A decomposition approach for the analysis of discrete-time queuing networks with finite buffers

Dr.-Ing. Judith Stoll, Judith.stoll@kit.edu, Institute for Material Handling and Logistics (IFL), Karlsruhe Institute of Technology (KIT)

Abstract: This paper describes an approximation procedure for determining the throughput time distribution of material handling and service systems modeled as queuing networks. We consider finite buffer capacities and general distributed processing times in terms of discrete probability functions. In order to quantify the influence of blocking caused by finite buffers, we present a decomposition approach. Two-server subsystems of the queuing network are analyzed subsequently to obtain the throughput time distribution of the whole queuing system. The quality of the presented approximation procedure is tested against the results of various simulation experiments.

1 Introduction and Problem Description

Health establishments are faced with growing health expenses, while the society tries more and more to reduce health expenses in order to guarantee its social welfare. This leads to the search of efficiency in order to limit the increase of expenses, which is a well known problem in the context of logistics systems. Therefore, some authors started with the performance analysis of support services in health establishments, e.g. the sterilization of medical devices. Starting with a simulation model of a generic sterilization process (see Di Mascolo et al. 2009), analytical methods which were previously used to analyse the material flow in production systems, were applied to analyze the sterilization service in health establishments (see Stoll and Di Mascolo 2013). Discrete time queuing models have been used as a method to achieve a fast and quite accurate way of determining performance figures of service systems. While with classical general queuing models characteristic values are calculated only on the basis of means and variances, in discrete-time modeling all input and output variables are described with discrete probability distributions.

Grassmann and Jain (1989) were the first who analyzed the waiting time distribution in a G|G|1 queuing system in discrete time domain. In addition to their model, several basic elements for network analysis have been treated analytically which can now be combined (Stochastic Finite Elements, see publications of Furmans, Zillus, Schleyer, Matzka/Stoll and Özden between 1996 and 2012). For an overview of the existing discrete-time models see Matzka (2011). The advantage of these models lies in the fact, that they allow the computation of not
only averages but also the distributions of e.g. waiting times. This enables the derivation of quantiles of performance measure, which are often needed for the design of logistics systems.

All the existing discrete time queuing models with general distributions assume infinite buffers. This assumption is not valid for many practical cases. In the sterilization area of health establishments for example, buffers are limited, too. The space for buffering medical devices between the single sterilization steps is limited as there is a given number of racks or boxes to buffer the devices, before the next process step can start (see Stoll and Di Mascolo 2013). If one of these buffers is full, this can cause blocking in upstream process steps. In real material handling and service systems, finite buffers cause blocking situations which can be classified in three main types of blocking: Blocking After Service (BAS), Blocking Before Service (BBS) and Repetitive Service Blocking (RS) (Onvural 1990). For an overview on queuing networks with blocking see Perros and Altioik (1994), Balsamo et al. (2001) and Manitz and Tempelmeier (2012). To our knowledge there is no paper dealing with the analysis of queuing networks with blocking with general distributed service times given in terms of discrete probability functions. Existing models are either dealing with two-parameter approximations or known discrete probability functions, e.g. the binomial distribution. Thus, we will now present a method that enables us to quantify the influence of blocking after service in discrete time queuing networks with general distributed service times.

The paper is organized as follows: First we will give a short introduction to discrete-time modeling in section 2. In section 3 we give an overview of discrete time modeling of the sterilization process in health establishments. In section 4, we present the approximation method for the calculation of the waiting time distribution of blocked queuing systems in discrete time domain. As the presented approach is an approximation we want to give some insights to the quality of this approximation comparing analytical results to simulation results (section 5). In section 6 we give a conclusion and an outlook on further research on this topic.

