PERFORMANCE ANALYSIS OF SEMIDYNAMIC SCHEDULING STRATEGIES IN DISCRETE-TIME DOMAIN

Phuoc Tran-Gia
Erwin Rathgeb

Reprinted from PROCEEDINGS OF THE IEEE INFOCOM '87
San Francisco, California, March 31-April 2, 1987

PERFORMANCE ANALYSIS OF SEMIDYNAMIC SCHEDULING STRATEGIES IN DISCRETE-TIME DOMAIN

Phuoc TRAN-GIA*, Erwin RATHGEB

Institute of Communications Switching and Data Technics, University of Stuttgart, FRG

Abstract. Semidynamic scheduling strategies used in load sharing, routing or job scheduling problems form a class of mechanisms, which lead to delay optimizations in distributed systems. A performance investigation for semidynamic scheduling strategies using discrete-time analysis methods is presented in this paper. The interarrival process as well as the service process are characterized by means of generally distributed random variables. The arising performance model can be decomposed into submodels in an exact manner. The analysis of submodels includes, e.g., the discrete-time investigation of the general class of G/G/1 models with general service and cyclic renewal input processes. In order to compare semidynamic scheduling strategies with random scheduling schemes, a model example is given, for which numerical results are provided to show the influences of a range of system parameters, e.g., the types of input and service processes, the traffic intensity, etc., on the system performance.

1. Introduction

In distributed systems the messaging delays are strongly influenced by the applied load sharing mechanism, routing strategy or scheduling scheme. In a common class of distributed processing systems a decentralized architecture is employed, where a number of heterogeneous processing units of different speeds and service characteristics share service of an incoming job stream. This architecture can be found, e.g., in file server systems, distributed databases, switching processors in stored program controlled (SPC) systems, etc. In such systems, the incoming traffic has to be distributed among the processing units (cf. Fig. 1) according to a predefined scheduling strategy, taking into account the load conditions and the properties of the dedicated servers. The aim of the scheduling strategy design often is an optimization of the delays for customers or jobs. A scheduler can be characterized by means of the following characteristics:

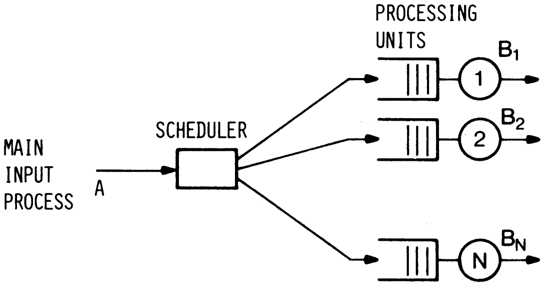

Fig. 1. Load distribution and scheduling.

(i) Load distribution scheme:
Defines the amount of load which is directed to each processing unit. Examples are:
— *Load balancing scheme:* Load is distributed in such a way, so that the heterogeneous servers will have the same utilization factor.
— *Load-driven or Dynamic:* Load is distributed depending on the actual system load, in order to symmetrize the actual load in all processing units (e.g., an incoming job will join the queue with lowest actual load level).

(ii) Scheduling scheme:
Determines the order how to distribute the load according to the applied load distribution scheme. Examples are:
— *Random scheduling:* The amount of load which is dedicated to the i-th processing unit will be formulated as a branching probability p_i, according to which the incoming traffic will be randomly split (Bernoulli branching).
— *Semidynamic:* The load distribution is implemented according to a deterministic scheduling cycle. The number of scheduling positions of a queue is proportional to the amount of traffic distributed to this queue.

The different load sharing strategies require different levels of information about the system state. An overview including an extensive classification of scheduling strategies has been given in Wang and Morris [17]. Buzen and Chen [5] presented an

* The author is now with the IBM Zurich Research Division, Zurich Research Laboratory, 8803 Rüschiikon, Switzerland

algorithm to define the optimal load distribution scheme assuming the random scheduling strategy, Poisson input processes and general service times. In Yum [16] semidynamic scheduling strategies have been presented and investigated in the context of routing problems in computer communication networks, whereby Markovian input and service processes are taken into account. Analysis and performance comparisons concerning the random and semidynamic scheduling schemes have been given in Agrawala and Tripathi [3,4] and Ephremides et al. [7].

The performance analysis of models for scheduling strategies, especially for semidynamic schemes leads to a number of submodels, for which a closed-form solution or a numerical algorithm in continuous-time domain and related transforms are not available. In most of the performance studies mentioned above, Markovian assumptions for arrival or service processes are often considered, where solutions in transform domain can be obtained (cf. [4]).

In this paper, general assumptions are made for arrival and service processes, by means of which effects appearing in real systems like the influence of heterogeneous servers and non-Poissonian arrival processes on the performance of the scheduling strategies can be investigated. The analysis employs methods developed for discrete-time queueing systems [1,8,15].

