SERIE RESEARCH MEMORANDA

GOOD PROBABILISTIC IDEAS ARE OFTEN SIMPLE

by

Henk Tijms

Research Memorandum 1989-83

December 1989

VRIJE UNIVERSITEIT

Faculteit der Economische Wetenschappen en Econometrie

A M S T E R D A M

GOOD PROBABILISTIC IDEAS ARE OFTEN SIMPLE

by

Henk Tijms
Instituut voor Econometrie, Vrije Universiteit, Amsterdam

The applied fields of inventory and queueing lie at the heart of stochastic operations research. This paper deals with two typical problems from these fields and shows how they can be solved by using simple probabilistic methods.

1. INTRODUCTION

Operations Research is a rather young discipline. From the very beginning stochastic inventory and queueing problems have played a prominent role in the development of Operations Research. It is no coincidence that the development of Operations Research took place in the track of a greater acceptance of probabilistic and statistical methods for solving problems from daily life.

Nowadays widely used tools in stochastic operations research are renewal-reward processes and Markov decision processes. The roots of these tools go back to the early sixties. In that time De Leve made already an extensive use of these probabilistic methods in his research and teaching. His approach was always very intuitive and usually led to surprisingly elegant solutions. This fascinated me as student who was used to formal thinking in probability theory. Chance plays a great role in one's life. My choice for stochastic operations research would not have been made without the motivating education I had in applied probability. A main lesson I learned is that good probabilistic ideas are often simple. In the remainder of this paper I will try to support this claim.

2. THE PERIODIC REVIEW (R,S) INVENTORY SYSTEM WITH RESTIRCTED ORDER SIZE

2.1 Model

A widely used inventory control system is the periodic review system where at each review a replenishment order is placed for the cumulative demand since the previous review. This control rule assumes that there is no limitation on the size of the replenishment order. In practice this assumption is not always satisfied. In this section we will consider the case of a restricted order size. The inventory control model is as follows. The demands for a single product in the successive periods t=1,2,... are independent random variables having a common probability density f(x) with mean μ and standard deviation σ. Demand in excess of the stock on hand is backordered until stock becomes available by the delivery of a replenishment order. The inventory position is reviewed every R periods, where R is a fixed positive integer. At each review the inventory position is ordered up to the level S provided that the order size does not exceed Q; otherwise, an amount of Q is ordered. Here Q is a given number, where it is assumed that

$$Q > R\mu.$$

This prevents the inventory position drifting to minus infinity. The goal is to compute the order-up-to-level S so that the following service level constraint is satisfied:

the fraction of demand satisfied directly from stock on hand $\geq \alpha$,

where α is a prespecified value (e.g., α=0.95). It will be shown in the next subsection how a computationally tractable method can be obtained using simple and basic probabilistic tools.

2.2 Analysis

For ease the analysis will assume that the lead time of any replenishment order is negligible. The outline of the analysis is as follows:

1. For a given control rule (R,S), it will be shown that the process describing the inventory position just after a review is probabilistically equivalent to the waiting-time process in the single-server D/G/1 queueing model with deterministic arrivals. The waiting-time distribution for the D/G/1 queueing model can be explicitly given for the class of Coxian-2 service time distributions.
2. The smallest order-up-to-level S_α achieving the service level α when the demand distribution is general will be approximated by the corresponding order-up-to-level for the case of a Coxian-2 demand distribution having the same fist two or first three moments as the original demand distribution. This approximation requires that the review-time demand has a squared coefficient of variation of at least ½. Otherwise, the order-up-to-level S_α is approximated by linear extrapolation of the corresponding levels for two special Coxian-2 distributions (exponential and Erlang-2) with the same means as the original demand distribution. Here the extrapolation is with respect to the squared coefficient of variation of the review-time demand.

Let us first show that the process describing the inventory just after a review is equivalent to the waiting-time process in a D/G/1 queue. Fix an (R,S) control rule and define the random variable

Δ_i = the difference between the order-up-to-level S and the inventory position just after the i^{th} review.

Letting the random variable ξ_k be defined by

ξ_k = the total demand between the k^{th} and $(k+1)^{st}$ review,

it follows that the inventory position just prior to the i^{th} review equals $S-\Delta_{i-1}-\xi_{i-1}$ and that an amount of $\min(Q,\Delta_{i-1}+\xi_{i-1})$ is ordered at the i^{th} review. It is now easily seen that

$$\Delta_i = \max(0,\Delta_{i-1}+\xi_{i-1}-Q) \qquad \text{for } i=1,2,\ldots,$$

where $\Delta_0=0$ (assuming that the initial stock equals S). The same recurrence relation arises for the single-server D/G/1 queue in which the deterministic interarrival times of the customers are equal to Q and the service time of the k^{th} customer is distributed as ξ_k. Assuming service in order of arrival, let

W_i = the waiting time of the i^{th} customer (excluding service time).

