Zeitschrift für Operations Research, Band 21, 1977, Seite 19–32. Physica-Verlag, Würzburg.

On a Switch-Over Policy for Controlling the Workload in a Queueing System with Two Constant Service Rates and Fixed Switch-Over Costs[1])

By *H.C. Tijms*, Amsterdam[2])

Eingegangen am 11. März 1975
Revidierte Fassung eingegangen am 22. Juni 1975

Summary: This paper considers a single-server system where jobs arrive in accordance with a Poisson process. Each job involves an amount of work which is known upon arrival and is sampled from an exponential distribution. The server has available two constant service rates 1 and 2 where rate 2 is faster than rate 1. The total work remaining to be processed in the system (= workload) is controlled by a switch-over policy which switches from rate 1 to rate 2 only when the workload exceeds the level y_1 and switches from rate 2 to rate 1 only when the workload falls to the level y_2 where $0 \leqslant y_2 \leqslant y_1$. The costs of this system consist of a linear holding cost, a service-cost rate and fixed switch-over costs. The purpose of this paper is to derive an explicit expression for the average cost of this policy.

Zusammenfassung: Es wird ein einzelner Bedienungskanal betrachtet, bei dem die Belastungen gemäß einem Poisson-Prozeß eintreffen und unabhängig voneinander dieselbe Exponentialverteilung besitzen. Der Bedienungskanal verfügt über zwei verschiedene konstante Bedienungsraten. Wenn die Gesamtbelastung über einen vorgegebenen Wert y_1 steigt, so wird von der kleineren auf die größere Bedienungsrate umgeschaltet. Auf die kleinere Bedienungsrate wird zurückgeschaltet, wenn die Gesamtbelastung unter den vorgegebenen Wert y_1 fällt, wobei $0 \leqslant y_2 \leqslant y_1$ sei. Jedes Umschalten von einer Bedienungsrate auf die andere verursache fixe Kosten. Außerdem entstehen Kosten pro Zeiteinheit für die Bedienung und die Gesamtbelastung.

Die Bedienungskostenrate hängt von der jeweiligen verwendeten Bedienungsrate ab während die durch die Belastung bedingte Kostenrateproportional zur Gesamtbelastung ist. Der Zweck der Arbeit ist die Angabe eines expliziten Ausdrucks für die entstehenden durchschnittlichen Gesamtkosten.

1. Introduction

We consider a service station with a single server where jobs arrive in accordance with a Poisson process with rate λ. Each job involves an amount of work. The amounts

[1]) This paper is registered as Mathematical Centre Report BW 45/75

[2]) Dr. *H.C. Tijms*, Department of Operations Research, Mathematisch Centrum, Amsterdam 2e Boerhavestraat 49, Amsterdam 1005.

of work of the jobs are known upon arrival and are independently sampled from an exponential distribution with mean $1/\mu$. At any time the server may choose between the service rates 1 and 2. When the server is in service and uses service rate i an amount of work σ_i will be processed per unit time, $i = 1, 2$. It is assumed that $\sigma_2 > \sigma_1 > \lambda/\mu$. Define the workload at time t as the total amount of work remaining to be processed in the system at time t, $t \geqslant 0$. The server provides service when the system is not empty and uses the following switch-over policy. The server switches from rate 1 to rate 2 only when the workload exceeds the level y_1 and switches from rate 2 to rate 1 only when the workload falls to the level y_2, where y_1 and y_2 are given numbers with $0 \leqslant y_2 \leqslant y_1$. It is assumed that it takes no time to switch from one service rate to another. We denote the above switch-over policy as the (y_1, y_2) policy.

The following costs are incurred. There is a holding cost of $h > 0$ per unit work in the system per unit time. When the server is busy and uses service rate i there is a service cost at rate $r_i \geqslant 0$, $i = 1, 2$. There is a service cost at rate $r_0 \geqslant 0$ when the system is empty. The cost of switching from rate 1 (2) to rate 2 (1) is K_1 (K_2), where $K_1\, K_2 \geqslant 0$.

The purpose of this paper is to derive an explicit expression for the average cost of the (y_1, y_2) policy[3]). Roughly, this will be done as follows. We first consider a Markov chain embedded at the epochs where the server switches from one rate to another and the epochs where the system becomes empty. It will be shown that this Markov chain has a unique stationary probability distribution which can be explicitly given. Because of the existence of this distribution, a formula familiar from the theory of semi-Markov reward processes applies to the average cost. From this formula we shall derive an alternative one which allows to give an explicit expression for the average cost. This analysis will be done in the sections 3 and 4 after we have given some preparatory results in section 2. Finally, section 5 discusses the minimization of the obtained expression for the average cost.

Related work was done by *Thatcher* [1968] who studied the (y_1, y_2) policy with $y_1 = y_2$ for an M/G/1 queue with no switch-over costs. Using busy period analysis he derived for the average cost of this policy a formula involving the stationary distribution of the workload under rate 1. Also, he proved that a policy of this type is average cost optimal among the class of all stationary policies (cf. also *Mitchell* [1973]). For M/G/1 queueing system in which the control is based on the queue size, a policy similar to the (y_1, y_2) policy has been studied amongst others by *Crabill* [1972 a, 1972 b] and *Meyer* [1971]. Other related references can be found in the two papers on the optimal control of queues given in *Clarke* [1974].