2. Models of Scheduling Strategies

In this section the queueing model and the according scheduling strategies will be defined.

2.1 Load Distribution and Scheduling Model

As illustrated in Fig. 1 the queueing system consists of a number N of single server queueing stations, which represent the processing units. The offered traffic of each queue results from the load distribution performed by the scheduler. As mentioned above, the main subject of interest is the quantitative load distribution among N heterogeneous servers of different capacities and service-time distributions.

The main arrival process is assumed to be a renewal process with generally distributed discrete interarrival time and the service processes to be general discrete-time processes. The following random variables (r.v.) are used:

A r.v. for the interarrival time of the main arrival process
B_i r.v. for the service time of server i.

The queues are thought of to be infinite and the service discipline is first-in, first-out (FIFO); thus, the waiting time of a job only depends on the amount of unfinished work in the system (queue and server) seen upon arrival. The following notations for functions belonging to a discrete-time random variable X will be used:

$$x(k) = \Pr\{X = k\}, \quad -\infty < k < +\infty \qquad \text{:distribution of } X \tag{2.1a}$$

$$X(k) = \sum_{i=-\infty}^{k} x(i), \quad -\infty < k < +\infty \qquad \text{:distribution function of } X \tag{2.1b}$$

$$x_{ZT}(z) = \sum_{k=-\infty}^{\infty} x(k)\, z^{-k} \qquad \text{:Z-transform of } x(k) \tag{2.1c}$$

EX mean of X

c_X coefficient of variation of X.

In the analysis, where the sum of independent random variables of the same type is used, the r.v. and the according distribution obtained by convolution are denoted as:

$$X^{(j)} = \underbrace{X + X + \dots + X}_{\text{j-times}} \tag{2.2a}$$

$$x^{(j)}(k) = \underbrace{x(k) \star x(k) \star \dots \star x(k)}_{\text{j-times}} . \tag{2.2b}$$

In the following, attention is devoted to the load balancing scheme used as load distribution principle, according to which all servers in the system have the same utilization factor (i.e., the normalized traffic intensity)

$$\rho_i = \frac{EB_i}{EA_i} = \rho, \; i = 1, \dots, N . \tag{2.3}$$

With the service time factors k_i defined by

$$k_i = \frac{EB_i}{EB_1}, \tag{2.4}$$

the mean interarrival time at queue i can be given as follows

$$EA_i = k_i EA_1 . \tag{2.5}$$

Considering the conservation of flows in the system we arrive at

$$\frac{1}{EA} = \sum_{i=1}^{N} \frac{1}{EA_i} . \tag{2.6}$$

Thus, we obtain for the mean interarrival time of the input process offered to queue i

$$\frac{EA_i}{EA} = k_i \sum_{j=1}^{N} \frac{1}{k_j} \, . \qquad (2.7)$$

2.2 Semidynamic vs Random Scheduling

The characterization of input processes at individual queues depends on the applied scheduling scheme, as specified in the following:

(i) Random Scheduling Scheme (RS):
According to this strategy the traffic offered to a queue i results of a Bernoulli branching of the main input process with the routing probability given by

$$p_i = \frac{EA}{EA_i} \, , \qquad (2.8)$$

the input process at each queue is again a renewal process. The calculation of these decomposed processes in discrete-time domain will be dealt with in section 3.1.

(ii) Semidynamic Scheduling Scheme (SD):
The jobs are distributed in a cyclic manner. In general, the input process can be described by means of a cyclic input process (or alternating renewal process [6]). Since, the performance investigation requires an algorithm to analyze general single server queues with cyclic inputs, as will be described in section 3.3.

2.3 Model Example

2.3.1 Model Parameters: For the numerical computations in section 4 a system with three servers (N = 3) is considered. According to the load balancing scheme the service time factors result to

$$k_1 = 1, k_2 = \frac{EB_2}{EB_1} = \frac{3}{2} \, , \; k_3 = \frac{EB_3}{EB_1} = 3 \, . \qquad (2.9)$$

From eqs. (2.7) and (2.9) the routing probabilities p_i are obtained

$$p_1 = \frac{1}{2}, \; p_2 = \frac{1}{3}, \; p_3 = \frac{1}{6} \, . \qquad (2.10)$$

For the semidynamic scheme, in order to fulfill the load distribution scheme given in eq. (2.8) the minimal cycle length is 6. During each cycle the servers 1, 2 and 3 will receive 3, 2 and 1 jobs, respectively. There exist ten alternatives to design such a cycle using different groupings for the appearances of the servers. According to this cycle length, two semidynamic scheduling schemes will be defined and investigated for comparison purposes:

- Semidynamic Scheduling Scheme 1 (SD1) defined by the sequence

  1 1 1 2 2 3 .