It easily follows that $W_i=W_{i-1}+\xi_{i-1}-Q$ if $W_{i-1}+\xi_{i-1}-Q > 0$ and $W_i=0$ otherwise. This yields the famous Lindley equation,

$$W_i = \max(0,W_{i-1}+\xi_{i-1}-Q) \quad \text{for } i=1,2,\ldots,$$

where $W_0=0$. Consequently, the probability distribution of the "inventory deficit" Δ_i is the same as that of the waiting-time of the i^{th} customer in the above D/G/1 queue. This observation goes back to De Kok [..] who also devised an interesting approximative method to solve the Lindley equation for the general GI/G/1 queue. His method can also be used to a useful approximation for the service level of an (R,S) policy when the demand has a general distribution. However, an alternative approach can be given using the special class of Coxian-2 distributions and the idea of extrapolation with respect to the squared coefficient of variation of the review time demand. A key observation is that the limiting distribution function

$$W(x) = \lim_{i\to\infty} P\{W_i \le x\}, \quad x \ge 0,$$

allows for a simple explicit expression when the service times have a Coxian-2 distribution. Before we proceed, let us describe the practically useful class of Coxian-2 distributions.

Coxian-2 distribution

A positive random variable S is said to have a Coxian-2 (C_2) distribution when S can be represented as

$$S = \begin{cases} U_1 & \text{with probability } 1-b \\ U_1+U_2 & \text{with probability } b \end{cases}$$

for some branching probability $0 \le b \le 1$, where U_1 and U_2 are independently distributed exponential random variables with respective means $1/\mu_1$ and $1/\mu_2$. Any C_2 distributed random variable S can be shown to have a squared coefficient of variation of at least $\frac{1}{2}$, where the squared coefficient of variation c_S^2 is defined by

$$c_S^2 = \frac{\sigma^2(S)}{E^2(S)}.$$

It is often convenient to fit a C_2-distribution to a positive random variable by matching its first two or first three moments. Let X be a positive random variable with $c_X^2 \ge \frac{1}{2}$ and denote by $m_i = E(X^i)$ the i^{th} moment of X. If a three-moment fit to X by a C_2-distribution exists, the three parameters μ_1, μ_2 and b of this unique fit are given by

$$\mu_{1,2} = \frac{1}{2}\left[\alpha_1 \pm \sqrt{\alpha_1^2 - 4\alpha_2}\right], \quad b = \frac{\mu_2}{\mu_1}(\mu_1 m_1 - 1),$$

where $\alpha_1 = (1+\frac{1}{2}m_2\alpha_2)/m_1$ and $\alpha_2 = (6m_1^2 - 3m_2)/(3m_2^2/2 - m_1 m_3)$. An infinite number of C_2-distributions can be fitted to X by matching only the first two moments. An appealing and very useful two-moment fit is the one with parameters

$$\mu_{1,2} = \frac{2}{m_1}\left[1 \pm \sqrt{\frac{c_x^2 - \frac{1}{2}}{c_x^2 + 1}}\right], \quad b = \frac{\mu_2}{\mu_1}(\mu_1 m_1 - 1).$$

This particular C_2-distribution has the same first three moments as a gamma distribution, cf. Tijms [4].

Let us now return to the stationary waiting-time distribution

function W(x) for the D/G/1 queue. In case the service times ξ_i has a C_2-distribution with parameters μ_1, μ_2, and b, then

$$W(x) = 1 - a_1 e^{-\eta_1 x} - a_2 e^{-\eta_2 x} \quad , \quad x \geq 0,$$

where η_1 and η_2 with $0<\eta_1<\min(\mu_1,\mu_2)\leq\eta_2$ are two real zeros of the equation

$$x^2 - (\mu_1+\mu_2)x + \mu_1\mu_2 - \{\mu_1\mu_2-(1-b)\mu_1 x\}e^{-xQ} = 0,$$

and the constants a_1 and a_2 are given by

$$a_1 = \frac{-\eta_1^2\eta_2+\eta_1\eta_2(\mu_1+\mu_2)-\eta_2\mu_1\mu_2}{\mu_1\mu_2(\eta_1-\eta_2)} \qquad a_2 = \frac{\eta_1\eta_2^2-\eta_1\eta_2(\mu_1+\mu_2)+\eta_1\mu_1\mu_2}{\mu_1\mu_2(\eta_1-\eta_2)}.$$

An elementary proof of this result can be found in Van Ommeren and Nobel [6].