2. Preliminaries

In this section we give some preparatory results. We first consider the M/G/1 queue in which jobs arrive in accordance with a Poisson process with rate λ and the amounts

[3]) The analysis given in this paper is also applicable when we assume an arbitrary distribution for the amount of work of a job. However, in this case no simple explicit results can be obtained.

of work involved by the jobs are independent, positive random variables having a common probability distribution function F with finite first moment β and finite second moment $\beta^{(2)}$. When there is work to be done in the system the server provides service where an amount of work σ is processed per unit time. It is assumed that $\lambda\beta / \sigma < 1$. Also, suppose that there is a holding cost of h per unit workload per unit time. For this queueing system, let $b(x)$ be the expectation of the first epoch at which the system is empty and let $h(x)$ be the expected total holding cost incurred up to that epoch when workload equals x at epoch 0.

Lemma 1.

For all $x > 0$,

$$b(x) = \frac{x}{\sigma(1 - \lambda\beta / \sigma)} \text{ and } h(x) = \frac{hx^2}{2\sigma(1 - \lambda\beta / \sigma)} + \frac{h\lambda\beta^{(2)}x}{2\sigma^2(1 - \lambda\beta / \sigma)^2}.$$

Proof.

For completeness, we briefly sketch a proof of these known results. For any $x > 0$ and $n = 0, 1, \ldots$, let $b_n(x)$ be the expectation of the first epoch at which te system becomes empty and let $h_n(x)$ be the expected total holding cost incurred up to that epoch given that the workload equals x at epoch 0 and that n jobs have arrived during the initial occupation time x/σ of the server. Then, for any $x > 0$,

$$b(x) = \sum_{n=0}^{\infty} b_n(x) e^{-\lambda x/\sigma} \frac{(\lambda x/\sigma)^n}{n!} \text{ and } h(x) = \sum_{n=0}^{\infty} h_n(x) e^{-\lambda x/\sigma} \frac{(\lambda x/\sigma)^n}{n!}. \quad (1)$$

Also, let:

$$B = \int_0^{\infty} b(x)\, dF(x) \text{ and } H = \int_0^{\infty} h(x)\, dF(x), \quad (2)$$

i.e. B is the expected length of a busy period generated by a single job and W is the expected total holding costs incurred during such a busy period. Now, using the fact that both the time during which the server is busy and the workload are independent of the order in which the jobs are served and observing that each job arriving during the initial occupation time x/σ generates a busy period, some reflections show that, for all $x > 0$ and $n > 0$,

$$b_n(x) = \frac{x}{\sigma} + nB \text{ and } w_n(x) = \frac{hx^2}{2\sigma} + \frac{hn\beta x}{2\sigma} + nW + h \sum_{k=1}^{n} (n - k)\beta B,$$

where the latter relation uses the fact that under the condition that n arrivals have occured in $(0, x/\sigma)$ each of the n arrival epochs has expectation $x/2\sigma$ as follows from Theorem 2.3 of *Ross* [1970]. Using (1) and (2) we obtain the desired results after some algebra.

We now return to the queueing system introduced in section 1. The state of this system can be described by a point in $\{(x, 1)|x \geqslant 0\} \cup \{(x, 2)|x \geqslant 0\}$, where state (x, i) corresponds to the situation that the workload equals x and the server is adjusted to rate i. We now introduce a number of functions that will be needed hereafter. These functions are defined independently of the (y_1, y_2) policy. For any $x > 0$ and $i = 1, 2$, define $t_i(x)$ as the expectation of the first epoch at which the system becomes empty and define $k_i(x)$ as the expected total cost incurred up to that epoch when the system is in state (x, i) at epoch 0 and the server always uses rate i. Using lemma 1 with $\beta = 1/\mu$ and $\beta^{(2)} = 2/\mu^2$, it follows that, for all $x > 0$ and $i = 1, 2$,

$$t_i(x) = \frac{\mu x}{(\sigma_i \mu - \lambda)} \text{ and } k_i(x) = \frac{h\mu x^2}{2(\sigma_i \mu - \lambda)} + \frac{h\lambda x}{(\sigma_i \mu - \lambda)^2} + \frac{r_i \mu x}{(\sigma_i \mu - \lambda)}. \tag{3}$$

We note that the assumption $\lambda/\sigma_1 \mu < 1$ is needed to ensure that the functions $t_1(x)$ and $k_1(x)$ are well defined and finite. The assumption $\lambda/\sigma_1 \mu < 1$ will only be used for this purpose. Next, let

$$\alpha_0 = \frac{r_0}{\lambda} + \int_0^\infty k_1(x)\, \mu e^{-\mu x} dx \text{ and } \beta_0 = \frac{1}{\lambda} + \int_0^\infty t_1(x)\, \mu e^{-\mu x} dx, \tag{4}$$

that is, β_0 is the expected time until the next return of the system to state (0, 1) and α_0 is the expected total cost incurred during this time when the initial state is (0, 1) and the server always uses rate 1. Then, by (3),

$$\alpha_0 = \frac{r_0 - r_1}{\lambda} + \frac{r_1 \sigma_1 \mu}{\lambda(\sigma_1 \mu - \lambda)} + \frac{h\sigma_1}{(\sigma_1 \mu - \lambda)^2} \text{ and } \beta_0 = \frac{\sigma_1 \mu}{\lambda(\sigma_1 \mu - \lambda)}.$$