  In this scheme the appearances of the servers are grouped together in blocks.

- Semidynamic Scheduling Scheme 2 (SD2) defined by the sequence

  3 1 2 1 2 1 .

  In this scheme the appearances are distributed as regular as possible over the cycle.

2.3.2 Description of Submodels

a) Random Scheduling
As mentioned in section 2.2, the processes resulting from a random decomposition still are renewal processes and therefore the waiting time distributions for the RS scheme can be evaluated using standard methods for discrete-time GI/G/1 systems as described in section 3.2.

b) Semidynamic SD1
In the SD1 scheme the jobs are distributed as depicted in Fig. 2, where the interarrival processes at individual processing units are illustrated.

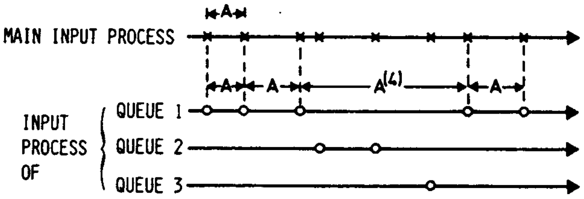

Fig. 2. Input process characteristics for the semidynamic scheduling SD1.

As a result we observe at processing unit 1 a cyclic input with an arrival cycle consisting of three segments. The interarrival interval is of length A for the first and second segment and of length $A^{(4)}$ for the third segment. These three segments yield to an arrival process which is obviously non-renewal and has to be analyzed using the methods described in section 3.3. Similarly, the cyclic input process at processing unit 2 consists of two segments of length A and $A^{(5)}$, respectively.

A special case of a cyclic input can be observed for processing unit 3, where only one segment of length $A^{(6)}$ occurs. This process still is a renewal process and can be obtained from the main arrival process by means of convolutions. Because of this characteristic we call this class of processes "convolved" inputs. Their renewal property makes convolved inputs amenable to the more efficient analysis with standard methods for discrete-time GI/G/1 systems.

c) Semidynamic SD2
The processing units 1 ($A^{(2)}$) and 3 ($A^{(6)}$) in the SD2 scheme fall into the class of convolved input submodels, whereby for processing unit 3 there is no

difference between the SD1 and the SD2 scheme. Processing unit 2 has to be analyzed using the algorithms for G/G/1 queues with cyclic inputs due to its two-segment cyclic arrival process.

2.4 System Characteristics

The performance measures of the whole model are obtained from the characteristics calculated in the submodel analysis. Denote W to be the waiting time for an arbitrary job entering the system and W_i the waiting time of jobs distributed to the i-th processing unit, the overall waiting-time distribution can be determined by a weighted summation

$$w(k) = \sum_{i=1}^{N} p_i \, w_i(k) \; . \tag{2.11}$$

The distribution of the sojourn time F of an arbitrary job of the main arrival process is given by:

$$f(k) = \sum_{i=1}^{N} p_i [w_i(k) \star b_i(k)] \; . \tag{2.12}$$

3. Discrete-Time Analysis of Submodels

In this section, analysis methods for the submodels arising out of the scheduling models described above will be presented.

3.1 Decomposition of Discrete-Time Renewal Processes

According to the random scheduling scheme (RS) the main arrival process is considered to branch randomly to the i-th processing unit with the probability p_i. The input process of queue i is a renewal process described by the r.v. A_i with the distribution given as follows (cf. [11]):

$$a_{i,ZT}(z) = \frac{p_i a_{ZT}(z)}{1-(1-p_i)a_{ZT}(z)} \; . \tag{3.1}$$

3.2 Discrete-Time Analysis of GI/G/1-Queues

For the discrete time as well as for the continuous time GI/G/1 system several approaches to compute the waiting-time distribution w(k) have been proposed [1,9], most of which are based on the Lindley Integral Equation [8,11]. Assuming the distributions for interarrival and service times to be of finite length according to

$$a(k) = \Pr\{A = k\}, k = 0, 1, ..., n_A - 1, n_A < \infty, \tag{3.2a}$$

$$b(k) = \Pr\{B = k\}, k = 0, 1, ..., n_B - 1, n_B < \infty, \tag{3.2b}$$

in the discrete-time domain an equivalent form of this equation is given for stationary conditions by

$$w(k) = \pi(w(k) \star c(k)) \; ,$$

where (3.3)

$$c(k) = a(-k) \star b(k) \; ,$$

and the discrete π-operator is defined by [1,8,14]

$$\pi(x(k)) = \begin{cases} x(k) & k > 0 \\ \sum_{i=-\infty}^{0} x(i) & k = 0 \; . \end{cases} \tag{3.4}$$

