University of Würzburg
Institute of Computer Science
Research Report Series

Analysis of the Departure Process of a Batch Service Queueing System

Mathias Dümmler

Report No. 210 September 1998

Universität Würzburg
Lehrstuhl für Informatik III
Am Hubland, 97074 Würzburg, Germany
duemmler@informatik.uni-wuerzburg.de

Analysis of the Departure Process of a Batch Service Queueing System

Mathias Dümmler
Universität Würzburg, Lehrstuhl für Informatik III, Am Hubland, 97074 Würzburg, Germany,
duemmler@informatik.uni-wuerzburg.de

Abstract

We investigate a queueing system with *batch service* or *bulk service*. Batch servers are common in many areas of manufacturing, for example in the manufacturing of semiconductors. In particular, we study a batch server operated under the *minimum batch size* rule. Under this rule, a service will be initiated as soon as the number of lots waiting for service reaches a certain threshold. This threshold can be arbitrarily chosen.

Using discrete–time analysis, we derive the interdeparture time distribution for batches and for single lots. It is shown that the interdeparture time distribution of a batch service system has characteristic properties that can not be approximated with sufficient accuracy by other commonly used distributions. We further study the distribution of the number of lots in the departing batches and the correlation between interdeparture times and batch sizes.

1 Introduction

In this paper, we investigate in the interdeparture time distribution of a *batch service* or *bulk service* process. A batch server is able to serve a number of lots simultaneously. Servers of this type are common in many areas of manufacturing, for example in semiconductor manufacturing.

The departure process of a batch server is of particular interest in the production of semiconductors. In this area of manufacturing the bottleneck workstation, i. e., the workstation that constrains the production rate of the facility, is commonly preceded by a batch service workstation. To fully exploit the capacity of the bottleneck workstation it is necessary to know the arrival process at the workstation, i. e., the departure process of the preceding batch server.

Furthermore, the interdeparture processes of batch service systems are needed for modeling manufacturing systems as networks of queues. It is necessary to derive the properties of the departure process of a batch server that serves as the input process for the downstream workstation.

Analytical approaches for single batch servers are well–known. Neuts [5] presented an analytical study of batch service systems that operate according to a *minimum batch size rule*. Chaudhry and Templeton [2] provide an extensive discussion of bulk service systems and their analysis. In [3] and [8], batch service systems in push and pull manufacturing environments are analyzed using embedded Markov chain techniques.

We use *discrete–time analysis* (DTA) to investigate in the bulk service process. DTA, as presented in [1] and [6], is a technique for the numerical analysis of single stage queueing systems. DTA is mainly applied to the analysis of high speed communication systems such as ATM (Asynchronous Transfer Mode) networks. In these networks, the transmitted data packets are of constant size and the medium of transmission can be modeled as being partitioned into units of discrete length. This suggests analyzing such systems using DTA. See, for example, [7] for an application of DTA to the analysis of ATM systems.

DTA can also be employed for the analysis of queueing systems if some or all quantities encountered are of continuous type, as is the case in manufacturing systems. For a sample application of DTA in the manufacturing context, see [9].

This paper is structured as follows. In the following paragraph, we give a short introduction to discrete–time analysis. The batch server model under investigation is presented in Section 2. Sections 3 and 4 are dedicated to the derivation of the state distribution of the batch server at departure instants and the computation of the interdeparture time distribution, respectively. Numerical results of our approach are discussed in Section 5 and the paper is concluded in Section 6.

Notation in discrete–time analysis

Random variables representing periods of time considered in this paper are of discrete–time nature, i. e., samples are multiples of Δt, the *unit length*. In other words, the time axis is divided into intervals of a fixed length Δt. Given a discrete–time random variable X, we denote its *distribution* (probability mass function) by

$$x(k) \quad = \quad \Pr\{X \ = \ k \cdot \Delta t\}\,, \quad k = \ldots, -2, -1, 0, 1, 2, \ldots\,. \tag{1}$$

The *mean* of a random variable X is defined as

$$\mathrm{E}[X] \quad = \quad \sum_{i=-\infty}^{\infty} i \cdot x(i)\,, \tag{2}$$

and its *coefficient of variation* as

$$c_X \quad = \quad \frac{\sqrt{\mathrm{Var}[X]}}{\mathrm{E}[X]}\,, \tag{3}$$

where

$$\mathrm{Var}[X] \quad = \quad \sum_{i=-\infty}^{\infty} (i - \mathrm{E}[X])^2 \cdot x(i) \tag{4}$$

is the *variance* of the random variable X.

