Contribution to the Tenth International Teletraffic Congress (ITC), June 1983, Montréal, Canada.

Subcall-Oriented Modelling of Overload Control in SPC Switching Systems

P. Tran-Gia *

Dept. of Communications, University of Siegen
Siegen, Federal Republic of Germany

ABSTRACT

The performance of stored program controlled (SPC) switching systems is strongly influenced by their subcall handling efficiency, especially in overload situations. The subcall generation and handling process depends mainly on the interaction between the actual system effectiveness and the customer behaviour. In this paper overload control aspects in SPC switching systems are investigated by means of two subcall-oriented queueing models.

In the first model the subcall and call handling processes are considered together, whereby an overload control strategy, the bad-call-interruption scheme (BCI), is presented and investigated. The interdependency between the subcall generation process, the call completion characteristics and the call waiting time according to the customer behaviour, is pointed out in the second model. A state-dependent call throttling method is analyzed, whereby the instationary system reaction on time varying overload traffic patterns and the instationary performance of the overload control scheme are investigated.

1. INTRODUCTION

In modern stored program controlled switching systems, a modular structure is often provided for both hard- and software. The system functions are divided and allocated among a number of autonomous processors and processes. Due to the modular structure the number of tasks or telephonic events *(subcalls)* generated per call attempt is larger in comparison with conventional, centralized controlled SPC switching systems. Thus, the overall performance of the switching system is strongly influenced by the subcall handling efforts, which become a critical factor in overload situations.

Subcall-oriented performance considerations in telephone switching systems are given in a number of papers [3-6]. Basic models for overload control mechanisms are discussed in [2, 6] and [12]. Detailed descriptions of subcall traffic are considered in [4, 5] and [11], where simulations studies are presented. Several aspects concerning the modelling of SPC switching system are outlined in [3].

The subcall-oriented modelling technique offers an appropriate tool for performance investigations in SPC switching systems. By means of this method, the call handling process can be modelled and investigated in more detail, where the system behaviour in overload situations and the effectiveness of overload control strategies can be realistically investigated. Furthermore, to account for the observed subcall generation process with respect to subscriber behaviour, models describing the interdependency between the call waiting time, the subcall traffic and the call completion rate can be developed.

2. COMPOUND MODEL FOR CALL AND SUBCALL HANDLING

2.1 Modelling approach and overload control strategy

The subcall traffic characteristics in a SPC switching system is strongly affected by the call arrival and handling process. In order to model this influence in detail, both the call and subcall processes must be taken into account. By means of the compound model for call and subcall handling developed in this chapter, the dependency of the call completion characteristics on the subcall handling efficiency will be investigated.

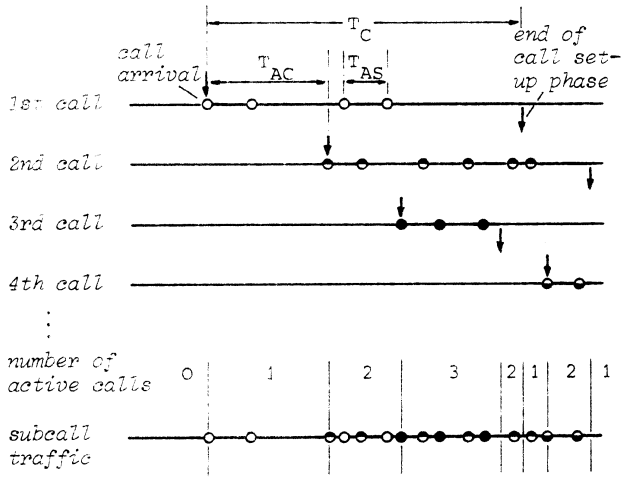

Fig. 2.1 Subcall generation and superposition in switching systems

*) The author is now with the University of Stuttgart, Stuttgart, FRG.

Fig. 2.1 illustrates the generation and the superposition of subcalls as well as the relationship between call and subcall traffic streams. It can be clearly seen here that the subcall stream constitutes a non-recurrent process. Its intensity is modulated by the number of active calls in the system.

