Faculteit der Economische Wetenschappen en Econometrie

Serie Research Memoranda

An exact Solution for a Finite Solution Server Model

Nico M. van Dijk

Research Memorandum 1991 - 10
January 1991

vrije Universiteit amsterdam

AN EXACT SOLUTION

FOR A FINITE SLOTTED SERVER MODEL

Nico M. van Dijk†
Bond University, Queensland, Australia

Abstract

A slotted service system is studied with a finite capacity, a general input, general services and a FCFS-discipline. A closed form insensitive expression is derived for the total workload distribution. The prooftechnique seems of interest for extension.

Keywords

Slotted Server, Packets, Finite Capacity, Discrete-time Queueing, Insensitive Product Form Type Expression.

†**Visiting Address:** School of Information and Computing Sciences, Bond University, Gold Coast, Queensland, 4229, Australia
Permanent Address: Department of Econometrics, Free University, The Netherlands

Contents

1 Introduction

Motivation

As present-day communication becomes more and more digitized tools from classical continuous-time queueing analysis are no longer applicable. A simple model for digital transmissions is a slotted server that transmits packets in a time-slotted manner. Multiple packets however may arrive during a time-slot such as according to a Poisson arrival process.

In particular each packet or message itself may generate a number of service quanta or segments of some fixed or random number.

For example, as per Potter and Zukerman (1989), (1990), the main capacity limitations of a "router" used for interconnection of "metropolitan area networks" (MANs) will be the number of packets, where, as per the IEEE 802.6 standard, a packet may consist of a large number of segments. This capacity constraint on batches of segments plays an essential role in the newly introduced "Cyclic Request Control (CYREC)" scheme for DQDB (Distributed Queue Dual Bus) protocols.

When the number of packets (or workload) is restricted to a finite capacity constraint no solution for this generic time-slotted system seems to be available.

Background

The dynamics of a slotted service system somewhat resembles that of a queueing system with batch arrivals.

Explicit results for continuous-time finite FCFS queueing systems with batch arrivals are limited to the case of Poisson arrivals and exponential services as based on a 'flow in equals flow out' principle for sets of the form $\{j, j+1, \ldots, K\}$ (cf. Kabak 1970, Mansfield and Tran-Gia 1982, Chaudhry and Templeton 1983, Takahashi and Katayama 1985 and Nobel 1987). Recently, in Van Dijk (1989) these results were relaxed to non-exponential finite source input but under the restrictive condition of a LCFS-Preemptive service discipline.

The discrete-time formulation herein allows a non-exponential (or rather as discrete analogue non-geometric) input and servicing. As a consequence, a direct recursive solution scheme as above is no longer applicable.

As most closely related discrete-time result in the literature, in Daduna and Schassberger (1981) an elegant insensitive product form expression, in analogy with standard continuous-time results, is established for a discrete-time queue under a LCFS-Round Robbin discipline and under the assumption of an infinite capacity. Further, the input is assumed to be Bernoulli so that only one customer or packet per time-slot can arrive. However, the combined features of multiple arrivals and a finite capacity constraint in this paper are the essential complications which lead to a different (non-standard) expression.

Result

An explicit expression will be obtained somewhat similar to that Van Dijk (1989). In contrast, though, general discrete packet lengths are allowed and the total workload distribution also applies to FCFS-disciplines. Furthermore, to obtain this result one has to regard the slotted server model in a convenient manner. Most essentially, this requires different technical details as

1. standard partial balance principles, which are known to be responsible for explicit product form type expressions, are no longer applicable. (Also see remark 3.2).
2. in the discrete-time mechanism probabilities for both arrivals and departures at the same time have to be taken into account. (Also see remark 3.3)

Technique

The result is obtained by an intermediate step of first establishing a detailed and insensitive steady state distribution under a LCFS-assumption. This result is of interest in itself as it is of non-standard form and based on a principle of balance per time-slot which seems promising for extension. For the FCFS-case of primary interest a recursive expression for the total workload distribution will then be derived.

Outline

First, in section 2 the model and the discrete-time mechanism are described. Next, in section 3 a detailed result for the LCFS-case is derived first and followed by the FCFS-total workload distribution. Section 4 gives a recursive relation for this distribution.

