Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

J.H. van Schuppen

Overload control for an SPC telephone exchange
An optimal stochastic control approach

Department of Operations Research and System Theory Report OS-R8404 February

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

OVERLOAD CONTROL FOR AN SPC TELEPHONE EXCHANGE - AN OPTIMAL STOCHASTIC CONTROL APPROACH

J.H. VAN SCHUPPEN
Centre for Mathematics and Computer Science, Amsterdam

The current stored program control (SPC) telephone exchanges are the operational units of the telephone networks. One of the problems with these exchanges is the performance degradation during time periods of peak demand. The problem of overload control is then to maximize the number of admitted and succesfully completed calls under technical constraints of which the main one is the available processor capacity. In the paper the processor load of a SPC telephone exchange is modelled as a hierarchical queueing system while the problem of overload control is formulated as an optimal stochastic control problem. The latter problem is solved. An implementation of the derived control law will be suggested.

1980 MATHEMATICS SUBJECT CLASSIFICATION: 93E20, 90B22, 60K30, 69L20.
KEY WORDS & PHRASES: stored program control exchange, overload control, queueing theory, stochastic control.
NOTE: This report will be submitted for publication elsewhere.

Report OS-R8404
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1. INTRODUCTION

The purpose of this paper is to present a mathematical model for the processor load of a telephone exchange, to formulate the problem of overload control as a stochastic control problem, to solve the latter problem, and to suggest ways to implement the solution.

Telephone exchanges are the main operational units of telephone networks. The currently installed telephone exchanges are computer controlled and are called *Stored Program Control (SPC) Exchanges*. In such an exchange the operations are executed by a processor according to a stored program. The operations of such an exchange may be summarized as follows. If a customer takes up the receiver this signal is noticed by the exchange. Possibly after some delay, the exchange answers by sending a dial tone. After the customer has dialed the desired number, the exchange establishes the connection with either the requested local phone or with another telephone exchange in the network.

The objective of the operation of a telephone exchange is to maximize the number of admitted and succesfully completed calls without too much delay. A call will be termed *succesful* if it reaches a ringing or busy signal, or if it reaches another telephone exchange in the network [7].

One of the problems with the operation of a telephone exchange is that its performance can degrade considerably during time periods of peak demand. During such time periods the response time of the exchange is relatively long. This causes impatient customers to dial prematurely, before a dial tone has been given, after which an incompletely received telephone number takes up processor capacity and ends up as an unsuccesful call. Other requests for connections that have been transmitted properly to the exchange, may encounter long processing delays. This then causes customers to abandon the call and, possibly, to attempt to redial. In this case too capacity of the exchange is wasted. Empirical data [7] on customer behaviour indicate that when delays are long prematurely dialed calls may exceed 20% of the call request, while abandoned calls may exceed 40% of the call requests.

The *problem of overload control* is then to maximize the number of admitted and succesful calls, especially during periods of peak demand.

A control variable is the access of a call request to the telephone exchange. A call request may be either admitted or refused access. A technical constraint in this problem is that the available processor capacity is limited. For references on this problem see [6,7,10,13,14,15].

Basically two approaches to the problem of overload control have been considered. The first approach is to propose a control algorithm based on engineering experience. This algorithm is then evaluated theoretically and through simulations. One such control algorithm is to limit access to the exchange when the number of calls being processes exceeds a certain number, and to start readmitting calls only if this quanity falls below another level. A second algorithm [7] in the same approach consists of a technical structure and a last-in-first-out control law.

The second approach to the problem of overload control is to synthesize a control algorithm via optimal control theory. This approach has been proposed by F.C. Schoute [13,14,15], and been worked out for a discrete-time queueing model. The simulation results for this case indicate a significant increase in succesfully handled calls, and encourage further research in this direction. The background for this approach is system and control theory in which the concept of state and synthesis of algorithms play a central role. This should be seen in contrast with queueing theory in which analysis dominates the discussion.

The approach presented in this paper is an extension of that of F.C. Schoute [14]. The load of the central processor of a telephone exchange will be modelled as a continuous-time hierarchical queueing system. The problem of overload control will then be formulated as an optimal stochastic control problem. The solution to this stochastic control problem will be derived. Ways to implement the solution will also be suggested. The approach of this paper differs from that of F.C. Schoute [14] in that continuous-time queueing systems are considered which makes some unverifiable assumptions of [14] unnecessary, and that an optimal stochastic control problem is formulated.

Throughout the paper use will be made of the theory of stochastic integrals and of stochastic differential equations; for references see [5,9]. For an elementary introduction see [4]. Knowledge of stochastic control theory is helpful, but necessary only for the proofs.

A brief summary of the paper follows. In the next section the model is

proposed and evaluated. The problem of overload control is formulated as a stochastic control problem in section 3. In section 4 this problem is solved and interpreted. Ways to implement the solution are suggested in section 5.

Acknowledgements are due to F.C. Schoute for useful discussions on the problem of overload control and comments on an earlier draft of the paper. Acknowledgements are also due to J.W. Cohen for suggesting the problem and useful comments.

2. THE MODEL

In this section the technical set up of a telephone exchange will be summarized and a mathematical model in the form of a hierarchical queueing system developed. In the next section the problem of overload control will be formulated as an optimal stochastic control problem.

An engineering model

To model the operation of a telephone exchange and, specifically, the dynamics of the processor load, it is necessary to describe the technical set up in some detail.

A customer who takes up the receiver sends thus a signal to the telephone exchange, to be called a *call request*. Call requests are on detection at the exchange placed in a buffer by the central processor. These buffered requests will be termed *calls-in-build-up*. During its presence at the buffer a call-in-build-up generates *tasks* which are again handled by the central processor. These tasks consist of a request for a dial tone, a request to establish a desired connection, and related actions.

It has been argued by F.C. Schoute [14] that the variables calls-in-build-up and tasks are the essential state variables that describe the dynamics of the processor load. This suggestion is followed below.

