PREPRINT

Proceedings
Teletraffic Congress
Torino, Italy, 1-8 June, 1988

ANALYSIS OF A LOAD-DRIVEN OVERLOAD CONTROL MECHANISM IN DISCRETE-TIME DOMAIN*

Phuoc TRAN-GIA

IBM Research Division, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland

A general class of overload control mechanism for communication and switching systems is modeled and analyzed in this paper. The overload indicator is considered to be the amount of unfinished work in the system. The overload control is based on a throttling mechanism for new arriving requests when the load status of the system is above a predefined threshold. The input process, modeling the subscriber requests, is assumed to be general and can be considered under stationary or nonstationary conditions. To model the main mechanism of this class of overload control strategy, a generic queueing system of type *G/G/1 with feedback*, in conjunction with a workload-controlled acceptance scheme is used. The analysis method works in the discrete-time domain and allows use of efficient discrete transform algorithms [e.g., Fast Fourier Transform (FFT)] to determine the system characteristics. In discussing dimensioning aspects of the overload control strategy, numerical results are given for different types of input processes and overload threshold parameters.

1. OVERLOAD CONTROL MODELING

In modern communication systems the overload phenomenon and its influence on system performance has become more critical and complex due to the increasing number of new system features and customer facilities provided. To guarantee proper system performance, sensitivity of communication systems against overload must be taken into account in the design and development phase as well as in the post-cutover phase of a switching system.

Modeling approaches for investigation of the overload phenomenon in communication systems, especially in telephone switching systems [1-9], can be classified in two categories:

i) Overload modeling: evaluation models for overload indicators and time-dependent requirements for overload detection which include models describing the dynamics of overload situations and those describing customer behavior.
ii) Overload control modeling: models for performance evaluation of overload control strategies, where aspects like timeliness, influence on other subsystems, etc., are considered.

The class of overload control strategy operating with mechanisms which throttle the input process is often implemented in communication systems, especially in stored program controlled (SPC) switching systems. The main subject of this paper is this class of control strategy in conjunction with a generic queueing model.

The main mechanism of these overload control strategies is based on workload-driven call-acceptance schemes. We consider two levels of traffic: i) *call traffic*, representing the process of subscriber requests and ii) *workload*, standing for the amount of tasks or atomic-processing activities generated during the lifetime of a call. In contrast to various overload control schemes known in the literature [1,6-9], the actual *workload* (i.e. the amount of unfinished work in the system) is taken as the overload indicator instead of the

*The major part of this work was done while the author was with the Institute of Communications Switching and Data Technics, University of Stuttgart, Stuttgart, FRG.

number of active calls in the system. When the workload exceeds a defined threshold, arriving customers are rejected.

The basic model of the class of overload control strategy considered in this paper is of type G/G/1 with a feedback-controlled input process. This generic model leads to a modified form of the Lindley integral equation [10], when the analysis is derived in the continuous-time domain. In this case, solutions are only given for some specific systems or as approximations [1]. To obtain a generally applicable exact calculation algorithm, analysis methods operating in the discrete-time domain are employed in this paper (cf. [11-14]). A further obvious justification of the discrete-time approach is the fact that the parameterization of model components is often based on measured data in terms of histograms.

An algorithm to calculate system characteristics (e.g., customer blocking probability) is developed to estimate the performance of the overload control strategy. The discrete-time analysis technique used here takes advantage of efficient convolution and transformation algorithms (e.g., the Fast Fourier Transform FFT [3,6,11]). The results obtained show system behavior under stationary and nonstationary load conditions.

2. BASIC MODEL FOR LOAD-DRIVEN OVERLOAD CONTROL

The basic structure and related parameters of the overload control model are illustrated in Fig. 1. In principle, the model has the structure of a *G/G/1-system with a feedback control path*. Note that although the model and the analysis presented can be more generally applied, we restrict ourselves to the context of call-acceptance and set-up procedures in switching systems. Some motivations for applications in call-control processes of switching systems are mentioned below.