Finally, define the functions $k(x)$ and $t(x)$ by

$$k(x) = k_2(x) - k_1(x) \text{ and } t(x) = t_2(x) - t_1(x) \text{ for } x > 0. \tag{5}$$

Then, by (3),

$$k(x) = \alpha_1 x^2 + \alpha_2 x \text{ and } t(x) = \beta_1 x \qquad \text{for } x > 0, \tag{6}$$

where

$$\alpha_1 = \frac{h\mu^2(\sigma_1 - \sigma_2)}{2(\sigma_1 \mu - \lambda)(\sigma_2 \mu - \lambda)}, \tag{7}$$

$$\alpha_2 = \frac{h\lambda}{(\sigma_2 \mu - \lambda)^2} - \frac{h\lambda}{(\sigma_1 \mu - \lambda)^2} + \frac{r_2 \mu}{(\sigma_2 \mu - \lambda)} - \frac{r_1 \mu}{(\sigma_1 \mu - \lambda)}, \tag{8}$$

$$\beta_1 = \frac{\mu^2(\sigma_1 - \sigma_2)}{(\sigma_1 \mu - \lambda)(\sigma_2 \mu - \lambda)} \tag{9}$$

By direct integration, for all $y > 0$,

$$\int_y^\infty k(x)\,\mu e^{-\mu x}dx = e^{-\mu y}\,\{k(y) + \alpha_3 y + (\alpha_2 + \alpha_3)/\mu\}, \tag{10}$$

$$\int_y^\infty t(x)\,\mu e^{-\mu x}dx = e^{-\mu y}\,\{t(y) + \beta_1/\mu\}, \tag{11}$$

$$\alpha_3 = \frac{h\mu(\sigma_1 - \sigma_2)}{(\sigma_1\mu - \lambda)(\sigma_2\mu - \lambda)}\,. \tag{12}$$

To end this section, we give some required results for a Markov chain with a general state space. Consider a Markov chain $\{X_n,\ n = 0, 1, \ldots\}$ with stationary transition probability function $P(\cdot,\cdot)$ on $(S, \mathcal{B})$, where the state space S is a Borel set of a finite dimensional Euclidean space and $\mathcal{B}$ is the class of all Borel sets in S. Suppose that this Markov chain satisfies the following assumption.

Assumption.

There is some state s^* (say) such that

$$Pr\,\{X_n = s^* \text{ for some } n \geqslant 1 \mid X_0 = s\} = 1 \qquad \text{for all } s \in S, \tag{13}$$

and

$$E\,(N|X_0 = s^*) < \infty \text{ where } N = \inf\,\{n \geqslant 1 \mid X_n = s^*\}. \tag{14}$$

We have the following theorem whose proof is included for completeness.

Theorem 1.

There is a unique stationary probability distribution function Q satisfying

$$Q(A) = \int_S P(s, A)\,Q(ds) \qquad \textit{for all } A \in \mathcal{B}. \tag{15}$$

Moreover, when the initial state $X_0 = s^*$, then

$$\lim_{n\to\infty} \frac{1}{n} E\,\{\sum_{k=0}^{n} f(X_k)\} = \int_S f(s)\,Q(ds) \tag{16}$$

for any real-valued Baire function f *such that* $\int |f(s)|\,Q(ds) < \infty$.

Proof.

For any $n \geqslant 0$, let $P^n(s, A) = Pr\{X_n \in A \mid X_0 = s\}$. Further, for any $n \geqslant 1$, let $\tilde{P}^n(s, A) = Pr\{X_n \in A, X_k \neq s^* \text{ for } 1 \leqslant k \leqslant n \mid X_0 = s\}$, and let $\tilde{P}^0(s, A) = P^0(s, A)$. Define $f_n(s) = Pr\{N = n \mid X_0 = s\}$ for $n \geqslant 1$, and define $f_0(s) = 0$. Then (cf. *Feller* [1966, p. 365]), for all s and A.

$$P^n(s, A) = \tilde{P}^n(s, A) + \sum_{k=0}^{n} P^{n-k}(s^*, A) f_k(s) \quad \text{for all } n \geqslant 0. \tag{17}$$

By (13), $\sum_{0}^{\infty} f_n(s^*) = 1$. Hence the relation (17) with $s = s^*$ is a renewal equation for any A. Further, for any A,

$$\sum_{n=0}^{\infty} \tilde{P}^n(s^*, A) \leqslant \sum_{n=0}^{\infty} \tilde{P}^n(s^*, S) = E(N \mid X_0 = s^*) = \sum_{n=0}^{\infty} n f_n(s^*), \tag{18}$$

so, by (14), both the first series and the last series in (18) are convergent. Now, by applying the Key Renewal Theorem (see *Feller* [1957, p. 292], for any A.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} P^k(s^*, A) = \sum_{n=0}^{\infty} \tilde{P}^n(s^*, A) \Big/ \sum_{n=0}^{\infty} n f_n(s^*). \tag{19}$$

