Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach

Stolletz, Raik

Published in:
European Journal of Operational Research

Link to article, DOI:
10.1016/j.ejor.2007.06.036

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 190 (2008) 478–493
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach ☆

Raik Stolletz *

University of Hannover, Department of Production Management, Königsworther Platz 1, 30167 Hannover, Germany

Received 18 December 2006; accepted 17 June 2007
Available online 30 June 2007

Abstract

This paper proposes a new approach for the time-dependent analysis of stochastic and non-stationary queueing systems. The analysis of a series of stationary queueing models leads to a new approximation of time-dependent performance measures.

Based on a stationary backlog-carryover (SBC) approximation of the time-dependent expected utilization, different approximations of the time-dependent expected queue length and the number of customers in the system are discussed. Limiting results are given for the case of constant rates.

The accuracy of the SBC approach is shown for non-stationary $M(t)/M(t)/c(t)$ queueing systems with time-dependent and piecewise constant arrival rates. In numerical experiments we demonstrate the reliability of this approach and compare it with the (lagged) stationary independent period by period (SIPP) approach. In addition, the approximation is applied to temporarily overloaded systems that cannot be analyzed by the variants of the SIPP approach.

Keywords: Queueing; Non-stationary approximation; Time-dependent analysis; Temporarily overloaded systems

1. Introduction

Many service and production systems operate under dynamic conditions which leads to time-dependent performance measures like average number of jobs or customers in system. For example, the arrival rate of new jobs in a machine center may change or resources may become temporarily unavailable, see for example Lin and Cochran (1990) or Stahlman and Cochran (1998). Kolesar et al. (1975) and Ingolfsson et al. (2002) describe an approach for scheduling police patrol cars using the analysis of a non-stationary queueing system. Applications for agent staffing in call centers are described for example in Gans et al. (2003), Green et al. (2003), and Testik et al. (2004). These systems

☆ The research was supported by the Niderzaksen Fund 2003/2004 (German Technion Society). I am grateful to Kirsten Henken for providing the simulation environment and to Stefan Helber for helpful comments on an earlier version of this paper. The author thanks the anonymous referees for their helpful comments and suggestions.
* Tel.: +49 511 7625649; fax: +49 511 7624863.
E-mail address: raik.stolletz@prod.uni-hannover.de

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2007.06.036

can often be analyzed using queueing models with a relatively simple structure, but the dynamic nature of the systems complicates the analysis. Another characteristic trait of these queueing systems is that the arrival rate may exceed the service capacity for extended periods of time, which we call temporal overloading. During overloaded periods queues build up and arriving customers are served only after a possibly large waiting time. Methods for the performance evaluation of such queueing systems have to take this temporal overloading into account.

Besides the evaluation of given staffing plans, computationally fast methods for evaluating the time-dependent performances are often necessary to create staffing requirements and workforce schedules, see for example Ingolfsson et al. (2002). For optimization tasks in many service systems, the cost function often consists of two components: costs of waiting and costs of server staffing, see Stolletz (2003) for different examples of such cost functions. Traffic intensities greater than one (overload situation) lead to waiting customers, but the servers are well utilized. In the case of a low traffic intensity (underload situation), the server utilization is low and so is the probability of delay. The optimal server allocation has to balance these two cost components. In many practical cases, the servers are moderately utilized at most times t, but operate at the critical load (between under- and overload) or are overloaded during extended periods of time. Therefore, accurate and computationally fast methods for the performance evaluation of such queueing systems are required.

Especially in contact center applications, stationary analysis is a common approach for dealing with time-varying rates and stochastic inter-arrival and processing times. The *stationary independent period by period* (SIPP) approach divides the time interval of interest (for example a day) into T small periods $i = 1, \ldots, T$ with constant arrival rates and numbers of servers within each period i, see Green et al. (2001). A series of independent stationary queueing models are solved for each period i to derive stationary performance measures. The SIPP approach works accurately if the delays in consecutive time intervals are statistically independent of each other, the system achieves a steady-state in each interval, and the arrival rate does not change during a period (see Green et al., 2001). For underloaded systems with a high quality of service, the SIPP approach approximates the performance measures of the dynamic system well, see Gans et al. (2003), Green et al. (2007), and Kwan et al. (1988). A problem of such a steady-state analysis is that for each period the demand rate must be strictly smaller than the service rate, i.e., overloading is prohibited. If some extended periods are overloaded stationarity may also fail in subsequent underloaded periods because of the possibly large quantity of waiting customers transferred to later periods, see Jimenez and Koole (2004). However, many service systems in reality can be temporarily overloaded and queues build up with substantial numbers of waiting customers.

For systems which operate near the critical load or which are temporarily overloaded, the SIPP approach may fail. This paper presents a new approximation of time-dependent performance measures through the analysis of a series of stationary queueing models for such systems with temporal overloading. Contrary to the SIPP approach, the models of consecutive periods are no longer independent of each other. A *backlog* b_i of work is measured in each period i and *carried over* into future periods $j > i$. For this reason the method is called the *stationary backlog-carryover* (SBC) approach. This approximation allows queues to build up in overloaded periods and waiting jobs can be transferred to a subsequent period. The proposed method is general and different queueing systems can be analyzed with the SBC approximation. This paper presents the SBC approach for the non-stationary $M(t)/M(t)/c(t)$ queueing system which is applicable to temporarily overloaded systems. The focus of this paper is to demonstrate the accuracy of the SBC approach by comparing it to other good approximations with stationary queueing models in underloaded situations. We also show the reliability of the SBC approach for systems that operate near the critical load or that are temporarily overloaded. However, the integration into optimization algorithms, for example for agent staffing in call centers, is outside the scope of the paper.

The remainder of the paper is organized as follows: Section 2 reviews literature on the time-dependent analysis via stationary queueing models and discusses their applicability to temporarily overloaded systems. Section 3 describes the SBC approach for the $M(t)/M(t)/c(t)$ queue and presents some limiting results for the approximation of the time-dependent behavior of the $M/M/c$ queue with constant arrival rates. Different approximations of the expected number of customers in system and in queue are compared. The accuracy of the SBC

approximation and the (lagged) SIPP approach is shown in numerical experiments in Section 4. Underloaded as well as temporarily overloaded systems are studied and the results are compared to simulations. A conclusion and extensions of this approximation method to other queueing models are given in Section 5.

2. Time-dependent analysis of $M(t)/M(t)/c(t)$ queues

This section reviews some methods for the time-dependent analysis of non-stationary $M(t)/M(t)/c(t)$ queueing systems. The assumptions of the model are described and the differential-difference equations for the exact solution for one model formulation are presented. The numerical solution of these differential-difference equations as well as different approximations are briefly discussed. Since our approximation uses stationary queueing models, the review of literature focuses on approximations with stationary queueing models. Possible applications of these methods to temporarily overloaded systems are discussed.

The $M(t)/M(t)/c(t)$ system has an inhomogeneous Poisson arrival process with instantaneous arrival rates $\lambda(t)$. The service time for a customer that starts service at time t corresponds to the time until the first arrival of a non-homogenous Poisson process that starts at time t and has rate $\mu(s)$ for $s \geqslant t$. For time-independent service rates μ it follows that the service times are exponentially distributed. The number of waiting positions is assumed to be infinite. Waiting customers will be served according to First-Come-First-Serve (FCFS). The number of servers $c(t)$ is time-varying. If servers are scheduled to leave an end-of-shift policy has to be specified for the customers they are serving. If a server is providing service when scheduled to leave a pre-emptive or an exhaustive discipline can be assumed, compare Ingolfsson (2005). In the pre-emptive discipline a customer in service is sent back to the queue if the server is scheduled to leave. For the exhaustive discipline it is assumed that the servers always complete the current job before leaving. Many service organizations work according to an exhaustive discipline. Although the exhaustive discipline cannot be considered explicitly in our approximation approach, we developed different simulation models for both disciplines. These simulation models are used to evaluate the accuracy of our approximation for the pre-emptive and the exhaustive discipline.

Let $\rho(t) = \frac{\lambda(t)}{c(t)\mu(t)}$ be the instantaneous traffic intensity. The traffic intensity function $\rho^*(t)$ is defined similarly to the function used in Mandelbaum and Massey (1995), through

$$\rho^*(t) = \sup_{0 \leqslant s < t} \frac{\int_s^t \lambda(r)\,\mathrm{d}r}{\int_s^t c(r)\mu(r)\,\mathrm{d}r} \tag{1}$$

for $t > 0$ and $\rho^*(0) = \rho(0)$ for $t = 0$. According to Mandelbaum and Massey (1995) the system operates in one of the following three regimes at time t: underloaded for $\rho^*(t) < 1$, critically loaded for $\rho^*(t) = 1$, and overloaded for $\rho^*(t) > 1$.