Given two independent random variables X_1 and X_2, their sum $X = X_1 + X_2$ is distributed according to the convolution of their individual distributions:

$$x(k) \quad = \quad \sum_{l=-\infty}^{\infty} x_1(l) \cdot x_2(k-l) \quad = \quad x_1(k) * x_2(k)\,. \tag{5}$$

The symbol '$*$' denotes the discrete convolution operation.

For the ease of notation, we introduce two operators on discrete distributions. The operator $\pi_m(.)$ "sweeps up" all elements $x(k)$ of a distribution with $k < m$ and adds their sum to the element $x(m)$. It is defined as

$$\pi_m(x(k)) \quad = \quad \begin{cases} 0 & , \quad \text{if} \quad k < m \\ \sum\limits_{i=-\infty}^{m} x(i) & , \quad \text{if} \quad k = m \\ x(k) & , \quad \text{if} \quad k > m\,. \end{cases} \tag{6}$$

The operator $\Delta_m(.)$ is defined as follows

$$\Delta_m(x(k)) \quad = \quad x(k+m)\,. \tag{7}$$

It is used to shift the elements of a distribution down by m units.

2 Model

We will consider the batch server model depicted in Figure 1.

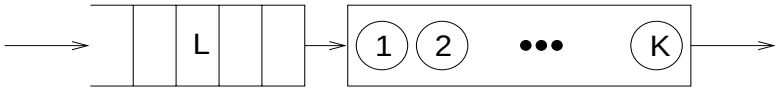

Figure 1: Batch server model

The model consists of a queue with infinite capacity and a batch server. The server has a maximum batch capacity of K lots. When a service period ends and there are less than L lots waiting in queue, where $1 \leq L \leq K$ can be arbitrarily chosen, the server remains idle until L lots

are accumulated in queue. The server is said to operate under a *minimum batch size* strategy since a minimum number of L lots per batch is guaranteed under this strategy.

If a batch of size S_n has formed, where $L \leq S_n \leq K$, this batch will proceed to the server and will be served. The service time for the nth batch is of length B_n, where the B_n are random variables following an arbitrary distribution with mean $\mathrm{E}[B]$. Service times are independent of the number of lots in the batch.

Lots finding the server busy upon their arrival and lots that are waiting for a batch to form are queued in FIFO order. The interarrival times of lots are geometrically distributed. To avoid batch arrivals, i. e., interarrival times of zero, we shift the geometric distribution by one unit to the right. Hence, the random variable A_n, denoting the interarrival time between the $(n-1)$th and the nth lot, is distributed as follows

$$a_n(k) \quad = \quad (1-p)^{k-1}\, p\,, \quad k = 1, 2, \ldots\,, \tag{8}$$

where p is the probability of an arrival in one discretized time unit. The mean interarrival time is $\mathrm{E}[A] = 1/p$.

We define the server load ρ to be the normalized offered traffic

$$\rho \quad = \quad \frac{\mathrm{E}[B]}{K \cdot \mathrm{E}[A]}\,. \tag{9}$$

3 State distribution at departure instants

We now derive the steady state distribution of the number of lots in queue at departure instants using the embedded Markov chain technique. For an introduction to this technique, see [4].

