UNIVERSITY OF MANNHEIM

Analysis of buffer allocations in time-dependent and stochastic flow lines

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Justus Arne Schwarz
Mannheim

Dekan: *Dr. Jürgen M. Schneider*
Referent: *Professor Dr. Raik Stolletz*
Korreferent: *Professor Dr. Moritz Fleischmann*

Tag der mündlichen Prüfung: *11. Dezember 2015*

To my grandmother Hildegard

Summary

Flow lines are production systems that can be found, e.g., in the food and automotive industries, for production at high volume. Modern information technology allows flexible and short-term changes of buffer capacities in flow lines by means of control mechanisms such as electronic Kanban cards. This thesis investigates the potential of time-dependent adjustments of buffer capacities to account for time-dependent changes in demand and machine characteristics. Performance evaluation approaches for given buffer allocations and algorithms to systematically derive time-dependent buffer allocations are developed. In contrast, the existing literature typically assumes constant buffer capacities and treats their allocation as a long-term design problem.

The first essay presents an overview of the existing literature on buffer capacity optimization in unreliable flow lines. All reviewed articles assume that flow lines operate under steady-state conditions. The second essay provides a survey and classification of performance evaluation approaches for queueing systems with time-dependent parameters and their applications. It is apparent that, even for single-stage systems with finite buffer capacities, there are no exact analytical solutions. The third essay introduces two sample-based approaches for the performance evaluation of flow lines with a time-dependent processing rate on the first machine and constant buffer capacities and constant rates for the subsequent machines. The equivalence of the two approaches has been established for a special case. The numerical study demonstrates that, given a time-dependent input, buffers can smooth the output over time. The fourth essay proposes a sample-based evaluation approach that accounts for time-dependent buffer capacities. The numerical study indicates the potential for influencing key performance measurements, such as work in process inventory (WIP) and throughput by time-dependent buffer allocations. The fifth essay reports monotonicity properties for the WIP and the service level of a flow line with respect to time-dependent buffer capacities based on a numerical study. A search algorithm utilizes these properties to find time-dependent buffer capacities. The numerical study indicates that the generated solutions lead to lower average WIP compared to constant buffer capacities while satisfying the same service level goal.

Future research should address the analysis of larger and more general systems. Their analysis may require the development of new evaluation and optimization methods.

Contents

Summary **V**

List of Figures **XI**

List of Tables **XIII**

1 Introduction **1**

2 Buffer Allocation Problems for stochastic flow lines with unreliable machines **7**
2.1 Introduction 8
2.2 Classification scheme for characteristics of flow lines 9
2.3 Classification scheme for decision problems 11
2.4 Conclusion and future research 14

3 Performance analysis of time-dependent queueing systems: Survey and classification **15**
3.1 Introduction 16
3.2 Scope and classification scheme 17
3.3 Performance evaluation approaches 20
3.3.1 Numerical and analytical solutions 20
3.3.2 Approaches based on models with piecewise constant parameters 27
3.3.2.1 Piecewise stationary models with independent periods 27
3.3.2.2 Piecewise stationary models with linked periods 31
3.3.2.3 Piecewise transient models 35
3.3.3 Approaches based on modified system characteristics 39
3.3.3.1 Modified number of servers 39
3.3.3.2 Modified job characteristics 41
3.4 Methodological relations and numerical comparisons 48

3.5 Areas of application 52
3.5.1 Service systems 52
3.5.2 Road and air traffic systems 55
3.5.3 IT systems 58
3.6 Conclusions and future research 59

4 Approximations of time-dependent unreliable flow lines with finite buffers 61
4.1 Introduction 62
4.2 Modeling of time-dependent flow lines 66
4.2.1 Model description 66
4.2.2 Performance measures 68
4.2.3 Sampling approach 68
4.3 MIP model for the evaluation of time-dependent flow lines . 69
4.4 Continuous model for the evaluation of time-dependent flow lines 72
4.4.1 Numerical solution of the continuous model 77
4.5 Link between MIP and continuous model 80
4.6 Numerical evaluation of the approximation approaches . . . 82
4.6.1 Performance measures 82
4.6.2 Case I: Increase of the release rate 83
4.6.3 Case II: Decrease of the release rate 85
4.6.4 Impact of the number of machines and buffer capacities 88
4.7 Conclusion 92

5 A sampling approach for the analysis of time-dependent stochastic flow lines 93
5.1 Introduction 94
5.2 Time-dependent stochastic flow lines 95
5.3 Evaluation by a discrete-time sampling approach 97
5.4 Numerical study 100
5.5 Conclusion and future research 106

6 A proactive approach to Kanban allocation in stochastic flow lines with time-dependent parameters 107
6.1 Introduction 108
6.2 Literature review 110
6.3 Proactive Kanban System 112
6.3.1 Flow line model 112
6.3.2 Proactive Kanban Card Setting Problem 114

6.4 Solving the Proactive Kanban Card Setting Problem 115
6.4.1 Observations from numerical tests 115
6.4.2 Dominance between time-dependent buffer allocations 117
6.4.3 Local search approach 117
6.4.4 Approaches based on steady-state models 120
6.5 Numerical study 121
6.5.1 Impact of the timing of buffer changes 121
6.5.2 Impact of the number of buffer changes 124
6.5.3 Lines with multiple stations and Erlang-k processing distributions 125
6.5.4 Performance of the search algorithm 126
6.6 Conclusion and further research 127

7 Conclusions and Outlook **129**
7.1 Conclusions 129
7.2 Further possible research directions 130

Bibliography **XV**

Appendix A **XLI**

Appendix B **XLV**

Curriculum vitae **XLVII**

List of Figures

2.1 Serial production line with K stations and $K-1$ buffers of capacity B_i 8

3.1 Classification of approaches 18
3.2 Transformation with $L^S = 0$ as initial condition 33
3.3 Methodological links between approaches for the performance evaluation of time-dependent systems 49

4.1 Relationships between the proposed approximations and discrete-event simulation 63
4.2 Flow line with time-dependent release rate and random breakdowns and repairs 66
4.3 Realization of a two-state process ω_m with values in $\mathbb{B}$ . . . 68
4.4 Flow line represented by unit intervals and zero space dimensional buffers 72
4.5 Flow function 73
4.6 Graphical interpretation of Equation (4.7) 75
4.7 Expected WIP over time for an increase of the release rate . 84
4.8 Expected cumulated output and its relative error over time for an increase of the release rate 85
4.9 Expected WIP over time for a decrease of the release rate . 86
4.10 Expected cumulated output and its relative error over time for a decrease of the release rate 87
4.11 Cumulated output $\mathrm{E}[TH^c(t)]$ obtained by the MIP, the continuous model, and the DES for $\mu_0(t)$ with $h = 1.0$, $l = 140$, $M = 10$, and $b_m = 5\ \forall m$ 91

5.1 Stochastic flow line with time-dependent buffer allocation and effective processing times 96
5.2 Time-dependent generation of production capacities $c_{k,t,s}$ for sample s on machine k in period t 97

5.3 Static and dynamic buffer allocations, with $t_{rc} = t_{bc} = 51$ for the dynamic allocation 102
5.4 Change of the buffer capacity from large to small (LS) and small to large (SL) 103
5.5 Dynamic buffer allocations for $t_{rc} = 51$ and $t_{bc} \in \{31, 51, 71\}$ 104
5.6 Stepwise dynamic buffer allocations with $b_{1,t}^{inter} \in \{3, 4, 5\}$. 105

6.1 Flow line representation of the Proactive Kanban System . . 113
6.2 Example of non-convexity of $\mathrm{E}[W(\mathbf{B})]$ in $B_{1,1}$ 116
6.3 Search procedure for subproblem with 2 decision variables $B_{M,i'}$ and $B_{M,i''}$ 119
6.4 Expected WIP over time for $I = 1, t_1^* = 960$ for PKA, constant, and steady-state based allocations 122
6.5 Impact of the timing of the buffer change t_1^* $(I = 1)$ 124
6.6 Impact of the number of buffer changes I and their timing t_i^* 125

B.1 Expected average WIP and γ-service level for $t_i^* = 480$ . . . XLV
B.2 Expected average WIP and γ-service level for $t_i^* = 960$ . . . XLV
B.3 Expected average WIP and γ-service level for $t_i^* = 1440$ . . XLVI
B.4 Expected average WIP and γ-service level for $M = 2$, $I = 1$, $t_i^* = 960$, $B_{1,0} = B_{1,1} = 1$ XLVI

List of Tables

2.1 Characteristics of unreliable flow lines 10
2.2 Characteristics of the decision problems 14

3.1 Notation 19
3.2 Numerical solution approaches 24
3.3 Analytical results and explicit solutions (EXPL) 27
3.4 Performance evaluation methods based on piecewise stationary models 28
3.5 Approaches based on piecewise stationary models (independent periods) 32
3.6 Approaches based on piecewise stationary models (linked periods) 34
3.7 Approaches based on piecewise transient models 36
3.8 Discrete-time approaches (DTA) 38
3.9 Infinite-server approximations (INFSA) 41
3.10 Approximations based on modified job characteristics 46
3.11 Numerical comparisons of time-dependent queueing systems 50
3.12 Applications in the area of service systems 54
3.13 Applications in the areas of road and air traffic 57
3.14 Applications in the area of IT systems 59

4.1 Notation for the discrete-time and discrete-material model . 70
4.2 Parameters for an increasing release rate $\mu_0(t)$ 84
4.3 Parameters for a decreasing release rate $\mu_0(t)$ 85
4.4 E[$TH^c(t)$] obtained with CON and MIP approach and relative deviation to DES for different values of h, l, b_m, M . . 89

5.1 Notation for the evaluation model 98

6.1 Notation for Proactive Kanban Systems 112
6.2 Comparison of allocations and resulting performance generated by different $M = 1, I = 1$ 122

6.3 Comparison of allocations and resulting performance for multi-stage systems 126
6.4 Computational efficiency of the local search algorithm . . . 127

1 Introduction

Flow lines are production systems which are typically installed for production involving high volumes and comparatively low costs. Amongst others, manufacturers in the automotive and food industries use this manufacturing concept (Liberopoulos and Tsarouhas, 2002; Li, 2013).

A flow line consists of multiple stations in series. Each station performs manufacturing operations on discrete workpieces that move individually along the line. The operations at each station are performed either manually by workers or automated. In both cases, the production processes are typically subject to both stochastic and time-dependent effects. The stochastic variability of the time workers need to perform a given manufacturing task can be represented by random variables. The processing time of automated machines is typically close to deterministic. However, they are subject to random breakdowns and successive repairs (Inman, 1999). Time-dependent effects, i.e., parameter changes over time, e.g., in the random variables, may occur in the production system itself or may be induced by customers. Time-dependent changes in the customer demand, e.g, as a result of seasonal patterns, are reported by Tardif and Maaseidvaag (2001) and Takahashi and Nakamura (2002). A less explored field is that of time-dependent changes in the production process. The changes are caused by learning effects during the ramp-up (Terwiesch and Bohn, 2001) and the introduction of new manufacturing technologies and machinery (Jaikumar and Bohn, 1992). These effects often improve both mean and variance of the production process.

Typically, buffers are allocated between stations to compensate for stochastic effects in the production process. They decouple the stations by storing processed workpieces in the event of a long processing time or a breakdown in the downstream station. If the buffer capacities are finite, blocking may occur, i.e., the condition in which a station stops processing as it cannot move the completed workpiece to the downstream buffer. Moreover, the workpieces stored in the buffers can prevent a station from starving, i.e., a station stops processing due to a lack of raw material. The adequate allocation of finite buffer capacities represents a crucial decision. Inadequate or misallocated

buffer capacities may lead to a reduction in throughput and ultimately to lost sales, whereas excessively buffer capacities lead to high installation costs and additional costs originating from WIP stored in the buffers (Gershwin and Schor, 2000). This WIP produces inventory holding costs as well as operational risks such as damage, theft, or, in the case of perishable goods, deterioration and consequently costs incurred due to scrapping. Burman et al. (1998) optimized the buffers for the printer production at Hewlett-Packard and increased revenues by $280 million. Liberopoulos and Tsarouhas (2002) increased the profit of a Greek food company by $19,150 per week. Their optimization of buffer capacities reduced costs for scrapping and overtime and also yielded additional revenue.

The physical limitation of the number of workpieces arising from finite buffer capacities can also be deliberately created by control mechanisms, such as Kanban. The maximum WIP at each stage is limited by the corresponding number of Kanban cards at each stage. Allocating Kanban cards to stations means in effect a decision on the buffer capacities. A formal equivalence of Kanban-controlled flow lines and flow lines with finite buffers is established by Berkley (1991). Originally introduced by Toyota, Kanban cards also serve as production authorization (Monden, 1983). Production at each station is only possible if both WIP and a matching Kanban card are available. Processed workpieces are moved jointly with the card to a downstream buffer. The cards are detached from the workpieces as soon as production on the following station starts. The cards are then moved to the upstream station to signal demand for replenishment.

The flexibility of the Kanban control mechanism also permits changes in Kanban allocations, i.e., buffer allocations, to account for time-dependent changes in the demand and production process. Tardif and Maaseidvaag (2001) propose to add and capture extra cards in relation to inventory thresholds. Takahashi and Nakamura (2002) use statistical analysis of the demand data to detect changes in the parameters of the demand distribution. If a parameter change is detected, the Kanban allocation is adapted in such a way that it matches the new demand parameters. Both approaches do not take into account information about future parameter changes. This information can be made available if parameters are under direct control, e.g., the introduction of new machinery, or if empirical data allow precise forecasts, e.g., by learning curves. Here there is an opportunity for further research with the objective of proactively changing the Kanban allocation to account for time-dependent parameter changes.

One way of exploiting the flexibility of Kanbans is to determine time-dependent Kanban allocations in such a way that they minimize the average WIP in the flow line while maintaining a required service level. This decision problem is hard to solve for two main reasons: First, the stochastic and time-dependent impacts lead to non-linear changes in performance with respect to changes of the Kanban allocation. Hence, even with practical experience, intuition is only of limited help in predicting the outcome of such changes. Consequently, adequate performance evaluation approaches need to be developed. Secondly, allowing a time-dependent change means that additional complexity is added to the already NP-hard Buffer Allocation Problem (Smith and Cruz, 2005). The large number of candidate allocations makes a complete enumeration of the solution space impossible even for small problems. All in all, the evaluation and determination of time-dependent buffer allocations in stochastic flow lines is a novel and challenging research field which requires the development of new analytical methods.

Initially, this thesis reviews and classifies the literature on the Buffer Allocation Problem under steady-state conditions and on performance evaluation approaches for queueing systems with time-dependent parameters. Subsequently, new performance evaluation approaches are developed. Finally, a local search algorithm for the derivation of time-dependent buffer allocations is proposed. The algorithm is based on numerically observed monotonicity properties of the system performance in the time-dependent buffer allocations. Numerical examples illustrate that time-dependent buffer allocations represent an adequate way of minimizing the average WIP in the flow line while achieving a desired service level.

Chapter 2 proposes a classification scheme for flow lines and different versions of the Buffer Allocation Problem. A survey with respect to flow lines with unreliable machines is conducted. The articles are categorized according to the introduced classification scheme. Common assumptions, existing test instances, and existing solution approaches are identified. In many cases the underlying assumptions of the model and the characteristics of the solutions obtained are described only insufficiently or not at all. The new classification scheme is designed to establish a set of characteristics which are required for systematic numerical comparisons of different solution algorithms. All of the reviewed flow line models assume steady-state conditions. This article was written jointly with Sophie Weiss and Raik Stolletz[1].

[1] Weiss, S., J. A. Schwarz, and R. Stolletz (2015). Buffer Allocation Problems for stochastic flow lines with unreliable machines. In Proceedings of the 10th Conference on Stochastic Models of Manufacturing and Service Operations, Volos, Greece, pages 271-277

Chapter 3 introduces a classification scheme for performance evaluation approaches of queueing systems with time-dependent parameters. The development of performance evaluation approaches is often motivated by real-world problems. Service, IT, and road and air traffic systems are identified as the main areas of application. The adaption of buffer capacities in response to time-dependent behavior is investigated only in a single article, in this case in a call center context. The developed classification scheme groups the existing literature in accordance with the key ideas of analysis into three main categories, (i) numerical and analytical solutions, (ii) approximations based on models with piecewise constant parameters, and (iii) approximations based on modified system characteristics. The survey reveals that exact analytical solutions are established only on the basis of restrictive assumptions. In particular, no exact results are available for systems with finite buffer capacities. Methodological links are established between the different approaches. These links exist for approaches from the same category of the classification scheme but also beyond category borders. Moreover, a list of existing numerical comparisons for different approaches is provided to enrich the picture of the relations between the approaches. The article is the result of joint work with Gregor Selinka and Raik Stolletz[2].

The fourth chapter presents two sampling approaches for the performance analysis of flow lines with a time-dependent production rate for the first machine but constant and finite buffer capacities. Sampling approaches replace random variables by sampled realizations of the random variables which then allow a deterministic analysis. Insights regarding the time-dependent and stochastic system can be obtained from sample averages. The first approach is based on a mixed-integer program (MIP). It captures the discrete nature of workpieces in the line but approximates the continuous time by discrete-time intervals. The second approach is based on partial and ordinary differential equations in continuous time. However, it approximates the discrete workpiece by establishing a continuum. References for both continuous and discrete material flow models are reviewed. In addition, it is shown that the two proposed approaches are equivalent, given certain linearity assumptions, and thereby the two literature streams can be linked. A numerical study demonstrates the accuracy of both approximations relative to a discrete-event simulation in continuous time. Moreover, it is shown that buffers do not only capture stochastic variations but also smooth the time-dependent output induced by a time-dependent processing rate for the first machine. This article

[2]Schwarz, J. A., G. Selinka, and R. Stolletz (2016). Performance analysis of time-dependent queueing systems: Survey and classification. Omega (DOI: 10.1016/j.omega.2015.10.013)

was written jointly with Simone Göttlich, Sebastian Kühn, and Raik Stolletz[3].

Chapter 5 proposes a sample-based evaluation approach for systems with time-dependent buffer allocations. It introduces the concept of adapting buffer capacities to account for time-dependent station parameters. Further, the particular characteristics of a buffer capacity reduction are discussed. A numerical study demonstrates the accuracy of the evaluation approach by comparison with a discrete-event simulation. In addition, it provides numerical evidence that time-dependent buffer allocation can be used to influence key performance characteristics of flow lines such as WIP and throughput. This article was written jointly with Raik Stolletz[4].

Chapter 6 investigates a serial flow line with finite buffers which serves a stochastic and time-dependent demand from a finished goods buffer. Each station of the line is characterized by generally distributed processing times with time-dependent parameters. We propose time-dependent changes in buffer capacities utilizing Kanban cards to minimize the required WIP, while maintaining a predefined γ-service level over a finite planning horizon. We first report monotonicity results for the service level and the expected average WIP with respect to time-dependent buffer capacities that are observed in a numerical study. Based on these observations, a local search algorithm is developed. The numerical study demonstrates that the generated time-dependent allocations reduce the required WIP, as compared to constant allocations. Moreover, we test allocation approaches based on steady-state models and demonstrate that they may lead to infeasibility. This article was written jointly with Raik Stolletz[5].

The Chapters 2 to 6 may be read independently. Each of the chapters includes an introduction, a review of the relevant literature, and concluding remarks for the chapter in question. The references for all chapters are listed in a joint bibliography. Conclusions and indications for future research based on this thesis as a whole are provided in Chapter 7.

[3] Göttlich, S., S. Kühn, J. A. Schwarz, and R. Stolletz (2016). Approximations of time-dependent unreliable flow lines with finite buffers. Mathematical Methods of Operations Research (DOI: 10.1007/s00186-015-0529-6)

[4] Schwarz, J. A. and R. Stolletz (2013). A sampling approach for the analysis of time-dependent stochastic flow lines. In Proceedings of the 9th Conference on Stochastic Models of Manufacturing and Service Operations, Seeon, Germany, 2013, pages 181-188

[5] Schwarz, J. A. and R. Stolletz (2015). A proactive approach to Kanban allocation in stochastic flow lines with time-dependent parameters. Working paper

2 Buffer Allocation Problems for stochastic flow lines with unreliable machines

Co-authors:
Sophie Weiss and **Raik Stolletz**
Chair of Production Management, Business School, University of Mannheim, Germany

Published in:
Proceedings of the 10th Conference on Stochastic Models of Manufacturing and Service Operations, Volos, Greece, 2015, pages 271-277

Abstract:
The Buffer Allocation Problem in serial production lines is solved for different objectives, constraints, and assumptions. The aim of this work is to characterize analyzed production lines with unreliable machines and the underlying decision problems. We investigate unreliable serial lines with finite intermediate buffers and a single machine per station that processes discrete material. Moreover, we review existing solution approaches.

2.1 Introduction

Flow lines process workpieces sequentially on multiple stations. These production systems usually have a finite buffer capacity and are frequently used in manufacturing, in particular in the automotive industry (Tempelmeier, 2003; Li, 2013). They often experience random processing times, stochastic failures, and successive repairs. This leads to blocking and starvation which reduce the throughput of the line. A station starves if it cannot produce due to a lack of material in the upstream buffer whereas a blocked machine stops production due to a full downstream buffer. The choice of the total buffer capacity and its allocation between machines is a key design decision. This is because buffer capacities are associated with the costs of the buffer itself and the related work-in-process inventory (WIP) stored in it. The decision on the buffer capacities and their allocation is well known as the Buffer Allocation Problem (BAP).

The BAP is a well-researched problem which is hard to solve. On the one hand, the exact performance evaluation of flow lines is only possible for small systems under specific assumptions, and on the other hand, the allocation of buffer capacities is an NP-hard combinatorial problem (Smith and Cruz, 2005). Therefore, exact solutions for the BAP exist only for special cases (Enginarlar et al., 2005). However, heuristic search algorithms in combination with approximative evaluation methods are frequently used. The solution quality of these approaches is typically investigated numerically. Gershwin and Schor (2000) provide a comprehensive overview of solution approaches for the BAP published prior to the year 2000.

We provide a survey of the characteristics of the lines for analyzed instances of the BAP. We focus on unreliable serial lines with finite intermediate buffers and a single machine per station that processes discrete material (Figure 2.1).

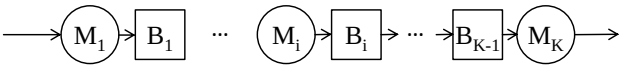


Figure 2.1: Serial production line with K stations (circles) and $K-1$ buffers of capacity B_i (rectangles)

Further, we discuss different problem formulations of the BAP and their solution approaches. We include references that have been published after the review of Gershwin and Schor (2000).

The remainder is organized as follows: Section 2.2 provides a classification of flow line characteristics. Section 2.3 addresses the different versions of the decision problem and the corresponding solution approaches. Concluding remarks and suggestions for future research are provided in Section 2.4.

2.2 Classification scheme for characteristics of flow lines

The key characteristics of serial lines are the number of stations, K, and the stations' stochastic properties. A station is characterized by the distribution of the processing times, the times to failure (TTF), and the times to repair (TTR). We found the following distributions in the literature: Deterministic (DET), Exponential (EXP), Erlang (ERL), Rayleigh (RA), Geometric (GEO), Uniform (U), Gamma (GAMMA), Normal (NORM), Lognormal (LOGN), and Bernoulli (BER). We distinguish whether all machines have the same (balanced line) or different properties (unbalanced line). We include references only if all of these key characteristics are clearly documented with published parameters for all distributions.

In addition to the key characteristics, a set of assumptions about the flow of workpieces in the line is required in order to reproduce the dynamics of a flow line (Dallery and Gershwin, 1992). An assumption has to be made on the supply of raw material in front of the first machine, which can be unlimited, i.e., saturated or limited. Similarly, the demand for finished goods can be a limiting factor or there is a saturated demand. Moreover, the type of blocking has to be defined. If a buffer is full, the upstream station may either process an additional workpiece which then remains on the station until space in the downstream buffer becomes available, i.e., blocking after service (BAS), or no workpiece enters the machine until a buffer space becomes available, i.e., blocking before service (BBS). Unreliable stations can experience operation-dependent (OD) or time-dependent (TD) failures. In the former case, a station fails only while it is processing workpieces, while in the latter case, breakdowns occur independent of the operational status. If a failure occurs while a workpiece is being processed, it has to be specified whether the progress on the workpiece is conserved or lost. The differentiation becomes obsolete for exponentially distributed processing times or discrete-time models with Bernoulli and Geometric failures if the processing time equals the time interval length. In several cases these detailed assumptions are not reported on in

the surveyed papers. We mark missing information by * and not applicable categories by - in the tables. Notably, many references lack the required information to reproduce the instance of the line. Other features receive only little or no attention and are therefore not included in the table. For example scrap is only considered by Han and Park (2002). Moreover, correlations in the processing times are addressed only by Weiss and Stolletz (2015). They demonstrate that correlations can have a substantial impact on the optimal buffer allocation. Table 2.1 shows unreliable lines reported in the literature after the review of Gershwin and Schor (2000).

Table 2.1: Characteristics of unreliable flow lines

Reference	No. of stations	Processing time distr.	TTF distr.	TTR distr.	Unbalanced	Saturated supply	Saturated demand	Blocking type	Failure type	Work-conserving
Alon et al. (2005)	3,5,6,10	EXP	EXP	EXP	x	x	x	*	TD	-
	5	ERL	EXP	EXP	x	x	x	*	TD	*
Bekker (2013)	5	EXP	EXP	EXP	x	x	x	*	OD	-
	5	LOGN	EXP	EXP	x	x	x	*	OD	*
Chiang et al. (2000)	15	DET	EXP	EXP	x	x	x	BBS	OD	*
Demir et al. (2011)	5,9,10,12,20,40	DET	GEO	GEO		x	x	*	*	-
Diamantidis and Papadopoulos (2004)	4-6,10	DET	BER	BER	x	x	x	*	OD	-
Dolgui et al. (2002)	5	DET	EXP	EXP	x	x	x	*	OD	*
Dolgui et al. (2007)	5	DET	EXP	EXP	x	x	x	*	OD	*
Enginarlar et al. (2002)	2-20	DET	EXP	EXP		x	x	BBS	*	*
	2-20	DET	ERL	ERL		x	x	BBS	*	*
	2-20	DET	RA	RA		x	x	BBS	*	*
Enginarlar et al. (2005)	3-30	DET	EXP	EXP		x	x	BBS	TD	*
Gershwin and Schor (2000)	5,10,12,20,30	DET	GEO	GEO	x	x	x	*	*	-
	3,20	DET	GEO	GEO		x	x	*	*	-
	7	DET	EXP	EXP	x	x	x	*	*	*
Han and Park (2002)	5,10	DET	GEO	GEO	x	x	x	*	*	-
	5,10	DET	GEO	GEO		x	x	*	*	-
Helber (2001)	6	DET	GEO	GEO		x	x	*	OD	-
Kim and Lee (2001)	3,8,10	EXP	EXP	EXP	x	x	x	BAS	OD	x
Kose and Kilincci (2015)	5,10	DET	GEO	GEO	x	x	x	*	*	-
	9,20,40	DET	GEO	GEO		x	x	*	*	-
Lee et al. (2009)	5	DET	GEO	GEO	x	x	x	BBS	OD	-
Lee and Ho (2002)	5,6	EXP	EXP	EXP	x		x	*	*	-
	5,6	EXP	EXP	EXP			x	*	*	-
Li (2013)	9,20	DET	EXP	EXP	x	x	x	*	*	*
Massim et al. (2010)	3,5,10	DET	EXP	EXP	x	x	x	*	OD	*
Matta et al. (2012)	5	DET	EXP	EXP	x	x	x	*	OD	*
	12	DET	GEO	GEO	x	x	x	*	OD	-
Nahas et al. (2006)	7	DET	EXP	EXP	x	x	x	*	*	*
Papadopoulos and Vidalis (2001a)	3-6	EXP	EXP	EXP	x	x	x	BAS	OD	x
Sabuncuoglu et al. (2006)	3,5,10	DET	EXP	EXP		x	x	*	OD	x
	4-6,8-10	EXP	EXP	EXP	x	x	x	*	OD	x
	4,5,7-10,12	DET	EXP	EXP	x	x	x	*	OD	x
Savsar (2006)	5	EXP	EXP	U		x	x	*	OD,TD	*
	7	DET	U/EXP/NORM/ ERL/GAMMA	U/NORM/ LOGN/DET	x	x	x	*	OD,TD	*
Shi and Gershwin (2009)	3-6,12	DET	GEO	GEO	x	x	x	*	OD	-
Shi and Gershwin (2014)	30,70	DET	GEO	GEO		x	x	*	OD	-
	20	DET	GEO	GEO	x	x	x	*	OD	-
Shi and Men (2003)	9	DET	GEO	GEO		x	x	*	*	-
Tempelmeier (2003)	8,19,23	DET	EXP	EXP	x	x	x	*	OD	*
	14	ERL	EXP	EXP	x	x	x	*	OD	*
	14	EXP	EXP	EXP	x	x	x	*	OD	-
Weiss and Stolletz (2015)	14,24	DET/ERL	EXP	EXP	x	x	x	BAS	OD	x

Two-thirds of the references consider flow lines that are balanced. Processing times are mostly deterministic with exponentially or geometrically distributed TTF and TTR. In almost all other cases processing times are exponentially or Erlang-distributed, again with exponentially distributed TTF and TTR. It can be observed that OD-failures dominate TD-failures. For the majority of the references the assumptions on conservation of work during failures is either not applicable or not addressed. With respect to the supply of the line, all but one of the articles assume unlimited supply. Lee and Ho (2002) assume random arrivals with exponentially distributed inter-arrival times. The blocking policy is often not defined. For the cases in which the blocking policy is defined, BBS occurs twice as often as BAS.

Some instances of flow lines are used by multiple authors. Kose and Kilincci (2015), Demir et al. (2011), Lee et al. (2009), and Nahas et al. (2006) use instances of Gershwin and Schor (2000). Instances proposed by Papadopoulos and Vidalis (2001a) are utilized by Sabuncuoglu et al. (2006). Furthermore, Bekker (2013), Dolgui et al. (2007), Alon et al. (2005), and Dolgui et al. (2002) base their choice of instances on Vouros and Papadopoulos (1998).

2.3 Classification scheme for decision problems

The literature encompasses three main versions of the BAP. They all share the decision on the vector $\mathbf{B} = (B_1, B_2, ..., B_i, ..., B_{K-1})$, where B_i represents the capacity of the buffer behind station i.

(i) *Primal Problem:*

$$\min \sum_{i=1}^{K-1} B_i \tag{2.1a}$$

s.t.

$$\mathrm{E}[Th(\mathbf{B})] \geq Th^* \tag{2.1b}$$

$$B_i \in \mathbb{N}_0, \quad 1 \leq i \leq K-1 \tag{2.1c}$$

The objective of the primal problem is to minimize the total buffer capacity in the line while ensuring that the expected throughput, $\mathrm{E}[Th(\mathbf{B})]$, equals or exceeds a given desired throughput, Th^*. Th^* is usually selected as percentage of the theoretically achievable throughput in a line with infinite buffers.

(ii) *Dual Problem:*

$$\max \mathrm{E}[\mathit{Th}(\mathbf{B})] \tag{2.2a}$$

s.t.

$$\sum_{i=1}^{K-1} B_i = B_{tot} \tag{2.2b}$$

$$B_i \in \mathbb{N}_0, \qquad 1 \leq i \leq K-1 \tag{2.2c}$$

The dual problem with respect to the introduced primal (2.1) is the maximization of the expected throughput subject to the total buffer capacity, B_{tot}, available in the line. The value of B_{tot} is usually given by space requirements on the shop floor. However, the dual problem may also be used to solve the primal problem by repetitively solving the dual for several values of total buffer capacities (Lee et al., 2009; Tempelmeier, 2003).

(iii) *Profit Problem:*

$$\max \text{Profit} = \alpha\, \mathrm{E}[\mathit{Th}(\mathbf{B})] - \beta\, \mathrm{E}[\mathit{WIP}(\mathbf{B})] - \gamma \sum_{i=1}^{K-1} B_i \tag{2.3a}$$

s.t.

$$\sum_{i=1}^{K-1} B_i \leq B_{tot} \tag{2.3b}$$

$$\mathrm{E}[\mathit{Th}(\mathbf{B})] \geq \mathit{Th}^* \tag{2.3c}$$

$$B_i \in \mathbb{N}_0, \qquad 1 \leq i \leq K-1 \tag{2.3d}$$

An attempt to directly balance the economic benefits of throughput with the buffer-related costs in the objective function is the profit problem. It uses weightings α, β, and γ to convert the technical measures of expected throughput, expected WIP, and buffer capacities into monetary units. The objective is to maximize the profit resulting from the gained revenue under the consideration of costs for the buffer capacities and the WIP stored in them. There is a constrained and an unconstrained version of the profit problem, i.e., Constraints (2.3b) and (2.3c) are not necessarily part of the decision problem. In the references considered, the parameters α, β, and γ are chosen without a direct link to empirical data.

(iv) *Other Problems:*

The works of Kim and Lee (2001) and Lee and Ho (2002) consider special cases of the BAP. Kim and Lee (2001) solely focus on the cost originating from the expected WIP, whereas Lee and Ho (2002) omit WIP-related costs and include costs for occurring throughput losses. Helber (2001) emphasizes that cash flows from revenue and investments in buffer capacities have different time scales. Thus, Helber (2001) suggests the use of a net present value approach. The problems introduced so far are all based on a single objective. Another idea is a multi-objective function. This approach delivers pareto-optimal solutions. Bekker (2013) employs this concept for the conflicting goals of throughput and WIP.

Table 2.2 lists the types of decision problems and the solution approaches that can be found in the literature. Most of the references address the primal or the dual problem. Both are addressed equally often. The minority of the references covers the optimization of profits.

The solution approaches for the BAP include a generative and an evaluative part. The generative method selects candidate solutions which have to be evaluated. The evaluation method determines the performance of the line, e.g., expected throughput or expected WIP, for a given buffer allocation. Sometimes integrated approaches are applied. Weiss and Stolletz (2015) use a Benders Decomposition approach which is based on a mixed-integer programming formulation. In this special case, the corresponding master- and subproblem divide the approach into an integer programming-based generative and an evaluative method. An approach only delivers exact solutions if the generative and the evaluative part are both exact. Note that the simulation result converges to the exact solution if the length of the simulation run or the number of replications is chosen large enough. We therefore mark simulation with (x) in the table. Exact results for both, the generative and the evaluative method, are obtained only for two-machine lines (Enginarlar et al., 2002, 2005). For long simulation runs, Weiss and Stolletz (2015) also provide exact results. Metaheuristics, such as Genetic algorithms (GA), tabu search (TS), and artificial neural networks (ANN) are developed mainly for the dual problem. In contrast, rule-based allocation strategies and search algorithms are often employed for the primal problem. Maximization of profit functions is mainly addressed by genetic algorithms and gradient methods. Evaluation approaches are typically based on simulation, decomposition, and aggregation.

Table 2.2: Characteristics of the decision problems

	Decision Problem				Solution Approach			
Reference	Primal	Dual	Profit	Others	Generative method	Exact	Evaluation method	Exact
Alon et al. (2005)		x			Alias method based on cross entropy		Simulation	(x)
Bekker (2013)				x	Cross entropy method		Simulation	(x)
Chiang et al. (2000)	x				Rule of thumb		Aggregation	
Demir et al. (2011)		x			TS		DDX	
	x				Binary search and TS		DDX	
Diamantidis and Papadopoulos (2004)		x			Dynamic Programming		Aggregation	
Dolgui et al. (2002)			x		GA		Aggregation	
Dolgui et al. (2007)			x		Hybrid GA and Branch and Bound		Aggregation	
Enginarlar et al. (2002)	x				Analytical solution	x	Analytical solution	x
	x				Buffer allocation rule		-	
Enginarlar et al. (2005)	x				Analytical solution	x	Analytical solution	x
	x				Analytical solution	x	Aggregation	
	x				Buffer allocation rule		-	
Gershwin and Schor (2000)	x				Search algorithm		DDX/ADDX	
		x			Gradient algorithm		DDX/ADDX	
			x		Gradient algorithm		DDX/ADDX	
Han and Park (2002)	x				Steepest descent with penalty function		Aggregation	
Helber (2001)				x	Gradient algorithm		Decomposition	
Kim and Lee (2001)				x	Local search		Decomposition	
Kose and Kilincci (2015)		x			Hybrid GA and Simulated Annealing		Simulation	(x)
Lee et al. (2009)		x			ANN and GA		Simulation	(x)
Lee and Ho (2002)				x	Modified responds surface methodology		Simulation	(x)
Li (2013)	x				Bottleneck-based iterative approach		Approx. analytical formula	
Massim et al. (2010)			x		Artificial immune algorithm		DDX	
Matta et al. (2012)	x				Numerical optimization technique		Kriging approximation	
Nahas et al. (2006)		x			Degraded ceiling approach		ADDX	
Papadopoulos and Vidalis (2001a)		x			Sectioning approach		Markovian state model	
Sabuncuoglu et al. (2006)		x			Search algorithm		Simulation	(x)
Savsar (2006)		x			Enumeration		Simulation	(x)
Shi and Gershwin (2009)			x		Gradient method		Decomposition	
Shi and Gershwin (2014)			x		Gradient method with segmentation		Decomposition	
Shi and Men (2003)		x			Hybrid nested partition and TS		DDX	
Tempelmeier (2003)	x				Search algorithm and gradient-based search		ADDX	
		x			Gradient-based search		ADDX	
Weiss and Stolletz (2015)	x				Integer program	x	Simulation	(x)

2.4 Conclusion and future research

We introduce a classification scheme that is used to describe existing unreliable flow lines for which the BAP is solved in its different problem formulations. Common assumptions are unlimited supply and an infinite last buffer. Failure type, conservation of work, and blocking type are only reported on insufficiently. Most of the references consider the primal and the dual problem. The maximization of a profit function is only considered in few cases. The corresponding solution approaches are mostly heuristic for both the generative and the evaluation part. Although some instances are used in several publications, there is a need for a library of sample instances with a complete description of the line characteristics and the allocations obtained with different solution approaches.

3 Performance analysis of time-dependent queueing systems: Survey and classification

Co-authors:
Gregor Selinka and **Raik Stolletz**
Chair of Production Management, Business School, University of Mannheim, Germany

Published in:
Omega, 2016, DOI: 10.1016/j.omega.2015.10.013, In Press, pages 1-20, reprinted with permission from Elsevier

Abstract:
Many queueing systems are subject to time-dependent changes in system parameters, such as the arrival rate or number of servers. Examples include time-dependent call volumes and agents at inbound call centers, time-varying air traffic at airports, time-dependent truck arrival rates at seaports, and cyclic message volumes in computer systems.

There are several approaches for the performance analysis of queueing systems with deterministic parameter changes over time. In this survey, we develop a classification scheme that groups these approaches according to their underlying key ideas into (i) numerical and analytical solutions, (ii) approaches based on models with piecewise constant parameters, and (iii) approaches based on modified system characteristics. Additionally, we identify links between the different approaches and provide a survey of applications that are categorized into service, road and air traffic, and IT systems.

3.1 Introduction

Many queueing systems feature time-dependent changes in parameters. Examples of non-stationary parameters, such as the arrival rate or number of servers, include time-dependent call volumes and agents at inbound call centers, time-varying air traffic at airports, non-stationary truck arrival rates at container terminals, and cyclic message volumes in IT systems. Because these time-dependent parameter changes can have a substantial impact on a queueing system's performance, they must be considered in the design and control of such systems.

In this article, we classify performance evaluation methods for single-stage queueing systems with time-dependent but deterministic parameter changes. While such systems are also called non-stationary, time-varying, time-inhomogeneous, or non-homogeneous queueing systems, we solely use the term time-dependent queueing systems.

The analysis of time-dependent queueing systems has a long tradition dating back to Kolmogorov (1931). Since then, the practical relevance of such systems has stimulated increasing interest in various research areas, including mathematics, computer science, and operations management. Such an analysis itself is difficult since common relations for steady-state queueing systems, such as Little's law, must be reformulated (Bertsimas and Mourtzinou, 1997).

The contribution of the present work is a survey and classification of the literature on performance evaluation approaches for time-dependent queueing systems and their applications. Additionally, links between different approaches are identified and discussed.

The remainder of this paper is organized as follows. The scope of the survey and the classification scheme are introduced in Section 3.2. In Section 3.3, approaches for the analytical treatment of time-dependent queueing systems are reviewed and classified according to the developed scheme. A visualization of links between the approaches and a review of numerical studies that compare several methods are provided in Section 3.4. Areas of application and their unique characteristics are described in Section 3.5. In Section 3.6, concluding remarks and areas for future research are provided.