Let $p_n(t)$ be the time-dependent probability that n customers or jobs are in the system at time t. The exact behavior of the system with a pre-emptive discipline is described by the set of the following differential-difference equations, see for example Kleinrock (1975):

$$\frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = -\lambda(t)p_0(t) + \mu(t)p_1(t), \tag{2}$$

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = -(\lambda(t) + n\mu(t))p_n(t) + \lambda(t)p_{n-1}(t) + (n+1)\mu(t)p_{n+1}(t) \quad \text{for } 0 < n < c(t), \tag{3}$$

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = -(\lambda(t) + c(t)\mu(t))p_n(t) + \lambda(t)p_{n-1}(t) + c(t)\mu(t)p_{n+1}(t) \quad \text{for } c(t) \leqslant n. \tag{4}$$

For the equations describing the system with an exhaustive discipline we refer to Ingolfsson et al. (2007). This set of differential-difference equations has a closed-form solution only for special cases.[1] For a numerical solution, the system size will be restricted, i.e., an $M(t)/M(t)/c(t)/K$ queue approximates the $M(t)/M(t)/c(t)$ queue, see for example Green et al. (1991). With a maximal system size of K customers, Eq. (4) holds for $c(t) \leqslant n < K$ and for $n = K$ we have

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = -c(t)\mu(t)p_n(t) + \lambda(t)p_{n-1}(t). \tag{5}$$

This truncation works well if the probabilities $p_K(t)$ are negligibly small at all times t. For an overview of numerical methods for the solution of these equations (and hints to choose the number K), see for example Stewart (1994) or de Souza e Silva and Gail

[1] For example, explicit expressions of the time-dependent probabilities $p_n(t)$ for an $M/M/c$ system with $c \leqslant 4$ servers are given in Parthasarathy and Sharafali (1989). They assume constant arrival and service rates.

(2000). The solution of these ordinary differential equations (ODE) can be numerically challenging and requires high computation times, especially for systems with a high offered load, compare Ingolfsson et al. (2007).

Now we will focus on literature on the *approximation through stationary queueing models* and discuss how they can be applied to temporarily overloaded systems. Well-known approaches are the simple stationary approximation (SSA), the pointwise stationary approximation (PSA), the stationary independent period by period (SIPP) approach, the average stationary approximation (ASA), and the effective arrival rate approximation (EAR).

The *simple stationary approximation* (SSA) assumes a constant number c of servers and a constant processing rate μ. The arrival rate is averaged over the whole time interval of interest (for example a day) and a stationary $M/M/c$ model is applied with this arrival rate. Therefore, any non-stationarity is ignored in this approach. Green et al. (1991) analyze the question of how non-stationary an arrival process can be before the SSA fails. They found that the SSA gives reasonable results for small systems with a low relative amplitude of the arrival rate function. If the aggregated stability condition

$$\rho = \frac{1}{T}\int_0^T \frac{\lambda(t)}{c\cdot\mu}\,\mathrm{d}t < 1 \tag{6}$$

is satisfied, the SSA can be applied to temporarily overloaded systems, but the time-dependent behavior is ignored.

For the *pointwise stationary approximation* (PSA) the performance measures are approximated by a stationary $M/M/c(t)$ model with instantaneous rates $\lambda(t)$ and $\mu(t)$ at each time t, see Green and Kolesar (1991) and Green et al. (1991). The PSA can only be applied to systems with a maximum traffic intensity less than one, i.e., temporal overloading in some periods is strictly forbidden. Green and Kolesar (1991) analyze the reliability of the PSA for approximations of the expected queue length and the expected waiting time. They found that the PSA improves as the event frequency increases or the service rate increases.[2] The PSA worsens as the maximum traffic intensity increases. Steckley and Henderson (2007) develop approximations for the error involved in using the PSA. Whitt (1991) proves that the PSA is asymptotically correct as the service and arrival rates increase, while the instantaneous traffic intensity and the number of servers are fixed.

The PSA and the SSA use an effective arrival rate which is averaged over an interval of length zero for the PSA and a length of the whole interval of interest for the SSA, see Jennings et al. (1996). The *stationary independent period by period* (SIPP) approach averages the rates for intervals of arbitrary length. For underloaded systems with a high quality of service, the SIPP approach approximates the performance measures of the dynamic system well, see Green et al. (2007). The performance can be improved using a lagged version of the approach, which shifts the planning period one average service time into the past (Green et al., 2003). For this shifted planning period the maximum or the average arrival rate is chosen. After these steps, the arrival rate λ_i, the processing rate μ_i, and the number c_i of servers are held constant during a single period i. The SIPP approach determines the steady-state performance measures for each period i in isolation, i.e., an $M/M/c_i$ model with rates λ_i and μ_i is analyzed for period i. A problem of such a steady-state analysis is that the demand rate may never exceed the service rate, i.e., the stability condition

$$\rho_i = \frac{\lambda_i}{c_i\cdot\mu_i} < 1 \tag{7}$$

has to be satisfied for each period $i = 1,\ldots,T$. Therefore, a system with temporal overloading in single periods cannot be analyzed with the (lagged) SIPP approach.

The *average stationary approximation* (ASA) averages the arrival rate for time t over the interval $[t-m,t]$, where m is chosen to be proportional to the mean service time, see Whitt (1991). The average performance over some intervals is computed using an $M/G/c$ queue with these average arrival rates. Temporal overloading is restricted to cases where the aggregated stability condition is satisfied for the intervals $[t-m,t]$ for all t.

The *effective arrival rate approximation* (EAR) of Thompson (1993) uses the SSA to derive the expected waiting time $E[W_q]$. Then he assumes that the waiting time of each customer equals $E[W_q]$. Under the additional assumption that the service times are deterministic, effective arrival rates are derived. The effective arrival rate for period i is the average original arrival rate of this period

[2] They analyze a system with a constant service rate and a constant number of servers during the whole time interval.

increased by the rate of customers who arrived in earlier periods $j < i$ but are served in period i. He decreases the original arrival rate by the number of customers arriving in i but who are served in later periods $k > i$. Then a stationary model with this effective arrival rate is applied in each period. The effective arrival rate has to fulfill the stability condition for each period i, which again restricts the analysis of temporarily overloaded systems.

Therefore all these approximations with stationary queueing models cannot be applied without restrictions to the analysis of temporarily overloaded systems.

Beside these approximations with stationary queueing models other approaches are discussed in literature.

Fluid models can be used to derive time-dependent performance measures for systems without stochastic variations. In these models, the discrete processes of customer arrivals and service completions are replaced by continuous processes. Such fluid models are also applicable to approximate the performance of stochastic systems, see for example Mandelbaum and Massey (1995). If the system is clearly under- or overloaded, i.e., the arrival rate is significantly smaller or greater than the maximum service rate, fluid approximations work well, see for example Mandelbaum et al. (1999) or Gans et al. (2003). Fluid approximations predict that queues do not begin to form until the traffic intensity exceeds one. Therefore fluid approximations often fail if the system operates near the critical load, see for example Altman et al. (2001) and Jimenez and Koole (2004).

Ingolfsson et al. (2007) give a literature review and compare the performance of different approximation methods in a comprehensive experimental study. In addition to some of the above-mentioned methods they study the closure approximation (see for example Rothkopf and Oren (1979) or Taaffe and Ong (1987)), the infinite-server approximations (see Jennings et al., 1996; Massey and Whitt, 1997), and the randomization method (see Grassmann, 1977). As a result, they found that the numerical solution of the ODE and the randomization method give accurate results but are numerically demanding. The infinite-server approximations, the effective arrival rate approximation, and the lagged SIPP approach were less accurate but computationally fast.

This overview shows that common approaches with stationary queueing models have restrictions for the application to temporarily overloaded systems. According to numerical results of Ingolfsson et al. (2007), the lagged SIPP approach is one of the preferable computationally fast approximation methods. In the numerical experiments in Section 4 this approximation is therefore used as a benchmark when underloaded systems are concerned. Simulation studies are used to judge the accuracy of both the lagged SIPP and the SBC approximations.

3. SBC approximation of the $M(t)/M(t)/c(t)$ queue

3.1. Main idea

This section describes the stationary backlog-carryover approach for the $M(t)/M(t)/c(t)$ queueing system. As in the SIPP approach the time interval is divided into small periods. The SBC approximation can be divided into two steps. In the first step the expected utilization is approximated. A loss model is applied for each period and backlog is generated and carried over into a subsequent period (Section 3.2). In the second step different methods for the approximation of the expected queue length and the expected number of customers in the system are discussed, see Section 3.3. Section 3.4 shows some limiting results for the case of constant rates and a constant number of servers and discusses the sensitivity of the period length of the SBC approach to the approximation results.