Let us next apply the above results to the periodic review inventory model with the (R,S) control rule. For any $v \leq S$, let

$$\Pi(v) = \lim_{i\to\infty} P\{\text{the inventory position just after the } i^{th} \text{ review is} \leq v\},$$

Assume for the moment that the review time demand ξ_i have a C_2-distribution with parameters μ_1, μ_2, and b. It is no restriction to assume that $\mu_1>\mu_2$ (otherwise, redefine μ_1, μ_2, and b as $\mu_1 := \mu_2$, $\mu_2 := \mu_1$, and $b := 1-(1-b)\mu_1/\mu_2$; then, the same density arises). The probability density of the ξ_i's is easily calculated as

$$f_R(x) = \begin{cases} p\mu_1 e^{-\mu_1 x} + (1-p)\mu_2 e^{-\mu_2 x} & \text{when } \mu_1>\mu_2 \\ p\mu_1 e^{-\mu_1 x} + (1-p)\mu_1^2 x e^{-\mu_1 x} & \text{when } \mu_1=\mu_2 \end{cases},$$

where $p=1-b\mu_1/(\mu_1-\mu_2)$ when $\mu_1\neq\mu_2$ and $p=1-b$ when $\mu_1=\mu_2$.

Since the inventory position after the i^{th} review equals $S-\Delta_i$ and the limiting distribution of the Markov process $\{\Delta_i\}$ is given by the function W(x) above, it follows that

$$\Pi(v) = \begin{cases} a_1 e^{-\eta_1(S-v)} + a_2 e^{-\eta_2(S-v)} & \text{for } v<S \\ 1 & \text{for } v=S \end{cases},$$

where the constants η_1, η_2, a_1, and a_2 are specified above. Note that the probability distribution function $\Pi(v)$ has a mass of $1-a_1-a_2=\eta_1\eta_2/(\mu_1\mu_2)$ at the point S. The distribution has a density $\pi(v)$ for $v<S$. We can now give a formula for the long-run franction of demand not satisfied directly from stock on hand. Since it is assumed that the lead time of any replenishment order is zero, the net stock (= on-hand stock minus backlog) is the same as the inventory position (= net stock plus stock on order). Thus, using Markov chain theory, it holds true with probability 1 that

the long-run fraction of demand not satisfied directly from stock on hand

$$= \frac{1}{R\mu}\left\{R\mu I(0) + \int_0^S \pi(v)dv \int_v^\infty (x-v)f_R(x)dx + \right.$$

$$\left. + (1-a_1-a_2)\int_S^\infty (x-S)f_R(x)dx\right\}.$$

To avoid technicalities, let us next assume that the parameters of the C_2 density of the review time demand satisfy $\mu_1\neq\mu_2$ as will be usually the case. Then, using the short-hand notation $\alpha(S)$ for the long-run fraction of demand not satisfied directly from stock on hand under the (R,S) policy, we find the analytical expression

$$\alpha(S) = \frac{1}{R\mu}\left\{ R\mu \sum_{i=1}^{2} a_i e^{-\eta_i S} + (1-a_1-a_2)\sum_{i=1}^{2} (p_i/\mu_i)e^{-\mu_i S} \right.$$

$$\left. + \sum_{i=1}^{2}\sum_{j=1}^{2} p_i a_j \eta_j (e^{-\mu_i S} - e^{-\eta_j S})/(\mu_i(\eta_j-\mu_i))\right\},$$

where $p_1=p$ and $p_2=1-p$. For any specified service level α, Newton-Raphson or bisection can be used to find the order-up-to-level $S=S_\alpha$ for which $\alpha(S)=1-\alpha$.

For the case of Coxian-2 distributed review time demand, we have a tractable method for the computation of the order-up-to-level achieving a prespecified service level. What to do for the case of a generally distributed review-time demand? To answer this question, we distinguish between two cases with respect to the squared coeffient of variation, c_ξ^2, of the review-time demand.

(a) $c_\xi^2 \geq \frac{1}{2}$. Then we suggest to fit a Coxian-2 density to the review time demand ξ by using a three-moment fit whenever possible or using otherwise a two-moment fit. This approximation step is justified by the empirical finding that the service level of an order-up-to-level policy is rather insensitive to more than the first two moments of the review time demand ξ provided that c_ξ^2 is not very large and the review-time demand density has a "reasonable" shape (e.g. a unimodal density).

(b) $c_\xi^2 < \frac{1}{2}$. Then we suggest to use the following approximation procedure. For a prespecified service level α, compute the order-up-to-levels S_α for the case of an exponentially distributed review-time demand and for the case of an Erlang-2 distributed review-time demand, where in both cases the mean of the review-time demand is equal to the mean $\mu_\xi = R\mu$ of the original distribution. This gives the respective order-up-to-levels $S_\alpha(1)$ and $S_\alpha(\frac{1}{2})$. Using the linear interpolation formula $f = f_0 \times (x-x_1)/(x_0-x_1) + f_1 \times (x-x_0)/(x_1-x_0)$, the order-up-to-level S_α for the actual review-time demand distribution is next approximated by

$$S_\alpha \approx 2(1-c_\xi^2)S_\alpha(\tfrac{1}{2}) + 2(c_\xi^2-\tfrac{1}{2})S_\alpha(1),$$

where the extrapolation is with respect to the squared coefficient of variation, c_ξ^2, of the review-time demand. It is an empirical finding that the linear interpolation approach cannot be applied directly to the service level of a given policy, but works satisfactorily for the critical points S_α provided that c_ξ^2 is not too small (say, $c_\xi^2 \geq 0.25$). The interpolation approach using "percentiles" rather than "probabilities" is generally useful, cf. Tijms [4].