2 Model

Consider a single service system or transmission device which processes service segments in a slotted manner as follows. Time is slotted in fixed intervals of length Δ. At the end of a time-slot a number (or batch) of packets (jobs, message requests or data frames) for processing may be generated, say k packets with probability $\alpha b(k)$ where $b(1) + b(2) + \ldots = 1$. Hence, with probability $(1 - \alpha)$ no packets arrive at all. Each packet itself in turn will require a batch of service segments or quanta to be processed, say a batch of j segments with probability $a(j)$.

During one time-slot exactly one segment is processed with probability ν, provided no new packets arrive. Otherwise (also see remark below) no segment is processed. Processing takes place in a FCFS-manner, where service segments that were generated in one and the same time-slot can be ordered in an arbitrary manner.

Finite capacity constraint

The system, furthermore, has a finite capacity to contain no more than a total number of K packets. To this end the 'total loss' protocol is in order. Under this protocol a new packet batch which leads to an excess is rejected in total. More precisely, when k packets are already present, all packets of a new packet batch of more than $K-k$ packets are rejected. Note that in view of this capacity restriction one is forced to keep track of both the residual number of packets and segments.

Remark 2.1 (Assumption) The assumption that in time-slots during which a batch of packets arrives no segment is transmitted may not be satisfied in general but does seem justified either

- due to technical restrictions such as a common device to handle or process both arrival and departures, or
- approximately as the number of transmission slots during which packet batches arrive can be relatively very small, recalling that the time-slot is the length of time required to process just one segment.

Remark 2.2 (Poisson input) As a special case a packet input according to a Poisson process with parameter λ is included by

$$\begin{aligned} \alpha b(k) &= e^{-\lambda\Delta}(\lambda\Delta)^k/k! \\ [1-\alpha] &= e^{-\lambda\Delta} \end{aligned}$$

Remark 2.3 (Finite segment constraint) A finite capacity constraint K on the number of segments rather than the number of packets is covered by the above model by identifying a packet as one segment and incorporating the segment size distribution $a(k)$ (by a convolution term) in the packet size distribution $b(k)$.

Objective

We are interested in the steady state total workload distribution, that is in the steady state probabilities $\pi(k)$ where k is the total residual number of segments that require to be processed.

Steps

In order to determine $\pi(k)$ we will first analyze the system under a LCFS-preemptive assumption in stead of the original FCFS discipline of interest. An explicit and insensitive steady state distribution will be obtained. Next, from this distribution the total workload distribution also for FCFS-disciplines is concluded.

3 Steady State Distribution

For convenience of analytical tractibility, we first assume that packets are processed in the following last-come first-served preemptive (LCFS) manner.

LCFS-description

When during a time-slot a number of packets arrive, no segment in that time-slot is transmitted and the ongoing transmission of segments of another packet is stopped. The next time slot a segment from one of the newly arrived packets is transmitted. More precisely, the newly arrived packets are numbered is some arbitrary order in which they have to be transmitted. The segments from the packet ordered first have to be transmitted first. Next, those from the second one in order etc. Further, with packets still present from n different time-slots, say k_i from the i-th time-slot in time-order, that is with k_n representing the remaining packets that have entered last and k_1 that have entered first, the transmission order of these packets, provided no new arrive meanwhile, is the k_n from the n-th ordered (that is, latest) arrival time-slot first down to the k_1 from the 1-st ordered (that is the earliest) arrival time-slot last.

Notation

With $k_1 + \ldots + k_n \leq K$, let the state

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

denote for each i-th ordered arrival time-slot, the number k_i of remaining packets, as defined above, and r_i the number of remaining segments still to be transmitted of the packet from this group in first next transmission order. (Note that segments from only this remaining packet may already have been transmitted). Further, for any u and j with $u + j \leq K$ and r, we introduce the notation

$$\begin{aligned} V(j|u) &= \sum_{k=j}^{K-u} b(k) \\ R(r) &= \frac{1}{\tau\nu} \sum_{k=r}^{\infty} a(k) \qquad (3.1) \\ \tau &= \sum_{k} \frac{k}{\nu} a(k) \end{aligned}$$

Note that the above state description provides sufficient information for the transition mechanism. More precisely, the corresponding process constitutes an irreducible Markov chain at the set of admissible state determined by $k_1 + \ldots + k_n \leq K$. The existence of a unique steady state distribution is thus guaranteed (eg. Kohlas 1982, p 93). In what follows, steady state expressions are denoted by $\pi(\cdot)$ and without mentioning restricted to admissible states only.