The random variable (r.v.) T_{AC} stands for the interarrival time of calls, which follow a Poisson process with the rate λ. Considering the fact that the most subcalls are produced during the set-up phase T_C of a call, this phase will be used in the model as the main subcall producing phase. Taking into account the mixture of call types in a switching system, (e.g. successful or abandoned calls, facilities calls,...) the r.v. T_C is assumed to be negative exponentially distributed. The intervals between generated subcalls, which belong to a call, are characterized by the independently identically distributed r.v. T_{AS}. Because of the large number of subcalls produced per call in SPC switching systems and considering the different types of subcalls produced by different call controlling processes (call handling, signalling, device handling, etc...), T_{AS} is assumed to be negative exponentially distributed. Thus, the subcall process corresponding to an active call is characterized by a Poisson process with the rate λ_S.

The compound model for call and subcall handling is depicted in Fig. 2.2. An accepted call request activates a subcall source (subcall generator), which produces subcalls with the rate λ_S during the call set-up phase. After this time, corresponding to the r.v. T_C mentioned above, the subcall source will be idled. According to the call handling capacity of the switching system, the number of subcall sources is limited by n.

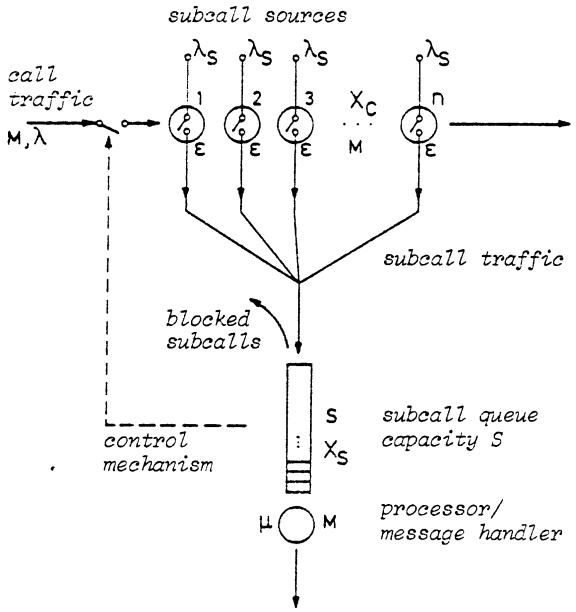

Fig.2.2 Compound model for call and subcall handling

Subcalls generated by active sources are superposed and form the subcall traffic stream offered to the subcall handler, which models the processor or the control unit. The subcall queue has the finite capacity S. Taking into account all subcall types requiring different processing times, the subcall handler is approximated to be a Markovian server with the mean service time $1/\mu$.

When the subcall queue is completely filled, further subcalls offered will be considered as blocked. A call which has produced a blocked subcall cannot be completed and is said to become a *bad-call*. The performance of the switching system will be degraded by bad-calls, since the subcall handler has spent an ineffective amount of processing time to handle subcalls belonging to bad-calls. Furthermore, according to signalling delays and customer reactions, a bad-call will produce further ineffective subcalls. These effects reduce in a dramatic way the system performance in overload situations.

The considerations discussed above lead to an overload control mechanism, the *bad-call-interruption* method (BCI), which allows a more effective use of the system real-time capacity. In accordance with the BCI strategy, a call is cleared immediately after producing a blocked subcall. This scheme can be implemented by hard- or software, e.g. by means of an instantaneous changing of the event scanning process for active calls. The aim of the BCI strategy is a reservation of processor capacity to handle completable calls (good-calls), in order to optimize the call completion rate of the switching system.

2.2 Analysis

The actual system state can be described by means of the two r.v. X_C and X_S, which represent the actual number of calls and subcalls in the system, respectively. According to the memoryless property of the model components, the analysis can be done by means of a two-dimensional Markov process, using standard techniques of queueing theory [1].

With the state probabilities

$$P(i,j)=Pr\{X_S=i,X_C=j\},i=0,..,S+1,j=0,..,n, \qquad (2.1)$$

the corresponding state transition diagram can be developed as shown in Fig. 2.3. In this diagram, Θ is an auxiliary variable indicating the presence of the BCI overload control strategy, where the values $\Theta=1$ and $\Theta=0$ stand for the two cases with and without bad-call-interruption, respectively.

According to the state transition diagram, a set of difference equations for the steady state is derived, with which the state probabilities are numerically calculated using an iteration method with overrelaxation (see [12]).

