STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

LEIDING: PROF. DR D. VAN DANTZIG
ADVISEUR VOOR STATISTISCHE CONSULTATIE: PROF. DR J. HEMELRIJK

Report SP 67

On queueing processes with a certain type of bulk service

by

A.R.Bloemena

1958

The Mathematical Centre at Amsterdam, founded the 11th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.

1. Introduction

In this paper queueing situations are considered in which the following assumptions are fulfilled:

a) The interarrivaltime $\underline{y}$ *) between two successive arrivals has the distributionfunction:

$$A(y) = 1 - e^{-\lambda y} , \quad y \geq 0 \qquad (1.1)$$

All $\underline{y}$ are mutually independent.

b) If the server is engaged, arriving customers line up in a queue in order of arrival.

c) At the moment a servicetime is finished, the server examines the queuelength. If the length is:

$> n$: service on the first n customers in the queue starts immediately;

$\leq n$, but > 0: service on all customers waiting in the queue starts immediately;

0 : the server interrupts service, resuming it on arrival of a customer.

d) The servicetime $\underline{s}$, which is the time to service a "batch" of customers, does not depend on the size of this batch. All $\underline{s}$ are mutually independent and also independent of the $\underline{y}$. The distribution function of $\underline{s}$ is $B(s) = P\{\underline{s} \leq s\}$.

$B(s) = 0$ for $s \leq 0$. The expectation of $\underline{s}$ is finite. We denote by $\beta(\rho)$ the Laplace transform of $B(s)$:

$$\beta(\rho) \overset{\text{def}}{=} \int_0^\infty e^{-\rho s} \, d\, B(s) ,$$

for $\mathrm{Re}(\rho) \geq 0$.

We define:

*) Random variables will be denoted by underlined symbols. The same symbols not underlined stands for values taken on by the random variables (e.g. observed values).

$$\rho = \lambda \mathcal{E}\,\underline{s} \,,$$

in which ρ is the usual traffic intensity. In the sequel λ will be taken equal to 1, which means that the timescale will be chosen in such a way that the parameter in (1.1) = 1.

The bulkservice situation we deal with can be used as a basis for the analysis of congestion problems connected with e.g, elevators, locks that can lock through more than one ship at a time, and small size ferries.

BAILEY [1], and DOWNTON [5], [6], dealt with a similar bulk service situation. However, they assumed the server to go on serving, also if no customers are waiting. In the situation dealt with in this paper, the server interrupts service in such a case. An example of a bulk server of first type is a locomotive, collecting trucks at a factoryrailwaymarshalling yard at more or less irregular time intervals. The trucks have been loaded or unloaded in the factory, and are waiting in the yard for the locomotive to come and take them away, though not more than a given number at a time. As a rule a locomotive will not wait if it does not find any waiting trucks. BAILEY gave another example, where this type of service can be found, viz. hospital out-patients departments.

2. The stationary state

We consider the instants, just before the server finishes the service of a batch of customers. For these instants we can describe the state of the system completely by a single number, viz. the number of customers waiting. We thus obtain a Markov chain M with a denumerable number of states. As every state of M can be reached from every other one in a finite number of steps and as M can remain in the same state after a single transition, the chain is irreducible and aperiodic. Therefore it follows from corrolary 1 of page 328 of FELLER [7], that a stationary distribution of the queuelength exists and is independent of the initial probability distribution.

To prove that M is ergodic, we apply a theorem by FOSTER-MOUSTAFA [8], [12]:

An irreducible aperiodic Markov chain with transitionmatrix (p_{ij}), $(i,j=0,1,\ldots)$ is ergodic if for some $\epsilon > 0$ and some integer i_o, a non-negative solution $\{y_i\}$ exists of:

$$\sum_j p_{ij} y_j \leqslant y_i - \epsilon \quad \text{for } i > i_o \, , \tag{2.1}$$

and

$$\sum_j p_{ij} y_j < \infty \quad \text{for } i \leqslant i_o \, . \tag{2.2}$$

In applying this theorem we take i_o=n and $\underline{y}$ equal to the number of customers waiting at the instants for which M is defined. Therefore $\sum_j p_{ij} y_j$ is the conditional expectation after one step of the number of waiting customers given $\underline{y}$=i. Starting from a state with number $i \leqslant n$, this conditional expectation is equal to ρ. In taking

$$\rho < n \tag{2.3}$$

we comply with (2.1) and (2.2). Therefore if ρ satisfies (2.3), M is ergodic.

In the following two sections the queue size distribution and the waiting time distribution will be studied under the assumption that the stationary state has been reached.