The derivation of eq. (3.3) can be found in [14]. This eq. can be solved by iteration in the time domain (probability domain) [1] or directly, without iteration in the frequency domain. The latter method has been found to be more effective for computation, especially when the distributions a(k) and b(k) are relatively long (> 2^{10} elements). To get eq. (3.3) into a suitable form for solutions in the frequency domain we introduce the discrete probability distribution function W(k) and express (3.3) as

$$W(k) = \begin{cases} 0 & k < 0 \\ c(k) \star W(k) & k \geq 0 \; . \end{cases} \tag{3.5}$$

Defining a sequence $W^-(k)$ similar to Kleinrock [9] as

$$W^-(k) = \begin{cases} c(k) \star W(k) & k < 0 \\ 0 & k \geq 0 \end{cases} \tag{3.6}$$

and using the Z-transform to get into the frequency domain we get

$$\frac{W^-_{ZT}(z)}{W_{ZT}(z)} = c_{ZT}(z) - 1 \tag{3.7}$$

or, replacing the probability distribution function by the probability distribution

$$\frac{W^-_{ZT}(z)}{w_{ZT}(z)} = \frac{c_{ZT}(z) - 1}{1 - z^{-1}} = S_{ZT}(z) \; . \tag{3.8}$$

For finite length sequences a(k) and b(k), c(k) is also a finite length sequence. Furthermore, the function $c_{ZT}(z)$ can be shown to have a single zero for $z = 1$. Taking into account these properties $S_{ZT}(z)$ has to be a finite polynomial in 1/z and for that reason has no poles. It can also be shown that the term $W^-_{ZT}(z)$ is a polynomial without poles for a finite length c(k). Applying the theorem of Eneström and Kakeya [1] to $W^-(-k)$ we find, that all zeros of $W^-_{ZT}(z)$ are located outside the unit circle. The function $w_{ZT}(z)$ is the Z-transform of a probability distribution and converges for $z = 1$. From that we can conclude, that $w_{ZT}(z)$ has only poles inside the unit circle and so all zeros of $1/w_{ZT}(z)$ have to be located inside the unit circle as well. Since $S_{ZT}(z)$ and $W_{\bar{Z}T}(z)$ have no poles and the latter function only has zeros outside the unit circle it is obvious, that $1/w_{ZT}(z)$ can have no poles. To obtain $w_{ZT}(z)$ from eq. (3.8) it is thus necessary to find the zeros of $S_{ZT}(z)$ and to separate them with respect to their location to the unit circle.

Two principles have been proposed to accomplish this separation numerically:

- The polynomial factorization algorithm as proposed by Konheim [10]. In this algorithm the zeros of the characteristic function have to be explicitly determined, which may be ineffective for interarrival and service-time distributions with a great number of elements. Furthermore, the results are given in the frequency domain only and further computations are required to get them into the probability domain.
- The complex cepstrum algorithm as presented by Ackroyd [1]. This algorithm takes advantage of the properties of the complex cepstrum [12] and all operations involved, e.g., convolutions and correlations, can be computed using highly effective Fast Fourier Transform algorithms. The result of this algorithm is the waiting-time distribution in the probability domain. An implementation of the Ackroyd algorithm has been used, in combination with a decomposition procedure implementing eq. (3.1), for the computation of the random scheduling scheme as well as for the parts of the semidynamic schemes with convolved inputs.

3.3 Discrete-Time Analysis of G/G/1-Systems with Cyclic Inputs

In the literature solutions of queueing systems with cyclic inputs can be found in continuous-time domain, mainly for systems with Poisson input [4]. In order to investigate the system with more general assumptions, e.g., heterogeneous service process and general cyclic input, an analysis approach in discrete-time domain will be presented in the following. It should be noted here that this class of problems can be dealt with using results and algorithms employing Levinson's method for systems with cyclo-stationary behavior [2]. In this chapter the alternative applying iterative convolutions will be described.

3.3.1 Algorithm in Discrete-Time Domain

We consider a single server with arbitrary distributed service times in discrete-time domain. The input process is cyclic and consists of a number n of interarrival intervals. With respect to the processing unit i these intervals will be denoted by means of the r.v. A_{ij}, $j=1,\ldots,n$. Within a cycle an interval A_{ij} is assumed to be started with the job j which experiences the waiting time W_{ij}. For the first-in, first-out service discipline, W_{ij} is the amount of unfinished work seen from job j upon arrival (cf. Fig. 3). The service time of job j is denoted by the r.v. B_{ij}, $j=1,\ldots,n$. Considering the process development of unfinished work during an arrival cycle, as depicted in Fig. 3, the following equation system can be obtained under stationary conditions for the random variables