Let X_{n-1} be the number of lots in queue immediately after the $(n-1)$th batch has left the server, not counting the lots in server. If $X_{n-1} \leq K$ the queue will be empty after the next service starts, because all lots waiting in queue form a single batch. If there are more than K lots waiting the number of lots in queue will be reduced by K immediately after beginning of the nth service period. We denote the number of customers in queue immediately after the nth service starts by Y_n. Hence we have

$$Y_n \quad = \quad \begin{cases} X_{n-1} - K & , \quad \text{if} \quad X_{n-1} > K \\ 0 & , \quad \text{if} \quad X_{n-1} \leq K\,. \end{cases} \tag{10}$$

As the respective distribution $y_n(k)$ of Y_n, we obtain

$$y_n(k) \quad = \quad \pi_0(\Delta_K(x_{n-1}(k)))\,. \tag{11}$$

Let Γ_n be the number of arrivals during the nth service. Since we assume geometrically distributed interarrival times, the distribution of the number of lots arriving during the nth service period can be derived as

$$\gamma_n(k) \quad = \quad \sum_{m=k}^{\infty} \binom{m}{k} p^k (1-p)^{m-k} \, b_n(m) \,, \quad k = 0, 1, \ldots \,, \tag{12}$$

where $b_n(m)$ is the distribution of the length of the nth service period.

Taking into account the number of arrivals during the nth service, the number X_n of lots in the queue observed immediately after the nth service period is

$$X_n \quad = \quad Y_n + \Gamma_n \,, \tag{13}$$

and the respective distribution

$$x_n(k) \quad = \quad y_n(k) * \gamma_n(k) \,. \tag{14}$$

Using eqns. (11) and (14), we obtain $x_n(k)$ from $x_{n-1}(k)$ as

$$x_n(k) \quad = \quad \pi_0(\Delta_K(x_{n-1}(k))) * \gamma_n(k) \,. \tag{15}$$

We now iterate eqn. (15) starting with an initial distribution $x_0(k)$, for example

$$x_0(k) \quad = \quad \begin{cases} 1 & , \quad \text{if} \quad k = 0 \\ 0 & , \quad \text{if} \quad k \neq 0 \,, \end{cases} \tag{16}$$

which corresponds to an initially empty system. If the system is stable, i. e., $\rho < 1$, the series of distributions $x_n(k)$ converges to the steady state distribution $x(k)$ of lots in queue observed immediately after a service period ends:

$$x(k) = \lim_{n \to \infty} x_n(k) \,. \tag{17}$$

4 Interdeparture time distribution

Let D_b be the random variable denoting the time between two consecutive batch departures. In order to compute D_b, we have to distinguish two cases: If there are at least L lots in queue at a departure instant, the time to the next departure only consists of the service time B of the batch. If there are not enough lots to form a batch, i. e., $X < L$, the interdeparture time consists of the interarrival times of the $(L-X)$ lots that have to arrive to form a batch and of the service time of this batch. Hence, we obtain the distribution of the batch interdeparture time as

$$d_b(k) \quad = \quad b(k) \cdot \sum_{i=L}^{\infty} x(i) \; + \; \sum_{i=0}^{L-1} (a^{*(L-i)}(k) * b(k)) \cdot x(i)\,, \tag{18}$$

where $a^{*i}(k)$ denotes the i–fold convolution of $a(k)$ with itself.

From $x(k)$, we can also derive the number S of lots in a batch. Its distribution is

$$s(k) \quad = \quad \begin{cases} \sum\limits_{i=0}^{L} x(i) & , \quad \text{if} \quad k = L \\ x(k) & , \quad \text{if} \quad L < k < K \\ \sum\limits_{i=K}^{\infty} x(i) & , \quad \text{if} \quad k = K \\ 0 & , \quad \text{else}\,. \end{cases} \tag{19}$$

We now compute the coefficient of correlation r between the interdeparture time D_b of a batch and the number S of lots in the batch. This coefficient is a measure of how strongly these two values depend on each other. It is defined as

$$r \quad = \quad \frac{\mathrm{E}[D_b \cdot S] - \mathrm{E}[D_b]\,\mathrm{E}[S]}{\sqrt{\mathrm{Var}[D_b]}\,\sqrt{\mathrm{Var}[S]}}\,. \tag{20}$$

The expectation of the product $D_b \cdot S$ is

$$\mathrm{E}[D_b \cdot S] \quad = \quad \sum_{i=0}^{L-1} ((L-i) \cdot \mathrm{E}[A] + \mathrm{E}[B]) \cdot L \cdot x(i) \; + \sum_{i=L}^{K} \mathrm{E}[B] \cdot i \cdot s(i)\,. \tag{21}$$

Note that eqn. (20) is valid for both discrete–time and continuous–time models.

