Dependency of Service Time on Waiting Time in Switching Systems—A Queueing Analysis with Aspects of Overload Control

PHUOC TRAN-GIA, MEMBER, IEEE, AND MICHIEL H. VAN HOORN

Abstract—**Performance degradation of switching systems when the load increases above the engineered load can be caused by system-dependent and customer-dependent factors. In this paper the dependency of the service time and the call completion rate on the waiting time of a customer is investigated. The problem is modeled by means of a queueing system of type $M^{[X]}/G/1$, where state-dependent batch size distributions are considered. Two analysis methods, the continuous Markov chain approach and the regenerative method, are used for Markovian and generally distributed service phases, respectively. Numerical results are given for system characteristics, in particular the call completion rate of the system. Finally, a basic overload control scheme is investigated which increases the throughput of completed calls at higher traffic levels.**

I. Introduction

In switching systems, especially in stored program controlled systems, overload situations are caused by various factors, e.g., customer behavior or lack of system resources. Interaction between customer and system is an important factor which affects the system performance very strongly.

Reactions of customers can influence the system in different ways. On the one hand, a customer may abandon his call with a certain probability when he is confronted with large delays during the call setup phase (e.g., waiting for dial tone, post-dialing delay, etc.). In this case, an ineffective amount of work has been offered to the processor, and hence the call completion rate of the system decreases. On the other hand, rejected customers may reattempt their call after a certain time. The repeated attempts will further inflate the overload.

In order to investigate the behavior of a switching system in overload situations, a queueing model is employed in which the service time of a customer depends on his waiting time before entering service. Although this queueing model is generally applicable, attention is devoted to the determination of the performance limitation of overloaded switching systems. Regarding the existing complexity of the model, the repeated attempts phenomenon is not considered in this modeling approach.

The most important performance measures in a switching system are the probability for call completion and the call completion rate. The probability for call completion is defined as the number of call attempts that have been performed successfully, compared to all call requests offered to the system.

There are a number of studies which consider the dependency of the service time on the waiting time with varying degrees of complexity. Posner [1] analyzes a single server queueing model with respect to the dependency of the service duration on the waiting time, where an example for two service levels is given. Forys [2] discusses a basic model for applications in telephone switching systems where customers contribute one of two exponentially distributed processing times, depending on their waiting time. Rosenshine [3] considers this dependency in modeling the service time of air traffic controllers where the imbedded Markov chain method is used for analysis. Doshi and Lipper [9] introduce a queue with delay-dependent service, where the last-come-first-served (LCFS) service discipline is used for overload control purposes in switching systems.

The modeling approach will be presented in Section II. In Section III the analysis method will be described and some numerical results are given to show the main effects for the considered essential system characteristics. Finally, in Section IV a control mechanism for overload situations will be presented and investigated which allows us to optimize the system performance above engineered load. It will be shown that a remarkable enhancement of the system performance can be obtained by introducing a very simple overload control scheme.

II. Modeling Approach with Service Time Discretization

In this section a queueing system is presented which allows us to describe the dependency of processor service time of a call on its waiting time in order to calculate the call completion rate in a switching system.

We observe a test call entering a switching system. The call sees an amount of work waiting for processing. Concretely, this work may stand for the number of subcalls or telephonic events (call handling tasks or messages to be transferred for interprocessor or interprocess communications) buffered in the processor queue. Based on this observation and in order to simplify the analysis without losing essential effects, we consider the amount of work in the processor queue as a discrete number of phases which are assumed to be independent and identically distributed random variables.

The number of phases the test call sees upon arrival corresponds to its waiting time before entering service. Depending on the duration of its waiting time, the test call decides to bring a number of phases into the system. These phases can be interpreted as the number of subcalls and the corresponding call handling effort the switching system must spend for the call attempt. From the analysis point of view, we can consider the decision to be taken at the arrival epoch of the call, although in reality it is taken at the instant the customer enters service (e.g., dialing phase).

Calls with incompleted dialing or abandoned calls usually offer a small number of phases to the system, while successful calls with completed dialing often have offered a larger number of phases to the system. Therefore, according to the

Paper approved by the Editor for Computer Communication of the IEEE Communications Society. Manuscript received September 6, 1982; revised August 12, 1985. This work was supported in part by the Netherlands Organization for the Advancement of Pure Research.

P. Tran-Gia is with the Institute of Communications Switching and Data Technics, University of Stuttgart, Stuttgart, West Germany.

M. H. van Hoorn is with the Free University of Amsterdam, Amsterdam, The Netherlands.

IEEE Log Number 8607741.

number of phases chosen by a call, we define the probability that it will become a bad call or a successful call.

Considering all arguments discussed above, we have modeled the system as a single server queueing system of type $M^{[X]}/G/1$ with state-dependent batch arrivals. In fact we have the discrete version of a single server queue with state-dependent service time.

In this model the following assumptions are made.

- Call arrivals follow a Poisson process with rate λ.
- A call that sees k phases in the system (including the phase in service) will offer $G^{(k)}$ phases to the system, where $\Pr\{G^{(k)} = j\} = g_j^{(k)}$.
- A call having chosen j phases becomes a successful call (completed call) with the conditional completion probability c_j.
- The service time for an arbitrary phase has the distribution function $F_S(t)$ with mean h.

As will be specified in Section III-B, the probability that a call chooses j phases decreases with increasing number k of phases in the system and the conditional completion probability c_j increases in j.