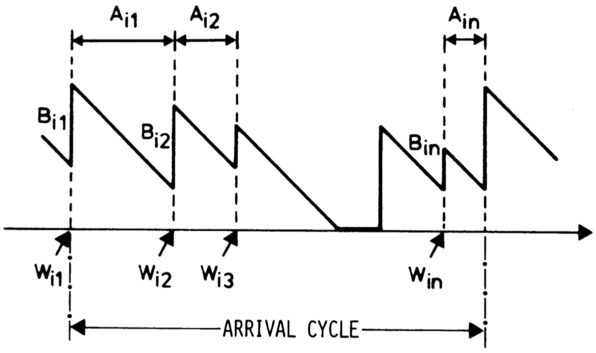

Fig. 3. A sample path of unfinished work in a single server queue with cyclic input.

$$\begin{aligned} W_{i,j+1} &= \max(W_{ij} + B_{ij} - A_{ij}, 0), \quad j = 1, \ldots, n-1, \\ W_{i1} &= \max(W_{in} + B_{in} - A_{in}, 0) \ , \end{aligned} \tag{3.9}$$

and accordingly, for the waiting time distributions

$$\begin{aligned} w_{i,j+1}(k) &= \pi\big(w_{ij}(k) \star b_{ij}(k) \star a_{ij}(-k)\big), \\ & \qquad j = 1, \ldots, n-1, \\ w_{i1}(k) &= \pi\big(w_{in}(k) \star b_{in}(k) \star a_{in}(-k)\big) \ . \end{aligned} \tag{3.10}$$

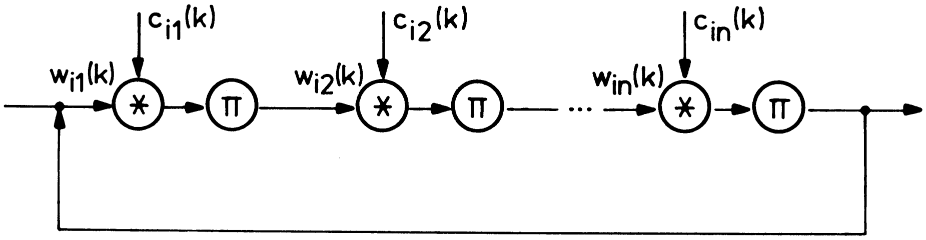

Fig. 4. Flow graph of the algorithm in time domain.

In accordance with eq. (3.10), the waiting time distributions of jobs in the next arrival cycle can successively be calculated from those of the current arrival cycle. Using this fact the equilibrium waiting-time distributions can be determined iteratively, as schematically depicted in Fig. 4 (cf. [15]). This iteration scheme is referred to as the method of iterative convolutions. For large vector sizes of the arrival and service distributions, the discrete convolution operation can effectively be implemented using standard algorithms, e.g., the Fast Fourier Transform (based on the Discrete Fourier Transform) [13].

3.3.2 Waiting-Time Distribution

The waiting-time distributions of jobs within a observed cycle $w_j(k), j = 1, ..., n$, which are obtained by means of the iterative convolution described in the previous subsection, form the basic requirements for the calculation of further system characteristics.

The waiting time distribution of an arbitrary job arriving at the observed processing unit i can be determined as follows:

$$w_i(k) = \frac{1}{n} \sum_{j=1}^{n} w_{ij}(k). \qquad (3.11)$$

4. Numerical Examples

In order to provide a quantitative comparison of the scheduling strategies discussed above, the model example as presented in section 2.3 will be investigated. The three scheduling mechanisms considered will be referred to as

- RS : Random Scheduling

- SD1: Semidynamic Scheduling with the sequence
  1 1 1 2 2 3

- SD2: Semidynamic Scheduling with the sequence
  3 1 2 1 2 1.

In order to investigate the influences of the random processes in a systematic way we will consider the random variables having distributions given by the first two moments. In this context, the arrival process and the service processes will be characterized by means of the negative binomial distribution:

$$x(k) = \binom{y+k-1}{k} p^y (1-p)^k, 0 < p < 1, y \text{ real}. \qquad (4.1)$$

The mean and the coefficient of variation are given by:

$$EX = \frac{y(1-p)}{p}, \quad c_X^2 = \frac{1}{y(1-p)} \qquad (4.2)$$

or

$$p = \frac{1}{EX \cdot c_X^2}, \quad y = \frac{EX}{EX \cdot c_X^2 - 1}, \qquad (4.3)$$

where

$$EX \cdot c_X^2 > 1 . \qquad (4.4)$$

The time measures will be given in a normalized form with the discrete time unit $\Delta t = 1$. For the mean waiting time EW_i of jobs at individual queues, Fig. 5 shows a comparison of the worst performing scheme (Random Scheduling) and the scheme with the best performance (SD2). The server coefficients of variation have been set to $c_B = 0.5$, which are equivalent to continuous-time Erlangian distribution of 4-th order ($EB_1 = 30, \rho = 0.5$). The mean-waiting times have been normalized to the mean service times of the corresponding servers. It is obvious from Fig. 5 that for any queue the normalized mean waiting time is higher for the RS scheme than for the SD2 scheme over the whole range of c_A.