Theorem 3.1 (Detailed distribution) For the LCFS case and with $\pi(0)$ a normalizing constant, we have

$$\pi\left([\bar{k}_n, \bar{r}_n]\right) = \pi(0)\tau^{[k_1+\ldots+k_n]}\left(\frac{\alpha}{1-\alpha}\right)^n \prod_{j=1}^{n}\left[R(r_j)V(k_j|k_1+\ldots+k_{j-1})\right] \tag{3.2}$$

Proof We need to verify the global balance equations for any state $[\bar{k}_n, \bar{r}_n]$ when substituting (3.2). To this end, for arbitrary vector $[\bar{k}_t, \bar{r}_t] = ((k_1, r_1), \ldots, (k_t, r_t))$ and (k, r) let

$$\left[[\bar{k}_t, \bar{r}_t], (k, r)\right] = ((k_1, r_1), \ldots, (k_t, r_t), (k, r))$$

First, consider a state $[\bar{k}_n, \bar{r}_n]$ with $n > 0$ and let $(k_n, r_n) = (k, r)$ as corresponding to the last ordered arrival time-slot. Recall that $(1-\alpha)$ is the probability of no arrivals and $b(k)$ the conditional probability of k arrivals given that arrivals occur in a time-slot. We now have to show that the probability (outrate)

$$\pi\left([\bar{k}_n, \bar{r}_n]\right) = \pi\left([\bar{k}_n, \bar{r}_n]\right)\{(1-\alpha)+\alpha\} \tag{3.3}$$

is equal to the total probability flux into this state given by

$$\begin{aligned}
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r))(1-\alpha)(1-\nu)+\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r+1))(1-\alpha)\nu+\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k+1, 1))(1-\alpha)\nu a(r)1_{\{k_1+\ldots+k_{n-1}+k<K\}}+\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r), (1, 1))(1-\alpha)\nu 1_{\{k_1+\ldots+k_{n-1}+k<K\}}+\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}])\alpha b(k)a(r)+\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r))\alpha[\sum_{j=K-(k_1+\ldots+k_n)+1}^{\infty} b(j)]
\end{aligned} \tag{3.4}$$

where the first four terms all correspond to no packets arriving, while the last two terms concern arriving packets and where the last term reflects the rejection of a total number of arriving packets when the level K is excessed. By substituting (3.2), writing $\ell = k_1 + \ldots + k_{n-1}$ and noting that $V(k|\ell) > 0$ if $k + \ell \leq K$ we obtain

$$\begin{aligned}
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r+1)) = \pi([\bar{k}_n, \bar{r}_n])R(r+1)/R(r)\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k+1, 1)) = \pi([\bar{k}_n, \bar{r}_n])V(k+1|\ell)\tau R(1)/[V(k|\ell)R(r)]\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}], (k, r), (1, 1)) = \pi([\bar{k}_n, \bar{r}_n])V(1|k+\ell)\tau R(1)\alpha/(1-\alpha)\\
&\pi([\bar{k}_{n-1}, \bar{r}_{n-1}]) = \pi([\bar{k}_n, \bar{r}_n])\tau^{-1}[V(k|\ell)R(r)]^{-1}[(1-\alpha)/\alpha]
\end{aligned} \tag{3.5}$$

By substituting these relations in (3.4) and noting that $R(1) = [\tau\nu]^{-1}$, we obtain for $k + \ell \leq K$:

$$\begin{aligned}
&\pi([\bar{k}_n, \bar{r}_n])(1-\alpha)\Big\{(1-\nu) + R(r)^{-1}\{\nu R(r+1) \\
(3.6)\qquad &[V(k+1|\ell)1_{\{k+\ell+1\leq K\}}a(r) + b(k)a(r)r^{-1}]/V(k|\ell)\}\Big\} + \\
&\pi([\bar{k}_n, \bar{r}_n])\alpha[V(1|k+\ell)1_{\{k+\ell+1\leq K\}} + \sum_{j=K-(k+\ell)+1}^{\infty} b(k)]
\end{aligned}$$

Now first note that the indicators $1_{\{k+\ell+1\leq K\}}$ can be deleted as by (3.1): $V(k + 1|\ell) = V(1|k+\ell) = 0$ for $k + \ell = K$. Further, from (3.1) and the fact that the conditional packet number probabilities $b(k)$ sum up to 1, we also conclude:

$$\begin{aligned}
& V(1|k+\ell) && = 1 - \sum_{j=K-(k+\ell)+1}^{\infty} b(j) \\
(3.7)\qquad & \nu R(r+1) && + a(r)\tau^{-1} = \nu R(r) \\
& V(k+1|\ell) && + b(k) = V(k|\ell)
\end{aligned}$$