Numerical illustration

Let us assume that the demand per period has a gamma distribution whose mean μ and squared coefficient of variation c^2 satisfy

$$\mu = 50, \quad c^2 \in \{1/4, 1/2, 1, 1.5, 2, 3, 4\}.$$

Further, take the length of the review interval equal to

$$R = 2.$$

The maximal order size Q and the service level α are varied as Q=125, 200 and α=0.95, 0.99. The mean and the squared coefficient of variation of the review-time demand are given by $\mu_\xi=2\mu$ and $c_\xi^2=\frac{1}{2}c^2$. In the table we give for the various parameter combinations the approximate order-up-to-level S_α which has been calculated by the procedure sketched above. For the cases with $c_\xi^2 \geq \frac{1}{2}$ a Coxian-2 density is fitted to the review-time demand by matching the first three moments. Also, we give in the table the actual service level of the S_α policy under the situation of gamma distributed demand. The actual service level and its 95% confidence interval are determined by computer simulation. The numerical results confirm that the approximative approach performs quite well, provided that c_ξ^2 is not too small.

Table 2.1 Numerical results

	Q=125				Q=200			
c_ξ^2	$S_{0.95}$	act.service	$S_{0.99}$	act.service	$S_{0.95}$	act.service	$S_{0.99}$	act.service
1/4	234	.944(±.002)	342	.988(±.002)	153	.940(±.001)	209	.986(±.001)
1/2	425	.952(±.004)	642	.990(±.002)	227	.950(±.001)	333	.991(±.001)
3/4	615	.950(±.005)	941	.991(±.002)	300	.950(±.002)	455	.990(±.001)
1	807	.952(±.005)	1240	.989(±.003)	376	.948(±.002)	576	.990(±.001)
1.5	1194	.950(±.007)	1842	.990(±.005)	534	.950(±.002)	828	.989(±.001)
2	1583	.950(±.009)	2444	.989(±.004)	695	.950(±.003)	1081	.991(±.001)

3. OVERFLOW PROBABILITIES IN BUFFERS WITH SERVICE INTERRUPTIONS

3.1 Model

Buffer overflow in communication and production systems is an important problem, particularly when those systems may be subject to random breakdowns. This section shows how simple probabilistic tools can be used to dimension the buffer size so that a very small overflow probability is achieved.

Let us consider a communication channel at which batches of packets arrive according to a Poisson process with rate λ. The batch-size has a general discrete distribution

$$P(\text{batch size is } j) = \beta_j \quad \text{for} \quad j=1,2,\ldots,$$

where $\beta=\Sigma\, j\beta_j$ denotes the average batch size. The packets are temporarily stored in a finite buffer to await transmission. Overflow occurs for those packets from an arriving batch which are excess of the remaining buffer capacity. The packet transmission is synchronous, that is, one packet is taken out of the buffer for transmission at discrete clock times $t=1,2,\ldots$, provided that the transmission channel is available. The packets have all a fixed length and the transmission time of each packet is one time slot. The channel is subject to random service interruptions. The on-times of the channel are assumed to have a geometric distribution

$$P(\text{on-time is } j) = (1-p)p^{j-1} \quad \text{for } j=1,2,\ldots ,$$

while the off-time of the channel have a general discrete distribution

$$P(\text{off-time is } j) = q_j \quad \text{for } j=1,\ldots,M$$

for some finite $M\geq 1$. The on-times and off-times form a sequence of independent random variables. In other words, the process of on- and off-times is modeled as an alternating renewal process in which the breakdowns occur according to a Bernoulli process. A special case of

this model is the simple model of random independent interruptions, where the channel fails in each time slot with a same probability. This simplest model with service interruptions was first studied in Heines [2] and was generalized in Tijms and Van Ommeren [5] and Woodside and Ho [6]. The analysis in these references will be refined and extended in this section. The goal is to find a computationally tractable method for the calculation of the buffer size so that the overflow probability of an arbitrary packet is less than a prespecified value α. In typical applications α is very small (e.g. $\alpha=10^{-9}$). To achieve a very small overflow probability, it must be assumed that the offered load to the channel is not excessively high. The precise assumption will be given in subsection 3.3.