3. The queue size distribution

To derive an expression for the queue size distribution, the method of collective marks by D. VAN DANTZIG will be used. This method, which has been described in [2], [3] and [4], has recently been applied to a number of congestion problems, e.g. [11] and [13], leading in an elegant and simple way to the desired results.

In order to apply this method, we introduce an event E, which happens with probability 1-X, with $0 \leqslant X \leqslant 1$, whenever a customer arrives, for which events D. VAN DANTZIG used the term "catastrophes". The events E are independent for all customers. In the sequel we only consider the probability that E does not occur, therefore it is immaterial to be more definite about E.

The generating function of the queue-size distribution in th stationary state is:

$$P(X) \stackrel{\text{def}}{=} \sum_{i=o}^{\infty} p_i X^i \, . \tag{3.1}$$

Obviously this is the probability that the queuelength $\underline{i}$=i, multiplied by the probability that with respect to none of the i waiting customers E happened, summed over i. Therefore P(X) is the probability that with respect to none of the customers waiting at the instant just before the server finishes serving a group of customers, E happened.

The probability that j customers arrive in a servicetime and with respect to none E happened, is easily seen to be

$$\int_0^\infty e^{-s} \frac{s^j}{j!} X^j \, dB(s) \ . \tag{3.2}$$

Therefore

$$\beta(1-X) \ , \tag{3.3}$$

which is obtained from (3.2) by summing over j, is equal to the probability that with respect to none of the customers arriving in a servicetime E happens.

If at the end of a servicetime the queue size is $\leqslant n$, no customers remain waiting after the service on the next batch of the customers has started. If the queue size $\underline{i}$ equals $i > n$, $i-n$ customers remain waiting after the start of the next servicetime. We obtain therefore as the probability that no E happens with respect to any customers remaining in the queue after the service on the next batch of customers started:

$$\sum_{i=o}^{n} p_i + \sum_{i=n+1}^{\infty} p_i X^{i-n} \ . \tag{3.4}$$

The probability that E happens neither with respect to any of the customers remaining in the queue after the beginning of a servicetime, nor to any of the customers arriving during this servicetime can, because of the independence of the events, be found from (3.3) and (3.4) as:

$$\left(\sum_{i=o}^{n} p_i + \sum_{i=n+1}^{\infty} p_i X^{i-n}\right) \beta(1-X) \ . \tag{3.5}$$

From the interpretation it is clear that (3.5) is equal to the probability P(X) that with respect to none of the customers waiting at the end of a servicetime E happened.

Therefore

$$P(X) = \beta(1-X)\left(\sum_{i=o}^{n} p_i + \sum_{i=n+1}^{\infty} p_i X^{i-n}\right)$$

or:

$$P(X) = \frac{\beta(1-X)\left(\sum_{i=o}^{n-1} p_i(X^i - X^n)\right)}{\beta(1-X) - X^n} \tag{3.6}$$

This expression has been derived for $0 \leqslant X \leqslant 1$. However, by analytic continuation the domain of X may be extended to $|X| \leqslant 1$. As P(X) is a generating function, it has to be regular in this domain. As we shall show, the denominator has n roots within or on the unit circle, if $\rho < n$. By choosing the n unknown probabilities $p_o, \ldots, p_{n-1}$, in such a way that the n roots of the denominator coincide with roots of the nominator, we can meet the regularity requirements for P(X).

In order to prove that the denominator of (3.6) has n roots with $|X| \leqslant 1$, we consider:

$$f_\nu(X) \overset{\text{def}}{=\!=} \beta(1-X)-(1+\frac{1}{\nu})\ X^n \ . \qquad (3.7)$$

On the unit circle:

$$\left|-(1+\frac{1}{\nu})\ X^n\right| = 1+\frac{1}{\nu}$$

and

$$|\beta(1-X)| \leqslant \int_0^\infty \left|e^{-(1-X)s}\right| dB(s) \leqslant \int_0^\infty dB(s) = -1 \ ,$$

so according to Rouché's theorem $f_\nu(x)$ has exactly n zero's within the unit circle. We note that it has no zero's on the unit circle.

According to a theorem by HURWITZ ([9], vol.I, p.269) the zero's of $f(x) = \lim_{\nu\to\infty} f_\nu(x)$ are the accumulation points of the zero's of $f_\nu(x)$ for $\nu \to \infty$. In order to apply the theorem, we first note that $f_\nu(x)$ has one root on the real axis between 0 and 1, as both $(1+\frac{1}{\nu})X^n$ and $\beta(1-X)$ are increasing functions of X, while for X=0 $\beta(1-X) > 0$ and for X=1 $\beta(1-X)=1$. For $\nu \to \infty$ the limitpoints of this zero is X=1. The zero at X=1 is a simple one; in case this were not true:

$$n\ X^{n-1} - \beta'(1-X)$$

would have a zero for X=1, which would mean $\rho = n$. This case has been excluded by hypothesis.