3.2 Scope and classification scheme

The survey presented in this paper reviews and classifies approaches for the time-dependent performance evaluation of single-stage queueing systems without spatial dimension, known as point queues, that include

- abandonments and retrials,
- arrivals from an infinite population that are served individually by a single server or one of multiple parallel servers (for a treatment of finite source systems, see e.g. Alfa (1979), Chung and Min (2014), and references within),
- waiting rooms larger or equal to one (i.e., waiting or loss-waiting systems; for a recent but incomplete survey of time-dependent loss queues, see Alnowibet and Perros (2006)),
- and deterministic system parameters that change over time (the transient analysis of systems with constant parameters is addressed, e.g., by Van de Coevering (1995), Tarabia (2000), and references within).

We survey approaches that allow for the performance analysis of arbitrary time instances. Discrete-event simulation is also applied for time-dependent performance evaluation. However, it is associated with a simulation error. This error can be reduced by an increase in the number of replications at the price of increasing run times (Nasr and Taaffe, 2013). Moreover, structural system properties remain intractable. Thus, the survey comprises only approaches which do not require the generation of random numbers.

We identify three main categories of evaluation approaches: the first category comprises numerical and analytical solution approaches for systems of equations that describe the time-dependent behavior of a queueing system (Section 3.3.1); the second category includes approaches that assume piecewise constant parameters and that apply stationary or transient models (Section 3.3.2); and the third category includes approximation methods that modify the number of servers or properties of the processed jobs (Section 3.3.3). Figure 3.1 presents our classification scheme including these categories and all evaluation approaches reviewed in this work. In the corresponding sections, each approach is described in terms of its key idea, its chronological development, and its advantages and limitations. These descriptions include only references that develop or methodologically extend an approach.

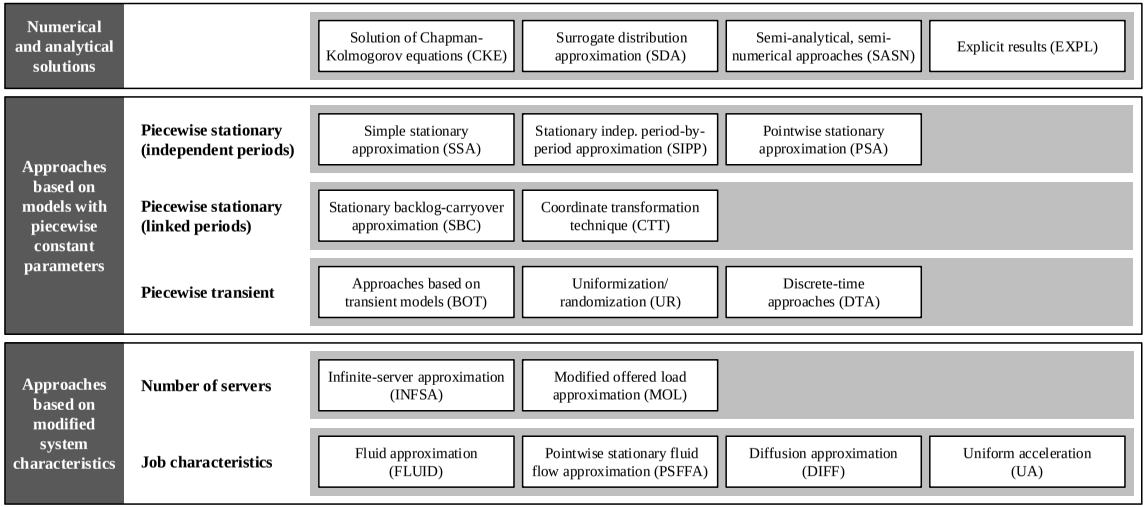


Figure 3.1: Classification of approaches

All surveyed references together with the characteristics of the analyzed queueing systems are listed in Tables 3.2, 3.3, and 3.5 to 3.10. For each reference that considers several queueing systems, the characteristics of the most general one are given. The references are sorted chronologically for each approach. The notation used in the following sections is provided in Table 3.1.

Table 3.1: Notation

Model description	
λ	Arrival rate
X	Arriving batch size distribution
c	Number of parallel servers
μ	Processing rate
Y	Batch service size distribution
s	Max. no. of jobs served by a batch server
$\rho = \frac{\lambda}{c\mu}$	Traffic intensity
K	Maximum no. of jobs in the system
PPrio	Preemptive priority
NPPrio	Non-preemptive priority
t	Time parameter
$(\cdot)$'	Derivatives with respect to time
Performance measures	
U	Utilization
L^Q	No. of jobs in the queue
L^S	No. of jobs in the system
W^Q	Waiting time of a job
W^S	Sojourn time of a job
Probabilities	
$P_n = P(L^S = n)$	Probability of n jobs in the system
$P_w = P(W^Q > 0)$	Probability of waiting
$\mathbf{P} = (P_0, P_1, ...)$	Vector of state probabilities

The development of approaches for the performance evaluation is often driven by real-world problems. Hence, many articles include both, an evaluation approach and its application to a real-world problem. The classification according to the area of application considers the references that include a detailed description of a specific application accompanied by a numerical study. The

reviewed applications of the approaches are divided into the areas of service systems (Section 3.5.1), road and air traffic systems (Section 3.5.2), and IT systems (Section 3.5.3).

3.3 Performance evaluation approaches

3.3.1 Numerical and analytical solutions

The **Chapman-Kolmogorov equations (CKEs)** compose a set of differential equations (DEs) that describes the dynamic behavior of a Markovian queueing system. For an $M(t)/M(t)/c$ system, the DEs are given as

$$\begin{aligned} P_0'(t) =& \mu(t)P_1(t) - \lambda(t)P_0(t), && n = 0 \\ P_n'(t) =& (n+1)\mu(t)P_{n+1}(t) + \lambda(t)P_{n-1}(t) \\ & - (\lambda(t) + n\mu(t))P_n(t), && 1 \leq n < c \\ P_n'(t) =& c\mu(t)P_{n+1}(t) + \lambda(t)P_{n-1}(t) - (\lambda(t) + c\mu(t))P_n(t), && n \geq c. \end{aligned} \tag{3.1}$$

Analytical solutions for these DEs exist only for special cases, e.g., $c = \infty$. However, solutions can be obtained numerically by using the Euler method or a Runge Kutta scheme. Systems with an infinite waiting room result in an infinite number of DEs. Kolesar et al. (1975) suggest approximating such systems by using a system with a finite but sufficiently large waiting room.

The CKEs are introduced by Kolmogorov (1931) for an $M(t)/M/c$ system. The numerical solution of the CKEs is used for the performance evaluation of an $M(t)/M/1/K$ system by Koopman (1972), an $M(t)/M(t)/1/K$ system with two separate queues and a common server by Bookbinder (1986), and a multi-class $M(t)/M(t)/1/K$/NPPrio system by Van As (1986). In addition, the numerical solution is used in the evaluation part of optimization algorithms, e.g., by Parlar (1984) and Nozari (1985), as well as in the dynamic programming approaches of Bookbinder and Martell (1979) and Jung and Lee (1989b).

The numerical solution for the CKEs has the advantage that the complete time-dependent distribution of the state probabilities is obtained. Thus, this solution can be used to calculate relevant quantiles (Ingolfsson et al., 2002). However, the main disadvantages are that the solution approach applies only to Markovian systems and has long computation times (Ingolfsson et al., 2007).

Approximation approaches are proposed to reduce computation times. Instead of solving the CKEs directly, Escobar et al. (2002) suggest first reducing the state space of an $M(t)/E_k(t)/c/K$ system as an approximation of the original state space and then solving the reduced number of DEs numerically.

Another widely used approach for reducing the computational effort is the **closure approximation** or **surrogate distribution approximation (SDA)**. In this approach, the large or infinite number of CKEs is replaced by a small number of DEs for the moments of the distribution of the number of jobs in the system. The k-th moment differential equation (MDE) is obtained by summation of the differential equations in (3.1), each multiplied by n^k. The differential equations for the first moment $\mathrm{E}[L^S(t)]$ and the variance $\mathrm{Var}[L^S(t)]$ of an $M(t)/M(t)/c$ system are given by

$$\mathrm{E}[L^S(t)]' = \sum_{n=0}^{\infty} n \cdot P'_n(t) = \lambda(t) - c \cdot \mu(t) + \mu(t) \cdot \sum_{n=0}^{c-1} (c-n) \cdot P_n(t), \tag{3.2}$$

$$\begin{aligned}\mathrm{Var}[L^S(t)]' &= \sum_{n=0}^{\infty} \left(n - \mathrm{E}[L^S(t)]\right)^2 \cdot P'_n(t) \\ &= \sum_{n=0}^{\infty} n^2 \cdot P'_n(t) - 2 \cdot \mathrm{E}[L^S(t)] \cdot \sum_{n=0}^{\infty} n \cdot P'_n(t) \\ &= \lambda(t) + c \cdot \mu(t) - \mu(t) \cdot \sum_{n=0}^{c-1} (c-n) \cdot P_n(t) \cdot \left(2 \cdot \mathrm{E}[L^S(t)] + 1 - 2 \cdot n\right).\end{aligned} \tag{3.3}$$

MDEs (3.2) and (3.3) are independent of the maximum number of jobs in the system. Hence, systems with a large or infinite waiting room can be analyzed efficiently. However, to solve these MDEs, the time-dependent state probabilities $P_n(t)$ must be known. They are assumed to follow a certain distribution that *closes* the set of MDEs. This *surrogate distribution* is always chosen such that its first and second moments match $\mathrm{E}[L^S(t)]$ and $\mathrm{Var}[L^S(t)]$, respectively. The closed MDEs can then be solved numerically.

The idea of focusing on the analysis of MDEs is used in an earlier study by Clarke (1956). However, Rider (1976) reports the first attempt to approximate the expected queue length of an $M(t)/M(t)/1$ system with a closure for the probability of an idling server. Since then, different types of distributions have been used as a surrogate distribution. For instance, Rothkopf

and Oren (1979) use the negative binomial distribution for the number of jobs in an $M(t)/M(t)/c$ system. Clark (1981) shows that the Polya Eggenberger distribution yields superior results for the same type of system. The Polya Eggenberger distribution is also used for priority queues by Taaffe and Clark (1988) and phase-type arrival and service processes by Ong and Taaffe (1988). These models go along with an increased state space. Taaffe and Ong (1987) introduce state-space partitioning to handle the growing state space. Instead of using a single SDA for the complete state space, the approximation quality is improved by introducing subspaces and allowing for different surrogate distributions depending on the respective subspaces. Lau and Song (2008) analyze a queue with multiple job classes by first aggregating the classes, then applying the model of Rothkopf and Oren (1979), and subsequently disaggregating the results again. Pender (2014a) uses the Poisson distribution as a surrogate distribution and obtains a closed-form solution for $\mathrm{E}[L^S(t)]$. To improve the quality of the approximation of $\mathrm{Var}[L^S(t)]$, he suggests a truncated Poisson-Charlier polynomial expansion. In an approach unlike the other approaches, Massey and Pender (2013) propose using a continuous distribution for the approximation of the queueing process. Thereby, they derive the so-called Gaussian-variance and Gaussian-skewness approximations. These approaches are complemented by the approach of Pender (2014b), who includes the fourth MDE, reflecting the kurtosis of the queue length distribution, in his approximation based on a Gram Charlier expansion.

A key idea of the SDA is to calculate only moments of a distribution. Thus, the performance analysis is limited to these moments. Typically, the first and second moments of the number of jobs in the system are calculated. The SDA requires the Markov property for the arrival and service process. However, the use of phase-type distributions allows for the analysis of different coefficients of variations (Taaffe and Clark, 1988). This comes at the cost of an increased, but still limited, number of DEs (Ong and Taaffe, 1988).

Markovian queueing systems are often described by generating functions that can often be reduced to an integral equation or formulations that include modified Bessel functions. These evaluation approaches are known as **semi-analytical, semi- numerical (SASN)** approaches (Tan et al., 2013). A survey and numerical comparison of early SASN approaches is provided by Leese and Boyd (1966).

Clarke (1956) obtains a Volterra-type integral equation for the probability of an empty system in an $M(t)/M(t)/1$ system. An explicit solution is found for the special case of a constant relation $\lambda(t)/\mu(t)$. The approaches

of Luchak (1956) and Luchak (1957) involve Taylor expansions to obtain the probability of jobs in the system and the busy period of an $M(t)/PH(t)/1$ system, respectively. Rosenlund (1976) studies the busy period of an $M(t)^X/G/1$ system with batch arrivals distributed according to X. The system also features balking, i.e., arriving jobs join the queue only with a certain probability. Lyubarskii (1982) obtains the busy period of a $G(t)/G(t)/1$ system as a two-dimensional Volterra integral equation. Wragg (1963) and Zhang and Coyle (1991) find the complete state probability distribution of an $M(t)/M(t)/1$ system as a solution of integral equations. Stadje (1990) develops a solution approach for the $M(t)/M(t)/2$ system that is similar to the approach of Clarke (1956). A multi-server $M(t)/M(t)/c$ system is analyzed by Margolius (1999). Margolius (2005) derives integral equations for the probability distribution of jobs in an $M(t)/M(t)/c(t)$ system. By considering quasi-birth-and-death processes, Margolius (2007) generalizes her results to phase-type distributions and establishes a connection with matrix analytic methods (Margolius, 2008). Al-Seedy et al. (2009) extend the analysis of $M(t)/M(t)/1$ systems by incorporating time-dependent balking. For a special structure of $\lambda(t)$ and $\mu(t)$, Al-Seedy and Al-Ibraheem (2003) and El-Sherbiny (2010) obtain the probability distribution of $L^S(t)$ in an $M(t)/M(t)/\infty$ system.

Nelson and Taaffe (2004) derive a quasi-closed form of MDEs that describes the expected number $\mathrm{E}[L^S(t)]$ and the variance $\mathrm{Var}[L^S(t)]$ of jobs in the system in a $PH(t)/PH(t)/\infty$ system and integrate them numerically. Similarly, Nasr and Taaffe (2013) derive quasi-closed MDEs for the first and second moments of the departure process of a $PH(t)/M(t)/c/K$ system. In contrast to the SDA, the partial-moment differential equations that are used to close the MDEs are exact.

Table 3.2 summarizes the references that use a numerical solution approach and links them to the considered queueing systems. It becomes apparent that the approaches are applicable to a wide range of queueing systems with features such as priorities and abandonments. However, the numerical solutions of the CKEs and the SDA exploit the Markovian property. Notably, Czachórski et al. (2009) use the CKE approach only for the special case of exponential distributions, and Rothkopf and Johnston (1982) analyze an $M(t)/M/1$ system and then scale the results according to the Polaczek-Kintchine formula to integrate general service times.

Table 3.2: Numerical solution approaches

Reference	Queueing system			
Numerical solution of the CKEs				
Kolmogorov (1931)	$M(t)$	/ M	/ c	
Leese and Boyd (1966)	$M(t)$	/ M	/ 1	
Koopman (1972)	$M(t)$	/ M	/ 1	/ K
Kolesar et al. (1975)	$M(t)$	/ $M(t)$	/ $c(t)$	
Rider (1976)	$M(t)$	/ $M(t)$	/ 1	
Bookbinder and Martell (1979)	$M(t)$	/ M	/ c	/ K
Rothkopf and Oren (1979)	$M(t)$	/ $M(t)$	/ c	
Clark (1981)	$M(t)$	/ $M(t)$	/ c	
Parlar (1984)	$M(t)$	/ $M(t)$	/ c	/ K
Nozari (1985)	$M(t)$	/ M	/ c	
Bookbinder (1986)	$M(t)$	/ $M(t)$	/ 1	/ K
Van As (1986)	$M(t)$	/ M	/ 1	/ K / NPPrio
Taaffe and Ong (1987)	$PH(t)$	/ $M(t)$	/ c	/ K
Ong and Taaffe (1988)	$PH(t)$	/ $PH(t)$	/ 1	/ K
Taaffe and Clark (1988)	$M(t)$	/ $M(t)$	/ 1	/ K / NPPrio
Jung and Lee (1989b)	$M(t)$	/ M	/ $c(t)$	
Tipper and Sundareshan (1990)	$M(t)$	/ M	/ 1	
Green and Kolesar (1991)	$M(t)$	/ M	/ c	
Green et al. (1991)	$M(t)$	/ M	/ c	
Jung (1993)	$M(t)$	/ M	/ c	
Green and Kolesar (1995)	$M(t)$	/ M	/ c	
Green and Kolesar (1997)	$M(t)$	/ M	/ c	
Massey and Whitt (1997)	$M(t)$	/ M	/ c	
Escobar et al. (2002)	$M(t)$	/ $E_k(t)$	/ c	/ K
Ingolfsson et al. (2002)	$M(t)$	/ M	/ $c(t)$	/ K
Ingolfsson et al. (2007)	$M(t)$	/ M	/ $c(t)$	
Czachórski et al. (2009)	$G(t)$	/ G	/ 1	/ K / PPrio
Gillard and Knight (2014)	$M(t)$	/ M	/ $c(t)$	
Jacquillat and Odoni (2015)	$M(t)$	/ $E_k(t)$	/ 1	
Surrogate distribution approximation (SDA)				
Rider (1976)	$M(t)$	/ $M(t)$	/ 1	
Rothkopf and Oren (1979)	$M(t)$	/ $M(t)$	/ c	
Clark (1981)	$M(t)$	/ $M(t)$	/ c	
Rothkopf and Johnston (1982)	$M(t)$	/ G	/ 1	
Taaffe and Ong (1987)	$PH(t)$	/ $M(t)$	/ c	/ K
Ong and Taaffe (1988)	$PH(t)$	/ $PH(t)$	/ 1	/ K
Taaffe and Clark (1988)	$M(t)$	/ $M(t)$	/ 1	/ K / NPPrio
Ingolfsson et al. (2007)	$M(t)$	/ M	/ $c(t)$	
Lau and Song (2008)	$M(t)$	/ M	/ c	
Massey and Pender (2013)	$M(t)$	/ M	/ $c(t)$ + M	
Pender (2014a)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$	
Pender (2014b)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$	
Semi-analytical, semi-numerical approaches (SASN)				
Clarke (1956)	$M(t)$	/ $M(t)$	/ 1	
Luchak (1956)	$M(t)$	/ $PH(t)$	/ 1	
Luchak (1957)	$M(t)$	/ $PH(t)$	/ 1	
Wragg (1963)	$M(t)$	/ $M(t)$	/ 1	
Leese and Boyd (1966)	$M(t)$	/ M	/ 1	
Rosenlund (1976)	$M(t)^X$	/ G	/ 1	
Lyubarskii (1982)	$G(t)$	/ $G(t)$	/ 1	
Stadje (1990)	$M(t)$	/ $M(t)$	/ 2	

Table 3.2: Numerical solution approaches - continued

Zhang and Coyle (1991)	$M(t)$	/ $M(t)$	/ 1	
Margolius (1999)	$M(t)$	/ $M(t)$	/ c	
Al-Seedy and Al-Ibraheem (2003)	$M(t)$	/ $M(t)$	/ ∞	
Nelson and Taaffe (2004)	$PH(t)$	/ $PH(t)$	/ ∞	
Margolius (2005)	$M(t)$	/ $M(t)$	/ c	
	$M(t)$	/ M	/ $c(t)$	
Margolius (2007)	$PH(t)$	/ $M(t)$	/ 1	
Margolius (2008)	$M(t)$	/ E_k	/ 1	
	$M(t)$	/ $M(t)$	/ 1	
Al-Seedy et al. (2009)	$M(t)$	/ $M(t)$	/ 1	
El-Sherbiny (2010)	$M(t)$	/ $M(t)$	/ ∞	
Nasr and Taaffe (2013)	$PH(t)$	/ $M(t)$	/ c	/ K

Analytical results and explicit solutions (EXPL) for time-dependent queueing systems exist only for special system configurations and usually cannot be generalized.

Palm (1943) and Khintchine (1969) show that in an $M(t)/M/\infty$ system, the number of jobs $L^S(t)$ is Poisson distributed for a queueing system which started operating in the distant past. Newell (1966) extends the results to general service times. Ramakrishnan (1980) provides a simple argument for these findings for the special case of deterministic service times. Sharma and Gupta (1983) consider an $M(t)/PH/\infty$ system and prove that $L^S(t)$ is Poisson distributed if it follows a Poisson distribution at the beginning of the time horizon. For exponentially distributed service times and a given number of jobs at $t = 0$, Thakur et al. (1972) derive the mean and variance of $L^S(t)$. Abol'nikov (1968) obtains the generating function for the number of jobs in an $M(t)^X/M/\infty$ system and uses it to derive $\mathrm{E}[L^S(t)]$. Shanbhag (1966) studies an $M(t)^X/G/\infty$ system and confirms that $L^S(t)$ is a Poisson process for the special case of an initially empty system and $P(X = 1) = 1$, i.e., all jobs arrive individually. Carrillo (1991) and Eick et al. (1993b) review reported analytical results with respect to the $M(t)/G/\infty$ system. In addition, Eick et al. (1993b) highlight that in contrast to the stationary case, $L^S(t)$ depends on the service time distribution beyond its mean.

Brown and Ross (1969), Purdue (1974a), and Purdue (1974b) extend the analysis to time-dependent service time distributions. Brown and Ross (1969) show that for the $M(t)^{X(t)}/G(t)/\infty$ system, $L^S(t)$ and the number of departures $D^{S,c}(t)$ up to time t follow a compound Poisson process. Purdue (1974b) and Foley (1982) demonstrate that $L^S(t)$ and $D^{S,c}(t)$ correspond to Poisson processes if the system is initially empty and if batch arrivals are omitted. McCalla and Whitt (2002) derive an explicit formula for $\mathrm{E}[L^S(t)]$ in a $G(t)^{X(t)}/G(t)/\infty$ system and propose an approximation for the distri-

bution of $L^S(t)$, since it is not a Poisson distribution.

Several authors focus on the analysis of the $M(t)/M(t)/\infty$ system. Purdue (1974a) obtains the mean and variance of $L^S(t)$, given the initial distribution of jobs $L^S(0)$. For the special case of an initially empty or Poisson-distributed number of jobs in the system, Collings and Stoneman (1976) confirm that $L^S(t)$ is Poisson distributed. Additionally, Kambo and Bhalaik (1979) obtain the joint probability distribution of $L^S(t)$ and $D^{S,c}(t)$, which is used to derive the time-dependent mean and variance of both $L^S(t)$ and $D^{S,c}(t)$. Seemingly unaware of the previous findings, Ellis (2010) derives the same formulas for $\mathrm{E}[L^S(t)]$ and $\mathrm{Var}[L^S(t)]$ as obtained earlier by Kambo and Bhalaik (1979). Both Ellis (2010) and Kambo and Bhalaik (1979) demonstrate that the expected number of jobs in the system can be described by DE (3.4) with solution (3.5)

$$\mathrm{E}[L^S(t)]' = \lambda(t) - \mu(t) \cdot \mathrm{E}[L^S(t)], \tag{3.4}$$

$$\mathrm{E}[L^S(t)] = \mathrm{E}[L^S(0)] \cdot e^{-\int_0^t \mu(\tau)\,d\tau} + e^{-\int_0^t \mu(\tau)\,d\tau} \cdot \int_0^t \lambda(\tau) e^{\int_0^\tau \mu(r)\,dr}\,d\tau. \tag{3.5}$$

Thakur and Rescigno (1978) establish that solution (3.5) for a stochastic system is equivalent to the solution for a $D(t)/D(t)/\infty$ system.

Dai (1998) derives bounds on the moment-generating function of $L^S(t)$ for an $M(t)/M(t)/1$ system and discusses bounds on $\mathrm{E}[L^S(t)]$. Knessl and Yang (2002) obtain explicit results for an $M(t)/M(t)/1$ system given a special form of the traffic intensity. Green and Soares (2007) find exact formulas for the probability $P(W^Q(t) > w)$ of waiting longer than w time units in an $M(t)/M/c(t)$ system under the assumption that the state probabilities $P_n(t)$ are known and that a maximum of one change in the number of servers occurs in the interval under consideration. For the case of more than one change, they propose approximation formulas. Kim and Ha (2012) exploit the property that the explicit solution for an $M(t)/M/\infty$ system can be used to model an $M(t)/M/c(t) + M$ system with Poisson abandonments if the abandonment rate equals the service rate.

Except for four references in Table 3.3, all the references report results for infinite-server systems. In addition, all but one of the analyzed queueing systems share the property of a Poisson arrival process with time-dependent rate.

Table 3.3: Analytical results and explicit solutions (EXPL)

Reference	Queueing system		
Palm (1943)	$M(t)$	/ M	/ ∞
Newell (1966)	$M(t)$	/ G	/ ∞
Shanbhag (1966)	$M(t)^X$	/ G	/ ∞
Abol'nikov (1968)	$M(t)^X$	/ M	/ ∞
Brown and Ross (1969)	$M(t)^{X(t)}$	/ $G(t)$	/ ∞
Khintchine (1969)	$M(t)$	/ M	/ ∞
Thakur et al. (1972)	$M(t)$	/ M	/ ∞
Purdue (1974a)	$M(t)$	/ $M(t)$	/ ∞
Purdue (1974b)	$M(t)$	/ $G(t)$	/ ∞
Collings and Stoneman (1976)	$M(t)$	/ $M(t)$	/ ∞
Thakur and Rescigno (1978)	$M(t)$	/ $M(t)$	/ ∞
Kambo and Bhalaik (1979)	$M(t)$	/ $M(t)$	/ ∞
Ramakrishnan (1980)	$M(t)$	/ D	/ ∞
Foley (1982)	$M(t)$	/ $G(t)$	/ ∞
Sharma and Gupta (1983)	$M(t)$	/ $PH(t)$	/ ∞
Eick et al. (1993a)	$M(t)$	/ G	/ ∞
Eick et al. (1993b)	$M(t)$	/ G	/ ∞
Dai (1998)	$M(t)$	/ $M(t)$	/ 1
Green and Kolesar (1998)	$M(t)$	/ G	/ ∞
Knessl and Yang (2002)	$M(t)$	/ $M(t)$	/ 1
McCalla and Whitt (2002)	$G(t)^{X(t)}$	/ $G(t)$	/ ∞
Buczkowski and Kulkarni (2006)	$M(t)$	/ G	/ ∞
Green and Soares (2007)	$M(t)$	/ M	/ $c(t)$
Ellis (2010)	$M(t)$	/ $M(t)$	/ ∞
Kuraya et al. (2011)	$M(t)$	/ M	/ ∞
Kim and Ha (2012)	$M(t)$	/ M	/ $c(t) + M$

3.3.2 Approaches based on models with piecewise constant parameters

3.3.2.1 Piecewise stationary models with independent periods

This set of approaches divides the overall time horizon T into intervals for which constant input parameters are assumed. The performance in each interval $[a_i, b_i]$ $(i = 1, 2, ..., I)$ is then analyzed independently by using steady-state formulas. The approaches differ in the length l of the analyzed intervals and the determination of the input parameters in the corresponding performance calculations (see Table 3.4 for the case of a time-dependent arrival rate $\lambda(t)$).

Table 3.4: Performance evaluation methods based on piecewise stationary models

Interval length	**Interval i**	**Method / Reference**	**Input in performance evaluation**
$l = T$	$t \in [0, T]$	SSA / Green et al. (1991)	$\tilde{\lambda}(i) = \frac{1}{T} \int_0^T \lambda(t)\, dt$
		SPEA / Green and Kolesar (1995)	$\tilde{\lambda}(i) = \max_{t \in [0,T]} \lambda(t)$
		SPHA / Green and Kolesar (1995)	$\tilde{\lambda}(i) = \frac{1}{b-a} \int_a^b \lambda(t)\, dt$ with peak interval $[a, b]$
$0 < l < T$	$t \in [a_i, b_i]$	SIPP Avg / Kolesar et al. (1975)	$\tilde{\lambda}(i) = \frac{1}{l} \int_{a_i}^{b_i} \lambda(t) dt$
		SIPP Max / Green et al. (2001)	$\tilde{\lambda}(i) = \max_{t \in [a_i, b_i]} \lambda(t)$
		SIPP Mix / Green et al. (2001)	$\tilde{\lambda}(i) = \begin{cases} \frac{1}{l} \int_{a_i}^{b_i} \lambda(t)\, dt & \text{if } \frac{d\lambda(t)}{dt} > 0 \quad \forall t \in [a_i, b_i] \\ \max_{t \in [a_i, b_i]} \lambda(t) & \text{otherwise} \end{cases}$
		Lag Avg / Green et al. (2001)	$\tilde{\lambda}(i) = \frac{1}{l} \int_{a_i - \frac{1}{\mu}}^{b_i - \frac{1}{\mu}} \lambda(t) dt$
		Lag Max / Green et al. (2001)	$\tilde{\lambda}(i) = \max_{t \in [a_i - \frac{1}{\mu}, b_i - \frac{1}{\mu}]} \lambda(t)$
		Lag Mix /Green et al. (2001)	$\tilde{\lambda}(i) = \begin{cases} \frac{1}{l} \int_{a_i - \frac{1}{\mu}}^{b_i - \frac{1}{\mu}} \lambda(t)\, dt & \text{if } \frac{d\lambda(t)}{dt} > 0 \; \forall t \in [a_i - \frac{1}{\mu}, b_i - \frac{1}{\mu}] \\ \max_{t \in [a_i - \frac{1}{\mu}, b_i - \frac{1}{\mu}]} \lambda(t) & \text{otherwise} \end{cases}$

Table 3.4: Performance evaluation methods based on piecewise stationary models - continued

$l = 0$	$t = a_i = b_i$	PSA / Newell (1979)	$\tilde{\lambda}(i) = \lambda(t)$
		Lagged PSA / Green and Kolesar (1997)	$\tilde{\lambda}(i) = \lambda(t - \frac{1}{\mu})$
		ASA / Whitt (1991)	$\tilde{\lambda}(i) = \mu \int_{t-\frac{1}{\mu}}^{t} \lambda(\tau)\, d\tau$
		EAA / Thompson (1993)	$\tilde{\lambda}(i) = \mu \int_{t-\frac{1}{\mu}-\mathrm{E}[W^Q]}^{t-\mathrm{E}[W^Q]} \lambda(\tau)\, d\tau$
		RA / Pang and Whitt (2012b)	$\tilde{\lambda}(i) = \int_{0}^{\infty} \lambda(t-\tau)\delta e^{-\delta\tau}\, d\tau$

The **simple stationary approximation (SSA)** averages the system parameters over the complete time horizon. Green et al. (1991) apply this approach to systems with periodic time-dependent input parameters. The *simple peak epoch approximation (SPEA)* approximates the time-dependent performance based on the stationary performance by using the instantaneous peak input parameters. In a similar way, the *simple peak hour approximation (SPHA)* divides the time horizon into intervals and uses the input parameters of the peak interval as inputs in the performance calculation. Both concepts are used by Green and Kolesar (1995) for an $M(t)/M/c$ system and by Green and Kolesar (1998) for an $M(t)/M/\infty$ system, each with a periodic input arrival rate.

Shorter intervals are used by the **stationary independent period-by-period approximation (SIPP)**. Therein, the *SIPP Avg* considers the average over an interval; the *SIPP Max*, the maximum; and the *SIPP Mix*, a combination of the mean and maximum as inputs in the performance calculations. The lagged versions of the SIPP incorporate a time lag of one expected service time between the input parameters and the resulting system performance.

The interval length is set to $l = 0$ in the **pointwise stationary approximation (PSA)**. Here, the instantaneous parameter values serve as inputs in the performance calculation. Similar to the Lag SIPP, the *Lagged PSA* considers a time lag between the input parameters and the resulting performance values. The *average stationary approximation (ASA)* uses the mean value over the preceding interval $[t - \frac{1}{\mu}, t]$ as the input in the performance calculation at time t. In the similar *effective arrival rate approximation (EAA)*, the considered interval of input parameters is additionally shifted backward in time by the expected waiting time. The *recent approximation (RA)* calculates a weighted average of the parameters up to time t with weight factor δ. This approach is applied to infinite-server queues with dependencies among successive service times indicated by superscript D in the Kendall notation (Pang and Whitt, 2012a). A peak epoch analysis of a periodic $M(t)/M/c$ system is performed by Green and Kolesar (1997) with a *lagged PSA*, which considers the difference between the time of the peak value of the probability of delay when the standard PSA is applied and the time of the real peak value.

The main advantage of the approaches described in this subsection is their low computational complexity, especially when closed-form steady-state solutions exist for the considered system configuration (Ingolfsson et al., 2007). However, any transient behavior within an evaluation interval is neglected, which results in approximation errors, especially for highly utilized systems

in which long transient phases occur until the steady state is reached (Green and Kolesar, 1991). Further approximation errors result from the independent analysis of consecutive intervals, as a high number of waiting jobs at the end of one interval, e.g., has a substantial impact on the expected waiting time in the subsequent interval. The approaches cannot be used for the analysis of overloaded systems if no steady state exists (Jiménez and Koole, 2004). Whitt (1991) shows that the PSA is asymptotically correct for an $M(t)/M(t)/c$ system if the arrival and service rates increase with constant traffic intensity (compare with uniform acceleration in Section 3.3.3.2). The accuracy of the PSA for an $M(t)/M/c$ system with and without abandonments is analyzed by Steckley and Henderson (2007). Eick et al. (1993a) analyze the SSA and the PSA for infinite-server queueing systems with periodic arrival rate and compare their results with the exact solutions.

An overview of the literature on the evaluation approaches described in this subsection is presented in Table 3.5. The approaches are applicable to a wide range of system characteristics, including abandonments and heterogeneities. However, most analyzed systems consider Poisson arrivals, and many consider exponentially distributed service times.

3.3.2.2 Piecewise stationary models with linked periods

Similar to the approaches described in Section 3.3.2.1, the **stationary backlog-carryover approximation (SBC)** divides the overall time horizon into intervals and applies steady-state formulas. However, backlogs of non-served arrivals within an interval are carried over to the succeeding interval and are then considered in its performance evaluation.

Each interval is analyzed in two steps. In the first step, a loss system is assumed to calculate a backlog of unserved arrivals based on the lost jobs. These unserved arrivals are carried over to the successive interval. In the backlog calculation, the actual arrivals and the backlog of arrivals carried over from the previous interval are used as the input. In the second step, the performance measures are calculated based on the steady-state model of the corresponding waiting system. Here, a modified arrival rate is chosen such that the utilization of the waiting system equals the utilization of the loss system, as approximated in the first step.

Table 3.5: Approaches based on piecewise stationary models (independent periods)

Reference	Queueing system
Kolesar et al. (1975)	$M(t)$ / $M(t)$ / $c(t)$
Foote (1976)	$M(t)$ / M / $c(t)$
Rider (1976)	$M(t)$ / $M(t)$ / 1
Curry et al. (1978)	$M(t)$ / M / c
Newell (1979)	$M(t)$ / D / 1
Kolesar (1984)	$M(t)$ / M / c / K
Sze (1984)	$M(t)$ / G / c
Kwan et al. (1988)	$M(t)$ / M / $c(t)$
Agnihothri and Taylor (1991)	$M(t)$ / PH / $c(t)$
Green and Kolesar (1991)	$M(t)$ / M / c
Green et al. (1991)	$M(t)$ / M / c
Whitt (1991)	$M(t)$ / G / c
Deng et al. (1992)	$M(t)$ / $M(t)$ / $c(t)$
Andrews and Parsons (1993)	$M(t)$ / $M(t)$ / $c(t)$
Eick et al. (1993a)	$M(t)$ / G / ∞
Thompson (1993)	$M(t)$ / $M(t)$ / $c(t)$
Green and Kolesar (1995)	$M(t)$ / M / c
Choudhury et al. (1997)	$M(t)$ / $G(t)$ / 1
Green and Kolesar (1997)	$M(t)$ / M / c
Green and Kolesar (1998)	$M(t)$ / G / ∞
Kolesar and Green (1998)	$M(t)$ / M / c
Green et al. (2001)	$M(t)$ / M / $c(t)$
Ingolfsson et al. (2002)	$M(t)$ / M / $c(t)$ / K
Green et al. (2003)	$M(t)$ / M / $c(t)$
Koole and van der Sluis (2003)	$M(t)$ / M / $c(t)$
Dietz and Vaver (2006)	$M(t)$ / M / $c(t)$
Green et al. (2006)	$M(t)$ / M / $c(t)$
de Bruin et al. (2007)	$M(t)$ / M / ∞
Ingolfsson et al. (2007)	$M(t)$ / M / $c(t)$
Steckley and Henderson (2007)	$M(t)$ / M / c + M
Wall and Worthington (2007)	$M(t)$ / G / c
Atlason et al. (2008)	$M(t)$ / M / $c(t)$
Liu and Wein (2008)	$M(t)$ / M / c / K
Singer and Donoso (2008)	$M(t)$ / M / $c(t)$
	$M(t)$ / G / $c(t)$
Stolletz (2008a)	$M(t)$ / $M(t)$ / $c(t)$
Kuraya et al. (2009)	$M(t)$ / M / ∞
Manohar et al. (2009)	$M(t)$ / $G(t)$ / ∞
Zhang (2009)	$M(t)$ / M / c
	$M(t)$ / G / c
Ingolfsson et al. (2010)	$M(t)$ / M / $c(t)$
Dietz (2011)	$M(t)$ / $M(t)$ / $c(t)$ + M
Stolletz (2011)	$M(t)$ / G / $c(t)$
Pang and Whitt (2012a)	$G(t)^X$ / G^D / ∞
Pang and Whitt (2012b)	$G(t)$ / G^D / ∞
Chassioti et al. (2014)	$M(t,n)$ / G / c
Chen and Yang (2014)	$M(t)$ / G / c
Vanberkel et al. (2014)	$M(t)$ / G / ∞
Selinka et al. (2016)	$M(t)$ / M / c

The SBC is introduced by Stolletz (2008a) for an $M(t)/M(t)/c(t)$ system. Stolletz (2008b) extends the SBC to the analysis of $M(t)/G(t)/1$ systems. $M(t)/G/c(t)$ systems are considered by Stolletz (2011). Stolletz and Lagershausen (2013) analyze $G(t)/G/1/K$ systems. To improve the accuracy of the approximation, these authors use a variable interval length that depends on the utilization of the system. Selinka et al. (2016) extend the SBC to the analysis of a queueing system with two job classes, two server classes, and a routing decision on arrival.

To account for overload situations, the **coordinate transformation technique (CTT)** uses a model partially based on a deterministic fluid approximation (Section 3.3.3.2) that offers an accurate performance approximation for overloaded periods.

An interval's performance is calculated by using a transformation of a steady-state queueing formula. The transformation is chosen such that it converges to both the performance according to the steady-state formula for decreasing traffic intensities and the performance according to a deterministic fluid approximation for increasing traffic intensities. Such a transformation used in the analysis of an $M(t)/M(t)/1$ system is shown in Figure 3.2. As the performance at the end of an interval is used as the initial condition in the fluid approximation of the succeeding interval, the CTT integrates dependencies between successive intervals.

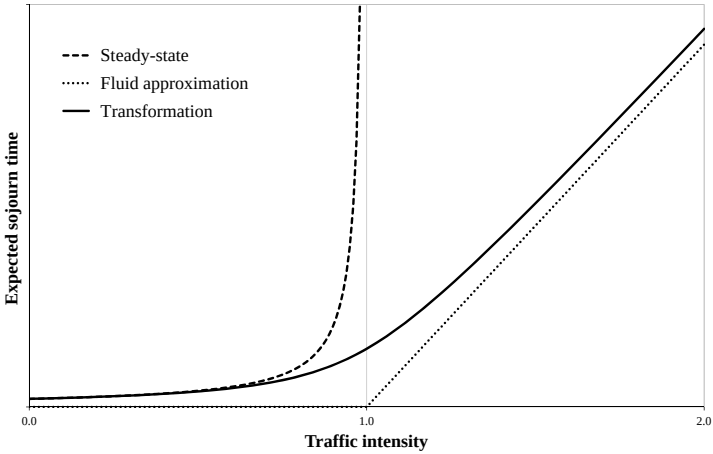


Figure 3.2: Transformation with $L^S = 0$ as initial condition (Kimber et al., 1977)

Kimber et al. (1977) introduce the CTT for an $M(t)/M(t)/1$ system. However, the shape of the time-dependent traffic intensity is restricted to a rectangular peak and adjacent-to-peak periods with a traffic intensity of zero. Kimber and Hollis (1978) extend this approach in analyzing peaks with non-zero adjacent-to-peak periods and considering general shapes of the peak traffic intensity. Catling (1977) analyzes an $M(t)/G/1$ system and allows for a general shape of the input arrival rate that is not restricted to a single peak. The CTT for $G(t)/G(t)/1$ systems with arbitrary input parameters is considered by Kimber and Daly (1986). Brilon and Wu (1990) derive a formula for the average queue length in an $M(t)/D/1$ system with a parabolic shape of the time-dependent arrival rate. Griffiths et al. (1991) expand the CTT to an $M(t)/G^{(0,s)}/1$ system with batch service up to a maximum of s jobs. However, in their version of the CTT, dependencies between successive time intervals are not considered. These dependencies are considered again by Holland and Griffiths (1999), who use the CTT to analyze the time-dependent performance of $M(t)/M^{(1,s)}/c$ systems.