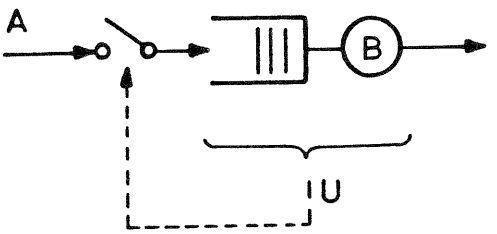

FIGURE 1
The basic load-driven overload control model.

The input process, which models the call arrival process and the service time which models the call set-up time are assumed to be general. In the following, a *call* will be also referred to as a *customer*. A further assumption made for these two processes is that they can be customer-dependent, i.e., the interarrival time and the service time can be individually chosen for each customer. Reasons for this general assumption is useful e.g., in the modeling of realistic call mixes in systems with integrated services and in the construction of nonstationary overload patterns offered to a system.

As indicated in Fig. 1, the following symbols and random variables (r.v.) are used:

- A_n r.v. for the interarrival time between the n-th and the (n + 1)-st customer.
- B_n r.v. for the service time of the n-th customer.
- U_n r.v. for the amount of unfinished work (workload) which remains in the system (e.g., the number of call handling tasks to be executed) immediately prior to the arrival instant of the n-th customer. This measure is used as overload indicator.
- L threshold value of the overload control mechanism.

The main mechanism of the overload control strategy discussed is based on the following workload-driven customer acceptance scheme:

$U_n < L$ the arriving call will be accepted

$U_n \geq L$ the arriving call will be rejected

3. ANALYSIS IN THE DISCRETE-TIME DOMAIN

3.1. Notation

In the context of discrete-time analysis, we consider the random variables to be of discrete-time nature, i.e., the time axis is divided into intervals of unit length Δt. As a consequence, samples of those random variables are integer multiples of Δt; the time discretization is equidistant.

The following notation is used for functions belonging to a discrete-time random variable (r.v.) X:

$x(k) = Pr(X = k), \quad -\infty < k < +\infty$ distribution (probability mass function) of X

$X(k) = \sum_{i=-\infty}^{k} x(i), \quad -\infty < k < +\infty$ distribution function of X

$EX, \; c_X$ mean and coefficient of variation of X

3.2. Outline of the Analysis

A sample of the state process development in the system is shown in Fig. 2. Observing the n-th customer in the system and the condition for customer acceptance upon arrival instant, the following conditional r.v. for the workload is introduced:

$$U_{n,0} = U_n | U_n < L, \qquad U_{n,1} = U_n | U_n \geq L, \tag{3.1}$$

$$U_{n+1,0} = U_{n+1} | U_n < L, \qquad U_{n+1,1} = U_{n+1} | U_n \geq L. \tag{3.2}$$

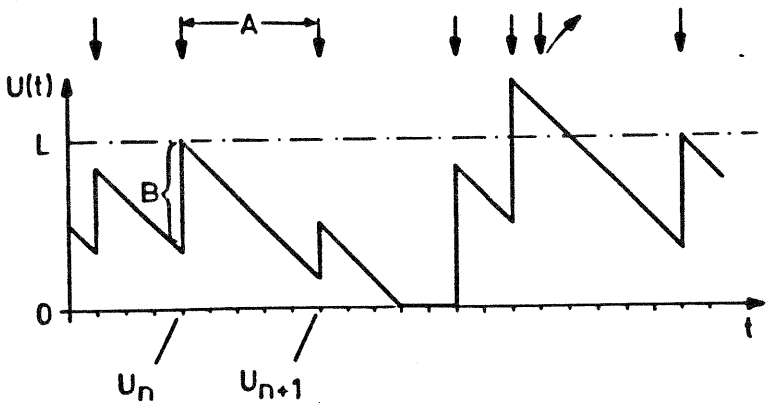

FIGURE 2
Sample path of unfinished work in the system.