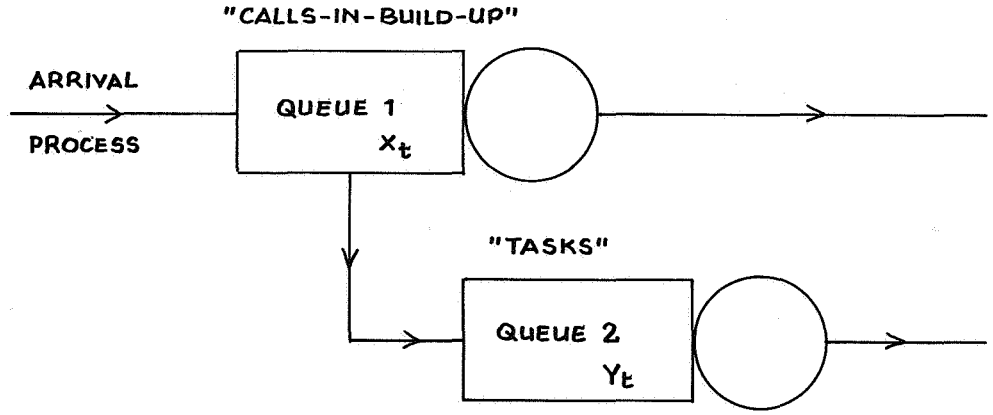

Fig. 1. An engineering model of a telephone exchange.

An engineering model for the dynamics of the processor load is the hierarchical queueing model depicted in figure 1. Call requests by customers are represented by the arrival process. The first queue represents the buffer with the calls-in-build-up. The first server unit is assumed to have an infinite number of servers. The calls-in-build-up in queue 1 generate tasks during their presence there. The second queue represents the tasks for the central processor. The second server unit is assumed to have only one server and to operate on a first-in-first-out basis. The combined queues will be called a *hierarchical queueing system* because the arrival process of tasks for queue 2 is assumed to be proportional to the number of calls-in-build-up present in queue 1.

A mathematical model

A mathematical model for the above defined hierarchical queueing system will be formulated next. It will be assumed that both queues have an infinite buffer, that queue 1 has an infinite number of servers, and that queue 2 has only one server. All arrival and server processes are assumed to be Poisson processes, possibly with time-varying intensity.

Assume given a complete probability space $\{\Omega,F,P\}$ and a time-index set $T = R_+$. The notation $(m_t,F_t,t\in T) \in M_l$ will be used to indicate that the stochastic process m is a martingale with respect to the σ-algebra family $(F_t,t\in T)$. Similarly the class of submartingales is denoted by $SubM_l$. For terminology on stochastic integrals and martingale theory consult [5]. Below $Z_+ = \{1,2,\ldots\}$, $N = \{0,1,2,\ldots\}$, and for any $n \in Z_+$, $Z_n = \{1,2,\ldots,n\}$, $N_n = \{0,1,2,\ldots,n\}$.

Let the arrival process $a: \Omega \times T \to R$ be a Poisson process with intensity $\lambda_1: T \to R_+$. Let the servers of queue 1 be represented by a sequence of independent Poisson processes $\{b_k, k \in Z_+\}$ $b_k: \Omega \times T \to R_+$, each with intensity process $\mu: T \to R_+$. Furthermore, let the task generating processes be represented by a sequence of independent Poisson processes $\{a_k, k \in Z_+\}$ $a_k: \Omega \times T \to R_+$, each with intensity process $\lambda: T \to R_+$. The server process of queue 2 is presented by a Poisson process $b: \Omega \times T \to R$ with intensity $\mu_0: T \to R_+$. Assume further that the Poisson processes $\{a, b_k, a_k, b, k \in Z_+\}$ are mutually independent.

Let further $x: \Omega \times T \to R_+$ represent the number of calls-in-build-up that are being served in queue 1, and $\lambda: \Omega \times T \to R_+$ the number of tasks in queue 2 that are waiting or being served. Let for all $t \in T$

$$F_t = \sigma(\{a_s, b_{ks}, a_{ks}, b_s, \forall s \leq t, \forall k \in Z_+\})$$

and let the σ-algebra family $(F_t, t \in T)$ be constructed such that it satisfies the usual conditions [5].

Under the above conditions one obtains the following representations for the dynamic behaviour of the processes x and y:

$$(2.1) \qquad dx_t = da_t - I_{[1,\infty)}(x_{t-}) \sum_{k=1}^{\infty} I_{[1,x_{t-}]}(k)\, db_{kt},$$

$$(2.2) \qquad dy_t = \sum_{k=1}^{\infty} I_{[1,x_{t-}]}(k)\, da_{kt} - I_{[1,\infty)}(y_{t-})\, db_t.$$

In (2.1) the second term on the right hand side presents the server process. The term $I_{[1,\infty)}(x_{t-})$ is present to let a completion of service effect x_t only if there are customers being serviced; thus if $x_{t-} > 0$. The infinite sum represents the infinite number of servers. Note that if there are x_t customers present that then a number of x_t servers are busy. That the infinite sum presents the situation where each customer is served by a separate server is due to the forgetting property of the exponential distribution and the assumption that the server processes $\{b_k, k \in Z_+\}$ are independent. The first term on the right hand side of (2.2) represents the arrival prodess at queue 2. This arrival process is constructed from the task generating process. Note that if there are x_t customers present in queue 1, that then there are x_t task generating processes that arrive at queue 2. That the infinite sum represents the situation where each customer has his own task generating process follows by an argument similar to the justification of the infinite sum in (2.1).