Including dependencies between consecutive intervals, the SBC and the CTT take the transient behavior of a system's performance into consideration. Moreover, they can be applied to the performance evaluation of temporarily overloaded systems (Kimber and Hollis, 1978; Stolletz, 2008a). The characteristics of the analyzed systems are quite different (Table 3.6). However, all but one of the cited references consider systems with an infinite waiting room.

Table 3.6: Approaches based on piecewise stationary models (linked periods)

Reference	**Queueing system**
Catling (1977)	$M(t)$ / G / 1
Kimber et al. (1977)	$M(t)$ / $M(t)$ / 1
Kimber and Hollis (1978)	$M(t)$ / $M(t)$ / 1
Kimber and Daly (1986)	$G(t)$ / $G(t)$ / 1
Brilon and Wu (1990)	$M(t)$ / D / 1
Griffiths et al. (1991)	$M(t)$ / $G^{(0,s)}$ / 1
Holland and Griffiths (1999)	$M(t)$ / $M^{(1,s)}$ / c
Stolletz (2008a)	$M(t)$ / $M(t)$ / $c(t)$
Stolletz (2008b)	$M(t)$ / $G(t)$ / 1
Stolletz (2011)	$M(t)$ / G / $c(t)$
Chen et al. (2013c)	$M(t)$ / E_k / $c(t)$
Stolletz and Lagershausen (2013)	$G(t)$ / G / 1 / K
Selinka et al. (2016)	$M(t)$ / M / c

3.3.2.3 Piecewise transient models

The approaches described in this paragraph are **based on transient models (BOT)** that are used to analyze consecutive intervals with constant input parameters. The system state at the end of an interval serves as initial condition for the performance evaluation of the subsequent interval.

The transient solution of a queueing system with a finite waiting room is used by Upton and Tripathi (1982) to approximate the performance of an $M(t)/M/1$ system with an infinite waiting room. Choudhury et al. (1997) use numerical transform inversion and apply transient models to analyze an $M(t)/G(t)/1$ system. Parthasarathy and Sudhesh (2006) derive the exact transient solution for an $M/M/1$ system by using generating functions and then apply it to an $M(t)/M(t)/1$ system with piecewise constant input parameters. This approach is extended by Griffiths et al. (2008) to the case of Erlang-distributed service times. Duda (1986), Czachórski et al. (2009), and Czachórski et al. (2010) use the diffusion approximation for the transient performance evaluation (see also Section 3.3.3.2).

A common approximation technique for the transient analysis of Markovian queueing systems is the **uniformization/randomization (UR)** approximation. This approach analyzes the transient performance of a continuous-time Markov chain (CTMC) by transformation in a discrete-time Markov chain (DTMC).

The transition probability matrix $\mathbf{A}$ of the DTMC is derived by the uniformization of the generator matrix of the original CTMC. If the overall outgoing transition rates are identical for all states in the CTMC, the probability $g(j)$ for j transitions within one evaluation interval of the DTMC follows a Poisson distribution. Thus, self-transitions are included to unify the overall transition rates out of every state in the original CTMC. Then, the state probability vector $\mathbf{P}(i)$ at the end of interval i can be calculated according to Equation (3.6), where $\mathbf{A}^j$ denotes a j-times multiplication of the matrix $\mathbf{A}$ by itself. Here, the transitions within an interval are randomized according to the Poisson distribution mentioned above. Only a maximum of m possible transitions within one interval is considered to preserve computational tractability. The chosen value of m must be sufficiently large to achieve a reasonable approximation quality of the infinite number of possible state transitions

$$\mathbf{P}(i) = \sum_{j=0}^{m} g(j) \cdot \mathbf{P}(i-1) \cdot \mathbf{A}^j. \tag{3.6}$$

To account for non-stationary input parameters, the time-dependent generator matrix and the resulting time-dependent transition probability matrix must be considered.

Grassmann (1977a) reports the first study to use the concept of the uniformization/ randomization approximation. Here, the transient behavior of an $M/M/1$ system is analyzed, but only constant input parameters are considered. However, the applicability of the approach to the analysis of time-dependent systems is mentioned by Gross and Miller (1984) and assessed for general Markovian systems by Van Dijk (1992). Dormuth and Alfa (1997) apply the uniformization/randomization approach in the performance analysis of an $MAP(t)/PH(t)/1/K$ system. Furthermore, they extend the approach by incorporating an online adaptation of the length of the discretization intervals to improve the performance approximation. Flexible interval lengths are also included in the modification of Arns et al. (2010). Although their approach is applicable to general Markovian systems, it is applied to the analysis of an $M(t)/M(t)/1/K$ system in their numerical study. Creemers et al. (2014) apply the uniformization/ randomization approximation in the analysis of $PH(t)/PH(t)/c(t)+PH(t)$ systems with limited and unlimited waiting rooms to approximate queueing systems with general distributions.

A major advantage of the uniformization/randomization approach is its applicability to any Markovian queueing system. Furthermore, the complete time-dependent distribution of the state probabilities is derived (Gross and Miller, 1984). However, the approach is characterized by high computation times (Grassmann, 1977b; Ingolfsson et al., 2007). Table 3.7 shows that piecewise transient models can be used in the performance evaluation of a wide range of system configurations. Such models require only a tractable method for the transient analysis with arbitrary initial conditions.

Table 3.7: Approaches based on piecewise transient models

Reference	**Queueing system**
Upton and Tripathi (1982)	$M(t)$ / M / 1
Gross and Miller (1984)	$M(t)$ / $M(t)$ / $c(t)$ / K
Mok and Shanthikumar (1987)	$M(t)$ / M / $c(t)$ / $K + M$
Choudhury et al. (1997)	$M(t)$ / $G(t)$ / 1
Dormuth and Alfa (1997)	$MAP(t)$ / $PH(t)$ / 1 / K
Hebert and Dietz (1997)	$M(t)$ / $PH(t)$ / 1
Parthasarathy and Sudhesh (2006)	$M(t)$ / $M(t)$ / 1
Ingolfsson et al. (2007)	$M(t)$ / M / $c(t)$
Griffiths et al. (2008)	$M(t)$ / E_k / 1
Arns et al. (2010)	$M(t)$ / $M(t)$ / 1 / K
Ingolfsson et al. (2010)	$M(t)$ / M / $c(t)$
Creemers et al. (2014)	$G(t)$ / $G(t)$ / $c(t) + G(t)$
	$G(t)$ / $G(t)$ / $c(t)$ / $K + G(t)$

The underlying idea of the **discrete-time approach (DTA)** is to replace the continuous time with discrete points in time at which the system state is observed. The use of this approach leads to an approximation error if the system does not operate with time slots. The state probabilities for the next observation point are obtained by multiplying the state probability vector of the current observation point with a time-dependent transition probability matrix. The evolution of the system performance over time is then obtained through recursive vector matrix multiplications. In contrast to the UR, which discretizes a CTMC via uniformization, the DTA directly assumes that time is discrete and that only one transition per interval is possible. Depending on the queueing system and the length of the discretization interval, one transition accounts for multiple arrivals and/or multiple service completions.

Galliher and Wheeler (1958) introduce the basic idea for an $M(t)/D(t)/c(t)$ system. Setting the interval length equal to the service time is reported to work well if the deterministic service time is rather short compared with the time interval of interest. Otherwise, Minh (1978) suggests modifying the interval length and introducing auxiliary state variables for the remaining service time in addition to the number of jobs in the system. This concept is also used by Alfa (1982), Omosigho and Worthington (1985), Omosigho and Worthington (1988), Brahimi and Worthington (1991a), Brahimi and Worthington (1991b), Mejía-Téllez and Worthington (1994), and Chassioti and Worthington (2004) to model a general service time distribution. Regarding the arrival process, these models require only that the number of arrivals in each discretization interval be an independent random variable. This assumption allows for time-dependent Poisson processes, potentially with batch arrivals, and $D^{X(t)}$ arrival processes where the inter-arrival time is equal to the interval length and where the batch size distribution X is time-dependent. Although Powell and Simão (1986) call their approach numerical simulation, they use the same technique of auxiliary variables in analyzing a discrete-time $M(t)^X/G(t)^Y/1/K$ bulk queue with a random number of Y jobs that can be served simultaneously under different dispatching rules for the server. Kahraman and Gosavi (2011) focus on stranded customers, i.e., unserved customers that remain in the queue directly after the visit of the server, and consider different dispatching rules. Alfa (1990) and Alfa and Chen (1991) avoid the use of computationally expensive auxiliary variables by approximating the probability of an empty system by using the Maximum Entropy Principle. The expected queue length is then obtained based on the probability of an empty system. Using an approach unlike the approaches discussed so far, Moore (1975) observes the system state at the departure time of the jobs

from the queue. At these observation points, the expected queue length of an $M(t)^{X(t)}/E_k/1$ system is computed. Worthington and Wall (1999) provide a survey of most of the existing DTAs for systems with time-dependent Markovian arrival processes, generally distributed service times, and single or multiple servers.

Table 3.8: Discrete-time approaches (DTA)

Reference	**Queueing system**			
Galliher and Wheeler (1958)	$M(t)$	/ $D(t)$	/ $c(t)$	
Koopman (1972)	$M(t)$	/ D	/ 1	/ K
Moore (1975)	$M(t)^{X(t)}$	/ E_k	/ 1	
Minh (1978)	$M(t)^{X}$	/ G	/ 1	
Alfa (1982)	$M(t)^{X}$	/ G^{Y}	/ 1	
	$M(t)^{X}$	/ G^{Y}	/ 1	/ K
Upton and Tripathi (1982)	$M(t)$	/ M	/ 1	
Omosigho and Worthington (1985)	$M(t)^{X(t)}$	/ G	/ 1	/ K
	$D^{X(t)}$	/ G	/ 1	/ K
Powell and Simão (1986)	$M(t)^{X}$	/ $G(t)^{Y}$	/ 1	/ K
	$D^{X(t)}$	/ $G(t)^{Y}$	/ 1	/ K
Omosigho and Worthington (1988)	$M(t)^{X(t)}$	/ G	/ 1	/ K
	$D^{X(t)}$	/ G	/ 1	/ K
Alfa (1990)	$M(t)$	/ D	/ 1	
Brilon and Wu (1990)	$M(t)$	/ D	/ 1	
	$D^{X(t)}$	/ D	/ 1	
Alfa and Chen (1991)	$M(t)$	/ G	/ 1	
Brahimi and Worthington (1991a)	$D^{X(t)}$	/ G	/ 1	
Brahimi and Worthington (1991b)	$M(t)$	/ G	/ c	/ K
	$D^{X(t)}$	/ G	/ c	/ K
Lackman et al. (1992)	$M(t)$	/ D	/ 1	+ D
	$M(t)$	/ D	/ 1	+ D / NPPrio
Mejía-Téllez and Worthington (1994)	$M(t)$	/ $G^{(0,s)}$	/ 1	
Daniel (1995)	$M(t)$	/ D	/ c	/ K
Bennett and Worthington (1998)	$D^{X(t)}$	/ G	/ 1	
Daniel and Pahwa (2000)	$M(t)$	/ D	/ c	/ K
Chassioti and Worthington (2004)	$M(t)$	/ G	/ $c(t)$	
	$M(t)$	/ G	/ $c(t)$	/ K
Wall and Worthington (2007)	$M(t)$	/ G	/ c	
	$D^{X(t)}$	/ G	/ c	
Alfa and Margolius (2008)	$M(t)$	/ $M(t)$	/ $c(t)$	
Daniel and Harback (2008)	$M(t)$	/ D	/ c	/ K
Daniel and Harback (2009)	$M(t)$	/ D	/ c	/ K
Viti and van Zuylen (2009)	$M(t)$	/ $D(t)$	/ 1	/ K
Viti and van Zuylen (2010)	$M(t)$	/ $D(t)$	/ 1	/ K
Kahraman and Gosavi (2011)	$M(t)$	/ G^{Y}	/ 1	
Blumberg-Nitzani and Bar-Gera (2014)	$D^{X(t)}$	/ D	/ 1	

The main advantage of the DTA is its flexibility with respect to the service time distribution (see Table 3.8) and the derivation of the time-dependent probability distribution for the complete state space. Hence, the approach allows one to obtain quantiles of the number of jobs in the system and the distribution of the virtual waiting time (Minh, 1978; Wall and Worthington, 2007). The major disadvantage of the DTA is the rapidly growing state space with an increasing waiting room and the need for an additional auxiliary variable for every additional server.

3.3.3 Approaches based on modified system characteristics

3.3.3.1 Modified number of servers

Although most real systems have a finite number of servers, the explicit results for infinite-server systems gain relevance in approximation approaches. An overview of the literature on these approaches is presented in Table 3.9.

In queueing systems with an infinite number of parallel servers, the time-dependent number of busy servers $L^B(t)$ is comparatively easy to determine (see Section 3.3.1). Thus, this number can be used to estimate performance measures of systems with a finite number of servers. Such an **infinite-server approximation (INFSA)** is applied by Jennings et al. (1996) to analyze the probability of waiting in a $G(t)/G(t)/c(t)$ system. They use a normal approximation to estimate the distribution of busy servers in the infinite-server system and apply their results to derive a staffing formula. Feldman et al. (2008) use such an INFSA to analyze $M(t)/G/c(t) + G$ systems. Liu and Whitt (2012a) derive a *delayed-infinite-server (DIS)* approximation model that is applied in a staffing algorithm for an $M(t)/G/c(t) + G$ system by decomposing the original system into two infinite-server systems – one representing the waiting jobs including abandonments and the other representing the jobs in service.

The key idea of the **modified offered load approach (MOL)** is the approximation of the time-dependent offered load in a queueing system by the number of busy servers in the corresponding system with an infinite number of servers. Based on this relation, a modified arrival rate is derived, and this arrival rate is then used to calculate the system's performance for every point in time by using steady-state models. In doing so, the MOL takes advantage of the known solution of the DE describing the number of jobs in an infinite-server system (see Section 3.3.1).

For an $M(t)/M(t)/\infty$ queueing system, the expected number $L^S(t) = L^B(t)$ of busy servers at time t is given by Equation (3.5). The modified arrival rate $\lambda^{MAR}(t) = \mathrm{E}[L^B(t)] \cdot \mu(t)$ is chosen such that the expected number of busy servers in the stationary $M/M/c$ system equals the expected number of busy servers in the time-dependent infinite-server system.

Jagerman (1975) introduces the MOL to analyze the blocking probability in $M(t)/M/c/c$ systems. The applicability of the MOL to the analysis of more general queueing systems with waiting rooms is mentioned by Jennings et al. (1996). Massey and Whitt (1997) apply the MOL to evaluate an $M(t)/M/c$ system and compare their results with the numerical solution of the CKEs. Feldman et al. (2008) extend the MOL to analyze $M(t)/G/c(t) + G$ systems. In addition to the DIS approach mentioned above, Liu and Whitt (2012a) develop the DIS-MOL, which is an extension in which the offered load in the queue, representing the jobs in service, is used as an input for a stationary $M/G/c + G$ system. Using an approach similar to the MOL, Yom-Tov and Mandelbaum (2014) use the time-dependent number of busy servers in an infinite-server system as the input in their staffing algorithm for an $M(t)/G/c(t)$ system.

In contrast to the methods described in Section 3.3.2.1, the MOL does not analyze the performance of intervals independently. Nevertheless, as the derivation of the modified arrival rate corresponds to the calculation of the exponentially weighted moving average over the period $[-\infty, t]$, the MOL is similar to the EAA and the RA described in Section 3.3.2.1, which also use a moving average as the input in the performance calculations according to steady-state formulas (Ingolfsson et al., 2007). The transient behavior and dependencies between intervals are taken into account in the derivation of the modified arrival rate. Thus, the MOL has a structure similar to the SBC (Section 3.3.2.2). Owing to the application of infinite-server systems, the approximation quality of the MOL renders this approach more suitable for systems with a decreasing probability of waiting, i.e., an increasing number of servers or decreasing traffic intensity (Jennings and Massey, 1997; Massey and Whitt, 1997). Additionally, the approximation quality decreases with increasing rate of change in the input arrival rate (Jagerman, 1975). Jennings and Massey (1997) show that the idea of the MOL is applicable to any time-dependent system if its state space is a subset of the state space of a larger system for which the performance is simpler to evaluate. Massey (2002) provides an overview through 2002 of the literature on approaches that use the explicit solution of infinite-server systems.

Table 3.9: Infinite-server approximations (INFSA)

Reference	Queueing system
Sze (1984)	$M(t)$ / G / c
Jennings et al. (1996)	$G(t)$ / $G(t)$ / $c(t)$
Green and Kolesar (1997)	$M(t)$ / M / c
Massey and Whitt (1997)	$M(t)$ / M / c
Ingolfsson et al. (2007)	$M(t)$ / M / $c(t)$
Feldman et al. (2008)	$M(t)$ / G / $c(t)$ + G
Liu and Wein (2008)	$M(t)$ / M / c / K / PPrio
Liu and Whitt (2012a)	$M(t)$ / G / $c(t)$ + G
Yom-Tov and Mandelbaum (2014)	$M(t)$ / G / $c(t)$

3.3.3.2 Modified job characteristics

The fluid approximation, the pointwise stationary fluid flow approximation, and the diffusion approximation replace discrete jobs with a continuum. These approaches differ in the way that they consider stochasticity. The fourth approach, the uniform acceleration, modifies the arrival and service rate of the jobs.

The key idea of the **fluid approximation (FLUID)** is to replace randomly arriving discrete jobs with a deterministic continuum. This continuum can be interpreted as a fluid that flows with rate $\lambda(t)$ into a reservoir. The service process is approximated by a deterministic outflow from the reservoir. The level of fluid in the reservoir then serves as an approximation for the number of jobs in the system. The derivative with respect to time of the fluid level for a queueing system with c parallel servers without a queue length limit is given by

$$\mathrm{E}[L^S(t)]' = \begin{cases} 0 & \text{if } \mathrm{E}[L^S(t)] = 0 \wedge \lambda(t) \leq \mu(t) \cdot c, \\ \lambda(t) - \mu(t) \cdot \min\{c; \mathrm{E}[L^S(t)]\} & \text{otherwise.} \end{cases} \tag{3.7}$$

The fluid approximation represents one of the first approaches for the analysis of time-dependent queueing systems. It is described in the book of Newell (1971) as engineering approach for the performance evaluation of systems for which temporary overload rather than randomness is the primary reason for the existence of queues. A direct application of the fluid approximation can be found in Horonjeff (1969), Koopman (1972), Wirasinghe and Shehata

(1988), and Janic (2009). It is also used by Harrison and Zeevi (2005) and Swaroop et al. (2012) within optimization approaches. Mandelbaum et al. (2002) investigate a fluid approximation for an $M(t)/M(t)/c(t)+M$(t) system with retrials. An adjusted fluid approximation for this system is proposed by Ko and Gautam (2013). Aguir et al. (2004) modify the queueing model and include the effect of balking but assume a time-invariant service rate. By choosing a balking probability of 0 if $L^S < K$ and 1 otherwise, their model can also approximate systems with a finite waiting room. Whitt (1999) develops the fluid approximation for an $M(t)/G/c$/PPrio system by reducing the service rate of a given class by the demand of all higher classes. Ridley et al. (2004) derive the fluid approximation for a two-class $M(t)/M/c$/PPrio system that is supported by a limit theorem. Ko and Gautam (2010) propose a Gaussian-based adjustment of the fluid approximation for a system with servers that switch between an active and an inactive pool.

Hampshire et al. (2009) combine the fluid approximation with the MOL approach (see Section 3.3.3.1) to analyze the abandonments and blocking probability in an $M(t)/M/c(t)/K(t)+M$ system. Liu and Whitt (2012b) introduce the fluid approximation for a $G(t)/G/c(t)+G$ system. They separately track the fluid in the queue and on the servers. For both parts, two-parameter functions $L^Q(t,y)$ and $L^S(t,y)$ describe the amount of fluid at time t that has spent at most y time units in the queue and on the server, respectively. Thus, abandonments can be treated as a proportion of the fluid that leaves the queue without being served depending on y. The authors develop an algorithm that generates approximations for $\mathrm{E}[L^Q(t)]$ and $\mathrm{E}[L^S(t)]$, as well as for the expected head of line and virtual waiting time. Liu and Whitt (2011) use this modeling approach to establish an asymptotic loss of memory property for a $G(t)/M(t)/c(t)+G(t)$ fluid approximation, i.e., the performance of the queue becomes asymptotically independent of the initial condition as time proceeds.

Apart from the use of the fluid approximation for performance evaluation, another literature stream establishes fluid limits for stochastic queueing systems. The existence of such fluid limits supports the use of the fluid approximation, particularly under heavy traffic. For a more in-depth discussion on fluid limits and their derivation, see Jiménez and Koole (2004), Liu and Whitt (2014), and the references therein. As Table 3.10 shows, the fluid approximation is often used for queueing systems with stochastic inter-arrival and service times. However, for periods of persistent underload, the fluid approximation predicts an empty system since it neglects randomness as a reason for the occurrence of queues. Pender (2014b) notes that a deterministic surrogate distribution

also leads to a fluid approximation. The fluid approximation gains additional relevance as integral part of other analytical approaches, namely, in the CTT (Section 3.3.2.2) and in the pointwise stationary fluid flow approximation, which is described in the next paragraph.

The **pointwise stationary fluid flow approximation (PSFFA)** combines the deterministic fluid approximation with steady-state queueing formulas to integrate stochasticity. The fluid flow described by Equation (3.7) is modified such that the outflow from the system depends on the server utilization. The utilization is approximated by the inverse of stationary queueing formulas such that the utilization becomes a function of the expected number of jobs in the system $\mathrm{E}[U(t)] = g^{-1}(\mathrm{E}[L^S(t)], c)$. All parameters are assumed to be constant, and the queueing system is assumed to be in the steady state at time t, i.e., pointwise stationary (Section 3.3.2.1). The resulting DE (3.8) can be integrated numerically to obtain the expected number of jobs in the system over time

$$\mathrm{E}[L^S(t)]' = \lambda(t) - \mu(t) \cdot c \cdot g^{-1}(\mathrm{E}[L^S(t)], c). \tag{3.8}$$

A finite waiting room causes blocking and reduces the effective arrival rate. Consequently, for finite waiting rooms, an additional function that relates the blocking probability to the expected number of jobs in the system is required. Chen et al. (2011) note that Equation (3.8) can be used directly for queueing systems with a finite waiting room by solving it subject to the constraint $\mathrm{E}[L^S(t)] \leq K$.

Agnew (1976) reports the first attempt to relate the outflow of a fluid queue to the expected number of jobs in the system. He derives general properties of the function $\mathrm{E}[U] = g^{-1}(\mathrm{E}[L^S(t)], 1)$ and notes that this function can be either determined through statistical analysis from real systems or determined analytically. Tipper and Sundareshan (1990) analyze a heterogeneous $M(t)/M/1$ system where arrivals originate from multiple independent sources. They introduce a DE similar to Equation (3.8) for the total number of jobs in the system and one additional equation for each job class. Coining term PSFFA, Wang et al. (1996) provide approximations for $M(t)/G/1$, $G(t)/G/1$, and $M(t)/M/1/K$ systems. Chen et al. (2013c) invert the approximation of Cosmetatos (1976) numerically with a bisection method and, thus, extend the approach to a multi-server $M(t)/E_k/c(t)$ system. Based on a data set, Xu et al. (2014) use polynomial curve fitting to derive $g^{-1}(\mathrm{E}[L^S(t)], 1)$.

The PSFFA delivers results only for the expected value of the number of jobs in the system. Higher moments remain intractable. $\mathrm{E}[L^S(t)]$ converges to exact steady-state values for constant parameters if the inverse function $g^{-1}(\mathrm{E}[L^S(t)], c)$ is exact. However, Tipper and Sundareshan (1990) and Wang et al. (1996) report that the PSFFA reaches the steady state too rapidly. This leads to an overestimation of peaks and an underestimation of valleys for quickly varying input rates. Although the PSFFA and the SDA (Section 3.3.1) are independently developed approaches, they share the same MDE as the starting point, i.e., for $c = 1$, Equation (3.2) simplifies to Equation (3.8) (Filipiak, 1984). The approaches differ in how they obtain the unknown probability $1 - P_0 = \mathrm{E}[U]$.

The **diffusion approximation (DIFF)** replaces the mathematically intractable discrete stochastic process $L^S(t)$ with a continuous stochastic process $\mathcal{X}(t)$ which is known as Brownian motion. The incremental changes $d\mathcal{X}(t) = \mathcal{X}(t + dt) - \mathcal{X}(t)$ are normally distributed with infinitesimal mean bdt and infinitesimal variance adt. For a non-empty system, the stochastic process $\mathcal{X}(t)$ is described by diffusion equation (3.9)

$$\frac{\partial f(x,t)}{\partial t} = \frac{a(t)}{2} \cdot \frac{\partial^2 f(x,t)}{\partial x^2} - b(t) \cdot \frac{\partial f(x,t)}{\partial x}. \tag{3.9}$$

Equation (3.9) is also known as the Kolmogorov or Fokker-Planck equation with probability density function $f(x, t)$ and x as a continuous approximation of the number of jobs in the system. Depending on $a(t)$ and $b(t)$ as well as on the boundary conditions, explicit solutions of the partial differential equation (3.9) exist; otherwise, it must be solved numerically. The diffusion approximation goes along with three key modeling decisions. (i) Depending on the system characteristics, $a(t)$ and $b(t)$ must be chosen. They either are a function of time (Newell, 1968a) or are assumed to be piecewise constant, and transient solutions are combined as described in Section 3.3.2.3 (Duda, 1986). (ii) Boundary conditions must be imposed to model the behavior of the stochastic process if the system is empty or, if applicable, if it reaches its waiting room limit K. For heavy traffic situations, Equation (3.9) is typically solved subject to boundary conditions. This leads to a reflected Brownian motion, i.e., trajectories of $\mathcal{X}(t)$ do not spend time at the boundary but are directly reflected. Such a condition does not work for underload situations, as idle times of the server must be taken into account. Thus, elementary return barriers are imposed, which ensure that $\mathcal{X}(t)$ stays in a boundary state for a certain period of time according to a stochastic holding time distribution. (iii) Only the moments of $L^S(t)$ may be directly approximated by their

equivalents of $\mathcal{X}(t)$. To obtain the state probabilities $P_n(t)$, the continuous density function $f(x,t)$ must be re-discretized (Duda, 1986).

In his pioneering works, Newell (1968a,b,c) proposes the diffusion approximation for $G(t)/G/1$ systems. He considers the case of a rush hour caused by an increasing arrival rate that eventually exceeds the service rate and then returns to values below the service rate. Knessl (2000) provides an exact solution of the diffusion process for $\rho \approx 1$ with an initially empty queue, which is generalized to arbitrary initial conditions by Knessl and Yang (2001). Both models assume that $\rho(t)$ either is linear in t or increases in a single step. Giorno et al. (1987) propose the diffusion approximation for the queue length distribution of the $M(t,n)/M(t,n)/1$ system. In their study, the arrival and service rates are time- and state-dependent such that they increase with the number of jobs n in the system. Their findings resemble the results of the special case discussed by Clarke (1956). Di Crescenzo and Nobile (1995) also analyze an $M(t,n)/M(t,n)/1$ system but use a more general arrival rate function that includes the model by Giorno et al. (1987) as a special case. Ko and Gautam (2010) obtain an adjusted diffusion model by using an adjusted fluid approximation. Mandelbaum et al. (2002) provide numerical results for the diffusion approximation of an $M(t)/M(t)/c(t)+M(t)$ system with retrials, which is based on a limit theorem established by Mandelbaum et al. (1998). The adjusted version of the limit theorem by Ko and Gautam (2013) improves the approximation quality of the diffusion approximation for systems with a small number of servers and for $\mathrm{E}[L^S(t)]$ close to $c(t)$. Massey and Pender (2013) show that their SDA is equivalent to the approach of Ko and Gautam (2013). Filipiak (1983) suggests moving the reflecting barrier from 0 to -1 to approximate underload situations in an $M(t)/M(t)/1$ system. Following the idea of a piecewise transient analysis (Section 3.3.2.3), Duda (1986) proposes an elementary return barrier diffusion approximation with Coxian distributed holding times at the return barrier for the transient solution of a $G/G/1$ system. Czachórski et al. (2009) and Czachórski et al. (2010) introduce another elementary return barrier with exponentially distributed holding times to model a finite waiting room for a $G/G/1/K/\mathrm{PPrio}$ system with multiple job classes and a simple $G/G/1/K$ system, respectively. Kimura (2004) provides a limited survey with respect to time-dependent and steady-state diffusion models.

The diffusion approximation can be used for non-Markovian systems since $a(t)$ and $b(t)$ depend on the means and variances of the arrival and service processes (Table 3.10). In addition, the use of the diffusion approximation for such systems results in an approximation of the complete probability dis-

tribution of $L^S(t)$. Further, it is rigorously supported by limit theorems and results of the uniform acceleration technique, which is described in detail in the next paragraph.

The **uniform acceleration (UA)** technique simultaneously increases the arrival and service rate such that their ratio remains fixed. UA may be regarded as the non-stationary analogue to steady-state analysis (Massey, 2002). Similar to the INFSA, which utilizes the tractability of infinite-server systems, the derived scaled queueing systems allow for an enhanced analytical understanding of the time-dependent behavior.

Keller (1982) reports the first attempt to use this scaling. Starting with the discrete stochastic process, he provides an attempt to rigorously derive Newell's diffusion approximation. Massey (1985) coins the term UA and proposes $\rho^*(t_0, t) = \sup_{0 \leq t_0 \leq t} \int_{t_0}^{t} \lambda(\tau) d\tau / \int_{t_0}^{t} \mu(\tau) d\tau$ as a modified parameter to differentiate between underloaded and overloaded queues. The results are refined by Yin and Zhang (2002). Mandelbaum and Massey (1995) apply UA to the sample path of an $M(t)/M(t)/1$ system and obtain an asymptotic expansion. Yang and Knessl (1997) correct the results of Keller (1982) and further extend them to general service processes. Flick and Liao (2010) extend the approach of Massey (1985) to queueing systems with more than one server. The case of finite waiting rooms is treated by Tan et al. (2013).

The results of the UA serve as rigorous justification for the PSA, the fluid approximation, and the diffusion approximation. These results suggest that the PSA works well for underloaded queues (Flick and Liao, 2010) and that overloaded queues are well approximated by the fluid approximation (Mandelbaum and Massey, 1995). In addition, these findings substantiate the core idea of the CTT.

Table 3.10: Approximations based on modified job characteristics

Reference	**Queueing system**			
Fluid approximation (FLUID)				
Newell (1968b)	$G(t)$	/ G	/ 1	
Gaver (1969)	$D(t)$	/ $D(t)$	/ 1	
Horonjeff (1969)	$D(t)$	/ $D(t)$	/ 1	
Paullin and Horonjeff (1969)	$D(t)$	/ D	/ c	
Newell (1971)	$D(t)$	/ $D(t)$	/ 1	
Koopman (1972)	$D(t)$	/ D	/ 1	/ K
Newell (1979)	$D(t)$	/ D	/ 1	
de Neufville and Grillot (1982)	$D(t)$	/ D	/ 1	
Wirasinghe and Shehata (1988)	$D(t)$	/ D	/ 1	
Jung and Lee (1989a)	$G(t)$	/ G	/ $c(t)$	
Wirasinghe and Bandara (1990)	$D(t)$	/ D	/ c	

Table 3.10: Approximations based on modified job characteristics - continued

Whitt (1999)	$M(t)$	/ G	/ c / PPrio
Mandelbaum et al. (2002)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$
Aguir et al. (2004)	$M(t)$	/ M	/ $c(t)$ + M
	$M(t)$	/ M	/ $c(t)$ / K + M
Jiménez and Koole (2004)	$M(t)$	/ M	/ c
Ridley et al. (2004)	$M(t)$	/ M	/ c / PPrio
de Barros and Tomber (2007)	$D(t)$	/ D	/ 1
Kuwahara (2007)	$D(t)$	/ D	/ 1
Hampshire et al. (2009)	$M(t)$	/ M	/ $c(t)$ / $K(t)$ +M
Janic (2009)	$D(t)$	/ $D(t)$	/ 1
Viti and van Zuylen (2009)	$M(t)$	/ $D(t)$	/ 1 / K
Bertsimas and Doan (2010)	$M(t)$	/ M	/ $c(t)$ + M
Chen and Yang (2010)	$D(t)$	/ D	/ 1
Ko and Gautam (2010)	$M(t)$	/ M	/ $c(t)$
Viti and van Zuylen (2010)	$M(t)$	/ $D(t)$	/ 1 / K
Liu and Whitt (2011)	$G(t)$	/ $M(t)$	/ $c(t)$ + $G(t)$
Stolletz (2011)	$M(t)$	/ G	/ $c(t)$
Liu and Whitt (2012b)	$G(t)$	/ G	/ $c(t)$ + G
Swaroop et al. (2012)	$D(t)$	/ D	/ 1
Ko and Gautam (2013)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$
Massey and Pender (2013)	$M(t)$	/ M	/ $c(t)$ + M
Chen and Yang (2014)	$M(t)$	/ G	/ c
Pender (2014a)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$
Pointwise stationary fluid flow approximation (PSFFA)			
Agnew (1976)	$G(t)$	/ G	/ 1
Filipiak (1984)	$M(t)$	/ M	/ 1
Tipper and Sundareshan (1990)	$M(t)$	/ M	/ 1
Wang et al. (1996)	$M(t)$	/ G	/ 1
	$G(t)$	/ G	/ 1
	$M(t)$	/ M	/ 1 / K
Chen et al. (2011)	$M(t)$	/ G	/ 1
Chen et al. (2013a)	$M(t)$	/ E_k	/ $c(t)$
Chen et al. (2013b)	$M(t)$	/ E_k	/ $c(t)$
Chen et al. (2013c)	$M(t)$	/ E_k	/ $c(t)$
Yang et al. (2013)	$M(t)$	/ E_k	/ $c(t)$
Chen and Yang (2014)	$M(t)$	/ G	/ c
Xu et al. (2014)	$D(t)$	/ D	/ 1
Diffusion approximation (DIFF)			
Newell (1968a)	$G(t)$	/ G	/ 1
Newell (1968b)	$G(t)$	/ G	/ 1
Newell (1968c)	$G(t)$	/ G	/ 1
Keller (1982)	$M(t)$	/ $M(t)$	/ 1
Filipiak (1983)	$M(t)$	/ $M(t)$	/ 1
Duda (1986)	$G(t)$	/ $G(t)$	/ 1
Giorno et al. (1987)	$M(t,n)$	/ $M(t,n)$	/ 1
Jung and Lee (1989a)	$G(t)$	/ G	/ $c(t)$ / K
Di Crescenzo and Nobile (1995)	$M(t,n)$	/ $M(t,n)$	/ 1
Knessl (2000)	$G(t)$	/ $G(t)$	/ 1
Knessl and Yang (2001)	$G(t)$	/ $G(t)$	/ 1
Mandelbaum et al. (2002)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$
Janic (2005)	$G(t)$	/ $G(t)$	/ 1
Czachórski et al. (2009)	$G(t)$	/ G	/ 1 / K / PPrio
Czachórski et al. (2010)	$G(t)$	/ G	/ 1 / K
Ko and Gautam (2010)	$M(t)$	/ M	/ $c(t)$
Ko and Gautam (2013)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$

Table 3.10: Approximations based on modified job characteristics - continued

Lovell et al. (2013)	$G(t)$	/ $G(t)$	/ 1	/ K
Massey and Pender (2013)	$M(t)$	/ M	/ $c(t)$ + M	
Pender (2014a)	$M(t)$	/ $M(t)$	/ $c(t)$ + $M(t)$	
Uniform acceleration (UA)				
Keller (1982)	$M(t)$	/ $M(t)$	/ 1	
Massey (1985)	$M(t)$	/ $M(t)$	/ 1	
Mandelbaum and Massey (1995)	$M(t)$	/ $M(t)$	/ 1	
Yang and Knessl (1997)	$M(t)$	/ G	/ 1	
Yin and Zhang (2002)	$M(t)$	/ $M(t)$	/ 1	
Flick and Liao (2010)	$M(t)$	/ $M(t)$	/ c	
Tan et al. (2013)	$M(t)$	/ M	/ 1	/ K

3.4 Methodological relations and numerical comparisons

The classification scheme introduced in Section 3.2 groups approaches that share a common idea in their analysis. In addition, we discuss links between approaches in Section 3.3. These links are summarized in Figure 3.3, which reveals that there are links between approaches not only within classification categories but also beyond category boundaries.

Besides the identified methodological links, some approaches are compared numerically in the literature. References that include numerical studies comparing two or more approaches for single-, multi-, and infinite-server systems are listed in Table 3.11.

The majority of these numerical studies focus on multi-server systems with Markovian properties. Further, some studies include general distributions, abandonments, or heterogeneities. From a methodological point of view, it becomes apparent that a popular benchmark is the numerical solution of the CKEs, as the approximation error then originates only from the numerical solution scheme. Ingolfsson et al. (2007) compare approaches from all three main categories of our classification scheme for an $M(t)/M/c(t)$ system. Apart from that study, most studies compare approaches within a single category. The quality of approximation approaches strongly depends on the system parameters. Nevertheless, some conclusions with respect to the applicability of a subset of the discussed approaches are provided by Stolletz (2008b) and Chen et al. (2013c).