3.2 Analysis via an $M^x/G/1/K$ Queueing Model

Assume for the moment that the buffer size is fixed and that the buffer has room for K packets including any packet in service. In this subsection it will be shown how the communication system with random service interruptions can be analyzed via an $M^x/G/1/K$ queueing model with exceptional first services. This translation step will be crucial in our analysis. In the $M^x/G/1/K$ queueing model batches of customers arrive according to a Poisson process with rate λ. The batch-size has the discrete distribution $\{\beta_j\}$. There is a single server and a finite waiting room with capacity K including any customer in service. An arriving batch whose size exceeds the remaining capacity in the buffer is partially lost due to overflow. The service times of the customers are independent random variables. The service of each customer is distributed as the generic random variable S, except for the first customer in each batch that finds upon arrival the system empty. The service times of those first customers are distributed as the generic random variable S_{exc} (exceptional first service). In addition, there is a warming-up time W before the server can start actual service after an idle period. Hence we are in fact considering a variant of the standard $M^x/G/1/K$ queueing model. In the next subsection we show how to compute the overflow probability for the versatile model with exceptional first services.

In the remaining part of this section we will translate the communication model with service interruptions to the $M^x/G/1/K$ queueing model with exceptional first services. To do so, we need to specify the "normal" service S, the "exceptional" service S_{exc}, and the warming-up time W. The arrival rate λ, the batch-size distribution (β_j), and the buffer capacity K are identical in both models. The translation step is easily understood by the following definition of the service time of a packet in the communication model:

> the service time of a packet = the number of time slots from the discrete clock time at which the packet comes in for its turn for transmission until the moment at which the transmission of the packet is successfully completed.

Further, we define for the communication model,

> the warming-up time = the time elapsed between the arrival of a batch finding the system empty and the beginning of the next time slot.

Let us first specify the probability distribution of the warming-up time W. This requires the calculation of the conditional probabilituy $P\{T_1>t \mid T_1\leq 1\}$, where T_1 is the first arrival epoch in a Poisson arrival process. Thus we find

$$P\{W\leq x\} = \frac{e^{-\lambda(1-x)}-e^{-\lambda}}{1-e^{-\lambda}} \quad \text{for} \quad 0\leq x\leq 1.$$

We have to distinguish between two types of services. First, the service time of a packet whose turn comes directly after the service completion of a preceding packet. Second, the service time of a packet which is served as first one from a batch finding upon arrival no other packets in the system. For the first type of service, the channel was necessarily on during the actual execution of the preceding service. Thus, at the beginning of the new time slot either the on-time continues for a next slot with probability p or an off-time starts with probability 1-p. Since the probability distribution of the

off-time is given by $\{q_j, 1 \le j \le M\}$, it follows that the "normal" service time S in the communication model has the probability distribution

$$P\{S=k\} = \begin{cases} p & \text{for } k=1, \\ (1-p)q_{k-1} & \text{for } 2 \le k \le M+1. \end{cases}$$

For the second type of service, the situation is more complicated. Then we need the distribution of the state of the system at the moment of the first arrival since the end of the last time slot at which a transmission was completed and the system was left empty. Let τ_0 be the time slot just following the latter time slot. At the beginning of time slot τ_0 the system is either in state (0,0) with probability p or in state (k,0) with probability $(1-p)q_k$ for k=1,...,M. Here state (0,0) means that the channel is on and no arrival occurred in the preceding slot, whereas state (k,0) means that the channel is off with a remaining off-time of k time slots while no arrival occurred in the preceding slot. In addition, the system is said to be in state (0,1) if the channel is on and one or more arrivals occurred in the preceding time slot, and the system is said to be in state (k,1) if the channel is off with a remaining off-time of k slots while one or more arrivals occurred in the preceding slot. Some reflection shows that a Markov chain can be used to describe the behavior of the system from the beginning of the particular time slot τ_0 until the beginning of the time slot τ_1 at which one of the (absorbing) states (0,1) or (k,1) for k=1,...,M-1 is reached for the first time. At the beginning of time slot τ_1 the system is no longer empty and an "exceptional" service S_{exc} of a new packet is ready to start. Thus,

$$P\{S_{exc}=k\} = \varphi_{k-1} \quad \text{for } k=1,\ldots,M,$$

where the (absorption) probability φ_j for j=0,...,M-1 is defined as

φ_j = the probability that the system is in state (j,1) at the beginning of time slot τ_1.

To calculate the φ_j's, we define for j=0,...,M-1 the absorption proba-

bilities

$f_{s,(j,1)}$ = the probability that the Markov chain will be absorbed in state (j,1) starting from state s,

where $s \in S_0 = \{(0,0),(1,0),\ldots,(M,0)\}$. Then

$$\varphi_j = pf_{(0,0)(j,1)} + (1-p)\sum_{i=1}^{M} q_i f_{(i,0)(j,1)} \quad \text{for } j=0,1,\ldots,M-1.$$

Using standard arguments from Markov chain theory, it is easily verified that for any fixed j the absorption probabilities $f_{s,(j,1)}$ for $s \in S_0$ can be calculated as the unique solution to a system of M+1 linear equations. For j=0, the linear equations are

$$f_{(0,0)(0,1)} = (1-e^{-\lambda})p + e^{-\lambda}pf_{(0,0)(0,1)} + e^{-\lambda}(1-p)\sum_{i=1}^{M} q_i f_{(0,i)(0,1)}$$

$$f_{(i,0)(0,1)} = (1-e^{-\lambda})U_{i-1} + e^{-\lambda}f_{(i-1,0)(0,1)} \quad \text{for} \quad i=1,\ldots,M,$$

where U_m is an abbreviation for

$$U_m = \begin{cases} 1 & \text{for } m=0 \\ 0 & \text{for } m\neq 0. \end{cases}$$