Thus, the distributions of these random variables can be obtained:

$$u_{n,0}(k) = \frac{\sigma^{L-1}[u_n(k)]}{Pr\{U_n < L\}} = \frac{\sigma^{L-1}[u_n(k)]}{\sum_{i=0}^{L-1} u_n(i)} \tag{3.3}$$

$$u_{n,1}(k) = \frac{\sigma_L[u_n(k)]}{Pr\{U_n \geq L\}} = \frac{\sigma_L[u_n(k)]}{\sum_{i=L}^{\infty} u_n(i)} \tag{3.4}$$

where $\sigma^m(.)$ and $\sigma_m(.)$ are operators which truncate a part of a probability distribution function. The results of these operations are unnormalized distributions as follows:

$$\sigma^m[x(k)] = \begin{cases} x(k) & k < m \\ 0 & k \geq m \end{cases} \tag{3.5}$$

$$\sigma_m[x(k)] = \begin{cases} 0 & k < m \\ x(k) & k \geq m \end{cases} . \tag{3.6}$$

Taking the development of the process (cf. Fig. 2) with the overload control strategy discussed above into account, the following relationships between the random variables and their distributions can be obtained:

i) $U_n < L$: customer acceptance

$$U_{n+1,0} = \max(0, U_{n,0} + B_n - A_n) \tag{3.7}$$

$$u_{n+1,0}(k) = \pi_0[u_{n,0}(k) \star b_n(k) \star a_n(-k)] , \tag{3.8}$$

where the operator π_m is defined as

$$\pi_m(x(k)) = \begin{cases} 0 & k < m \\ \sum_{i=-\infty}^{m} x(i) & k = m \\ x(k) & k > m \end{cases} \tag{3.9}$$

and the $\star$-symbol denotes the discrete convolution operation:

$$a_3(k) = a_1(k) \star a_2(k) = \sum_{j=-\infty}^{+\infty} a_1(k-j) \cdot a_2(j) . \tag{3.10}$$

ii) $U_n \geq L$: customer rejection

$$U_{n+1,1} = \max(0, U_{n,1} - A_n) \tag{3.11}$$

$$u_{n+1,1}(k) = \pi_0[u_{n,1}(k) \star a_n(-k)] . \tag{3.12}$$

The distribution of the workload seen by the $(n+1)$-st customer can be written as:

$$u_{n+1}(k) = Pr\{U_n < L\} \cdot u_{n+1,0}(k) + Pr\{U_n \geq L\} \cdot u_{n+1,1}(k) . \tag{3.13}$$

From Eqs. (3.8), (3.12) and (3.13), we finally arrive at a recursive relation to calculate the workload at arrival epochs of customers:

$$\begin{aligned} u_{n+1}(k) &= \pi_0\left[\sigma^{L-1}[u_n(k)] \star b_n(k) \star a_n(-k)\right] + \pi_0\left[\sigma_L[u_n(k)] \star a_n(-k)\right] \\ &= \pi_0\left[\left(\sigma^{L-1}[u_n(k)] \star b_n(k) + \sigma_L[u_n(k)]\right) \star a_n(-k)\right] \end{aligned} . \tag{3.14}$$

Using Eq. (3.14), an algorithm for the calculation of the workload prior to customer arrivals can be found for both stationary and nonstationary traffic conditions. The computational diagram is shown in Fig. 3.

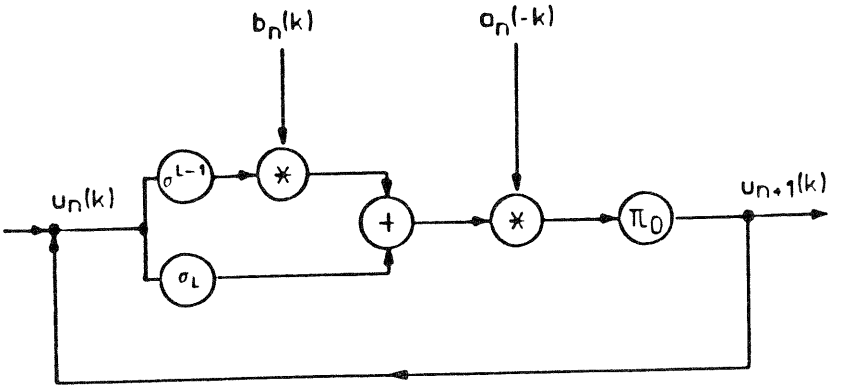

FIGURE 3
Computational diagram for G/G/1-system with feedback control.