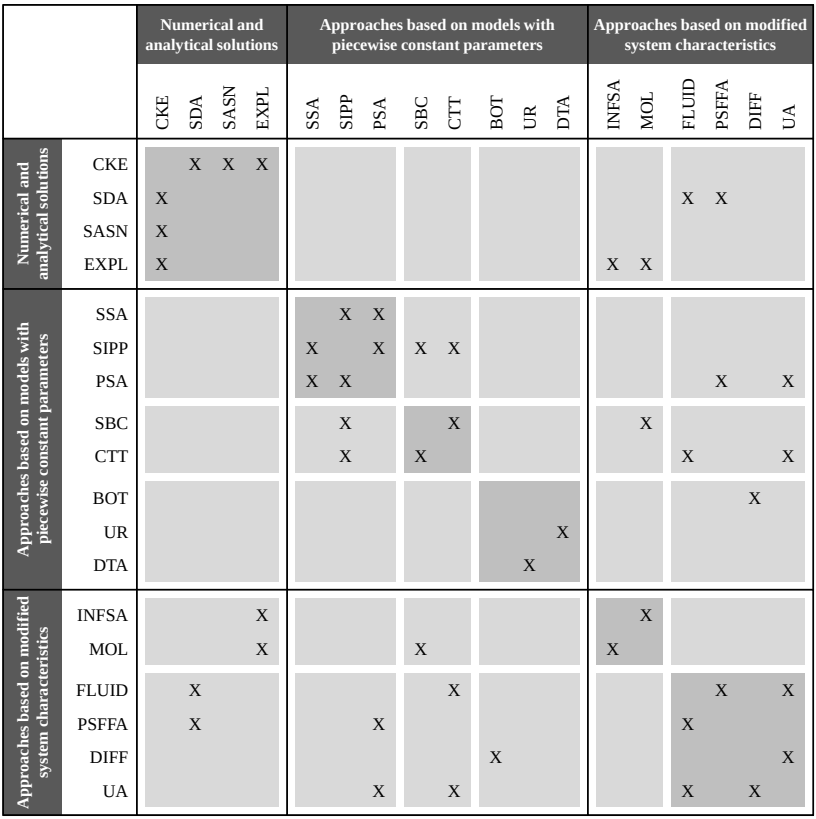


Figure 3.3: Methodological links between approaches for the performance evaluation of time-dependent systems

Table 3.11: Numerical comparisons of time-dependent queueing systems

	Reference	Compared methods	Queueing system	Performance measures
$c=1$	Leese and Boyd (1966)	CKE, SASN	$M(t)/M/1$	$L^Q(t)$
	Newell (1968b)	DIFF, FLUID	$G(t)/G/1$	E$[L^S(t)]$
	Koopman (1972)	CKE, DTA, FLUID	$M(t)/G(t)/1/K$	E$[L^Q(t)]$
	Rider (1976)	CKE, SDA, PSA	$M(t)/M(t)/1$	E$[L^S(t)]$, $P_0(t)$
	Upton and Tripathi (1982)	BOT, DTA	$M(t)/M/1$	E$[L^S(t)]$
	Ong and Taaffe (1988)	CKE, SDA	$PH(t)/PH(t)/1/K$	E$[L^S(t)]$, SD$[L^S(t)]$
	Taaffe and Clark (1988)	CKE, SDA	$M(t)/M/1/K$/NPPrio	E$[L^S(t)]$
	Brilon and Wu (1990)	CTT, DTA	$M(t)/D/1$	E$[L^Q(t)]$
	Tipper and Sundareshan (1990)	CKE, PSFFA	$M(t)/M/1$	E$[L^S(t)]$
	Choudhury et al. (1997)	BOT, SIPP	$M(t)/G(t)/1$	E[Workload(t)]
	Czachórski et al. (2009)	CKE, DIFF	$M(t)/M/1/K$/PPrio	E$[L^S(t)]$, $P_0(t)$, $P_K(t)$
	Viti and van Zuylen (2009)	DTA, FLUID	$M(t)/D(t)/1/K$	E$[L^S(t)]$, SD$[L^S(t)]$
	Viti and van Zuylen (2010)	DTA, FLUID	$M(t)/D(t)/1/K$	E$[L^S(t)]$, SD$[L^S(t)]$
$1<c<\infty$	Rothkopf and Oren (1979)	CKE, SDA	$M(t)/M/c$	E$[L^S(t)]$, SD$[L^S(t)]$
	Clark (1981)	CKE, SDA	$M(t)/M/c$	E$[L^S(t)]$, Var$[L^S(t)]$, E$[W^Q(t)]$
	Sze (1984)	SIPP, MOL	$M(t)/G/c$	E$[W^Q(t)]$
	Taaffe and Ong (1987)	CKE, SDA	$PH(t)/M(t)/c/k$	E$[L^S(t)]$, SD$[L^S(t)]$
	Jung and Lee (1989a)	DIFF, FLUID	$G(t)/G/c(t)$	E$[L^S(t)]$
	Green and Kolesar (1991)	CKE, PSA, SSA	$M(t)/M/c$	E$[L^Q(t)]$, $P_w(t)$
	Green et al. (1991)	CKE, SSA	$M(t)/M/c$	E$[L^Q(t)]$, E$[W^Q(t)]$, $P_w(t)$
	Green and Kolesar (1995)	CKE, PSA, SPHA,	$M(t)/M/c$	$P_w(t)$, E$[W^Q(t)]$
	Green and Kolesar (1997)	CKE, INFSA, Lagged PSA, SPEA	$M(t)/M/c$	$P_w(t)$
	Massey and Whitt (1997)	CKE, MOL	$M(t)/M/c$	$P_w(t)$, E$[W^Q(t)]$
	Green et al. (2001)	SIPP avg/max/mix, Lag avg/max/mix	$M(t)/M/c(t)$	$P_w(t)$
	Ingolfsson et al. (2002)	CKE, SIPP	$M(t)/M/c(t)/K$	$P_w(t)$
	Ingolfsson et al. (2007)	CKE, INFSA, Lag avg, MOL, SDA, UR	$M(t)/M/c(t)$	$P(W^Q \le \alpha)(t)$
	Wall and Worthington (2007)	DTA, PSA, SSA	$M(t)/G/c$	E$[W^Q(t)]$, Quant$^{0.95}[W^Q(t)]$
	Whitt (2007)	PSA, Lagged PSA, MOL	$M(t)/M/c(t)+M$	E$[L^S(t)]$, $P_w(t)$
	Atlason et al. (2008)	SIPP avg/max/mix, Lag avg/max/mix	$M(t)/M/c(t)$	$P_w(t)$
	Stolletz (2008a)	SBC, SIPP	$M(t)/M(t)/c(t)$	E$[U(t)]$, E$[L^S(t)]$, E$[L^Q(t)]$

Table 3.11: Numerical comparisons of time-dependent queueing systems - continued

∞	Stolletz (2011)	FLUID, SBC, SIPP	$M(t)/G/c(t)$	E$[U(t)]$, E$[L^Q(t)]$, E$[W^Q(t)]$
c	Chen et al. (2013c)	PSFFA, SBC	$M(t)/E_k/c$	E$[L^Q(t)]$
$<$	Massey and Pender (2013)	FLUID, DIFF, SDA	$M(t)/M/c(t)+M$	E$[L^S(t)]$, Var$[L^S(t)]$
$\leq\|$	Chen and Yang (2014)	FLUID, PSA, PSFFA	$M(t)/G/c$	E$[L^Q(t)]$
1	Pender (2014a)	FLUID, DIFF, SDA	$M(t)/M(t)/c(t)+M(t)$	E$[L^S(t)]$, Var$[L^S(t)]$
	Selinka et al. (2016)	SBC, SIPP	$M(t)/M/c$	E$[U(t)]$, E$[L^S(t)]$, E$[L^Q(t)]$, E$[W^S(t)]$, E$[W^Q(t)]$
∞	Eick et al. (1993a)	EXPL, PSA, SSA	$M(t)/G/\infty$	E$[L^S(t)]$
$=$	Green and Kolesar (1998)	EXPL, SPEA, SPHA	$M(t)/G/\infty$	E$[L^S(t)]$
c	Pang and Whitt (2012b)	PSA, RA	$M(t)/G^D/\infty$	Var$[L^S(t)]$

3.5 Areas of application

3.5.1 Service systems

Many customer service systems experience time-dependent arrival rates and numbers of servers (Thompson, 1993; Whitt, 2013). Surveys on time-dependent queueing models that are used for staffing decisions in service systems are provided by Green et al. (2007), Whitt (2007), Defraeye and Van Nieuwenhuyse (2011), and Defraeye and Van Nieuwenhuyse (2016). Hampshire and Massey (2010) integrate the performance analysis of time-dependent queueing systems in the optimization of multiple aspects of the communications industry. The applications can be categorized into the areas of telephone call centers, health care facilities, emergency services, service counters, and repair facilities.

Inbound **telephone call centers** are often characterized by a time-dependent arrival rate and a time-dependent number of agents (Gans et al., 2003). Whereas Sze (1984), Aguir et al. (2004), and Ridley et al. (2004) describe the performance evaluation of call centers, all other references cited in this paragraph concentrate on the development of staffing algorithms for call centers. Kolesar and Green (1998) focus on the analysis of the peak hour in their staffing analysis. Most of the models for call centers apply queueing systems with Poisson arrivals and exponentially distributed service times (Andrews and Parsons, 1993; Kolesar and Green, 1998; Green et al., 2001, 2003; Koole and van der Sluis, 2003; Ridley et al., 2004; Dietz and Vaver, 2006; Atlason et al., 2008; Hampshire et al., 2009; Ingolfsson et al., 2010). Abandonments are considered by Feldman et al. (2008), Hampshire et al. (2009), Bertsimas and Doan (2010), Dietz (2011), and Kim and Ha (2012). Customers who reenter the system after abandonment (retrials) are analyzed by Sze (1984) and Aguir et al. (2004). Sze (1984) considers abandonments as part of the arriving jobs that require a service time of zero. All of the models mentioned above consider systems with an infinite waiting room. In contrast, Mok and Shanthikumar (1987) consider a system with a limited waiting room. A call center that can be modeled as an $M(t,n)/G/c$ queueing system with state-dependent balking is considered by Chassioti et al. (2014). Mok and Shanthikumar (1987) consider a heterogeneous queueing system with two server classes, i.e., scheduled servers and standby servers that are used only if the queue exceeds a predetermined threshold. Different job classes and job class-dependent priorities are considered by Ridley et al. (2004) and Bertsimas and Doan (2010).

The request for medical services at **health care** facilities, such as emergency departments, can vary significantly over time (Bhattacharjee and Ray, 2014). Consequently, the number of medical personnel is often also time-dependent. Applications include the performance analysis of emergency facilities (Collings and Stoneman, 1976; de Bruin et al., 2007), staffing in clinical wards (Agnihothri and Taylor, 1991; Gillard and Knight, 2014; Yom-Tov and Mandelbaum, 2014), and ambulance management (Singer and Donoso, 2008). Yom-Tov and Mandelbaum (2014) derive a model that includes reentrant patients/repetitive service in clinical wards. Brahimi and Worthington (1991a) and Bennett and Worthington (1998) use the DTA to analyze outpatient appointment systems. The optimal patient mix with respect to patient service requirements is analyzed by Vanberkel et al. (2014).

Similarly, **emergency services** providers, such as police or fire fighters, face time-dependent service requests. Such systems are considered in the staffing and scheduling algorithm of Green et al. (2006). Green and Kolesar (1995) evaluate peak hour effects for emergency service systems. Bookbinder and Martell (1979) minimize the damage potential of forest fires by considering the allocation of available helicopters. Alfa and Margolius (2008) evaluate the queue of requests for police patrol cars, and Kolesar et al. (1975) and Ingolfsson et al. (2002) apply time-dependent queueing systems as part of scheduling algorithms for police patrol cars.

Service counters and facilities in airport terminals, such as check-in counters, security checks, departure lounges, and baggage claim facilities, experience time-dependent traveler arrivals. A detailed description of these applications is provided in the survey by Tošić (1992). The approaches of Horonjeff (1969), Paullin and Horonjeff (1969), de Neufville and Grillot (1982), Wirasinghe and Shehata (1988), Wirasinghe and Bandara (1990), and de Barros and Tomber (2007) rely on the fluid approximation. Stolletz (2011) uses the SBC to analyze the performance of check-in counters. As another type of service counter with time-dependent arrivals, a fast food restaurant is studied by Kwan et al. (1988). Foote (1976) evaluates the performance of a drive-in banking facility with multiple lines involving jockeying. Kolesar (1984) analyzes the expected number of waiting customers in front of automated teller machines to evaluate different layouts for a bank lobby. The staffing at border crossings in the form of a stationary congestion-based policy is considered by Zhang (2009). Liu and Wein (2008) derive a model to determine the number of necessary beds for the detention and removal of illegal aliens at border crossings.

The demand for repairs at **repair facilities** is also often time-dependent. The analysis of such systems provides insights into the required inventory level of spare parts over time. Jung (1993) analyzes a repair facility for expensive aircraft parts. Jung and Lee (1989a) and Lau and Song (2008) optimize stocking levels of spare parts in repair facilities in a military context. Buczkowski and Kulkarni (2006) use the explicit solution of an $M(t)/G/\infty$ system to model the time-dependent number of items under warranty to determine the optimal funding of a warranty reserve.

All references reporting an application in the area of service systems are listed in Table 3.12. The second column (Emb.) shows that the performance evaluation is often embedded within optimization algorithms, especially for call centers and emergency services. The third column indicates whether real-world data are used in the numerical study. The fluid approximation and methods based on stationary models are most frequently used.

Table 3.12: Applications in the area of service systems

Reference	Emb.	Real data	Eval. method
Telephone call centers			
Sze (1984)			SIPP
Mok and Shanthikumar (1987)	x	x	UR
Andrews and Parsons (1993)	x	x	SIPP
Kolesar and Green (1998)	x	x	SPHA
Green et al. (2001)	x	x	SIPP
Green et al. (2003)	x	x	SIPP
Koole and van der Sluis (2003)	x	x	SIPP
Aguir et al. (2004)		x	FLUID
Ridley et al. (2004)		x	FLUID
Dietz and Vaver (2006)	x		SIPP
Atlason et al. (2008)	x		SIPP
Feldman et al. (2008)	x		INFSA, MOL
Hampshire et al. (2009)	x		FLUID
Bertsimas and Doan (2010)	x	x	FLUID
Ingolfsson et al. (2010)	x		SIPP, UR
Dietz (2011)	x	x	SIPP
Kim and Ha (2012)	x	x	EXPL
Chassioti et al. (2014)	x		PSA
Health care			
Collings and Stoneman (1976)			EXPL
Agnihothri and Taylor (1991)	x	x	SIPP
Brahimi and Worthington (1991a)		x	DTA
Bennett and Worthington (1998)		x	DTA
de Bruin et al. (2007)		x	SIPP
Singer and Donoso (2008)		x	SIPP
Gillard and Knight (2014)	x		CKE
Vanberkel et al. (2014)	x	x	PSA
Yom-Tov and Mandelbaum (2014)	x	x	MOL

Table 3.12: Applications in the area of service systems - continued

Reference	Emb.	Real data	Eval. method
Emergency services			
Kolesar et al. (1975)	x	x	CKE, SIPP
Bookbinder and Martell (1979)	x	x	CKE
Green and Kolesar (1995)			SPHA
Ingolfsson et al. (2002)	x	x	CKE, SIPP
Green et al. (2006)	x	x	SIPP
Alfa and Margolius (2008)		x	DTA
Service counters			
Horonjeff (1969)		x	FLUID
Paullin and Horonjeff (1969)		x	FLUID
Foote (1976)		x	SIPP
de Neufville and Grillot (1982)		x	FLUID
Kolesar (1984)		x	SIPP
Kwan et al. (1988)	x	x	SIPP
Wirasinghe and Shehata (1988)	x		FLUID
Wirasinghe and Bandara (1990)	x	x	FLUID
Thompson (1993)	x		SIPP
de Barros and Tomber (2007)		x	FLUID
Liu and Wein (2008)		x	MOL, PSA
Zhang (2009)			SIPP
Stolletz (2011)			SBC
Repair facilities			
Jung and Lee (1989a)	x		DIFF, FLUID
Jung (1993)	x		CKE
Buczkowski and Kulkarni (2006)	x		EXPL
Lau and Song (2008)	x		SDA

3.5.2 Road and air traffic systems

Road traffic systems, such as roads, bridges, and intersections, are often analyzed by using time-dependent queueing systems to model rush hour and off-peak effects in traffic flows. Catling (1977) applies the CTT for an $M(t)/G/1$ system to analyze the delay at road junctions. The same approximation method is used by Kimber et al. (1977), Kimber and Hollis (1978), and Kimber and Daly (1986) to analyze the performance of three-arm major/minor priority junctions. Kimber et al. (1977) and Kimber and Hollis (1978) apply an $M(t)/M(t)/1$ model and restrict their analysis to artificial data, whereas Kimber and Daly (1986) consider real-world data and apply a $G(t)/G(t)/1$ system. The CTT is applied by Griffiths et al. (1991) to analyze the Channel Tunnel between France and England modeled as an $M(t)/G^{(0,s)}/1$ system. Brilon and Wu (1990) compare the results of a discrete-time approach with empirical data for a one-lane street with a traffic light. Viti and van Zuylen (2009) and Viti and van Zuylen (2010) develop a DTA to evaluate the queue

length at the end of and within a green/red cycle of intersections. Blumberg-Nitzani and Bar-Gera (2014) obtain within-cycle results through an interpolation between the results of the end-of-cycle model. Griffiths et al. (2008) analyze a 24-hour flow pattern on the Severn Bridge between England and Wales as an $M(t)/E_k/1$ system based on piecewise transient models. Gaver (1969) applies the fluid approximation to analyze the time-dependent queueing delays that occur during and after accidents on freeways. Departure time choice and commuting problems also often rely on the deterministic fluid approximation (see Kuwahara (2007) and the references therein).

Time-dependent demand is also distinct for **car and truck handling facilities**. Chen and Yang (2010), Chen et al. (2011), Chen et al. (2013a), Chen et al. (2013b), Chen et al. (2013c), Yang et al. (2013), and Chen and Yang (2014) analyze truck handling facilities at seaports. Based on these analyses, several optimization techniques are proposed to optimize the time-dependent truck arrival process. Selinka et al. (2016) apply the SBC to the performance evaluation of the truck handling system at an air cargo hub with heterogeneous jobs and heterogeneous servers.

Curry et al. (1978) analyze the performance of the queueing process at an airport's taxi stand. In their analysis, they consider the exponentially distributed clearing of a queue that corresponds to a context in which busses collect all customers waiting for a taxi. Deng et al. (1992) develop a model for the optimal allocation of taxis to service zones.

Air traffic is also often time-dependent. Early models for the analysis of runways are proposed by Galliher and Wheeler (1958), Koopman (1972), and Omosigho and Worthington (1985). Bookbinder (1986) analyzes a Markovian queueing system with two separate queues for landing and departing aircrafts and a single runway as a common server. Stolletz (2008b) analyzes a similar model with generally distributed service times. Hebert and Dietz (1997) use the uniformization/randomization approximation and Lovell et al. (2013) use the diffusion approximation to evaluate the time-dependent performance of a single runway. Janic (2009) investigates delays on airport runways under heavy snowfall. In the analysis, the runway's service rate depends on a second queue representing the accumulated snow at airports. Jacquillat and Odoni (2015) use a queueing model in their algorithm to control departure and arrival service rates to maximize the efficiency of an airport's runway system. Jung and Lee (1989b) propose a dynamic programming approach with an embedded time-dependent queueing model to staff air traffic controllers. Congestion-based prices for airport capacity are determined based on the dif-

fusion approximation by Janic (2005), whereas Daniel (1995), Daniel and Pahwa (2000), Daniel and Harback (2008), and Daniel and Harback (2009) use the basic DTA of Galliher and Wheeler (1958). Swaroop et al. (2012) include the fluid approximation in their derivation of slot-controlled flight schedules.

Table 3.13: Applications in the areas of road and air traffic

Reference	Emb.	Real data	Eval. method
Road traffic			
Gaver (1969)			FLUID
Catling (1977)			CTT
Kimber et al. (1977)			CTT
Kimber and Hollis (1978)			CTT
Kimber and Daly (1986)		x	CTT
Brilon and Wu (1990)		x	DTA
Griffiths et al. (1991)			CTT
Kuwahara (2007)	x		FLUID
Griffiths et al. (2008)		x	BOT
Viti and van Zuylen (2009)			DTA, FLUID
Viti and van Zuylen (2010)			DTA,FLUID
Blumberg-Nitzani and Bar-Gera (2014)			DTA
Car and truck handling facilities			
Curry et al. (1978)		x	SIPP
Deng et al. (1992)	x	x	SIPP
Chen and Yang (2010)	x		FLUID
Chen et al. (2011)	x		PSFFA
Chen et al. (2013a)	x	x	PSFFA
Chen et al. (2013b)	x	x	PSFFA
Chen et al. (2013c)	x		PSFFA
Yang et al. (2013)	x		PSFFA
Chen and Yang (2014)		x	FLUID, PSA,PSFFA
Selinka et al. (2016)		x	SBC
Air traffic			
Galliher and Wheeler (1958)		x	DTA
Koopman (1972)		x	FLUID
			CKE, DTA
Omosigho and Worthington (1985)		x	DTA
Bookbinder (1986)		x	CKE
Jung and Lee (1989b)	x		CKE
Daniel (1995)	x	x	DTA
Hebert and Dietz (1997)		x	BOT
Daniel and Pahwa (2000)	x	x	DTA
Janic (2005)	x	x	DIFF
Daniel and Harback (2008)	x	x	DTA
Stolletz (2008b)			SBC
Daniel and Harback (2009)	x	x	DTA
Janic (2009)		x	FLUID
Swaroop et al. (2012)	x	x	FLUID
Lovell et al. (2013)		x	DIFF
Jacquillat and Odoni (2015)	x	x	CKE

All references considered in this section are included in Table 3.13. The number of servers in road and air traffic systems cannot be adjusted over time. Instead, arrival patterns are optimized, e.g., at truck handling facilities.

3.5.3 IT systems

Computer and communication networks transfer data packets whose arrival rates often significantly vary over time (Tripathi and Duda, 1986). The amount of data that can be stored at a certain node is limited. Full buffers may lead to serious performance degradations owing to delays from waiting for transmission capacity or packet retransmissions. Lackman et al. (1992) develop a DTA for a statistical multiplexer that processes real-time and non-real-time traffic. Van As (1986) compares a common-buffer configuration with a foreground-background congestion control mechanism. Tipper and Sundareshan (1990) demonstrate how the PSFFA can be used to find optimal time-dependent arrival rates to individual nodes. The PSFFA is also used by Xu et al. (2014) to evaluate the performance of nodes in multihop wireless networks with constant bit rate traffic. Czachórski et al. (2009) and Czachórski et al. (2010) use the diffusion approximation to model nodes in a wireless network based on the IEEE 802.11 standard and the impact of an adaptive increase and decrease in TCP flow. The fluid and diffusion approximations are used by Ko and Gautam (2010) for the performance evaluation of queues that occur for peer-based multimedia content delivery. The number of active nodes of two different classes in a peer-to-peer (P2P) internet telephony system is modeled with an $M(t)/M/\infty$ system and analyzed via the SIPP and explicit solutions by Kuraya et al. (2009) and Kuraya et al. (2011). McCalla and Whitt (2002) evaluate the volume of lines in private line telecommunication services by using the explicit solution of a $G(t)^{X(t)}/G(t)/\infty$ system.

Rothkopf and Johnston (1982) apply the SDA to predict the queues in front of printers for which the arrival rate of jobs is time-dependent. The coverage process on a straight line in a sensor field is analyzed by Manohar et al. (2009), who show that this process can be modeled as a time-dependent $M(t)/G(t)/\infty$ system. In such a system, the time corresponds to the location in the sensor field, and the number of jobs in the system corresponds to the number of sensors that cover the associated area.

All references reporting applications with IT systems are presented in Table 3.14. In contrast to studies on other areas of application, most references focus on the performance evaluation only.

Table 3.14: Applications in the area of IT systems

Reference	Emb.	Real data	Eval. method
Rothkopf and Johnston (1982)		x	SDA
Van As (1986)			CKE
Tipper and Sundareshan (1990)	x		PSFFA
Lackman et al. (1992)			DTA
McCalla and Whitt (2002)		x	EXPL
Kuraya et al. (2009)			SIPP
Czachórski et al. (2009)			DIFF
Manohar et al. (2009)			PSA
Czachórski et al. (2010)			DIFF
Ko and Gautam (2010)			DIFF, FLUID
Kuraya et al. (2011)			EXPL
Xu et al. (2014)			PSFFA

3.6 Conclusions and future research

This paper provides a structured overview of approaches for the performance evaluation of time-dependent queueing systems (Section 3.3). We discuss links between the different approaches and demonstrate that numerical comparisons exist only for a subset of the existing approaches (Section 3.4). Thus, a research gap remains for a comprehensive numerical study comparing the approximation quality of approaches within all three categories for different types of queueing systems with various levels of stochasticity and different time-dependent patterns for the system parameters. Moreover, a methodological extension of some approaches is needed to analyze general systems. An opportunity for the development of new approaches lies in the combination of existing ideas concerning approximation. For instance, a transformation, as suggested by the CTT, could be integrated into approaches that currently rely on regular steady-state queueing formulas.

Section 3.5 demonstrates the wide range of areas of application for time-dependent queueing systems, including service, road and air traffic, and IT systems. The currently used evaluation methods are often based on stationary models, discrete-time approaches, or fluid approximations. Notably, some evaluation methods are used only within a single area of application. For example, the CTT is used only for the analysis of road traffic systems, and the PSFFA is used mainly for truck handling facilities. In general, for all areas of application, a systematic test of other evaluation approaches may represent a worthwhile investigation. Most of the optimization algorithms that use embedded time-dependent queueing formulas involve decisions regarding the

number of servers in service systems. In the area of truck handling and IT systems, the arrival rate is treated as a decision variable. The optimization of service rates is addressed only by the theoretical work of Parlar (1984) and is a potential field of future research. Another open field is the time-dependent decision regarding the provision of waiting rooms, which is introduced in a call center context by Hampshire et al. (2009). In summary, this review shows that there are numerous areas of application for time-dependent queues. A promising field of research is the extensive use of time-dependent performance evaluation approaches as embedded with optimization procedures.

4 Approximations of time-dependent unreliable flow lines with finite buffers

Co-authors:
Simone Göttlich and **Sebastian Kühn**
School of Business Informatics and Mathematics, University of Mannheim, Germany

Raik Stolletz
Chair of Production Management, Business School, University of Mannheim, Germany

Published in:
Mathematical Methods of Operations Research, 2016,
DOI: 10.1007/s00186-015-0529-6, In Press, pages 1-29, reprinted with permission from Springer

Abstract:
Flow lines process discrete workpieces on consecutive machines, which are coupled by buffers. Their operating environment is often stochastic and time-dependent. For the flow line under consideration, the stochasticity is generated by random breakdowns and successive stochastic repair times, whereas the processing times are deterministic. However, the release rate of workpieces to the line is time-dependent, due to changes in demand. The buffers between the machines may be finite or infinite. We introduce two new sampling approaches for the performance evaluation of such flow lines: one method utilizes an approximation based on a mixed-integer program in discrete time with discrete material, while the other approximation is based on partial and ordinary differential equations in continuous time and with a continuous flow of material. In addition, we sketch a proof that these two approximations are equivalent under some linearity assumptions. A computational study demonstrates the accuracy of both approximations relative to a discrete-event simulation in continuous time. Furthermore, we reveal some effects occurring in unreliable flow lines with time-dependent processing rates.

4.1 Introduction

Flow lines consist of several, serial machines that perform consecutive processing tasks. The flow line considered in this paper includes buffers located between two adjacent machines. It processes discrete workpieces, which can be moved independently along the line, i.e., the material flow is asynchronous. The operating environment of such flow lines is stochastic due to random machine failures and the associated downtime needed to implement repairs. This stochasticity may lead to blocking and starvation. A machine starves if it idles due to a lack of workpieces in the upstream buffer. Conversely, for finite buffer capacities, blocking may occur if a processed workpiece cannot leave a machine immediately due to a full downstream buffer.

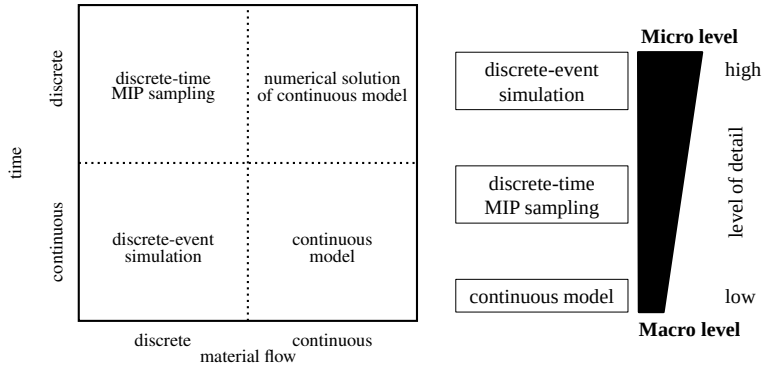
The literature on flow lines commonly assumes that flow lines operate under steady-state conditions (Dallery and Gershwin, 1992). In particular, the setting of the flow line is constant over time and the flow line is observed after a sufficient amount of time has passed. Thus, the probability distributions that describe the flow line behavior are time-invariant. However, in practice the setting of the line may change over time. Hence, a steady state is not reached or even does not exist. Dynamic changes of the setting arise from learning effects associated with production ramp-ups, newly introduced production technologies, or seasonal demand patterns (Terwiesch and Bohn, 2001; Jaikumar and Bohn, 1992; Takahashi and Nakamura, 2002). This paper focuses on time-dependent changes of the workpieces' release rate to the line, which typically occur after changes in the demand pattern. The supply of raw material for the first machine, as well as the storage capacity for finished goods behind the last machine, are assumed to be unlimited. The buffer capacities between the machines may be finite or infinite.

Related to flow lines two literature streams can be distinguished: optimization and evaluation approaches. Optimization approaches aim to find a setting of the line, e.g., the buffer allocation, to fulfill a given objective, e.g., maximization of the throughput. A comprehensive survey pertaining to the optimization of buffer allocations under steady-state conditions is provided by Demir et al. (2014). For evaluation approaches the line setting is given and the approaches provide exact or approximative results of the line performance. The performance evaluation of flow lines under steady-state conditions is studied in numerous articles. These evaluations are often based on Markov chains, decomposition approaches, or discrete-event simulation (Dallery and Gershwin, 1992). Evaluation models add value as they reveal the relationship between the setting of the line, e.g., buffer capacities, and the

line performance. Moreover, they can be used as integral part of optimization approaches (Demir et al., 2014).

We propose two new sampling approaches for the performance evaluation of time-dependent, unreliable flow lines that produce discrete workpieces in continuous time: The first approximation is based on a discrete-time, mixed-integer program (MIP), which maintains the property of discrete workpieces. This approximation replaces the continuous time with equal-length, discrete-time intervals. The second proposed approximation preserves the property of continuous time but approximates the discrete material using a continuous flow model. The continuous flow model is described by partial differential equations, and the numerical solution of the continuous model utilizes numerical methods that include a discretization of time. A discrete-event simulation (DES) in continuous time with discrete material is used to evaluate the accuracy of both approximations. Figure 4.1a illustrates the relationship between the different approaches.

DES is a very powerful and common approach for the modeling of unreliable flow lines. It can be interpreted as a *microscopic level*, which means that the computation highly depend on the number of workpieces to be considered. We introduce an *intermediate* and a *macroscopic* level. The different levels of detail regarding the workpieces leads to a model hierarchy which is depicted in Figure 4.1b.



(a) Approximations of time and material flow

(b) Model hierarchy

Figure 4.1: Relationships between the proposed approximations and discrete-event simulation

The basic idea is to transfer the detailed DES model with information about individual workpieces and their dynamical behavior to other scales. For instance, the MIP still includes discrete workpieces but only traces aggregated values which describe the number of workpiece in the buffers. Based on the intermediate approach, we are able to derive a macroscopic model avoiding the dependency on single workpieces. This continuous model relies on continuous time and approximates individual workpieces by a density function.

The advantages of the proposed models over DES are: First, both new approaches do not trace individual workpieces and thereby reduce the dependency on the workpieces' release rate. Second, the MIP approach for evaluation bears the potential to be converted into an optimization approach for system parameters, e.g., the buffer capacities. This two-step approach of first deriving an evaluation formulation and subsequently using the power of the MIP to optimize key system parameters has been successfully applied by, e.g., Alfieri and Matta (2012) and Helber et al. (2011) for the buffer allocation problem under stationary conditions.

Our literature survey focuses on articles related to the two approximation approaches under consideration. First, approaches that maintain the discrete flow of material are reviewed. Subsequently, existing approaches that use continuous flow models are presented.

Even under steady-state conditions, exact solutions for the performance evaluation of unreliable flow lines with finite buffers are known only for three machine lines for which repair times and the time between failures are geometrically distributed (Gershwin and Schick, 1983). For time-dependent systems with finite buffer capacities exist no exact analytical approaches, even with restriction to a single stage (Schwarz et al., 2016). For an approximation of a single-stage $G(t)/G/1/k$ queueing model, see Stolletz and Lagershausen (2013). Nasr and Taaffe (2013) proposed a time-dependent decomposition approach for networks consisting of $Ph(t)/M(t)/s/k$ queues. However, this approach is limited to exponential processing times and assumes that workpieces are lost in case of a full buffer. Hence, the characteristic effect of blocking of upstream machines in flow lines is neglected. For the analysis of time-dependent flow lines only DES techniques are applied (Fan, 1976; Takahashi and Nakamura, 2002).

Chan and Schruben (2008) proposed the idea of modeling stochastic discrete-event systems as optimization problems, and it is noted that this concept is methodologically related to our discrete-time approach. In this field, Helber et al. (2011) evaluated and optimized the performance of the buffer alloca-

tion in a flow line under stationary conditions. They used the realizations of random variables as a deterministic input for a discrete-time MIP. Alfieri and Matta (2012) optimized the buffer allocation using a sampling approach in continuous time. Contrary to these sampling approaches, our analysis considers changes of the rates over time and obtains measures for the resulting time-dependent performance.

The second literature stream is related to the proposed continuous-flow approach. Recently, Tan (2015) applied the sample-based optimization approach to a continuous flow line. Other approaches use partial differential equations to model the transport of workpieces through the flow line. These equations describe the density of workpieces as a continuous function of time and space, where the latter may be seen as the degree of completion of the workpieces. Early approaches for flow lines with infinite buffer capacities and smaller networks can be found in Vandergraft (1983), whereas recent developments in the field are presented, e.g., by Göttlich et al. (2005) or by D'Apice et al. (2010). The efficient description of large but deterministic production networks and their analysis have been considered by Fügenschuh et al. (2008). They establish the equivalence of a discretized model based on partial differential equations and a MIP model that both assume continuous material flow. Furthermore, optimization issues, such as maximization of throughput, have been considered by Kirchner et al. (2006) and other researchers. The stochastic influences of random breakdowns have been investigated by Degond and Ringhofer (2007) and by Göttlich et al. (2011). The former focused on the derivation of a time-recursion from which a partial differential equation model was obtained. However, the latter directly used partial differential equations to describe the evolution of a whole network. In this case, piecewise deterministic processes (PDP), originally invented by Davis (1984, 1993), were used to efficiently solve the model. Furthermore, Göttlich et al. (2011) presented a modified version of the stochastic simulation algorithm (SSA) of Gillespie (2001). All previously introduced models based on partial differential equations include infinite buffer capacities. In contrast, our model assumes finite buffer capacities. This is a novel and mathematically challenging restriction to the model.

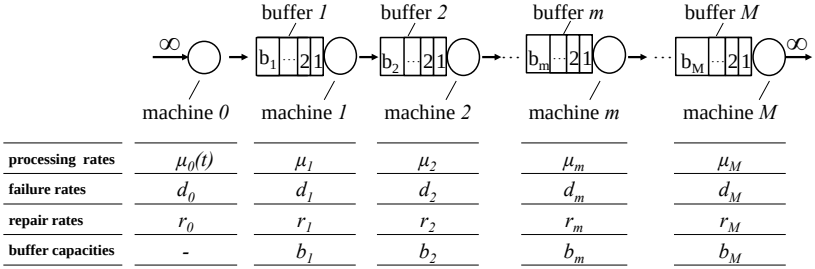
The contribution of this paper is threefold: First, two new analytical approaches for the evaluation of time-dependent, unreliable flow lines are proposed. Second, the paper provides a model hierarchy and a discussion on the relationship between the two approximation approaches. By establishing the equivalence of the MIP model with discrete material flow and the continuous material flow model, we connect two literature streams. Third, insights re-

garding the time-dependent behavior of unreliable flow lines are provided via a numerical study. Moreover, the approximation quality of both approaches is demonstrated via a comparison with a DES that models discrete material while maintaining continuous time.

The remainder of the article is organized as follows. In Section 4.2, the considered flow line is described. The discrete-time, discrete-flow approach is presented in Section 4.3. Alternatively, Section 4.4 includes the derivation of a continuous-time, continuous-flow approximation. In addition, a model reduction is proposed in Section 4.5, which relates both approximation approaches to one another. The numerical study presented in Section 4.6 demonstrates the approximation quality and provides insights regarding the time-dependent behavior of flow lines. Concluding remarks and suggestions for future research are provided in Section 4.7.

4.2 Modeling of time-dependent flow lines

4.2.1 Model description



processing rates	$\mu_0(t)$	μ_1	μ_2	μ_m	μ_M
failure rates	d_0	d_1	d_2	d_m	d_M
repair rates	r_0	r_1	r_2	r_m	r_M
buffer capacities	-	b_1	b_2	b_m	b_M

Figure 4.2: Flow line with time-dependent release rate and random breakdowns and repairs

The flow line model consists of $m = 0, 1, 2, ..., M$ machines coupled by buffers with capacity b_m in front of machine m (see Figure 4.2). Collectively, buffer m and machine m form station m. This notation complies with the common notation for continuous-flow networks, i.e., the machine and its preceding buffer share the same index. We assume that the flow line produces a single product. This assumption is common for flow lines as they are typically used for mass production (Dallery and Gershwin, 1992). The line

starts empty at time $t = 0$. Discrete workpieces enter the line via (artificial) machine $m = 0$. This machine has an infinite supply of raw materials and therefore never starves. Furthermore, we assume that there is unlimited space behind machine $m = M$. Transportation times between machines and buffers are assumed to be negligible. The finite capacity of the buffers $m = 1, ..., M$ may cause blocking of upstream machines whenever a buffer is completely filled with workpieces. For the following analysis, blocking after service is assumed. This implies that a workpiece remains on machine m after being processed until an empty space at buffer $m + 1$ becomes available. Additionally, the processing rate $\mu_0(t)$ of machine $m = 0$ is time-dependent and represents the release rate of workpieces into the line, starting at time $t = 0$.

All machines process the workpieces on a first-come first-serve basis. The raw processing times of machine m are deterministic with rate μ_m. This is a common assumption for flow lines, justified by the behavior of automated machines with little or no variability with respect to raw processing times (Dallery and Gershwin, 1992; Dolgui et al., 2002). However, the effective processing times are stochastic due to unreliable machines, i.e., machines are subject to random breakdowns. There are two failure modes which can be distinguished: run-based and time-based failures. A run-based failure can occur only if the machine is processing a workpiece whereas time-based failures can occur independently of the machine's state (operating, idling, or blocked) (Wu, 2014). We model a flow line with time-based failures. This kind of failure may be caused by faulty electronic parts and/or software controlling the machines (Buzacott and Hanifin, 1978), power outages, or by preventive maintenance which is deliberately started during an idling period (Wu et al., 2011). Time-based failures may also be used to approximate run-based failures but according to Mourani et al. (2007) this results in an underestimation of the throughput rate. Nevertheless, Li and Meerkov (2009) report that their numerical study revealed only minor difference of 1% to 3% between the two failure modes for throughput and work in process (WIP).

The time between failures is exponentially distributed with failure rate d_m for machine m. If machine m breaks down, the repair process starts immediately with an exponentially distributed repair time with rate r_m. The processing is continued without loss of previous work after the completion of the repair time. In the special case that a machine breaks down during a blocking period, the workpiece can leave the blocked machine as soon as the blocking is resolved, regardless of the machine's repair status.

4.2.2 Performance measures

Time-dependent performance measures of interest are related to WIP and the output of the flow line. We focus on risk neutral expected values, as it is common in the literature on flow line evaluation and optimization (Demir et al., 2014; Weiss et al., 2015). $\mathrm{E}[WIP_m(t)]$ measures the expected number of workpieces in buffer m and on machine m at time t. The line throughput is equivalent to that of the last machine $m = M$. The expected cumulated output $\mathrm{E}[TH^c(t)]$ of the flow line at time t equals the expected number of workpieces produced up to t on machine $m = M$.

4.2.3 Sampling approach

We choose a sampling approach to account for randomness of the unreliable machines. Our mathematical formulation deploys a two-state stochastic process (Göttlich et al., 2011)

$$\begin{aligned} \omega_m : \mathbb{R}_{\geq 0} \times \mathcal{S} &\longrightarrow \mathbb{B} \\ t \times s &\longmapsto \omega_m(t, s) \end{aligned} \tag{4.1}$$

for all machines $m = 1, \ldots, M$. Note that $\omega_m(t, s) = 0 \in \mathbb{B} = \{0, 1\}$ indicates a broken machine, while $\omega_m(t, s) = 1$ implies that a machine is operating properly. The state process ω_m depends both on the time and on the random sample $s \in \mathcal{S}$. Thus for a fixed time $t \geq 0$, $\omega_m(t, \cdot)$ is a binary random variable, whereas for a fixed random sample, $\omega_m(\cdot, s)$ is a realization of the state process (see Figure 4.3).

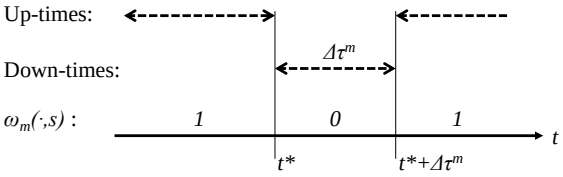


Figure 4.3: Realization of a two-state process ω_m with values in $\mathbb{B}$

To model the state switching of a machine (independent of the inventory level in its buffer, on the machine, or the state of other machines), we use the failure rate d_m and the repair rate r_m for each machine m. The former describes the switching rate from $\omega_m = 1$ to $\omega_m = 0$, while the latter defines the

rate of switching from $\omega_m = 0$ to $\omega_m = 1$. Then, for each machine, the time $\Delta\tau^m$ between two state transitions occurring at the switching points $t^\star$ and $t^\star + \Delta\tau^m$ is randomly chosen according to those rates. We assume an exponential distribution with the density function $Exp(t;\lambda)$ and the rate parameter

$$\lambda = \lambda(\omega_m(t^\star)) = \begin{cases} d_m & \omega_m(t^\star) = 1, \\ r_m & \omega_m(t^\star) = 0. \end{cases} \tag{4.2}$$

4.3 MIP model for the evaluation of time-dependent flow lines

The core idea of the discrete-time approximation is to maintain discrete material but to approximate the continuous time by discrete intervals i of equal length $\tau_i = \tau \ \forall i$. We propose a MIP for the time-dependent performance evaluation which is based on a formulation by Helber et al. (2011) for the evaluation of flow lines under stationary conditions. The proposed MIP solely decides on the production quantities and the storage of workpieces in the buffers. The values of the decision variables in the optimal solution serve as approximation of the flow line performance. The buffer capacities are given as integer values exogenously, as the model in the presented form is used for performance evaluation only. In principle, buffer capacities may be converted into decision variables. However, this requires the development of efficient solution techniques which is out of the scope of the present paper.

The MIP integrates the randomness of breakdowns and repairs by sampling of production capacities. Helber et al. (2011) use a single sample to approximate the stationary system behavior based on a time average. However, information regarding the time-dependent behavior is lost due to this time averaging. Thus, the new approach presented in the following derives the performance of the system over time from averages of $s = 1, 2, ..., S$ samples. For sample s, the production capacity $c_{m,i,s}$ equals the number of workpieces that machine m can produce during interval i if it were to operate in isolation. It is derived from samples of the failure and repair times represented by the stochastic process ω_m. The values of $c_{m,i,s}$ are obtained in two steps. We first generate a list with the continuous completion times of the processing, inducing sampled up and down times, for every machine in isolation. In a second step, the completion times are converted into discrete production capacities $c_{m,i,s}$ per machine m, interval i, and sample s.

