position p in the facility. We will separately verify (3.3) when source s is idle and when source s is busy under (i) and (ii) below.

(i) Source s is idle. Here we assume that $[\bar{s}, k, \bar{r}, n] = [0, 0, r]$ for some r . The rate out of state $[0, 0, r]$ due to source s is given by

$$(3.4) \quad \pi([0, 0, r])\alpha$$

The rate into this state $[0,0,r]$ due to source s is:

$$(3.5) \quad \pi([0,0,r+1])\alpha + \\ \pi([1,n+1,1])\mu a(r) + \\ \pi([0,0,1])\alpha[1 - \sum_{k=1}^K (k_1 + \dots + k_n) b(r)]a(r)$$

where in the second term source s is required to be at the last entered position with one remaining job to be completed (note that n other sources are assumed to be busy), and where the last term reflects that a batch from source s upon completion of its think time is rejected so that another think time is restarted. However, from assuming (3.2) we conclude:

$$\begin{aligned} \pi([0,0,1]) &= \pi([0,0,r])[u(1)/u(r)] \\ \pi([0,0,r+1]) &= \pi([0,0,r])[u(r+1)/u(r)] \\ \pi([1,n+1,1]) &= \pi([0,0,r])[(\mu\sigma)^{-1}/u(r)] V(1|k_1, \dots, k_n) \end{aligned}$$

By substituting these relations in (3.5), noting that $u(1)=1/[\alpha\sigma]$ by virtue of (2.3) and recalling (3.1), we can rewrite (3.5) as:

$$\begin{aligned} \pi([0,0,r])\alpha(u(r+1) + \\ [\alpha\sigma]^{-1}V(1|k_1, \dots, k_n)a(r) + \\ [\alpha\sigma]^{-1}[1-V(1|k_1, \dots, k_n)]a(r))/u(r). \end{aligned}$$

By virtue of the renewal equation: $u(r)=u(r+1)+[\alpha\sigma]^{-1}a(r)$ as according to (2.3), equality of (3.4) and (3.5) is proven. We have thus verified (3.3) when source s is idle.

(ii) Source s is busy. Here we assume that $[\bar{s}, \bar{r}, \bar{k}, \bar{n}]=[1, j, k]$ for some $j \leq n$ and $k \leq K$. First note that for $j < n$ (that is, source s not having the last entered position), by definition both the rate out of state $[1, j, k]$ and the rate into state $[1, j, k]$ are equal to 0, so that (3.3) holds. Therefore assume $j=n$ (that is, source s has the last entered position and

service is assigned to its batch) while $k \leq K - (k_1, \dots, k_{n-1})$. Then the rate out of state $[1, n, k]$ due to source s equals:

$$(3.6) \quad \pi([1, n, k])\mu.$$

The rate into state $[1, n, k]$ due to source s is:

$$(3.7) \quad \pi([0, 0, 1])\alpha b(k) + \\ \pi([1, n, k+1])\mu 1(k_1 + \dots + k_{n-1} + k + 1 \leq K)$$

where $1(t \leq K) = 1$ if $t \leq K$ and 0 else. As before, by assuming (3.2) we find

$$\pi([0, 0, 1]) = \pi([1, n, k])[\mu\sigma]u(1)/V(k|k_1, \dots, k_{n-1}) \\ \pi([1, n, k+1]) = \pi([1, n, k])V(k+1|k_1, \dots, k_{n-1})/V(k|k_1, \dots, k_{n-1})$$

By substituting these relations in (3.7), noting again that $u(1) = 1/[\alpha\sigma]$ and observing that $V(k+1|k_1, \dots, k_{n-1}) = 0$ when $k_1 + \dots + k_{n-1} + k + 1 = K + 1$, we can rewrite (3.7) as:

$$\pi([1, n, k])\mu(b(k) + V(k+1|k_1, \dots, k_{n-1}))/V(k|k_1, \dots, k_{n-1})$$

which by (3.1) proves equality of (3.6) and (3.7) for any $k \leq K - (k_1 + \dots + k_{n-1})$. We have thus also verified (3.3) when source s is busy. This completes the proof of the theorem.

Let the state vector $\vec{k}_n = (k_1, \dots, k_n)$ denote that n sources are busy of which the j -th batch still consists of k_j jobs to be completed. To calculate its steady state distribution $\pi(\cdot)$ first conclude from renewal theory (cf. Kohlas 1982) or standard calculation that

$$\sum_{r=1}^{\infty} u(r) = 1.$$

As a consequence, by summing over all possible numbers of residual exponential phases for idle sources and disregarding which sources are

actually idle or busy, the following theorem immediately results from theorem 3.1 and elementary combinatorics. This theorem shows that the steady state batch size distribution is insensitive to the input distribution function A (i.e. it depends only upon its mean) and has a geometric type form.

Theorem 3.2. With c the normalizing constant from (3.2), we have

$$(3.8) \quad \pi(\vec{k}_n) = c[M!/(M-n)!][\mu\sigma]^{-n} \prod_{j=1}^n V(k_j | k_1, \dots, k_{j-1})$$

Remark 3.3. For stochastic networks in which only one job or individual component can change at a time, it is well-known (cf. Cohen 1979, Kelly 1979, Schassberger 1978, Hordijk and Van Dijk 1983a, 1983b) that notions of partial or local balance per job or individual component separately are responsible for product form expressions and insensitivity results. In this respect it seems worthwhile noting that the source balance equations (3.3) actually require local balance per batch rather than per job. In fact, balance per job can be shown to fail. The notion of batch local balance therefore seems of interest for further investigation.

