stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE
(DEPARTMENT OF OPERATIONS RESEARCH)

BW 97/79 JANUARI

A. FEDERGRUEN & H.C. TIJMS

COMPUTATION OF THE STATIONARY DISTRIBUTION OF THE QUEUE SIZE IN AN M/G/1 QUEUEING SYSTEM WITH VARIABLE SERVICE RATE

Preprint

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

AMS(MOS) subject classification scheme (1970): 60K25

Computation of the stationary distribution of the queue size in an M/G/1 queueing system with variable service rate *)

by

A. Federgruen [1] & H.C. Tijms [2]

ABSTRACT. This paper presents a simple and computationally tractable method which recursively computes the stationary probabilities of the queue size in an M/G/1 queueing system with variable service rate. For each service two possible service types are available and the service rule is characterized by two switch-over levels. The computational approach discussed in this paper can be applied to a variety of queueing problems.

KEY WORDS & PHRASES: *M/G/1 queueing system, variable service rate, switch-over service rule, queue size, stationary distribution, computational method.*

---

*) This paper will be submitted for publication elsewhere.

1) Mathematisch Centrum, Amsterdam; temporarily Graduate School of Management, University of Rochester.

2) Free University, Amsterdam.

1. INTRODUCTION

An important queueing model arising in various practical situations is one in which the service rate can be varied depending on the queue length. Consider a single server system where customers arrive in accordance with a Poisson process with rate λ. Upon arrival a customer is immediately served if the server is idle and he waits in line if the server is busy. Each customer is served by using one of a finite number of available service types. For ease of presentation we assume two possible service types $k = 1,2$ where under service type k the service time of a customer is distributed as the positive random variable S_k having probability distribution function F_k. It is assumed that $ES_1 < \infty$ and $\lambda ES_2 < 1$. The service rule is characterized by two switch-over levels R_1 and R_2 where R_1 and R_2 are given integers with $0 \leq R_2 \leq R_1$. For a new service the server switches from service type 1 to service type 2 only when the number of customers present is larger than R_1 and switches from service type 2 to service type 1 only when the number of customers present is smaller than or equal to R_2.

This queueing model with linear waiting costs and switch-over costs was studied in Tijms (1978) where an algorithm was derived for computing the values of R_1 and R_2 for which the long-run average expected costs per unit time are minimal. In this paper we are concerned with the computation of the stationary distribution of the queue size under the above service rule with given parameters R_1 and R_2. In Loris-Teghem (1978) Laplace transform results for this stationary distribution were obtained. However, these analytical results seem to be of little practical value for computational purposes. In this paper we shall derive a computationally tractable method which computes recursively the stationary probabilities. The derivation is simple and based on a very useful idea introduced in Hordijk and Tijms (1976) where a recursive relation was found for the stationary distribution of the queue size in the standard M/G/1 queue, cf. also Gavish (1978) and Neuts (1977b) for another proof of this recursive relation. The approach given in this paper may be useful in the algorithmic analysis of a variety of queueing problems, cf. also Neuts (1977a).

2. THE COMPUTATIONAL METHOD

We first introduce some notation. Unless stated otherwise, we assume for ease that epoch 0 a customer arrives who finds the server idle. Define now the following random variables.

T = the next epoch at which an arriving customer finds the server idle.

N = the number of customers served in the busy cycle $(0,T)$.

$T_k(i)$ = the amount of time during which i customers are in the system and service type k is used in the busy cycle $(0,T)$, $i = 1,2,\ldots$; $k = 1,2$.

$N_k(i)$ = the number of service completion epochs at which the departing customer leaves i customers behind and was served by service type k in the busy cycle $(0,T)$, $i = 0,1,\ldots$; $k = 1,2$.

Define $p_k(i,t)$ as the probability that at epoch t there are i customers in the system and service type k is used ($i \geq 1$, $k=1,2$) and let $p(0,t)$ be the probability that the server is idle at epoch t. By the theory of regenerative processes (cf. Theorem 1 in Stidham (1972)) and the finiteness of ET, we have that the limits

$$p_k(i) = \lim_{t\to\infty} p_k(i,t) \quad \text{and} \quad p(0) = \lim_{t\to\infty} p(0,t)$$

exist and are given by

$$p(0) = \frac{1/\lambda}{ET} \quad \text{and} \quad p_k(i) = \frac{ET_k(i)}{ET} \quad \text{for } i \geq 1 \text{ and } k=1,2, \tag{1}$$