Now, for any A, define $Q(A)$ as the right side of (19). Then, by (18), Q is a probability measure. Next observe that, by (13), $\sum_{0}^{\infty} f_n(s) = 1$ and $\tilde{P}^n(s, A) \to 0$ as $n \to \infty$ for all s and A. Using this we obtain from (17) and (19) that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} P^k(s, A) = Q(A) \qquad \text{for all } s \in S \text{ and } A \in \mathcal{B},$$

from which it is easy to derive that Q satisfies the steady state equation (15) (cf. *Breiman* [1968, pp. 133–134]). Since the Markov chain X_n has no two disjoint closed sets, we have by Theorem 7.16 in *Breiman* [1968] that Q is the unique probability distribution satisfying (15). To prove (16), let m be a finite measure on $(S, \mathcal{B})$ such that $m(A) > 0$ if and only if $s^* \in A$. Then, by (13), $m(A) > 0$ implies $Pr\{X_n \in A \text{ for some } n \geqslant 1 \mid X_0 = s\} = 1$ for all $s \in S$. Consequently, the Markov chain $\{X_n\}$ satisfies the so-called recurrence condition of Harris (cf. *Jain* [1966, pp. 206–207]). Relation (15) now follows from Theorem 3.3 in *Jain* [1966].

3. An Embedded Markov Chain

In this section we shall determine the stationary probability distribution of the Markov chain embedded at the epochs where the server switches from one rate to another and the epochs where the system becomes empty.

Consider the queueing system which is controlled by an (y_1, y_2) policy with $0 < y_2 \leqslant y_1$ (the (y_1, y_2) policy with $y_2 = 0$ will be considered separately in the next section). For ease we assume from now on that the system is empty at epoch 0. Let $T_0 = 0$, and, for $n \geqslant 1$, let T_n be the n-th epoch at which either the server switches from one rate to another or the system becomes empty. For any $n \geqslant 0$, define X_n as the state of the system at epoch T_n with the convention that we take X_n equal to $(x, 1)$ $[(x, 2)]$ when at epoch T_n the workload equals x and the server switches from rate 1 [2] to rate 2 [1]. Observe that $X_0 = (0, 1)$. The embedded process $\{X_n, n \geqslant 0\}$ is a Markov chain with state space

$$S = \{(0, 1)\} \cup \{(x, 1) \mid x > y_1\} \cup \{(y_2, 2)\}.$$

Denote by $P(\cdot, \cdot)$ the one-step transition probability function of this Markov chain, that is, $P(s, A) = Pr\{X_n \in A \mid X_{n-1} = s\}$. For the above Markov chain the assumption of Theorem 1 is satisfied for $s^* = (0, 1)$, so, this Markov chain has a unique stationary probability distribution $Q(\cdot)$ (say) satisfying (15). This stationary distribution Q will now be determined explicitly. To do this, define, for all $0 < x < y_1$ and $v \geqslant y_1$,

$p(x, v)$ = probability that the state of the first entry of the system into the set of states $\{(0, 1)\} \cup \{(u, 1) \mid u > y_1\}$ belongs to the set $\{(u, 1) \mid u > v\}$ when the initial state is $(x, 1)$.

Further, let $p_0(x) = 1 - p(x, y_1)$ for $0 < x < y_1$. For shortness we write $Q_0 = Q(\{(0, 1)\})$, $Q(v) = Q(\{(u, 1) \mid u > v\})$ and $Q_2 = Q(\{(y_2, 2)\})$. Then, (15) gives

$$Q_0 = Q_2 p_0(y_2) + Q_0 \int_0^{y_1} p_0(y)\, \mu e^{-\mu y} dy,\ Q_2 = Q(y_1), \text{ and}$$

$$Q(v) = Q_2 p(y_2, v) + Q_0 \{e^{-\mu v} + \int_0^{y_1} p(y, v)\, \mu e^{-\mu y} dy\} \text{ for all } v \geqslant y_1.$$

Further, by $Q_0 + Q(y_1) + Q_2 = 1$, we have $Q_2 = (1 - Q_0)/2$. We shall now determine $p(x, v)$. Using a standard argument, we get for any $v \geqslant y_1$ and Δx very small,

$$p(x + \Delta x, v) = \frac{\lambda \Delta x}{\sigma_1} \left\{ \int_0^{y_1 - x} p(x + y, v)\, \mu e^{-\mu y} dy + e^{-\mu (v - x)} \right\} +$$

$$+ \left(1 - \frac{\lambda \Delta x}{\sigma_1}\right) p(x, v) + o(\Delta x) \quad \text{for } 0 < x < y_1,$$

which implies that, for all $v \geqslant y_1$,

$$\frac{\partial p(x, v)}{\partial x} = \frac{\lambda}{\sigma_1}\left[-p(x, v) + \int_0^{y_1 - x} p(x + y, v)\, \mu e^{-\mu y} dy + e^{-\mu(v-x)}\right]$$
$$\text{for } 0 < x < y_1.$$