For the processes x and y a special semi-martingale representation is derived:

$$
\begin{aligned}
dx_t &= da_t - I_{[1,\infty)}(x_{t-}) \sum_{k=1}^{\infty} I_{[I,x_{t-}]}(k)db_{kt} \\
&= da_t - I_{[1,\infty)}(x_{t-}) \sum_{k=1}^{\infty} I_{[1,x_{t-}]}(k)db_{kt} \\
&= [\lambda_1(t) - I_{[1,\infty)}(x_{t-}) \sum_{k=1}^{\infty} I_{[1,x_{t-}]}(k)\mu(t)]dt + dm_{1t} \\
&= [\lambda_1(t) - x_t\mu(t)]dt + dm_{1t}
\end{aligned}
$$

where $(m_{1t}, F_t, t \in T) \in M_1$. Furthermore

$$
\begin{aligned}
dy_t &= \sum_{k=1}^{\infty} I_{[1,x_{t-}]}(k)da_{kt} - I_{[1,\infty)}(y_{t-})db_t \\
&= [\sum_{k=1}^{\infty} I_{[1,x_{t-}]}(k)\lambda(t) - I_{[1,\infty)}(y_t)\mu_0(t)]dt + dm_{2t} \\
&= [x_t\lambda(t) - I_{[1,\infty)}(y_t)\mu_0(t)]dt + dm_{2t},
\end{aligned}
$$

where $(m_{2t}, F_t, t \in T) \in M_1$. It follows from these calculations that $\langle m_1, m_2 \rangle = 0$.

The representation of the dynamic behaviour of the processor load is thus given by

$$
dx_t = [\lambda_1(t) - x_t\mu(t)]dt + dm_{1t},\ x_0, \tag{2.3}
$$

$$
dy_t = [x_t\lambda(t) - I_{[1,\infty)}(y_t)\mu_0(t)]dt + dm_{2t}, y_0. \tag{2.4}
$$

This representation is analogous to that of a model for software reliability developed in [11].

To allow readers to evaluate the hierarchical queueing system the differential equation for the probability distribution of the queueing systems will be presented.

2.1 PROPOSITION. *Given the hierarchical queueing system specified above. Let* $p: T \times N \times N \to R$

$$p(t,k,m) = P(\{x_t = k\} \cap \{y_t = m\}).$$

Then p *is a solution of the differential equation*

$$\begin{aligned}(2.5)\qquad \dot{p}(t,k,m) = {} & [p(t,k-1,m)I_{[1,\infty)}(k) - p(t,k,m)]\lambda_1(t) \\ & + [(k+1)p(t,k+1,m) - kp(t,k,m)]\mu(t) \\ & + [p(t,k,m-1)I_{[1,\infty)}(m) - p(t,k,m)]k\lambda(t) \\ & + [p(t,k,m+1) - p(t,k,m)I_{[1,\infty)}(m)]\mu_0(t), \quad p(0,k,m).\end{aligned}$$

PROOF. The elementary calculation is omitted. □

3. THE PROBLEM FORMULATION

In this section the problem of overload control will be formulated as a stochastic control problem.

An engineering model for the dynamics of the processor load has been presented in section 2, see figure 1. This model is now modified to account for the fact that access to the telephone exchange can be controlled, see figure 2. With the switch S a call request may be admitted to the telephone exchange, or be refused access. It will be assumed that customers that have been refused access will not try to regain access. This assumption is a modelling approximation. However, it can be circumvented at the cost of additional complexity of the model.

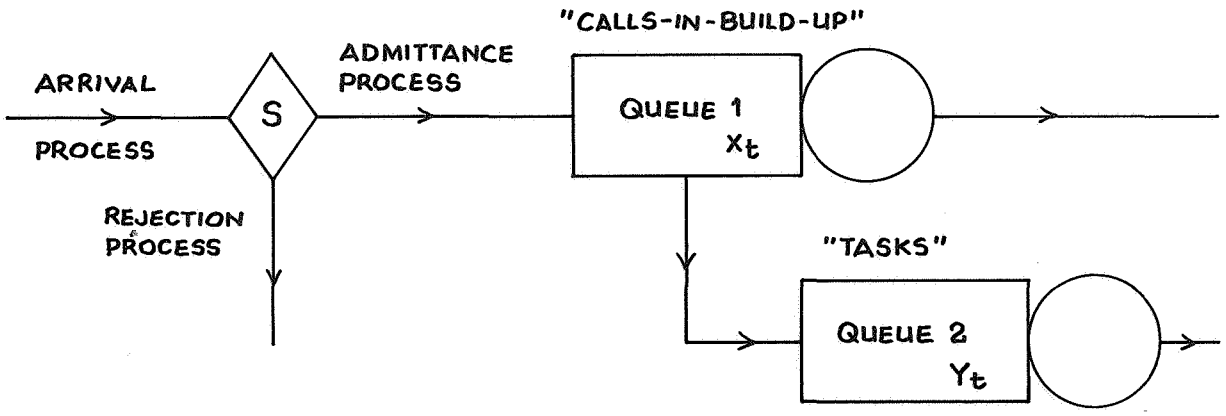

Fig. 2. An engineering model of a controlled telephone exchange.

A mathematical model for the above specified engineering model will be formulated below. The class of admissable control policies will consist of controls for the switch S that at each time moment are based on the past of all processes of the queueing system. The formulation of the stochastic control problem is completed by the specification of the cost function for which a discounted cost is taken of the form

$$(3.1) \qquad J_1(u) = E\Big[\int_{t_0}^{t_1} g(x_s,y_s)\exp(-cs)ds\Big]$$

where $g: N \times N \to R$ and $c \in R_+$.

A preliminary formulation of the stochastic control problem is then to determine a control in the class of admissable controls, if one exists, that minimizes the cost function $J_1(u)$. Below a formal problem formulation is given that follows [2] rather closely. The objective for the following discussion is to construct, for any control policy, a measure on the given measurable space.