For the generation of continuous completion times we consider that production may take place only during up-times, i.e., if $\omega_m(t,s) = 1$ holds for sample s. Thus, as long as an isolated machine m is operating, it produces at a given rate $\mu_0(t)$ and μ_m, respectively. However, if a breakdown occurs, processing is interrupted. Under the assumption of work conservation, the processing is continued after the repair process has been completed. For the time-dependent machine $m = 0$ the processing times are determined at the start of production according to the processing rate $\mu_0(t)$. Hence, a change in the rate $\mu_0(t)$ becomes effective with the first workpiece after the rate change.

In the second step, we obtain the potential production capacity $c_{m,i,s}$ for each machine m, discrete interval i, and sample s from the list of continuous completion times by counting the number of finished workpieces during the interval i. Thus, the capacity $c_{m,i,s}$ always takes integer values. For the special case of $\tau = 1/\mu_m$ the processing of workpieces prior to the first breakdown may finish immediately at the end of interval i. In this case the workpiece is assigned to interval i.

Table 4.1 lists the additional notation for the MIP. All undefined variables, such as $TH_{-1,i,s}$ and $WIP^b_{m,0,s}$, are omitted from the respective constraints.

Table 4.1: Notation for the discrete-time and discrete-material model

Indices	
$i = 1, 2, ..., I$	discrete intervals
$s = 1, 2, ..., S$	samples
Parameters	
b_m	exogenously given capacity of the buffer before machine m
$c_{m,i,s}$	potential processing capacity of machine m in interval i for sample s
Integer decision variables	
$WIP^b_{m,i,s}$	end-of-interval inventory level of buffer m in interval i for sample s
$TH_{m,i,s}$	production quantity of machine m in interval i for sample s

$$\max \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{i=1}^{I} (I - i)\, TH_{m,i,s} \tag{4.3a}$$

s.t.

$$WIP^b_{m,i,s} = WIP^b_{m,i-1,s} + TH_{m-1,i,s} - TH_{m,i+1,s}, \quad m = 1, ..., M, \forall i, \forall s \tag{4.3b}$$

$$TH_{m,i,s} \leq c_{m,i,s}, \quad \forall m, \forall i, \forall s \tag{4.3c}$$

$$WIP^b_{m,i,s} \leq b_m, \quad m = 1, ..., M, \forall i, \forall s, \tag{4.3d}$$

$$WIP^b_{m,i,s},\, TH_{m,i,s} \geq 0, \text{and integer} \quad \forall m, \forall i, \forall s \tag{4.3e}$$

The objective function (4.3a) maximizes the total production on all stages. The weight factor $(I - i)$ decreases with increasing values of i. Hence, it favors production in early intervals and consequently ensures that all workpieces are moved through the line as fast as possible. Constraints (4.3b) are inventory balance equations. The inventory in buffer m at the end of interval i equals the inventory at the end of the previous interval $i - 1$ increased by the production quantities $TH_{m-1,i,s}$ of the upstream machine in interval i and decreased by the production quantities $TH_{m,i+1,s}$ of the downstream machine in the next interval $i + 1$. This implies that workpieces produced from machine $m - 1$ flow into buffer m within an interval. From buffer m to machine m however, workpieces are moved only at the end of intervals. According to Constraints (4.3c), the production quantity in each interval may not exceed the sampled potential production quantity. It can, however, be lower in the case of blocking or starvation. Nevertheless, it is guaranteed by Constraints (4.3d) that the inventory in the buffers does not exceed the buffer restriction. Finally, all decision variables must be non-negative integers (4.3e). Note that the model can be solved independently for each sample s.

When the production capacities are such that $c_{m,1,s} \geq 1$ for all machines m, Constraints (4.3b) allow for a positive throughput at the last machine in the first interval. To avoid this type of initial condition, we generate the sampled production capacities of the first periods with respect to the real (not discretized) schedule for the first workpiece. The sampled production capacities $c_{m,i,s}$ are set to zero for all intervals before the first arrival of a workpiece. Helber et al. (2011) implemented a minimum lead-time of one interval between adjacent machines. In contrast to this approach, the modified sampling considers the actual length of the processing and repair times with respect to the first workpiece.

The proposed MIP model permits two types of approximation errors. In particular, a simulation error that decreases as the number of samples S increases. In addition, discretization errors originate from the inventory balance equation (4.3b) which allows the transfer of workpieces within and at the end of an interval. The magnitude of these discretization errors depends on the length of the intervals τ. Preliminary tests show that setting the length of the discretization interval equal to $\tau = \min_m\{1/\mu_m, \min_t\{1/\mu_0(t)\}\}$ yields small approximation errors, such that this τ is also used in the numerical study.

We use a standard solver to obtain solutions of the MIP (4.3). Potentially, specialized algorithms or heuristics can also be utilized to solve the problem.

4.4 Continuous model for the evaluation of time-dependent flow lines

In this section we present a continuous-time modeling approach as an alternative to the one presented in Section 4.3. We first concentrate on the derivation of a deterministic model. Subsequently, we explain how random breakdowns and repairs can be included. We close the section with a brief introduction to the numerical methods used to solve this stochastic model.

The approach presented in this section allows for continuous time but does not trace each workpiece individually. Instead, we consider the density of workpieces, which is the non-integer number of workpieces per unit length. We assume that each machine has a spatial extension, which may be interpreted as continuous degree of completion. Furthermore, we model the processing on each machine m as a continuous flow along each machine with a constant velocity v_m. As the density of workpieces is non-integer, the transfer to and from machines is a continuous variable as well.

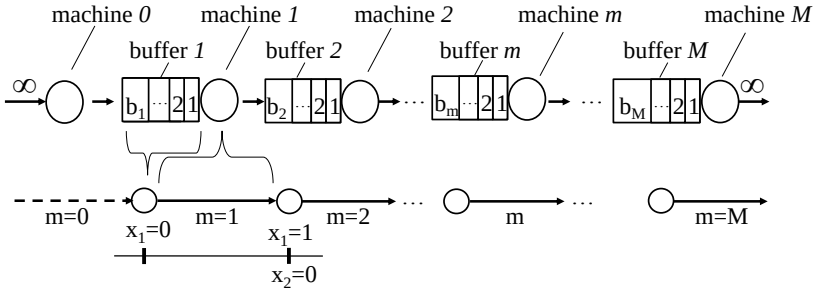


Figure 4.4: Flow line represented by unit intervals and zero space dimensional buffers

We model each machine as an one-dimensional arc and use a single coordinate $x \in [0, 1]$ to uniquely determine the location (the degree of completion) within a machine (see Figure 4.4). We define the density $\rho_m(x, t)$ of workpieces on machine m at position x at time t as the number of workpieces on machine m per unit length. With this definition of density, we are able to analyze the flow f_m on each machine m. The flow corresponds to the number of workpieces per time unit moving along the machine. It depends on the density $\rho_m(x, t)$. If the density $\rho_m(x, t)$ is in the free flow regime, i.e., machine m is working below its flux limit and is not congested, then the flow

is given by $v_m\rho_m$ (see Figure 4.5). In other words, the more workpieces are available the greater is the flow. This holds up to the maximal flux $\mu_m^{\max}$ of machine m. This $\mu_m^{\max}$ corresponds to the rate μ_m in the original flow line model presented in Section 4.2. On the other hand, if the downstream buffers are full, we observe congestion on machine m. The information about the congestion is passed backwards through machine m with the velocity v_m, too. The material flow is then reduced to $2\mu_m^{\max} - v_m\rho_m$ (see Figure 4.5). This marks a difference between the continuous model and the discrete model (4.3). For the discrete model, blocking interrupts material flow, whereas in the continuous model, congestion only reduces the flow on machine m to the flow allowable by the downstream machine $m+1$.

We define $\sigma_m = \mu_m^{\max}/v_m$ as the maximal density which is allowed for a free flow being attained on machine m and thus we get

$$f_m(\rho_m(x,t)) = \begin{cases} v_m\rho_m(x,t) & 0 \le \rho_m(x,t) \le \sigma_m, \\ 2\mu_m^{\max} - v_m\rho_m(x,t) & \sigma_m \le \rho_m(x,t) \le 2\sigma_m. \end{cases} \tag{4.4}$$

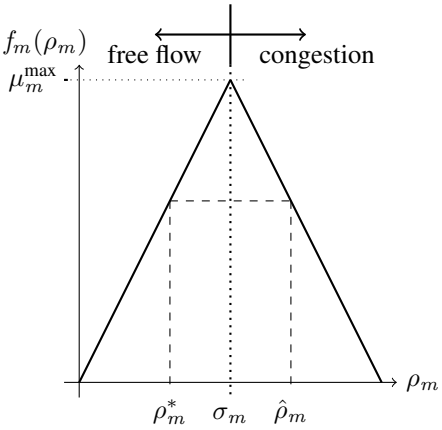


Figure 4.5: Flow (or clearing) function (4.4) (solid) and indicated two density values yielding the same flow value (dashed line). The relation is $\rho_m^* = 2\sigma_m - \hat{\rho}_m$, where $\rho_m^* < \sigma_m$ is the free flow density and $\hat{\rho}_m > \sigma_m$ the congested density of the corresponding flow value

As $\rho_m(x,t)$ describes the density of workpieces at position x at time t on machine m, the number of workpieces in a section $[x_a, x_b]$ of machine m

equals the integral of the density between the points $0 \leq x_a < x_b \leq 1$ at time t

$$\int_{x_a}^{x_b} \rho_m(x,t)\,\mathrm{d}x\,. \tag{4.5}$$

In the same way, we can express the amount of workpieces flowing through position x within a time interval $0 \leq t_1 < t_2$

$$\int_{t_1}^{t_2} f_m(\rho_m(x,t))\,\mathrm{d}t\,. \tag{4.6}$$

Assuming sufficient regularity on $\rho_m(x,t)$ and $f_m(\rho_m(x,t))$ we obtain material balance equation (4.7), because no parts are lost or generated within each single machine. This states, that the number of workpieces in section $[x_a, x_b]$ at time t_2 equals the number of workpieces, which were there at time t_1 plus the number of workpieces, which entered the section at x_a in time interval $[t_1, t_2]$ minus the workpieces, which left the section at x_b in this time interval (see Figure 4.6)

$$\int_{x_a}^{x_b} \rho_m(x,t_2)\,\mathrm{d}x = \int_{x_a}^{x_b} \rho_m(x,t_1)\,\mathrm{d}x + \int_{t_1}^{t_2} f_m(\rho_m(x_a,t))\,\mathrm{d}t - \int_{t_1}^{t_2} f_m(\rho_m(x_b,t))\,\mathrm{d}t\,. \tag{4.7}$$

From the fundamental theorem of calculus we obtain

$$\rho_m(x,t_2) - \rho_m(x,t_1) = \int_{t_1}^{t_2} \partial_t \rho_m(x,t)\,\mathrm{d}t \tag{4.8a}$$

$$f_m(\rho_m(x_a,t)) - f_m(\rho_m(x_b,t)) = \int_{x_a}^{x_b} \partial_x f_m(\rho_m(x,t))\,\mathrm{d}x \tag{4.8b}$$

for sufficiently smooth derivatives. Combining the balance equation (4.7) and Equations (4.8) yields

$$\int_{x_a}^{x_b}\int_{t_1}^{t_2} \partial_t \rho_m(x,t)\,\mathrm{d}t\,\mathrm{d}x + \int_{t_1}^{t_2}\int_{x_a}^{x_b} \partial_x f_m(\rho_m(x,t))\,\mathrm{d}x\,\mathrm{d}t = 0\,. \tag{4.9}$$

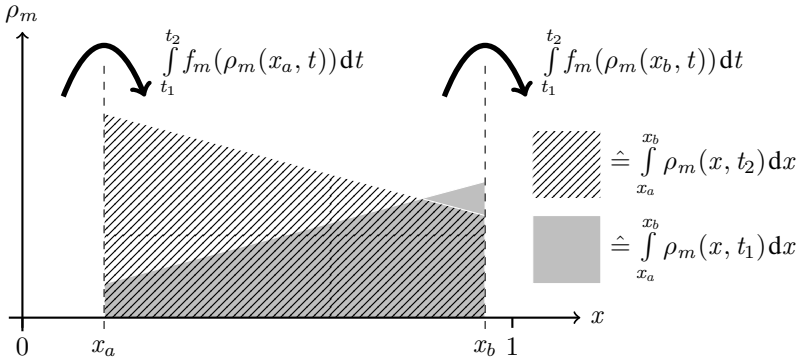


Figure 4.6: Graphical interpretation of Equation (4.7). Workpieces flowing into and out of the region $x_a < x_b$ are marked by arcs and computed according to (4.6). The number of workpieces within the region are calculated by (4.5) and marked as light gray (lines) for time t_1 (t_2)

If we assume that Equation (4.9) holds for all sections $[x_a, x_b]$ in machine m and all time intervals $[t_1, t_2]$, then we get the scalar transport equation (or conservation law) (LeVeque, 1992). This conservation law describes the transport of the density of workpieces within each machine

$$\partial_t \rho_m(x, t) + \partial_x f_m(\rho_m(x, t)) = 0\,. \tag{4.10}$$

The machines are coupled by buffers without spatial extension. Buffer m is placed in front of machine m at $x_m = 0$ and directly after machine $m-1$ at $x_{m-1} = 1$ (see Figure 4.4). For each buffer m, we introduce a variable $q_m(t)$ describing its inventory level at time t. The derivative $\partial_t q_m(t)$ expresses the change in the number of workpieces in the buffer m at time t (in workpieces per time). It is positive if workpieces enter the buffer and negative if the buffer is (partially) cleared. We obtain a rate equation for the buffer m, i.e., the change of inventory in buffer m, $\partial_t q_m(t)$, is the difference of the inflow $\gamma_m^{\text{in}}(t)$ to buffer m and the outflow γ_m^{out} of buffer m (see, e.g., Coclite et al. (2005); Garavello and Goatin (2012))

$$\partial_t q_m(t) = \gamma_m^{\text{in}}(t) - \gamma_m^{\text{out}}(t)\,. \tag{4.11}$$

Both the inflow $\gamma_m^{\text{in}}(t)$ to and the outflow $\gamma_m^{\text{out}}(t)$ of buffer m are given in workpieces per time unit. For the computation of the outflow, we have to

distinguish two cases. If the buffer is filled ($0 < q_m(t) \leq b_m$), machine m can produce at its current maximal flux, which we define by $\hat{\gamma}_m(t)$. This flux depends on the density of machine m. If it is in free flow regime, $\hat{\gamma}_m(t) = \mu_m^{\max}$ and if it is congested (see Figure 4.5, right part), the current maximal flux is reduced to the maximal flow being able to be processed, $\hat{\gamma}_m(t) = 2\mu_m^{\max} - v_m\rho_m(0,t)$, yielding $\hat{\gamma}_m(t) = \min\{\mu_m^{\max}, 2\mu_m^{\max} - v_m\rho_m(0,t)\}$. On the other hand, if buffer m is empty ($q_m(t) = 0$), then there are two subcases: either machine $m-1$ provides less than machine m is able to process ($f_{m-1}(\rho_{m-1}(1,t)) < \hat{\gamma}_m(t)$) or machine $m-1$ makes more workpieces available to machine m than it can process. The latter results in a rise of inventory in buffer m. This yields the following relation for the outflow

$$\gamma_m^{\text{out}}(t) = \begin{cases} \min\left\{f_{m-1}\left(\rho_{m-1}(1,t)\right), \hat{\gamma}_m(t)\right\} & q_m(t) = 0, \\ \hat{\gamma}_m(t) & 0 < q_m(t) \leq b_m. \end{cases} \tag{4.12}$$

This choice of the buffer outflow ensures that the throughput is as large as possible and that the buffer is emptied as quickly as possible (see Coclite et al. (2005); Garavello and Goatin (2012)).

Similar to the outflow, we have to distinguish two cases considering the inflow $\gamma_m^{\text{in}}(t)$. If there is space left in the buffer ($q_m(t) < b_m$), the inflow to the buffer is given by the output of the upstream machine $m-1$, i.e., the flow at $x_{m-1} = 1$. On the other hand, if the buffer is at its maximal capacity, it is not possible to transfer more flow into the buffer than machine m is able to process. Accordingly, this may lead to congestion on machine $m-1$. Thus the flow is transferred from machine $m-1$ to buffer m according to the equation

$$\gamma_m^{\text{in}}(t) = \begin{cases} f_{m-1}\left(\rho_{m-1}(1,t)\right) & 0 \leq q_m(t) < b_m, \\ \min\left\{f_{m-1}\left(\rho_{m-1}(1,t)\right), \hat{\gamma}_m(t)\right\} & q_m(t) = b_m. \end{cases} \tag{4.13}$$

Summarizing, the following model has been generated to represent a flow line with finite buffers in continuous time and continuous space

$$0 = \partial_t \rho_m(x,t) + \partial_x f_m(\rho_m(x,t)) \tag{4.14a}$$

$$\partial_t q_m(t) = \gamma_m^{\text{in}}(t) - \gamma_m^{\text{out}}(t) \tag{4.14b}$$

$$\gamma_m^{\text{in}}(t) = \begin{cases} f_{m-1}\left(\rho_{m-1}(1,t)\right) & 0 \le q_m(t) < b_m \\ \min\left\{f_{m-1}\left(\rho_{m-1}(1,t)\right), \hat{\gamma}_m(t)\right\} & q_m(t) = b_m \end{cases} \tag{4.14c}$$

$$\gamma_m^{\text{out}}(t) = \begin{cases} \min\left\{f_{m-1}\left(\rho_{m-1}(1,t)\right), \hat{\gamma}_m(t)\right\} & q_m(t) = 0 \\ \hat{\gamma}_m(t) & 0 < q_m(t) \le b_m \end{cases} \tag{4.14d}$$

$$\rho_m(x,0) = \rho_m^0(x),\ q_m(0) = q_m^0,\ f_0(\rho_0(0,t)) = \min\left\{\mu_0(t), \hat{\gamma}_0(t)\right\}, \tag{4.14e}$$

where q_m^0 and $\rho_m^0(x)$ describe the initial values of the buffer m and machine m, respectively. The inflow (release rate) $\mu_0(t)$ to the flow line represents the boundary value for the inflow. Note that breakdowns and repairs can be incorporated into model (4.14) by allowing the flux $\mu_m^{\max}$ to be dependent on the state of the machine and time as shown in the following equation

$$\mu_m^{\max}(t,s) = \mu_m^{\max} \cdot \omega_m(t,s). \tag{4.15}$$

This yields a modified flow function as well as a modification to the in- and out-flux of the buffer, implying that all deterministic capacities $\mu_m^{\max}$ in (4.14) are replaced according to (4.15). Model (4.14) generates a unique solution for the flow line problem. As we will show in Section 4.5, this solution corresponds to the one computed by the MIP model (4.3).

4.4.1 Numerical solution of the continuous model

In this section, we describe how the stochastic flow line model is solved numerically. First, we consider the deterministic flow line model (4.14). In order to solve the differential equations within the flow line model (4.14) numerically, we need to discretize both space and time and consequently, we introduce a spatial grid $0 = x_{m,0}, \ldots, x_{m,j}, \ldots, x_{m,J} = 1$ with $x_{m,j} = x_{m,0} + j\Delta x$ for each machine m and a time grid $0 = t_0, ..., t_i, ..., t_I = T$ with $t_i = t_0 + i\Delta t$, respectively. We define the deterministic time grid

$\mathcal{T} = (t_i)_i$ with

$$\Delta t = \min_{m=0,\dots,M} \left\{ \frac{\Delta x}{|f'_m(\rho)|} \right\} ,$$

where the time step is chosen to fulfill the stability condition (LeVeque, 1992)

$$\frac{\Delta t}{\Delta x} |f'_m(\rho)| \leq 1 \tag{4.16}$$

for all machines $m = 0, \dots, M$. Note that the resolution of each grid may be arbitrarily small as long as stability condition (4.16) is fulfilled and that f_m are piecewise differentiable functions for each m. We choose to use an explicit Euler scheme to solve the buffer equation (4.14b). For the discretization of the transport equation (4.14a) we use a Godunov scheme (LeVeque, 1992). This is a necessary extension to the approaches of Gillespie (1976, 2001) and Göttlich et al. (2011) as the modeling of finite buffer capacities leads to waves of negative speed, which have to be covered by the numerical scheme.

We now turn to the solution of the stochastic flow line model, i.e., model (4.14) with $\mu_m^{\max}(t, s)$ according to (4.15). Randomness occurs only at the points in time, where the states of a machine switch, the so-called switching points. The random variable $\omega_m(\cdot, s)$ that describes the switches between machine states is a step-function with a finite number of jumps in $[0 = t_0, T]$ for almost every sample s. Thus, for each sample $s \in \mathcal{S}$, we obtain a sequence $\overline{\mathcal{T}}^s = \{\bar{t}_i^s\}_i$ of switching points $\bar{t}_i^s$ such that almost every time there is a finite number $\Omega = \Omega(s)$ of state transitions, i.e., $t_0 = \bar{t}_0^s < \bar{t}_1^s < \dots < \bar{t}_\Omega^s = T$. Furthermore, $\omega(\cdot, s)$ is constant on each interval $\left[\bar{t}_i^s, \bar{t}_{i+1}^s\right)$, namely $\omega(t, s) = \omega(\bar{t}_i^s, s)$ for all $\bar{t}_i^s \leq t < \bar{t}_{i+1}^s$ and for all $i \geq 0$. Between the switching points the flow line behaves in a completely deterministic manner. Consequently, the solution of stochastic version of model (4.14) can be computed with a deterministic solver within an interval $\left[\bar{t}_i^s, \bar{t}_{i+1}^s\right)$ (see Algorithm 4.4.1).

This concept is known as a piecewise deterministic process (PDP), which was first formulated by Davis (1984, 1993). For one sample, such a PDP can be solved by the following stochastic simulation algorithm (SSA) (Gillespie (1976, 2001); Göttlich et al. (2011), see Algorithm 4.4.2). The algorithm first samples the next switching point and then uses a deterministic numerical solver to compute the solution of model (4.14) between two switching points. The initial values for each step are given by the final values of the former step.

To calculate a solution for the set of $\mathcal{S}$ samples, the deterministic time grid $\mathcal{T}$ is used as a common grid for all realizations. For each single realization

we refine the time grid by adding the sampled switching times. In this way, each switching point $\bar{t}_i^s \in \bar{\mathcal{T}}^s$ is included in the time grid and the stability condition (4.16) is satisfied.

Algorithm 4.4.1 Deterministic solution of model (4.14) in time interval $\left[\bar{t}_i^s, \bar{t}_{i+1}^s\right)$

Require: Sampled switching times $\bar{t}_i^s$ and $\bar{t}_{i+1}^s$; densities $\rho_m(\cdot, \bar{t}_i^s)$ and queues $q_m(\bar{t}_i^s)$ for all $m \in M$; global time grid $\mathcal{T} = (t_i)_i$

Ensure: $\rho_m(\cdot, t)$ and $q_m(t)$ for $\bar{t}_i^s \leq t \leq \bar{t}_{i+1}^s$ and for all $m \in M$

Compute local time grid $(\tilde{t}_l)_{0 \leq l \leq L} = \left\{\bar{t}_i^s, \bar{t}_{i+1}^s\right\} \cup \mathcal{T} \mid \left[\bar{t}_i^s, \bar{t}_{i+1}^s\right] = \left\{\bar{t}_i^s, t_i, t_{i+1}, \ldots, t_{i+L-2}, \bar{t}_i^s\right\}$

for $l = 0$ **to** $L - 1$ **do**

  for $m = 0$ **to** M **do**

    $\hat{\gamma}_m(t_l) = \min\left\{\mu_m^{\max}, 2\mu_m^{\max} - v_m \rho_m(0, t_l)\right\}$, $\Delta t_l = t_{l+1} - t_l$

    Compute $\gamma_m^{\text{out}}(t_l)$ according to (4.14d)

    Compute $\gamma_m^{\text{in}}(t_l)$ according to (4.14c)

    Do Euler step for queue: $q_m(t_{l+1}) = q_m(t_l) + \Delta t_l \left(\gamma_m^{\text{in}}(t_l) - \gamma_m^{\text{out}}(t_l)\right)$

    Set in-/outflow of queue as boundary conditions at t_l:

$$\rho_m(x_{-1}, t_l) = \begin{cases} (2\mu_m^{\max} - \gamma_m^{\text{out}}(t_l))/v_m & \rho_m(0, t_l) > \sigma_m \vee f_m(\rho_m(0, t_l)) = \gamma_m^{\text{out}}(t_l) \\ \gamma_m^{\text{out}}(t_l)/v_m & \text{otherwise} \end{cases}$$

$$\rho_m(x_{J+1}, t_l) = \begin{cases} \gamma_{m+1}^{\text{in}}(t_l)/v_m & \rho_m(1, t_l) < \sigma_m \vee f\left(\rho_m(1, t_l)\right) = \gamma_{m+1}^{\text{in}}(t_l) \\ \left(2\mu_m^{\max} - \gamma_{m+1}^{\text{in}}(t_l)\right)/v_m & \text{otherwise} \end{cases}$$

    for $j = 0$ **to** J **do**

$$\rho_m(x_j, t_{l+1}) = \rho_m(x_j, t_l) - \Delta t_l / \Delta x \times \left(F\left(\rho_m(x_j, t_l), \rho_m(x_{j+1}, t_l)\right) - F\left(\rho_m(x_{j-1}, t_l), \rho_m(x_j, t_l)\right)\right)$$

with the so-called *Godunov Flux* $F(\rho_l, \rho_r) = \begin{cases} \min_{\rho \in [\rho_l, \rho_r)} f(\rho) & \rho_l \leq \rho_r \\ \max_{\rho \in [\rho_r, \rho_l)} f(\rho) & \rho_l \geq \rho_r \end{cases}$

    end for

  end for

end for

Algorithm 4.4.2 Stochastic simulation algorithm

Require: $[t_0, T]$ real interval. Initial data for $t = t_0$.

Ensure: One realization of stochastic flow line model (4.14) on $[t_0, T]$.

while $\bar{t}_i^s < T$ **do**

  Sample next switching point $\bar{t}_{i+1}^s$

  Compute solution in the interval $\left[\bar{t}_i^s, \bar{t}_{i+1}^s\right)$ according to Algorithm 4.4.1.

  Set $i = i + 1$.

end while

4.5 Link between MIP and continuous model

We now explain the connection between the MIP and the numerical solution of (4.14). The key idea is to link the buffer equation (4.14b) to Equation (4.3b) by freezing the stochastic terms. Note that the MIP is piecewise deterministic and thus it suffices to consider the freezed stochastic terms, provided the discrete-time interval is chosen sufficiently small, such that every switching point is also a discretization point. Recalling the need for a discretization in space and time for the numerical solution of (4.14), we reuse the equidistant spatial grid defined in Section 4.4.1, i.e., $x_{m,j} = j \cdot \Delta x$ with $j = 0, \ldots, J$ for machine m with $J \geq 1$. To simplify the notation of the continuous-time model (4.14), we assume that the velocity on each machine m is fixed at $v_m = v = 1$.

For the following discussion, we use a coarse discretization with $\Delta x = 1$ for each machine, i.e., $x_0 = 0$ and $x_1 = 1$ are the only two discretization points. We abbreviate the density in these points by $\rho^i_{m,0} = \rho_m(0, t_i)$ and $\rho^i_{m,1} = \rho_m(1, t_i)$. For this discretization, we must distinguish between a free flow and a congested flow line. We first consider the case of free flow, $\rho^i_{m,j} \leq \sigma_m$, $j = 0, 1$. Using a left-handed upwind scheme with the stability condition (4.16), where $f'_m(\rho) = \pm v$, we obtain from Equation (4.14a)

$$\begin{aligned} &\frac{\rho^{i+1}_{m,1} - \rho^i_{m,1}}{\Delta t} + v \, \frac{\rho^i_{m,1} - \rho^i_{m,0}}{\Delta x} = 0 \\ &\qquad \iff \quad \rho^{i+1}_{m,1} = \rho^i_{m,1} - \frac{\Delta t}{\Delta x} v \left(\rho^i_{m,1} - \rho^i_{m,0} \right). \end{aligned}$$

We now choose the time step $\Delta t = \Delta x / v$ fulfilling (4.16) for the equality case. Doing so, the above equations becomes

$$\rho^{i+1}_{m,1} = \rho^i_{m,0}. \tag{4.17}$$

Second, we consider the congested flow line, i.e., $\rho^i_{m,j} > \sigma_m$, $j = 0, 1$. Using a right-handed, upwind scheme (LeVeque, 1992), we obtain

$$\begin{aligned} &\frac{\rho^{i+1}_{m,0} - \rho^i_{m,0}}{\Delta t} - v \, \frac{\rho^i_{m,1} - \rho^i_{m,0}}{\Delta x} = 0 \\ &\qquad \iff \quad \rho^{i+1}_{m,0} = \rho^i_{m,0} + \frac{\Delta t}{\Delta x} v \left(\rho^i_{m,1} - \rho^i_{m,0} \right). \end{aligned}$$

Again, the stability condition (4.16) must hold, and by choosing $\Delta t = \Delta x / v$,

we find that

$$\rho_{m,0}^{i+1} = \rho_{m,1}^{i}. \tag{4.18}$$

Straightforward Euler discretization (LeVeque, 1992) of Equation (4.14b) to model the filling of the buffer yields

$$\frac{q_m^{i+1} - q_m^i}{\Delta t} = \gamma_m^{\text{in},i} - \gamma_m^{\text{out},i}, \tag{4.19}$$

where $q_m^i = q_m(t_i)$, $\gamma_m^{\text{in},i} = f_{m-1}\left(\rho_{m-1,1}^i\right)$ and $\gamma_m^{\text{out},i} = f_m\left(\rho_{m,0}^i\right)$. Using (4.17) we can substitute the last term in Equation (4.19) for the free flow case as

$$\gamma_m^{\text{out},i} = f_m\left(\rho_{m,1}^{i+1}\right). \tag{4.20}$$

In order to find a representation for $\gamma_m^{\text{out},i}$ when congestion is present on machine m at time t_i (i.e., $\rho_{m,0}^i > \sigma_m$), there are two cases to be discussed: Congestion being resolved and congestion enduring. From Equation (4.18), we know that $\rho_{m,1}^{i-1} = \rho_{m,0}^i > \sigma_m$ holds. In the first case the congestion is resolved in time step t_i, and thus $\rho_{m,1}^i = 2\sigma_m - \rho_{m,1}^{i-1} < \sigma_m$ (see Figure 4.5). The update for the next time step is done according to (4.17), yielding $\rho_{m,1}^{i+1} = \rho_{m,0}^i$. Thus, the flow values are found as

$$f_m\left(\rho_{m,1}^{i+1}\right) = f_m\left(\rho_{m,0}^i\right). \tag{4.21}$$

The second case is that congestion persists on in time step t_i, and hence, we get $\rho_{m,1}^i = \rho_{m,1}^{i-1} > \sigma_m$ and $\rho_{m,0}^{i+1} = \rho_{m,1}^i$ (see (4.18)). Consequently, the density $\rho_{m,0}^{i+1}$ at time step t_{i+1} has to be considered. This yields two subcases: Either the congestion is still present in time step t_{i+1}, implying that $\rho_{m,1}^{i+1} = \rho_{m,1}^i = \rho_{m,1}^{i-1} = \rho_{m,0}^i$ and yielding $f_m\left(\rho_{m,1}^{i+1}\right) = f_m\left(\rho_{m,0}^i\right)$, or the congestion is resolved in time step t_{i+1}. In the latter subcase, we obtain $\rho_{m,1}^{i+1} = 2\sigma_m - \rho_{m,1}^i < \sigma_m$, and, consequently, the flow values are related as follows

$$f_m\left(\rho_{m,1}^{i+1}\right) = f_m\left(\rho_{m,1}^i\right) = f_m\left(\rho_{m,1}^{i-1}\right) = f_m\left(\rho_{m,0}^i\right). \tag{4.22}$$

Hence, for any case, we may rewrite (4.19) in terms of the flow values, yielding

$$q_m^{i+1} = q_m^i + \Delta t\left(f_{m-1}\left(\rho_{m-1,1}^i\right) - f_m\left(\rho_{m,1}^{i+1}\right)\right). \tag{4.23}$$

With respect to the balance equation (4.3b), the following theorem is stated.

Theorem 4.5.1. *Let* $WIP^{b}_{m,i} = q^{i+1}_{m}$ *and* $\tau^{-1} \cdot TH_{m,i} = f\left(\rho^{i}_{m,1}\right)$ *and* $\Delta x = 1$. *Then, each numerical solution of the time-continuous model* (4.14) *is also a solution for the discrete-time model* (4.3).

4.6 Numerical evaluation of the approximation approaches

4.6.1 Performance measures

Performance measures of interest are the expected work in process $\mathrm{E}[\,WIP_m(t)]$ at each station m and the expected cumulated output of the line $\mathrm{E}[\,TH^{c}(t)]$.

The performance calculation for the discrete-time model is as follows: The WIP on machine $m = 0$ is always 1 as it does not suffer from starvation. For the remaining stations $m = 1, 2, \ldots, M$ (machine and buffer in front of the machine), the expected WIP at the end of interval i can be determined by the difference of the cumulated production of machine $m-1$ and the cumulated production of machine m up to the end of interval i

$$\mathrm{E}[\,WIP_{m,i}] = \begin{cases} 1 & m = 0, \\ \frac{1}{S} \sum\limits_{s=1}^{S} \sum\limits_{i'=1}^{i} \left(TH_{m-1,i',s} - TH_{m,i',s} \right) & m = 1, 2, \ldots, M. \end{cases} \tag{4.24}$$

The expected number of produced workpieces up to the end of interval i (i.e., the expected cumulated output of the line) is given by

$$\mathrm{E}[\,TH^{c}_{i}] = \frac{1}{S} \sum_{s=1}^{S} \sum_{i'=1}^{i} TH_{M,i',s}. \tag{4.25}$$

In contrast, the performance measures for the continuous-time model are calculated in the following way: The expected WIP on station m at time t is given as the inventory level in the buffer plus the difference of the workpieces having entered machine m up to time t and those having already left machine m by time t, yielding

$$\mathrm{E}[WIP_m(t)]$$
$$= \frac{1}{S}\sum_{s=1}^{S}\left(q_m(t,s) + \int_0^t f_m(\rho_m(0,\tilde{t},s))\ \mathrm{d}\tilde{t} - \int_0^t f_m(\rho_m(1,\tilde{t},s))\ \mathrm{d}\tilde{t}\right). \tag{4.26}$$

Furthermore, the expected number of workpieces produced by time t is computed as the cumulated outflow from the last machine M

$$\mathrm{E}[TH^c(t)] = \frac{1}{S}\sum_{s=1}^{S}\int_0^t f_M(\rho_M(1,\tilde{t},s))\ \mathrm{d}\tilde{t}. \tag{4.27}$$

For $t = t_i$ being the end of interval i, we are able to compare (4.24) to (4.26) as well as (4.25) to (4.27).

4.6.2 Case I: Increase of the release rate

To evaluate the accuracy of the proposed approximations and to gain initial insights regarding the time-dependent performance of unreliable flow lines, we first analyze two cases of three-machine lines in detail and subsequently 54 cases with lines of different length in Section 4.6.4.

For both cases with $M = 2$, the buffer $m = 1$ is infinite, and the buffer $m = 2$ is finite with capacity $b_2 = 5$. Machine $m = 0$ is assumed to be reliable, whereas machines $m > 0$ fail randomly with exponentially distributed time between failures and repair times. The system is observed for a finite horizon of 200 time units.

In the first case, the rate of the reliable machine $m = 0$ increases from 0.75 for $t < 100$ to 0.9 for $100 \leq t \leq 200$. Machine $m = 0$ starts as the bottleneck. After the change in the processing rate at $t = 100$, all machines are bottlenecks if the unreliability of machines $m = 1, 2$ is considered (see Table 4.2).

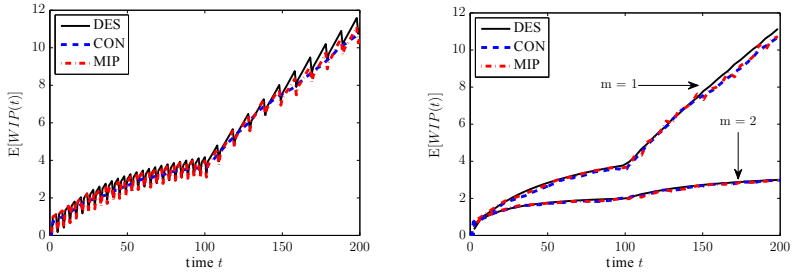
For the continuous model the velocity and maximal capacity is set equal to 1 on all machines. Four discretization points in space and time are used, giving $\Delta x = \Delta t = 1/4$. The length of the discretization interval for the discrete-time approach is $\tau = 1$ and $\alpha = 10$. A set of $S = 1,000$ samples is used for both of the approximation approaches and all test cases. The DES includes 1 million replications, which ensures tight confidence intervals.

Table 4.2: Parameters for an increasing release rate $\mu_0(t)$

	Test case parameters				Approximation	
m	b_m	d_m	r_m	μ_m	v_m	$\mu_m^{\max}$
0	—	0	∞	0.75 | 0.9	1	1
1	∞	$1/45$	$1/5$	1	1	1
2	5	$1/45$	$1/5$	1	1	1

Figure 4.7a depicts the expected work in process $\mathrm{E}[WIP_1(t)]$ of station 1. The oscillations in the simulation are caused by correlations of the output of the machines for different samples due to the deterministic and discrete material outflow of machine $m = 0$. The MIP approach exhibits oscillations similar to those of the DES. These types of oscillations do not occur for the continuous model (CON) as the density of workpieces constantly leaves the machines. For the sake of comparability to the continuous model, we compute a time average over multiple periods for the MIP and DES. We average values over four time units, which is the least common multiple of all processing times. It also represents the longest processing time of all machines.

Figure 4.7b demonstrates that the averaged work in process $\mathrm{E}[WIP_m(t)]$ is well approximated for both stations. Workpieces start to accumulate at both stations and do not reach a steady state during the first 100 time units. The maximum absolute deviation to the simulation is 0.6782 workpieces (0.1667 workpieces) for the MIP and 0.4561 workpieces (0.1086 workpieces) for the continuous model in station 1 (station 2).

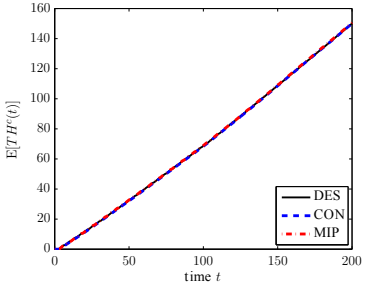


(a) Expected work in process $\mathrm{E}[WIP_1(t)]$ at station 1 without averaging

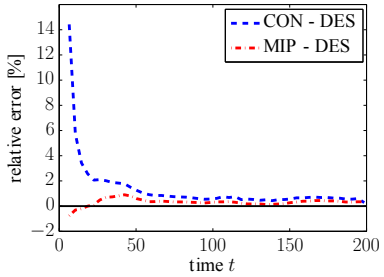
(b) Expected work in process $\mathrm{E}[WIP_m(t)]$ at both stations with averaging

Figure 4.7: Expected WIP over time for an increase of the release rate

Figure 4.8a reveals that both approximations of the expected cumulated output $\mathrm{E}[TH^c(t)]$ match very well with the simulated results. The relative errors of the approximations to the simulation in Figure 4.8b show that the continuous model slightly overestimates the throughput for the first periods. This initial overestimation is due to the approximation of continuous flow. The density of workpieces constantly leaves the machine, even before the minimum lead time of $\sum_{m'=1}^{m} \frac{1}{\mu_{m'}}$.



(a) Expected cumulated output of last machine $\mathrm{E}[TH^c(t)]$ over time



(b) Relative errors of $\mathrm{E}[TH^c(t)]$

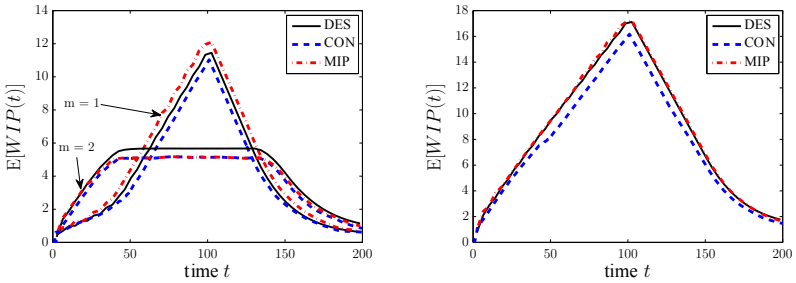
Figure 4.8: Expected cumulated output and its relative error over time for an increase of the release rate

4.6.3 Case II: Decrease of the release rate

For the second case, we consider a decrease of the release rate μ_0. The remaining parameters are chosen such that the bottleneck shifts from machine $m = 2$ to machine $m = 0$ after $t = 100$ (see Table 4.3).