The modeling arguments described below will help to simplify the calculation algorithm.

- Considering the observation of subcalls in switching systems, the number j of service phases chosen by a call may vary between fixed numbers N_0 and N_1 ($N_0 \leqslant j \leqslant N_1$).
- A call that sees upon arrival a very large number of phases in the system, say at least k_0 phases, will tend to become a bad call, and it will add N_0 phases for service. For the sake of convenience, we let $G = G^{(k)}$ and $g_j = g_j^{(k)}$ for $k \geqslant k_0$. This assumption corresponds to the observation that a customer who waits too long often tends to abandon his call after producing few subcalls.

Using the state-dependent batch size distribution, the effect of dependency between customer service time and waiting time can be described. In the next section, based on the calculation of the steady-state probabilities of the queueing system, the call completion rate for customers and the effective system throughput can be derived under different call traffic conditions.

III. Performance Analysis

In this section, the steady-state analysis of the $M^{[X]}/G/1$ queueing system with state-dependent batch arrivals, as described in the previous section, is presented. For ease of presentation, we shall refer to customers or calls consisting of phases and describe the state of the system by the number of phases present, including the phase in service.

Let the random variable X denote the number of phases in the system at an arbitrary epoch and define

$$p_n = \Pr\{X = n\}, \qquad n \geqslant 0.$$

The assumption is made that conditions for statistical equilibrium are satisfied. A sufficient condition for a stable queue is that the normalized call traffic intensity $\rho_0 = \lambda \cdot h \cdot E[G]$ is less than 1.

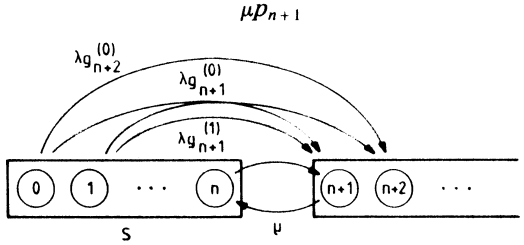
In Section III-A, Markovian service phases are considered ($F_S(t) = 1 - e^{-\mu t}$). In this case the queueing process is a birth-and-death process with multiple births, and the analysis is substantially easier than in the general case which will be dealt with in Section III-D.

A. Analytic Algorithm for Markovian Service Phases

To derive a set of equations for the state probabilities, we use the well-known balance property of Markov processes that the transition rate into some macro state S equals the transition rate out of S for any subset S of the state space. Consider the choice $S = \{0, 1, \cdots, n\}$; the transition rate out of S is given by

$$\lambda p_0 \Pr\{G^{(0)} \geqslant n+1\} + \lambda p_1 \Pr\{G^{(1)} \geqslant n\} + \cdots + \lambda p_n \Pr\{G^{(n)} \geqslant 1\}$$

and the transition rate into S by



Hence, we obtain

$$\mu p_{n+1} = \lambda \sum_{k=0}^{n} p_k \Pr\{G^{(k)} \geqslant n+1-k\}, \qquad n \geq 0. \quad (3.1)$$

The following generating functions are defined:

$$P(z) = \sum_{n=0}^{\infty} p_n z^n$$

$$\Gamma^{(k)}(z) = \sum_{n=0}^{\infty} g_n^{(k)} z^n$$

$$\Gamma(z) = \sum_{n=0}^{\infty} g_n z^n. \quad (3.2)$$

Multiplying (3.1) with z^{n+1} and summing over n, we obtain

$$\mu(P(z) - p_0) = \lambda z \sum_{k=0}^{\infty} p_k z^k \frac{1-\Gamma^{(k)}(z)}{1-z}$$

$$= \lambda z \frac{1-\Gamma(z)}{1-z} P(z) + \lambda z \sum_{k=0}^{k_0-1} p_k z^k \frac{1-\Gamma^{(k)}(z) - (1-\Gamma(z))}{1-z}$$

or

$$P(z) = \left[\mu p_0 + \lambda z \sum_{k=0}^{k_0-1} p_k z^k \frac{1-\Gamma^{(k)}(z) - (1-\Gamma(z))}{1-z}\right] \cdot \left[\mu - \lambda z \frac{1-\Gamma(z)}{1-z}\right]^{-1}. \quad (3.3)$$

Substituting $p_k = p_k^* p_0$ ($0 \leq k \leq k_0 - 1$) with $p_0^* = 1$, we can compute p_i^* ($1 \leq i \leq k_0 - 1$) with (3.1). Inserting $z = 1$ in (3.3) we obtain

$$p_0 = (\mu - \lambda E[G]) \cdot \left(\mu + \lambda \sum_{k=0}^{k_0-1} p_k^* [E[G^{(k)}] - E[G]]\right)^{-1} \quad (3.4)$$

by noting that

$$\lim_{z \to 1} \frac{1-\Gamma^{(k)}(z)}{1-z} = E[G^{(k)}].$$

The algorithm to calculate p_k is summarized in the following steps.

1) Set $p_0^* = 1$ and compute recursively p_k^*, $1 \leq k \leq k_0 - 1$ with (3.1).

2) Calculate p_0 with (3.4) and renormalize $p_k = p_k^* p_0$ for $0 \leq k \leq k_0 - 1$.

3) Compute further state probabilities p_k, $k \geq k_0$ recursively with (3.1).