3.2. Approximation of the expected utilization

For the stationary backlog-carryover approach, the entire time interval is divided into T small periods $(t_{i-1}, t_i]$ for $i = 1, \ldots, T$. The number of servers $c(t)$ is a discontinuous piecewise constant function. In our examples it is assumed that the length of the periods can be chosen so that the number of servers c_i is constant during each single period i, i.e., $c_i = c(t)$ holds for $t \in (t_{i-1}, t_i]$. The original arrival rate function $\lambda(t)$ is replaced by a piecewise constant function. Green et al. (2001) discuss different methods to set the constant arrival rate for the SIPP approach. For the SBC approach the average arrival rate during period i is used, i.e., we set

$$\lambda_i = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \lambda(s)\,\mathrm{d}s \quad \text{for } i = 1, \ldots, T. \tag{8}$$

Constant processing rates μ_i can be generated in a similar way through

$$\mu_i = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \mu(s)\,\mathrm{d}s \quad \text{for } i = 1,\ldots,T. \tag{9}$$

For the lagged SIPP approach the planning period is shifted one average service time into the past, i.e, the arrival rate for period i is

$$\lambda_i^{\text{lag}} = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}-\mu^{-1}}^{t_i-\mu^{-1}} \lambda(s)\,\mathrm{d}s \quad \text{for } i = 1,\ldots,T. \tag{10}$$

This results in a constant arrival rate λ_i, a constant processing rate μ_i, and a constant number of servers c_i in each period i. The SIPP approach applies the $M/M/c_i/\infty$ waiting system with rates λ_i and μ_i in each period i and the periods are analyzed independently of each other. Applying this approach, a large expected queue length in period i does not influence the performance in following periods $j > i$.

Contrary to the SIPP approach, the SBC approximation analyzes each period i with an $M/M/c_i/c_i$ Erlang-loss system instead of the delay system. The loss model is applied with an artificial arrival rate $\tilde{\lambda}_i$. This artificial arrival rate $\tilde{\lambda}_i$ consists of the average original arrival rate λ_i and the backlog rate b_{i-1} of the former period $i-1$. The rate b_{i-1} is the steady-state rate of customers leaving the $M/M/c_{i-1}/c_{i-1}$ system due to blocking in the former period, i.e., the rate b_{i-1} defines the *backlog generated through artificial blocking* in period $i-1$. Let $P_{i-1}(B)$ be the steady-state probability of blocking for the $M/M/c_{i-1}/c_{i-1}$ model in period $i-1$ with arrival rate $\tilde{\lambda}_{i-1}$. The backlog rate b_{i-1} is given by

$$b_{i-1} = \tilde{\lambda}_{i-1} \cdot P_{i-1}(B). \tag{11}$$

These blocked customers of period $i-1$ are *carried over* into period i as additional artificial arrivals which results in the artificial arrival rate $\tilde{\lambda}_i$. Starting with $\tilde{\lambda}_1 = \lambda_1$ and $b_0 = 0$ for the first period, we get these artificial arrival rates recursively through

$$\tilde{\lambda}_i = \lambda_i + b_{i-1} = \lambda_i + \tilde{\lambda}_{i-1} \cdot P_{i-1}(B). \tag{12}$$

Using this method of carryover, we evenly spread customers not served in period $i-1$ over period i. Because of this artificial increase of the arrival rates, waiting customers may "arrive" again and again. Applying Erlang's loss formula for $P_{i-1}(B)$ (see for example Gross and Harris, 1998), the SBC approximation results directly in the expected utilizations $E[U_i]$ and the backlog rates b_i for each period i with

$$b_i = \tilde{\lambda}_i \cdot P_i(B) = \tilde{\lambda}_i \cdot \frac{(\tilde{\lambda}_i/\mu_i)^{c_i}}{c_i! \sum_{k=0}^{c_i} \frac{(\tilde{\lambda}_i/\mu_i)^k}{k!}} \quad \text{and} \tag{13}$$

$$E[U_i] = \frac{\tilde{\lambda}_i(1 - P_i(B))}{c_i\mu_i} = \frac{\tilde{\lambda}_i - b_i}{c_i\mu_i} = \frac{\lambda_i + b_{i-1} - b_i}{c_i\mu_i}. \tag{14}$$

The expected utilization $E[U(t_i)]$ of the original system at time t_i is approximated by this steady-state utilization $E[U_i]$ of the ith period.

3.3. Approximation of the time-dependent expected number of customers in the system and the expected queue length

3.3.1. Modified arrival rate approximation

The loss model presented in Section 3.2 cannot be used to approximate expected queue lengths or the expected number of customers in the system. The basic idea of our modified arrival rate (MAR) approximation is to analyze in a second step an $M/M/c_i/\infty$ waiting model with the same utilization as the loss model analyzed in the first step. To do so, for each period i a modified arrival rate λ_i^{MAR} is chosen so that the approximated utilization $E[U_i]$ of the first step would be reached in the $M/M/c_i/\infty$ model in the second step. The $M/M/c_i/\infty$ model can then be applied with this modified arrival rate to derive different performance measures for each period i.

The modified arrival rate λ_i^{MAR} can be derived as follows. In Section 3.2 the processing rate μ_i, and the number c_i of servers and approximated expected utilization $E[U_i]$ are derived for each period i. Now consider a stationary $M/M/c_i/\infty$ model with the approximated expected utilization $E[U_i]$. To reach this steady-state utilization $E[U_i]$ for given parameters c_i and μ_i the modified arrival rate λ_i^{MAR} must be

$$\lambda_i^{\text{MAR}} = E[U_i]c_i\mu_i. \tag{15}$$

Applying (14) results in

$$\lambda_i^{\text{MAR}} = \lambda_i + b_{i-1} - b_i. \tag{16}$$

Different steady-state performance measures can be calculated from the probability distribution of the $M/M/c_i/\infty$ model with the modified arrival rate λ_i^{MAR}. These are, for example, the expected queue

length $E^{\mathrm{MAR}}[L_i]$, the expected number of customers in the system $E^{\mathrm{MAR}}[N_i]$, or the probability of delay $P^{\mathrm{MAR}}(W_i > 0)$. We approximate the time-dependent performance measures of the original system at time t_i with these values, i.e., set $E[L(t_i)] = E^{\mathrm{MAR}}[L_i]$, $E[N(t_i)] = E^{\mathrm{MAR}}[N_i]$, and $P(W(t_i) > 0) = P^{\mathrm{MAR}}(W_i > 0)$, respectively.

The idea of our above-mentioned modified arrival rate (MAR) approach is similar to the modified offered load (MOL) approach, see Ingolfsson et al. (2007). Contrary to the MOL approach, our MAR approach uses the approximated expected utilization of a loss model instead of the expected number of busy servers from an infinite-server approximation.

3.3.2. Approximations based on $E[U_i]$ and b_i

A simple approach is the approximation of the expected queue length with the number of backlogged customers during a period (approximation $A1$). If the backlog rate b_i is multiplied by the period length $t_i - t_{i-1}$, this value can be seen as the number of waiting customers at the end of period i. In a first approach let the approximated expected queue length at the end of the ith period be

$$E^{A1}[L_i] = b_i(t_i - t_{i-1}). \tag{17}$$

The expected number of customers in the system is approximated through

$$E^{A1}[N_i] = E^{A1}[L_i] + c_i E[U_i]. \tag{18}$$

Numerical experiments show that the expected queue length is often overestimated through the approximation $E^{A1}[L_i]$. A better approximation $A2$ is given if the approximated queue length $E^{A1}[L_i]$ is reduced by the expected number of non-busy servers $c_i(1 - E[U_i])$, i.e., the approximation $A2$ is

$$E^{A2}[L_i] = \max\{0, b_i(t_i - t_{i-1}) - c_i(1 - E[U_i])\}. \tag{19}$$

The expected number of customers in the system is approximated by

$$E^{A2}[N_i] = E^{A2}[L_i] + c_i E[U_i]. \tag{20}$$

Both approximations $A1$ and $A2$ can be seen as simple fluid approximations for each period with different initial conditions, incoming flow rates, and processing rates. We also tested other inner-period fluid approximations, but in almost all numerical experiments the approximations described above performed best. That is why the numerical analysis concentrates on the approximations MAR, $A1$, and $A2$.

3.4. Limiting results

A special case of time-dependent analysis is to analyze the behavior of queueing models with constant rates. If the traffic intensity $\rho = \frac{\lambda}{c\mu}$ is smaller than one, these systems reach a steady-state as time approaches infinity. This section shows that the SBC approximation of $M/M/c$ queues with a constant arrival rate λ, a constant processing rate μ, and a constant number of servers c approaches these steady-state values if the stability condition $\lambda < c\mu$ is satisfied. This limiting result of the approximated performance does not depend on the length $l = t_i - t_{i-1}$ of the periods.