The blocking probability of the n-th arriving customer is

$$B_{A,n} = \sum_{i=L}^{\infty} u_n(i). \tag{3.15}$$

4. SOME NUMERICAL RESULTS

Some results for the dimensioning of the overload control threshold where the discrete time axis is scaled to $\Delta t = 1$ will be discussed briefly. Time variables are normalized to the mean service time $EB = 20\ \Delta t$. The offered traffic intensity is denoted by $\rho = EB/EA$.

To obtain a parametric representation of random process types, we consider the interarrival and service times having distributions given by their two parameters, e.g., the mean and the coefficient of variation, whereby the negative binomial distribution is employed (cf. [14]). Thus, for a r.v. X with mean EX and coefficient of variation c_X

$$x(k) = \binom{y+k-1}{k} p^y (1-p)^k, \quad 0 \le p < 1, \quad y \text{ real},$$
$$p = \frac{1}{EX \cdot c_X^2}, \quad y = \frac{EX}{EX \cdot c_X^2 - 1}, \quad EX \cdot c_X^2 > 1. \tag{4.1}$$

Note here that the coefficients of variation of the discrete-time processes are intentionally chosen to be equivalent to the deterministic ($c_A, c_B = 0$), the Erlangian of 4-th order ($c_A, c_B = 0.5$), the Markovian ($c_A, c_B = 1$), and the hyper-exponential ($c_A, c_B = 1.5$) distributions.

The impact of the call arrival process on blocking probability is illustrated in Fig. 4 where customer blocking depends strongly on the coefficient of variation of the interarrival time. Customer blocking probability is depicted in Fig. 5 as a function of the threshold of the overload control strategy for different types of interarrival processes and traffic conditions. Figures 4 and 5 show that, for a proper dimensioning of the threshold value of the overload control strategy, the arrival process characteristics must carefully be taken into account.

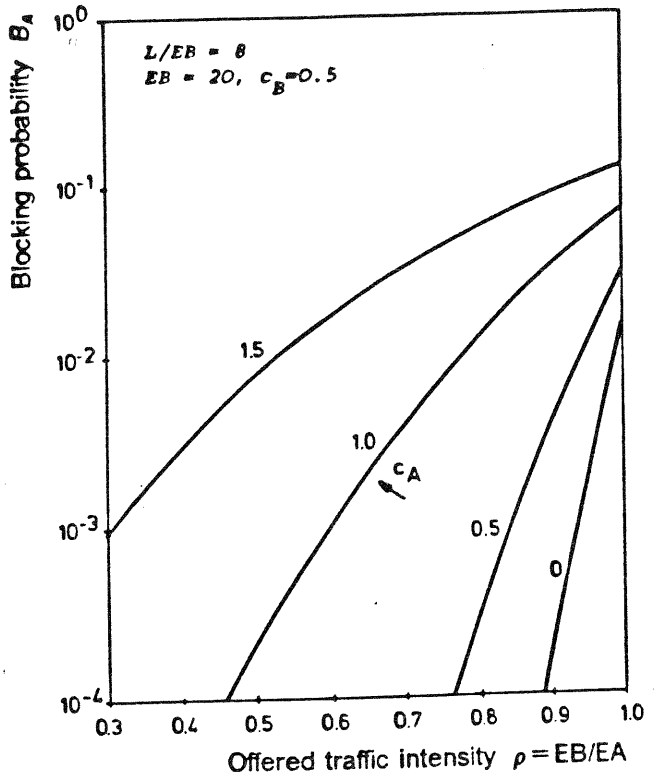

FIGURE 4
Impact of arrival process on blocking probability.