After differentiation of (3.3) and setting $z = 1$, the following expression for the mean number of phases in the system is found:

$$E[X] = P'(1) = \sum_{k=1}^{\infty} k p_k$$

$$= \frac{\lambda}{2(\mu - \lambda E[G])} \left[E[G] + E[G^2] + \sum_{k=0}^{k_0-1} (2k+1) p_k (E[G^{(k)}] - E[G]) + \sum_{k=0}^{k_0-1} p_k (E[G^{(k)2}] - E[G^2]) \right]. \quad (3.5)$$

B. System Characteristics

The derivation of the steady-state distribution of the number of phases in the system forms the basic requirements to obtain the following performance characteristics:

P_{COMPL}	completion probability for an arbitrary call
Y	call completion rate
Y_0	$= Y N_0$: normalized call completion rate
$E[X]$	mean number of phases in the system
MBS	mean batch size, i.e., mean number of phases brought into the system by an arbitrary customer.

The performance characteristics are expressed in terms of p_n in the following way:

$$P_{COMPL} = \sum_{k=0}^{\infty} p_k \sum_{j=0}^{\infty} c_j g_j^{(k)}$$

$$Y = \lambda P_{COMPL}$$

$$E[X] = \sum_{k=1}^{\infty} k p_k$$

$$MBS = \sum_{k=0}^{\infty} p_k E[G^{(k)}]. \quad (3.6)$$

With the assumptions $g_j^{(k)} = 0, j < N_0$ or $j > N_1$, and $g_j^{(k)} = g_j$ for $k \geq k_0$, P_{COMPL} can be rewritten as

$$P_{COMPL} = \sum_{j=N_0}^{N_1} c_j \left(g_j + \sum_{k=0}^{k_0-1} p_k (g_j^{(k)} - g_j) \right). \quad (3.7)$$

Further note that $\lambda \cdot MBS$ is the average arrival rate of phases and $\lambda \cdot MBS \cdot h$ is the workload offered to the system per time unit, which is equal to $1 - p_0$, the fraction of time the server is busy. So we have

$$MBS = \frac{1 - p_0}{\lambda h}. \quad (3.8)$$

The mean system size $E[X]$ is given by (3.5).

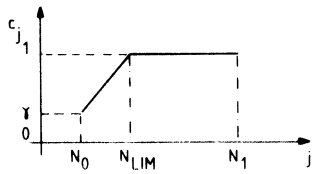


Fig. 1. Conditional completion probabilities.

For the probabilities c_j we have made the following choice containing the parameters γ and N_{LIM} as degrees of freedom (cf. Fig. 1).

$$C_j = \begin{cases} \gamma + (1-\gamma) \dfrac{j - N_0}{N_{LIM} - N_0} & N_0 \leq j \leq N_{LIM} \\ 1 & N_{LIM} \leq j \leq N_1 \\ 0 & \text{otherwise.} \end{cases} \quad (3.9)$$

In practical situations, the number of subcalls produced by a completed call will vary between certain limits, here presented by N_{LIM} and N_1. If the number of subcalls produced by a call is less than N_{LIM}, the probability to be completed decreases, but need not be zero.

The batch size distribution is the factor that takes into account the dependency between the service time of a customer and his waiting time. If a customer sees k phases in the system upon arrival, his waiting time has an Erlang-k distribution corresponding to the negative exponential phases. He is supposed to have a certain patience, i.e., he is willing to wait a reasonable time, say τ, before entering service. If his waiting time is short, he will choose a service time consisting of a relatively large number of phases, corresponding to a large number of subcalls. If his waiting time is longer than τ, he will tend to bring a smaller number of phases into the system, because he abandons his call sooner. As discussed in Section II, it is realistic to assume that the number of phases a customer chooses lies between certain numbers N_0 and N_1. However, for the analysis this assumption is not essential. The length of the patience τ could be obtained by measurement in a real system. Here, we choose $\tau = 3N_1 h$ (cf. Forys [2]). The above reasoning allows the following choice for the batch size distribution:

$$\Pr\{G^{(k)} = N_1\} = \Pr\{W_k \leq \tau\}$$

$$\Pr\{G^{(k)} = j\} = \Pr[\tau + (N_1 - j - 1)h \leq W_k < \tau + (N_1 - j)h\} \quad N_0 < j < N_1$$

$$\Pr\{G^{(k)} = N_0\} = \Pr\{W_k > \tau + (N_1 - N_0 - 1)h\}. \quad (3.10)$$

The random variable W_k denotes the waiting time of a customer seeing k phases in the system on his arrival epoch. In Fig. 2 the average number of phases chosen by a customer is shown. Also, the effect of the patience of customers is clearly illustrated.

C. Some Numerical Results

In this subsection, numerical results are presented which show system characteristics under different traffic conditions. For all the results, time is normalized by the mean service time of phases $h = 1/\mu = 1$, and the offered traffic intensity is standardized by $\rho_0 = \lambda N_0/\mu$.

Fig. 3 shows the completion probability for an arbitrary call as a function of the offered traffic intensity. The curves are drawn for different values of γ. It should be recalled that γ represents the completion probability for calls which have a relative long waiting time and choose the minimum number N_0 of phases. It can be seen here that the call completion

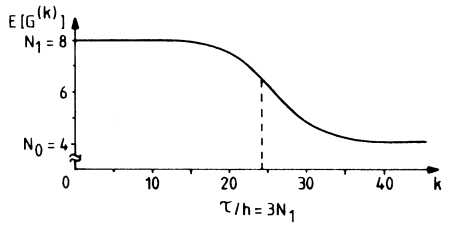


Fig. 2. Modeling aspects of customer behavior.