Let the random variable D_l denote the time between two consecutive lot departures. This random variable has the value 0 if the two lots are in the same batch, else it equals the interdeparture time D_b of two consecutive batches. This yields the distribution

$$d_l(k) \quad = \quad \delta(k) \cdot \frac{\mathrm{E}[S]-1}{\mathrm{E}[S]} + d_b(k) \cdot \frac{1}{\mathrm{E}[S]}\,, \tag{22}$$

where

$$\delta(k) \quad = \quad \begin{cases} 1 & , \quad \text{if} \quad k=0 \\ 0 & , \quad \text{else}\,. \end{cases} \tag{23}$$

5 Numerical results

First, we consider the mean interdeparture time of batches and the mean number of lots in a batch for the following system: Maximum batch size is $K = 4$, service time is deterministic with $E_B = 30$. The server load ρ is varied from 0.1 to 0.9. Results are given for starting thresholds $L = 1, \ldots, 4$.

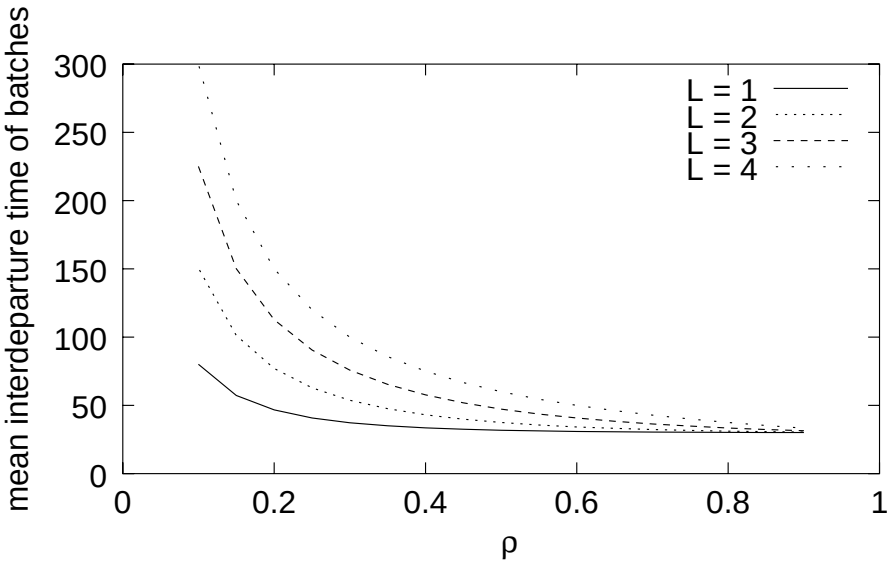

Figure 2: Mean interdeparture time of batches

In Figure 2, we see that the mean interdeparture time quickly approaches E_B as the load ρ is increased. However, the interdeparture times are larger for high starting thresholds, since it takes longer for a batch to form if L is large.