Routine analysis using Laplace transforms and the boundary condition $p(x, v) \to 0$ as $x \to 0$ yields after some algebra

$$p(x, v) = \lambda e^{-(\sigma_1 \mu v - \lambda y_1)/\sigma_1}\left[e^{(\sigma_1 \mu - \lambda)x/\sigma_1} - 1\right]\left[\sigma_1 \mu - \lambda e^{-(\sigma_1 \mu - \lambda) y_1/\sigma_1}\right]^{-1}$$

for all $0 < x < y_1$ and $v \geqslant y_1$. From $p_0(x) = 1 - p(x, y_1)$,

$$p_0(x) = \left[\sigma_1 \mu - \lambda e^{-(\sigma_1 \mu - \lambda)(y_1 - x)/\sigma_1}\right]\left[\sigma_1 \mu - \lambda e^{-(\sigma_1 \mu - \lambda) y_1/\sigma_1}\right]^{-1}$$
$$\text{for } 0 < x < y_1.$$

The formula for $p_0(x)$ was also found in *Keilson* [1963]. Using these results we get after some algebra

Theorem 2.
The stationary distribution Q *is given by*

$$Q_0 = c^{-1}\left\{\sigma_1 \mu - \lambda e^{-(\sigma_1 \mu - \lambda)(y_1 - y_2)/\sigma_1}\right\}$$

$$Q_2 = c^{-1}(\sigma_1 \mu - \lambda) e^{-(\sigma_1 \mu - \lambda) y_1/\sigma_1}$$

$$q(v) = c^{-1}\mu(\sigma_1 \mu - \lambda) e^{-(\sigma_1 \mu v - \lambda y_1)/\sigma_1}$$

for all $v \geqslant y_1$, *where* $q(v) = -\partial Q(v)/\partial v$, *and*

$$c = \sigma_1 \mu + (2\sigma_1 \mu - 2\lambda) e^{-(\sigma_1 \mu - \lambda) y_1/\sigma_1} - \lambda e^{-(\sigma_1 \mu - \lambda)(y_1 - y_2)/\sigma_1}$$

Remark 1.
For the case where the amount of work of a job has an arbitrary distribution function F the resulting differential equation for $p(x, v)$ can be converted into a delayed renewal equation by integration, and this fact allows to give a closed expression for $p(x, v)$ in which the renewal function of the defective distribution function $(\lambda/\sigma_1 \int_0^y \{1 - F(u)\}\, du$ appears, cf. *Cohen* [1974]. Hence Q can be explicitly given in terms of this renewal function.

4. The Average Cost of the (y_1, y_2) Policy

In this section we shall derive an explicit expression for the average cost of the (y_1, y_2) policy. To get to this expression, we first establish a formula which is familiar from the theory of semi-Markov processes with a cost structure. Next we derive from this formula an alternative one which allows to give an explicit for the average cost of the (y_1, y_2) policy.

Consider in the first instance the (y_1, y_2) policy with $0 < y_2 \leqslant y_1$. Let $Z(t)$ be the total cost incurred during $[0, t)$, $t \geqslant 0$. For any $n \geqslant 0$, let $\tau_n = T_{n+1} - T_n$, i.e., τ_n is the length of the time interval between the n-th and the $(n+1)$st epoch at which either the server switches from one rate to another or the system becomes empty. Further, for any $n \geqslant 0$, denote by Z_n the total cost incurred during $[T_n, T_{n+1})$, where Z_n includes the appropriate switch-over cost when at epoch T_n the server switches from one rate to another. Finally, let $\tau(s) = E(\tau_n | X_n = s)$, and let $c(s) = E(Z_n | X_n = s)$ for $s \in S$.

Lemma 2.

$$\lim_{t \to \infty} \frac{1}{t} EZ(t) = \int_S c(s)\, Q(ds) \Big/ \int_S \tau(s)\, Q(ds). \tag{20}$$

Proof.

We first observe that the process describing the behaviour of the state of the system is regenerative where the epochs at which the system becomes empty are regeneration epochs. There is a cost structure imposed on the process. Now, since both the expected time until the first return of the system to state (0, 1) and the expected cost incurred during this time are finite, we have by the renewal theoretic argument used in the proof of Theorem 7.5 of *Ross* [1970],

$$\lim_{t \to \infty} \frac{1}{t} EZ(t) = \lim_{n \to \infty} \frac{1}{n} E\left\{\sum_{k=0}^{n-1} Z_k\right\} \Big/ \lim_{n \to \infty} \frac{1}{n} E\left\{\sum_{k=0}^{n-1} \tau_k\right\}.$$

Next the Lemma follows from Theorem 1 (using formula (3) it is immediate from their definitions that the functions $\tau(s)$ and $c(s)$ are bounded by a linear and quadratic function, respectively, so , by Theorem 2, both integrals in (20) are absolutely convergent). □

Remark 2.

By Theorem 3.16 of *Ross* [1970], we also have that, with probability 1, $Z(t)/t$ converges to the right side of (20) as $t \to \infty$.