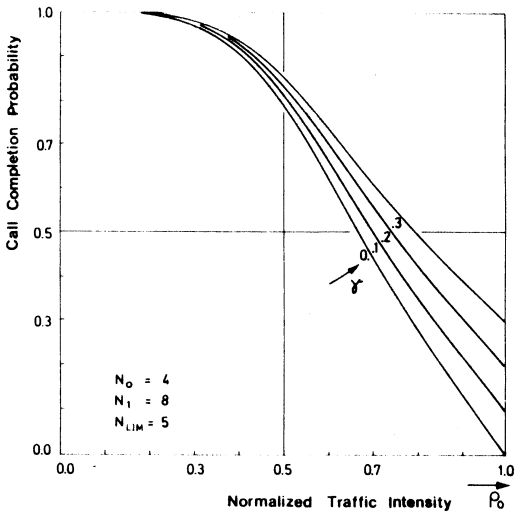


Fig. 3. Completion probability for calls versus normalized traffic intensity.

probability decreases rapidly above a certain level of the offered traffic. A degradation of the system performance is said to have occurred. This effect is shown more clearly in Fig. 4, where the call completion rate is depicted for different traffic intensities.

For $\rho_0 \geq 1$ the system becomes unstable and the queue increases to infinity. However, according to the modeling approach, the call completion rate is constant with value γ.

The mean number of phases in the system is shown in Fig. 5 as a function of the offered traffic intensity, where different values of the ratio N_1/N_0 are considered. For higher values of N_1/N_0 the curve can be clearly recognized as a superposition of two segments. The first segment of the curve corresponds to lower traffic levels where the batch size is approximately N_1; the second segment corresponds to higher traffic intensities, where the majority of customers chooses N_0 phases.

D. General Phase Distribution

The analysis of the $M^{[X]}/G/1$ queue with a general distribution of the service time of phases is more complicated than the $M^{[X]}/M/1$ case. In van Hoorn [4] the analysis is done by means of the regenerative method. Using up and down crossing arguments, a complete set of equations is derived to obtain the steady-state probabilities at arbitrary and at departure epochs. We shall summarize below the main aspects of the analysis.

Assuming the system is empty at epoch 0, we define the following random variables:

- T the next epoch at which the system becomes empty
- T_n amount of time during which n phases are in the system in the busy cycle $(0, T]$, $n \geq 0$
- N number of phases served in $(0, T]$
- N_n number of service completion epochs at which the phase served leaves n other phases behind in the system in $(0, T]$, $n \geq 0$

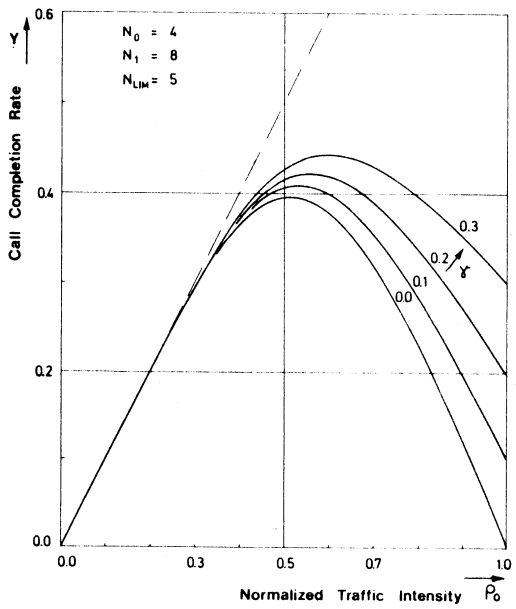


Fig. 4. Call completion rate versus normalized traffic intensity.

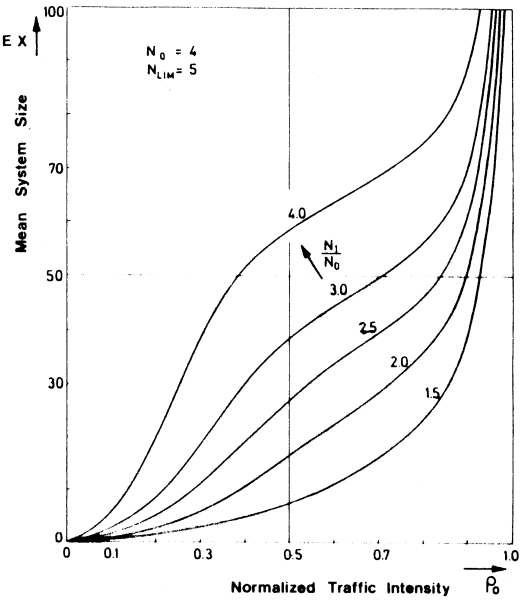


Fig. 5. Average number of phases in system versus normalized traffic intensity.

and the quantities

- A_{kn} expected amount of time during which n phases are in the system until the next service completion epoch, given that at epoch 0 a service is completed with k phases left behind in the system.

By partitioning the busy cycle by means of the service completion epochs and using Wald's theorem (cf. Ross [7]), we find

$$E[T_n] = \sum_{k=0}^{n} E[N_k] A_{kn}, \qquad n \geq 1. \tag{3.11}$$

Note that $E[N_k]$ equals the average number of times in a busy cycle that a service starts with k phases present.