## 2 Discrete-time modeling

Analysis in discrete time domain assumes that time is not continuous but discrete. This means, that events are only recorded at discrete moments which are multiples of a constant increment \( t_{inc} \). These events occur, when items are moved or when they change their status, for example by entering a queue, by being served, by merging with a stream of other items or at a split of a stream. In our analysis, events are described by a discrete random variable. When we have given a discrete random variable \( X \), we denote its distribution, which is also called probability mass function (pmf), by

\[
P(X = i \cdot t_{inc}) = x_i \quad \forall \ i = 0, 1, ..., i_{max} \quad (1)
\]

As a simplification we reduce this notation to

\[
P(X = i) = x_i \quad \forall \ i = 0, 1, ..., i_{max} \quad (2)
\]

When we talk about a distribution in the subsequent sections, we refer to the probability mass function.
3 Sterilization process in health establishments

In a sterilization process, reusable medical devices are re-injected in the process after their use in the operation room. When we integrate the use step, the sterilization process becomes a sterilization loop, with the following steps: use, pre-disinfection (including the transfer from the operating rooms to the sterilization service), rinsing, washing, verification, packing, sterilization, transfer from the sterilization service to the operating rooms, storage, before a new use (Di Mascolo et al. (2006)).

In order to improve the performance of the system, different scenarios for the transfer of medical devices to the sterilization area can be analyzed. In practice, the transport is normally not following a certain rule. This unsteadiness can cause a duration of pre-disinfection that does not lead to the desired effect on the material. Modeling the different possibilities of transport organizations would enable us to compare their performance and especially show the impact that a modification of the transport would have.

In Stoll and Di Mascolo (2013), we used a discrete time queueing network model to analyze some performance figures of a particular sterilization process. In a previous work, Di Mascolo et al. (2006) analyzed a specific health establishment via simulation. We used the same input data to compare our queueing network model to the accordant simulation model.

From the queueing model, we obtained two important performance figures that we compared to the simulation results. One of them is the average duration of the pre-disinfection step. The ideal duration of pre-disinfection, to guarantee an optimal impact of the disinfection liquid to the medical devices, is about 15 minutes. On the other hand, the sojourn time in the liquid should not exceed 50 minutes, because the disinfection product attacks the material, and thus causes a premature ageing. We thus want to know the average pre-disinfection time as well as the percentage of medical devices that stay in the liquid more than 50 minutes. Compared to simulation, our queueing network model obtains quite good results for the average pre-disinfection time (simu: 29.40 min, analysis: 30.50 min) and the percentage of medical devices, that stay in the liquid more than 50 minutes (simu: 92.90 %, analysis: 91.46 %). The deviations are caused by the fact that we had to make some assumptions in the queueing network model.

A second important parameter is the throughput time of medical devices through the sterilization process before they are ready to be used again. From a survey in the Rhone-Alpes region, we know that many health establishments are not able to estimate the duration of their sterilization process (see Di Mascolo et al. (2006)). The average throughput time calculated by our queueing network model can give them an idea of the cycle time of medical devices. This figure also allows us to compare different loading policies for the washers and autoclaves and the influence of the number of parallel machines to the throughput time in discrete time queueing networks with general distributed processing times.

All the existing discrete time queueing models with general distributions we used for the sterilization model assume infinite buffers. This assumption is not valid in many cases. In the sterilization area of health establish-
ments, buffers are limited, too. The space for buffering medical devices between the single sterilization steps is limited as there is a given number of racks or boxes to buffer the devices, before the next process step can start. Thus, it is our intention to get an approach that helps us to quantify the influence of blocking on the throughput time.

4 Decomposition approach

Let us regard a queuing network, consisting of several queuing systems in series. For each node of the network, the service time distribution is given in terms of a general discrete probability function by

$$P(B = \beta) = b_\beta \quad \forall \beta = 0, 1, ..., \beta_{\text{max}} \quad (3)$$

If the buffers in front of the servers would be infinite, we could use the methods of Grassmann and Jain (1988 and 1989) to calculate the waiting time distribution for each queuing system and the inter departure time distribution that connects the nodes of the network. The distribution of the number of customers in the system could be calculated according to the method of Furmans and Zillus (1996). As we assume the buffers to be finite, we have to consider blocking of servers if a succeeding buffer is full. This has an influence on the waiting time of jobs in the buffer in front of a blocked server.