We shall now convert formula (20) into an alternative form which allows to give an explicit expression for the average cost of the (y_1, y_2) policy. To do this, recall that $k_2(x)$ has been defined as the expected cost incurred until the system is empty when the initial state is $(x, 2)$ and the server always uses rate 2 (see section 2), and, so ,

$K_1 + k_2\ (x)$ represents the expected cost incurred until the system is empty when in the initial state $(x, 1)$ the server was to switch to rate 2 and always remains using rate 2. From these interpretations and the definition of $c\ ((x, 1))$ it now follows that

$$K_1 + k_2\ (x) = c\ ((x, 1)) + k_2\ (y_2) \qquad \text{for all } x > y_1. \tag{21}$$

Let $S_1 = \{(x, 1) \mid x > y_1\}$. Similarly, it is easily seen that

$$K_2 + k_1\ (y_2) = c\ ((y_2, 2)) + \int_{S_1} k_1\ (x)\, P((y_2, 2), ds) \tag{22}$$

$$\frac{r_0}{\lambda} + \int_0^\infty k_1\ (y)\, \mu e^{-\mu y} dy = c\ ((0, 1) + \int_{S_1} k_1\ (x)\, P((0, 1), ds) \tag{23}$$

where $P(\cdot\ , \cdot)$ is the one-step transition probability function of the embedded Markov chain introduced in section 3. For notational convenience, we now introduce functions $h_1\ (s)$ and $h_2\ (s)$, $s \in S$. Let $h_1\ ((x, 1)))$ be equal to the left side of (21) for $x > y_1$, let $h_1\ ((y_2, 1))$ be equal to the left side of (22), and let $h_1\ ((0,1))$ be equal to the left side of (23). Further, let $h_2\ ((x, 1)) = k_1\ (x)$, let $h_2\ ((y_2, 2)) = k_2\ (y_2)$, and let $h_2\ ((0, 1)) = 0$. Then, together (21) – (23) can be summarized as

$$h_1\ (s) = c\ (s) + \int_S h_2\ (w)\, P\ (s, dw) \qquad \text{for all } s \in S. \tag{24}$$

Integrating both sides of (24) with respect to the stationary distribution Q and using (15), we get after an interchange of the order of integration (it is immediate to verify that all integrals are absolutely convergent, since any function involved is bounded by a quadratic function),

$$\int_S h_1\ (s)\, Q\ (ds) = \int_S c\ (s)\, Q\ (ds) + \int_S h_2\ (w)\, Q\ (dw),$$

from which we get by using (4), (5) and Theorem 2,

$$\int_S c\ (s)\, Q\ (ds) = \int_S \{h_1\ (s) - h_2\ (s)\}\, Q\ (ds) =$$

$$= \alpha_0 0_0 + \int_{y_1}^\infty \{K_1 + k\ (x)\}\, q\ (x)\, dx + \{K_2 - k\ (y_2)\}\, Q_2. \tag{25}$$

In the same way, we obtain

$$\int_S \tau\ (s)\, Q\ (ds) = \beta_0 Q_0 + \int_{y_1}^\infty t\ (x)\, q\ (x)\, dx - t\ (y_2)\, Q_2 \tag{26}$$

(this relation can also be directly obtained from (25) by putting $r_0 = r_1 = r_2 = 1$ and $K_1 = K_2 = h = 0$, and noting that for these values the cost functions $c(\cdot)$ and $k_1(\cdot)$ reduce to the corresponding time functions $\tau(\cdot)$ and $t_1(\cdot)$). Now, by Lemma 2 and the relations (25) and (26), the average cost of the (y_1, y_2) policy with $0 < y_2 \leqslant y_1$ is given by the formula[4])

$$g(y_1, y_2) = \frac{\alpha_0 Q_0 + \int_{y_1}^{\infty} \{K_1 + k(x)\} q(x)\, dx + \{K_2 - k(y_2)\} Q_2}{\beta_0 Q_0 + \int_{y_1}^{\infty} t(x)\, q(x)\, dx - t(y_2)\, Q_2}$$

Using the relations (6), (10) and (11) and Theorem 2, we find

Theorem 3.

For any (y_1, y_2) policy with $0 < y_2 \leqslant y_1$ *the average (expected) cost per unit time is given by*

$$g(y_1, y_2) = \frac{\alpha_0 R(y_1, y_2) + \alpha_1 (y_1^2 - y_2^2) + \alpha_2 (y_1 - y_2) + \alpha_3 y_1 + (\alpha_2 + \alpha_3)/\mu + K}{\beta_0 R(y_1, y_2) + \beta_1 (y_1 - y_2) + \beta_1/\mu} \tag{27}$$

where $K = K_1 + K_2$ and

$$R(y_1, y_2) = (\sigma_1 \mu - \lambda)^{-1} \left\{ \sigma_1 \mu e^{(\sigma_1 \mu - \lambda) y_1/\sigma_1} - \lambda e^{(\sigma_1 \mu - \lambda) y_2/\sigma_1} \right\}.$$

Remark 3.

The above formula for the average cost holds also for the (y_1, y_2) policy with $y_2 = 0$. This result which will be intuitively clear from continuity considerations follows by considering the process embedded at points in time where either the server switches from rate 1 to rate 2 or the system becomes empty and by repeating the above analysis with obvious modifications.

Remark 4.

Consider the case of $K = 0$. Denote an (y_1, y_1) policy with $y_1 = y_2$ by the y-policy. Then, the average cost of an y-policy is given by

$$g(y) = \frac{\alpha_0 e^{(\sigma_1 \mu - \lambda) y/\sigma_1} + \alpha_3 y + (\alpha_2 + \alpha_3)/\mu}{\beta_0 e^{(\sigma_1 \mu - \lambda) y/\sigma_1} + \beta_1/\mu} \tag{28}$$

This formula agrees with the results in *Thatcher* [1968, p.78].