Table 4.3: Parameters for a decreasing release rate $\mu_0(t)$

	Test case parameters				Approximation	
m	b_m	d_m	r_m	μ_m	v_m	$\mu_m^{\max}$
0	—	0	∞	0.625 | 0.25	1	1
1	∞	1/45	1/5	1	1	1
2	5	1/95	1/5	0.5	1	0.5



(a) Expected work in process $\mathrm{E}[WIP_m(t)]$ for stations 1 and 2

(b) Expected work in process $\sum_{m=1}^{2} \mathrm{E}[WIP_m(t)]$ on station 1 and 2

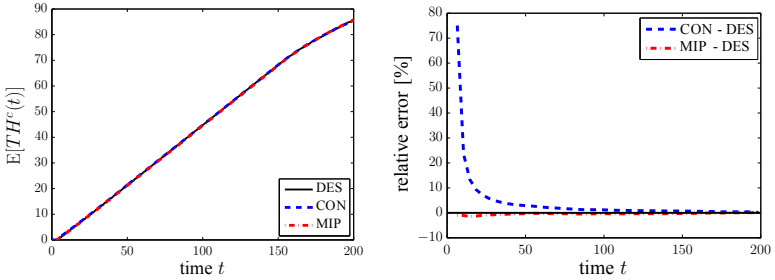
Figure 4.9: Expected WIP over time for a decrease of the release rate

Figures 4.9a and 4.9b depict the expected work in process $\mathrm{E}[WIP_m(t)]$ for the last two stations. The bottleneck is located at machine $m = 2$ for $t < 100$, which leads to a high WIP for buffer $m = 2$. This causes a high probability of blocking for machine $m = 1$, and hence, an increasing WIP for station 1. The decrease of the rate μ_0 resolves the blocking issue in this case for machine $m = 1$, and the inventory levels in both buffers decrease as a result.

The MIP approach overestimates $\mathrm{E}[WIP_1(t)]$ and underestimates $\mathrm{E}[WIP_2(t)]$, while the overall WIP is well approximated. Notwithstanding the approximation errors mentioned in Section 4.3, this shift in the WIP may be explained by the underestimation of the WIP on machine 2 in this example. Based on the assumption of the inventory balance equation (4.3b), a workpiece is taken from the buffer at the end of the period before the production finishes. The time that workpieces spend on $m = 2$ is underestimated by exactly one time unit as the length of the discretization interval $\tau = 1$ is just half of a processing time $1/\mu_2$. If the buffer capacity b_2 is reached, the underestimation of the time that workpieces spend on machine $m = 2$ leads to an overestimation of the time workpieces spend on station 1. This in turn leads to the observed overestimation of the inventory at station 1.

The continuous model results in a similar error for machine $m = 2$. It reduces the maximum flow through the machine $m = 2$, such that the model predicts that fewer workpieces will be present on the machine compared to the actual system.

The cumulated output of the flow line is well approximated by both approaches (see Figure 4.10a). The cumulated output increases at a constant rate shortly after the start of production. The throughput rate of machine $m = 2$ drops after $t > 150$. This late reaction relative to the decrease of the rate $\mu_0(t)$ is based on a high expected WIP level at time $t = 100$. We observe again an overestimation for the continuous model at the beginning of the time horizon $(t < 40)$; see the relative errors in Figure 4.10b. For later time values, this error tends to zero. The MIP approximates the cumulated output with a maximum relative error of only 1.4%.



(a) Expected cumulated output of last machine $\mathrm{E}[TH^c(t))]$

(b) Relative errors of $\mathrm{E}[TH^c(t)]$

Figure 4.10: Expected cumulated output and its relative error over time for a decrease of the release rate

4.6.4 Impact of the number of machines and buffer capacities

In the following, we evaluate the impact of the number of machines M in the flow line, the buffer capacities b_m, and different time-dependent release rates $\mu_0(t)$ on the expected cumulated output E[$TH^c(t)$] of the line. We consider a set of stepwise constant rate functions $\mu_0(t)$, which increases to a maximum rate h and subsequently decreases to 0 in the time interval $[0, 700]$. They are characterized by step lengths $l \in \{100,\ 140,\ 233.33\}$ and maximal release rates $h \in \{0.8,\ 1.0,\ 1.5\}$. Figure 4.11 depicts an example with $l = 140$ and $h = 1.0$. Two different buffer capacities $b_m \in \{5,\ 20\}$ and three different numbers of machines $M \in \{3,\ 10,\ 20\}$ are tested. For all configurations, machines $m > 0$ have the processing rate $\mu_m = 1$, failure rate $d_m = 1/45$, and repair rate $r_m = 1/5$. All in all, 54 test cases are considered.

The discretization interval for the discrete-time approach is set to $\tau = \min_m\{1/\mu_m(t)\}$. For the continuous model, we set the velocity and maximal flux $v_m = \mu_m^{\max} = 1$ for all machines $m > 0$. The numerical solution scheme uses discretization intervals of $\Delta x = \Delta t = 1/4$.

Table 4.4 summarizes the results of the expected cumulated output at times $t \in \{100,\ 350,\ 700,\ 800\}$ for all 54 test cases. The relative deviations of the MIP and the continuous approach to a DES with $S = 1,000,000$ replications are given in brackets. With respect to approximation quality, Table 4.4 suggests a close match of both approximations to the DES. The average absolute deviations to the DES are 0.81% for the MIP (4.3) and 0.47% for the continuous model (4.14), respectively.

Table 4.4: E[$TH^c(t)$] obtained with CON and MIP approach and relative deviation to DES for different values of h, l, b_m, M

Test cases				E[TH^c](100)		E[TH^c](350)		E[TH^c](700)		E[TH^c](800)	
h	l	b_m	M	CON	MIP	CON	MIP	CON	MIP	CON	MIP
0.8	100	5	3	18.75 (1.14%)	18.62 (0.44%)	148.46 (- 0.14%)	146.87 (- 1.21%)	310.33 (0.14%)	307.65 (- 0.73%)	311.61 (0.05%)	309.12 (- 0.75%)
0.8	100	5	10	16.18 (0.78%)	16.24 (1.17%)	129.78 (- 0.68%)	128.94 (- 1.32%)	304.58 (0.19%)	300.57 (- 1.13%)	308.49 (0.08%)	304.76 (- 1.14%)
0.8	100	5	20	12.45 (1.27%)	12.69 (3.18%)	110.43 (- 0.81%)	110.42 (- 0.82%)	299.82 (0.15%)	295.96 (- 1.14%)	308.38 (0.15%)	304.22 (- 1.21%)
0.8	100	50	3	18.77 (1.23%)	18.65 (0.58%)	149.57 (- 0.43%)	148.47 (- 1.16%)	318.58 (0.01%)	317.54 (- 0.31%)	319.85 (- 0.05%)	319.00 (- 0.31%)
0.8	100	50	10	16.12 (0.39%)	16.19 (0.82%)	131.26 (- 0.62%)	130.84 (- 0.94%)	316.05 (0.07%)	314.90 (- 0.30%)	319.86 (- 0.04%)	319.00 (- 0.31%)
0.8	100	50	20	12.47 (1.38%)	12.64 (2.81%)	110.80 (- 1.43%)	111.12 (- 1.15%)	311.34 (- 0.08%)	310.69 (- 0.29%)	319.86 (- 0.04%)	319.00 (- 0.31%)
0.8	140	5	3	24.88 (1.63%)	24.47 (- 0.07%)	155.94 (0.53%)	155.14 (0.00%)	324.57 (0.30%)	322.92 (- 0.21%)	326.32 (0.11%)	324.68 (- 0.39%)
0.8	140	5	10	21.21 (- 0.24%)	21.38 (0.56%)	136.55 (0.22%)	136.43 (0.13%)	317.40 (0.46%)	314.51 (- 0.45%)	322.79 (0.26%)	319.90 (- 0.63%)
0.8	140	5	20	15.97 (0.07%)	16.29 (2.03%)	116.72 (- 0.13%)	117.31 (0.38%)	309.80 (0.17%)	306.90 (- 0.77%)	321.93 (0.13%)	318.29 (- 1.00%)
0.8	140	50	3	24.97 (1.93%)	24.59 (0.35%)	156.96 (0.24%)	156.72 (0.09%)	334.17 (0.25%)	334.26 (0.28%)	335.87 (- 0.04%)	336.00 (0.00%)
0.8	140	50	10	21.23 (- 0.21%)	21.42 (0.65%)	137.81 (0.10%)	138.13 (0.33%)	330.40 (0.21%)	330.50 (0.24%)	335.87 (- 0.04%)	336.00 (0.00%)
0.8	140	50	20	15.84 (- 0.76%)	16.25 (1.81%)	118.22 (0.13%)	119.11 (0.89%)	323.90 (0.18%)	324.42 (0.34%)	335.88 (- 0.03%)	336.00 (0.00%)
0.8	233.33	5	3	36.90 (0.29%)	36.68 (- 0.29%)	172.24 (- 0.05%)	171.74 (- 0.34%)	357.14 (0.13%)	355.02 (- 0.47%)	360.03 (- 0.01%)	358.88 (- 0.33%)
0.8	233.33	5	10	30.82 (- 0.25%)	31.31 (1.31%)	150.19 (0.13%)	150.36 (0.24%)	343.61 (0.21%)	340.39 (- 0.73%)	352.87 (0.13%)	349.96 (- 0.69%)
0.8	233.33	5	20	21.73 (- 1.31%)	22.59 (2.57%)	127.60 (- 0.87%)	128.62 (- 0.08%)	330.45 (0.00%)	327.54 (- 0.88%)	351.61 (0.08%)	347.64 (- 1.05%)
0.8	233.33	50	3	37.12 (0.52%)	37.01 (0.23%)	174.17 (0.09%)	174.21 (0.12%)	370.43 (0.14%)	370.04 (0.03%)	373.29 (0.08%)	373.99 (0.27%)
0.8	233.33	50	10	30.99 (- 0.03%)	31.51 (1.64%)	152.15 (0.36%)	152.83 (0.81%)	363.76 (- 0.03%)	364.02 (0.04%)	373.30 (0.08%)	374.00 (0.27%)
0.8	233.33	50	20	21.54 (- 2.41%)	22.46 (1.76%)	129.46 (- 0.33%)	130.85 (0.75%)	350.27 (- 0.14%)	351.33 (0.16%)	373.26 (0.08%)	373.97 (0.27%)
1	100	5	3	23.42 (0.44%)	23.43 (0.48%)	177.37 (- 0.24%)	176.26 (- 0.86%)	372.24 (- 0.09%)	369.79 (- 0.75%)	373.88 (- 0.17%)	371.48 (- 0.81%)
1	100	5	10	20.10 (0.88%)	20.18 (1.30%)	151.65 (- 0.21%)	151.21 (- 0.49%)	354.37 (0.23%)	349.72 (- 1.08%)	359.29 (0.13%)	354.65 (- 1.17%)
1	100	5	20	15.08 (- 0.04%)	15.45 (2.46%)	126.12 (- 1.07%)	126.54 (- 0.74%)	339.43 (- 0.16%)	335.65 (- 1.28%)	356.73 (0.28%)	349.85 (- 1.65%)
1	100	50	3	23.37 (0.20%)	23.41 (0.35%)	179.62 (- 0.28%)	178.94 (- 0.66%)	398.19 (- 0.01%)	397.33 (- 0.23%)	399.76 (- 0.05%)	398.94 (- 0.26%)
1	100	50	10	19.95 (0.12%)	20.07 (0.74%)	153.81 (- 0.47%)	153.80 (- 0.48%)	393.68 (0.01%)	392.93 (- 0.18%)	399.78 (- 0.05%)	398.96 (- 0.25%)
1	100	50	20	15.06 (- 0.18%)	15.45 (2.41%)	128.62 (- 0.62%)	129.41 (- 0.01%)	366.25 (- 0.43%)	367.23 (- 0.16%)	399.56 (- 0.04%)	398.80 (- 0.22%)

Table 4.4: E[$TH^c(t)$] obtained with CON and MIP approach and relative deviation to DES for different values of h, l, b_m, M - continued

1	140	5	3	31.04	(0.18%)	31.04	(0.19%)	185.19	(- 0.01%)	185.00	(- 0.11%)	387.29	(0.00%)	385.90	(- 0.35%)	389.66	(- 0.13%)	388.94	(- 0.31%)
1	140	5	10	26.05	(- 0.75%)	26.38	(0.50%)	159.06	(- 0.06%)	159.36	(0.12%)	367.75	(0.44%)	364.22	(- 0.53%)	375.05	(0.32%)	372.08	(- 0.48%)
1	140	5	20	19.22	(0.04%)	19.89	(3.53%)	133.50	(- 0.69%)	134.72	(0.22%)	348.01	(- 0.23%)	345.30	(- 1.01%)	370.25	(- 0.04%)	365.02	(- 1.45%)
1	140	50	3	31.06	(0.03%)	31.16	(0.36%)	187.45	(- 0.24%)	187.85	(- 0.02%)	417.37	(0.06%)	417.85	(0.18%)	419.74	(- 0.05%)	420.89	(0.23%)
1	140	50	10	26.12	(- 0.60%)	26.54	(0.98%)	162.25	(0.09%)	163.29	(0.73%)	410.08	(0.06%)	410.80	(0.23%)	419.79	(- 0.03%)	420.94	(0.24%)
1	140	50	20	19.03	(- 1.07%)	19.66	(2.21%)	135.24	(- 1.04%)	136.99	(0.24%)	374.47	(- 0.45%)	376.48	(0.09%)	418.55	(- 0.13%)	419.83	(0.18%)
1	233.33	5	3	45.68	(- 0.18%)	45.73	(- 0.07%)	201.72	(- 0.08%)	201.85	(- 0.01%)	419.78	(0.13%)	418.33	(- 0.22%)	423.55	(0.03%)	421.93	(- 0.35%)
1	233.33	5	10	36.75	(- 1.03%)	37.05	(- 0.23%)	172.86	(- 0.32%)	173.27	(- 0.08%)	389.87	(0.08%)	389.31	(- 0.06%)	402.81	(- 0.02%)	402.23	(- 0.16%)
1	233.33	5	20	24.85	(- 1.98%)	25.76	(1.62%)	148.39	(- 0.57%)	149.50	(0.17%)	364.49	(- 0.10%)	362.26	(- 0.72%)	399.84	(0.16%)	393.80	(- 1.36%)
1	233.33	50	3	45.69	(- 0.16%)	45.96	(0.43%)	204.46	(- 0.15%)	205.32	(0.27%)	462.01	(- 0.04%)	462.43	(0.05%)	466.10	(- 0.08%)	466.46	(- 0.01%)
1	233.33	50	10	36.84	(- 0.94%)	37.54	(0.94%)	176.60	(- 0.32%)	177.95	(0.44%)	440.16	(- 0.12%)	441.47	(0.18%)	466.12	(- 0.08%)	466.50	(0.00%)
1	233.33	50	20	24.51	(- 2.19%)	25.53	(1.88%)	150.80	(- 0.56%)	152.71	(0.70%)	389.99	(- 0.29%)	392.52	(0.35%)	456.86	(- 0.18%)	458.32	(0.14%)
1.5	100	5	3	34.79	(0.10%)	34.85	(0.26%)	220.74	(0.09%)	219.99	(- 0.25%)	458.31	(0.03%)	454.35	(- 0.84%)	460.81	(- 0.13%)	457.23	(- 0.91%)
1.5	100	5	10	29.07	(- 0.20%)	29.66	(1.84%)	183.28	(0.19%)	182.25	(- 0.37%)	414.26	(0.18%)	404.56	(- 2.16%)	424.63	(0.18%)	413.73	(- 2.39%)
1.5	100	5	20	20.70	(- 1.49%)	21.84	(3.93%)	151.69	(- 0.29%)	152.00	(- 0.09%)	373.92	(- 0.33%)	366.41	(- 2.33%)	414.09	(0.10%)	398.91	(- 3.57%)
1.5	100	50	3	34.81	(- 0.18%)	34.98	(0.30%)	225.28	(0.14%)	225.71	(0.33%)	514.85	(- 1.38%)	523.18	(0.22%)	523.09	(- 3.32%)	541.24	(0.04%)
1.5	100	50	10	29.15	(- 0.14%)	29.84	(2.22%)	188.56	(- 0.17%)	189.82	(0.50%)	460.02	(- 0.21%)	461.21	(0.05%)	520.02	(- 2.15%)	531.44	(0.00%)
1.5	100	50	20	20.91	(- 0.70%)	22.12	(5.01%)	155.67	(- 0.36%)	157.89	(1.07%)	404.30	(- 0.27%)	406.67	(0.31%)	479.91	(- 0.04%)	481.86	(0.36%)
1.5	140	5	3	45.66	(- 0.22%)	45.81	(0.11%)	230.45	(- 0.22%)	229.35	(- 0.70%)	477.25	(- 0.04%)	472.82	(- 0.97%)	480.94	(- 0.10%)	476.23	(- 1.07%)
1.5	140	5	10	36.85	(- 0.76%)	37.62	(1.31%)	191.08	(- 0.01%)	189.70	(- 0.73%)	425.33	(0.16%)	415.81	(- 2.08%)	440.96	(0.30%)	429.04	(- 2.41%)
1.5	140	5	20	24.78	(- 2.26%)	26.07	(2.86%)	158.43	(- 0.76%)	158.53	(- 0.69%)	381.51	(- 0.39%)	373.97	(- 2.36%)	428.22	(0.09%)	412.61	(- 3.56%)
1.5	140	50	3	45.92	(- 0.39%)	46.37	(0.59%)	235.64	(- 0.30%)	235.87	(- 0.20%)	533.35	(- 0.46%)	534.79	(- 0.19%)	554.25	(- 2.21%)	563.67	(- 0.54%)
1.5	140	50	10	36.83	(- 1.36%)	37.80	(1.25%)	196.86	(- 0.55%)	197.95	(0.00%)	470.37	(- 0.22%)	471.42	(0.00%)	543.94	(- 0.67%)	546.70	(- 0.16%)
1.5	140	50	20	24.71	(- 2.96%)	26.05	(2.31%)	164.02	(- 0.27%)	166.00	(0.94%)	414.12	(- 0.14%)	416.40	(0.41%)	489.68	(0.00%)	491.44	(0.36%)
1.5	233.33	5	3	64.92	(- 0.39%)	65.36	(0.28%)	251.04	(- 0.32%)	250.05	(- 0.71%)	515.10	(- 0.13%)	511.08	(- 0.91%)	522.23	(- 0.14%)	517.66	(- 1.01%)
1.5	233.33	5	10	46.77	(- 1.41%)	47.71	(0.58%)	211.13	(- 0.77%)	209.12	(- 1.71%)	450.05	(- 0.47%)	441.35	(- 2.39%)	482.21	(- 0.06%)	468.89	(- 2.82%)
1.5	233.33	5	20	29.05	(- 3.47%)	30.53	(1.45%)	177.26	(- 0.76%)	176.37	(- 1.26%)	399.65	(- 0.80%)	390.97	(- 2.95%)	460.03	(- 0.34%)	445.27	(- 3.54%)
1.5	233.33	50	3	65.35	(- 0.75%)	66.13	(0.43%)	258.26	(- 0.40%)	259.19	(- 0.04%)	554.51	(- 0.53%)	556.97	(- 0.09%)	582.30	(- 3.82%)	605.02	(- 0.06%)
1.5	233.33	50	10	47.14	(- 1.56%)	48.30	(0.86%)	222.27	(- 0.31%)	223.91	(0.43%)	494.73	(- 0.12%)	496.46	(0.23%)	569.97	(- 0.79%)	575.69	(0.20%)
1.5	233.33	50	20	29.39	(- 3.00%)	30.97	(2.20%)	185.69	(- 0.46%)	188.20	(0.88%)	438.32	(- 0.11%)	441.15	(0.54%)	514.04	(0.05%)	516.41	(0.51%)

Figure 4.11 depicts the detailed development in the time interval $[0, 800]$ for the cases of $b_m = 5, M = 10, l = 140$, and $h = 1.0$. Both approaches approximate $\mathrm{E}[TH^c(t)]$ very well.

It becomes apparent that the time-dependent behavior of the release rate does not directly translate to the expected cumulated output of the flow line. The discontinuities in the release rate are marked by kinks in the cumulated release rate $\mu_0^c(t) = \int_0^t \mu_0(\tilde{t})d\tilde{t}$. These kinks are less pronounced in the cumulated output curve. Comparisons with the cases $M \in \{3, 20\}$, and $b_m = 50$ reveal that this smoothing effect increases with the number of machines M and with buffer capacities b_m. This demonstrates the impact of the flow line's design on its time-dependent performance.

Next, we compare the case of the time-dependent release rate to a constant release rate. We set $\mu_0(t) = \mu_0^{const} = \int_0^T \frac{\mu_0(t)}{T} dt = 0.6$, such that it corresponds to the average release rate of the time-dependent case. The constant release rate yields a 8.7% higher cumulated output at time $t = 800$ compared to the time-dependent case, which illustrates the relevance of the time-dependent effects in the flow line.

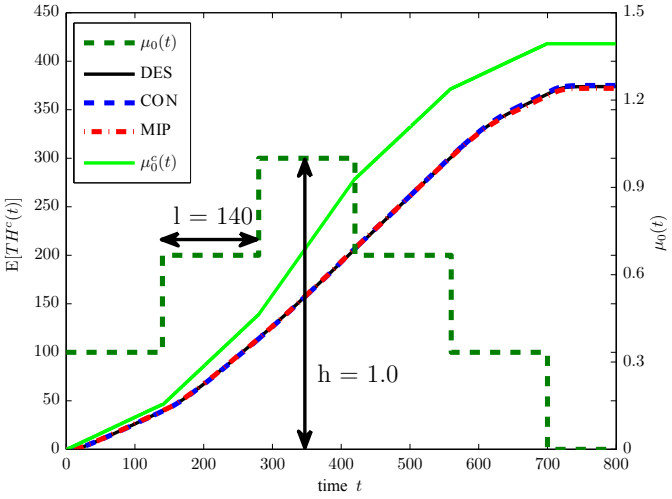


Figure 4.11: Cumulated output $\mathrm{E}[TH^c(t)]$ obtained by the MIP, the continuous model, and the DES for $\mu_0(t)$ with $h = 1.0$, $l = 140$, $M = 10$, and $b_m = 5\ \forall m$

4.7 Conclusion

In this work, two new approximation approaches for the performance evaluation of time-dependent, unreliable flow lines are introduced. For the first time, the MIP sampling approximation is used as an approach to evaluate a non-stationary flow line. The continuous model extends the existing literature on continuous supply chain models via the incorporation of finite buffers and congestion. We demonstrate that both approaches coincide on a deterministic level, which links the two corresponding literature streams.

The discussion on the approximation quality is supported by a numerical study. The numerical study also provides new insights regarding time-dependent effects in flow lines. In particular, the impact of a time-dependent release rate on the cumulated output is quantified. Moreover, the numerical study reveals a time-dependent smoothing effect with respect to the expected cumulated output.

A potential extension to the model is the integration of multiple product types. The proposed model can be used as a basis for this extension if no or only minor setups are required between the processing of different product types. Another direction of future research is the extension to performance measure which reflect, e.g., risk aversion to account for empirical research which suggests that operations managers are not necessarily risk neutral (Bendoly et al., 2006). Both approximation approaches are methodologically closely linked to performance optimization. Thus, future work should be directed towards the optimization of the design and control of flow lines under a time-dependent and stochastic operating environment. Potential objectives include the minimization of necessary buffer capacities or the optimization of the release rate. Moreover, from a methodological point of view, the integration of the stochastic effects without generation of random numbers is a potential field of future research.

5 A sampling approach for the analysis of time-dependent stochastic flow lines

Co-authors:
Raik Stolletz
Chair of Production Management, Business School, University of Mannheim, Germany

Published in:
Proceedings of the 9th Conference on Stochastic Models of Manufacturing and Service Operations, Seeon, Germany, 2013, pages 181-188

Abstract:

Flow lines often operate under stochastic and time-dependent influences, for example during the ramp-up phase. The considered model assumes stochastic processing times with arbitrary distributions which may change over time. Finite buffers are allocated between the machines to compensate for these stochastic effects. A discrete-time sampling approach for the analytical performance evaluation is proposed. The accuracy of the approach is demonstrated by comparison to a discrete-event simulation. Moreover, different time-dependent buffer allocation strategies are proposed and evaluated.

5.1 Introduction

Flow lines often operate under stochastic and time-dependent influences. In the literature, stochastic impacts from random demand, processing times, machine failures, and subsequent repair times are widely acknowledged (Dallery and Gershwin, 1992). In unpaced flow lines usually buffers are placed between adjacent machines in order to limit the negative effects of blocking and starving. A machine starves if it idles due to a lack of raw material and is blocked if a processed workpiece cannot leave the machine due to a full downstream buffer. Jaikumar and Bohn (1992) first describe and discuss that production systems often operate under non-stationary conditions. The time-dependent behavior occurs if the parameters of the random distributions change over time. Reasons are learning effects during the ramp-up (Terwiesch and Bohn, 2001), introduction of new manufacturing technologies, or seasonal demand patterns.

In contrast to the buffer allocation for flow lines under steady-state conditions (Demir et al., 2014), the time-dependent buffer allocation for non-stationary operating environments has received little attention so far. In practice, time-dependent changes of the buffer space configuration are easily realizable within a certain range. This holds especially for many modern production systems that are controlled by a pull mechanism. In pure pull systems the number of Kanban cards corresponds to the available buffer space. Hence, the buffer space is easily adaptable by means of the cards. This paper provides an initial approach for the systematic analysis of time-dependent buffer allocations for stochastic flow lines. This includes a description of unique characteristics for buffer changes over time. Moreover, an sample-based evaluation approach for the performance analysis is proposed, which bears the potential for modifications towards an optimization model. A numerical study demonstrates its accuracy and provides first insights regarding the system behavior.

There is a rich body of literature with respect to the analytical performance evaluation of stochastic flow lines under steady-state conditions (Dallery and Gershwin, 1992). However, the performance evaluation of time-dependent stochastic flow lines with finite buffers has not yet been addressed in an analytical way. Instead, engineering related literature suggests simulation as modeling tool (Fleischer et al., 2004). Fluid and diffusion approximations account for non-stationarity but assume infinitely large buffers (Vandergraft, 1983; Duda, 1986). This assumption neglects the effects of blocking which is characteristic for many manufacturing systems. The fluid approach approximates the discrete flow of goods by a continuous one and assumes deter-

ministic service times. For heavy traffic situations these approximations are justified by the functional law of large numbers. Diffusion approximations further utilize the functional central limit theorem and describe the network behavior by a one dimensional reflected Brownian monition. For the evaluation of a finite buffer followed by a single machine with a time-dependent interarrival time distribution a Stationary Backlog Carry Over (SBC) approach is proposed by Stolletz and Lagershausen (2013).

For a comprehensive survey of solution approaches for the optimization of buffer allocations under steady-state conditions, the reader is referred to Demir et al. (2014). Due to the combinatorial complexity of the problem a variety of heuristics have been applied. Helber et al. (2011) develop a discrete-time sampling approach. They propose a mixed-integer program (MIP) for the optimization of buffer allocations. Related publications utilize a continuous-time sampling approach (Matta, 2008; Alfieri and Matta, 2012). Thereby, optimal buffer allocations with respect to the analyzed samples can be obtained. To the best of our knowledge, in analytical models, the buffer allocation is treated solely as static and strategic decision problem for production systems under steady-state conditions. Merely, practitioners from the semiconductor industry suggest to change the CONWIP level over time to account for time-dependent effects (Haller et al., 2003).

Consequently, we propose an evaluation method for the performance evaluation as an initial step towards the systematic analysis of buffer spaces under non-stationary operating conditions. The remainder is structured as follows: Section 5.2 describes the analyzed system in detail and states the underlying assumptions. In Section 5.3, a sampling approach for the performance evaluation of time-dependent stochastic flow lines is proposed. The numerical study in Section 5.4 demonstrates the accuracy of the model and the potential of time-dependent buffer allocation strategies. Finally, concluding remarks and further directions regarding the optimization of time-dependent buffer allocations are given in Section 5.5.

5.2 Time-dependent stochastic flow lines

In the following, a flow line is considered which produces a single product. The line consists of K consecutive machines and $K-1$ buffers. Between machine k and $k+1$ the buffer space is assumed to be $b_{k,t}$ at time period t. The index t indicates that the buffer space is allowed to change in certain

intervals over time. Stochastic effects of demand and supply are omitted as we assume an infinite supply of raw material for machine 1 and an infinitely large buffer capacity behind machine K. We assume blocking after service, i.e., a processed workpiece remains on machine k if the buffer k is full until a buffer space becomes available. This corresponds to a configuration where the number of Kanban cards at stage k equals $b_{k,t} + 1$. While each of the $b_{k,t}$ cards corresponds to a buffer space, the additional card is used for a workpiece on the machine. We assume that the effective processing times are stochastic and time-dependent with an effective production rate of $\mu_{k,t}$. Figure 5.1 depicts the considered flow line.

Three basic cases can be distinguished if the buffer allocation is allowed to change over time. First, in the case of $b_{k,t} = b_{k,t+1}$ the buffer setting does not change. Second, for $b_{k,t} < b_{k,t+1}$ the available buffer space is increased. This can be realized by adding more Kanban cards to stage k of the flow line. Additional care is needed for the third case: $b_{k,t} > b_{k,t+1}$, of a buffer space reduction. The inventory level at the end of period t may be greater than the allowed buffer space in period $t + 1$. In this case we assume that the inventory may exceed the buffer space temporarily. However, the production of the upstream machine is stopped until the inventory level falls below the new buffer limit. This can be implemented by removing the Kanban cards from the excessive bins. The alternative of scrapping excessive inventory is economically unfavorable. In addition, a selective reduction of the production ahead of time is practically infeasible because of the stochastic environment.

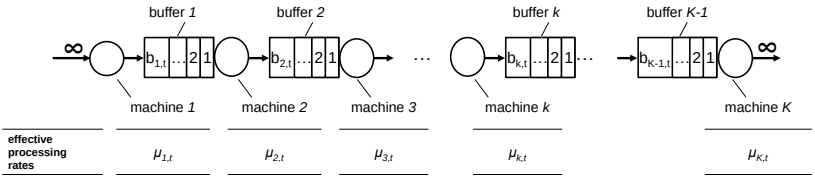


Figure 5.1: Stochastic flow line with time-dependent buffer allocation and effective processing times

5.3 Evaluation by a discrete-time sampling approach

This section presents a discrete-time sampling approach for the evaluation of the model described in the preceding section. The related approach of Helber et al. (2011) approximates the stationary behavior by deliberately ignoring the warm-up phase and building a time average for the remaining part of the sample. To evaluate the dynamic system behavior, we consider a set of S independent samples instead of a single sample. Each of them is generated as described by Helber et al. (2011). Random realizations of the processing time distributions are used to describe the isolated behavior for each machine k while ignoring blocking and starving caused by other machines. The sampled processing capacities $c_{k,t,s}$ are then given by the number of finished workpieces during discrete periods of equal length T_{length}. The processing time is determined when the workpiece enters the machine. Hence, the processing is completed according to this time regardless of a changing distribution during the processing. Figure 5.2 illustrates the case of a change regarding the processing rate over time with $\mu_{k,1} < \mu_{k,2}$. The dashed intervals represent realizations of the random variable describing the processing time distribution during phase 1 and the solid intervals, respectively, for the second phase.

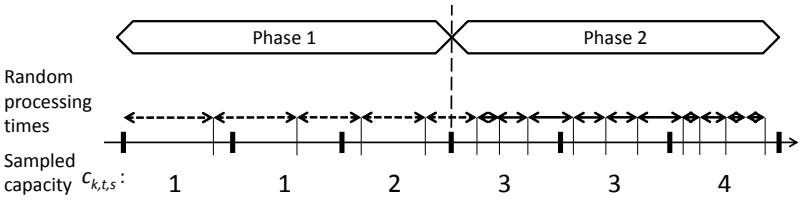


Figure 5.2: Time-dependent generation of production capacities $c_{k,t,s}$ for sample s on machine k in period t

The sampled processing capacities serve as deterministic input of a MIP. Compared to the existing model of Helber et al. (2011), the buffer space parameter has now a period index t. To allow for decreasing buffer spaces, the binary decision variable $Z_{k,t,s}$ indicates whether the allowed inflow to buffer k is positive or 0 in period t for sample s. It is 0 if the inventory present at the buffer is greater than or equal to the buffer space. The complete notation for the MIP evaluation model can be found in Table 5.1.

Table 5.1: Notation for the evaluation model

Indices	
$k = 1, ..., K$	machines in the flow line
$t = 1, ..., T$	periods
$s = 1, ..., S$	samples
Parameters	
$b_{k,t}$	exogenously given capacity of the buffer behind machine k at period t
$c_{k,t,s}$	potential processing capacity of machine k in period t for sample s
M	sufficiently large number
Integer and binary decision variables	
$Y_{k,t,s}$	end-of-period inventory level of buffer k in period t for sample s
$Q_{k,t,s}$	production quantity of machine k in period t for sample s
$Z_{k,t,s}$	binary variable, 1 if allowed flow into buffer k for sample s is positive, 0 otherwise

$$\max \quad \sum_{s=1}^{S} \sum_{k=1}^{K} \sum_{t=1}^{T} (10\,T - t)\, Q_{k,t,s} \tag{5.1}$$

s.t.

$$Y_{k,t,s} = Y_{k,t-1,s} + Q_{k,t,s} - Q_{k+1,t+1,s}, \quad k = 1, ..., K, t = 1, ..., T, s = 1, ..., S, \tag{5.2}$$

$$Q_{k,t,s} = 0, \quad k = 1, ..., K, t < k, s = 1, ..., S, \tag{5.3}$$

$$Q_{k,t,s} \leq c_{k,t,s} Z_{k,t,s}, \quad k = 1, ..., K, t = 1, ..., T, s = 1, ..., S, \tag{5.4}$$

$$Y_{k,t,s} \leq b_{k,t} + M(1 - Z_{k,t,s}), \quad k = 1, ..., K-1, t = 1, ..., T, s = 1, ..., S, \tag{5.5}$$

$$M Z_{k,t,s} \geq b_{k,t} - Y_{k,t-1,s}, \quad k = 1, ..., K-1, t = 1, ..., T, s = 1, ..., S, \tag{5.6}$$

$$-M(1 - Z_{k,t,s}) \leq b_{k,t} - Y_{k,t-1,s}, \quad k = 1, ..., K-1, t = 1, ..., T, s = 1, ..., S, \tag{5.7}$$

$$Y_{k,t,s}, Q_{k,t,s} \geq 0, \text{ and integer} \quad k = 1, ..., K, t = 1, ..., T, s = 1, ..., S, \tag{5.8}$$

$$Z_{k,t,s} \in \{0, 1\}, \quad k = 1, ..., K, t = 1, ..., T, s = 1, ..., S. \tag{5.9}$$

The objective (5.1) is to maximize the production on every stage of the flow line in every period over all samples. The weight factor $(10\,T - t)$ causes a prioritization of early production. Hence, every item is moved downstream as early as possible.

Constraints (5.2), (5.3), and (5.8) are identical to the formulation of Helber et al. (2011), besides the additional index s. Constraints (5.2) represent inventory balance equations. The inventory at the end of period t equals the inventory level at the end of the preceding period $t-1$ increased by the amount of processed workpieces of machine k in t and reduced by the processed workpieces on machine $k+1$ in the succeeding period $t+1$. It is assumed that the workpieces are transferred to the next machine at the end of each period. Further, the workpieces used at $k+1$ in $t+1$ are not included in the end of period inventory of buffer k. Please note that variables not defined (e.g., $Q_{K+1,\,T+1}$) are omitted from the respective constraints. We assume that the flow line starts completely empty such that workpieces are neither present on the machines nor in the buffers. Constraints (5.3) ensure this initial condition. Constraints (5.4) limit the production to the sampled capacity or to 0, respectively, if the allowed inflow is 0. According to Constraints (5.5), the inventory level at the end of t needs to be smaller than the available buffer space if the allowed inflow to buffer k is positive. For this case $Z_{k,t,s} = 1$ holds and the constraints are binding. The inventory may be larger if there is excessive inventory from a recent buffer reduction and the allowed inflow is 0. As in this case $Z_{k,t,s} = 0$ holds, Constraints (5.5) become redundant. It should be noted that the buffer space does not depend on the sample s in contrast to the other decision variables and parameters. Constraints (5.6) makes sure that $Z_{k,t,s}$ is set to 1 if the available space in buffer k is positive at the beginning of period t.The complementary constraints (5.7) ensure that $Z_{k,t,s}$ is 0 if the inventory level at the beginning of t exceeds the available buffer space. Finally, according to (5.8), $Y_{k,t,s}$ and $Q_{k,t,s}$ are non-negative integer values and all $Z_{k,t,s}$ are binary variables (5.9). As all decision variables depend on the samples, the model may be solved for each sample independently.

Based on the solution of the proposed MIP, a set of key performance measures can be obtained. The work in process WIP at the end of period t for machine $k+1$ and buffer k can be calculated by subtracting the cumulated production up to t of machine $k+1$ from the one at machine k

$$\mathrm{E}[\mathit{WIP}_{k,t}] = \frac{1}{S}\sum_{s=1}^{S}\Big(\sum_{\tau=1}^{t} Q_{k,\tau,s} - \sum_{\tau=1}^{t} Q_{k+1,\tau,s}\Big). \tag{5.10}$$

The expected line throughput $\mathrm{E}[Th_t]$ in period t can be obtained from the production rate at machine K

$$\mathrm{E}[Th_t] = \frac{1}{S \cdot T_{length}} \sum_{s=1}^{S} Q_{K,t,s}. \tag{5.11}$$

The expected cumulated output of the flow line over the planning horizon is given by

$$\mathrm{E}[Q_{cumulated}] = \frac{1}{S} \sum_{t=1}^{T} \sum_{s=1}^{S} Q_{K,t,s}. \tag{5.12}$$

Such a discrete-time sampling approach goes along with two kinds of approximation errors. On the one hand, a simulation error arises as the behavior of the random variables is approximated by a finite set of samples. This error decreases with an increased number of S. On the other hand, a discretization error may occur due to the assumption of constraints (5.2) that a workpiece which finishes processing on machine k during period t can be processed at the earliest in period $t+1$ on machine $k+1$. This leads to artificial blocking. Another discretization error is made regarding the end of period inventory. As described above, the inventory $Y_{k,t,s}$ does not account for workpieces which are used in the period $t+1$ on machine $k+1$.

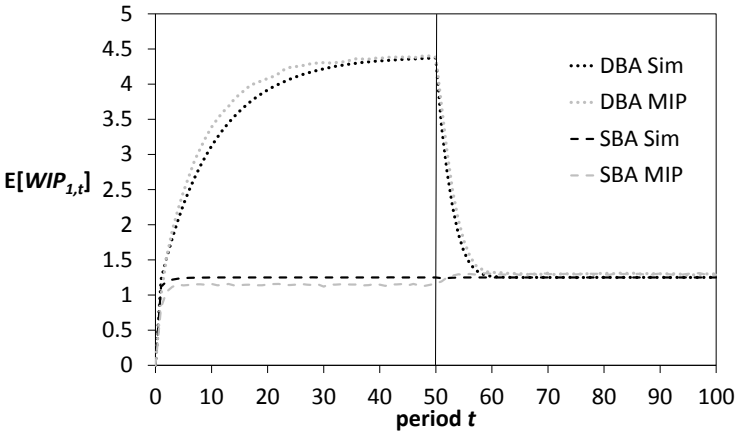
5.4 Numerical study

The following numerical study demonstrates the accuracy of the proposed evaluation approach and provides first insights regarding the system behavior for time-dependent buffer allocations. The analyzed flow line has $K = 2$ machines with a buffer in between. The processing times are assumed to be exponentially distributed. This rather simple configuration is chosen in order to maintain a distinct analysis of the otherwise potentially overlapping effects. A discrete-event simulation serves as benchmark for the approximation approach. Each simulation run consists of $1,000,000$ replications. The sample size S for the MIP is $10,000$. For the MIP the period length T_{length} is set to the largest expected processing time of all machines over time. The analysis focuses on the performance measures WIP (5.10), throughput (5.11), and cumulated output (5.12). The outline of the numerical study is as follows. First, the system behavior is analyzed for a change of the buffer space while all other system parameters remain the same over time. Second, it is assumed

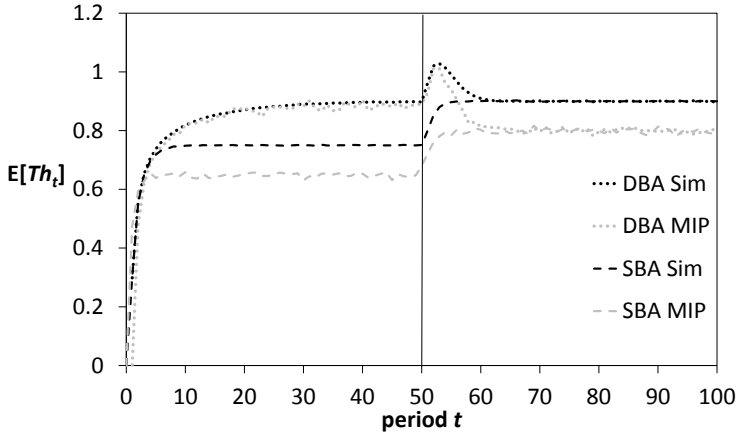
that processing rates of both machines increase at a point in time. The strategy of a Static Buffer Allocation (SBA) is compared to a Dynamic Buffer Allocation (DBA). Third, we investigate the impact of the point in time when the buffer allocation is changed for the DBA strategy. Fourth, a stepwise adjustment from one buffer capacity via an intermediate buffer allocation is investigated.