Let $E[U] = \frac{\lambda}{c\mu}$ be the expected utilization of the original $M/M/c$ model in steady-state. Firstly, we show that the approximated expected utilization $E[U_i]$ approaches this steady-state utilization $E[U]$, i.e., that

$$\lim_{i\to\infty} E[U_i] = E[U] \tag{21}$$

is valid. To prove the limit (21) the following properties of the SBC approximation for an $M/M/c$ model with constant arrival rate λ and without initial backlog ($b_0 = 0$) are useful:

(i) The artificial arrival rate $\tilde{\lambda}_i$ strictly monotonically increases in i. The proof is by induction on the number i of periods. The initial condition for the first two periods $\tilde{\lambda}_1 = \lambda$ and $\tilde{\lambda}_2 = \lambda + b_1$ clearly satisfies the statement $\tilde{\lambda}_1 < \tilde{\lambda}_2$. Because of this relation of the artificial arrival rates, the probability of blocking has a monotone increase, i.e., the relation $P_1(B) < P_2(B)$ holds. Assume that the statements are also true for period $i-1$, i.e., the conditions $\tilde{\lambda}_{i-2} < \tilde{\lambda}_{i-1}$ and therefore $P_{i-2}(B) < P_{i-1}(B)$ hold. Now consider period i with

$$\tilde{\lambda}_i = \lambda + b_{i-1} \tag{22}$$

$$= \lambda + \tilde{\lambda}_{i-1} \cdot P_{i-1}(B) \tag{23}$$

$$> \lambda + \tilde{\lambda}_{i-2} \cdot P_{i-2}(B) = \tilde{\lambda}_{i-1}. \tag{24}$$

Therefore, the artificial arrival rate $\tilde{\lambda}_i$ and the probability of blocking $P_i(B)$ monotonically increase in i. That is why the backlog rate b_i and the expected utilization $E[U_i]$ show a monotone increase as well.

(ii) The artificial arrival rate $\tilde{\lambda}_i$ is bounded. To show an upper bound Eq. (14) is applied with constant rates, i.e., we derive

$$E[U_i] = \frac{\lambda + b_{i-1} - b_i}{c\mu} = E[U] + \frac{b_{i-1} - b_i}{c\mu}. \tag{25}$$

From the monotone increase of b_i (proven in (i)) follows that $\frac{b_{i-1}-b_i}{c\mu}$ is negative and the relation $E[U_i] < E[U]$ holds. Therefore, the artificial arrival rate $\tilde{\lambda}_i$ must be bounded. The artificial arrival rate $\tilde{\lambda}_i$ is monotonically increasing and bounded. Therefore, the artificial arrival rate $\tilde{\lambda}_i$ and the backlog rate b_i converge if i approaches infinity.

The approximated expected utilization $E[U_i]$ is monotonically increasing and bounded. Therefore, $E[U_i]$ converges and from Eq. (25) follows:

$$\lim_{i\to\infty} E[U_i] = E[U] + \lim_{i\to\infty} \frac{b_{i-1} - b_i}{c\mu} = E[U], \tag{26}$$

i.e., the approximated utilization $E[U_i]$ approaches the steady-state utilization $E[U]$.

If the expected queue length and other performance measures are approximated using our MAR approach of Section 3.3.1, these measures converge to the steady-state measures of the original system as well.

Now consider a system with an initial backlog $b_0 > 0$, a constant arrival rate λ, a constant service rate μ, and a constant number of servers c. The system reaches a steady state if the stability condition $\lambda < c\mu$ holds. The steady-state performance measures are independent of the amount of initial backlog b_0. To show that the SBC approximation reaches this steady state we distinguish two cases with an initial backlog $b_0 > 0$. In the first case assume that the artificial arrival rate $\tilde{\lambda}_1$ of the first period is smaller than the artificial arrival rate $\tilde{\lambda}_2$ of the second period, i.e. $\lambda + b_0 < \lambda + b_1$ holds. The above prove of the limit (21) is valid for this case. In the second case we assume that $\lambda + b_0 \geqslant \lambda + b_1$ holds. The monotone decrease of the artificial arrival rate $\tilde{\lambda}_i$ can be shown using similar arguments as in part (i) of the proof above. The artificial arrival rate $\tilde{\lambda}_i$ has a lower bound of zero. Therefore, the artificial arrival rate $\tilde{\lambda}_i$ and the backlog rate b_i converge if i approaches infinity and limit (21) is valid for systems with an initial backlog.

This analysis shows that the SBC approximation would reach steady-state values if the arrival rate, the service rate, and the number of servers remain constant for long time intervals. These steady-state values are independent of the amount of initial backlog in the first period.

4. Numerical examples

4.1. Impact of the period length

This section discusses how the accuracy of the SBC approach depends on the chosen period length l. In the case of constant arrival and service rates for the whole time horizon, the performance measures of the ith period are independent of the length $l = t_i - t_{i-1}$, but the performance measures depend on the number i of the period. For different period lengths l the time t may fall into different periods. Therefore, applying the SBC approach with different period lengths l leads to different approximations of the performance measures for time t. This is important especially during the initial transient phase. As shown in Section 3.4, the artificial arrival rate and the approximated expected utilization monotonically increase in the number of periods. Therefore, for small period lengths l, the queues build up faster than for a large one. Therefore, the approximated expected utilization and the expected number of customers in the system at time t increase while the period length decreases.

The following examples demonstrate this numerically. We assume a system with $c = 32$ servers, each operating with an average processing time of $\mu^{-1} = 240$ seconds. An arrival rate of $\lambda = 7.8$ jobs per minute leads to a traffic intensity of $\rho = 0.975$. Fig. 1 compares the simulated average utilization with the approximated expected utilization $E^{\mathrm{SBC}}[U(t)]$ for three different period length

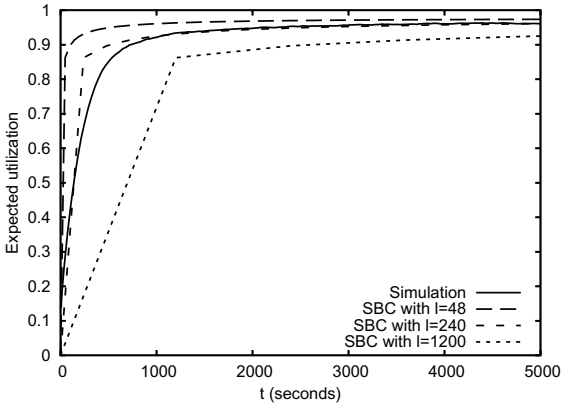

Fig. 1. SBC approximations of the expected utilization with different period lengths l.

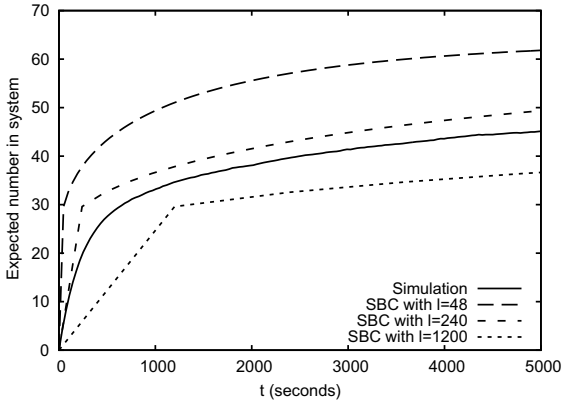

Fig. 2. SBC approximations of the expected number in system with different period lengths l.

$l = 0.2 \cdot \mu^{-1} = 48$, $l = \mu^{-1} = 240$, and $l = 5 \cdot \mu^{-1} = 1200$ seconds.[3] The SBC approximation with the small period length of $l = 48$ seconds reaches the steady-state value of $E[U] = 97.5\%$ faster than other approximations. Therefore, the approximated expected number of customers in the system $E^{\mathrm{SBC}}[N(t)]$ is higher for small period lengths, as shown in Fig. 2.

To find out an appropriate period length l for the SBC approximation we analyze the above system with different constant arrival rates λ between 5.6 and 7.92 arrivals per minute. This leads to traffic intensities ρ between 0.7 and 0.99. The results of the SBC approximation with different period lengths l are compared to the simulated average utilization for the first 9600 seconds. To calculate the reliability of the SBC approximation for a particular performance measure, the absolute deviation of the simulated and approximated values are computed in each period i. Averaged over all periods, this leads to the mean absolute deviation Δ^{abs}. A division by the mean simulated value results in the mean relative deviation Δ^{rel}.

Table 1 compares the deviation of the simulated and approximated utilization for SBC approximations with period lengths l between 120 and 420 seconds. For each traffic intensity ρ the second column gives the arrival rate λ per minute. The mean relative deviation is given in percent and the absolute deviation in brackets is given in percentage points.

[3] The simulation results are based on 4000 replications. This large number ensures that the standard deviation of the number of jobs in the system is smaller than 10^{-6} in all considered examples in this paper.

As the traffic intensity ρ increases the best found period length l decreases. To analyze the SBC approximation case-independent a period length l has to be chosen which works well in all examples. Because absolute deviations are higher in systems with high traffic intensities, we choose a period length l equal to the average processing time μ^{-1}. In all considered examples in Table 1 the simulated average utilization is well approximated by $E^{\mathrm{SBC}}[U(t)]$ using a period length $l = \mu^{-1}$. Assuming a period length of $l = \mu^{-1}$, Table 2 shows the resulting deviations of the expected number of customers in system and the expected queue length for the three approximations $A1$, $A2$, and MAR. These approximations work well, while the approximation error increases as the traffic intensity increases. In all cases the MAR approximation has the smallest deviation. Although the relative deviations are fairly large, the examples in the next sections demonstrate that the SBC approximation gives better results than the lagged SIPP approach.

