Modeling Independent Production Buffers in Discrete Time Queueing Networks

Kai Furmans and Andrea Zillus *
University of Karlsruhe, IFK

Manufacturing systems with manual or automatic material handling usually have only limited buffers. In order to configure the buffers the system is modeled by queueing network models in discrete time domain. The interarrival processes as well as the service processes are approximated by general renewal processes. The characterization of the manufacturing system by a queueing network provides means for analyzing the system's parameters, such as work in process, sojourn time and the resulting space requirements, e.g. for buffers.
In this paper, the impacts of buffers and their allocations are discussed and possible buffer configurations are presented. Then, after modeling the system by queuing network techniques, its performance is evaluated. Analytical methods are developed for the computation of the system's steady state probabilities, which lead to the buffer configuration.

Key words: buffer configuration, queueing networks, discrete time, production system, performance measure

1 Problem Description

During the planning process of manufacturing systems and their material handling systems, queueing network models could provide data about transport patterns, expected work in process, sojourn times and average buffer requirements (see [1], [2], [3]). In these models, the manufacturing lots are modeled as jobs which are processed on one or more groups of machines, which are represented by queueing systems, where the number of machines equals the number of servers of the queueing system.

The jobs that are waiting to be processed on a machine are kept in a buffer which is usually adjacent to the respective machine. In queueing models, the capacity of these buffers is very often considered to be infinite. This might be an appropriate modeling assumption, when very small parts are manufactured, as is the case of semiconductor manufacturing. In most manufacturing systems however, the space allocated to buffers is limited, and therefore should be modeled as being finite. When designing a layout for a manufacturing system, the amount of space which is allocated to buffers can have a significant impact on the total space requirement. Therefore it is desirable to evaluate the effects of buffer allocation decisions on the production output and the material handling requirements in an early planning stage. For queueing networks with finite buffers, so called blocking networks, several models with different blocking strategies have been developed.

*The authors can be reached by e-mail over kai.furmans@mach.uni-karlsruhe.de or andrea.zillus@mach.uni-karlsruhe.de

Akyildiz [4] differentiates between three types of blocking, namely transfer blocking, communication blocking and rejection blocking. All these blocking strategies could lead to a reduction in total system throughput because a machine will stop working when the destination buffer of the current or next job is already completely occupied.

This could be avoided if so called independent buffers are added which are not directly connected to a specific group of machines, but are shared between several groups. Jobs which could not be stored in the local buffer of their destination are temporarily stored in the independent buffer until space in the local buffer is available. Most flexible manufacturing systems provide an independent central buffer and local buffers at every machine to avoid the above mentioned throughput reductions. For flexible manufacturing systems, Tempelmeier and Kuhn [5] give a very complete overview about queueing models which allow the evaluation of flexible manufacturing systems with a central buffer. Most of these models rely on closed queueing networks and exponentially distributed service times. In job shop like manufacturing systems, other buffer configurations could be found as well which consist of several buffers linked in several levels.

In [6], an open queueing network model was developed which captured more general buffer configurations, but required exponentially distributed interarrival and service time for numerical evaluations. In this paper an extension is presented which allows the approximate calculation of performance measures for manufacturing systems with interarrival and service time which are described by general discrete time distributions. Network configurations with *Single-staged Buffers*, *Shared Single-staged Buffers* and *Multiple-*

staged Buffers are subsequently described and formulas for their performance evaluation are derived.

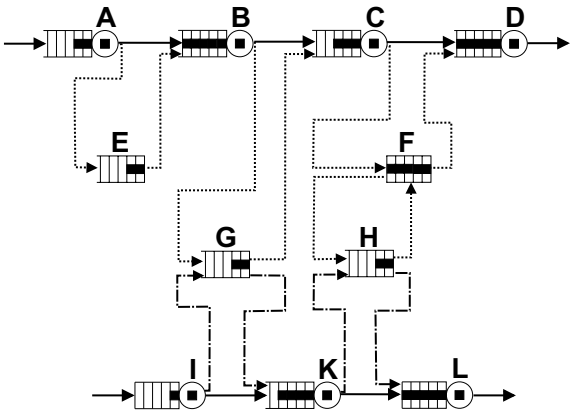

Figure 1: Different buffer configurations and resulting job routings. Solid line: planned route for jobs; dashed line: route via independent buffer

1.1 Single-staged Buffers

In Figure 1 the simplest buffer configuration is shown between queueing systems A and B with buffer E. Jobs enter the queueing network at queueing system A and are stored in the unlimited buffer of A if they cannot be processed directly. After completing the service at A, they proceed to B. If the limited local buffer of B is full, the job is transferred instead to the independent buffer E, and remains there until space in the local buffer of B is available. The transfer to B is done in a FCFS discipline.