Let $\{\Omega,F\}$ be a measurable space on which are defined counting processes $n: \Omega \times T \to R_+$ and $b: \Omega \times T \to R_+$, and families of counting processes $\{a_k,k \in Z_+\}$, $\{b_k,k \in Z_+\}$, $a_k: \Omega \times T \to R_+$, $b_k: \Omega \times T \to R_+$. For $m \in Z_+$ let $\tau_m: \Omega \to TU\{+\infty\}$ be the m-th jump time of the counting process n. Let further $\{z_m,m \in Z_+\}$ be random variables $z_m: \Omega \to \{0,1\}$ and $a: \Omega \times T \to R_+$

$$(3.2) \qquad a_t = \sum_{m=1}^{\infty} z_m I_{\{\tau_m \le t\}}.$$

Define the σ-algebra families, for $t \in T$

$$G_t = \sigma(\{b_{ks},a_{ks},b_s,\forall s \le t,k \in Z_+\}),$$

$$G_\infty = \bigvee_{t\in T} G_t,$$

$$F_t = \sigma(\{a_s,n_s,\forall s \le t\}) \vee G_t,$$

and for $m \in Z_+$

$$H_m = \sigma(\{\tau_i,z_i,\forall i \in Z_m\}),$$

$$K_m = \sigma(\{\tau_i,z_j,\forall i \in Z_m,j \in Z_{m-1}\}),$$

where $Z_m = \{1,2,\ldots,m\}$.

Let λ_0: $T \to R_+$ and

$$(3.3)\qquad \underline{\bar{U}} = \{\underline{\bar{u}}_m, K_m v G_{\tau_m}, m \in Z_+) \,|\, \forall m \in Z_+$$

$$\underline{\bar{u}}_m\colon \Omega \to [0,1],\ \bar{u}_m \text{ is } K_m v G_{\tau_m} \text{ measurable}\}.$$

For any $\bar{u} \in \underline{\bar{U}}$ the conditions

$$(3.4)\quad 1.\ \bar{E}_{\underline{u}}[I_{\{\tau_{m+1}-\tau_m \le t\}} \,|\, H_m v G_\infty]$$

$$= 1 - \exp\Big(-\int_{\tau_m}^{\tau_m+t} \lambda_0(s)ds\Big);$$

$$2.\ \bar{E}_{\underline{u}}[I_{\{z_m=1\}} \,|\, K_m v G_{\tau_m}] = \bar{u}_m;$$

3. b_k, a_k, b, for $k \in Z_+$ are mutually independent Poisson processes with the intensities specified in section 2;

determine a probability measure $\bar{P}_{\underline{u}}$ on $\{\Omega,F\}$. The proof of this fact is analogous to that of the remark in [2, p. 224].

The interpretation of the above construction is that with respect to $\bar{P}_{\underline{u}}$ n is a Poisson process with intensity λ_0, and that when a call request arrives or the process n jumps, that then $\bar{u}_m$ determines the probability that this request is admitted into the arrival process a; see (3.2) and (3.3(2)).

3.1 PROBLEM. Determine $\bar{u}^* \in \underline{\bar{U}}$ such that

$$\bar{J}(\bar{u}^*) = \bar{E}_{\underline{u}^*}\Big[\int_{t_0}^{t_1} g(x_s,y_s)\exp(-cs)ds\Big] \le \bar{J}(\bar{u}),$$

for all $\bar{u} \in \underline{\bar{U}}$. If such a $\bar{u}^*$ exists it is called an *optimal control*.

As in [2, lemma 1] one can then show that there exists a predictable process $u\colon \Omega \times T \to [0,1]$ $\{u_t,F_t,t \in T\}$ such that with respect to $\bar{P}_{\underline{u}}$ the counting process a has the intensity process $\{u_t\lambda_0(t),F_t,t \in T\}$, or that

$$(3.5)\qquad da_t = u_t\lambda_0(t)dt + dm_t,\quad a_0,$$

where $(m_t,F_t,t\in T) \in M_1(\bar{P}_u)$. Moreover, for $m \in Z_+$, is $u_{\tau_m} = \bar{u}_m$ a.s. $\bar{P}_u$.

One can then reformulate problem 3.1. Let the class of admissable controls be

(3.6) $\underline{U} = \{u: \Omega \times T \to [0,1] \mid (u_t,F_t,t\in T)$ a predictable process, such that there exists a probability measure P_u on $\{\Omega,F\}$ such that with respect to P_u and $(F_t,t\in T)$:

1. n has the intensity process $\lambda_0(t)$;
2. a has the intensity process $u_t\lambda_0(t)$;
3. a_k,b_k,b have the intensity processes as given in section 2;
4. while n,a_k,b_k,b are mutually independent;

(3.7) 5. $E_u\left|\int_{t_0}^{t_1} \exp(-cs)g(x_s,y_s)ds\right| < \infty\}$.

Under certain integrability conditions it can be shown that given a predictable process $(u_t,F_t,t\in T)$ there exists a probability measure P_u, see [8]. The only restrictive condition in the class of admissable controls is condition 5. However, this is necessary for the comparison of cost functions. It will be assumed that $\underline{U}$ is not empty. The queueing processes x and y are specified by the representation

(3.8) $$dx_t = [u_t\lambda_0(t) - x_t\mu(t)]dt + dm_{1t},x_0,$$

(3.9) $$dy_t = [x_t\lambda(t) - I_{[1,\infty)}(y_t)\mu_0(t)]dt + dm_{2t},y_0.$$

This representation follows from the discussion of section 2 and (3.5).

3.2 PROBLEM. a. Determine a $u^* \in \underline{U}$ such that

$$J(u^*) = E_u\left[\int_{t_0}^{t_1} g(x_s,y_s)\exp(-cs)ds\right] \le J(u)$$

for all $u \in \underline{U}$. If such a control exists it is called an *optimal control* for this problem

b. Determine a control $u^* \in \underline{U}$ such that for all $t \in T$

(3.10) $$E_{u^*}\left[\int_{t_0}^{t_1} g(x_s,y_s)\exp(-cs)ds \mid F_t\right]$$
$$\le E_u\left[\int_{t_0}^{t_1} g(x_s,y_s)\exp(-cs)ds \mid F_t\right],$$

for any control $u \in \underline{U}$ such that for all $s \in T$, with $s \leq t$, $u_s = u^*_s$. If such a control exists it is called a *conditionally optimal control.*

3.3 PROPOSITION. *Assume that* $u^* \in \underline{U}$ *is an optimal control for problem 3.2. Define for* $m \in Z_+$ $\bar{u}_m = u_{\tau_m}$, *where* τ_m *is the* m*-th jump time of* n. *Then* $\{\bar{u}_m, m \in Z_+\}$ *is an optimal control for problem 3.1.*

PROOF. This is analogous to that of [2, Prop. 2]. □

4. STOCHASTIC CONTROL

The optimal stochastic control problem 3.2 posed in section 3 will be solved below.