The above examples indicate that the average processing time gives an appropriate period length l for the SBC approximation. In the next sections the SBC approximation with this period length is applied for systems with different numbers of servers, different mean processing times, and different traffic intensities. The intent of these sections is to demonstrate that the SBC approach is a general approximation method. Therefore, the period length is not varied in the numerical examples, even though for the most examples the SBC approximations can be improved by calibrating the period length l for each special case. To determine general rules for setting the period length better than $l = \mu^{-1}$ could be a fruitful task for future research.

4.2. Underloaded systems with a constant number of servers

In Section 4.2 the SBC approach is studied for $M(t)/M/c$ systems with time-varying but piecewise constant arrival rates and a constant number of servers in underloaded situations. Examples with time-dependent numbers of servers and temporal overload are studied in Section 4.3. In the first set of examples in this section (Table 3) we show how the approximation reacts to changes in traffic intensity. We then study these systems by varying the number of servers or the mean processing time.

We assume a time interval of 9600 seconds, which is divided into four phases. The arrival rate

Table 1
Relative and absolute deviations (in parentheses) of the approximated expected utilization $\Delta_{E^{SBC}[U]}$ from simulation for different period lengths l ($c = 32$ servers, mean processing time $\mu^{-1} = 240$ seconds, λ given in arrivals per minute)

ρ	λ	$l = 120$	$l = 180$	$l = 240$	$l = 300$	$l = 360$	$l = 420$
0.7	5.6	1.9% (1.3)	1.7% (1.1)	1.4% (1.0)	1.3% (0.9)	**1.2**% (0.8)	1.3% (0.9)
0.8	6.4	1.8% (1.4)	1.6% (1.2)	1.3% (1.0)	1.1% (0.9)	**1.1**% **(0.8)**	1.1% (0.9)
0.9	7.2	1.8% (1.6)	1.4% (1.3)	1.1% (1.0)	**0.8**% **(0.7)**	**0.8**% **(0.7)**	0.9% (0.8)
0.925	7.4	1.8% (1.6)	1.3% (1.2)	0.9% (0.8)	**0.8**% **(0.7)**	0.9% (0.8)	1.2% (1.1)
0.95	7.6	1.9% (1.7)	1.3% (1.2)	**0.8**% **(0.7)**	**0.8**% **(0.7)**	1.1% (1.0)	1.5% (1.4)
0.975	7.8	1.9% (1.8)	1.1% (1.1)	**0.8**% **(0.8)**	1.0% (1.0)	1.4% (1.3)	1.9% (1.8)
0.99	7.92	1.6% (1.6)	**1.0**% **(0.9)**	1.0% (1.0)	1.3% (1.2)	1.7% (1.6)	2.3% (2.1)

Table 2
Relative and absolute deviations (in parentheses) of the expected number of customers in system and in queue for $l = \mu^{-1}$ ($c = 32$ servers, mean processing time $\mu^{-1} = 240$ seconds)

ρ	$\Delta_{E^{A1}[N]}$	$\Delta_{E^{A2}[N]}$	$\Delta_{E^{MAR}[N]}$	$\Delta_{E^{A1}[L]}$	$\Delta_{E^{A2}[L]}$	$\Delta_{E^{MAR}[L]}$
0.7	2.3% (0.5)	1.6% (0.3)	**1.5**% **(0.3)**	187.1% (0.2)	100.0% (1.0)	**38.1**% **(0.0)**
0.8	4.5% (1.2)	3.2% (0.8)	**1.6**% **(0.4)**	144.1% (0.8)	100.0% (0.6)	**14.5**% **(0.1)**
0.9	9.4% (2.9)	2.4% (0.8)	**2.6**% **(0.8)**	80.5% (2.7)	17.0% (0.6)	**15.1**% **(0.5)**
0.925	11.3% (3.8)	3.8% (1.3)	**3.4**% **(1.2)**	68.7% (3.6)	19.7% (1.0)	**17.1**% **(0.9)**
0.95	14.3% (5.4)	9.1% (3.4)	**5.7**% **(2.1)**	62.6% (5.1)	38.9% (3.2)	**23.0**% **(1.9)**
0.975	18.8% (8.0)	15.3% (6.5)	**9.9**% **(4.2)**	62.5% (7.8)	50.6% (6.3)	**32.5**% **(4.1)**
0.99	19.7% (9.2)	17.0% (8.0)	**11.1**% **(5.2)**	55.5% (9.2)	47.8% (7.9)	**31.3**% **(5.2)**

is assumed to be constant in each phase. We consider a system with $c = 32$ servers and a mean processing time of $\mu^{-1} = 240$ seconds. Such parameters can be found, for example, in call centers with patient customers. To show how the approximation works for changing traffic intensities, these examples assume that the traffic intensity is higher in one phase than in the remaining phases, similar to the numerical experiments in Jimenez and Koole (2004). This high traffic intensity occurs in one of the first three phases. The traffic intensities $\rho(t)$ for the four phases are summarized in Table 3. In each example the arrival rates are adjusted to meet the different traffic intensities. For instance, in Example 5 the arrival rate is $\lambda(t) = 7.2$ customers per minute ($\rho(t) = 0.9$) for the first 40 minutes (Phase I). In the next 40 minutes the arrival rate increases by about 10%, i.e., for $\lambda(t) = 7.92$ holds $\rho(t) = 0.99$ (Phase II). For $4800 < t \leqslant 9600$ we again assume the lower arrival rate of $\lambda(t) = 7.2$ customers per minute (Phases III and IV).

Table 3
Instantaneous traffic intensities $\rho(t)$ for different examples with $c = 32$ servers and a mean processing time of $\mu^{-1} = 240$ seconds

Ex.	Phase I $0 \leqslant t \leqslant 2400$	Phase II $2400 < t \leqslant 4800$	Phase III $4800 < t \leqslant 7200$	Phase IV $7200 < t \leqslant 9600$
1	0.95	0.8	0.8	0.8
2	0.8	0.95	0.8	0.8
3	0.8	0.8	0.95	0.8
4	0.99	0.9	0.9	0.9
5	0.9	0.99	0.9	0.9
6	0.9	0.9	0.99	0.9
7	0.99	0.95	0.95	0.95
8	0.95	0.99	0.95	0.95
9	0.95	0.95	0.99	0.95

Tables 4 and 5 present the relative deviations of the expected utilization, the expected number of customers in the system, and the expected queue length for the three SBC approximations. The average absolute deviations are given in brackets. To compare our results to the SIPP approach we initially evaluated different versions of this approach. For period lengths of 20 and 40 minutes the original SIPP approach and the lagged SIPP approach were tested. In the lagged SIPP approach a lag of one average service time (4 minutes) is used. For the examples presented in this section the lagged-average SIPP approach with a period length of 20 minutes gives the best results compared to the simulation study. This lagged SIPP approach is therefore used as a benchmark in the following examples. The deviations of the expected utilization, the expected number of customers in the system,

Table 4
Relative and absolute deviations (in parentheses) of the expected utilization and expected number of customers in system for the SBC approximations and the lagged SIPP approximation (Examples of Table 3)

Ex.	$\Delta_{E^{SBC}[U]}$	$\Delta_{E^{SIPP}[U]}$	$\Delta_{E^{A1}[N]}$	$\Delta_{E^{A2}[N]}$	$\Delta_{E^{MAR}[N]}$	$\Delta_{E^{SIPP}[N]}$
1	**1.4**% **(1.1)**	2.5% (2.0)	8.2% (2.3)	5.2% (1.4)	**3.7**% **(1.0)**	6.5% (1.8)
2	**1.6**% **(1.3)**	2.8% (2.3)	7.1% (2.0)	4.6% (1.3)	**3.0**% **(0.8)**	7.8% (2.2)
3	**1.7**% **(1.4)**	2.5% (2.1)	7.0% (2.0)	4.3% (1.2)	**2.8**% **(0.8)**	7.3% (2.1)
4	**1.4**% **(1.2)**	2.3% (2.0)	13.9% (4.7)	7.9% (2.7)	**5.8**% **(2.0)**	34.5% (11.6)
5	**1.6**% **(1.5)**	3.0% (2.7)	12.0% (4.2)	6.3%(2.2)	**4.8**% **(1.7)**	42.5% (14.8)
6	**1.6**% **(1.4)**	3.0% (2.7)	12.1% (4.2)	5.4% (1.9)	**4.2**% **(1.5)**	42.0% (14.5)
7	**0.9**% **(0.8)**	2.1% (2.0)	16.6% (6.6)	12.4% (4.9)	**7.7**% **(3.0)**	32.4% (12.8)
8	**1.1**% **(1.1)**	2.5% (2.4)	16.4% (6.6)	12.2% (4.9)	**7.7**% **(3.1)**	41.6% (16.7)
9	**1.1**% **(1.0)**	2.7% (2.5)	16.5% (6.5)	12.1% (4.8)	**7.8**% **(3.1)**	44.2% (17.5)