Summarizing: $f_\nu(X)$ has exactly n roots within the unit circle; for $\nu \to \infty$ one of these tends to X=1. Moreover as can be seen from (3.7) for $\nu < \infty$ and $\nu = \infty$ there are no other roots with $|X| = 1$, therefore in going to the limit for $\nu \to \infty$, no roots can cross the unit circle. From this it follows that f(X) has n-1 roots within the unitcircle and one at X=1.

Formula (3.6) is the same as found by BAILEY; in his paper he proves for several special cases of $\beta(\varphi)$ this to be a nonnull solution of the queue size distribution.

4. The waitingtime distribution

In this section we derive an expression for the Laplace transform $\gamma(\varphi)$ of C(w), the distribution function in the stationary state of the waitingtime $\underline{w}$ of a customer.

The probability that during the waitingtime of a customer j customers arrive, with respect to whom E does not happen, is easily seen to be equal to:

$$\int_0^\infty e^{-w} \frac{w^j}{j!} X^j \, dC(w) \quad . \tag{4.1}$$

Summing over j gives $\gamma(1-X)$. This quantity can therefore be interpreted as the probability that during a waitingtime no customers arrive with respect to whom E happens. The probability that during a waiting and service time no customers arrive with respect to whom E happens is therefore

$$\gamma(1-X)\beta(1-X) \quad . \tag{4.2}$$

The customers waiting at the end of a servicetime have all arrived during the waiting and servicetime of the last customer in the batch on which service is about to be completed. The probability that no E happens to the customers arriving during the waiting and service time of a customer, under the condition that this customer is the last one in a batch, is therefore:

$$P(X) \quad , \tag{4.3}$$

viz., the probability that with respect to none of the customers waiting at the end of a servicetime E happens.

Consider again the queuelength $\underline{i}$ at the end of a servicetime. With probability p_o $\underline{i}=0$. In this case the next batch to be served only consists of a first customer. The probability that with respect to neither him nor with respect to any customer arriving in his (zero-) waitingtime E happens is in this case equal to p_oX.

The probability that one or more customers are waiting at the end of the servicetime, and that with respect to none of these E happens is equal to $p(X)-p_o$. This again is the probability that neither with respect to the first customer in the batch nor to any of the customers arrived in his waitingtime E happened. From this we gather that

$$(P(X) - p_o + p_oX)\ \beta(1-X) \qquad (4.4)$$

can be interpreted as the probability that under the condition that a customer in the first one is a batch, neither to him nor to any of the customers arriving in his waiting + service time E happens.

We shall now consider two customers entering the system consecutively. For definiteness we shall label them as customer r and r+1 respectively and denote the Laplace transforms of their waitingtime distribution by $\gamma_r(\xi)$ and $\gamma_{r+1}(\xi)$ and the Laplace transforms of their servicetimes by $\beta_r(\xi)$ and $\beta_{r+1}(\xi)$ respectively. The servicetime of customer r+1 may or may not be the same as the one during which r will be served. We denote the probability that two consecutive customers are not served in the same batch by C. C will be determined later. If customers r and r+1 are not served in the same batch, customer r necessarily is a last customer in a batch and customer r+1 a first one in the next batch. It follows from the definition of C and from (4.3) that the probability that customer r is the last one in a batch and that with respect to none of the customers arriving in his waiting and service time E happens is equal to

$$CP(X)\ . \qquad (4.5)$$

Comparing the interpretation of (4.5) with those of (4.2) we find

$$\gamma_r\ (1-X)\ \beta_r(1-X)\ -\ CP(X) \qquad (4.6)$$

to be equal to the probability that customer r is not the last one in a batch and with respect to none of the customers arriving in his waiting and service time E happens.

The probability that customer r+1 is the first one in the batch, that E does not happen either with respect to himself or to any customer arriving during his waiting and service time, follows from the definition of C and of (4.4) as:

$$C\left\{P(X)\ -\ p_o(1-X)\right\}\beta_{r+1}(1-X)\ . \qquad (4.7)$$

Therefore

$$X\,\gamma_{r-1}(1-X)\ \beta_{r+1}(1-X)\ -\ C\left\{P(X)-p_o(1-X)\right\} \qquad (4.8)$$

can be interpreted as (cf. (4.2)) the probability that customer r+1 is not a first one in a batch, and that E happens neither with respect to himself nor to any of the customers arriving in his waiting and service time. From the interpretation of (4.6) and (4.8) in terms of probabilities, it follows that both expressions are equal. Under the assumption that the stationary state has been reached, the subscripts may be dropped, leading to:

$$\gamma(1\text{-}X)\,\beta(1\text{-}X) - CP(X) = X\gamma(1\text{-}X)\,\beta(1\text{-}X) - C\beta(1\text{-}X)\left\{P(X)-p_0(1\text{-}X)\right\}$$

or:

$$\gamma(1\text{-}X) = C\left[\frac{p(x)}{1-X}\left\{\frac{1}{\beta(1\text{-}X)} - 1\right\} + p_0\right] \qquad (4.9)$$