4.1 THEOREM. *Given the stochastic control problem 3.2. Assume that there exists a function* $v: T \times N \times N \to R$ *that satisfies the system of differential equations*

$$
\begin{aligned}
(4.1)\qquad & \dot{v}(t,k,m) - cv(t,k,m) + g(k,m) \\
& + [v(t,k-1,m) - v(t,k,m)]k\mu(t) \\
& + [v(t,k,m+1) - v(t,k,m)]k\lambda(t) \\
& + [v(t,k,m-1) - v(t,k,m)]\mu_0(t)I_{[1,\infty)}(m) \\
& + [v(t,k+1,m) - v(t,k,m)]\lambda_0(t).I_{R-}(v(t,k+1,m)-v(t,k,m)) \\
& = 0,\ v(t_1,k,m) = 0.
\end{aligned}
$$

Define the control law $u^*: T \times N \times N \to R$

$$(4.2)\qquad u^*_t = I_{R-}(v(t,x_{t-}+1,y_{t-}) - v(t,x_{t-},y_{t-})).$$

Assume that

$$(4.3)\qquad E_{u^*}\left|\int_{t_0}^{t_1} \exp(-cs)g(x_s,y_s)ds\right| < \infty.$$

a. *Then* $u^* \in \underline{U}$ *is an admissable conditional optimal control for problem* 3.2.

b. *Let for* $m \in Z_+$, $\bar{u}^*_m = u^*_{\tau_m}$. *Then* $(\bar{u}^*_m, m \in Z_+)$ *is an optimal control for problem 3.1.*

4.2 REMARKS. 1. The solution presented in 4.1 has a nice interpretation. From the proof of 4.1 it follows that $v(t,x_t,y_t)$ is an estimate of the future cost based on the current information at time $t \in T$. At any time $t \in T$.

$$v(t,x_{t-}+1,y_{t-}) - v(t,x_{t-},y_{t-})$$

is then the increase or decrease of the estimated future cost if a new call request is admitted to the exchange. The optimal control

$$u_t^* = I_{R-}(v(t,x_{t-}+1,y_{t-}) - v(t,x_{t-},y_{t-}))$$

has thus the interpretation that a call request should be admitted if and only if the estimated future cost is reduced by doing so.

2. Whether a solution to the system of differential equations (4.1) exists is an open question. The main difficulty is that this system is doubly infinite because $k,m \in N$. It seems likely that there exist $\bar{k}(t)$, $\bar{m}(t) \in N$, depending on $t \in T$, such that for $k > \bar{k}(t)$, $m > \bar{m}(t)$ no call requests should be admitted. This divides the set $N \times N$ in two areas in which (4.1) should be solved. However the determination of $\bar{k}(t)$ and $\bar{m}(t)$ is not clear. In section 5 a finite and time-invariant version of the stochastic control problem will be investigated for which one can hope to obtain more explicit results.

4.3 PROOF of 4.1. Let $h: \Omega \times T \times \underline{U} \to R$

$$(4.3) \qquad h_t(u) = \int_{t_0}^{t} \exp(-cs)g(x_s,y_s)ds + v(t,x_t,y_t)\exp(-ct).$$

It will be shown that u^* as defined in 4.1 belongs to $\underline{U}$, that for any $u \in \underline{U}$ the process $h(u) = (h_t(u),F_t,t \in T) \in \mathrm{SubM}_1(P_u)$, and that $h(u^*) \in M_1(P_{u^*})$.

2. By the differentiablity assumption on v, the stochastic calculus rule, the representation (3.8,3.9), and some calculations, it follows that

$$v(t,x_t,y_t) = v(0,x_0,y_0)$$
$$+ \int_{t_0}^{t} [\dot{v}(s,x_s,y_s)$$
$$+ [v(s,x_s+1,y_s) - v(s,x_s,y_s)]u_s\lambda_0(s)$$
$$+ [v(s,x_s-1,y_s) - v(s,x_s,y_s)]I_{[1,\infty)}(x_s)x_s\mu(s)$$
$$+ [v(s,x_s,y_s+1) - v(s,x_s,y_s)]x_s\lambda(s)$$
$$+ [v(s,x_s,y_s-1) - v(s,x_s,y_s)]I_{[1,\infty)}(y_s)\mu_0(s)]ds + m_t,$$

where $(m_t,F_t,t\in T) \in M_1(P_u)$. Then

$$dh_t(u) = \exp(-ct)[g(x_t,y_t)-cv(t,x_t,y_t)+\dot{v}(t,x_t,y_t)$$
$$+ [v(t,x_t-1,y_t)-v(t,x_t,y_t)]I_{[1,\infty)}(x_t)x_t\mu(t)$$
$$+ [v(t,x_t,y_t+1)-v(t,x_t,y_t)]x_t\ \lambda(t)$$
$$+ [v(t,x_t,y_t-1)-v(t,x_t,y_t)]I_{[1,\infty)}(y_t)\mu_0(t)$$
$$+ \inf_{k_t\in[0,1]} k_t\lambda_0(t)[v(t,x_t+1,y_t)-v(t,x_t,y_t)]]dt$$
$$+ \exp(-ct)[u_t\lambda_0(t)[v(t,x_t+1,y_t)-v(t,x_t,y_t)]$$
$$- \inf_{k_t\in[0,1]} k_t\lambda_0(t)[v(t,x_t+1,y_t)-v(t,x_t,y_t)]]dt + \exp(-ct)dm_t.$$

Let for any $u \in \underline{U}$ $r(u)\colon \Omega \times T \to R$

$$r_t(u) = \exp(-ct)u_t\lambda_0(t)[v(t,x_t+1,y_t)-v(t,x_t,y_t)]$$
$$- \inf_{k_t\in[0,1]} \{k_t\lambda_0(t)[v(t,x_t+1,y_t)-v(t,x_t,y_t)]\}.$$