1.2 Shared Single-staged Buffers

Shopfloor space can be saved if an independent buffer is used by several queueing systems to temporarily store the jobs that could not be directly transferred to their destination's local buffer. In Figure 1 the independent buffer G is used to keep jobs that could not proceed directly from B to C as well as those which could not be brought from I to K. The transfer to C or K respectively is also done FCFS separately for each destination as soon as space in the local buffer at the destination is available.

1.3 Multiple-staged Buffers

In the previous sections it was assumed that the capacity of the independent buffers is unlimited. In a job shop environment or a flexible manufacturing system this is usually not the case. Therefore it may be necessary to transfer jobs from a first level independent buffer to a second level independent buffer with infinite capacity. This is shown in Figure 1. Here the independent buffer F has a limited capacity, therefore jobs that have been routed to F because they could not be stored in the local buffer of D are rerouted to the buffer H with infinite capacity. H also serves as first stage buffer for L in this example.

2 Model Description

The manufacturing system is modeled by an open queueing network, where each machine is represented by a queueing system with a single server. The network consists of M queueing systems and P independent buffers. The size of buffer i is denoted by c_i, the total capacity of a queueing system is $c'_i = c_i + 1$. The elements of matrix Q, $q_{i,j}$ describe the routing probabilities between the queueing systems i and j. The adjacency relations between queueing systems and buffers are defined by an $M+P$-dimensional overflow-adjacency matrix R. An element $r_{i,j}$ of matrix R assumes the value '1', if blocked jobs with destination i are transferred to buffer j instead, otherwise $r_{i,j} = 0$.

To calculate performance measures, it is required that the queueing system satisfies the subsequently described assumptions:

- there is one class of jobs
- all external interarrival and service times are described by renewal processes with discrete time distributions and fixed interval length t_{inc}
- the service order is FCFS
- the adjacency relations defined in R are circular free
- the routing probabilities defined in matrix Q are independent of the current state of the system
- all queueing systems either possess a buffer with infinite capacity or have direct or indirect access to a buffer with infinite capacity.

3 Computing Performance Measures

To compute performance measures for a given buffer configuration two steps are necessary. First performance measures are computed for each queueing system in the network. The queueing systems are treated as having unlimited buffer capacities. The emphasis is set on the number of jobs in the queueing system at arrival instants. Based on these results, a subsequent evaluation of the effects of the buffer space assignment can be done.

3.1 Number of Jobs in Queueing System at Arrival Instants

The first step of the suggested evaluation method requires the computation of the distribution of the number of jobs in system.

It is assumed that all involved stochastic variables A (interarrival times), S (service times) and W (waiting times) describing a queueing system are identically and independently distributed for all jobs in a finite discrete time domain, with a constant t_{inc} being the increment between two adjacent steps. No two

jobs arrive in the same instant, thus the distributions of the service time S_n for an arbitrary job n, as well as the interarrival times A_n between the n-th and the $n+1$st job are given as:

$$\begin{aligned} P\{S_n = i \cdot t_{inc}\} &= s_i \quad \forall i = 0, \ldots, s_{max} \\ P\{A_n = i \cdot t_{inc}\} &= a_i \quad \forall i = 1, \ldots, a_{max} \end{aligned}$$

Using the results described in [7] to compute the waiting time and interdeparture time distributions for $G|G|1$-queueing systems and linking these systems, and applying the results of [8] to compute the effects of splitting and merging of streams, it is possible to compute good approximations for the waiting time distribution of jobs in the buffers of the queueing systems. It is assumed that the system achieves steady state and the waiting time distribution for the jobs exists and is described by

$$P\{W_n = i \cdot t_{inc}\} = w_i \quad \forall i = 0, \ldots, w_{max}$$

To calculate the number of jobs in the queueing system at arrival instants along the lines of [9] we proceed as follows.

Apparently, the distribution of the waiting times is closely linked to the distribution of number of jobs in queue at arrival instants. The task is now to derive from the distribution of W, the distribution of the number of N jobs in queue at arrival instants.