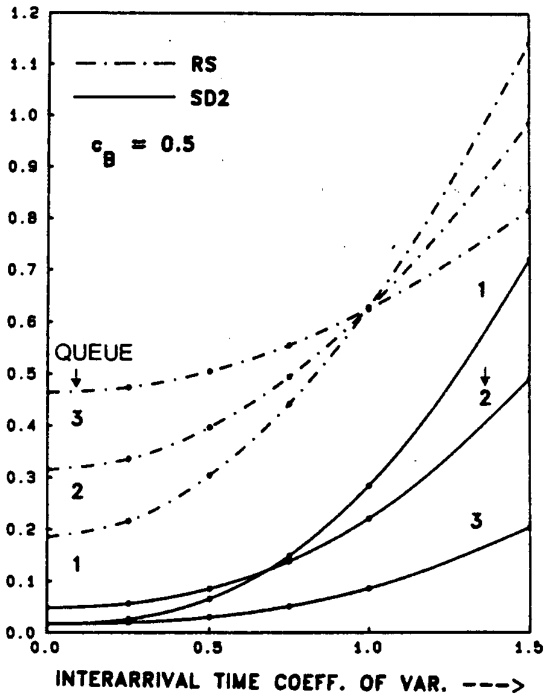

Fig. 5. Influence of scheduling strategies on queue individual waiting times.

The coefficient of variation of the interarrival process at queue i for the RS scheme is known as [11]:

$$c_{Ai}^2 = p_i \cdot c_A^2 + (1 - p_i) . \qquad (4.5)$$

According to this formula $c_{A3,RS}$ is the highest of the $c_{Ai,RS}$ for $c_A < 1$, for $c_A = 1$ there is a cross-over point with $c_{A1,RS} = c_{A2,RS} = c_{A3,RS} = 1$ and for $c_A > 1$, $c_{A3,RS}$ is lower than $c_{A1,RS}$ and $c_{A2,RS}$.

Since a higher interarrival time coefficient of variation implying a higher normalized mean waiting time, the same behavior can be observed for these mean-waiting times.

As far as the SD2 scheme is concerned, queue 1 and queue 3 have got convolved inputs. In this context it shall be mentioned here, that mean value and coefficient of variation of a convolved input process are defined by

$$E[A^{(j)}] = j \cdot E A \qquad (4.6)$$

$$c_{A^{(j)}} = \frac{c_A}{j} . \qquad (4.7)$$

According to eq. (4.6) the normalized mean-waiting time of queue 1 $(A_1 = A^{(2)})$ has to be higher compared to the one of queue 3 $(A_3 = A^{(6)})$ over the full range except for $c_A = 0$, where $c_{A1,SD2} = c_{A3,SD2} = 0$.

For the cyclic input queue 2, due to the alternating input process, $c_{A2,SD2}$ is greater than 0 even for $c_A = 0$, which results in a higher normalized waiting time compared to the queues 1 and 3 at this point.

Figure 6 shows the waiting-time coefficients of variation of the queues for the same parameters and Fig. 7 shows the complementary waiting-time probability distribution functions for the case where $c_A = c_B = 0.5$. Figures 8 to 10 show a comparison of the mean overall waiting times for the three scheduling schemes. The mean-waiting times have

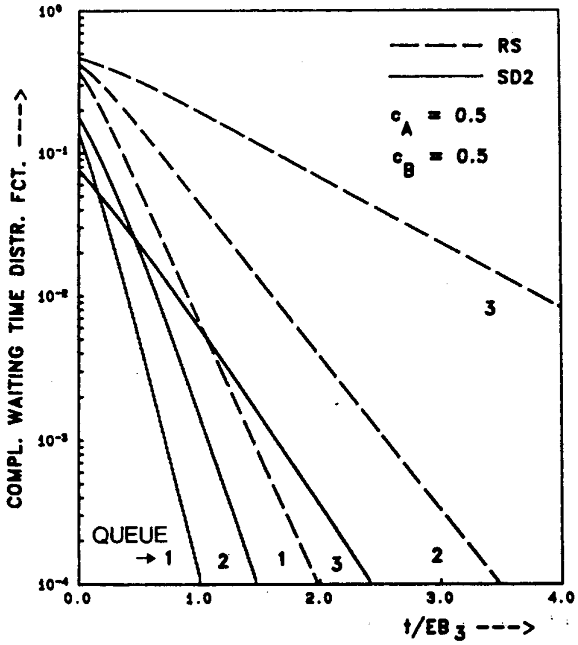

Fig. 7. Queue individual complementary waiting time distribution functions.