3. Take any $u \in \underline{U}$. Because v satisfies the differential equation (4.1), one has

$$dh_t(u) = r_t(u)dt + \exp(-ct)dm_t. \tag{4.4}$$

By definition of r(u) one has that for any $t \in T$ $r_t(u) \geq 0$. Hence $h(u) \in \text{SubM}_1(P_u)$. Because $v(t_1,x_{t_1},y_{t_1}) = 0$, $h_{t_1}(u)$ satisfies the terminal condition

$$h_{t1}(u) = \int_{t_0}^{t_1} \exp(-cs)g(x_s,y_s)ds.$$

4. Consider now the control policy u^* specified by (4.2)

$$u_t^* = I_{R-}(v(t,x_{t-}+1,y_{t-})-v(t,x_{t-},y_{t-})).$$

By the measure transformation construction, the processes x and y that satisfy (3.8, 3.9) exist. Then $(u_t^*,F_t,t \in T)$ is a predictable process. It is now necessary to assume (4.3). Then $u^* \in \underline{U}$ is an admissable control. From the definition of r(u) it is seen that $r(u^*) = 0$, hence that $h(u^*) \in M_1(P_{u^*})$. From [1,16] then follows that u^* is conditional optimal for problem 3.2. From 3.3 then follows that $(\bar{u}_m^*, m \in Z_+)$ is optimal for problem 3.1. □

5. TOWARDS AN IMPLEMENTATION

Practical application of results from optimal control theory to overload control demands a control law that is time-invariant. The reason for this is that one is generally interested in long term control of overload. In additions a time-invariant control law is much easier to implement.

An approach to determine a time-invariant control law is to use the solution of the optimal stochastic control problem of section 4 and to let the starting time t_0 go to $-\infty$. This comes down to taking the limit $\lim_{t\to-\infty} v(t,k,m) = w(k,m)$, and to use as optimal control law the structure of (4.2) with w instead of v. However, there is a difficulty with this approach. The system of differential equations for v(t,k,m) is doubly infinite because it is indexed by $N \times N$. This makes convergence analysis rather involved.

Below another approach is followed that proceduces an algorithm that is relatively easy to implement. Specifically, it will be assumed that the server units have finite waiting rooms and that, if the waiting rooms are filled, additional customers or tasks are not admitted or produced.

5.1 ASSUMPTIONS. Consider the hierarchical queueing system developed in the sections 2 and 3.

1. In queue 1 there can be maximally k_1 customers in the system that are being served.
2. If the serving room of queue 1 is filled then newly arriving customers are turned away and are assumed not to return.
3. In queue 2 there can be maximally m_1 customers in the system that are waiting or being served.
4. If the waiting room of queue 2 is filled then the servers of queue 1 stop serving completely.
5. The intensities of the arrival and server processes do not depend on time explicitly. Below these intensities will be denoted by $\lambda_0, \mu, \lambda, \mu_0 \in R$.

Under the assumptions 5.1 one can derive the following representation for the controlled hierarchical queueing system

$$(5.1) \qquad dx_t = [\lambda_0 u_t I_{N_{k_1-1}}(x_t) - \mu x_t I_{N_{m_1-1}}(y_t)]dt + dm_{1t},\ x_{t_0}$$

$$(5.2) \qquad dy_t = [\lambda x_t I_{N_{m_1-1}}(y_t) - \mu_0 I_{[1,\infty)}(y_t)]dt + dm_{2t},\ y_{t_0}.$$

In (5.1) one notices that if there are k_1 or more calls-in-build-up in the first queue then the arrival process to the first queue is stopped. Similarly, if there are m_1 or more tasks in queue 2, then the servers of queue 1 stop serving, see (5.1), with as a consequence that the arrival process to queue 2 stops also, see (5.2). In the following it is assumed that $x_{t_0} : \Omega \to N_{k_1}$, $y_{t_0} : \Omega \to N_{m_1}$.

5.2 PROBLEM. Consider the stochastic control system described by (5.1,5.2), with the assumptions 5.1 and cost function (3.1).

a. Assume that the class of admissable controls is specified by (3.3). The problem is then to determine an optimal control $\bar{u}^* \in \bar{U}$.
b. As in section 3 one can reformulate the above problem. Assume that the class of admissable controls is given by (3.6). The problem is then to determine an optimal control $u^* \in U$.

5.3 THEOREM. *Given the stochastic control problem 5.2 with the assumption 5.1.*

a. *There exists a solution to the system of differential equations*

$$v: T \times N_{k_1} \times N_{m_1} \to R$$

$$\begin{aligned}(5.3)\qquad & \dot{v}(t,k,m)-cv(t,k,m)+g(k,m)\\ & + [v(t,k-1,m)-v(t,k,m)]k\mu I_{N_{m_1-1}}(m)\\ & + [v(t,k,m+1)-v(t,k,m)]k\lambda I_{N_{m_1-1}}(m)\\ & + [v(t,k,m-1)-v(t,k,m)]\mu_0 I_{Z_+}(m)\\ & + [v(t,k+1,m)-v(t,k,m)]\lambda_0 I_{N_{k_1-1}}(k)*\\ & \quad *I_{R-}(v(t,k+1,m)-v(t,k,m)) = 0,\ v(t_1,k,m) = 0.\end{aligned}$$

b. *The solution to problem 5.2.b. is the control law*

$$(5.4)\qquad u_t^* = I_{R-}(v(t,x_{t-}+1,y_{t-})-v(t,x_{t-},y_{t-})).$$

c. *The solution to problem 5.2.a is given by the control* $(\bar{u}_m, m\in Z_+)$, *where with* u^* *specified by b.,*

$$\bar{u}_m^* = u^*_{\tau m}$$

for all $m \in Z_+$.