as is intuitively clear by interpreting $p_k(i)$ as the long-run expected fraction of time during which i customers are present and service type k is used. Observe that $\sum_{i=0}^{\infty} p(i) = 1$ where $p(i) = p_1(i)+p_2(i)$ for $i \geq 1$. Further we have

$$p(i) = \frac{EN_1(i) + EN_2(i)}{EN} \quad \text{for } i = 0,1,\ldots \tag{2}$$

To explain this, we first note that by Theorem 3 in Stidham (1972) we can

interprete p(i) as the long-run expected fraction of customers who find upon arrival i other customers in the system (roughly speaking, Poisson arrivals see time averages). Clearly, the right side of (2) can be interpreted as the long-run expected fraction of customers who leave upon departure i customers behind in the system. Since in any time interval the number of times the queue size increases from i to i+1 cannot be differ more than one from the number of times the queue size decreases from i+1 to i, we get the equality (2). By (1) and (2), we have the relation

$$EN = \lambda ET. \tag{3}$$

We shall now relate $ET_k(i)$ and $EN_k(i)$. To do this, define for k=1,2, $j \geq 1$ and $i \geq j$,

$A_k(i,j)$ = expected amount of time during which i customers are present in one service time by type k given that this service starts when j customers are present.

Since the Poisson arrival process has the well-known property that under the condition of n arrival epochs in (0,t) the joint probability distribution of these n arrival epochs is the same as that of the order statistics of n independent random variables uniformly distributed on (0,t) (cf. Theorem 2.3 in Ross(1970)), we find

$$A_k(i,j) = \int_0^\infty dF_k(t)\{\sum_{n=i-j}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \frac{t}{n+1}\}.$$

Note that $A_k(i,j)$ depends on i and j only through i-j. Define for k=1,2,

$$a_k(j) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dF_k(t) \quad \text{for } j=0,1,\ldots, \tag{4}$$

i.e. $a_k(j)$ is the probability of j arrivals during a service time S_k. Then

$$A_k(i,j) = \frac{1}{\lambda} \sum_{n=i-j+1}^{\infty} a_k(n). \tag{5}$$

We note that the probabilities $a_k(j)$, $j \geq 0$ can be recursively computed when

F_k is a phase-type distribution function, cf. Neuts (1977b). Letting $\iota(x)=1$ for $x\geq 0$ and $\iota(x)=0$ for $x<0$, it is easily seen that

$$ET_1(i) = A_1(i,1) + \sum_{j=1}^{\min(i,R_1)} EN_1(j)A_1(i,j) + \iota(i-R_2)EN_2(R_2)A_1(i,R_2) \quad \text{for } i\geq 1 \tag{6}$$

and

$$ET_2(i) = \sum_{j=R_2+1}^{\min(i,R_1)} EN_2(j)A_2(i,j) + \sum_{j=R_1+1}^{i} \{EN_1(j)+EN_2(j)\}A_2(i,j) \quad \text{for } i\geq R_2 \tag{7}$$

where $\sum_a^b = 0$ if $a>b$. Observing that $EN_2(i) = 0$ for $i<R_2$ and

$$p(i) = p_1(i) \quad \text{for} \quad 1\leq i\leq R_2 \tag{8}$$

we get from (1)-(3) and (6)-(7) that

$$p(i)ET = A_1(i,1) + \sum_{j=1}^{i} \lambda A_1(i,j)p(j)ET \quad \text{for } 1\leq i\leq R_2 \tag{9}$$

$$p_1(i)ET = A_1(i,1) + \sum_{j=1}^{R_2} \lambda A_1(i,j)p(j)ET + \sum_{j=R_2+1}^{\min(i,R_1)} EN_1(j)A_1(i,j) \quad \text{for } i>R_2 \tag{10}$$

$$p_2(i)ET = \sum_{j=R_2+1}^{\min(i,R_1)} EN_2(j)A_2(i,j) + \sum_{j=R_1+1}^{i} \lambda A_2(i,j)p(j)ET \quad \text{for } i>R_2. \tag{11}$$

By writing $A_1(i,j) = A_2(i,j) + A_1(i,j) - A_2(i,j)$ in the last term of the right side of (10) and adding the equations (10) and (11), we get

$$p(i)ET = A_1(i,1) + \sum_{j=1}^{R_2} \lambda A_1(i,j)p(j)ET + \sum_{j=R_2+1}^{i} \lambda A_2(i,j)p(j)ET + \sum_{j=R_2+1}^{\min(i,R_1)} EN_1(j)\{A_1(i,j)-A_2(i,j)\} \quad \text{for } i>R_2. \tag{12}$$