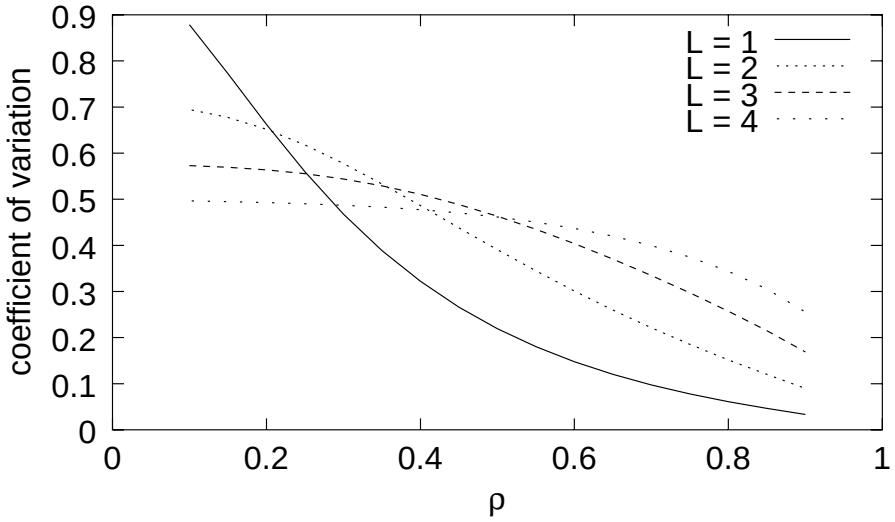

Figure 3: Coefficient of variation of batch interdeparture time

The coefficient of variation of the interdeparture times, displayed in Figure 3, shows significant differences for different server loads and different values of L. For $L = 1$, the coefficient of variation ranges from values close to 0.9 for lower server loads to values close to zero for high server loads. This can be explained as follows: For $\rho = 0.1$, the system is empty for most of the time. Since $L = 1$, the probability that a single lot per batch will be served is high. Hence, the variance of the interdeparture times mainly depends on the variance of the interarrival times of the lots at the server, which is 0.99 for this case. On the other hand, for $\rho = 0.9$, the probability that there are enough lots waiting in queue to form a full batch is high. Hence, the variance of the interdeparture time mainly depends on the variance of the service time. Since we assume deterministic service times in this case, the coefficient of variation of the interdeparture times is close to zero for high server loads.

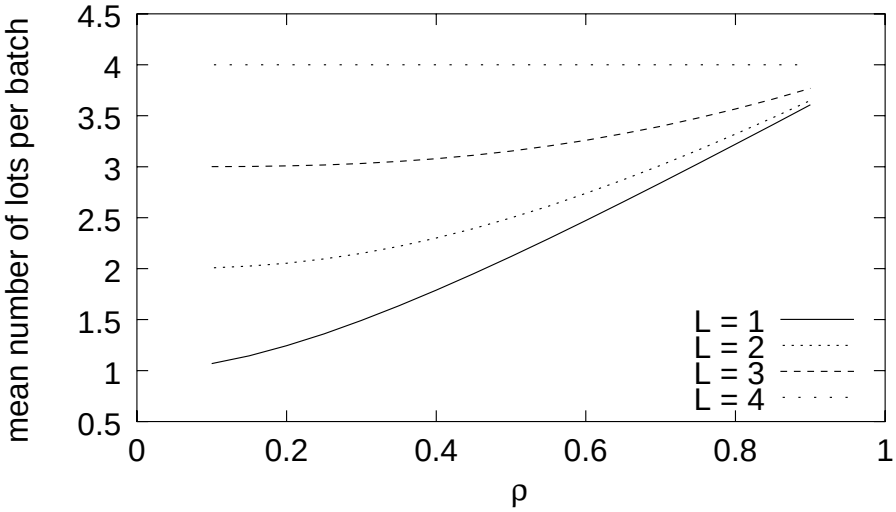

Figure 4: Mean number of lots per batch

Similarly, it can be seen in Figure 4 that the mean batch size approaches K if server load is increased. Note, that the utilization of the server depends strongly on the choice of the starting threshold L for low server loads.