PROOF. a. Let $M = (k_1+1).(m_1+1)$ $f: T \to R^M$

$$f(t)^T = (v(t,0,0),v(t,0,1),\ldots,v(t,0,m_1),v(t,1,0),\ldots,v(t,k_1,m_1)).$$

Denote the elements of f(t) by

$$f(t)^T = (f_{0,0}(t),f_{0,1}(t),\ldots,f_{0,m_1}(t),f_{1,0}(t),\ldots,f_{k_1,m_1}(t)).$$

Let further $g_1 \in R^M$,

$$g_1 = (g(0,0),g(0,1),\ldots,g(0,m_1),g(1,0),\ldots,g(k_1,m_1)),$$

and $A \in R^{M\times M}$, $F: R^M \to R^M$ be such that the differential equation (5.3) is represented by

(5.5) $\dot{f}(t) = G(f(t)) - g_1, \quad f(t_1) = 0,$

where

$$G(x) = (A+cI)x + F(x),$$

$$F_{k,m}(f(t)) = \lambda_0[f_{k+1,m}(t)-f_{k,m}(t)]I_{R-}(f_{k+1,m}(t)-f_{k,m}(t)),$$

$$\text{if } k \leq k_1 - 1,$$

$$0, \text{ otherwise.}$$

The components of F are indicated as those of f. It is then a calculation to show that G is Lipschitz, of which the key step is to show that if $x,y \in R$ then

$$|xI_{R-}(x)-yI_{R-}(y)| \leq |x-y|.$$

This is easily done. The existence of a solution f of (5.5), and hence of the solution v of (5.3), follows then from standard analysis.
b. The proof of this statement is analogous to that of 4.1. □

Although the optimal control law of 5.3.b is indexed by a finite number of values, it is still time-varying. Implementation of this control law is therefore difficult. In the following attention is restricted to stationary control laws.

5.4 PROBLEM. Consider the stochastic control system (5.1,5.2) and the class of stationary control laws

(5.6) $\underline{U}_s = \{(u(x_{t-},y_{t-}),t \in T) | u: N_{k_1} \times N_{m_1} \to [0,1],$
and (x,y) are determined by (5.1,5.2)
with $u_t = u(x_{t-},u_{t-})\}$

Consider further the cost function

(5.7) $$w(x_{t_0},y_{t_0},u) = E_u[\int_{t_0}^{\infty} \exp(-c(s-t_0))g(x_s,y_s)ds | F_{t_0}^{x,y}].$$

The problem is then to determine a control $u^* \in \underline{U}_s$, to be called an *optimal stationary control*, such that for any $u \in \underline{U}_s$, $x_{t_0} \in N_{k_1}$, $y_{t_0} \in N_{m_1}$

$$w(x_t, y_t, u^*) \leq w(x_{t0}, y_{t0}, u).$$

Because x and y are finite valued, the conditional expectation in (5.7) is well defined. It is easily seen that the expression for w does not depend on $t \in T$ explicitly. Because only controls dependent on (x_t,y_t) are admitted in $\underline{U}_s$, is (x,y) a Markov process. Hence w depends only on (x_{t_0},y_{t_0}) and u.

Define $w^*: N_{k_1} \times N_{m_1} \to R$

$$(5.8) \qquad w^*(k,m) = \inf_{u \in \underline{U}_s} w(k,m,u).$$

Problem 5.4 can be solved by methods from the theory of Markov decision processes [12, Ch.6]. The key steps are outlined below

5.5 THEOREM. *Consider the stochastic control problem* 5.4. *Assume that the class* $\underline{U}_s$ *is relatively complete.*

a. *For any* $s,t \in T$, $s < t$,

$$(5.9) \qquad w^*(x_s,y_s) = \inf_{u \in \underline{U}_s} E_u[\int_s^t \exp(-c(\tau-s))g(x_\tau,y_\tau)d\tau + \exp(-c(t-s))w^*(x_t,y_t)\,|\,F_s^{x,y}].$$

b. *If there exists a* $u^* \in \underline{U}_s$ *such that for all* $s,t \in T$, $s < t$,

$$(5.10) \qquad E_{u^*}[\int_s^t \exp(-c(\tau-s))g(x_\tau,y_\tau)d\tau + \exp(-c(t-s))w^*(x_t,y_t)\,|\,F_s^{x,y}]$$

$$= \inf_{u \in \underline{U}_s} E_u[\int_s^t \exp(-c(\tau-s))g(x_\tau,y_\tau)d_\tau + \exp(-c(t-s))w^*(x_t,y_t)\,|\,F_s^{x,y}]$$

then u^* *is an optimal stationary control.*

c. w^* *as defined by* (5.8) *is the unique solution of equation* (5.9).

PROOF. The proof is analogous to that of the discrete-time setting presented in [12,6.2] and therefore omitted. □

The class of admissable controls $\underline{U}_s$ is called *relatively complete* [17] if for all $\varepsilon \in (0,\infty)$, $x_{t_0} \in N_{k_1}$, $y_{t_0} \in N_{m_1}$ there exists a $u \in \underline{U}_s$ such that $w(x_{t0},y_{t0},u) < w^*(x_{t0},y_{t0}) + \varepsilon$.

5.6 THEOREM. *Consider the stochastic control problem* 5.4. *Assume that the class* $\underline{U}_s$ *is relatively complete. Assume further that there exists a* $w: N_{k_1} \times N_{m_1} \to R$ *such that*

$$
\begin{aligned}
(5.11)\quad & g(k,m)-cw(t,m) \\
& + [w(k-1,m)-w(k,m)]\mu k I_{N_{m_1}-1}(m) \\
& + [w(k,m+1)-w(k,m)]\lambda k I_{N_{m_1}-1}(m) \\
& + [w(k,m-1)-w(k,m)]\mu_0 I_{[1,\infty)}(m) \\
& + [w(k+1,m)-w(k,m)]\lambda_0 I_{N_{k_1}-1}(k) I_{R-}(w(k+1,m)-w(k,m)) = 0.
\end{aligned}
$$

a. *Then an optimal stationary control law for problem* 5.4 *is given by*

$$
(5.12)\quad u^*(x_{t-},y_{t-}) = I_{R-}(w(x_{t-}+1,y_{t-})-w(x_{t-},y_{t-}))I_{N_{k_1}-1}(x_{t-}).
$$

b. *The total cost is given by* $w(x_{t0},y_{t0})$.