In order to consider the influence of finite buffers in discrete-time queuing networks we propose the following decomposition approach. We define subsystems of a queuing network consisting of two queuing stations in series (see figure 1). Starting from the most downstream server, we quantify the influence of blocking to the waiting time at the upstream server(s).

![Two-server subsystem with blocking after service](image)

**Figure 1:**
Two-server subsystem with blocking after service

The first server (upstream server) is modeled as a discrete time G|G|1 queuing system with infinite buffer and the second server (downstream server) is modeled as a G|G|1-k queuing system with a finite buffer capacity k, which means that a maximum of k units can wait in the queue. When the buffer of the second queuing system is full, a material unit that is finished cannot leave server 1 after service. The upstream server is blocked and stops further processing. We analyze the influence of blocking to the sojourn time of a job in the upstream server, and thus the waiting time of jobs in the upstream buffer. We define sojourn time in this context as the time, the job spends in the server, that is service time and blocking time.

We distinguish three cases for the sojourn time distribution of jobs in an upstream server:
1. If the downstream buffer is not full (number of customers in the queuing system \(N < k + 1\)), when a job of the upstream server is finished and wants to leave the server, there is no blocking and the job enters the downstream buffer. The sojourn time is then equal to the service time of server 1, and the distribution of sojourn time \((\hat{s})\) is equal to the distribution of the service time of server 1 \((\hat{b}_1)\): 
\[ \hat{s} = \hat{b}_1. \]

2. If at the departure of a job, the succeeding buffer is full \((N = k + 1)\), the server is blocked and the sojourn time of the job increases by the residual service time of server 2 and the distribution of the sojourn time can be calculated by convoluting the service time distribution of server 1 and residual service time distribution of server 2 \((\hat{r}_2)\): 
\[ \hat{s} = \hat{b}_1 \otimes \hat{r}_2. \]

3. After blocking, the next job in the upstream server and the next job in the downstream server start simultaneously. If the service time \(\beta_1\) of the job in server 1 is shorter than the service time \(\beta_2\) of the job in server 2 \((\beta_1 \leq \beta_2)\), the upstream server is blocked again and the sojourn time of the job in server 1 is equal to the service time in server 2 \((\hat{s} = \hat{b}_2)\). In the opposite case, server 1 is not blocked and the sojourn time distribution in server 1 is equal to his service time distribution \((\hat{s} = \hat{b}_1)\).

The three cases build a closed markov chain, where the states are given by the three different cases to calculate the sojourn time of a customer in the upstream server (see figure 2). The transitions between the system states can be interpreted as follows:

If a job was finished without blocking, his sojourn time was \(\hat{s} = \hat{b}_1\). The following job also has a sojourn time of \(\hat{s} = \hat{b}_1\), if the number of customers \(N\) in the succeeding queuing system is smaller than \(k + 1\) in his departure moment. If there are \(k + 1\) customers in the system, the queue is full and the sojourn time increases to \(\hat{s} = \hat{b}_1 \otimes \hat{r}_2\).

Note that we calculated the number of customers in the system using the method of Furmans and Zillus (1996) for infinite buffers, and thus have to normalize the probabilities with \(P(N \leq k + 1)\). If a job was blocked and had a sojourn time of \(\hat{s} = \hat{b}_1 \otimes \hat{r}_2\), the following job starts his service simultaneously with the job in server 2. If the service time of the job in server 1 is smaller than the service time in server 2 \((\beta_1 \leq \beta_2)\), the sojourn time of server 1 equals the service time of server 2. In the opposite case, the sojourn time is equal to the service time of server 1. When the sojourn time of server 1 equals the service time of server 2, the following job also starts his service simultaneously with the job in server 2 and the same transitions are valid.
Figure 2:
Markov-chain with sojourn times for succeeding customers