Observe that the last term in the right side of (12) vanishes when $R_1=R_2$. Suppose for the moment that we have numerically evaluated ET and $EN_1(j)$ for $R_2<j\le R_1$. Then we recursively compute p(i) for i = 0,1,... by (2),(9) and (12). Next we can compute $p_1(i)$ and hence also $p_2(i)$ by (10). We note that we need not to numerically evaluate ET beforehand when for any $\varepsilon>0$ we can give an integer M such that $\Sigma_{i=0}^{M} p(i) \ge 1-\varepsilon$. Then we can recursively compute $q(i) = p(i)ET$ for i=0,1,... by (2),(9) and (12) and next we can approximate p(i) by $q(i)/\Sigma_{i=0}^{M} q(i)$ in any desired accuracy.

We shall now give a method for the computation of ET and $EN_1(j)$ for $R_2<j\le R_1$. Assume that $R_1\ge 1$. Otherwise $R_1=R_2=0$ and hence $ET=1/\lambda(1-\lambda ES_2)$. Define

$\alpha(i)$ = expected time until the first epoch at which system becomes empty given that at epoch 0 a service of type 1 starts when i customers are present ($1\le i\le R_1$),

and, for any j with $R_2<j\le R_1$, let

$\beta^{(j)}(i)$ = expected number of departing customers who are served by service type 1 and leave j customers behind in the system until the first epoch at which the system becomes empty given that at epoch 0 a service of type 1 starts when i customers are present ($1\le i\le R_1$).

Then

$$ET = \alpha(1) + \frac{1}{\lambda} \quad \text{and} \quad EN_1(j) = \beta^{(j)}(1) \text{ for } R_2<j\le R_1. \tag{13}$$

Noting that in the standard M/G/1 queue with arrival rate λ and service time S_2 the expected length of one busy period equals $ES_2/(1-\lambda ES_2)$, we find

$$\alpha(i) = ES_1 + \sum_{k=0}^{R_1-i+1} a_1(k)\alpha(i-1+k) + \\ + \sum_{k=R_1-i+2}^{\infty} a_1(k)\{(i-1+k-R_2)\frac{ES_2}{1-\lambda ES_2} + \alpha(R_2)\}, \quad 1\le i\le R_1, \tag{14}$$

where $\alpha(0)=0$. Similarly, letting $\delta(a,a)=1$ and $\delta(a,b)=0$ for $a\neq b$, we have

for any $R_2<j\leq R_1$,

$$\beta^{(j)}(i) = \sum_{k=0}^{R_1-i+1} a_1(k)\{\delta(i-1+k,j)+\beta^{(j)}(i-1+k)\} +$$

$$+ \beta^{(j)}(R_2) \sum_{k=R_1-i+2}^{\infty} a_1(k), \qquad 1\leq i\leq R_1, \qquad (15)$$

where $\beta^{(j)}(0) = 0$. Each of the relations (14)-(15) has the structured form

$$x(i) = a(i) + \sum_{k=0}^{R_1-i+1} b(k)x(i-1+k)+c(i)x(R_2), \qquad 1\leq i\leq R_1, \qquad (16)$$

where $a(i)$, $b(i)$ and $c(i)$ are non-negative and known and $x(i)$ is unknown with $b(0)>0$ and $x(0)=0$. From this system of linear equations we can easily compute $x(1)$. Successively for $i=R_1-1, \ldots, 1$ we can express $x(i)$ as a linear combination of $x(R_1)$ and $x(R_2)$ by using the equation for $x(i+1)$. Then, by the linear combination for $x(R_2)$ and the final equation for $x(1)$, we can solve for $x(R_1)$ and $x(R_2)$ and hence for $x(1)$.

REMARK. Consider the case where $R_1=R_2=R$ and $\lambda ES_1<1$. For this special case an alternative method for computing ET can be given. This method is in particular suited when ET must be evaluated for a range of values for R. Define for parameter μ the function $\alpha(i,\mu)$ for $0\leq i\leq R$ by $\alpha(0,\mu)=0$ and

$$\alpha(i,\mu) = ES_1 + \sum_{k=0}^{R-i+1} a_1(k)\alpha(1-i+k,\mu) + \alpha(R,\mu) \sum_{k=R-i+2}^{\infty} a_1(k) +$$

$$+ g(\mu) \sum_{k=R-i+2}^{\infty} (i-1+k-R)a_1(k), \qquad 1\leq i\leq R, \qquad (17)$$