For any fixed j=1,...,M-1, the linear equations are

$$f_{(0,0)(j,1)} = (1-e^{-\lambda})(1-p)q_j + e^{-\lambda}p\, f_{(0,0)(j,1)} + e^{-\lambda}(1-p)\sum_{i=1}^{M} q_i\, f_{(i,0)(j,1)}$$

$$f_{(i,0)(j,1)} = (1-e^{-\lambda})\, U_{i-j-1} + e^{-\lambda}\, f_{i-1,0)(j,1)} \quad \text{for} \quad i=1,\ldots,M.$$

For any fixed j, the corresponding system of M+1 linear equations is very easy to solve. Each system is upper-diagonal and can be simply solved by backwards substitution. The details are easily worked out.

Remark 3.1 Though the calculations of the absorption probabilities for the general case are not difficult, they become extremely simple for the geometric case. If the off-times have the geometric distribution

$$P\{\text{off-time} = j\} = (1-r)r^{j-1} \quad \text{for} \quad j=1,2,\ldots,$$

we can restrict to a Markov chain with four states (i,j) with $i,j=0,1$. The component $i=0$ (1) means that the channel is on (off), and the component $j=0$ (1) means that no (one or more) arrivals occurred in the preceding time slot. The states $(0,1)$ and $(1,1)$ are absorbing. Now we only have to solve two systems of two linear equations each. The absorption probabilities $f_{(0,0)(0,1)}$ and $f_{(1,0)(0,1)}$ are the unique solution to

$$f_{(0,0)(0,1)} = (1-e^{-\lambda})p + e^{-\lambda}pf_{(0,0)(0,1)} + e^{-\lambda}(1-p)f_{(1,0)(0,1)}$$

$$f_{(1,0)(0,1)} = (1-e^{-\lambda})(1-r) + e^{-\lambda}rf_{(1,0)(0,1)} + e^{-\lambda}(1-r)f_{(0,0)(0,1)}.$$

Similarly, the other two absorption probabilities follow from

$$f_{(0,0)(1,1)} = (1-e^{-\lambda})(1-p) + e^{-\lambda}pf_{(0,0)(1,1)} + e^{-\lambda}(1-p)f_{(1,0)(1,1)}$$

$$f_{(1,0)(1,1)} = (1-e^{-\lambda})r + e^{-\lambda}rf_{(1,0)(1,1)} + e^{-\lambda}(1-r)f_{(0,0)(1,1)}.$$

The two systems of two linear equations each can be explicitly solved. The probability distribution of the "exceptional" service time S_{exc} is now calculated as

$$P\{S_{exc}=1\} = pf_{(0,0)(0,1)} + (1-p)f_{(1,0)(0,1)}$$

$$P\{S_{exc}=j\} = [pf_{(0,0)(1,1)} + (1-p)f_{(1,0)(1,1)}](1-r)r^{j-2} \quad \text{for } j\geq 2.$$

3.3 The Overflow Probability

The overflow probability will be calculated using the method of regenerative processes. This intuitive and powerful probabilistic

approach got its popularity in the OR community after the appearance of the book of Ross [3] in 1970. However, in the sixties, De Leve made already extensive use of the same approach under the name of "herhalingsprogrammering".

Numerous stochastic processes arising, for example, in queueing and inventory systems have the property that they regenerate themselves at certain points in time. Then the behavior of the process after each regeneration epoch is a probabilistic replica of the behavior starting at time zero and is independent of the behavior before the regeneration epoch. It will be intuitively clear that the long-run behavior of a regenerative stochastic process can be studied in terms of its behavior during a single regeneration cycle.

For the communication model, let us say that a cycle starts each time an arriving batch finds the system idle. For the model with a buffer capacity of K packets, define

$N(K)$ = the number of packets served during one cycle.

In particular, $N(\infty)$ denotes the number of packets served during one cycle for the model with $K=\infty$ (i.e., the infinite capacity model). To ensure that $N(\infty)$ has a proper probability distribution, we need the assumption

$$\rho = \lambda\beta E(S) < 1,$$

where $E(S) = 1+(1-p)\ \Sigma_k\ kq_k$ denotes the mean of a "normal" service. Without this assumption we cannot guarantee a very small overflow probability. Also, for the communication model with capacity K, define

$\pi_{loss}(K)$ = the long-run fraction of packets that overflow.

We now derive the following lemmas.