As a consequence, (3.6) now reduces to:

$$\pi([\bar{k}_n, \bar{r}_n])\{(1-\alpha) + \alpha\}$$

by which the equality of (3.3) and (3.4) is proven for $n > 0$. For $n = 0$, the global balance equations in the empty state 0 lead to the boundary condition

$$\pi(0) = \pi(0)(1-\alpha) + \pi((1,1))(1-\alpha)\nu + \pi(0)\alpha\left[\sum_{k=K+1}^{\infty} b(k)\right]$$

This in turn is satisfied as substitution of (3.2), $R(1) = 1/[\tau\nu]$ and $V(1|0)$ gives:

$$\begin{aligned}
\pi(1,1) &= \pi(0)\tau[\alpha/(1-\alpha)]R(1)V(1|0) \\
&= \pi(0)\frac{\alpha}{\nu(1-\alpha)}\left[1 - \sum_{k=K+1}^{\infty} b(k)\right]
\end{aligned}$$

The proof of the theorem is hereby completed.

As an immediate consequence, with $\bar{k} = (k_1, \ldots, k_n)$ only denoting that k_i packets from an i-th ordered arrival time-slot are still present the following insensitivity result is obtained. This result shows that the steady state packet distribution depends on the segment distribution $a(k)$ only through its mean. Herein, let $k = k_1 + \ldots + k_n$ be the total number of remaining packets (workload).

Theorem 3.2 (Insensitivity result) For the LCFS-case, with $\pi(0)$ a normalizing constant and $k = k_1 + \ldots + k_n$ the total number of packets, we have

$$\pi(\bar{k}) = \pi(0)\tau^k \left(\frac{\alpha}{1-\alpha}\right)^n \prod_{j=1}^{n} V(k_j|k_1 + \ldots + k_{j-1}) \tag{3.8}$$

Proof First note that the capacity restriction K allows an arbitrarily large number of segments (Note that this does not conflict with remark 2.3). For any given i-th time slot and number k_i we can thus arbitrarily vary the number r_i. The result now immediately follows using

$$\sum_{r=1}^{\infty} R(r) = \frac{1}{\tau}\sum_{r=1}^{\infty}\sum_{k=r}^{\infty}\frac{a(k)}{\nu} = \frac{1}{\tau}\sum_{k=1}^{\infty}\sum_{r=1}^{k}\frac{a(k)}{\nu} = \frac{1}{\tau}\sum_{k=1}^{\infty}\frac{k}{\nu}a(k) = 1 \tag{3.9}$$

Remark 3.1 (Interpretation). The term $R(r)$ exactly corresponds to the steady state excess probability of a residual number of r time units in a discrete-time renewal process with renewal probabilities $a(\cdot)$. The terms $V(k|\ell)$ would have a similar interpretation with renewal probabilities $b(\cdot)$ when $K = \infty$. For the finite case $K < \infty$, one could thus roughly think of state-dependent truncated steady state excess probabilities $V(k|\ell)$.

Remark 3.2 (Insensitivity and partial balance). In continuous-time frameworks, explicit insensitive expressions are known to be related to special notions of partial balance as opposed to global balance relations, such as most notably: 'local balance'(cf. Schassberger 1978), 'job-local balance' (cf. Hordijk and Van Dijk 1983), 'detailed balance' (cf. Barbour 1976, Kelly 1979) or 'partial balance' (cf. Whittle 1985). The present result relates to the principle of balance per batch generated per source as used for the finite source models in Van Dijk 1989, if one identifies each time-slot as a source. The number of 'sources', though, has now become unlimited. Further, in a finite source model the generation of new arrivals is stopped when the source becomes busy while here the arrival process always continues.

Remark 3.3 (Simultaneous arrivals and departures). As a technical complication, note, though we assume upon packet arrivals no segment to be transmitted, that the probabilities for no joint arrivals and departures have to be taken into account in the balance equations.

FCFS-case

Now let us return to the original FCFS-case of interest. To this end, observe that the total number of packets present (workload) is determined only by the number of packets and thus segments that arrive per time unit or time slot and the service capacity of the system at a probability ν per segment per time-slot. The actual packet precedence or order in which the packets are transmitted is hereby not relevant. (The conservation of workload principle).