---

[4]) The idea used to derive this formula from (2) is generally applicable and a sophisticated use of it has been made in the Markov decision model considered in *De Leve and Tijms* [1974].

Remark 5.
The average cost of the policy that always uses rate i equals

$$g_i = r_0 \left(1 - \frac{\lambda}{\sigma_i \mu}\right) + \frac{r_i \lambda}{\sigma_i \mu} + \frac{h\lambda}{\mu(\sigma_i \mu - \lambda)} \qquad \text{for } i = 1, 2 \tag{29}$$

as follows by putting $\sigma_1 = \sigma_2$, $r_1 = r_2$ and $K = 0$ in (27). Observe that $g(0) = g_2$, however $g(0, 0) > g_2$ when $K > 0$. Also, observe that, for any y_2, $g(y_1, y_2)$ converges to $\alpha_0 / \beta_0 = g_1$ as $y_1 \to \infty$.

5. Minimization of the Average Cost

This section discusses the determination of the numbers y_1^* and y_2^* for which the average cost function $g(y_1, y_2)$ is minimal. We shall distinguish between the cases $K = 0$ and $K > 0$. First we consider

Case 1.
$K = 0$. For this case of no switch-over costs we only consider the y-policies (as shown by *Thatcher* [1968] a policy of this type is average cost optimal among the class of all stationary policies). We find after some algebra that the derivative of the average cost function $g(y)$ has the same sign as the function

$$h(y) = y + \lambda(\sigma_1 - \sigma_2)(\sigma_1 \mu - \lambda)^{-1} (\sigma_2 \mu - \lambda)^{-1} \{1 - e^{-(\sigma_1 \mu - \lambda) y / \sigma_1}\} - a,$$

where $a = (h\sigma_1 \mu)^{-1} (\sigma_1 \mu - \lambda) \{r_0 + (r_2 \sigma_1 - r_1 \sigma_2)/(\sigma_2 - \sigma_1)\}$. It is immediate to verify that $h(0) = -a$ and that $h(y)$ is strictly increasing for $y \geq 0$ with $h(y) \to \infty$ as $y \to \infty$. Hence, if $a > 0$, then the average cost is minimal for the y^*-policy where y^* is the unique positive root to the equation $h(y) = 0$. If $a \leq 0$, then $g(y)$ is minimal for $y^* = 0$, that is, the policy that always uses rate 2 minimizes the average cost. In table 1 we give the optimal y^* and $g(y^*)$ for a number of numerical examples.

Case 2.
$K > 0$. For this case we find after some algebra

$$\frac{\partial g(y_1, y_2)}{\partial y_1} = \phi(y_1, y_2) \Big[\mu e^{(\sigma_1 \mu - \lambda) y_1 / \sigma_1} \left\{ \gamma_1 (y_1^2 - y_2^2) + \gamma_2 (y_1 - y_2) - \frac{K \sigma_1 (\sigma_2 \mu - \lambda)}{\lambda \mu} \right\}$$

$$+ \frac{\lambda}{(\sigma_1 \mu - \lambda)} \left\{ e^{(\sigma_1 \mu - \lambda) y_2 / \sigma_1} - e^{(\sigma_1 \mu - \lambda) y_1 / \sigma_1} \right\} \left\{ 2\gamma_1 y_1 + \frac{2\gamma_1}{\mu} + \gamma_2 \right\}$$

$$+ \frac{\gamma_3}{2} (y_1 - y_2)^2 + \frac{\gamma_3}{\mu} (y_1 - y_2) - K(\sigma_1 - \sigma_2) \Big]$$

and

$$\frac{\partial g(y_1, y_2)}{\partial y_2} = \phi(y_1, y_2)\Big[\frac{-\lambda}{\sigma_1} e^{(\sigma_1\mu-\lambda)y_2/\sigma_1}\Big\{\gamma_1(y_1^2 - y_2^2) + \gamma_2(y_1 - y_2) + \frac{2\gamma_1}{\mu} y_1$$

$$+ \frac{2\gamma_1}{\mu^2} + \frac{\gamma_2}{\mu} - \frac{K\sigma_1(\sigma_2\mu-\lambda)}{\lambda\mu}\Big\} + R(y_1, y_2)(2\gamma_1 y_1 + \gamma_2) + \frac{\gamma_3}{2}(y_1 - y_2)^2$$

$$+ \frac{\gamma_3}{\mu}(y_1 - y_2) + \frac{\gamma_3}{\mu^2} + K(\sigma_1 - \sigma_2)\Big],$$

where

$$\phi(y_1, y_2) = \mu^2(\sigma_1\mu-\lambda)^{-1}(\sigma_2\mu-\lambda)^{-1}\{\beta_0 R(y_1, y_2) + \beta_1(y_1 - y_2) + \beta_1/\mu\}^{-2}$$

$$\gamma_1 = \frac{-h\sigma_1\mu(\sigma_1-\sigma_2)}{2\lambda(\sigma_1\mu-\lambda)}, \quad \gamma_2 = \frac{(r_0 - r_1)(\sigma_1-\sigma_2)}{\lambda} + \frac{(r_1 - r_2)\sigma_1}{\lambda} - \frac{h\sigma_1(\sigma_1-\sigma_2)}{(\sigma_1\mu-\lambda)(\sigma_2\mu-\lambda)}$$

$$\gamma_3 = \frac{h\mu^2(\sigma_1-\sigma_2)^2}{(\sigma_1\mu-\lambda)(\sigma_2\mu-\lambda)}.$$