For a second relation between the $E[T_n]$ and $E[N_n]$, we use a similar up and down crossing argument as in Section III-A. However, now we equate the number of transitions into S and out of S in a busy cycle. Noting that $E[N_n]$ is the average number of transitions from state $n+1$ to n and $\lambda E[T_k]$ the average number of arriving batches, given state k, we get

$$E[N_n]=\lambda \sum_{k=0}^{n} E[T_k] \Pr\{G^{(k)} \geq n+1-k\}, \qquad n \geq 0. \tag{3.12}$$

Together, (3.11) and (3.12) allow the computation of the $E[T_n]$ and $E[N_n]$, as shown below.

1) Evaluate the constants A_{kn}.

2) Put $E[N_0] = 1$,

$$E[T_0]=\frac{1}{\lambda \cdot \Pr\{G^{(0)} \geq 1\}}.$$

3) Given that $E[T_0], \cdots, E[T_{n-1}], E[N_0], \cdots, E[N_{n-1}]$ are computed, solve

$$E[T_n]=E[N_n]A_{nn} + \text{func}\,(E[N_0], \cdots, E[N_{n-1}]) \quad \text{[cf. (3.11)]}$$

$$E[N_n]=\lambda E[T_n] \Pr\{G^{(n)} \geq 1\} + \text{func}\,(E[T_0], \cdots, E[T_{n-1}]). \quad \text{[cf. (3.12)]}$$

4) Return to step 3 if necessary.

5) Compute $E[T]=\sum_{n=0}^{\infty} E[T_n]$ and $E[N]=\sum_{n=0}^{\infty} E[N_n]$. Define

p_n steady-state distribution of the number of phases at an arbitrary epoch

q_n steady-state distribution of the number of phases at a (phase) departure epoch.

Then by the theory of the regenerative processes (cf. Stidham [8] and Ross [7]),

$$p_n=\frac{E[T_n]}{E[T]} \quad \text{and} \quad q_n=\frac{E[N_n]}{E[N]} \qquad \text{for all } n \geq 0.$$

Below, we specify some schemes for the evaluation of the constants A_{kn} in the case of exponential, hyperexponential, and Erlang service time distributions of phases. These schemes can be extended to more general phase type service time distributions.

In van Hoorn [4] a scheme is given to compute A_{kn} in the case of general service time distributions which is, however, less efficient than the scheme below.

Case 1: $F_S(t) = 1 - e^{-\mu t}$: Using the memoryless property of the exponential distribution and the property that with probability $\lambda/(\lambda+\mu)$ a batch of phases arrives before the completion of a service, we find

$$A_{kn}=\frac{\lambda}{\lambda+\mu} \sum_{i=0}^{n-k} g_i^{(k)} A_{k+i,n}, \qquad 1 \leq k < n.$$

$$A_{nn}=\frac{\lambda}{\lambda+\mu} g_0^{(n)} A_{nn}+\frac{1}{\lambda+\mu}, \qquad n \geq 1.$$

Starting with A_{nn}, the A_{kn} can be computed recursively for $k = n-1, \cdots, 1$.

Case 2: $F_S(t) = 1 - p_1 e^{-\mu_1 t} - p_2 e^{-\mu_2 t}$. We apply case 1 twice to compute $A_{kn}^{(1)}$ and $A_{kn}^{(2)}$, with μ replaced by μ_1 and μ_2, respectively, and then find

$$A_{kn}=p_1 A_{kn}^{(1)}+p_2 A_{kn}^{(2)}.$$

Case 3: $F_S(t) = 1 - (1+\mu t)e^{-\mu t}$: We define the auxiliary quantities:

B_{kn} = expected amount of time that during the second phase of the service n phases are present, given that the second phase of the service starts with k phases present.

Again, applying Case 1, B_{kn} can be computed and then we get

$$A_{kn}=\frac{\lambda}{\lambda+\mu} \sum_{i=0}^{n-k} g_i^{(k)} A_{k+i,n}+\frac{\mu}{\lambda+\mu} B_{kn}, \qquad 1 \leq k < n$$

$$A_{nn}=\frac{\lambda}{\lambda+\mu} g_0^{(n)} A_{nn}+\frac{1}{\lambda+\mu}+\frac{\mu}{\lambda+\mu} B_{nn}, \qquad n \geq 1.$$

For the numbers A_{0n} the following relation holds.

$$A_{0n}=\sum_{i=1}^{n} \frac{g_i^{(0)}}{1-g_0^{(0)}} A_{in}, \qquad n \geq 1.$$

Remark that $g_i^{(0)}/(1-g_0^{(0)})$ is the probability that an arriving batch initiating a busy period consists of j phases.

Remark 1: Putting $E[N_0] = 1$ in step 2 of the algorithm is motivated by the fact that in every busy cycle, the system is left behind empty only once after the completion of a service.