LEMMA 3.1. *For any* $K\geq 1$,

$$\pi_{loss}(K) = \frac{\beta - \rho + \lambda\beta\{E(W)+E(S_{exc})\} - (1-\rho)EN(K)}{\beta - \rho + \lambda\beta\{E(W)+E(S_{exc}\} + \rho EN(K)}\ .$$

Proof. The proof is based on simple probabilistic arguments. First,

the long-run average input of accepted packets
= the long-run average output of accepted packets.

Since the average arrival rate of packets is $\lambda\beta$, we have

the long-run average input of accepted packets = $\lambda\beta(1-\pi_{loss}(K))$.

By the theory of regenerative processes,

$$\text{the long-run average output of accepted packets} = \frac{E[\text{number of packets served during one cycle}]}{E[\text{length of one cycle}]}.$$

The numerator of this ratio is by definition equal to EN(K). Further,

$$E[\text{length of one cycle}] = E(W) + E(S_{exc}) + \{EN(K)-1\}E(S) + \frac{1}{\lambda},$$

using Wald's equation to justify the third term in the right-hand side of this equation. Combining the above relations, the desired result follows.

Next we prove

LEMMA 3.2. *For any* $K\geq 1$,

$$EN(K) = \sum_{j=0}^{K-1} q_j(\infty) \times EN(\infty),$$

where for the infinite capacity model $q_j(\infty)$ *is defined as the long-run fraction of service completion epochs at which* j *packets are left behind in the system.*

Proof. The assumption of a Poissonian arrival process of batches is crucial in the proof. Using the memoryless property of the Poisson process and the fact that the packets are served one at a time, it can be seen that for any fixed $0\leq j\leq K-1$ the probability distribution of the number of service completion epochs at which j packets are left behind

in one cycle for the model with finite capacity K is identical to the corresponding probability distribution for the infinite capacity model. In the latter model, the ratio of the expected number of service completions at which j packets are left behind in one cycle and the expected number of service completions in one cycle is equal to $q_j(\infty)$, by the theory of regenerative processes. Hence $EN(K)/EN(\infty)$ equals $\sum_{j=0}^{K-1} q_j(\infty)$, yielding the desired result.

As a consequence of the Lemmas 3.1 and 3.2, we have expressed the loss probability $\pi_{loss}(K)$ for the finite capacity model in terms of quantities for the infinite capacity model. The latter model is well-studied. We first note

LEMMA 3.3 $$EN(\infty) = \frac{\beta - \rho + \lambda\beta(E(W)+E(S_{exc}))}{1-\rho} .$$

This result is also an immediate corollary of Lemma 3.1 and the fact that $\pi_{loss}(\infty)=0$. The probability distribution $\{q_j(\infty)\}$ is the equilibrium distribution of an embedded Markov chain $\{X_n\}$, where

X_n = the number of packets left behind at the n^{th} service completion epoch in the infinite capacity model.

Using standard arguments from Markov chain theory, we have

$$q_n(\infty) = \sum_{k=1}^{n+1} q_k(\infty) a_{n+1-k} + q_0(\infty) \sum_{k=1}^{n+1} \beta_k \, a^*_{n+1-k}, \quad n=0,1,\ldots,$$

where

$a_k(a^*_k)$ = the probability that a total of k packets arrive during the normal service time S (during the sum of the warming-up time W and the exceptional service time S_{exc}).

Note that

$$q_0(\infty) = \frac{1}{EN(\infty)} ,$$

where an explicit expression for $EN(\infty)$ is given above. Next we apply the basic technique of generating functions. It is a matter of routine

algebra to verify that

$$\sum_{n=0}^{\infty} q_n(\infty) z^n = q_0(\infty) \frac{\{A(z)-\beta(z)A^*(z)\}}{A(z)-z} ,$$

where $\beta(z)$, $A(z)$ and $A^*(z)$ are the generating functions of the probability distributions $\{\beta_j\}$, $\{a_j\}$ and $\{a_j^*\}$. These generating functions can be calculated from

$$\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j , \quad A(z) = \sum_{t=1}^{\infty} e^{\lambda t\{\beta(z)-1\}} P\{S=t\},$$

$$A^*(z) = \frac{(e^{\lambda\beta(z)}-1)e^{-\lambda}}{\beta(z)(1-e^{-\lambda})} \times \sum_{t=1}^{\infty} e^{\lambda t\{\beta(z)-1\}} P\{S_{exc}=t\}.$$

Note that $A^*(z)$ is the product of the generating functions of the probability distributions of the number of arrivals during W and the number of arrivals during S_{exc}. Using the specific structure of the distributions of S and S_{exc}, the expressions for $A(z)$ and $A^*(z)$ can be further simplified.

A practically useful asymptotic expansion for the $q_j(\infty)$'s can be obtained from the generating function under the following assumption.

ASSUMPTION. *The convergence radius R of the power series* $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ *is larger than* 1.