Table 1: $\mu = 2, \sigma_1 = 4, \sigma_2 = 5, h = 1, r_0 = 0, r_1 = 5$ and $r_2 = 10$

	λ	6	6.5	7	7.5	7.75
$K = 0$	y^*	4.418	3.747	3.146	2.605	2.353
	$g(y^*)$	5.168	5.925	6.812	7.855	8.450
$K = 10$	y_1^*	11.066	9.509	8.194	7.097	6.606
	y_2^*	3.108	2.209	1.463	0.878	0.636
	$g(y_1^*, y_2^*)$	5.237	6.121	7.226	8.541	9.270
$K = 25$	y_1^*	14.678	12.462	10.611	9.143	8.520
	y_2^*	3.024	2.016	1.155	0.496	0.234
	$g(y_1^*, y_2^*)$	5.247	6.181	7.429	8.979	9.838
	g_2	6.750	7.429	8.167	9.000	9.472

Observe that $\partial g(\bar{y}, \bar{y})/\partial y_1 < 0$ for all $\bar{y}$, so, for each point (y_1^*, y_2^*) minimizing the function $g(y_1, y_2)$ for $0 \leqslant y_2 \leqslant y_1$ holds $y_2^* < y_1^*$. Also observe that, for each y_2, the partial derivative $\partial g(y_1, y_2)/\partial y_1$ is positive for all y_1 sufficiently large and, so, $g(y_1, y_2)$ converges to g_1 from below as $y_1 \to \infty$ which proves that the policy which always uses rate 1 is not average cost optimal (of course, this conclusion also applies to the case of $K = 0$). For the numerical computation of the minimum of the function $g(y_1, y_2)$ for $0 \leqslant y_2 \leqslant y_1$, we have used a computer program based on the variable metric algorithm of *Fletcher* [1970] for unconstrained minimization. In table 1 we give for a number of numerical examples the numbers y_1^* and y_2^* for which the func-

tion $g(y_1, y_2)$ is minimal for $0 \leqslant y_2 \leqslant y_1$ (numerical computations indicate that the function $g(y_1, y_2)$ has a single minimum, although this function is not convex). We note that $g(y_1^*, y_2^*)$ should be compared with g_2, since the average cost of the policy that always uses rate 2 may be less than that of any (y_1, y_2) policy. Finally, we note that it is reasonable to conjecture that either an (y_1, y_2) policy with $y_2 < y_1$ or the policy that always uses rate 2 is a average cost optimal among the class of all possible policies.

Acknowledgment

I would like to thank Mr. *R. van der Horst*, computer programmer at the Mathematisch Centrum, for numerical assistance.

References

Breiman, L.: Probability, Addison-Wesley, Reading, Massachusetts 1968.

Clarke, A.B. (ed.): Mathematical Methods in Queuing Theory, Lecture Notes in Economics and Mathematical Systems 98, Springer-Verlag, Berlin 1974.

Cohen, J.W.: A Simple Derivation of the Distribution of the Supremum of the Virtual Waiting Time during a Busy Period of an M/G/1 Queue, Report of the Mathematical Institute, University of Utrecht, Utrecht 1974.

Crabill, T.B.: Optimal Control of a Service Facility with Variable Exponential Service Time and Constant Arrival Rate, Management Science 18, 560–566, 1972.

–: Optimal Hysteretic Control of a Stochastic Service System with Variable Service Rates and Fixed Switch-Over Costs, University of North Carolina 1972.

De Leve, G., and *H.C. Tijms: A* General Markov Decision Method, with Applications to Controlled Queueing Systems, Report BN 24/74, Mathematisch Centrum, Amsterdam 1974.

Feller, W.: An Introduction to Probability Theory and its Applications Vol. 1 (2nd ed.), Wiley, New York 1957.

–: An Introduction to Probability Theory and its Applications Vol. 2, Wiley, New York 1966.

Fletcher, R.: A New Approach to Variable Metric Algorithms, Computing Journal 13, 317–322, 1970.

Jain, N.C.: Some Limit Theorems for General Markov Processes, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 6, 206–223, 1966.

Keilson, J.: A Gambler's Ruin Problem in Queueing Theory, Operations Research 11, 570–576, 1963.

Meyer, K.H.F.: Wartesysteme mit Variabler Bearbeitungsrate, Lecture Notes in Economics and Mathematical Systems Vo. 61, Springer-Verlag, Berlin 1971.

Mitchell, B.: Optimal Service-Rate Selection in an M/G/1 Queue, Siam Journal of Applied Mathematics 24, 19–35, 1973.

Ross, S.M.: Applied Probability Models with Optimization Applications, Holden-Day, Inc., San Francisco 1970.

Thatcher, R.M.: Optimal Single-Channel Service Policies for Stochastic Arrivals, Report ORC 68–16, Operations Research Center, University of California, Berkeley 1968.