From the above shown markov chain, we get a set of linear equations that leads to the following system state probabilities:

\[
P(\vec{s} = \vec{b}_1) = 1 - \frac{x}{x+y} \left(1 + \frac{x}{x+y}\right)
\]

(4)

\[
P(\vec{s} = \vec{b}_1 \otimes \vec{b}_2) = \frac{x}{1+x} (1 + \frac{x}{x+y})
\]

(5)

\[
P(\vec{s} = \vec{b}_2) = \frac{x}{x+y}
\]

(6)

with

\[
x = \frac{P(N=k+1)}{P(N \leq k+1)}
\]

(7)

and

\[
y = P(\beta_1 > \beta_2)
\]

(8)

Using the analytical method of Furmans and Zillus, we can calculate the system state probabilities (system state = number of customers in the system at the arrival of a customer) of a G|G|1 queuing system with infinite buffers. Thus, we can approximate the blocking probability \(P(N = k + 1)\).

The probability that the service time of server 1 is higher than the service time of server 2 can be calculated according to the following formula:

\[
P(\beta_1 > \beta_2) = \sum_{i=0}^{\beta_{1,\max}} \sum_{j=0}^{i-1} b_{1,i} b_{2,j}
\]

(9)
Knowing the probabilities for the sojourn times of a job in a server in each of the three cases, we can calculate a service time distribution of server 1 as follows:

\[
\vec{b}_{mod} = P(\xi = \vec{b}_1) \cdot \vec{b}_1 + P(\xi = \vec{b}_1 \otimes \vec{f}_2) \cdot \vec{b}_1 \otimes \vec{f}_2 + P(\xi = \vec{b}_2) \cdot \vec{b}_2
\]  

(10)

With this modified service time distribution we can calculate the waiting time distribution of queueing system 1 according to the method of Grassmann and Jain (1989). The number of customers in the system can again be calculated by the method of Furmans and Zillus (1996). We then declare server 1 as the downstream server of a new subsystem. These steps are repeated until the first node of the network is reached. Knowing the waiting time distributions and service time distributions of each node, the throughput time distribution of the complete network can be calculated.

5 Analysis of the approximation quality

As the presented approach is an approximation we want to give some insights to the quality of this approximation. For an example with two succeeding servers, we can compare the waiting time distribution calculated with the analytical approach to the distribution obtained by simulation as shown in figure 3.

![Figure 3: Waiting time distribution in queueing system 1 obtained by analytical approach and simulation](image)

In order to make the experiments comparable, we calculate the deviation of the mean waiting time of analysis and simulation and show its dependence to the blocking probability. The blocking probability is influenced by the following parameters: number of buffer spaces k, variability of departure process from server 1, variability of the service process of server 2. Table 1 shows the results for a couple of experiments.
We can see that the approximation quality is quite high for low blocking probabilities. The higher the blocking probability gets, the higher are the deviations between analytical results and simulation results. In practice, blocking probabilities of more than 5% would not be tolerated. The buffer spaces would be increased or the process would be improved to be more stable. Thus, for practical applications, the analytical approach is useful to analyze the waiting time distribution in front of a server with blocking.

### 6 Conclusion and Outlook

In this paper we presented an analytical approach for the analysis of blocking after service in discrete time queuing networks. For the analysis we build two-server subsystems and approximate the influence of blocking by a modified service time distribution for the blocked server. The approximation quality is good for low blocking probabilities and therefore applicable for practical cases. Our next step will be to apply the new method to our model of the sterilization process in health establishments replacing the model elements with infinite buffers by finite buffer.

### 7 Acknowledgment

This research is supported by the research project “Analytische Berechnung von Durchlaufzeiten in komplexen Stetigfördersystemen”, funded by the Deutsche Forschungsgemeinschaft (DFG) (reference number FU 273/12-1)

### References


Schleyer, M.: *An analytical method for the calculation of the number of units at the arrival instant in a discrete time G/G/1-queueing system with batch arrivals*. OR Spectrum 34(1), pp 293-310, 2012