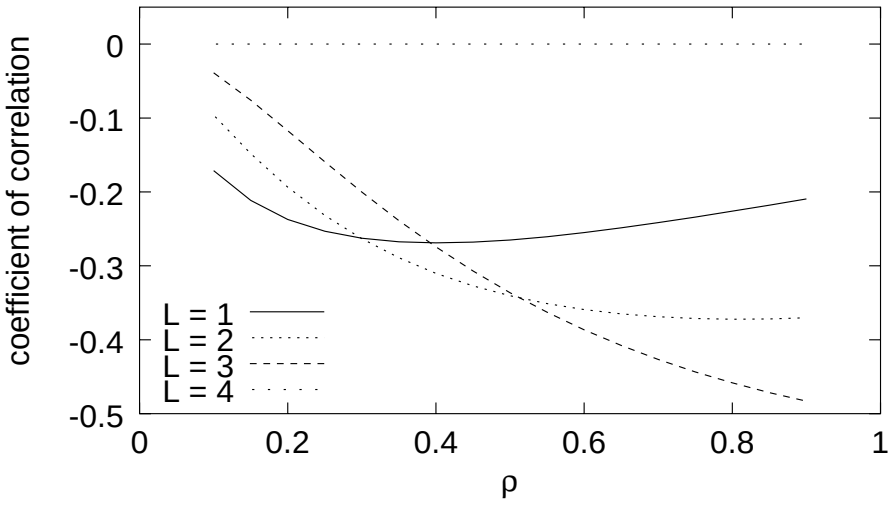

Figure 5: Coefficient of correlation of batch size and interdeparture time

For the same parameters, the coefficients of correlation between batch interdeparture time and batch size are displayed in Figure 5. Note, that the coefficient of correlation r is zero for $L = 4$, since the batch size is constant in this case. For $L = 1, \ldots, 3$, the coefficient of correlation is negative. This means that long interdeparture times correspond to small batch sizes. This is due to the fact that long waiting times occur if there are less than L lots in queue. Hence, the batch that will be served next has the minimum batch size L.

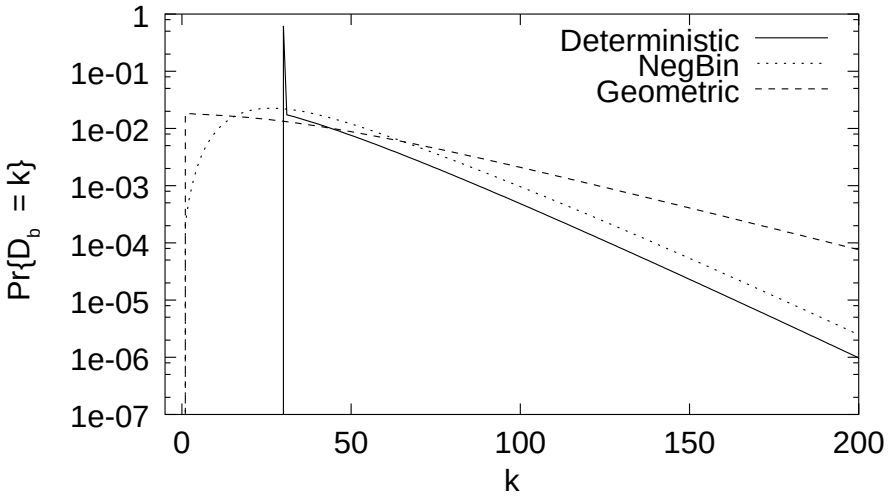

Figure 6: Batch interdeparture time distribution

The distribution of the interdeparture time of batches is displayed in Figure 6 for a system with batch capacity $K = 4$, starting threshold $L = 2$ and server load $\rho = 0.5$. Mean service time

is E[B] = 30. Three different service time distributions are considered, namely deterministic, geometric and negative binomial distribution with coefficient of variation $c_B = 0.5$.