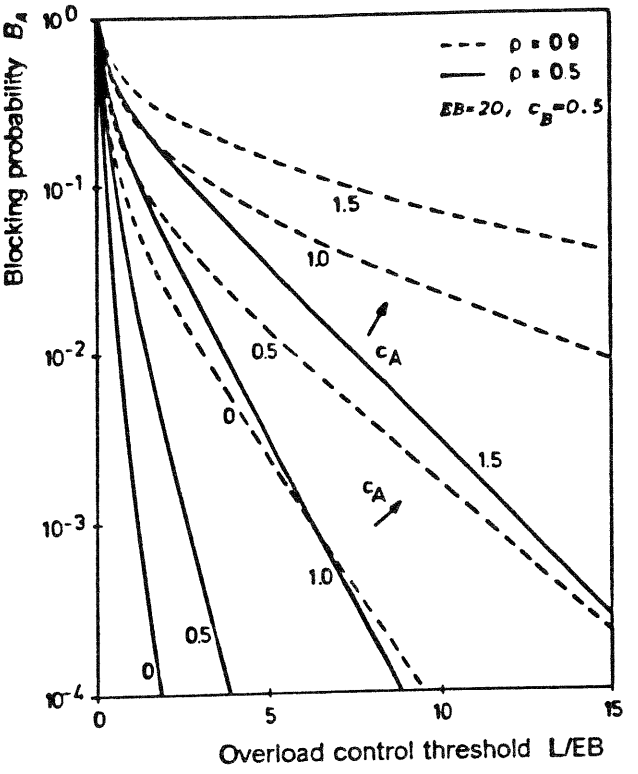

FIGURE 5
Impact of threshold value on blocking probability.

Typical nonstationary behavior of the load-driven overload control strategy is depicted in Fig. 6. Due to the stochastic nature of the arrival process the *customer number* is marked on the x-axis of this diagram instead of *time*. We investigate the system reaction in terms of customer-individual blocking probability to short-term overload patterns. These patterns are parameterized as follows:

$EA_n = 2EB, \quad n = 0,...,14, \quad (\rho = 0.5)$
$EA_n = EB, \quad n = 15,...,39, \quad (\rho = 1.0)$
$EA_n = 2EB, \quad n = 40,..., \quad (\rho = 0.5)$.

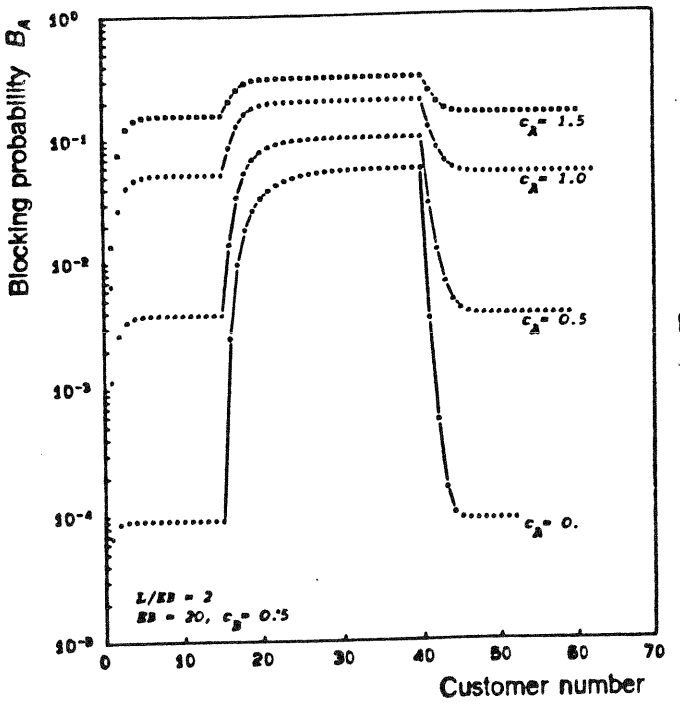

FIGURE 6
Nonstationary behavior of the overload control mechanism.

According to this, the mean duration of the overload pattern is 25.EB. We assume further that the first customer (n = 0) finds an empty system ($B_{A,0} = 0$). For the given parameters, the speed of the dynamic system reaction to the nonstationary overload pattern is relatively independent of the type of the arrival process.