Remark 2: $E[T]$ and $E[N]$ can be computed as follows. Note that $\sum_{n=k}^{\infty} A_{kn} = h$ and $\sum_{n=k}^{\infty} \Pr\{G^{(k)} \geq n+1-k\} = E[G^{(k)}]$. By summing (3.11) for $n \geq 1$ and (3.12) for $n \geq 0$, we get

$$E[T]-E[T_0]=E[N] \cdot h$$

$$E[N]=\sum_{k=0}^{\infty} \lambda E[T_k] E[G^{(k)}]. \tag{3.13}$$

Using $E[G^{(k)}] = E[G]$ for $k \geq k_0$, (3.13) is rewritten as

$$E[N]=\lambda E[T] \cdot E[G]+\lambda \sum_{k=0}^{k_0-1} E[T_k](E[G^{(k)}]-E[G]).$$

So, $E[N]$ and $E[T]$ can be found after computing $E[T_k]$, $0 \leq k \leq k_0 - 1$. A comparison of different service phase distributions is given in Table I, where numerical results for the call completion rate are listed. In general, above a certain value of the mean batch size, the system performance is dominated by the batch size statistics, and the system is relatively insensitive to the service time distribution (cf. [5]).

IV. Investigation of an Overload Control Scheme

A. Description of the Control Scheme

In the previous section it can be seen that the system performance, say the call completion rate, has decreased rapidly above a critical level of the offered load. At these high load levels, the queue becomes large and customers must wait for a long time before they enter service. They then become impatient, tend to abandon their calls and, as a result, the call completion rate decreases.

In order to avoid this effect, the system may stop accepting all calls at a predefined load level. The idea behind it is that, if the switching system accepts fewer calls, it is able to handle them more effectively. As illustrated in Fig. 6, we can save

TABLE I
COMPARISON OF THE CALL COMPLETION RATE FOR DIFFERENT PHASE SERVICE TIME DISTRIBUTIONS ($N_0 = 4$, $N_1 = 8$, $N_{LIM} = 5$, $\gamma = 0.1$)

Offered Traffic Intensity	Phase Service Time Distribution		
ρ_0	E_3	M	H_2 ($c^2 = 3$)
0.1	0.099981	0.099975	0.099953
0.2	0.199375	0.199231	0.198698
0.3	0.294200	0.293248	0.290196
0.5	0.411585	0.406709	0.394015
0.7	0.356940	0.352679	0.342012
0.9	0.198035	0.196382	0.192602

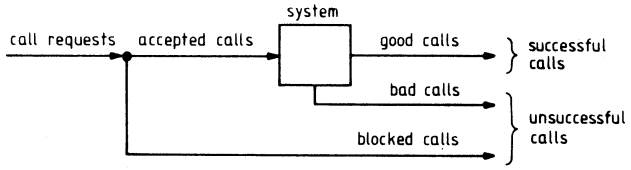


Fig. 6. On the call completion in a switching system.

processor time and increase the amount of good calls, if we allow the system to reject calls according to a scheme which will be described in the following. It should be noted here that the phenomenon of repeated attempts of blocked calls is not taken into account.

Two levels L_1 and L_2 are defined for the call blocking scheme. A call, seeing upon arrival k phases in the system, will be blocked with probability B_k, where we choose

$$B_k = \begin{cases} 0 & 0 < k \le L_1 \\ \dfrac{k - L_1}{L_2 - L_1} & L_1 < k < L_2 \\ 1 & k \ge L_2. \end{cases} \tag{4.1}$$

According to this scheme, the maximum number of phases the system can have is $L_2 + N_1 - 1$. The system attains this maximum when an accepted call sees $L_2 - 1$ phases in the system, and then adds N_1 phases for service. Hence, we have a queueing system with finite capacity $L_2 + N_1 - 1$.

To show the performance of the overload control method, we have chosen the linear characteristic of B_k in (4.1). In principle, from the analysis point of view, we could choose any other gradual blocking scheme for B_k between L_1 and L_2.

In the case of *one-level control* we can choose $L_2 = L_1 + 1$, and all call arrivals seeing at most L_1 phases will be accepted; otherwise they will be blocked.

B. Model Modification and Analysis

The overload control scheme, described by the two levels L_1 and L_2, reduces our queueing system to a finite capacity $M^{[X]}/G/1$ queue. The blocking probability, gradually increasing with the queue size a customer sees, equals 1 when there are more than $L_2 - 1$ phases in the system. As discussed above, the system has the finite capacity $M = L_2 + N_1 - 1$.

The most simple way to model blocking is to allow a customer to bring a "batch of size zero" into the system. Blocked customers do not affect the system by having a batch without phases.

The modified batch size distribution for the overload control scheme is denoted by $\bar{g}_j^{(k)} = \Pr\{\bar{G}^{(k)} = j\}$. We have the following relations between the probabilities $\bar{g}_j^{(k)}$ and $g_j^{(k)}$:

$$0 \le k \le L_1$$

$$\begin{cases} \bar{g}_0^{(k)} = 0 & j = 0 \\ \bar{g}_j^{(k)} = g_j^{(k)} & N_0 \le j \le N_1 \end{cases}$$

$$L_1 < k < L_2$$

$$\begin{cases} \bar{g}_0^{(k)} = B_k & j = 0 \\ \bar{g}_j^{(k)} = g_j^{(k)}(1 - B_k) & N_0 \le j \le N_1 \end{cases}$$

$$L_2 \le k \le M$$

$$\begin{cases} \bar{g}_0^{(k)} = 1 & j = 0 \\ \bar{g}_j^{(k)} = 0 & N_0 \le j \le N_1 \\ \bar{g}_j^{(k)} = 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

For the computation of the state probabilities, we again use (3.1) with the modified batch size distribution. The algorithm to calculate p_k, $0 \le k \le M$ is as follows.

1) Set $p_0 = 1$.
2) Compute $p_1, \cdots, p_M$ recursively with (3.1).
3) Renormalize $p_0, p_1, \cdots, p_M$.