Table 5
Relative and absolute deviations (in parentheses) of the expected queue length for the SBC approximations and the lagged SIPP approximation (Examples of Table 3)

Ex.	$\Delta_{E^{A1}[L]}$	$\Delta_{E^{A2}[L]}$	$\Delta_{E^{MAR}[L]}$	$\Delta_{E^{SIPP}[L]}$
1	117.5% (1.9)	74.2% (1.2)	**44.0**% **(0.7)**	75.9% (1.3)
2	82.2% (1.7)	51.8% (1.0)	**27.7**% **(0.6)**	75.5% (1.5)
3	82.3% (1.7)	49.2% (1.0)	**25.1**% **(0.5)**	71.9% (1.5)
4	83.7% (4.3)	46.1% (2.4)	**31.1**% **(1.6)**	214.9% (11.1)
5	61.3% (3.9)	31.4% (2.0)	**22.9**% **(1.4)**	222.8% (14.0)
6	64.4% (3.9)	27.8% (1.7)	**19.7**% **(1.2)**	227.3% (13.7)
7	64.1% (6.4)	47.2% (4.7)	**28.5**% **(2.8)**	124.0% (12.3)
8	60.0% (6.4)	44.1% (4.7)	**26.8**% **(2.9)**	151.0% (16.0)
9	63.0% (6.3)	45.5% (4.6)	**28.6**% **(2.9)**	167.0% (16.7)

and the expected queue length for this lagged SIPP approach are given as well.

Table 4 shows that the relative deviations $\Delta^{rel}_{E^{SBC}[U]}$ are small in all nine examples. The expected utilizations of the agents are well approximated in all examples shown. The lagged SIPP approach results in relative deviations $\Delta^{rel}_{E^{SIPP}[U]}$ of the expected utilization between 2.1% and 3.0%, which are significantly higher than the deviations with the SBC approximation. The SBC approximations A2 and MAR of the expected number of customers in the system are better than the lagged SIPP approximation $E^{SIPP}[N(t)]$. The approximation A1 is better than the lagged SIPP approach in 8 out of 9 cases. In all examples of Table 3, our MAR approximations have the smallest deviations.

Fig. 3 shows the approximated number of customers in the system for Example 5. The MAR approach is better than other SBC approximations and significantly better than the lagged SIPP approximation. The deviation from the simulated average number of customers in the system is only $\Delta^{rel}_{E^{MAR}[N]} = 4.8\%$ ($\Delta^{abs}_{E^{MAR}[N]} = 1.7$ customers). The SBC approach also gives a better approximation

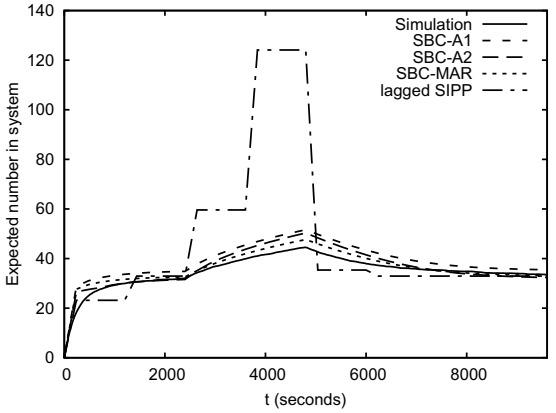

Fig. 3. Approximations of the expected number of customers in the system (SBC and lagged SIPP) vs. simulated average number of customers in the system for Example 5.

Table 6
Relative and absolute deviations (in parentheses) of the expected utilization and expected number of customers in the system for the SBC approximations and the lagged SIPP approximation for different numbers of servers and different processing rates

Ex.	c	μ^{-1}	$\Delta_{E^{SBC}[U]}$	$\Delta_{E^{SIPP}[U]}$	$\Delta_{E^{A2}[N]}$	$\Delta_{E^{MAR}[N]}$	$\Delta_{E^{SIPP}[N]}$
3	16	240	**3.1**% **(2.5)**	3.6% (2.9)	8.0% (1.2)	**3.8**% **(0.6)**	15.6% (2.4)
6	16	240	**2.1**% **(1.9)**	3.7% (3.2)	9.6% (1.9)	**7.0**% **(1.4)**	75.6% (15.1)
9	16	240	**1.5**% **(1.4)**	4.1% (3.7)	16.7% (3.9)	**11.5**% **(2.7)**	85.4% (19.8)
3	64	240	**1.4**% **(1.2)**	2.5% (2.0)	2.3% (1.3)	**2.2**% **(1.2)**	3.8% (2.0)
6	64	240	**1.6**% **(1.4)**	2.6% (2.3)	5.4% (3.5)	**2.7**% **(1.7)**	21.8% (13.8)
9	64	240	**1.2**% **(1.1)**	2.5% (2.4)	7.3% (5.2)	**4.7**% **(3.3)**	23.5% (16.6)
3	32	120	**1.1**% **(0.9)**	1.8% (1.5)	4.6% (1.3)	**2.4**% **(0.7)**	6.2% (1.8)
6	32	120	**1.1**% **(1.0)**	2.2% (2.0)	6.7% (2.4)	**4.2**% **(1.5)**	43.4% (15.9)
9	32	120	**0.8**% **(0.7)**	1.8% (1.7)	11.4% (4.9)	**6.3**% **(2.8)**	41.9% (18.2)
3	32	480	**4.5**% **(3.5)**	5.3% (4.1)	4.3% (1.2)	**4.0**% **(1.1)**	9.4% (2.5)
6	32	480	**3.2**% **(2.7)**	5.5% (4.7)	**5.3**% **(1.7)**	5.6% (1.8)	48.7% (15.7)
9	32	480	**2.6**% **(2.3)**	6.2% (5.5)	11.8% (4.2)	**9.0**% **(3.2)**	59.3% (20.9)

of the number of customers in the queue than the lagged SIPP approach. With a relative deviation of $\Delta^{rel}_{E^{MAR}[L]} = 22.9\%$ and an average absolute deviation of only $\Delta^{abs}_{E^{MAR}[L]} = 1.4$ customers, the MAR approximation gives the best results for the approximation of the expected queue length. Especially in the second phase with the high instantaneous traffic intensity $\rho(t) = 0.99$, the approximation with the lagged SIPP approach results in a massive over-estimation of the expected queue length, which results in high deviations of $\Delta^{rel}_{E^{SIPP}[L]} = 222.8\%$ and $\Delta^{abs}_{E^{SIPP}[L]} = 14.0$ customers.

Table 5 shows the deviations of the approximated expected queue length to the simulated average queue length for the other examples of Table 3. Again, our MAR approach gives the best approximation and is significantly better than the lagged SIPP approach.

In the next examples we again analyze systems with traffic intensities as described in the Examples 3, 6, and 9 in Table 3, but the number of servers or the processing rate changes. Instead of $c = 32$ servers we study these examples with $c = 16$ and $c = 64$ servers. The mean processing time is assumed to be $\mu^{-1} = 240$ seconds in the first six examples and the arrival rates are adjusted to meet the instantaneous traffic intensities $\rho(t)$ of Table 3. The first parts of Tables 6 and 7 present the results for these examples.[4] The SBC approximation gives reasonable results for the approximation of the expected utilization of both numbers of servers. The best approximation of the expected number in system and in queue is given by the MAR approach. Similar to the examples with $c = 32$ servers, the lagged SIPP approach has a substantially higher deviation from the simulation results than our MAR approach for all reported performance measures.

The mean processing times vary in the examples of the second parts of Tables 6 and 7. With $\mu^{-1} = 120$ seconds, half of the original average processing time is assumed. Therefore, the period length for the SBC approach is only 120 seconds. The number of servers is assumed to be constant with $c = 32$ and the arrival rates are adjusted to meet the instantaneous traffic intensities $\rho(t)$ of Table 3. Table 6 shows that the expected utilization is well approximated for both mean processing times. In almost all cases the best approximation of the expected number of customers in the system is the MAR approach. Again, the relative deviations for the SBC approximations of the expected queue length are fairly large but significantly lower than those for the lagged SIPP approach, see Table 7.

All these examples show that underloaded systems with time-varying and piecewise constant arrival rates can be analyzed with the stationary backlog-carryover approach. The approximation of the expected utilization, the expected number of customers in the system, and the expected queue length works well. The best results are reached using the modified arrival rate (MAR) approximation. The MAR approximation of time-dependent per-

[4] The deviations for the approximation $A1$ are not summarized because these deviations are always higher than the deviations for the approximation $A2$.