Remark 3.4. Based upon the expressions (3.2) and (3.8) also other steady state distributions or performance measures of interest can in principle be computed by simple substitution and enumeration. For instance, one might be interested in just the total number of jobs still to be completed, the loss probability of a batch, or the server utilization. Clearly, enumeration of detailed states is computationally most inefficient. In the next section therefore it will be shown how the busy source distribution and normalizing constant can be efficiently computed by recursion. Similar recursions can be derived also for other distributions or quantities.

4. Recursive computation

This section is concerned with an efficient computation of the busy source distribution and normalizing constant. To this end, a recursion will be derived from (3.8). Let state n denote that n sources are busy.

Then, by (3.8) its steady state distribution $\pi(\cdot)$ can be calculated from:

$$(4.1) \quad \pi(n) = c[\mu\sigma]^{-n} [M!/(M-n)!] \sum_{k_1+\dots+k_n \leq K} \prod_{j=1}^n V(k_j | k_1, \dots, k_{j-1}).$$

For convenience define for all $k_1+\dots+k_n \leq t \leq K$:

$$(4.2) \quad U_t(k_n | k_1, \dots, k_{n-1}) = \sum_{k=k_n}^{t-(k_1+\dots+k_{n-1})} b(k)$$

as well as for all $n \leq t \leq K$:

$$(4.3) \quad \Phi^n(t) = \sum_{k_1+\dots+k_n \leq t} \prod_{j=1}^n U_t(k_j | k_1, \dots, k_{j-1}).$$

Then we can write

$$(4.4) \quad \pi(n) = c[\mu\sigma]^{-n} [M!/(M-n)!] \Phi^n(K).$$

Now, first conclude from (4.2):

$$(4.5) \quad U_t(k_n | k_1, \dots, k_{n-1}) = U_{t-k_1}(k_n | k_2, \dots, k_{n-1}).$$

By the reduction (4.5), and writing out $\Phi^n(t)$ we find

$$\begin{aligned} (4.6) \quad \Phi^n(t) &= \sum_{k_1=1}^t U_t(k_1) \sum_{k_2+\dots+k_n \leq t-k_1} \prod_{j=2}^n U_t(k_j | k_1, \dots, k_{j-1}) \\ &= \sum_{k_1=1}^t U_t(k_1) \sum_{k_2+\dots+k_n \leq t-k_1} \prod_{j=2}^n U_{t-k_1}(k_j | k_2, \dots, k_{j-1}) \\ &= \sum_{k_1=1}^t U_t(k_1) \Phi^{n-1}(t-k_1). \end{aligned}$$

From (4.4) and (4.6) we thus derive the following recursive scheme for computing $\pi(n)$ and c :

$$\pi(n) = c[\mu\sigma]^{-n} \Phi^n(K)$$

$$c^{-1} = 1 + \sum_{n=1}^K [\mu\sigma]^{-n} \Phi^n(K)$$

where

$$\Phi^n(t) = \sum_{k_1=1}^t \left[\sum_{k=k_1}^t b(k) \right] \Phi^{n-1}(t-k_1),$$

$$\Phi^1(t) = \sum_{k_1=1}^t \left[\sum_{k=k_1}^t b(k) \right] \quad \text{and} \quad \Phi^n(.) = 0.$$

References

- Barbour, A. (1976), Networks of queues and the method of stages, Adv. Appl. Prob. 8, 584-591.
- Burke, P.J. (1975), Delays in single-server queues with batch input, Oper. Res. 23, 830-833.
- Chaundry, M.L. and Templeton, J.G.C. (1983), A first course in bulk queues, Wiley, New York.
- Cohen, J.W. (1976), On a single server queue with group arrivals, J. Appl. Prob. 13, 619-622.
- Cohen, J.W. (1979), The multiple phase service network with generalized processor sharing, Acta Informatica 2, 245-284.
- Eikeboom, A.M. and Tijms, H.C. (1987), Waiting time percentiles in the multiserver Mx/G/c queue with batch arrivals, Prob. in the Eng. and Inf. Sciences 1, 163-174.
- Hordijk, A. and Schassberger R. (1982). Weak convergence of generalized semi-Markov processes. Stochastic Process. Appl. 12, 271-291.
- Hordijk, A. and Van Dijk, N. (1983a) Networks of queues. Part I: Job-local-balance and the adjoint process. Part II: General routing and service characteristics. Lecture notes in control and informational sciences, Springer Verlag, Vol. 60, 158-205.
- Hordijk, A. and Van Dijk, N. (1983b), Adjoint processes, job-local-balance and insensitivity of stochastic networks, Bull. 44-th Session Int. Stat. Inst. 50, 776-788.
- Kabak, I.W. (1970), Blocking and delays in $M^{(x)}|M|c$ -bulk arriving queueing systems, Mn. Sci. 17, 112-115.

- Kelly, F.P. (1979), Reversibility and stochastic networks, Wiley.
- Kohlas, J. (1982), Stochastic methods of operations research, Cambridge University Press.
- Manfield, D.R. and Tran-Gia, P. (1982), Analysis of a finite storage system with batch input arising out of message packetization, IEEE Trans. Comm. 30, 456-463.
- Nobel, R. (1987), Practical approximations for finite buffer queueing models with batch arrivals, Research Report, Free University, Amsterdam.
- Van Ommersen, J.C.W. (1987), Exponential expansion for the tail of the waiting time probability in the single server queue with batch arrivals. To appear: Applied Probability.
- Schassberger, R. (1978), The insensitivity of stationary probabilities in networks of queues, Adv. Appl. Prob. 10, 906-912.
- Takahashi, Y. and Katayama, T. (1985), Multi-server system with batch arrivals of queueing and non-queueing calls, ITC 11, Elsevier Science Publishers, 3.2 A. 4.1-4.7.
- Whitt, W. (1980), Continuity of generalized semi-Markov processes, Math. Oper. Res. 5, 494-501.