The mean number E_C of calls and E_S of subcalls in system can be determined by :

$$E_C = \sum_{i=0}^{S+1} \sum_{j=0}^{n} j.P(i,j), \qquad (2.2a)$$

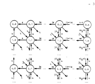

Fig.2.3 The state transition diagram

and

$$E_S = \sum_{i=0}^{S+1} \sum_{j=0}^{n} i.P(i,j). \qquad (2.2b)$$

Using the Little's law, the mean sojourn time T_S for accepted subcalls is calculated :

$$T_S = E_S/\lambda_{AS} \qquad (2.3a)$$

where λ_{AS} is the mean arrival rate for accepted subcalls :

$$\lambda_{AS} = \sum_{i=0}^{S} \sum_{j=0}^{n} (j\lambda_S+\lambda).P(i,j). \qquad (2.3b)$$

Subsequently, the blocking probability B_C for calls and B_S for subcalls can be computed :

$$B_S = \frac{1}{\lambda+\lambda_S E_C} \sum_{j=0}^{n} (j\lambda_S+\lambda).P(S+1,j) \qquad (2.4)$$

$$B_C = \sum_{i=0}^{S} P(i,n) + \sum_{j=0}^{n} P(S+1,j) + B\frac{\lambda_S}{\lambda} \sum_{j=1}^{n} jP(S+1,j). \qquad (2.5)$$

In eqn. (2.5) it can be seen that the call blocking probability consists of three components corresponding to: the shortage of subcall sources, the call blocking owing to rejected first subcalls, and the call clearance in consequence of the BCI overload control strategy.

The call completion rate can be calculated for the case with BCI as follows:

$$Y = \lambda(1-B_C) \qquad (2.6)$$

In the case without BCI, the call completion rate Y cannot be derived directly from the state probabilities, since a bad-call producing several blocked subcalls cannot be identified using the steady state analysis discussed above. Therefore, for comparison purposes, the call completion rate is obtained by means of simulations.

It should be noted here, that the r.v. G for the number of subcalls per call - without BCI and under the assumptions for T_C and T_{AS} mentioned above - is geometrically distributed :

$$g_x = P[G=x] = \int_0^{\infty} \frac{(\lambda_S t)^{x-1}}{(x-1)!} e^{-\lambda_S t} . \varepsilon e^{-\varepsilon t} dt$$

$$= \left[1 - \frac{\lambda_S}{\lambda_S+\varepsilon}\right] \left[\frac{\lambda_S}{\lambda_S+\varepsilon}\right]^{x-1} , \quad x=1,2,\ldots \qquad (2.7)$$

with the mean

$$E[G] = 1 + \lambda_S/\varepsilon . \qquad (2.8)$$

2.3 Effectiveness of the BCI overload control method

In this section, numerical results are presented to point out the efficiency of the BCI method. The traffic intensities are normalized by $\rho_C=\lambda/(n\varepsilon)$ and $\rho_S=n\lambda_S/\mu$ for calls and subcalls, respectively. A comparison of the two cases with and without BCI is shown in Fig. 2.4 where the subcall blocking probability B_S is depicted as a function of the subcall queue capacity S. It can be clearly seen here that by means of the BCI strategy B_S can be remarkably reduced.

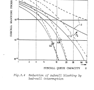

Fig.2.4 Reduction of subcall blocking by bad-call interruption

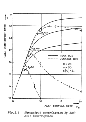

Fig.2.5 Throughput optimization by bad-call interruption

Fig. 2.5 shows the performance of the BCI strategy in overload situations, where the dependency of the call completion rate Y on the call traffic intensity is depicted. The performance decline at higher load levels can be effectively prevented by means of the BCI strategy. It should be noted here that an assumption is made in this investigation, that the overhead spent for BCI tasks is very small in comparison with the subcall handling efforts.

Simulation studies were done for a more realistic subcall generation model, with which the discussed results are validated. The results for the case without BCI shown in Fig. 2.5 are given with their 95 percent confidence intervals.

3. MODELLING WITH RESPECT TO CUSTOMER BEHAVIOUR

3.1 Backgrounds and modelling approach

The characteristics of the subcall generation process and the performance of a switching system are strongly affected by the interaction between the system efficiency and the reaction of customers. On the one hand, when a customer is confronted with large delays during the call set-up phase (e.g. waiting for dial tone, post-dialling delay, etc...), he may abandon his call. In this case, an ineffective amount of processing time has been provided for handling subcalls belonging to the unsuccessful call. Hence, the call completion rate of the switching system will decrease. On the other hand, unsuccessful or rejected customers may reattempt their call requests and further inflate the overload.

There are a number of studies which consider waiting time dependent service time with varying degrees of complexity. [7] and [8] analyze single server queueing systems with respect to the waiting time dependency. In [3] a basic model is discussed for applications in telephone switching systems where customers can have two exponentially distributed call durations, depending on their waiting time. The patience of customers, in accordance with call waiting time, is also considered in the overload control strategy described in [10], which is designed for No. 1 ESS switching system type.