Table 7
Relative and absolute deviations (in parentheses) of the expected queue lengths for the SBC approximations and the lagged SIPP approximation for different numbers of servers and different processing rates

Ex.	c	μ^{-1}	$\Delta_{E^{A2}[L]}$	$\Delta_{E^{MAR}[L]}$	$\Delta_{E^{SIPP}[L]}$
3	16	240	49.6% (1.3)	**18.2% (0.5)**	80.6% (2.1)
6	16	240	27.9% (1.7)	**19.1% (1.1)**	245.9% (14.6)
9	16	240	42.2% (3.7)	**28.1% (2.4)**	221.2% (19.2)
3	64	240	39.8% (0.7)	**33.4% (0.5)**	54.8% (0.9)
6	64	240	48.6% (2.9)	**19.3% (1.2)**	205.5% (12.4)
9	64	240	41.4% (4.6)	**24.9% (2.8)**	137.3% (15.2)
3	32	120	49.7% (1.2)	**23.1% (0.6)**	55.5% (1.3)
6	32	120	31.3% (2.4)	**19.0% (1.4)**	202.7% (15.3)
9	32	120	36.0% (4.8)	**19.6% (2.6)**	131.6% (17.7)
3	32	480	55.2% (1.1)	**26.2% (0.5)**	86.1% (1.7)
6	32	480	28.5% (1.4)	**20.1% (1.0)**	300.5% (14.3)
9	32	480	49.4% (3.5)	**35.1% (2.5)**	276.9% (19.4)

formance measures is significantly better than the approximation with the lagged SIPP approach.

4.3. Critically loaded and temporarily overloaded systems with time-dependent numbers of servers

This section studies the SBC approximation for a system with time-varying arrival rates and time-dependent number of servers as they may occur in service centers. Table 8 gives the piecewise constant arrival rates per hour and three configurations of the number of servers. The resulting traffic intensities $\rho(t)$ are given in brackets.

In all three configurations the mean service time is $\mu^{-1} = 240$ seconds. This results in an instantaneous traffic intensity of $\rho_{\mathrm{I}}(t) = 0.95$ at all times in Configuration I. Configurations II and III limit the maximum number of servers to 50 and 48 servers, respectively. The maximum traffic intensity of Configuration II is 0.988 and the system works near the critical load between 11:30 and 12:00. In Configuration III the system is overloaded at several times with a maximum traffic intensity of 1.029 and the (lagged) SIPP approach cannot be applied.

Even though the traffic intensity $\rho(t)$ is smaller than one for each t in Configuration I, the lagged SIPP approach with a period length of 20 minutes cannot be applied. Due to the varying number of servers, the resulting arrival rate λ^{lag} (see Eq. 10) may exceed the capacity $c_t \cdot \mu_t$ in some periods. For example, in the period 14:00-14:20 the arrival rate for the lagged SIPP is $\lambda^{\mathrm{lag}} = 0.2 \cdot 641.25 + 0.8 \cdot 498.75 = 527$ calls per hour. With $c = 35$ servers, the traffic intensity $\rho = \frac{\lambda^{\mathrm{lag}}}{c\mu} = 1.00429$ violates

Table 8
Arrival rates and numbers of servers for three configurations

Time t	λ_t [arrivals/hour]	$c_{\mathrm{I},t}$ ($\rho_{\mathrm{I}}(t)$)	$c_{\mathrm{II},t}$ ($\rho_{\mathrm{II}}(t)$)	$c_{\mathrm{III},t}$ ($\rho_{\mathrm{III}}(t)$)
8:00–8:40	114	8 (0.95)	8 (0.95)	8 (0.95)
8:40–9:20	142.5	10 (0.95)	10 (0.95)	10 (0.95)
9:20–10:00	142.5	10 (0.95)	10 (0.95)	10 (0.95)
10:00–10:40	185.25	13 (0.95)	13 (0.95)	13 (0.95)
10:40–11:20	356.25	25 (0.95)	25 (0.95)	25 (0.95)
11:20–12:00	641.25	45 (0.95)	45 (0.95)	45 (0.95)
12:00–12:40	712.5	50 (0.95)	50 (0.95)	48 (0.990)
12:40–13:20	741	52 (0.95)	50 (0.988)	48 (1.029)
13:20–14:00	641.25	45 (0.95)	45 (0.95)	45 (0.95)
14:00–14:40	498.75	35 (0.95)	35 (0.95)	35 (0.95)
14:40–15:20	712.5	50 (0.95)	50 (0.95)	48 (0.990)
15:20–16:00	498.75	35 (0.95)	35 (0.95)	35 (0.95)
16:00–16:40	427.5	30 (0.95)	30 (0.95)	30 (0.95)
16:40–17:20	427.5	30 (0.95)	30 (0.95)	30 (0.95)
17:20–18:00	356.25	25 (0.95)	25 (0.95)	25 (0.95)

the stability condition for the SIPP approach. Therefore, the lagged SIPP with a period length of 40 minutes was tested, but the approximation results are worse than those of the original (non-lagged) SIPP approach. This problem of the lagged SIPP caused by significant changes in the number of servers also arises in Configuration II. Therefore the non-lagged SIPP approach is used as a benchmark for the SBC approximations of Configuration I and II.

Tables 9 and 10 show the deviations from the simulation results for the expected utilization and the expected number of jobs in system and in queue. For the results of the upper parts of both tables, an exhaustive discipline is simulated if a server is scheduled to leave. The deviation in the lower parts are related to a pre-emptive discipline. The accuracy of the SBC approximation is higher for the pre-emptive end-of-shift policy in the presented examples.

Figs. 4–6 show the approximations of the expected number in system of the SBC and the SIPP approximations compared to the simulation results for the exhaustive and the pre-emptive discipline. The simulated average number in system for the pre-emptive discipline is greater than or equal to this number for an exhaustive discipline at any time. As other approaches with stationary models, our SBC approach do not differentiate between the two disciplines. A fruitful direction of future research could be to refine the SBC approach such that the end-of-shift policies are explicitly considered.

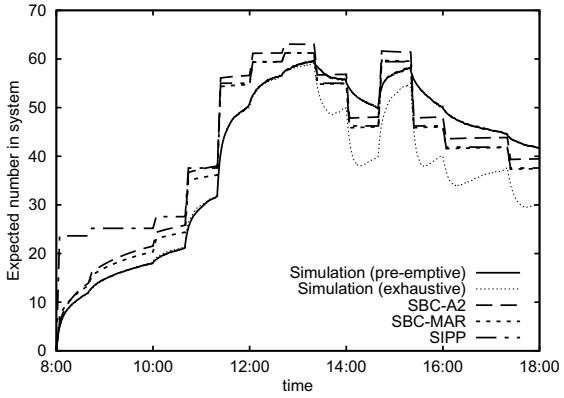

Fig. 4. Expected number of customers in the system for Configuration I.

These three examples show that the approximation error is fairly large for all approximation methods. However, the new SBC approximation gives smaller errors for the expected utilization than the SIPP approach in Configurations I and II. Again, the best approximation of the expected number of customers in the system and the expected queue length is found via the MAR approximation.

Table 9
Relative and absolute deviations (in parentheses) of the expected utilization and expected number of customers in system for the SBC and the SIPP approximations

Discipline	Configuration	$\Delta_{E^{SBC}[U]}$	$\Delta_{E^{SIPP}[U]}$	$\Delta_{E^{A2}[N]}$	$\Delta_{E^{MAR}[N]}$	$\Delta_{E^{SIPP}[N]}$
Exhaustive	I	**4.0**% **(3.7)**	5.2% (4.8)	18.3% (6.5)	**13.7**% **(4.9)**	20.1% (7.1)
Exhaustive	II	**4.1**% **(3.7)**	5.4% (4.9)	20.8% (7.5)	**15.6**% **(5.7)**	29.6% (10.7)
Exhaustive	III	**4.3**% **(4.0)**	–	34.6% (13.4)	**27.7**% **(10.7)**	–
Pre-emptive	I	**1.6**% **(1.5)**	2.6% (2.5)	10.2% (4.1)	**9.9**% **(4.0)**	15.1% (6.1)
Pre-emptive	II	**1.5**% **(1.4)**	2.8% (2.6)	10.9% (4.5)	**9.9**% **(4.1)**	25.6% (10.7)
Pre-emptive	III	**1.1**% **(1.0)**	–	12.4% (5.9)	**10.0**% **(4.8)**	–

Table 10
Relative and absolute deviations (in parentheses) of the expected queue length for the SBC and the SIPP approximations

Discipline	Configuration	$\Delta_{E^{A2}[L]}$	$\Delta_{E^{MAR}[L]}$	$\Delta_{E^{SIPP}[L]}$
Exhaustive	I	96.2% (6.4)	**71.7**% **(4.7)**	104.0% (6.9)
Exhaustive	II	101.7% (7.4)	**75.8**% **(5.5)**	143.0% (10.4)
Exhaustive	III	135.6% (13.3)	**108.4**% **(10.6)**	–
Pre-emptive	I	35.3% (3.9)	**34.4**% **(3.8)**	52.3% (5.8)
Pre-emptive	II	35.0% (4.4)	**31.7**% **(4.0)**	81.8% (10.3)
Pre-emptive	III	31.3% (5.8)	**25.5**% **(4.7)**	–

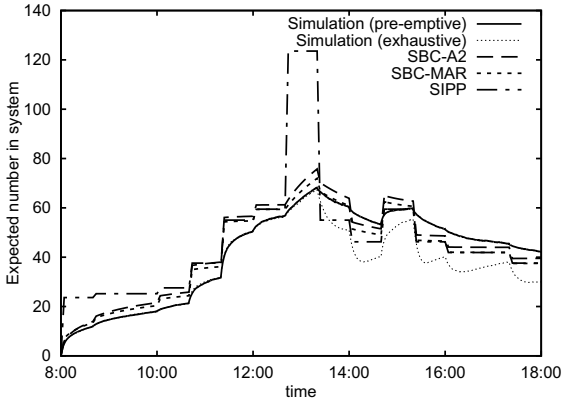

Fig. 5. Expected number of customers in the system for Configuration II.