The jobs $C_j (j = 0, 1, \ldots)$, are coming successively to the regarded queueing system. When job C_n enters the system at instant T_n it has to wait W_n time units until it is about to be served. It leaves the system at the departure instant D_n. The next instant after D_n, $D_n + t_{inc}$, shall be denoted by D_n+. Job C_{n+k+1} enters the system at instant $T_{n+k+1} > T_n$. The probability of it finding not more than k jobs in the system equals the probability of job C_n leaving after its departure instant D_n not more than k jobs behind.

$$P\{N(T_{n+k+1}) \leq k\} = P\{N(D_n+) \leq k\} \quad (1)$$

It can be shown, that the conditions $N(D_n+) \leq k$ and $N(T_{n+k+1}) \leq k$ are equivalent:

Proof: Assume that $N(T_{n+k+1}) \leq k$. Then the n-th job has already been served before job C_{n+k+1} arrives. Therefore just after D_n, no more than k jobs could have been present at the queueing system. On the other hand, in case C_n leaves at D_n+ no more than k jobs in the system, C_{n+k+1} could have found no more than k jobs in the system.

$$\begin{aligned} &\text{if} \quad N(T_{n+k+1}) \leq k \quad \Rightarrow \quad N(D_n+) \leq k \\ &\text{if} \quad N(D_n+) \leq k \quad \Rightarrow \quad N(T_{n+k+1}) \leq k \\ \Rightarrow \quad &N(D_n+) \leq k \quad \Leftrightarrow \quad N(T_{n+k+1}) \leq k \end{aligned} \quad (2)$$

The distribution of the difference $T_{n+k+1} \Leftrightarrow T_n$ is the sum of $k+1$ independent realizations of A, and can be computed by using a $k+1$-fold convolution of the interarrival distribution which shall be denoted $A^{(k+1)}$. The element $a_l^{(k+1)*}$ at position l of the probability vector represents the probability, that the sum of $k+1$ subsequent arrival intervals add up to length $l \cdot t_{inc}$.

$$P\{T_{n+k+1} \Leftrightarrow T_n = l \cdot t_{inc}\} = a_l^{(k+1)*} \quad (3)$$

Using (1) and the independence of T_{n+k+1}, T_n, W_n, and S_n, another important relation can be established. If C_{n+k+1} finds k or less jobs present, then the interval between T_{n+k+1} and T_n must have been longer than it took C_n to wait and be serviced.

$$\begin{aligned} &P\{N(T_{n+k+1}) \leq k\} = P\{N(D_n+) \leq k\} \\ &= P\{T_{n+k+1} \Leftrightarrow T_n > W_n + S_n\} \end{aligned} \quad (4)$$

Considering the $n+1$-st job instead of the n-th, a simpler formula can be obtained.

$$\begin{aligned} &P\{T_{n+k+1} \Leftrightarrow T_n > W_n + S_n\} = \\ &P\{T_{n+k+1} \Leftrightarrow T_{n+1} > W_n + S_n \Leftrightarrow (T_{n+1} \Leftrightarrow T_n)\} \\ &= P\{T_{n+k+1} \Leftrightarrow T_{n+1} > W_{n+1}\} \end{aligned} \quad (5)$$

The service time of the n-th job is not relevant and it is sufficient to take into account the waiting time distribution. When using (5) it has to be considered however, that one job less has arrived. Thus for an arbitrary job n we get subsequent described probabilities for all $k \geq 0$:

$$\begin{aligned} P\{N(D_n+) \leq k\} &= \sum_{i=0}^{\infty} w_i \left[\sum_{j=i+1}^{\infty} a_j^{(k-1)*} \right] \\ &= \sum_{i=0}^{\infty} w_i (1 \Leftrightarrow \bar{a}_i{}^{(k-1)*}) \end{aligned} \quad (6)$$

with $\bar{a}_i$ denoting the i-th element of the cumulative probabilities vector. The initial condition

$$P\{N(D_n+) \leq 0\} = P\{W_{n+1} = 0\} \quad (7)$$

allows the computation of the probabilities for all values of $k \geq 1$ at arrival instants iteratively:

$$\begin{aligned} &P\{N(D_n+) = k\} = \\ &P\{N(D_n+) \leq k\} \Leftrightarrow P\{N(D_n+) \leq k \Leftrightarrow 1\} \end{aligned} \quad (8)$$

Taking the limit for $n \to \infty$, the probability $p(k)$ for finding $k \geq 1$ jobs in a queueing system at the arrival instant results from:

$$p(k) = \sum_{i=0}^{\infty} w_i \left[\bar{a}_i{}^{(k-2)*} \Leftrightarrow \bar{a}_i{}^{(k-1)*} \right] \quad (9)$$

and $p(0) = w_i$.