In this chapter, a model will be developed and analyzed, which takes into account the waiting time dependency of the service time and the subcall generating process of a call, in order to estimate the call completion capacity of SPC switching systems in overload situations [9, 12]. On account of the model complexity the phenomenon of repeated attempts is disregarded.

The model components describing the dependency between the waiting time, the number of subcalls generated and the completion probability of a call will be discussed below.

A test call request enters a switching system and sees an amount of work in front of it waiting for processing. This amount stands for the number of subcalls buffered in the processor queue. Again, the subcall service time, which will be called in the following a phase, is assumed here to be negative exponentially distributed with the mean $h=1/\mu$. Thus, the waiting time W of the test call before entering service can be determined by the number k of phases seen upon arrival, which leads to an Erlangian distribution of order k:

$$P\{W \le t\} = E_k(t) = 1 - e^{-\mu t} \sum_{i=0}^{k-1} \frac{(\mu t)^i}{i!}, \quad t \ge 0. \qquad (3.1)$$

According to the customer behaviour, the number j of subcalls (phases) which will be generated by the test call depends on the call waiting time, i.e. on the actual value of k. This dependency is described here by means of a state-dependent batch arrival process. A call, which arrives according to a time-dependent Poisson process (generalized Poisson process) with the rate $\lambda(t)$ and meets k service phases in the system upon arrival, will generate j subcalls (phases) with the probability $g_j^{(k)}$.

Based on observations in SPC switching systems, where calls with incompleted dialling or abandoned calls usually produce less subcalls than successful calls with completed dialling, a relationship between the call completion probability and the number of subcalls generated by a call can be provided. This will be modelled by the conditional call completion probability c_j, with which a call having produced j subcalls will be completed.

Hence, the subcall-oriented model developed here is of type $M^{[X]}(t)/M/1$ with time-varying, state-dependent batch arrivals. Although the analysis

presented in the next part is provided for arbitrary characteristics of $g_j^{(k)}$ and c_j, these functions will be defined in the following according to observations made in switching systems.

The number of subcalls produced by a call may vary between a lower limit N_0 and an upper limit N_1. An additional assumption is made, that a call seeing a very large number of phases, say at least k_0 phases, will tend to become a bad-call and generate the minimum number N_0 of subcalls or phases. Taking additionally into account the call waiting time according to eqn. (3.1), the probability $g_j^{(k)}=Pr\{G^{(k)}=j\}$ ($G^{(k)}$ denotes the r.v. for the state-dependent subcall number) is assumed to be:

$$k=0 \quad : g_{N_1}^{(0)}=1 \ , \ g_j^{(0)}=0 \ \text{ otherwise},$$

$$0<k<k_0 : g_{N_0}^{(k)}=1-E_k[(N_1-N_0)h],$$

$$g_j^{(k)}=E_k[(N_1-j+1)h]-E_k[(N_1-j)h], N_0<j\le N_1,$$

$$k=k_0 \quad : g_{N_0}^{(k_0)}=1 \ , \ g_j^{(k_0)}=0 \ \text{ otherwise}. \quad \ldots(3.2)$$

The reasoning for the choice in eqn. (3.2) is illustrated in Fig. 3.1, where the state-dependent mean number of subcalls produced by a call is shown. It can be seen here that the observation, after which calls with short waiting times tend to produce more subcalls, is contained in the definition in eqn. (3.2).

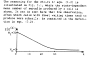

Fig.3.1 State-dependent mean number of subcalls generated per call

For the numerical studies discussed in this chapter, the completion probability c_j for a call which produces j subcalls is defined by means of the characteristic illustrated in Fig. 3.2. The simple form of the function for c_j takes into account the modelling argument, that successful calls often generate more subcalls than blocked or abandoned calls.

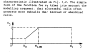

Fig.3.2 Example for conditional completion probability

3.2 Stationary analysis

The analysis of the presented model under stationary conditions is given in [9] and [10], where a general phase distribution is considered. The stationary state probabilities are calculated by means of a numerical algorithm. In order to analyze the system reaction on short-term overload patterns, the model analysis discussed below will be restricted to the instationary case. However, to reduce the numerical efforts, the algorithm mentioned above [9, 10] is used for the calculation of initial state probabilities required for the instationary analysis.