PROOF. Let $s,t \in T$, $s < t$. Let w be a solution to the equation (5.11). A calculation then shows that for any $u \in \underline{U}_s$

$$
\begin{aligned}
& E_u\Big[\int_s^t \exp(-c(\tau-s))g(x_\tau,y_\tau)d\tau + \exp(-c(t-s))w(x_t,y_t)\,\Big|F_s^{x,y}\Big] \\
& = E_u[w(x_s,y_s) + \int_s^t \exp(-c(\tau-s))[g(x_\tau,y_\tau)-cw(x_\tau,y_\tau) \\
& + [w(x_\tau,y_\tau+1)-w(x_\tau,y_\tau)]\lambda x_\tau I_{N_{m_1}-1}(y_\tau) \\
& + [w(x_\tau-1,y_\tau)-w(x_\tau,y_\tau)]\mu x_\tau I_{N_{m_1}-1}(y_\tau) \\
& + [w(x_\tau,y_\tau-1)-w(x_\tau,y_\tau)]\mu_0 I_{[1,\infty)}(y_\tau)
\end{aligned}
$$

$$+ [w(x_\tau+1,y_\tau)-w(x_\tau,y_\tau)]\lambda_0 u_\tau I_{N_{k_1-1}}(x_\tau)]d\tau \,|\, F_s^{x,y}]$$

$$= w(x_s,y_s) + E_u[\int_s^t \exp(-c(\tau-s))$$

$$[w(x_\tau+1,y_\tau)-w(x_\tau,y_\tau)]\lambda_0 I_{N_{k_1-1}}(y_\tau)$$

$$[u_\tau - I_{R-}(w(x_\tau+1,y_\tau)-w(x_\tau,y_\tau))]d\tau \,|\, F_s^{x,y}].$$

The expression on the right-hand side of the above equation is minimized for $u_t = u^*(x_{t-},y_{t-})$ given by (5.12), and then vanishes. Thus w is a solution of the equation (5.9), and by 5.5.c. the solution of (5.9); thus $w = w^*$. Furthermore u^* specified by (5.12) achieves the minimum value as shown above, and from 5.5.b. then follows that u^* is an optimal stationary control. Because $w = w^*$, by (5.8) the total cost is given by $w(x_{t_0},y_{t_0})$. □

The equation (5.11) for w can be solved by a policy improvement method, by a successive approximation method, or a combination of these methods; see [12,6.2].

Implementation of the control law (5.12) can now be considered. The equation (5.11) for w can be solved, and the control law is then given by (5.12).

6. OPEN QUESTIONS

The problem of overload control of a SPC telephone exchange has been formulated as a stochastic control problem. The latter problem has been solved. A stationary control law has been derived that may be considered for implementation. Simulations of the controlled queueing system have not yet been made.

The approach of this paper can still be made more realistic. In practice one may observe y but not x. In this case one obtains a stochastic filtering problem and a partially observed stochastic control problem. These problems may be considered in the future. In practice one does not know the values of the parameters of the queueing system, for example of λ_0, μ, λ, μ_0. Thus one encounters a system indentification problem for point process systems and the adaptive control problem for point process observations. Much research remains to be done.

REFERENCES

[1] BOEL, R.K. & P. VARAIYA, *Optimal control of jump processes*, SIAM J. Control Optim., 15 (1977), pp. 92-119.

[2] BREMAUD, P., *Optimal thinning of a point process*, SIAM J. Control Optim., 17 (1979), pp. 222-229.

[3] BRÉMAUD, P., *Point processes and queues.- Martingale dynamics*, Springer-Verlag, Berlin, 1981.

[4] BRÉMAUD, P. & J. JACOD, *Processes ponctuels et martingales: revue des résultats récents sur la modélisation et le filtrage*, Adv. in Appl. Probab., 9 (1977), pp. 362-416.

[5] DELLACHERIE, C. & P.A. MEYER, *Probabilités et potentiel*, Chapitres I à IV, Hermann, Paris, 1975; *Probabilités et potentiel*, Chapitres V à VIII, Théorie des martingales, Hermann, Paris, 1980.

[6] FORYS, L.J. & H. ZUCKER, *A characterization of traffic variability for SPC systems*, paper presented at the 9th International Teletraphic Congress (ITC), 1981.

[7] FORYS, L.J., *Performance analysis of a new overload strategy*, 10th ITC, 1983.

[8] JACOD, J., *Multivariate point processes: predictable projection, Radon-Nikodyn derivatives, representation of martingales*, Z. Wahrsch. Verw. Gebiete, 31 (1975), pp. 235-253.

[9] JACOD, J., *Calcul stochatique et problèmes de martingales*, Lecture Notes in Mathematics, volume 714, Springer-Verlag, Berlin, 1979.

[10] KARLANDER, B., *Control of central processor load in an SPC system*, Ericsson Technics, 30 (1977), pp. 221-243.

[11] KOCH, G. & P.J.C. SPREIJ, *Software reliability as an application of martingale and filtering theory*, IEEE Trans. Reliability, 32 (1983), pp.

[12] ROSS, S.M., *Applied probability models with optimization applications*, Holdon-Day, San Francisco, 1970.

[13] SCHOUTE, F.C., *Optimal control and call acceptance in a* SPC *exchange*, 9th-ITC, 1981.

[14] SCHOUTE, F.C., *The technical queue: A model for definition and estimation for processor loading*, 9th Int. Symposium on Computer Performance Modelling, Measurement, and Evaluation, University of Maryland, May, 1983.

[15] SCHOUTE, F.C., *Adaptive overload control for an* SPC *exchange*, 10th-ITC, 1983.

[16] STRIEBEL, C., *Martingale conditions for the optimal control of continuous-time stochastic systems*, preprint, 1974.

[17] RISHEL, R., *Necessary and sufficient dynamic programming conditions for continuous-time stochastic control*, SIAM J. Control, 8 (1979), pp. 559-571.