The first analysis examines the impact of changing buffer allocations. The processing rates are given by $\mu_{1,t} = \mu_{2,t} = 1$ and are constant over time. Two time-dependent buffer allocations are tested. Both of them include a change of the buffer capacity at the beginning of period $t_{bc} = 51$. We investigate a change from a small capacity to a large one (SL) and a change from a large to a small capacity (LS). Figure 5.4 depicts the development of the WIP and the throughput obtained by the MIP evaluation model and the discrete-event simulation (Sim). As expected for the LS case, the WIP exceeds the buffer and machine capacity temporarily. The graphs in Figure 5.4 reveal that the transient as well as the steady-state values are well approximated by the evaluation model. Due to the artificial blocking caused by the discretization error, the throughput is underestimated by the MIP approach. Consistently with the observations of Helber et al. (2011), the approximation quality is better for larger buffer capacities.

Changes to the buffer allocation are primarily motivated by non-stationary system parameters. For the subsequent examples we assume that the processing rates of both machines increase by 20% to 1.2 at the beginning of period $t_{rc} = 51$. Next, the SBA strategy is compared to the DBA strategy. The SBA sets a constant buffer allocation over the whole planning horizon. Dynamic changes of the system over time are ignored and only the system parameters at the end of the planning horizon are taken into account. The buffer allocation is determined such that it guarantees a goal throughput at a minimum buffer capacity assuming the steady state. The DBA strategy changes the buffer allocation if the system configuration changes. The buffer allocation is then set to the optimal steady-state value for each phase of constant system parameters. For the considered case and a goal throughput of 0.9, the steady-state buffer capacity is $b_{1,t} = 7$ for $t < t_{bc} = 51$ and $b_{1,t} = 1$ for $t \geq t_{bc} = 51$. The observed throughput peak in Figure 5.3b for the DBA strategy originates from the excessive inventory of the first phase (Figure 5.3a) which is then processed with the increased rate. The expected cumulated production generated by the DBA is 7.9% higher than the one of the SBA.

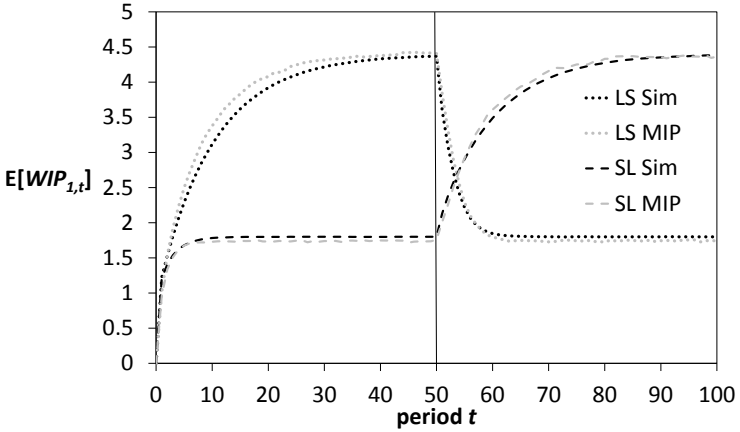


(a) Expected WIP over time

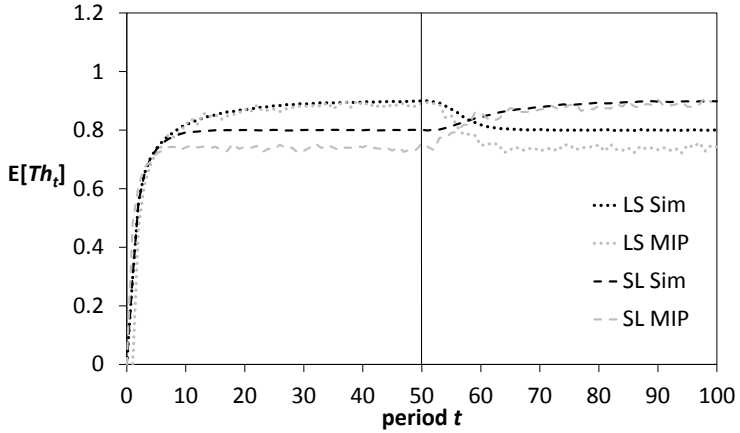


(b) Expected throughput over time

Figure 5.3: Static and dynamic buffer allocations, with $t_{rc} = t_{bc} = 51$ for the dynamic allocation

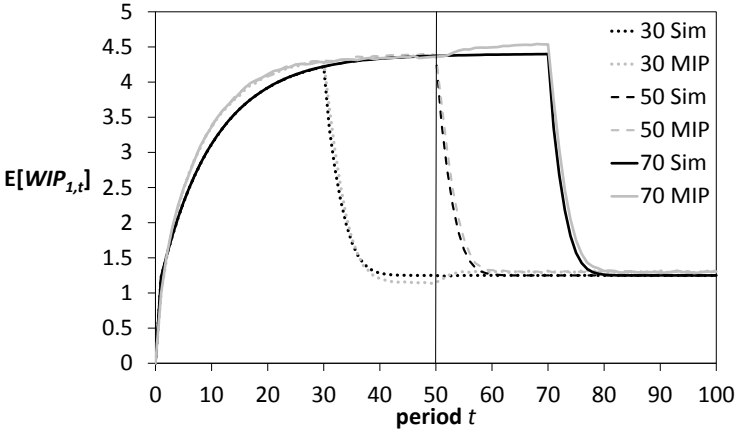


(a) Expected WIP over time

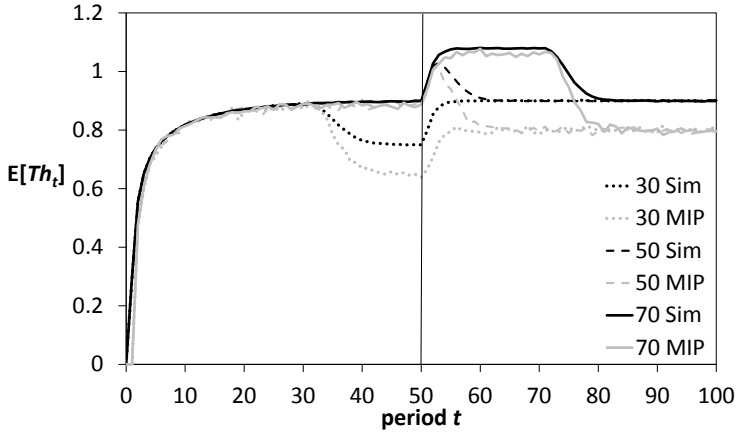


(b) Expected throughput over time

Figure 5.4: Change of the buffer capacity from $b_{1,t} = 7$ to $b_{1,t'} = 2$ (LS) and $b_{1,t} = 2$ to $b_{1,t'} = 7$ (SL) for $t < t_{bc} \leq t'$

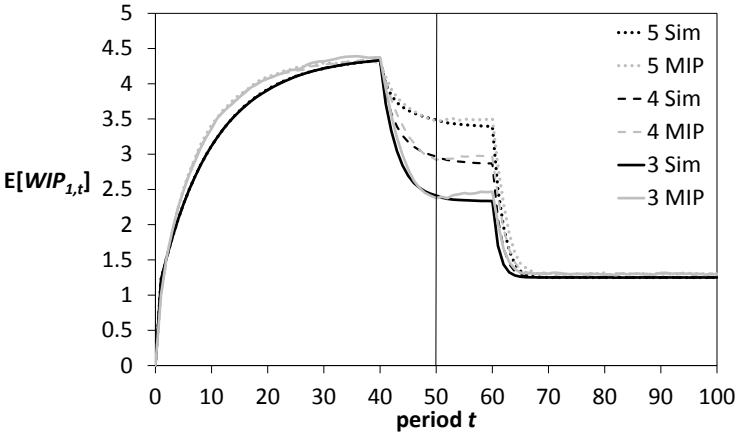


(a) Expected WIP over time

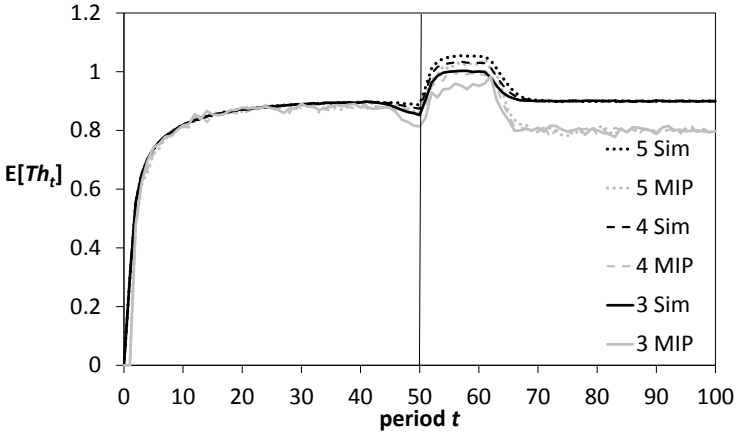


(b) Expected throughput over time

Figure 5.5: Dynamic buffer allocations for $t_{rc} = 51$ and $t_{bc} \in \{31, 51, 71\}$



(a) Expected WIP over time



(b) Expected throughput over time

Figure 5.6: Stepwise dynamic buffer allocations with $b_{1,t}^{inter} \in \{3, 4, 5\}$

Production managers usually aim at a constant production rate close to their goal production rate. Hence, the observed throughput peak of the DBA is undesirable. Subsequently, we examine the impact of the time t_{bc} when the buffer allocation is changed. So far, it is assumed that the buffer allocation is changed simultaneously with the rate change ($t_{bc} = 51$). Figure 5.5 depicts the results for the DBA with changed buffer configurations at $t_{bc} \in \{31, 51, 71\}$. The throughput peak is even longer and higher for $t_{bc} = 71$ as the large buffer configuration is maintained also for situations with the increased processing rates. For $t_{bc} = 31$, there is no throughput peak but as the buffer capacities are decreased too early, the throughput falls below the desired level prior to the rate change. The value of t_{bc}, which minimizes the deviation from the goal throughput, is therefore expected to be between 31 and 51.
For the analysis depicted in Figure 5.6, the DBA strategy is extended by an intermediate buffer capacity $b_{k,t}^{inter}$. This capacity is used between the actual DBA buffer configurations. The underlying idea is to take also the transient behavior of the buffer change into account. For the considered example $b_{1,t}^{inter}$ is valid for $40 < t \leq 60$. Figure 5.6b reveals that the height of the throughput peak can be controlled by $b_{1,t}^{inter}$. The higher $b_{1,t}^{inter}$ the higher is the throughput peak.

5.5 Conclusion and future research

This paper presents an initial approach towards time-dependent buffer allocations for flow lines under non-stationary conditions. Therefore, the unique characteristics of time-dependent buffer allocations are discussed. Furthermore, a sampling evaluation approach is presented. The numerical study demonstrates its accuracy and the potential arising from changes of the buffer allocation over time. Further research is needed to develop decision models for the optimization of time-dependent buffer allocations. The sampling approach bears the potential to combine the optimizing power of linear programming with the modeling flexibility of simulation. Future research should explore the option of converting the so far exogenously given buffer capacities $b_{k,t}$ into decision variables. The key decisions to be made are when and in what kind of steps the buffer allocation has to be changed in order to react adequately to changing system parameters.

6 A proactive approach to Kanban allocation in stochastic flow lines with time-dependent parameters

Co-authors:
Raik Stolletz
Chair of Production Management, Business School, University of Mannheim, Germany

Working paper

Abstract:
Flow lines operate under stochastic and time-dependent influences. Stochasticity originates from random demand and processing times. Time-dependent changes in parameters are caused, e.g., by seasonal demand or due to replacement of machinery. We propose a time-dependent change in buffer capacities by utilizing Kanban cards to minimize the required work-in-process (WIP) inventory while maintaining a predefined gamma service level over a finite planning horizon.

We report observations regarding the monotonicity of the gamma service level and the expected average WIP in the line with respect to time-dependent buffer capacities. Based on these observations, a local search algorithm is developed. The numerical study indicates that the algorithm evaluates only a small fraction of all possible allocations. Moreover, we provide examples that demonstrate the advantages of time-dependent, as compared to constant allocations. Tests of allocation approaches based on steady-state models indicate that they may lead to infeasibility and poor performance.

6.1 Introduction

Flow lines are production systems for high volume production, e.g., used by manufacturers from the automotive industry (Li, 2013). They operate under stochastic and time-dependent influences. Stochastic impacts of random demand, processing times, machine failures, and subsequent repairs are widely acknowledged in the literature (Dallery and Gershwin, 1992). Time-dependent changes in parameters lead to non-stationarity of the flow line. For instance, changes in demand over time on the supply chain level cause time-dependent demand arriving at the flow line (Takahashi et al., 2004; Shang, 2012). Replacement of machinery or learning effects during the production ramp-up cause time-dependent effects in the production system itself (Jaikumar and Bohn, 1992; Terwiesch and Bohn, 2001). If these effects occur in flow lines, they directly impact their performance.

We consider a serial flow line with finite buffer capacities which serves a stochastic and time-dependent demand from a finished goods buffer. The flow line is controlled by Kanban cards with continuous review policy, i.e, cards without attached workpieces are immediately transferred to upstream stations to signal demand. The number of Kanban cards limits the work in process (WIP) inventory and is equal to the required buffer capacity plus one (Berkley, 1991). Each station in the line is characterized by generally distributed processing times with time-dependent parameters. Whereas the actual processing times are random, the time-dependent changes in the random variables' parameters are assumed to be deterministic and known in advance. This is a valid assumption particularly for parameters of the flow line. These parameters are either under direct control, e.g., for the introduction of new machinery, or empirical data allow for reliable forecasts, e.g, from learning curves (Lapre et al., 2000).

The minimization of WIP while maintaining a given service level is a common goal for demand-driven flow lines and production systems (Gaury et al., 2000; Liu et al., 2004). A γ-service level is selected to reflect both, the number of backorders and the time the backorders persist (Schneider, 1981). We propose a time-dependent change of the buffer capacities to minimize the expected average WIP while maintaining a predefined γ-service level over a finite planning horizon. In practice, time-dependent changes of buffer capacities are easy to realize by means of Kanban cards. This new approach is termed *Proactive Kanban*, as it changes the number of Kanban cards before inventory thresholds are met or changes of system parameters are detectable by statistical analyses.

The majority of the literature on flow lines assumes constant parameters and steady-state conditions. Hence, the allocation of Kanban cards is typically treated as a design decision. An exception is the analysis of time-dependent demand processes. To the best of our knowledge, changes in parameters of the flow line itself have not been analyzed. The existing approaches that use the flexibility of Kanban cards to adapt the buffer capacities for time-dependent demand are based on historic data or inventory thresholds (Takahashi, 2003). These approaches only react to parameter changes. Hence, the potential of planning for known parameter changes remains unused.

To solve the new decision problem we suggest a local search algorithm. The algorithm uses discrete-event simulation to evaluate the performance of a given time-dependent Kanban allocation. In addition, two approaches that rely on steady-state models are developed.

The goals of this study are: (i) to provide a time-dependent card setting that delivers the desired service level at a minimal average WIP level for flow lines that are subject to stochasticity and time-dependent parameter changes and (ii) to generate insights into how time-dependent card settings differ from standard approaches based on steady-state assumptions. Our contributions can be summarized as follows:

1. We introduce a new card setting approach and the resulting decision problem which accounts for time-dependent changes of flow line parameters by means of changes in buffer capacities.
2. The numerical study demonstrates that the proposed approach improves the system performance compared to constant allocations. Moreover, we test approaches based on stationary models and show that they may lead to infeasibility and poor performance. Further, the numerical study provides an example that a change of buffer capacities before the parameter change can lead to lower expected average WIP than a simultaneous or delayed change of buffer capacities.

The remainder of the paper is organized as follows. The related literature is reviewed in Section 6.2. The flow line model and the resulting decision problem are described in Section 6.3. Subsequently, we develop a local search algorithm in Section 6.4. It is based on numerical observations of the monotonicity of the service level and the expected average WIP in the line with respect to time-dependent buffer capacities. Section 6.5 numerically investigates the benefits of the new card setting approach. Concluding remarks and directions for future research are provided in Section 6.6.

6.2 Literature review

We first review performance evaluation approaches for time-dependent and stochastic systems. Second, existing Kanban card setting approaches in the literature are presented. Moreover, we emphasize the differences to time-dependent supply chain models. Finally, we provide a brief summary of structural properties with respect to buffer capacities in flow lines.

No exact analytical performance evaluation approaches are available for production systems with finite and time-dependent buffer capacities, even with restriction to single-stage systems (Schwarz et al., 2016). An approximative approach for a single-stage system is introduced by Hampshire et al. (2009). Existing approaches for the analysis of time-dependent multi-stage systems lack key properties of flow lines. They assume infinite buffer capacities (Vandergraft, 1983), infinite number of servers (Massey and Whitt, 1993), or that customers are lost if they arrive at a full downstream buffer (Nasr and Taaffe, 2013). Hence, the characteristic effect of blocking is neglected. Recently, Meerkov and Zhang (2008) and Zhang et al. (2013) proposed approaches for the transient performance evaluation of flow lines with constant parameters. These approaches can in principle be used for the approximation of piecewise constant parameters (Seki and Hoshino, 1999). However, they are restrictive in their assumptions on the used probability distributions. In the absence of analytical models, discrete-event simulation is the common approach to performance evaluation of Kanban systems with parameters and number of Kanban cards that change over time (Tardif and Maaseidvaag, 2001; Takahashi and Nakamura, 2002).

The literature on Kanban card allocation can be classified according to the external parameters that describe the system and the card setting. Both can be either constant over time or time-dependent. *Traditional Kanban* systems are designed for stochastic production systems with constant parameters and have a constant card setting (Diaz and Ardalan, 2010). A survey of different variants of traditional Kanban systems is provided by Berkley (1992). For flow lines with constant parameters, time-dependent increases of buffer capacities during an initial transient phase are evaluated by Anderson and Moodie (1968). They conclude that it is best to start with the allocation that is suitable for the steady state.

Time-dependent changes of system parameters can to some extend be covered by constant buffer capacities (Göttlich et al., 2016). However, if the magnitude of changes is too high or non-cyclic, a change of the number of Kanban cards, i.e., of the buffer capacities, can be beneficial. The approaches

that change the Kanban allocations over time can be further distinguished by the information the reallocation is based on. Tardif and Maaseidvaag (2001) suggest an *Adaptive Kanban* concept. It releases and captures extra cards depending on inventory thresholds. The *Reactive Kanban* approach by Takahashi and Nakamura (2002) uses statistical analysis to detect a change of the demand distribution. Subsequently, the number of cards is adapted to meet the new demand parameters, while the transient transition phase in-between is neglected. For both, adaptive and reactive Kanban, metaheuristics such as genetic algorithms and parameter variations are used to determine the number of Kanban cards. Takahashi (2003) numerically compares adaptive and reactive Kanban. Both approaches outperform the traditional Kanban system. However, none of the approaches uses information about future parameter changes for the reallocation of cards.

The impact of time-dependent demand changes is also investigates by the inventory theory literature. Serial inventory systems with time-dependent demand are analyzed by Clark and Scarf (1960) and recently by Shang (2012). Both models include echelon base-stock policies, i.e., for replenishment decisions on each stage m they consider the inventory at stage m and all downstream stages. In contrast, Kanban is an installation stock policy for which a replenishment at stage m is triggered solely by the inventory level at stage m (Axsäter and Rosling, 1993). Additionally, the Kanban control is not equivalent to a base-stock policy as the reordering differs in the case of empty buffers (Spearman, 1992). Iida (2002) and Bollapragada and Rao (2006) consider time-dependent and stochastic production capacity for single-stage inventory systems. However, all of the reviewed inventory models assume a periodic review policy whereas the flow line under consideration operates with continuous review.

Structural properties of flow lines with respect to buffer capacities are so far only established for parameters and buffer capacities that are constant over time. A good overview of the available results is provided by Glasserman and Yao (1996). There are no proven structural results for flow lines with time-dependent and generally distributed processing times with respect to the relation of time-dependent buffer capacities and the resulting service level or expected average WIP. To the best of our knowledge, even for flow lines with constant parameters and buffer capacities, structural properties for the γ-service level and the expected average WIP are not addressed in the literature. Based on numerical studies So (1997) and Papadopoulos and Vidalis (2001b) provide some observations regarding the WIP under steady-state conditions for a given total buffer capacity in the line.

6.3 Proactive Kanban System

6.3.1 Flow line model

We consider a Kanban system with $m = 1, ..., M$ stages that produces a single product and operates for a finite planning horizon of length T. The required notation can be found in Table 6.1.

The transport of workpieces is assumed to occur instantaneous and with negligible transportation times. At each stage, Kanban cards circulate between buffer m and station m. Each workpiece that completes processing at station m is stored together with an attached card. The cards are detached from the workpieces that start processing on station $m+1$ and are transferred back to station m where they are collected and serve as production authorization signals. Station m starts processing a workpiece only if a Kanban card from

Table 6.1: Notation for Proactive Kanban Systems

Indices	
$m = 1, ..., M$	Stages in the flow line
$i = 0, ..., I$	Changes of buffer allocations over time
Parameters	
T	Length of planning horizon
γ^*	Goal γ-service level
b_m^{max}	Maximum buffer capacity at stage m
$0 = t_0^*, t_1^*, ..., t_i^*, ..., t_I^*$	Time of the ith change of the buffer allocation
Decision variables	
$\mathbf{B} = \begin{pmatrix} B_{1,0} & \cdots & B_{1,i} & \cdots & B_{1,I} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ B_{m,0} & \cdots & B_{m,i} & \cdots & B_{m,I} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ B_{M,0} & \cdots & B_{M,i} & \cdots & B_{M,I} \end{pmatrix}$	Buffer capacities at stage m after the ith change
Dependent variables	
$W(\mathbf{B})$	Average WIP in the line over the planning horizon
$W^-(\mathbf{B})$	Backorder level over the planning horizon
$SL^\gamma(\mathbf{B})$	γ-service level over the planning horizon

stage m and a workpiece in buffer $m-1$ are available. The formal equivalence of such a Kanban system with constant card setting and a flow line with finite and constant buffer capacities and blocking after service mechanism is established by Berkley (1991).

The distinctive feature of the Proactive Kanban System is that the number of cards at each stage can change over time. The number of Kanban cards may be changed at predefined time instances t_i^*. If the number of cards at stage m is increased at time t_i^* ($B_{m,i} < B_{m,i+1}$), the additional cards are directly added at station m to authorize production. In case of a card reduction ($B_{m,i} > B_{m,i+1}$), cards without an attached workpiece are retrieved. If this is insufficient to attain the desired number of cards, the remaining cards are removed from workpieces in the buffer, starting with the workpiece that is processed next at the downstream station. Consequently, until all workpieces without cards are served, the number of workpieces may temporarily exceed the number of cards at a given stage.

Figure 6.1 depicts a flow line representation of the considered system, including the synchronization point between customer demand and finished goods as final stage. The supply of raw material to station 1 is assumed to be unlimited. The processing times of all stations are generally distributed and characterized by their time-dependent rate $\mu_m(t)$ and the coefficient of variation $cv_m(t)$ at time t. We assume piecewise constant rates. The time-dependent changes in rates and coefficients of variation are given by forecasts. They are independent of the timing of buffer capacity changes. Potentially, the type of the distribution may also change over time. We further assume that all parameter changes are effective immediately, i.e, the residual time of a workpiece that is processed on a station is adapted accordingly. At the final stage customer demand arrives with generally distributed inter-arrival times with time-dependent rate $\lambda_c(t)$ and coefficient of variation $cv_c(t)$. It is served from a finished goods buffer behind station M. Orders that cannot be served immediately are backlogged.

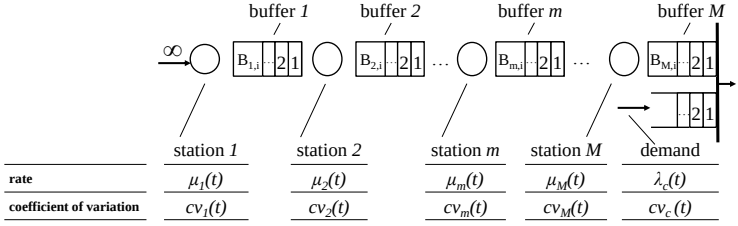


Figure 6.1: Flow line representation of the Proactive Kanban System

6.3.2 Proactive Kanban Card Setting Problem

The Proactive Kanban Card Setting Problem is given by the non-linear mixed integer program (6.1).

$$\min \ \mathrm{E}[W(\mathbf{B})] \tag{6.1a}$$

s.t.:

$$SL^{\gamma}(\mathbf{B}) \geq \gamma^{*} \tag{6.1b}$$

$$0 \leq B_{m,i} \leq b_{m}^{max} \qquad \forall m, \forall i \tag{6.1c}$$

The key decision is to determine the time-dependent allocation of Kanban cards, i.e., the buffer capacities $\mathbf{B}$. The objective (6.1a) is to minimize the expected average WIP, $\mathrm{E}[W(\mathbf{B})]$, in the line, which is a function of $\mathbf{B}$. The WIP accounts for all workpieces in line. This includes the workpiece on station $m = 1$, but not the infinite amount of workpieces in front of it. By Constraints (6.1c) we enforce a maximum buffer capacity, b_m^{max}, at every stage m that must not be exceeded during the planning horizon. Hence, the model accounts for physical limitations of the buffers between the stations. The model can support greenfield planning by selecting b_m^{max} sufficiently large. According to Constraint (6.1b), the flow line has to fulfill a goal γ-service level γ^{*}. The γ-service level relates the backorder level, $W^{-}(\mathbf{B})$, to the total demand, N_T, over the planning horizon. The achieved expected γ-service level also depends on $\mathbf{B}$. Based on the discussion of Chen et al. (2003) we adapt the standard γ-service level definition to a finite planning horizon

$$SL^{\gamma}(\mathbf{B}) = 1 - \mathrm{E}\left[\frac{W^{-}(\mathbf{B})}{N_T}\right]. \tag{6.2}$$

In the following, we assume that at time $t = 0$ there are no workpieces in the line. However, for long planning horizons the decision model may serve as a building block in a rolling planning horizon approach. In this case, initial values of workpieces in the flow line have to be set. Furthermore, only a subset of the allocation decisions is fixed. Hence, the rolling horizon approach allows for the integration of updated information regarding the system status and forecasts of parameter changes as time passes.

6.4 Solving the Proactive Kanban Card Setting Problem

Motivated by analytical results for a Markovian single-stage system in steady state we present insights on the behavior of the expected average WIP and γ-service level in Proactive Kanban Systems based on numerical experiments. In the absence of analytical solutions, a discrete-event simulation is used to evaluate given allocations. Based on the two key observations we propose dominance criteria for time-dependent buffer allocations. Subsequently, a local search algorithm which systematically evaluates potential candidate allocations is introduced. It exploits the observed properties from Section 6.4.1 to reduce the number of required evaluations to find solutions for problem (6.1).

6.4.1 Observations from numerical tests

It is possible to establish the following properties for a Markovian system, given constant buffer capacities over time and steady-state conditions.

Theorem 6.4.1. $\mathrm{E}[W(B_M)]$ *is strictly increasing and convex in the buffer capacity,* B_M*, given* $M = 1$*, exponentially distributed processing times, Poisson demand, and steady-state conditions.*

Theorem 6.4.2. $SL^{\gamma}(B_M)$ *is strictly increasing and concave in the buffer capacity,* B_M*, given* $M = 1$*, exponentially distributed processing times, Poisson demand, and steady-state conditions.*

The proofs are based on a closed-form solution which can be obtained from a birth and death process representation of the system (see Appendix A).

Motivate by this insights for the steady-state systems, a set of numerical experiments is conducted to gain insights into how the performance measures expected average WIP and γ-service level in Proactive Kanban Systems are influenced by the time-dependent buffer allocation. The experiments include varying numbers of stations M, processing distributions, numbers of buffer changes I, and timings of buffer changes t_i^*. Exemplary results of the conducted numerical study can be found in Appendix B.

With respect to the expected average WIP we make the following observation.

Observation 6.4.3. *The expected average WIP,* $\mathrm{E}[W(\mathbf{B})]$*, is non-decreasing in* $B_{m,i}$ $\forall m, \forall i$*.*

Notably, the expected average WIP in a Proactive Kanban System is not necessarily convex in the buffer capacities. A numerical counter example can be created by buffer capacity increases close to the end of the planning horizon. This increase of the buffer capacity causes decreasing marginal increases of the expected average WIP as the remaining time during the planning horizon may not be sufficient to fill the buffer to its steady-state value. We provide such an example for a single station system ($M = 1$), a planning horizon of $T = 1000$, exponentially distributed processing times with rate $\mu_1(t) = 2/3,\ t \in [0; 1000]$, and Poisson demand with rate $\lambda_c(t) = 0.5$ $\forall t \in [0, 1000]$. The expected average WIP in the system, $\mathrm{E}[W(\mathbf{B})]$, for a buffer capacity change from $B_{1,0} = 5$ at $t = 980$ to varying values of $B_{1,1}$ is depicted in Figure 6.2. Clearly $\mathrm{E}[W(\mathbf{B})]$ is not convex in $B_{1,1}$.

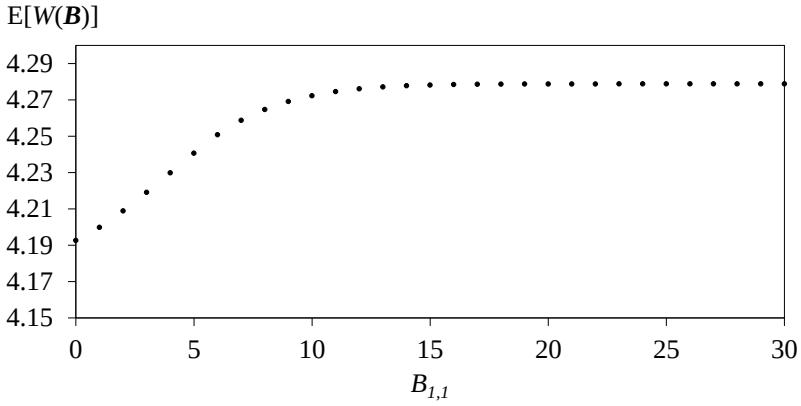


Figure 6.2: Example of non-convexity of $\mathrm{E}[W(\mathbf{B})]$ in $B_{1,1}$

The second observation is made with respect to the service level.

Observation 6.4.4. *The γ-service level, $SL^{\gamma}(\mathbf{B})$, is non-decreasing in $B_{m,i}$ $\forall i, m$.*

Intuitively, larger buffer capacities allow a higher compensation of the stochastic and time-dependent influences. Hence, the cumulated throughput increases, whereas the arriving customer demand is independent of the buffer capacities. Hence, the backlog decreases and reversely the service level increases.

6.4.2 Dominance between time-dependent buffer allocations

Every evaluation of an allocation $\mathbf{B}$ provides information about $\mathrm{E}[W(\mathbf{B})]$ and $SL^{\gamma}(\mathbf{B})$. To exclude a set of other candidate allocations from future evaluations we use this information and the Observations 6.4.3 and 6.4.4.

Every *feasible* allocation $\mathbf{B}$ is used to exclude all allocations $\mathbf{B'}$ with $B'_{m,i} \geq B_{m,i}\ \forall m, \forall i$, i.e., all allocations with componentwise greater or equal buffer capacities. Observation 6.4.3 suggests that for these allocations $\mathrm{E}[W(\mathbf{B'})] \geq \mathrm{E}[W(\mathbf{B})]$ holds. Consequently, the allocations $\mathbf{B'}$ are not candidates for a lower objective value than $\mathrm{E}[W(\mathbf{B})]$. Moreover, the resulting expected average WIP, $\mathrm{E}[W(\mathbf{B})]$, of the feasible allocation provides an upper bound on the objective value. In the following, UB_W denotes the best upper bound on the objective value found so far by the algorithm.

Every *infeasible* allocation $\mathbf{B}$ with $SL^{\gamma}(\mathbf{B}) < \gamma^*$ is used to exclude all allocations $\mathbf{B'}$ with $B'_{m,i} \leq B_{m,i}\ \forall m, \forall i$. According to Observation 6.4.4 the service level of the allocations $\mathbf{B'}$ is expected to be lower than the one of allocation $\mathbf{B}$, i.e., $SL^{\gamma}(\mathbf{B'}) \leq SL^{\gamma}(\mathbf{B})$. Hence, the allocations $\mathbf{B'}$ are also infeasible. If in addition $\mathrm{E}[W(\mathbf{B})] > UB_W$ holds, we also exclude all allocations $\mathbf{B'}$ with $B'_{m,i} \geq B_{m,i}\ \forall m, \forall i$. According to Observation 6.4.3 the allocations $\mathbf{B'}$ cannot yield an improved objective value because any increase in the buffer capacities will lead to an at least as high expected average WIP value as UB_W.

6.4.3 Local search approach

The proposed algorithm exploits the dominance criteria to avoid a complete enumeration of all allocations. Given that Observations 6.4.3 and 6.4.4 hold true, the algorithm delivers optimal results. It divides the decision problem into subproblems with two decision variables $B_{m',i'}$ and $B_{m'',i''}$ and fixed values for all other decision variables $B_{m,i}, m \in \{1, 2, ..., M\}\backslash\{m', m''\}$, $i \in \{0, 1, ..., I\}\backslash\{i', i''\}$. A subproblem is solved if all allocations are either evaluated or excluded by a dominating allocation. In the following we first describe the case $M = 1$ and then explain the required extensions of the algorithm to cases with $M > 1$.

The algorithm solves all subproblems that are generated by iterating over all possible configurations of the decision variables $B_{M,i}$, $i \in \{0, 1, ..., I\} \backslash \{i', i''\}$. Information about dominated allocations obtained from previously solved subproblems are exploited to reduce the number of required evaluations. For all $i \in \{0, 1, ..., I\} \backslash \{i', i''\}$ the algorithm divides the interval of not evaluated values into two subintervals. Starting with the evaluation of $B_{M,i} = \lfloor b_M^{max}/2 \rfloor \; \forall i \in \{0, 1, ..., I\} \backslash \{i', i''\}$ the algorithm continues recursively with the subintervals $[0; \lfloor b_M^{max}/2 \rfloor - 1]$ and $[\lfloor b_M^{max}/2 \rfloor + 1; b_M^{max}]$. This recursive procedure systematically evaluates allocations which potentially dominate infeasible and feasible allocations of other subproblems.
The best upper bound, UB_W, on the objective value is constantly updated during the search process. Thus, at termination of the algorithm, the obtained solution is the allocation which generates the best upper bound found.

Each subproblem is solved in three steps as illustrated in Figure 6.3. In part (a) of step 1 the value of $B_{M,i''}$ is fixed to b_M^{max} and the smallest value for $B_{M,i'}$ which results in a feasible allocation is determined ($B_{M,i'}^{1a}$). For the search a bisection method is applied. It exploits the observation that the service level increases in the buffer capacities. After step 1 (a) all allocations with $B_{M,i'} < B_{M,i'}^{1a}$ are excluded from future evaluations within subproblem due to infeasibility. Moreover, according to the dominance criterion for feasible solutions, all allocations with $B_{M,i''} = b_M^{max} \wedge B_{M,i'} \geq B_{M,i'}^{1a}$ are excluded from being evaluated within the subproblem.

Step 1 (b) applies a bisection method to determine $B_{M,i''}^{1b}$, i.e, the smallest value of $B_{M,i''}$ that ensures feasibility with given $B_{M,i'} = B_{M,i'}^{1a}$. The advantage of step 1 is that it quickly determines if according to Observations 6.4.3 and 6.4.4 no feasible solution exists to the subproblem ($B_{M,i'}^{1a} > b_M^{max}$), or if all allocations are feasible ($B_{M,i'}^{1a} = B_{M,i''}^{1b} = 0$). If neither of this is the case the algorithm continues with step 2.

In step 2 the feasible allocation in the subproblem with the lowest value of $B_{M,i''}$ and the corresponding lowest possible value of $B_{M,i'}$ is determined. We therefor fix the value of $B_{M,i'}$ to b_M^{max} and determine the smallest value of $B_{M,i''}$ that ensures feasibility ($B_{M,i''}^{2a}$). The search of the bisection method is limited to allocations that are not dominated by allocations that were evaluated in step 1. In step 2 (b) we fix $B_{M,i''}$ to the value obtained for $B_{M,i''}^{2a}$ in step 2 (a) and search for $B_{M,i'}^{2b}$, i.e., the smallest value of $B_{M,i'}$ that provides a feasible allocation given $B_{M,i''}^{2a}$. After completion of step 2 all allocations with $B_{M,i''} \geq B_{M,i''}^{2a} \wedge B_{M,i'} \geq B_{M,i'}^{2b}$ can be excluded from future evaluations due to the dominance criterion for feasible allocations. Furthermore, the

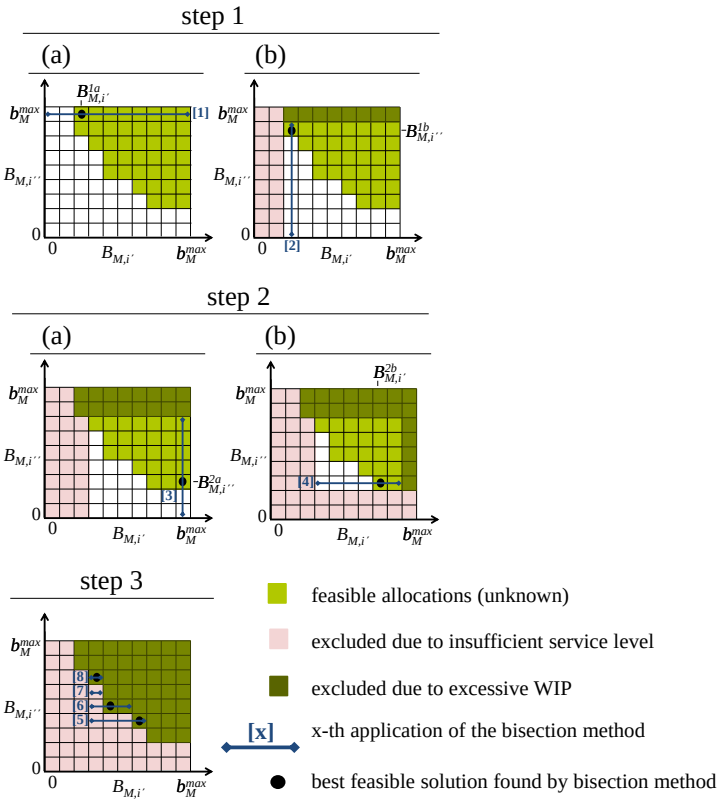


Figure 6.3: Search procedure for subproblem with 2 decision variables $B_{M,i'}$ and $B_{M,i''}$

allocations with lower buffer capacities than in the allocation $(B^{2b}_{M,i'}, B^{2a}_{M,i''})$ are expected to be infeasible, i.e., all allocations with $B_{M,i''} < B^{2a}_{M,i''}$ and $B_{M,i'} < B^{2b}_{M,i'} \wedge B_{M,i''} = B^{2a}_{M,i''}$ are excluded from future evaluations.

To complete the search in the subproblem we iteratively increase the value of $B_{M,i''}$ (step 3). For each iteration the corresponding smallest value of $B_{M,i'}$, which ensures feasibility, is obtained. Dominated allocations from previous iterations are excluded from the search with the bisection method. The algorithm terminates the search in the subproblem if all allocations are dominated by already evaluated allocations.

For $M > 1$ it has to be specified for which buffers m' and m'' the subproblems are solved. The iteration then has to be performed over all values of the remaining decision variables $B_{m,i}, m \in \{1, 2, ..., M\}\backslash\{m', m''\}$, $i \in \{0, 1, ..., I\}\backslash\{i', i''\}$.

There are three reasons that prevent the algorithm from completely iterating over all allocations, (i) the efficient solution of the subproblems in three steps, as outlined above, (ii) that infeasible allocations for one subproblem are assumed to be also infeasible for all other subproblems with smaller fixed buffer capacities $B_{m,i}, m \in \{1, 2, ..., M\}\backslash\{m', m''\}, i \in \{0, 1, ..., I\}\backslash\{i', i''\}$, and (iii) that allocations which are dominated by a feasible allocation with smaller expected average WIP are also dominated in all subproblems with larger fixed buffer capacities.