The system characteristics can be written as follows.

$$P_{\text{COMPL}} = \sum_{j=N_0}^{N_1} c_j \sum_{k=0}^{M} p_k \bar{g}_j^{(k)}$$

$$E[X] = \sum_{k=0}^{M} k p_k$$

$$\text{MBS} = \frac{(1 - p_0)}{\lambda h (1 - P_{\text{BLOCK}})}$$

$$P_{\text{BLOCK}} = \sum_{k=L_1+1}^{M} p_k \bar{g}_0^{(k)}$$

$$P^*_{\text{COMPL}} = \frac{P_{\text{COMPL}}}{1 - P_{\text{BLOCK}}}. \tag{4.3}$$

P^*_{COMPL} is defined as the completion probability for accepted calls (cf. Fig. 6).

C. Results and Comparison

The performance of the overload control strategy will be discussed in the subsection.

Fig. 7 shows the interference between call blocking and call completion in the system. The dashed line stands for the case without overload control. With the simple control mechanism ($L_1 = 3N_1$, $L_2 = 4N_1$, linear call blocking between L_1 and L_2) the call blocking probability increases rapidly with higher offered traffic intensity, while the call completion probability P_{COMPL} lies above the curve without overload control. As expected, the system accepts fewer calls but is able to perform them well. For the accepted calls, P^*_{COMPL} gives us an idea about the fraction of good calls served by the system.

The call completion rate with overload control is depicted in Fig. 8. It can be seen clearly that the choice of the control levels (with $L_2 = L_1 + 1$) affects the system performance very strongly in overload situations. For $L_1 = N_1$, the system performance is worse in the case of lower traffic levels but becomes better for higher traffic intensities. Above a level of L_1/N_1 the call completion rate is always higher with overload control.

Table II compares the call completion rate using the

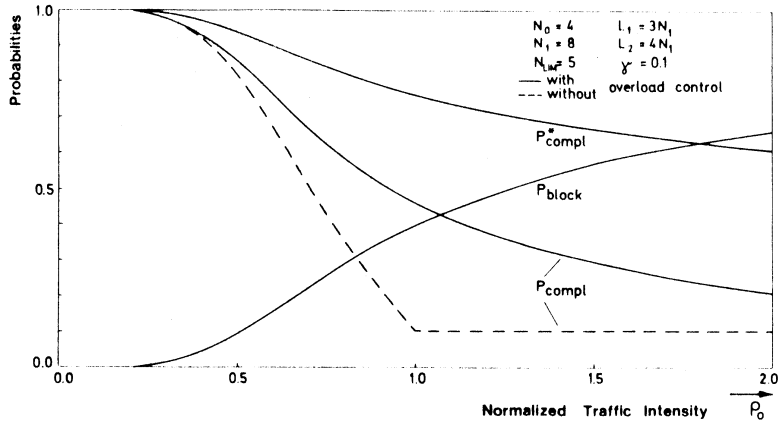


Fig. 7. Call completion with overload control.

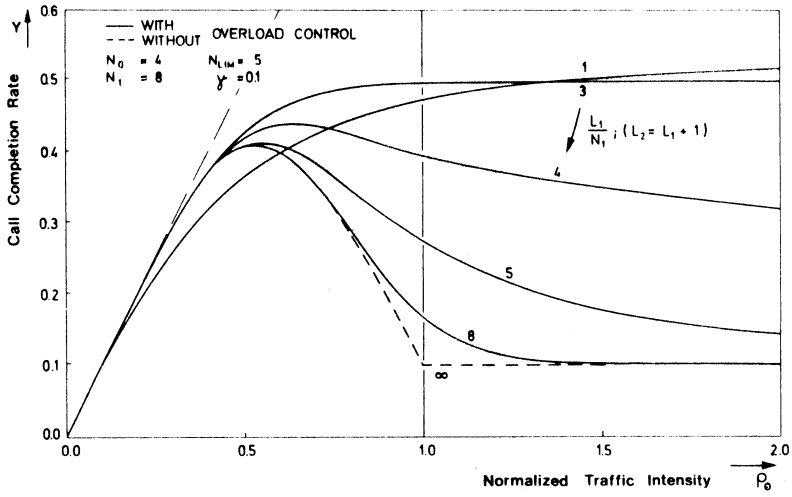


Fig. 8. Performance of the overload control strategy: call completion rate versus normalized traffic intensity.

TABLE II
CALL COMPLETION RATE WITH OVERLOAD CONTROL—A COMPARISON FOR DIFFERENT PHASE SERVICE TIME DISTRIBUTIONS ($N_0 = 4$, $N_1 = 8$, $N_{LIM} = 5$, $\gamma = 0.1$, CONTROL LEVELS: $L_1 = 3N_1$, $L_2 = 5N_1$)

Offered Traffic Intensity ρ_0	Phase Service Time Distribution		
	E_3	M	H_2 ($c^2 \geq 3$)
0.1	0.099980	0.099974	0.099949
0.3	0.294192	0.293271	0.290380
0.5	0.425019	0.421388	0.412192
0.7	0.445664	0.443856	0.439294
1.0	0.405236	0.407232	0.412044
1.2	0.375811	0.379266	0.387623
1.5	0.338031	0.343039	0.354784
2.0	0.290323	0.297062	0.312029

overload control strategy for different phase service time distributions. For the given batch statistics, the difference caused by the phase distributions is not essential. This argument justifies the Markovian phase modeling approach, which requires a simple analysis and less computing effort without losing the essential effects.