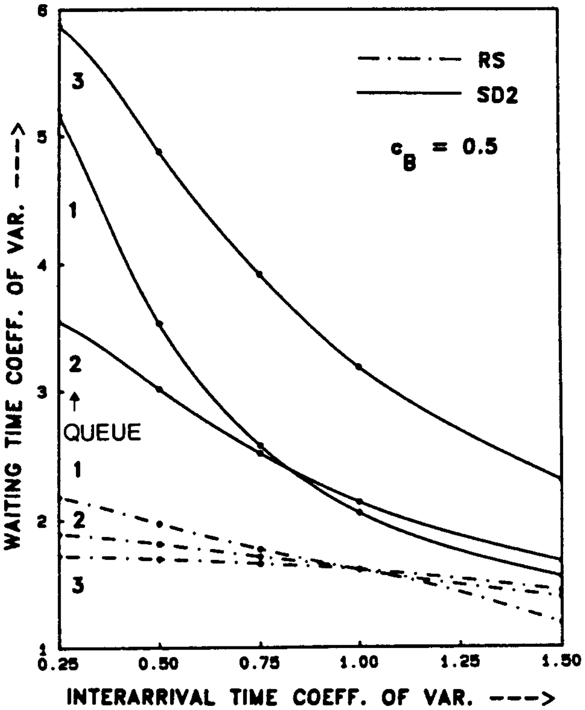

Fig. 6. Influence of scheduling strategies on the waiting time coefficients of variation.

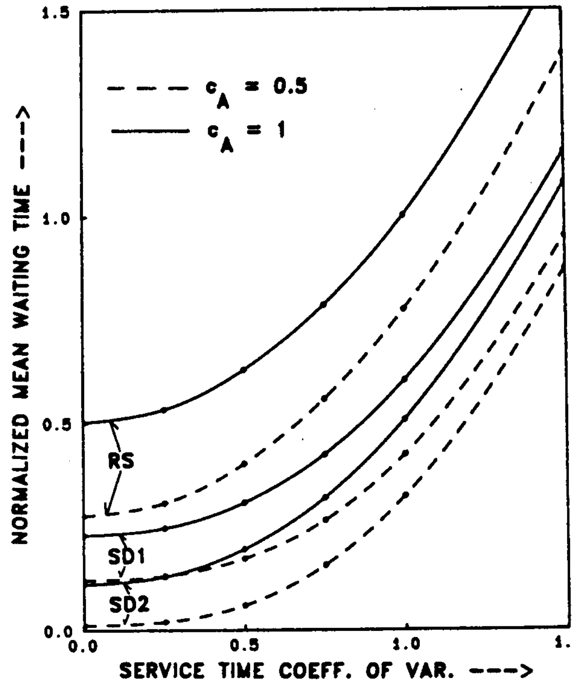

Fig. 8. Influence of scheduling strategies and types of servers on waiting times.

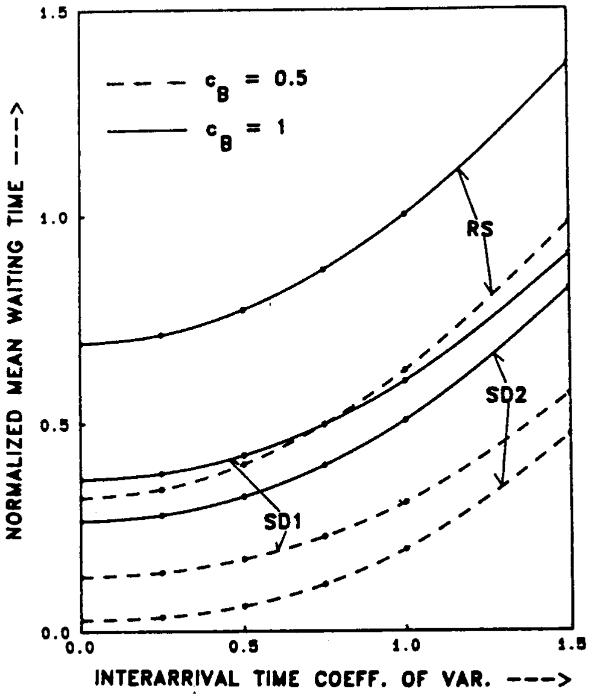

Fig. 9. Influence of scheduling strategies and types of arrival processes on waiting times.