5. CONCLUSIONS

An exact analysis of a general class of overload control strategy applied in communication switching systems is presented. Using the analysis derived in this paper system behavior, under steady state load conditions as well as transient behavior of systems under short-term overloads, can be investigated. The performance modeling is done using a generic queueing system of type G/G/1 with controlled input process. The modeling approach takes two main principles of these overload control strategies into account: i) instead of the number of active calls in the system the actual workload (i.e., the amount of unfinished work) is taken as the overload indicator and ii) the call-acceptance mechanism is directly workload-driven, (i.e., when the workload exceeds a defined threshold, arriving customers are rejected).

To obtain a generally applicable exact analysis algorithm, methods operating in the discrete-time domain are employed in this paper, allowing us to take advantage of efficient convolution and transformation algorithms like the FFT. The discrete-time analysis approach is advantageous as the parameterization of model components can be done directly based on measured data in terms of histograms. An algorithm to calculate system characteristics is derived to estimate the performance of the overload control strategy. The results presented show system behavior under stationary and nonstationary load conditions.

ACKNOWLEDGEMENTS

The author would like to thank Professor P.J. Kühn for his interest in this study and R. Sieglen for helpful discussions and valuable programming efforts during the course of this work.

REFERENCES

[1] Daisenberger, G., Oehlerich, J. and Wegmann, G., "STATOR - Statistical Overload Regulation - and TAIL - Time Account Input Limitation - Two Concepts for Overload Regulation in SPC Systems," Proc. 11th Int. Teletraffic Congress (ITC), Kyoto, 1985, paper 2.1B-4, pp.1-7.

[2] Doshi, B.T. and Heffes, H., "Analysis of Overload Control for a Class of Distributed Switching Machines," Proc. 10th Int. Teletraffic Congress (ITC), Montreal, 1983, paper 5.2.2, pp. 1-6.

[3] Forys, L.J., "Performance Analysis of a New Overload Strategy," Proc. 10th Int. Teletraffic Congress (ITC), Montreal, 1983, paper 5.2.4, pp. 1-8.

[4] Heffes, H., "Analysis of Overload Performance for a Class of M/D/1 Processor Queueing Disciplines," Proc. 11th Int. Teletraffic Congress (ITC), Kyoto, 1985, paper 2.1B-1, pp. 1-7.

[5] Manfield, D.R., Denis, B. and Basu, K., "Overload Control in a Hierarchical Switching System," Proc. 11th Int. Teletraffic Congress (ITC), Kyoto, 1985, paper 5.1B-4, pp. 1-7.

[6] Schoute, F.C., "Adaptive Overload Control for an SPC Exchange," Proc. 10th Int. Teletraffic Congress (ITC), Montreal, 1983, paper 5.2.3, pp.1-6.

[7] Schoute, F.C., "Overload Control in SPC Processors," Philips Telecommun. Rev., 41(1983) 300-310.

[8] Tran-Gia, P., "Subcall-oriented Modeling of Overload Control in SPC Switching Systems," Proc. 10th Int. Teletraffic Congress (ITC), Montreal, 1983, paper 5.2.1, pp. 1-7.

[9] Tran-Gia, P. and Van Hoorn, M.H., "Dependency of Service Time on Waiting Time in Switching Systems - A Queueing Analysis with Aspects of Overload Control," IEEE Trans. Commun., COM-34 (1986) 357-364.

[10] Lindley, D.V., "The Theory of Queues with a Single Server," Proc. Cambridge Philos. Soc., 48 (1952) 277-289.

[11] Ackroyd, M.H., "Computing the Waiting Time Distribution for the G/G/1 Queue by Signal Processing Methods," IEEE Trans. Commun., COM-28 (1980) 52-58.

[12] Henrici, P., "Fast Fourier Methods in Computational Complex Analysis," Siam Review, 21 (1979) 481-527.

[13] Konheim, A.G., "An Elementary Solution of the Queueing System GI/G/1," SIAM J. Comp., 4 (1975) 540-545.

[14] Tran-Gia, P., "Discrete-Time Analysis for the Interdeparture Distribution of GI/G/1 Queues," Proc. Seminar on Teletraffic Analysis and Computer Performance Evaluation, Amsterdam, 1986, pp. 341-357.