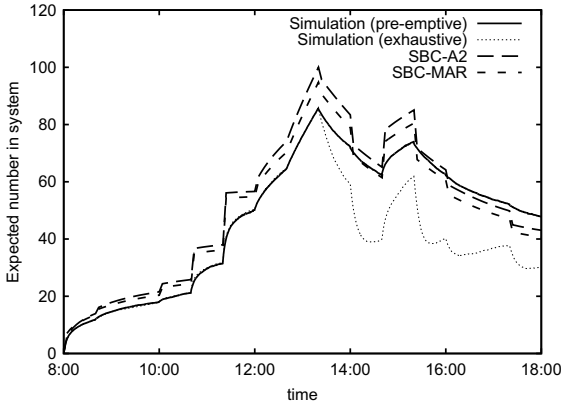

Fig. 6. Expected number of customers in the system for Configuration III.

5. Conclusion and suggestions for further research

We presented an approximation method for non-stationary queueing systems using a series of stationary queueing models. Consecutive periods are linked through the carryover of backlog. For the $M(t)/M(t)/c(t)$ model with piecewise constant arrival rates, the reliability of the stationary backlog-carryover (SBC) approach is demonstrated in several numerical experiments. The approximation of the expected utilization works accurately for underloaded as well as temporarily overloaded systems. The expected number of customers in the system and the expected queue length are well approximated by the modified arrival rate (MAR) approximation in almost all cases. For underloaded systems the SBC approach approximates the time-dependent expected utilization, the time-dependent number of customers in the system, and the time-dependent queue length significantly better than the lagged SIPP approach. For temporarily overloaded systems a competing approximation with stationary models is not available.

Future research could be directed towards more complex queueing systems. In general, in order to apply the SBC approach to a queueing system, it needs to be specified (i) which stationary queueing model is used in each single period, (ii) how the backlog can be measured, and (iii) how the backlog is carried over from one period to its successors. For example, the method could be extended to models with different classes of customers and heterogeneous servers or systems with impatient customers. To analyze systems with impatient customers and retrials an $M/M/c+M$ model (with or without retrials) could be used in the second step of the approximation. If the average time a customer waits before he retries is significantly longer than the period length l of the SBC approximation, these retrials generate additional backlog. Based on the distribution of the retrial time this additional backlog will be carried over into different future periods. Especially for the analysis of such queueing models it would be interesting to get time-dependent expected waiting times or service levels. The approximations of these waiting-based measures have to cope with time-varying numbers of servers.

The insensitivity of the $M/G/c/c$ model to the shape of the service time distribution suggests that the SBC approximation may work for general service time distributions as well as for exponential service time distribution. Additional numerical studies are necessary to explore the reliability of the approach for such queueing models.

A productive direction for further research would be integrating the SBC approach into optimization procedures, for example, agent staffing in call centers.

References

Altman, E., Jimenez, T., Koole, G., 2001. On the comparison of queueing systems with their fluid limits. Probability in the Engineering and Informational Sciences 15, 165–178.

de Souza e Silva, E., Gail, H.R., 2000. Transient solutions for Markov chains. In: Grassmann, W.K. (Ed.), Computational Probability. Kluwer Academic Publishers, Boston, pp. 43–79.

Gans, N., Koole, G., Mandelbaum, A., 2003. Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management 5 (2), 79–141.

Grassmann, W.K., 1977. Transient solutions in Markovian queueing systems. Computers & Operations Research 4 (1), 47–53.

Green, L., Kolesar, P., 1991. The pointwise stationary approximation for queues with nonstationary arrivals. Management Science 37 (1), 84–97.

Green, L., Kolesar, P., Svoronos, A., 1991. Some effects of nonstationarity on multi-server Markovian queueing systems. Operations Research 39 (3), 502–511.

Green, L.V., Kolesar, P.J., Soares, J., 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research 49 (4), 549–564.

Green, L.V., Kolesar, P.J., Soares, J., 2003. An improved heuristic for staffing telephone call centers with limited operating hours. Production and Operations Management 12 (1), 46–61.

Green, L.V., Kolesar, P.J., Whitt, W., 2007. Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management (POMS) 16 (1), 13–39.

Gross, D., Harris, C.M., 1998. Fundamentals of Queueing Theory, third ed. John Wiley & Sons., New York.

Ingolfsson, A., 2005. Modelling the $M(t)/M/c(t)$ queue with an exhaustive discipline. Technical Report, School of Business, University of Alberta, Canada.

Ingolfsson, A., Haque, M.A., Umnikov, A., 2002. Accounting for time-varying queueing effects in workforce scheduling. European Journal of Operational Research 139 (3), 585–597.

Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y., Wu, X., 2007. A survey and experimental comparison of service level approximation methods for non-stationary $M(t)/M/s(t)$ queueing systems. INFORMS Journal of Computing 19 (2), 201–214.

Jennings, O.B., Mandelbaum, A., Massey, W.A., Whitt, W., 1996. Server staffing to meet time-varying demand. Management Science 42 (10), 1383–1394.

Jimenez, T., Koole, G., 2004. Scaling and comparison of fluid limits of queues applied to call centers with time varying parameters. OR Spectrum 26 (3), 413–422.

Kleinrock, L., 1975. Queueing Systems Volume I: Theory. John Wiley & Sons, Inc., New York.

Kolesar, P.J., Rider, K.L., Crabill, T.B., Walker, W.E., 1975. A queuing-linear programming approach to scheduling police patrol cars. Operations Research 23 (6), 1045–1061.

Kwan, S.K., Davis, M.M., Greenwood, A.G., 1988. A simulation model for determining variable worker requirements in a service operation with time-dependent customer demand. Queueing Systems 3 (3), 265–276.

Lin, L., Cochran, J.K., 1990. Metamodels of production line transient behaviour for sudden machine breakdowns. International Journal of Production Research 28 (10), 1791–1806.

Mandelbaum, A., Massey, W.A., 1995. Strong approximations for time-dependent queues. Mathematics of Operations Research 20 (1), 33–64.

Mandelbaum, A., Massey, W., Reiman, M., Rider, B., 1999. Time varying multiserver queues with abandonment and retrials. Teletraffic Science and Engineering 3a, 355–364.

Massey, W.A., Whitt, W., 1997. Peak congestion in multi-server service systems with slowly varying arrival rates. Queueing Systems 25, 157–172.

Parthasarathy, P.R., Sharafali, M., 1989. Transient solution to the many-server Poisson queue: A simple approach. Journal of Applied Probability 26 (3), 584–594.

Rothkopf, M.H., Oren, S.S., 1979. A closure approximation for the nonstationary M/M/s queue. Management Science 25 (6), 522–534.

Stahlman, E.J., Cochran, J.K., 1998. Dynamic metamodelling in capacity planning. International Journal of Production Research 36 (1), 197–210.

Steckley, S.G., Henderson, S.G., 2007. The error in steady-state approximations for the time-dependent waiting time distribution. Stochastic Models 23 (2), 307–332.

Stewart, W.J., 1994. Introduction to the Numerical Solution of Markov Chains. Princeton University Press, Princeton, NJ.

Stolletz, R., 2003. Performance Analysis and Optimization of Inbound Call CentersLecture Notes in Economics and Mathematical Systems, vol. 528. Springer, Berlin.

Taaffe, M., Ong, K.L., 1987. Approximating nonstationary $Ph(t)/M(t)/s/c$ queueing systems. Annals of Operations Research 8, 103–116.

Testik, M.C., Cochran, J.K., Runger, G.C., 2004. Adaptive server staffing in the presence of time-varying arrivals: A feed-forward control approach. Journal of the Operational Research Society 55 (3), 233–239.

Thompson, G.M., 1993. Accounting for the multi-period impact of service when deter- mining employee requirements for labor scheduling. Journal of Operations Management 11 (3), 269–287.

Whitt, W., 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. Management Science 37 (3), 307–314.