where $g(\mu)=\mu/(1-\lambda\mu)$ and $0\leq\mu<1/\lambda$. For any μ the function $\alpha(i,\mu)$ is uniquely determined by the system of linear equations (17). Observe that, by (14), $\alpha(i,ES_2)=\alpha(i)$ for $1\leq i\leq R$. Inverting the system of linear equations (17), it follows that for some functions $c(i)$ and $d(i)$ independent of μ we have $\alpha(i,\mu)=c(i)+d(i)g(\mu)$ for $1\leq i\leq R$. In particular, there are constants $c(1)$

and d(1) such that

$$\alpha(1,\mu) = c(1) + d(1)g(\mu) \qquad \text{for } 0 \le \mu < 1/\lambda.$$

Hence $\alpha(1)=\alpha(1,ES_2)$ follows by determining c(1) and d(1). By (17) we clearly have that $\alpha(1,ES_1)$ is the expected length of one busy period in the standard M/G/1 queue with arrival rate λ and service time S_1. Hence

$$c(1)+d(1)\frac{ES_1}{1-\lambda ES_1} = \frac{ES_1}{1-\lambda ES_1}.$$

Further, it is readily seen that $\alpha(1,0)$ can be interpreted as the expected length of one busy period in the above standard M/G/1 queue with finite capacity R+1 for the number of customers in the system. For this finite capacity M/G/1 queue denote by $\pi_0(R)$ the stationary probability that the server is idle. Then, by $\pi_0(R)=(1/\lambda)/\{\alpha(1,0)+1/\lambda\}$, we have

$$c(1) = \{1-\pi_0(R)\}/\lambda\pi_0(R).$$

The stationary probabilities of the queue size in the finite capacity M/G/1 queue can be expressed in those of the infinite capacity M/G/1 queue, e.g. see Cooper (1972). In particular we have

$$\pi_0(R) = \pi_0/\{\pi_0+\lambda ES_1 \sum_{i=0}^{R} \pi_i\}$$

where $\{\pi_i, i\ge 0\}$ is the stationary probability distribution of the queue size in the infinite capacity M/G/1 queue with arrival rate λ and service time S_1. As a special case of (12) with $R_1=R_2=0$ and S_2 replaced by S_1, we find that the probabilities π_i for i=0,1, ... can be recursively computed from $\pi_0=1-\lambda ES_1$ and

$$\pi_i = \lambda(1-\lambda ES_1)A_1(i,1) + \sum_{j=1}^{i} \lambda\pi_j A_1(i,j) \qquad \text{for } i\ge 1.$$

Consequently for a range of values for R the quantity $\pi_0(R)$ and so ET can be computed by a simple forward recursion.

We further remark that the average expected queue size may be directly computed in a very similar way as ET, cf. also Tijms (1978). We conclude

by remarking that the above analysis may be routinely extended to the case of more than two service types, group arrivals and switch-over times. Also, by using the approach discussed in this paper, we have obtained promising approximations for the stationary distribution of the queue size in multi-server queueing systems which we are currently investigating.

REFERENCES

1. COOPER, R.B. (1972), *Introduction to Queueing Theory*, MacMillan, New York.
2. GAVISH, B. (1978), A direct method for computing stationary state probabilities for M/G/1 queues with setup times and group arrivals, Working Paper Series Number 7828, Graduate School of Management, University of Rochester.
3. HORDIJK, A. AND TIJMS, H.C. (1976), A simple proof of the equivalence of the limiting distributions of the continuous-time and the embedded process of the queue size in the M/G/1 queue, *Statist.Neerlandica* 30, 97-100.
4. LORIS-TEGHEM, J. (1978), An M/G/1 queueing system in which the service time depend on the queue length, University of Mons (paper presented at the Bolyai Janos Mathematical Society Colloquium on Point Processes and Queueing Theory, Keszthely, Hungary, September 4-8).
5. NEUTS, M.F. (1977a), *Algorithmic Methods in Probability*, Vol. 7 in Studies in the Management Sciences, North-Holland, Amsterdam.
6. NEUTS, M.F. (1977b), Algorithms for the waiting time distributions under various queue disciplines in the M/G/1 queue with service time distributions of phase type, pp. 177-197 in NEUTS (1977a).
7. ROSS, S.M. (1970), *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco.
8. STIDHAM, S., Jr. (1972), Regenerative processes in the theory of queues with applications to the alternating priority queue, *Adv.Appl.Prob.* 4, 542-577.
9. TIJMS, H.C. (1978), An algorithm for average costs denumerable state semi-Markov decision problems with applications to controlled production and queueing systems, Report BW 94/78, Mathematisch Centrum, Amsterdam (to appear in: D.J. WHITE (ed.) *Markov Decision Theory*, Academic Press, New York).