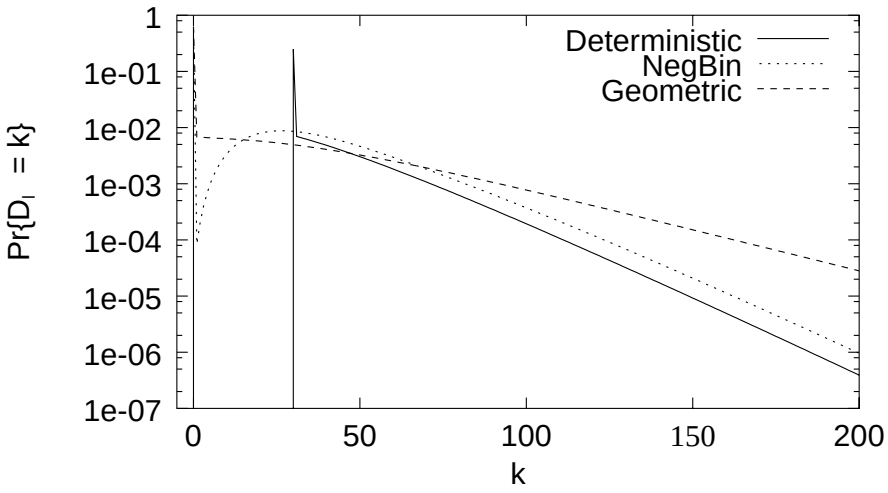

Figure 7: Lot interdeparture time distribution

The interdeparture time distribution has a clear peak at E[B] for the deterministic service time, whereas for the geometric and the negative binomial service time distribution, the interdeparture times are spread over a wider range. Note, that the distribution is truncated in this figure at $k = 200$.

In Figure 7, the interdeparture time of single lots is given for the same system. We see a clear peak at $k = 0$ for the three different service time distributions. This peak corresponds to the $S - 1$ lots in a batch of size S that have interdeparture time zero.

The proposed method for the analysis of the departure process is computationally efficient. The computation of a given parameter set takes about one second on a PC using MATLAB. Simulating the same model using MODSIM requires about thirty seconds.

6 Conclusion

In this paper, we investigated in the interdeparture time distribution of a batch service process. Using discrete–time analysis, we derived the interdeparture time distribution of batches and of individual lots. We also computed the distribution of the number of lots in a batch and the coefficient of correlation of batch interdeparture time and batch size.

The proposed approach is computationally efficient. Furthermore, this approach allows to model more complicated systems. It will be the object of further research to observe how other batching strategies, like bounded idle time or lookahead strategies, affect the departure process of a batch server.

We observed that the departure process of the batch server under the minimum batch size strategy has properties that can not be captured accurately by modelling the process using standard distributions like the exponential distribution. The variance of the interdeparture times and the correlation between batch size and interdeparture time has to be taken into account. Therefore, it is necessary to find processes that allow to model the departure process of batch servers accurately.

References

[1] Martin H. Ackroyd. Computing the waiting time distribution for the G/G/1 queue by signal processing methods. *IEEE Transactions on Communication*, COM-28(1):52–58, January 1980.

[2] M. L. Chaudhry and J. G. C Templeton. *A First Course in Bulk Queues.* Wiley, New York, 1983.

[3] Hermann Gold. *Performance Modelling of Batch Service Systems with Push and Pull manufacturing Management Policies.* PhD thesis, Universität Würzburg, Fakultät für Mathematik und Informatik, 1992.

[4] Leonard Kleinrock. *Queueing Systems, Vol. 1: Theory.* Wiley, New York, 1975.

[5] Marcel F. Neuts. A general class of bulk queues with poisson input. *Ann. Math. Stat.*, 38:759–770, 1967.

[6] Phuoc Tran-Gia. Discrete-time analysis for the interdeparture distribution of GI/G/1 queues. In Onno J. Boxma, Jacob W. Cohen, and Henk C. Tijms, editors, *Teletraffic Analysis and Computer Performance Evaluation*, pages 341–357. Elsevier (North-Holland), 1986.

[7] Phuoc Tran-Gia. Discrete–time analysis technique and application to usage parameter control modelling in ATM systems. In *Proceedings of the 8th Australian Teletraffic Research Seminar*, Melbourne, Australia, December 1993.

[8] Phuoc Tran-Gia, Herman Gold, and Heidrun Grob. A batch service system operating in a pull production line. *Archiv für Elektronik und Übertragungstechnik*, 47(5/6):379–385, 1993.

[9] Phuoc Tran-Gia and Alexander Schömig. Discrete–time analysis of batch servers with bounded idle time. In *Modelling and Simulation 1996. (Proceedings of the European Simulation Multiconference)*, pages 999–1003, Budapest, June 1996.