Substituting X=1 gives:

$$C = \frac{1}{\rho + p_0} \qquad (4.10)$$

As C is a probability we find from (4.10)

$$p_0 \geq 1 - \rho \qquad (4.11)$$

As has been shown in section 2, the stationary state will be reached if $\rho < n$. For $1 \leq \rho \leq n$ (4.11) does not contain any new information, as p_0, being a probability itself, will always be ≥ 0. For $\rho < 1$ (4.11) gives a non trivial bound for p_0. It can be found from (4.9) that the probability of a zero-waiting time is equal to:

$$P\{\underline{w} = 0\} = \frac{p_0}{\rho + p_0} \qquad (4.12)$$

In case n=1, all formulae reduce to the well-known results for M/G/1, where $p_0 = 1 - \rho$.

The situation that has been studied in this report can be extended to the case where the server only starts, if m (>1) or more customers are waiting. The stationary condition $\rho < n$ remains unchanged, and (3.6) is still valid.

I want to thank Prof. Dr D. van DANTZIG and J.Th. RUNNENBURG for their continuous advice and stimulating criticism.

References

1 BAILEY, N.T.J., On queueing processes with bulk service, J.R.S.S. B 16 (1954) 80-87.

2 VAN DANTZIG, D., Kadercursus Statistiek, Stencilled notes of lectures given at the University of Amsterdam, 1947.

3 VAN DANTZIG, D., Sur la méthode des fonctions génératrices, Colloques internationaux du Centre National de la Recherche Scientifique 13 (1948) 29-45.

4 VAN DANTZIG, D., Chaînes de Markof dans les ensembles abstraits et applications aux processus avec régions absorbantes et au problème des boucles, Ann. de l'Inst. H. Poincaré 14 (fasc.3) 1955, 145-199.

5 DOWNTON, F., Waiting times in bulk service queues, J.R.S.S. B 17 (1955) 256-261.

6 DOWNTON, F., On limiting distributions arising in bulk service queues, J.R.S.S. B 18 (1956), 265.

7 FELLER, W., An introduction to probability theory and its applications, New York 1950.

8 FOSTER, F.G., On the stochastic matrices associated with certain queueing processes, A.M.S. 24 (1953) 355-360.

9 HURWITZ, A., Mathematische Werke, Basel 1932.

10 Kendall, D.G., Stochastic processes occuring in the theory of queues and their analysis by the method of embedded Markof chains, A.M.S. 24 (1953) 338-354.

11 KESTEN, H. and J.Th. RUNNENBURG, Priority in waiting line problems, Proceedings Kon. Ned. Akad.v.Wet. A 60, Indagationes Mathematicae 19 (1957) 312-336.

12 MOUSTAFA, M.D., Input-output Markof processes, Proceedings Kon. Ned. Akad.v.Wet. A 60, Indagationes Mathematicae 19 (1957) 112-118.

13 RUNNENBURG, J.Th., Probabilistic interpretation of some formulae in queueing theory, Report SP 66 of the Statistical Department of the Mathematical Centre.

Résumé

Dans cet article la situation est étudiée òu des visiteurs, qui se présentent à un guichet, sont servis par un employé, qui peut traiter pendant d'une durée d'opération plus d'une, mais au plus n, visiteurs en même temps. BAILEY et DOWNTON ont ainsi étudiée cette situation dans l'hypothèse que l'employé continuera à servir, même quand il n'y a pas de visiteurs qui attendent pour être servis. Quelquesfois cela ne sera pas conformément la situation qu'on trouve dans les cas òu le mecanisme a un capacité de plus d'un visiteur, comme par example: des ascenceurs, des écluses par lesquelles on peut faire passer plus d'un bateau. Dans ces cas l'employé qui venait d'être libre, examine le nombre des visiteurs dans la queue d'attente. Si ce nombre est egal zéro, il attend qu'un visiteur arrive, et puis il commence à le servir. Si ce nombre est ≥ 1, mais $\leq n$, tous les visiteurs sont admis pour être servis. Si ce nombre est $> n$, n visiteurs sont admis pour être servis.

Les fonctions de répartition du longueur de la queue et de la délais d'attente sont dérivés par interpreter les fonctions géneratrices et les fonctions caractéristiques comme probabilités.