As a consequence, the (workload) distribution for the total number of packets for both the LCFS and FCFS-case (as well as possible other disciplines such as processor sharing) is given by

$$\pi(k) = \sum_{\{\bar{k}:k_1+\ldots+k_n=k\}} \pi(\bar{k}) \tag{3.10}$$

with $\pi(\bar{k})$ as per (3.8).

We have thus obtained an exact expression for the total workload distribution which is insensitive for the packet length distribution and which in principle can be computed straightforward. To possibly further reduce the computational complexity of (3.10), in the next section a recursive computational scheme will also be presented.

4 Recursive Computation

The recursive scheme below is related to the one in section 4 of van Dijk (1989) but given in detail for selfcontainedness as now also the workload distribution is involved.

First note that nor the total number of packets k nor n can ever exceed K. For any $k_1 + \ldots + k_n \leq t \leq K$ and any n define

$$\begin{aligned} P(t) &\leq \sum_{k\leq t} \pi(k) \\ U_t(k_n|k_1,\ldots,k_{n-1}) &= \sum_{j=k_n}^{t-(k_1+\ldots+k_{n-1}} b(j) \\ \phi^n(t) &= \sum_{k_1+\ldots+k_n\leq t} \tau^{[k_1+\ldots+k_n]} \prod_{j=1}^{n} U_t(k_j|k_1,\ldots,k_{j-1}) \end{aligned} \tag{4.11}$$

Then by (3.10) and the normalization condition $P(K) = 1$:

$$\begin{aligned} P(t) &= \pi(0)[1 + \sum_{n=1}^{K}[\alpha/(1-\alpha)]^n \phi^n(t)] \\ \pi(0)^{-1} &= 1 + \sum_{n=1}^{K}[\alpha/(1-\alpha)]^n \phi^n(K) \end{aligned} \tag{4.12}$$

By recursively expressing $\phi^n(\cdot)$ in $\phi^{n-1}(\cdot)$ we will hereby establish a recursive computational scheme for the cumulative workload distribution $P(t)$. To this end, we write using (4.1) for $n \geq 1$:

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

Hence

$$\begin{aligned}\phi^n(t) &= \sum_{k_1=1}^{t} \left[\sum_{k=k_1}^{t} b(k) \right] \phi^{n-1}(t-k_1) \qquad (n \geq 1) \\ \phi^0(t) &= 1 \qquad (t \leq K)\end{aligned} \tag{4.13}$$

Acknowledgement The author would like to thank Bond University and Telecom Australia for arranging a short visit to this university during which this paper was written. He would also like to thank Moshe Zukerman from Telecom Australia for his mentioning the related possible DQDB application.

5 References

[1] Barbour, A. (1976), "Networks of queues and the method of stages", Adv. Appl. Prob. 8, 584-591.

[2] Chaudhry, M.L. and Templeton, J.G.C. (1983), "A first course in bulk queues", Wiley, New York.

[3] Daduna, H. and Schassberger, R. (1981), "A discrete-time round robbin queue with Bernoulli input and general arithmetic service time distributions", Acta Informatican 15, 251-263.

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

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

[6] Kelly, F.P. (1979), "Reversibility and stochastic networks", Wiley, London.

[7] Kohlas, J. (1982), "Stochastic methods of operations research", Cambridge University Press, Cambridge.

[8] Mansfield, 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.

[9] Nobel, R. (1987), "Practical approximations for finite buffer queueing models with batch arrivals", Res. Rept, Free University, Amsterdam, To appear: European J. Oper Res.

[10] Potter, P.G. and Zukerman, M. (1989), "A discrete shared processor model for DQDB", To appear: Special issue 'Computer Networks and ISDN Systems' on the 7th ITC Seminar, Adelaide, Australia, September 1989.

[11] Potter, P.G. and Zukerman, M. (1990), "Cyclic request control for provision of guaranteed bandwidth within the DQDB framework", Proceedings ISS '90, Paper No A4.1, Stockholm, Sweden.

[12] Schassberger, R. (1978), "The insensitivity of stationary probabilities in networks of queues", Adv. Appl. Prob. 10, 906-912.

[13] Van Dijk, N.M. (1989), "A LCFS finite buffer model with finite source batch input" J. Appl. Prob. 26, 372-380.

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

[15] Whittle, P. (1985), "Partial balance and insensitivity", J. Appl. Prob. 22, 168-176.