Loosely put, this assumption requires that the batch-size distribution has no extremely long tail. For example, the assumption is satisfied with $R=\infty$ when $\{\beta_j\}$ has finite support. The generating functions above were originally defined only for $|z|\leq 1$. However, under the assumption, they can be analytically extended beyond the unit circle. From complex function analysis it is known that the smallest zero of $A(z)-z$ in the domain beyond the unit circle determines the asymptotic expansion of $q_j(\infty)$ for j large. A minor modification of the proof of Theorem 1 in Tijms and Van Ommeren [5] yields the important result

THEOREM 3.4. *For large* K,

$$\sum_{j=K}^{\infty} q_j(\infty) \sim q_0(\infty) \frac{[\beta(z_0)A^*(z_0)-A(z_0)]}{\{A'(z_0)-1\}(z_0-1)} z_0^{-K} ,$$

where $z_0 \in (1,R)$ *is the unique number satisfying*

$$\sum_{t=1}^{\infty} e^{\lambda t\{\beta(z_0)-1\}} P\{S=t\} = z_0 .$$

We are now in a position to state our main result. Therefore, we first note that by the Lemmas 3.1-3.3,

$$\pi_{loss}(K) = \frac{(1-\rho)[1 - \sum_{j=0}^{K-1} q_j(\infty)]}{1 - \rho + \rho \sum_{j=0}^{K-1} q_j(\infty)} .$$

Hence the following final result is obtained from Theorem 3.4.

THEOREM 3.5. *For* α *small enough, the minimal buffer size* K *satisfying* $\pi_{loss}(K) \leq \alpha$ *can be approximately calculated from*

$$K(\alpha) \approx \frac{1}{\ln(z_0)} \ln\left\{\frac{\gamma(1-\rho+\rho\alpha)}{\alpha}\right\} ,$$

where the constant γ *is given by*

$$\gamma = \frac{(1-\rho)[\beta(z_0)A^*(z_0) - A(z_0)]}{[\beta - \rho + \lambda\beta\{E(W) + E(S_{exc})\}][A'(z_0)-1](z_0-1)} .$$

REFERENCES

1. A.G. DE KOK (1989). A moment-iteration method for approximating the waiting-time characteristics of the GI/G/1 queue. *Probability in the Engineering and Informational Sciences* 3.

2. T.S. HEINES (1983). On the behavior of buffers with random service interruptions. *IEEE Transactions on Communication* 28, 573-576.

3. S.M. ROSS (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, Inc., San Francisco.

4. H.C. TIJMS (1986). *Stochastic Modeling and Analysis: A Computational Approach*, Wiley, Chichester.

5. H.C. TIJMS, J.C.W. VAN OMMEREN (1989). Asymptotic analysis for buffer behavior in communication systems, *Probability in the Engineering and Informational Sciences* 3, 1-12.

6. J.C.W. Van Ommeren, R.D. Nobel (1989), On the Waiting-time distribution in a GI/G/1 queue with Coxian-2 services, *Statistica Neerlandica* 43, 85-90.

7. C.M. WOODSIDE, E.D.S. HO (1987). Engineering calculation of overflow probabilities in buffers with Markov interrupted service. *IEEE Transactions on Communication* 35, 1272-1277.

1989-1	O.J.C. Cornielje	A time-series of Total Accounts for the Netherlands 1978-1984
1989-2	J.C. van Ours	Self-Service Activities and Legal or Illegal Market Services
1989-3	H. Visser	The Monetary Order
1989-4	G.van der Laan A.J.J. Talman	Price Rigidities and Rationing
1989-5	N.M. van Dijk	A Simple Throughput Bound For Large Closed Queueing Networks With Finite Capacities
1989-6	N.M. van Dijk	Analytic Error Bounds For Approximations of Queueing Networks with an Application to Alternate Routing
1989-7	P.Spreij	Selfexciting Counting Process Systems with Finite State Space
1989-8	H.Visser	Rational Expectations and New Classical Macroeconomics
1989-9	J.C. van Ours	De Nederlandse Boekenmarkt tussen Stabiliteit en Verandering
1989-10	H. Tieleman A. Leliveld	Traditional "Social Security Systems" and Socio-economic Processes of Change: The Chase of Swaziland; opportunities for research
1989-11	N.M. van Dijk	"Stop = Recirculate" for Exponential Product Form Queueing Networks with Departure Blocking
1989-12	F.A.G. den Butter	Modelbouw en matigingsbeleid in Nederland
1989-13	N.M. van Dijk	Simple performance estimates and error bounds for slotted ALOHA loss systems
1989-14	H. Clemens J.P. de Groot	Sugar Crisis, a Comparison of two Small Peripheral Economies
1989-15	I.J.Steyn	Consistent Diffuse Initial Conditions in the Kalman Filter
1989-16	I.J.Steyn	Als Estimation of Parameters in a State Space Model
1989-17	B.Vogelvang	Dynamic Interrelationships between Spot Prices of some Agricultural Commodities on Related Markets
1989-18	J.C. van Ours	Zoeken naar nieuwe medewerkers
1989-19	H. Kox	Integration of Environmental Externalities in International Commodity Agreements