V. Conclusion and Outlook

In this paper a queueing model is presented in which the service time of a customer depends on his waiting time in the queue. The model is used to investigate the influence of customer behavior on the call completion characteristics of switching systems. It is shown that under overload conditions, the impatience of customers causes a serious degradation of the system performance. The influence of this effect can be controlled and minimized by using a very simple overload regulation scheme which is presented and discussed in Section IV. However, the improvement of the call completion rate by the overload control method depends strongly on the statistics of the customer behavior and on the call handling mechanism of the switching system, which is modeled by the number of subcalls according to a successful or unsuccessful call.

The modeling approach discussed in this paper can also be used for investigations of system performance under instationary conditions. In Tran-Gia [10] the system responses to

short-term, time-dependent overload patterns, and the transient behavior of the overload control method are considered. Furthermore, the analysis methods used in this paper can be applied for a wide range of systems and modeling approaches, using the freedom of the state-dependent batch size and the call service time distributions.

ACKNOWLEDGMENT

The authors would like to thank Prof. H. C. Tijms and Prof. P. J. Kuehn for helpful discussions.

REFERENCES

[1] M. Posner, "Single server queues with service time dependent on waiting time," *Oper. Res.*, vol. 21, pp. 610-616, 1973.

[2] L. J. Forys, "Modelling of SPC switching systems," in *Proc. 1st ITC Sem. Modelling of SPC Exchanges and Data Networks,* Delft, The Netherlands, 1977, pp. 83-100.

[3] M. Rosenshine, "Queues with state-dependent service times," *Transport. Res.*, pp. 97-104, 1967.

[4] M. H. van Hoorn, "Algorithms for the state probabilities in a general class of single server queueing sytems with group arrivals," *Management Sci.*, vol. 27, pp. 1178-1187, 1981.

[5] D. R. Manfield and P. Tran-Gia, "Analysis of a finite storage system with batch input arising out of message packetization," *IEEE Trans. Commun.*, vol. COM-30, pp. 456-463, 1982.

[6] C. M. Harris, "Some results for bulk-arrival queues with state-dependent service times," *Management Sci.*, vol. 5, pp. 313-326, 1970.

[7] S. M. Ross, *Applied Probability Models with Optimization Applications.* San Francisco, CA: Holden-Day, 1970.

[8] S. Stidham, Jr., "Regenerative processes in the theory of queues with applications to the alternating priority queue," *Adv. Appl. Prob.*, vol. 4, pp. 542-577, 1972.

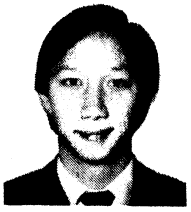
[9] B. T. Doshi and E. H. Lipper, "Comparison of service disciplines in a queueing system with delay-dependent customer behavior," in *Proc. ORSA-TIMS Conf.*, Appl. Prob. Comput. Sci., The Interface, 1981.

[10] P. Tran-Gia, "Subcall-oriented modelling of overload control in SPC switching systems," in *Proc. 10th Int. Teletraffic Congr.*, Montreal, P.Q., Canada, 1983.

[11] ———, "Modelling overload control in SPC switching systems" (36th Rep. Studies in Congestion Theory), Inst. Commun. Switching and Data Technics, Univ. Stuttgart, Stuttgart, West Germany, 1982.

[12] M. H. van Hoorn, "Algorithms and approximations for queueing systems," Ph.D. dissertation, Free Univ. Amsterdam, Amsterdam, The Netherlands, 1983.

★

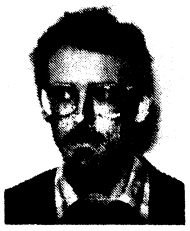


Phuoc Tran-Gia (M'80) was born in Vietnam. He received the M.S. (Dipl.-Ing.) degree from the University of Stuttgart, Stuttgart, West Germany, in 1977 and the Ph.D. (Dr.-Ing.) degree from the University of Siegen, Siegen, West Germany, in 1982, both in electrical engineering.

In 1977 he joined Standard Elektrik Lorenz (ITT), Stuttgart, where he was working in software development of digital switching systems. From 1979 to 1982 he worked as an Assistant Professor of Communications at the University of Siegen. Since 1983 he has been Head of a Research Group in the Institute of Communications Switching and Data Technics, University of Stuttgart. In 1985 he was a Lecturer for Computer Networks at the University of Würzburg, Würzburg, West Germany. His current research activities are in the field of environment simulations for communication systems, as well as queueing theory and its application in performance analysis for communication and computer systems.

Dr. Tran-Gia is a member of the German Communications Society (NTG).

★



Michiel H. van Hoorn was born in Amsterdam, The Netherlands. He received the M.S. degree from the University of Amsterdam in 1979 and the Ph.D. degree from the Free University of Amsterdam in 1983, both in applied mathematics and operations research.

In 1979 he joined the Free University of Amsterdam, where he worked as an Assistant Professor of Actuarial Sciences and Econometrics. There he worked together with Prof. Henk Tijms on numerical solution methods for single and multiserver queueing systems. He is a coauthor of the book *Tables for Multiserver Queues* (Amsterdam, The Netherlands: North-Holland, 1985). Since 1984, he has been working at the consultancy firm PACT Networks, which specializes in design and management of data communication networks.