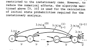

Fig.3.3 State transition diagram of the $M^{[X]}(t)/M/1$ queue

3.3 Instationary analysis

With X(t) as the r.v. for the number of phases at time t and $P_i(t)=Pr\{X(t)=i\}$, $i=0,\ldots,\infty$, the time-dependent state probabilities, the state transition diagram of the queuing system $M^{[X]}(t)/M/1$ can be developed as shown in Fig. 3.3. Observing the interval $(t,t+dt)$, $dt\to 0$, and taking into account the balance property of the process X(t) during dt, the following set of differential state equations is derived:

$$\frac{d}{dt}P_0(t) = -\lambda(t)P_0(t)+\mu P_1(t),$$

$$\frac{d}{dt}P_i(t) = -(\mu+\lambda(t))P_i(t)+\lambda(t)\sum_{n=0}^{i-1}P_n(t)g_{i-n}^{(n)} +\mu P_{i+1}(t), \quad i=1,2,\ldots \qquad (3.3)$$

The time-dependent state probabilities are determined by solving the differential equation system given in (3.3), where a proper truncation of the state space is provided. The obtained state probabilities form the basic requirement for the calculation of the time-dependent system characteristics.

A call request arriving at time t will become a successful call with the call completion probability

$$C(t) = \sum_{k=0}^{\infty} P_k(t) \sum_{j=N_0}^{N_1} c_j g_j^{(k)}. \qquad (3.4)$$

The time-dependent call completion rate Y(t) is

$$Y(t) = \lambda(t).C(t) \ . \qquad (3.5)$$

In the following, the call arrival rate $\lambda(t)$ and the call completion rate Y(t) are standardized by:

$$\Omega_0(t) = \lambda(t).\ N_0.h \ , \qquad (3.6a)$$

$$Y_0(t) = Y(t).\ N_0.h \ . \qquad (3.6b)$$

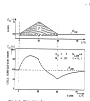

Fig.3.4 Time-dependent system response to a triangular overload pattern

3.4 System reaction on time-dependent overload patterns

In the following, overload traffic will be modelled by means of short-term, time-continuous pulse patterns. Fig.3.4 shows the response of the system to a triangular overload traffic pattern, where the time-dependent call completion rate $Y_0(t)$ is depicted. The overload pattern is characterized by the length T_P and the amount P of additional load.

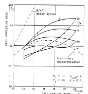

Fig.3.5 Trajectories of time-dependent system responses

It can be clearly seen here that during the recovery time, in which the overload traffic decreases and returns to the normal load condition ρ_{0S}, the transient call completion rate is temporary beneath the value Y_{0S} of the stationary case. The reason for this effect is that during the recovery time the previous overload is still in the system and the call waiting time is longer on average than in the stationary case.

The trajectories of time-dependent system responses to triangular overload patterns are shown in Fig. 3.5 for different length T_P of the overload situation. It is obvious that for $T_P = \infty$ the trajectory will be identical with the stationary call completion characteristics.

3.5 Overload control by means of call throttling

In order to improve the system performance under overload, the waiting time for calls before entering service has to be minimized. A basic scheme is to control the call arrival stream as well as the call acceptance, where the number of subcalls (phases) in the system is used as overload indicator. The number of subcalls in the system will be limited by means of a state-dependent call throttling method. In the following the effectiveness of this class of overload control mechanisms will be discussed by means of an example shown in Fig. 3.6. According to this scheme, a call will be rejected with the probability B_k, where k is the actual system state k at arrival epoch.

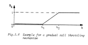

Fig.3.6 Example for a gradual call throttling mechanism

The analysis described in subsection 3.3 can also be applied for the case with overload control, whereby a modification for the state-dependent number of subcalls produced by a call is made. The probabilities $q_j^{(k)}$ are replaced by $\tilde{q}_j^{(k)}$:

$$\tilde{q}_0^{(k)} = B_k \;, \quad \tilde{q}_j^{(k)} = q_j^{(k)} (1-B_k) \quad j > 0 \qquad (3.7)$$

The influence of the gradual call blocking scheme on the system response to the triangular overload pattern in Fig.3.4 is shown in Fig.3.7, where an improvement of the system reaction can be recognized.