3.2 Transformation of Steady State Probabilities

Using the above described results, the subsequently described approximation method is proposed to compute the steady state probabilities of the number of jobs in the limited and unlimited buffers at queueing systems and independent buffers. The approximations assume that the steady state probabilities in

the open queueing network are independent for each queueing system.

The formulas of the previous section give probabilities $p(k_i)$ for the the number k_i of jobs in queueing system i at the arrival instant. These probabilities now have to be transformed in all cases where the number in queue surpasses the buffer size. The subsequently described transformations are necessary.

For each queueing system the transformed probabilities $p(k_i')$ are computed as follows. If $c_i' = \infty$ no transformation is necessary. For finite values of c_i' the transformation is defined by:

$$p(k_i') = \begin{cases} p(k_i) & \text{if} \quad k_i' < c_i' \\ 1 \Leftrightarrow \sum_{i=0}^{c_i'-1} p(k_i) & \text{if} \quad k_i' = c_i' \end{cases} \tag{10}$$

To deal with the independent buffers, all buffers (local and independent) are topologically sorted according to the adjacency matrix R. Then the modified probabilities $p(k_i')$ for queueing systems with limited buffers are computed according to (10). It is then possible to compute for all buffers, the probability of an overflow ($k_i > c_i'$) in the topological sorted order.

The variable o_i denotes the number of jobs that could not be stored in the local buffer of system i and have to be transferred to the adjacent independent buffer. The associated probabilities $p(o_i)$ are computed as follows:

$$p(o_i) = \begin{cases} p(k_i = c_i' + o_i) & \text{if} \quad o_i > 0 \\ \sum_{h=0}^{c_i'} p(k_i = h) & \text{if} \quad o_i = 0 \end{cases} \tag{11}$$

The probability vector P_i^o of the overflows is given by $P_i^o = (p(o_i = 0), p(o_i = 1), \ldots)^T$ The number of jobs in an independent buffer now is computed as the probability of the sum of the o_i of the adjacent queueing systems and buffers using the convolution operator $\otimes$ on the vectors P_i^o.

$$P(j) = \bigotimes_{j:r_{i,j}=1} P_i^o \tag{12}$$

3.3 Computing the Additional Flow

The additional flow from a queueing system to a buffer is necessary when a job arrives and the local buffer is already completely occupied. Based on the probabilities $p(k_i)$, which give the probability of number of jobs in queueing system i at arrival instants, the additional flow λ_i^o is easily calculated.

$$\lambda_i^o = \lambda_i p(k_i' = c_i') \tag{13}$$

The flow from an independent buffer h with limited capacity to another buffer l has to be traced back to the probability of a job arriving at i being rerouted to h and immediately further on to l. This probability is denoted as $p^{i \to h \to l}$. It can be derived from the probability that in the equivalent network with unlimited buffers, the local buffer of i already contains $k_i \geq c_i'$ jobs and the sum of jobs swapped out to h already exceeds $c_h' \Leftrightarrow (k_i \Leftrightarrow c_i')$. The set $\mathcal{J}$ contains all queueing systems and buffers that have a direct adjacency relation to h.

$$\mathcal{J} = \{j | r_{jh} = 1\}$$

Then probability $p^{i \to h \to l}$ can be expressed as

$$p^{i \to h \to l} = \sum_{k_i} p(k_i) \cdot p \left[\sum_{j \in \mathcal{J}} o_j \geq c_h' \Leftrightarrow (k_i \Leftrightarrow c_i') \right] \tag{14}$$

Using $p^{i \to h \to l}$ λ_h^o can be expressed as:

$$\lambda_h^o = \sum_{r_{ih}} \lambda_i \cdot p^{i \to h \to l} \tag{15}$$

This method can be extended to more than two stages by determining all nodes i from where an independent buffer h is reachable.