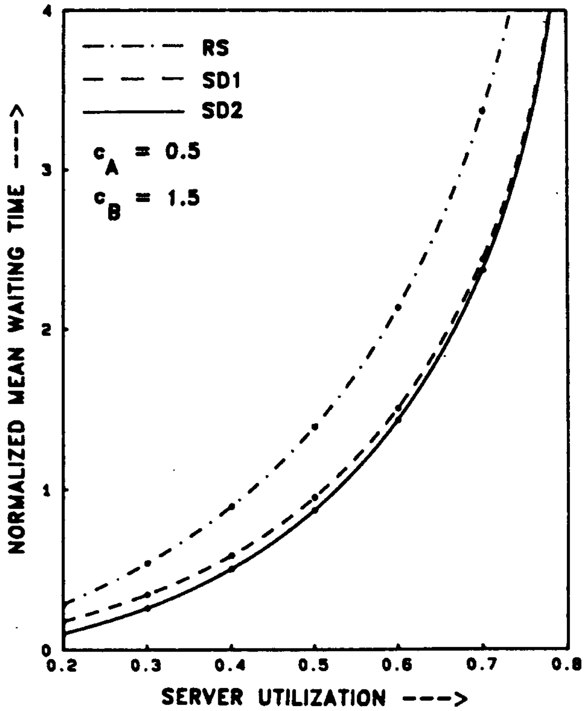

Fig. 10. Influence of scheduling strategies and server utilizations on waiting times.

been normalized to the overall mean-service time EB = 45.

These figures demonstrate that the SD2 scheme performs better than the SD1 and the RS scheme, independent of c_A (Fig. 8), c_B (Fig. 9) and the traffic intensity (Fig. 10).

5. Conclusion and Outlook

In this paper, a performance investigation for scheduling strategies using discrete-time analysis methods is presented, whereby attention is devoted to the class of semidynamic scheduling strategies which are used in load sharing, routing and job scheduling problems in distributed systems. General assumptions are made for arrival and service processes, by means of which effects appearing in real systems like the influence of heterogeneous servers and non-Poissonian arrival processes on the performance of the scheduling strategies are investigated.

The model is decomposed into submodels in an exact manner. The analysis of the the arising submodels includes a new class of models with non-renewal processes, i.e., G/G/1 models with general service and cyclic input processes.

Different variations of semidynamic scheduling strategies are compared with the random scheduling scheme using a model example, for which numerical results are provided to show the influences of system parameters on the overall system performance.

Acknowledgement

The authors would like to express their thanks to Dieter Schmidt and Bozo Cesar for the valuable programming efforts.

References

[1] M. H. Ackroyd, "Computing the waiting time distribution for the G/G/1 queue by signal processing methods," *IEEE Trans. Commun.*, vol. COM-28, pp. 52-58, 1980.

[2] M. H. Ackroyd, "Stationary and cyclostationary finite buffer behavior computation via Levinson's method," *AT&T Bell Lab. Techn. J.*, vol. 63, pp. 2159-2170, 1984.

[3] A. K. Agrawala and S. K. Tripathi, "On the optimality of semidynamic routing scheme," *Information Proc. Lett.*, vol. 13, pp. 20-22, 1981.

[4] A. K. Agrawala and S. K. Tripathi, "On an exponential server with general cyclic arrivals," *Acta Informatica*, vol. 18, pp. 319-334, 1982.

[5] J. P. Buzen and P. S. Chen, "Optimal Load Balancing in Memory Hierarchies," Proc. Information Processing (IFIP). Amsterdam: North Holland, 1974.

[6] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. Chapman and Hall, 1965.

[7] A. Ephremides, P. Varaiya and J. Walrand, "A simple dynamic routing problem," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 690-693, 1980.

[8] P. Henrici, "Fast fourier methods in computational complex analysis," *SIAM Rev.*, vol. 21, pp. 481-527, 1979.

[9] L. Kleinrock, *Queueing Systems*, Vol. I: Theory, Vol. II: Computer Applications. New York: Wiley, 1975.

[10] A. G. Konheim, "An elementary solution of the queueing system GI/G/1," *SIAM J. Comp.*, vol. 4, pp. 540-545, 1975.

[11] P. J. Kuehn, "Approximate analysis of general queueing networks by decomposition," *IEEE Trans. Commun.*, vol. COM-27, pp. 113-126, 1979.

[12] D. V. Lindley, "The Theory of Queues with a Single Server," Proc. of the Cambridge Philosophical Society, vol. 48, pp. 277-289, 1952.

[13] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[14] W. L. Smith, "On the Distribution of Queueing Times," Proc. of the Cambridge Philosophical Society, vol. 49, pp. 449-461, 1953.

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

[16] T.-S. P. Yum, "The design and analysis of a semidynamic deterministic routing rule," *IEEE Trans. Commun.*, vol. COM-29, pp. 498-504, 1981.

[17] Y. T. Wang and R. J. T. Morris, "Load sharing in distributed systems," *IEEE Trans. Comput.*, vol. C-34, pp. 204-217, 1985.