In order to enable comparisons concerning the overload performance, a measure has to be defined. For this purpose, the quantity C_{AD} is introduced, which stands for the completion probability of additional traffic :

$$C_{AD} = \frac{\{\text{amount of completed additional traffic}\}}{\{\text{total amount of additional traffic}\}} \qquad (3.8)$$

Fig.3.7 System response to a triangular overload pattern with overload control (cf. Fig.3.4)

A comparison of the dynamical overload performance is given in Fig. 3.8, where C_{AD} is depicted as a function of the total amount F of additional overload traffic. It can be seen that above a level of F, say F>10, a remarkable improvement of the overload handling efficiency is obtained by means of the overload control method.

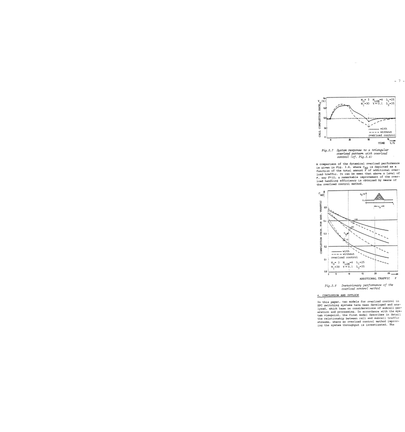

Fig.3.8 Instationary performance of the overload control method

4. CONCLUSION AND OUTLOOK

In this paper, two models for overload control in SPC switching systems have been developed and analyzed, which base on considerations of subcall generation and processing. In accordance with the system viewpoint, the first model describes in detail the relationship between call and subcall traffic streams, where an overload control method improving the system throughput is investigated. The model can be applied in a wide range of SPC switching systems with centralized or distributed control structures, where dimensioning aspects of the subcall handler or the control unit are of interest.

In the second model, the customer behaviour is taken into account, whereby the waiting time dependency of the subcall generation and processing is considered. Investigations of time-dependent system reaction to time-varying overload traffic patterns are presented and an overload control strategy is analyzed. The performance considerations under instationary load conditions represent a technique, which allows investigations of the effectiveness of overload control strategies in a more realistic way. The hysteresis of the system throughput corresponds to observations made in switching systems in short-term overload situations.

The subcall-oriented modelling technique presented in this paper forms a basic requirement for the investigations of the overload handling efficiency in SPC switching systems. These studies will be continued by system-dependent environment simulations, which are used for investigations of particular overload control schemes, towards the design of a general overload control plan.

ACKNOWLEDGEMENTS

The author would like to thank Professor P.J.Kuehn for helpful discussions and for his continued interest in this study. The programming efforts of Dipl.-Ing. E. Radant and Dipl.-Ing. H. Juechter are greatly appreciated.

REFERENCES

[1] Cooper, R.B., "Introduction to Queueing Theory". Mac Millan Co., New York, 1972.

[2] Romoeuf, L., "Modelling a Stored Program Controlled Telephone Switching System: Evaluating a Regulation Method for Traffic Overloads". Symp. on Meas., Mod. and Eval. of Comp. Systems, North Holl. Publ. Co., 1977.

[3] Forys, L.J., "Modelling of SPC Switching Systems". Proc. of the 1st ITC Seminar on Modelling of SPC Exchanges and Data Networks, Delft, 1977, pp. 83-100.

[4] Dietrich, G., Salade, R., "Subcall-Type Control Simulation of SPC Switching Systems". Proc. 8th ITC, Melbourne, 1976.

[5] Dietrich, G., "Telephone Traffic Model for Common Control Investigations". Proc. 7th ITC, Stockholm, 1973.

[6] Zaboste, F. C., "Optimal Control and Call Acceptance in a SPC Exchange". Proc. 9th ITC, Torremolinos, 1979.

[7] Posner, M., "Single Server Queues with Service Time Dependent on Waiting Time". Operations Research, 21 (1973), pp. 610-616.

[8] Rosenshine, M., "Queues with State-dependent Service Times". Transpn. Res., 1 (1967), pp. 97-104.

[9] Tran-Gia, P., Van Hoorn, M.H., "Dependency of Service Time on Waiting Time in Switching Systems - A Queueing Analysis with Aspects of Overload Control". Internal Report, University of Siegen - Free University Amsterdam, 1982.

[10] Borcheding, J.W. et al., "Coping with Overloads". Bell Lab. Record, 7/8 (1981), pp. 183-185.

[11] Wizgall, M., "On Architecture, Operating Mode and Traffic Behaviour of the Control of an SPC Switching Exchange". Ph. D. Thesis, University of Stuttgart, 1980 (in German).

[12] Tran-Gia, P., "Modelling Overload Control in SPC Switching Systems". Ph. D. Thesis, University of Siegen, 1982 (in German).