3.4 Examples

3.4.1 Example I

The first example is chosen in order to verify the obtained results by a simple example based on M/M/1 queueing systems. In this case the probabilities for the waiting times and number of elements in the system are easy to calculate. A manufacturing system consists of four sequentially arranged queueing systems with local buffers and one independent buffer. The interarrival and service times of each system are exponentially distributed with the arrival rate $\lambda = 8$ [jobs/time units] respectively, the service rate $\mu = 10$ [jobs/time units], the throughput was $\rho = \lambda/\mu = 0.8$. For the first time units the well known probabilities for number of jobs in the M/M/1 system $p(k) = \rho^k \cdot (1 \Leftrightarrow \rho)$ are listed in Table 1.

jobs	M/M/1 System	G/G/1 System	M/M/1 Buffer	G/G/1 Buffer
0	0.20	0.24	0.39	0.37
1	0.16	0.10	0.08	0.08
2	0.13	0.12	0.07	0.08
3	0.10	0.10	0.06	0.07
4	0.08	0.09	0.06	0.06
5	0.07	0.07	0.01	0.05
6	0.05	0.06	0.01	0.05
7	0.04	0.05	0.01	0.04
8	0.03	0.04	0.00	0.03
9	0.03	0.03	0.00	0.03

Table 1: Probabilities for the number of jobs in the M/M/1 and G/G/1 queueing systems and in the independent buffers

It is assumed that the buffer space of each queueing system is limited to five places, i.e. with the job being served, it is possible for six jobs to stay in the system.

The queueing systems are modelled as G/G/1 systems and analyzed in the discrete time domain. The discrete values representing the exponentially distributed interarrival and service times are obtained by the following method. The time is split up into several equi-distant time intervals. For each interval the corresponding discrete value is the difference of the two boundary values of the continuous exponential function. In our case, the time intervals have the length 1/20 [time units], the arrival and service rate referring to the intervals are now $\lambda = 0.4$ and $\mu = 0.5$. The waiting time probabilities for the M/M/1 system is obtained through the waiting time distribution (16) (see [10]).

$$w_i = 1 \Leftrightarrow \rho \cdot e^{-(\mu-\lambda)\cdot i} \tag{16}$$

The results for some discrete values of the M/M/1 system and the obtained results for the waiting time probabilities of the G/G/1 system are listed in Table 2.

i	0	1	2	3
$w_{i,M/M/1}$	0.200	0.076	0.069	0.062
$w_{i,G/G/1}$	0.237	0.071	0.065	0.059
i	4	$\cdots$	20	21
$w_{i,M/M/1}$	0.057	$\cdots$	0.011	0.010
$w_{i,G/G/1}$	0.053	$\cdots$	0.011	0.010

Table 2: Waiting time probabilities for the M/M/1 and G/G/1 system

The average waiting time of the M/M/1 system equals $\overline{w}_{M/M/1} = \rho/(\mu \cdot (1 \Leftrightarrow \rho))$=8.0 [time units]. The G/G/1 system has the calculated average of $\overline{w}_{G/G/1} = 7.779$ [time units], which is very close to $\overline{w}_{M/M/1}$.

Having obtained the waiting time probabilities, the distribution of the number of jobs in each queueing system is calculated with formula (9). The results for the M/M/1 system, shown in the table above, can be compared with the results of the G/G/1 system in Table 1.

The average number of jobs in a system equals $\overline{k}_{G/G/1} = 4.1$ which is comparable to the expected number of jobs in M/M/1 systems:
$\overline{k}_{M/M/1} = \lambda \cdot \overline{w}_{M/M/1} + \rho = 4.0$.

Finally, all jobs that does not fit into the appropriate system are transferred into the independent buffer. The results for the independent buffer in the M/M/1 queueing network, as well as for the buffer for the G/G/1 cases, are shown in Table 1.
It can be stated that the probability of having an empty independent buffer is slightly higher in the case of the M/M/1 systems. On the other side, the probabilities with at least one job in the buffer are smaller and with at least five jobs in the buffer, the difference to the G/G/1 system is enormous. This is explained by the fact that the calculation for the M/M/1 system is an approximation, i.e. the probabilities of number of jobs in the system is limited to 10 jobs (see [11]) which becomes especially apparent when the distributions of number of jobs that do not fit into the queueing systems are convoluted in order to get the probabilities for the buffer occupation.

3.4.2 Example II

The next example is chosen to demonstrate the effects of differently distributed service and interarrival times in a queueing network, and the consequences of independent buffers.

The network consists of three queuing systems, see Figure 2. The external interarrival time distribution of system A is shown in Table 3, the service time distributions are documented in Table 4.

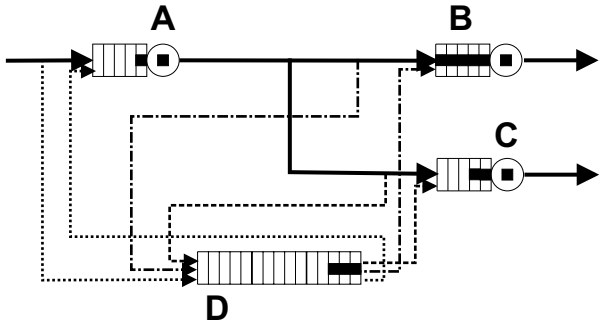

Figure 2: Queueing network

interarrival time	0	1	2	3	4
System A	0.0	0.2	0.5	0.8	1.0

Table 3: Interarrival time distribution for A

service time	0	1	2	3	4
System A	0.0	0.3	0.6	1.0	1.0
System B	0.0	0.5	0.9	1.0	1.0
System C	0.0	0.4	0.7	0.9	1.0

Table 4: Service time distributions for the queueing network

The resulting probabilities for number of jobs in each system are shown in Table 5.

Analyzing the results, it appears convenient to allocate more buffer spaces to the systems B and C, say two and three places, and let one place belong to system A. Then the number of jobs in the shared independent buffer D would have the probabilities shown in Table 5.

This means that it would be sufficient to allocate two places to the independent buffer D in order to catch all jobs that do not fit into their designated buffers.

Another solution for the buffer configuration is to direct all transferred jobs to the local buffer of a sys-

jobs	Sys. A	Sys. B	Sys. C	Buff. D	A (new)
0	0.795	0.778	0.567	0.929	0.733
1	0.138	0.167	0.252	0.070	0.193
2	0.060	0.049	0.130	0.001	0.070
3	0.000	0.005	0.032	0.000	0.004
4	0.000	0.000	0.006	0.000	0.000
5	0.000	0.000	0.001	0.000	0.000

Table 5: Probabilities for the number of jobs in the systems A.B.C, in the independent buffer and in system A for the new configuration

tem, system A would be convenient. The new distribution of number of jobs in system A would have the probabilities shown in Table 5.

The addition of two further places to system A would have similar performance results as installing an independent buffer with two places. There are other aspects that influence the decision where to place the buffers, e.g. if the costs of separate buffer space would be greater than those of the local buffer, the latter solution would be preferred. Another point to take into consideration is the distance of the manufacturing systems to the independent buffer, which results in different costs for the transport for all systems to the shared buffer.

References

[1] B. Connors, G. Feigin, D. Yao, *A Queueing Network Model for Semiconductor Manufacturing System*, Submitted to IEEE Transactions on Semiconductor Manufacturing, 1994

[2] V. Dörrsam, K. Furmans, *Application Case Study of a Queueing Network Simulation Tool for Analyzing and Optimizing a Manufacturing System*, European Simulation Symposition, ESS'95, Erlangen, 1995

[3] H. Chen, J. M. Harrison, A. Mandelbaum, A. van Ackere, L. M. Wein, *Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication*, in: Operations Research, Vol. 36, No. 2, 1988, p. 202 - 215

[4] I. Akyildiz, *Product Form Approximations for Queueing Networks with Multiple Servers and Blocking*, IEEE Transactions on Computers, Vol. 38, No. 1

[5] H. Tempelmeier, H. Kuhn, *Flexible Manufacturing Systems*, Decision Support for Design and Operation, Wiley, New York, 1993

[6] K. Furmans, *An Open Queueing Network Model of Manufacturing Systems with Independent Production Buffers*, in: W. Krug, A. Lehmann (Eds.) *Simulation and AI in Computer Aided Techniques*, Proceedings of the European Simulation Symposium, Dresden, 1992, p. 560 - 565

[7] W. K. Grassman, J. L. Jain, *Numerical Solutions for the Waiting Time Distribution and Idle Time Distribution of the Arithmetic GI/G/1 Queue*, in: Operations Research 37, 1989

[8] K. Furmans, *A Discrete Time Model of a Car Assembly System*, in: *Emerging Technologies in Manufacturing and Automation*, Proceedings of the INRIA/IEEE Conference, Paris, 1995.

[9] R. Schassberger, *Warteschlangen*, Springer-Verlag, Wien, 1973

[10] B. W. Gnedenko, D. König, *Handbuch der Bedienungstheorie*, Band 2, Formeln und andere Ergebnisse, Akademie-Verlag, Berlin, 1984

[11] K. Furmans, *Ein Beitrag zur theoretischen Behandlung von Materialflußpuffern in Bediensystemnetzwerken*, Dissertation am Institut für Fördertechnik, Karlsruhe, 1992