Preliminary numerical test show that the obtained allocation is not affected by the choice of m',m'',i' and i''. However, the numerical results indicate that buffers and phases for which low buffer capacities are expected in the solution should be selected as m',m'',i' and i''. They tend to reduce the number of required evaluations because infeasible allocations with high buffer capacities $B_{m,i}, m \in \{1, 2, ..., M\}\backslash\{m', m''\}, i \in \{0, 1, ..., I\}\backslash\{i', i''\}$ are found early in the search process.

6.4.4 Approaches based on steady-state models

Steady-state models dominate the flow line literature. Moreover, Anderson and Moodie (1968) suggest that for a transient phase, given constant parameters, the buffer allocation that is best for the steady state should be implemented. Hence, we introduce two approaches that demonstrate how steady-state models can be applied to heuristically solve the Proactive Kanban Card Setting Problem. Both approaches incorporate the information about future parameter development but differ in the way how they aggregate this information. The two approaches are originally developed by Kolesar et al. (1975) and Green et al. (1991) for staffing in single-stage service systems.

(i) **Simple stationary approximation (SSA)**

Time-averages of the parameter values are calculated over the complete planning horizon. The averages are then used as constant parameters for a single steady-state model. The approach delivers a constant allocation over the planning horizon.

(ii) **Stationary independent period by period approximation (SIPP)**

The decision problem is decomposed in time. For each phase with constant buffer capacities, time-averages of the parameter values are used as input for the steady-state performance evaluation. This approach allows for time-dependent allocations by solving $I + 1$ independent allocation problems.

The advantage of the two approaches is that they make the rich body of literature on methods for the performance evaluation in steady state accessible for the optimization of time-dependent systems. However, both approaches neglect some of the time-dependent behavior induced by the parameter changes. Consequently, neither feasibility nor optimality with respect to the original problem can be guaranteed.

6.5 Numerical study

In the following, the proposed solution approaches from Section 6.4 are applied to solve the Proactive Kanban Card Setting Problem for different parameter configurations. We first focus on the case with $M = 1$ and compare the proposed local search algorithm to card settings based on steady-state models and constant allocations. Subsequently, the effects of changing processing time distributions in lines with $M > 1$ are investigated. Finally, we comment on the runtime performance of the local search algorithm for all investigated cases. Moreover, we compare the obtained solution to optimal results generated by a complete enumeration.

The planning horizon of all tests is $T = 32\,\text{h} = 1920$ minutes. We set the restriction of the buffer capacity to $b_m^{max} = 20$. The demand is assumed to arrive according to a Poisson process with rate $\lambda_c(t) = 0.5,\ t \in [0, 1920]$ (orders per minute).
For the search algorithm we use 20,000 replications in the performance evaluation with a Java-based discrete-event simulation.

6.5.1 Impact of the timing of buffer changes

We first consider a single-stage system ($M = 1$) and a goal service level of $\gamma^* = 0.85$. The processing times are exponentially distributed. Motivated by a replacement of machinery we assume a time-dependent processing rate that increases from $\mu_1(t) = 2/3,\ t \in [0; 960)$ at $t = 960$ minutes to $\mu_1(t) = 1$, $t \in [960; 1920)$ jobs per minute. We first analyze the case of a single

change in the buffer allocation ($I = 1$) synchronously with the rate change at $t_1^* = 960$ in detail, before we turn to a sensitivity analysis of cases with different t_1^*.

The Proactive Kanban Allocation (PKA) obtained with the algorithm introduced in Section 6.4.3 and the results of the three benchmark approaches are provided in Table 6.2. The time-dependent behavior of the expected WIP, $\mathrm{E}[W(\mathbf{B}, t)]$, in the system at time t is depicted in Figure 6.4. This detailed time-dependent performance evaluation is based on discrete-event simulation with 100,000 replications.

Table 6.2: Comparison of allocations and resulting performance for $M = 1, I = 1, t_1^* = 960, \gamma^* = 0.85, \mu_1(t) = 2/3,\ t \in [0; 960)$, $\mu_1(t) = 1,\ t \in [960; 1920)$, and $\lambda_c(t) = 0.5\ \forall t \in [0, 1920]$

Approach	$B_{1,0}$	$B_{1,1}$	$SL^{\gamma}(\mathbf{B})$	$\mathrm{E}[W(\mathbf{B})]$
PKA	13	4	0.854	7.946
SSA	5	5	0.440	4.872
SIPP	12	3	0.809	7.018
CONST	12	12	0.869	11.442

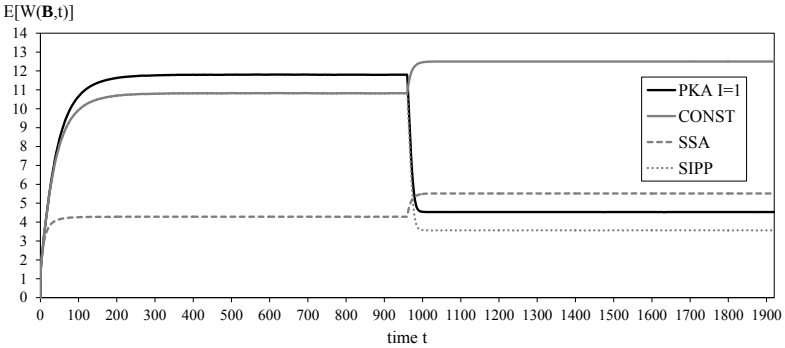


Figure 6.4: Expected WIP over time for $I = 1, t_1^* = 960$ for PKA, constant, and steady-state based allocations

Both allocations based on steady-state models, SSA and SIPP, lead to infeasible solutions. The averaging of the processing rates over time by the SSA leads to a drastic underestimation of the required buffer capacity. Even

though the SIPP approach gets closer to the desired service level it still underestimates the required buffer capacities, as it neglects the transient effects.

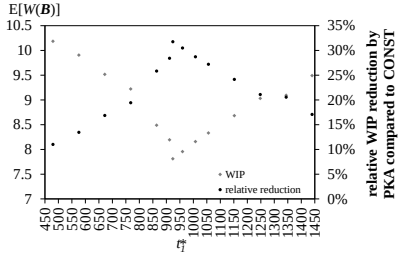
To create a benchmark for the objective value based on a feasible allocation, we determine the allocation given a constant allocation (CONST). To obtain this allocation we solve the original problem with the additional constraint $B_{M,0} = B_{M,1}$. Thus, this approach captures the time-dependent system behavior. The constant allocation has a lower buffer capacity in the first half of the planning horizon compared to the PKA. However, in the second half of the planning horizon the capacity of the constant allocation is three times greater than in the PKA. This leads to a substantial accumulation of WIP for the constant allocation (see Figure 6.4). A 30.5% reduction of the expected average WIP can be achieved by applying the time-dependent allocation.

So far we assumed a synchronous change of the buffer capacities and the processing rate. In the following, we investigate the sensitivity of the solutions with respect to the timing of the buffer change t_1^*.

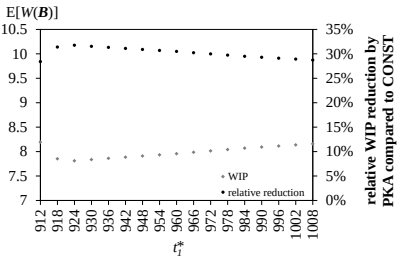
Figure 6.5a depicts the expected average WIP, $\mathrm{E}[W(\mathbf{B})]$, obtained from the PKA and the relative reduction of the expected average WIP compared to the constant allocation for values of t_1^* ranging from 480 to 1440. By definition the results for the constant allocation are independent of t_1^* and can be found in Table 6.2. Figure 6.5b depicts the respective results for t_1^* in the range from 912 to 1008, i.e., values of t_1^* close to the rate change of $\mu_1(t)$ at $t = 960$ minutes.

For the tested cases the lowest expected average WIP and respectively the highest reduction is achieved for a change of the buffer capacity at $t_i^* = 924$. This illustrates the potential of a planned buffer change, even before the change in the processing rate of $\mu_1(t)$ at $t = 960$ minutes occurs. Comparing the results of $t_1^* < 924$ with the results for $t_1^* > 924$, it becomes apparent that the expected average WIP grows faster for buffer changes before $t = 924$ minutes compared to changing the allocation after $t = 924$ minutes (see Figure 6.5a).

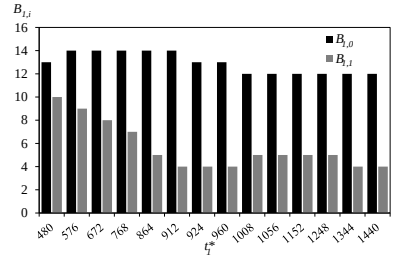
The corresponding PKAs are depicted in Figures 6.5c and 6.5d. It can be observed from Figure 6.5c that smaller values of t_1^* tend to require lager values of $B_{M,1}$. This is because the buffer capacity $B_{M,1}$ has to be adapted before the rate change, i.e, when still a larger capacity is required. If t_1^* is set close to the rate change at $t = 960$, the PKA is independent of the timing of the buffer change t_1^*, (see Figure 6.5d).



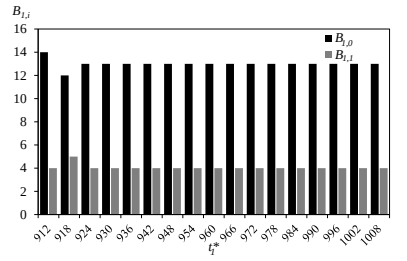
(a) Expected average WIP and relative WIP reduction compared to CONST allocations for $t_1^* \in [480; 1440]$



(b) Expected average WIP and relative WIP reduction compared to CONST allocations for $t_1^* \in [912; 1008]$



(c) Allocations for different $t_1^* \in [480; 1440]$



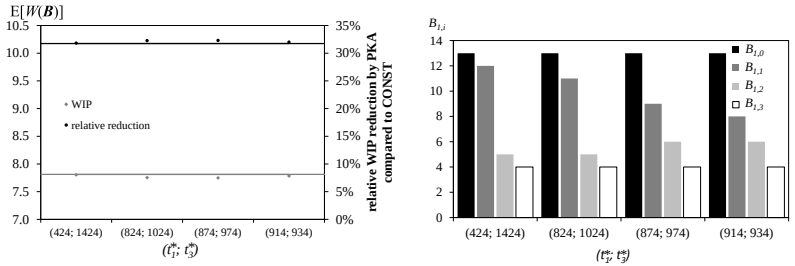
(d) Allocations for different $t_1^* \in [912; 1008]$

Figure 6.5: Impact of the timing of the buffer change t_1^* ($I = 1$)

6.5.2 Impact of the number of buffer changes

Next we consider the potential of additional buffer capacity changes. The expected average WIP reductions if two additional changes are allowed ($I = 3$) are depicted in Figure 6.6a.

The best allocation that is obtained with only a single change ($I = 1$) at time $t = 924$ serves as a benchmark, represented by the grey and black lines for the expected average WIP and the relative WIP improvement compared to a constant allocation, respectively. To ensure that the benchmark allocation is also a potential allocation for the case with $I = 3$ changes we set $t_2^* = 924$ minutes and only vary the timing of the first and the third buffer change, t_1^* and t_3^*.



(a) Expected average WIP and relative WIP reduction compared to CONST allocations for $I = 3, t_2^* = 924$

(b) Allocations for different t_i^* and $I = 3$, $t_2^* = 924$

Figure 6.6: Impact of the number of buffer changes I and their timing t_i^*

Figure 6.6b reveals that the capacity of the buffer in the beginning, $B_{1,0} = 13$, and at the end of the planning horizon, $B_{1,3} = 4$, equals the capacities from the PKAs allocation in the single change ($I = 1$) case for all tested combinations of t_1^* and t_3^*. The additional flexibility of setting values for $B_{1,1}$ and $B_{1,2}$ is used to reduce the buffer capacity in smaller steps.

From the tested configurations the one with additional changes 50 minutes before and 50 minutes after $t_2^* = 924$ provides the smallest expected average WIP value. However the additional reduction of the expected average WIP compared to a single change ($I = 1$) amounts only to 1%.

6.5.3 Lines with multiple stations and Erlang-k processing distributions

In this section we investigate the benefits of time-dependent buffer allocations in multi-stage systems. A process improvement at the bottleneck station $m = 1$ is considered. The processing distribution changes at $t = 960$ from an exponential distribution with rate $\mu_1(t) = 2/3, t \in [0, 960)$, to an Erlang-k distribution with $\mu_1(t) = 1, t \in [960, 1920]$, job per minute and $cv_1^2(t) = 0.5, t \in [960, 1920]$. Stations $m = 2$ and, if applicable $m = 3$, have an exponential processing distribution with constant rate $\mu_m(t) = 1$, $t \in [0, 1920], m > 1$ job per minute. We allow for a single change of the allocation, i.e., $I = 1$, at time $t_1^* = 960$ and set the goal service level to $\gamma^* = 0.8$.

Table 6.3 includes the results of the PKAs and a constant card setting. For both tested cases $M = 2$ and $M = 3$ a reduction of the expected average WIP by about 17% is achieved by a time-dependent compared to a constant allocation.

To further emphasize the need for the joint consideration of all buffer capacities, we investigate the case for which only the finished goods buffer capacity, $B_{M,i}$, is allowed to be changed (PKA- $B_{M,i}$). For the case of $M = 3$ stations, a change of the buffer capacity $B_{M,i}$ allows for a 3.6% reduction compared to a 17.7% reduction of the expected average WIP by changing the complete buffer allocation. The results illustrate that in order to take full advantage of the improved processing time distribution at the bottleneck station $m = 1$, a change of all buffer capacities in the line is required.

Table 6.3: Comparison of allocations and resulting performance for multi-stage systems $I = 1, t_1^* = 960$

	Approach	$B_{1,0}$	$B_{1,1}$	$B_{2,0}$	$B_{2,1}$	$B_{3,0}$	$B_{3,1}$	$SL^{\gamma}(\mathbf{B})$	$\mathrm{E}[W(\mathbf{B})]$
$M = 2$	CONST	5	5	7	7	-	-	0.801	12.056
	PKA- $B_{M,i}$	3	3	12	5	-	-	0.803	11.287
	PKA	5	1	9	5	-	-	0.805	9.940
$M = 3$	CONST	5	5	3	3	9	9	0.803	17.294
	PKA- $B_{M,i}$	5	5	3	3	11	6	0.805	16.680
	PKA	6	1	3	2	9	7	0.800	14.232

6.5.4 Performance of the search algorithm

Table 6.4 provides insights regarding the efficiency of the proposed local search algorithm based on the test cases described in Sections 6.5.1 to 6.5.3. We sort the different problem instances by the number of decision variables, i.e., $M \cdot (I + 1)$ for the Proactive Kanban Card Setting Problem. The test cases from Section 6.5.1 include two decision variables, the cases from Section 6.5.2 and the two-station case from Section 6.5.3 both have four decision variables, whereas six decision variables are included in the three-station case. The algorithm was also used for the determination of the CONST allocation and the PKA-$B_{M,i}$ in Section 6.5.3. This results in additional problem instances with two, three, and four decision variables. Table 6.4 includes the average number of evaluated allocations for the different numbers of de-

cision variables. Moreover, it provides the average percentage of evaluated allocations of all possible allocations. Even though the number of evaluations increases with the number of decision variables, the share of evaluated allocations decreases. This indicates the efficiency of the algorithm also for problems with over 85 million possible allocations, e.g., for the discussed three station case.

Table 6.4: Computational efficiency of the local search algorithm

No. of decision variables	2	3	4	6
Average no. of evaluated alloc.	16.5	156	808	47,637
Average % of all alloc. evaluated	3.736	1.684	0.415	0.056

We investigate the quality of the obtained solutions by comparison to a complete enumeration. A comparison is only possible for problems with a maximum of four decision variables due to run times that otherwise exceed several weeks. For all tested cases the local search algorithm terminates with the optimal solution. Moreover, both Observations 6.4.3 and 6.4.4 hold for all of these cases.

6.6 Conclusion and further research

Proactive Kanban Allocations are proposed as an approach that accounts for time-dependent parameters by means of time-dependent changes of buffer capacities. The objective is to minimize the required expected average WIP while maintaining a predefined γ-service level over a finite planning horizon. Monotonicity of the service level and the expected average WIP with respect to time-dependent buffer capacities is observed. The observations are used to establish dominance relations between time-dependent buffer allocations. A search algorithm that is based on these dominance criteria obtains accurate solutions while evaluating only a small percentage of all possible allocations. The numerical study demonstrates that the proposed approach reduces the required WIP compared to constant allocations. Moreover, allocation approaches based on steady-state models are tested. It is shown that they may lead to insufficient service levels. Further we provide an example for which a proactive change in the buffer allocation before the rate change is advantageous. For multi-stage systems the numerical study provides an example for which a time-dependent change of all buffer capacities in the line is re-

quired to take full advantage of the improved processing time distribution at a bottleneck station.

The proposed approach leaves the potential for further methodological enhancements. For instance, a proof of the observed monotonicity results is of interest. In addition, the development of more advanced heuristics to obtain solutions for larger systems is another field for future research.
Moreover, variations of the investigated decision problem are also worth to be considered. For the finite-horizon problem, the use of service level goals for subperiods of the planning horizon could be considered. Furthermore, extensions of other control policies such as CONWIP to a time-dependent setting can be developed.

7 Conclusions and Outlook

7.1 Conclusions

This thesis suggests and investigates approaches for the analysis of buffer allocations in flow lines under stochastic and time-dependent influences.

Chapter 2 introduces a classification scheme which characterizes unreliable flow lines and reviews various different formulations of the Buffer Allocation Problem. It is observed that in many cases not all the characteristics which are required to reproduce the models are reported in the literature. Moreover, all of the reviewed articles assume steady-state conditions.

A structured overview of approaches for the performance evaluation of time-dependent queueing systems is provided in Chapter 3. Links between the different approaches are established and discussed. It can be observed that numerical comparisons exist for only a subset of the reviewed approaches. Moreover, there are no exact analytical solutions for time-dependent queueing systems with finite buffers. In the literature only a single approximate approach for time-dependent buffer capacities in a call center exists.

Chapter 4 includes two new sampling approaches for the performance evaluation of time-dependent unreliable flow lines with constant and finite buffer capacities. One is based on a mixed-integer program in discrete time with discrete material, while the other approximation is based on partial and ordinary differential equations in continuous time and with a continuous flow of material. It can be demonstrated that both approaches coincide on a deterministic level which links the two corresponding literature streams. The numerical study demonstrates the accuracy of both approaches. Moreover, increased buffer capacities help to smooth the cumulative output over time, given a time-dependent release rate to the flow line.

Chapter 5 presents an initial approach towards time-dependent buffer capacities in flow lines. The MIP sampling approach of the previous chapter is modified to allow the buffer capacities to change over time. A numerical study indicates that changes of buffer capacities over time can be applied to

ensure a stable throughput, given time-dependent changes in the production rate.

Finally, Chapter 6 proposes a time-dependent change in buffer capacities by utilizing Kanban cards to minimize the required expected average WIP while maintaining a predefined service level over a finite planning horizon. A numerical study suggest monotonicity properties for the service level and the expected average WIP with respect to time-dependent buffer capacities. Based on these observed properties, a local search algorithm is developed. The algorithm allows to obtain solutions while evaluating only a small percentage of all possible allocations. The numerical study illustrates that the proposed approach reduces the required WIP to fulfil a given service level when compared to constant allocations over time. Moreover, allocation approaches based on steady-state models are tested. These may lead to insufficient service levels.

7.2 Further possible research directions

This thesis sheds a first light on time-dependent changes in buffer capacities to account for time-dependent parameters in flow lines. An initial attempt has been made to provide performance evaluation approaches and a systematic solution of the trade-off between service level and WIP in the line. However, there is still room for further methodological improvements.

The performance evaluation approaches used in Chapters 4 to 6 require the generation of random numbers. This produces a simulation error which can only be reduced by increasing the number of replications, which in turn increases the computation times. Future research should explore how existing ideas reviewed in Chapter 3 can be combined to further increase the quality of the approximation, and how these ideas can be adapted for the analysis of time-dependent and stochastic flow lines.
A rigorous proof for the observed monotonicity results is also of interest. Even under steady-state conditions, theoretical insights on how the buffer allocations influence the WIP in flow lines with multiple stations are missing. Consequently, theoretical support for the observations of So (1997) and Papadopoulos and Vidalis (2001b) is desirable. Moreover, the identification of additional structural properties which result in bounds on the objective value or the required buffer capacities could contribute to a speeding-up of the solution process.
The development of advanced heuristics to obtain solutions for larger systems is another field for future research. The proposed local search algorithm might

be one starting point for their development, e.g., by deliberately not solving all subproblems. Another direction to be explored is the use of metaheuristics such as Simulated Annealing or Genetic Algorithms. They have been successfully applied to solve the Buffer Allocation Problem under steady-state conditions (see Chapter 2). Due to their generic structure they are also applicable for the Proactive Kanban Card Setting Problem.

Proactive Kanban is a first approach to address the trade-off between the desired customer service level and WIP in flow lines with time-dependent parameters. Variations of this decision problem are also worth considering.

In the flow line literature, the use of expected values of performance measures is common (see Chapter 2). The use of higher moments and quantiles of the distribution of performance measures can be a meaningful extension. One example for such an additional requirement could be a service level goal that has to be met with a certain probability. In addition, the use of service level goals for subperiods of the planning horizon could be considered. This prevents solutions in which poor performance during a given period is compensated by overachieving the performance in another period.
By allowing the removal of all Kanban cards from a stage, the Proactive Kanban Card Setting Problem can be extended to solve an order release problem. By removing all Kanban cards from the first stage, the inflow of new workpieces is stopped. This offers the potential to further reduce the WIP in the line. This extension makes the approach also applicable to problems with constant parameters over time, but with additional constraints on the inventory level. This would be the case, for instance, if a line starts empty and all workpieces have to be cleared from the line again at the end of the planning horizon, e.g., when processing perishable goods which cannot be allowed to stay in the buffers over night.
Chapter 6 describes a Proactive Kanban policy and focuses on determining its parameters. Extensions of other control policies such as CONWIP and Extended Kanban to a time-dependent setting are another potential field for future research. Ng et al. (2012) and Xanthopoulos and Koulouriotis (2014) compare numerically different policies under steady-state conditions. Whether their findings can be translated to the time-dependent setting or not is well worth investigation.

The consideration of new inventory control policies and decision problems may also require methodological advances, such as establishing new structural properties and developing new analytical performance evaluation and optimization algorithms.

Bibliography

Abol'nikov, L. M. (1968). A nonstationary queueing problem for a system with an infinite number of channels for a group arrival of requests. *Problems of Information Transmission*, 4(3):82–85.

Agnew, C. E. (1976). Dynamic Modeling and Control of Congestion-Prone Systems. *Operations Research*, 24(3):400–419.

Agnihothri, S. R. and Taylor, P. F. (1991). Staffing a Centralized Appointment Scheduling Department in Lourdes Hospital. *Interfaces*, 21(5):1–11.

Aguir, S., Karaesmen, F., Aksin, O. Z., and Chauvet, F. (2004). The impact of retrials on call center performance. *OR Spectrum*, 26(3):353–376.

Al-Seedy, R. O. and Al-Ibraheem, F. M. (2003). New transient solution to the $M/M/\infty$ queue with varying arrival and departure rate. *Applied Mathematics and Computation*, 135(2-3):425–428.

Al-Seedy, R. O., El-Sherbiny, A. A., El-Shehawy, S. A., and Ammar, S. I. (2009). The transient solution to a time-dependent single-server queue with balking. *The Mathematical Scientist*, 34(2):113–118.

Alfa, A. S. (1979). A Numerical Method for Evaluating Delay to a Customer in a Time-Inhomogeneous, Single Server Queue with Batch Arrivals. *The Journal of the Operational Research Society*, 30(7):665–667.

Alfa, A. S. (1982). Time-Inhomogenous Bulk Server Queue in Discrete Time: A Transportation Type Problem. *Operations Research*, 30(4):650–658.

Alfa, A. S. (1990). Approximating queue lengths in $M(t)/D/1$ queues. *European Journal of Operational Research*, 44(1):60–66.

Alfa, A. S. and Chen, M. (1991). Approximating queue lengths in $M(t)/G/1$ queue using the maximum entropy principle. *Acta Informatica*, 28(8):801–815.

Alfa, A. S. and Margolius, B. H. (2008). Two classes of time-inhomogeneous Markov chains: Analysis of the periodic case. *Annals of Operations Research*, 160(1):121–137.

Alfieri, A. and Matta, A. (2012). Mathematical programming formulations for approximate simulation of multistage production systems. *European Journal of Operational Research*, 219(3):773–783.

Alnowibet, K. A. and Perros, H. (2006). The Nonstationary Loss Queue: A Survey. In Barria, J. A., editor, *Communication Networks and Computer Systems - A Tribute to Professor Erol Gelenbe*, pages 105–125. Imperial College Press.

Alon, G., Kroese, D. P., Raviv, T., and Rubinstein, R. Y. (2005). Application of the cross-entropy method to the buffer allocation problem in a simulation-based environment. *Annals of Operations Research*, 134(1):137–151.

Anderson, D. R. and Moodie, C. L. (1968). Optimal buffer storage capacity in production. *International Journal of Production Research*, 7(3):233–240.

Andrews, B. and Parsons, H. (1993). Establishing Telephone-Agent Staffing Levels through Economic Optimization. *Interfaces*, 23(2):14–20.

Arns, M., Buchholz, P., and Panchenko, A. (2010). On the Numerical Analysis of Inhomogeneous Continuous-Time Markov Chains. *INFORMS Journal on Computing*, 22(3):416–432.

Atlason, J., Epelman, M. A., and Henderson, S. G. (2008). Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. *Management Science*, 54(2):295–309.

Axsäter, S. and Rosling, K. (1993). Installation vs. Echelon Stock Policies for Multilevel Inventory Control. *Management Science*, 39(10):1274–1280.

Bekker, J. (2013). Multi-objective buffer space allocation with the cross-entropy method. *International Journal of Simulation Modelling*, 12(1):50–61.

Bendoly, E., Donohue, K., and Schultz, K. L. (2006). Behavior in operations management: Assessing recent findings and revisiting old assumptions. *Journal of Operations Management*, 24(6):737–752.

Bennett, J. C. and Worthington, D. J. (1998). An Example of a Good but Partially Successful OR Engagement: Improving Outpatient Clinic Operations. *Interfaces*, 28(5):56–69.

Berkley, B. J. (1991). Tandem queues and kanban-controlled lines. *International Journal of Production Research*, 29(10):2057–2081.

Berkley, B. J. (1992). A review of the kanban production control research literature. *Production and Operations Management*, 1(4):393–411.

Bertsimas, D. and Doan, X. V. (2010). Robust and data-driven approaches to call centers. *European Journal of Operational Research*, 207(2):1072–1085.

Bertsimas, D. and Mourtzinou, G. (1997). Transient laws of non-stationary queueing systems and their applications. *Queueing Systems*, 25(1-4):115–155.

Bhattacharjee, P. and Ray, P. K. (2014). Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. *Computers & Industrial Engineering*, 78:299–312.

Blumberg-Nitzani, M. and Bar-Gera, H. (2014). The effect of signalised intersections on dynamic traffic assignment solution stability. *Transportmetrica A: Transport Science*, 10(7):622–646.

Bollapragada, R. and Rao, U. S. (2006). Replenishment planning in discrete-time, capacitated, non-stationary, stochastic inventory systems. *IIE Transactions*, 38(7):583–595.

Bookbinder, J. H. (1986). Multiple queues of aircraft under time-dependent conditions. *INFOR*, 24(4):280–288.

Bookbinder, J. H. and Martell, D. L. (1979). Time-dependent queueing approach to helicopter allocation for forest fire initial-attack. *INFOR*, 17(1):58–70.

Brahimi, M. and Worthington, D. J. (1991a). Queueing Models for Out-Patient Appointment Systems – A Case Study. *Journal of the Operational Research Society*, 42(9):733–746.

Brahimi, M. and Worthington, D. J. (1991b). The finite capacity multi-server queue with inhomogeneous arrival rate and discrete service time distribution - and its application to continuous service time problems. *European Journal of Operational Research*, 50(3):310–324.

Brilon, W. and Wu, N. (1990). Delays at fixed-time traffic signals under time-dependent traffic conditions. *Traffic Engineering & Control*, 31(12):623–631.

Brown, M. and Ross, S. M. (1969). Some Results for Infinite Server Poisson Queues. *Journal of Applied Probability*, 6(3):604–611.

Buczkowski, P. S. and Kulkarni, V. G. (2006). Funding a warranty reserve with contributions after each sale. *Probability in the Engineering and Informational Sciences*, 20(3):497–515.

Burman, M. H., Gershwin, S. B., and Suyematsu, C. (1998). Hewlett-Packard uses operations research to improve the design of a printer production line. *Interfaces*, 28(1):24–36.

Buzacott, J. A. and Hanifin, L. E. (1978). Models of automatic transfer lines with inventory banks a review and comparison. *AIIE Transactions*, 10(2):197–207.

Carrillo, M. J. (1991). Extensions of Palm's Theorem: A Review. *Management Science*, 37(6):739–744.

Catling, I. (1977). A time-dependent approach to junction delays. *Traffic Engineering & Control*, 18(11):520–526.

Chan, W. K. and Schruben, L. (2008). Optimization Models of Discrete-Event System Dynamics. *Operations Research*, 56(5):1218–1237.

Chassioti, E., Worthington, D., and Glazebrook, K. (2014). Effects of state-dependent balking on multi-server non-stationary queueing systems. *Journal of the Operational Research Society*, 65(2):278–290.

Chassioti, E. and Worthington, D. J. (2004). A new model for call centre queue management. *Journal of the Operational Research Society*, 55(12):1352–1357.

Chen, G., Govindan, K., and Golias, M. M. (2013a). Reducing truck emissions at container terminals in a low carbon economy: Proposal of a queueing-based bi-objective model for optimizing truck arrival pattern. *Transportation Research Part E: Logistics and Transportation Review*, 55:3–22.

Chen, G., Govindan, K., and Yang, Z. (2013b). Managing truck arrivals with time windows to alleviate gate congestion at container terminals. *International Journal of Production Economics*, 141(1):179–188.

Chen, G., Govindan, K., Yang, Z.-Z., Choi, T.-M., and Jiang, L. (2013c). Terminal appointment system design by non-stationary $M(t)/E_k/c(t)$ queueing model and genetic algorithm. *International Journal of Production Economics*, 146(2):694–703.

Chen, G. and Yang, Z. (2010). Optimizing time windows for managing export container arrivals at Chinese container terminals. *Maritime Economics & Logistics*, 12(1):111–126.

Chen, G. and Yang, Z.-Z. (2014). Methods for estimating vehicle queues at a marine terminal: A computational comparison. *International Journal of Applied Mathematics and Computer Science*, 24(3):611–619.

Chen, J., Lin, D. K. J., and Thomas, D. J. (2003). On the single item fill rate for a finite horizon. *Operations Research Letters*, 31(2):119–123.

Chen, X., Zhou, X., and List, G. F. (2011). Using time-varying tolls to optimize truck arrivals at ports. *Transportation Research Part E: Logistics and Transportation Review*, 47(6):965–982.

Chiang, S.-Y., Kuo, C.-T., and Meerkov, S. M. (2000). DT-bottlenecks in serial production lines: Theory and application. *IEEE Transactions on Robotics and Automation*, 16(5):567–580.

Choudhury, G. L., Lucantoni, D. M., and Whitt, W. (1997). Numerical Solution of Piecewise-Stationary $M_t/G_t/1$ Queues. *Operations Research*, 45(3):451–463.

Chung, K. and Min, D. (2014). Staffing a service system with appointment-based customer arrivals. *Journal of the Operational Research Society*, 65(10):1533–1543.

Clark, A. J. and Scarf, H. (1960). Optimal Policies for a Multi-Echelon Inventory Problem. *Management Science*, 6(4):475–490.

Clark, G. M. (1981). Use of Polya Distributions in Approximate Solutions to Nonstationary $M/M/s$ Queues. *Communications of the ACM*, 24(4):206–217.

Clarke, A. B. (1956). A Waiting Line Process of Markov Type. *The Annals of Mathematical Statistics*, 27(2):452–459.

Coclite, G. M., Garavello, M., and Piccoli, B. (2005). Traffic Flow on a Road Network. *SIAM Journal on Mathematical Analysis*, 36(6):1862–1886.

Collings, T. and Stoneman, C. (1976). The $M/M/\infty$ Queue with Varying Arrival and Departure Rates. *Operations Research*, 24(4):760–773.

Cosmetatos, G. P. (1976). Some Approximate Equilibrium Results for the Multi-Server Queue $(M/G/r)$. *Operational Research Quarterly*, 27(3):615–620.

Creemers, S., Defraeye, M., and van Nieuwenhuyse, I. (2014). G-RAND: A phase-type approximation for the nonstationary $G(t)/G(t)/s(t) + G(t)$ queue. *Performance Evaluation*, 80:102–123.

Curry, G. L., De Vany, A., and Feldman, R. M. (1978). A queueing model of airport passenger departures by taxi: Competition with a public transportation mode. *Transportation Research*, 12(2):115–120.

Czachórski, T., Grochla, K., Nycz, T., and Pekergin, F. (2010). A diffusion approximation model for wireless networks based on IEEE 802.11 standard. *Computer Communications*, 33:S86–S92.

Czachórski, T., Nycz, T., and Pekergin, F. (2009). Diffusion Approximation Models for Transient States and their Application to Priority Queues. *International Journal On Advances in Networks and Services*, 2(2&3):205–217.

Dai, L. (1998). Performance Bounds for Nonhomogeneous Queues. *IEEE Transactions on Automatic Control*, 43(5):700–705.

Dallery, Y. and Gershwin, S. B. (1992). Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems*, 12(1-2):3–94.

Daniel, J. I. (1995). Congestion Pricing and Capacity of Large Hub Airports: A Bottleneck Model with Stochastic Queues. *Econometrica*, 63(2):327–370.

Daniel, J. I. and Harback, K. T. (2008). (When) Do hub airlines internalize their self-imposed congestion delays? *Journal of Urban Economics*, 63(2):583–612.

Daniel, J. I. and Harback, K. T. (2009). Pricing the major US hub airports. *Journal of Urban Economics*, 66(1):33–56.

Daniel, J. I. and Pahwa, M. (2000). Comparison of Three Empirical Models of Airport Congestion Pricing. *Journal of Urban Economics*, 47(1):1–38.

D'Apice, C., Göttlich, S., Herty, M., and Piccoli, B. (2010). *Modeling, Simulation and Optimization of Supply Chains: A Continuous Approach*. SIAM, Philadelphia.

Davis, M. H. A. (1984). Piecewise-deterministic Markov Processes: A General Class of Non-Diffusion Stochastic Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):353–388.

Davis, M. H. A. (1993). *Markov Models and Optimization*. Monographs on Statistics and Applied Probability. Chapman & Hall, Boca Raton.

de Barros, A. G. and Tomber, D. D. (2007). Quantitative Analysis of Passenger and Baggage Security Screening at Airports. *Journal of Advanced Transportation*, 41(2):171–193.

de Bruin, A. M., van Rossum, A. C., Visser, M. C., and Koole, G. M. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137.

de Neufville, R. and Grillot, M. (1982). Design of Pedestrian Space in Airport Terminals. *Transportation Engineering Journal of ASCE*, 108(1):87–102.

Defraeye, M. and Van Nieuwenhuyse, I. (2011). Setting staffing levels in an emergency department: opportunities and limitations of stationary queueing models. *Review of Business and Economics*, 56(1):73 – 100.

Defraeye, M. and Van Nieuwenhuyse, I. (2016). Staffing and scheduling under non-stationary demand for service: A literature review. *Omega*, 58:4–25.

Degond, P. and Ringhofer, C. (2007). Stochastic Dynamics of Long Supply Chains with Random Breakdowns. *SIAM Journal on Applied Mathematics*, 68(1):59–79.

Demir, L., Tunali, S., and Eliiyi, D. T. (2014). The state of the art on buffer allocation problem: A comprehensive survey. *Journal of Intelligent Manufacturing*, 25(3):371–392.

Demir, L., Tunali, S., and Løkketangen, A. (2011). A tabu search approach for buffer allocation in production lines with unreliable machines. *Engineering Optimization*, 43(2):213–231.

Deng, C. C., Ong, H. L., Ang, B. W., and Goh, T. N. (1992). A Modelling Study of a Taxi Service Operation. *International Journal of Operations & Production Management*, 12(11):65–78.

Di Crescenzo, A. and Nobile, A. G. (1995). Diffusion approximation to a queueing system with time-dependent arrival and service rates. *Queueing Systems*, 19(1-2):41–62.

Diamantidis, A. C. and Papadopoulos, C. T. (2004). A dynamic programming algorithm for the buffer allocation problem in homogeneous asymptotically reliable serial production lines. *Mathematical Problems in Engineering*, 3:209–223.

Diaz, R. and Ardalan, A. (2010). An Analysis of Dual-Kanban Just-In-Time Systems in a Non-Repetitive Environment. *Production and Operations Management*, 19(2):233–245.

Dietz, D. C. (2011). Practical scheduling for call center operations. *Omega*, 39(5):550–557.

Dietz, D. C. and Vaver, J. G. (2006). Synergistic modeling of call center operations. *Journal of Applied Mathematics and Decision Sciences*, 2006(2):1–13.

Dolgui, A., Eremeev, A., Kolokolov, A., and Sigaev, V. (2002). A Genetic Algorithm for the Allocation of Buffer Storage Capacities in a Production Line with Unreliable Machines. *Journal of Mathematical Modelling and Algorithms*, 1(2):89–104.

Dolgui, A., Eremeev, A. V., and Sigaev, V. S. (2007). HBBA: Hybrid algorithm for buffer allocation in tandem production lines. *Journal of Intelligent Manufacturing*, 18(3):411–420.

Dormuth, D. W. and Alfa, A. S. (1997). Two finite-difference methods for solving $MAP(t)/PH(t)/1/K$ queueing models. *Queueing Systems*, 27(1-2):55–78.

Duda, A. (1986). Diffusion Approximations for Time-Dependent Queueing Systems. *IEEE Journal on Selected Areas in Communications*, 4(6):905–918.

Eick, S. G., Massey, W. A., and Whitt, W. (1993a). $M_t/G/\infty$ Queues with Sinusoidal Arrival Rates. *Management Science*, 39(2):241–252.

Eick, S. G., Massey, W. A., and Whitt, W. (1993b). The Physics of the $M_t/G/\infty$ Queue. *Operations Research*, 41(4):731–742.

El-Sherbiny, A. A. (2010). Transient Solution to an infinite Server Queue with Varying Arrival and Departure Rate. *Journal of Mathematics and Statistics*, 6(1):1–3.

Ellis, P. M. (2010). The Time-Dependent Mean and Variance of the Non-Stationary Markovian Infinite Server System. *Journal of Mathematics and Statistics*, 6(1):68–71.

Enginarlar, E., Li, J., and Meerkov, S. M. (2005). How lean can lean buffers be? *IIE Transactions*, 37(4):333–342.

Enginarlar, E., Li, J., Meerkov, S. M., and Zhang, R. Q. (2002). Buffer capacity for accommodating machine downtime in serial production lines. *International Journal of Production Research*, 40(3):601–624.

Escobar, M., Odoni, A. R., and Roth, E. (2002). Approximate solution for multi-server queueing systems with Erlangian service times. *Computers & Operations Research*, 29(10):1353–1374.

Fan, W. (1976). Simulation of queueing network with time varying arrival rates. *Mathematics and Computers in Simulation*, 18(3):165–170.

Feldman, Z., Mandelbaum, A., Massey, W. A., and Whitt, W. (2008). Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*, 54(2):324–338.

Filipiak, J. (1983). Diffusion-equation model of slightly loaded $M/M/1$ queue. *Operations Research Letters*, 2(3):134–139.

Filipiak, J. (1984). Dynamic Routing in a Queueing System with a Multiple Service Facility. *Operations Research*, 32(5):1163–1180.

Fleischer, J., Weule, H., and Lanza, G. (2004). Quality Simulation for Optimization During Production Ramp-up. *Production Engineering*, 11(2):147–150.

Flick, A. and Liao, M. (2010). A queuing system with time varying rates. *Statistics & Probability Letters*, 80(5-6):386–389.

Foley, R. D. (1982). The non-homogeneous $M/G/\infty$ queue. *OpSearch*, 19(1):40–48.

Foote, B. L. (1976). A Queueing Case Study of Drive-In Banking. *Interfaces*, 6(4):31–37.

Fügenschuh, A., Göttlich, S., Herty, M., Klar, A., and Martin, A. (2008). A discrete optimization approach to large scale supply networks based on partial differential equations. *SIAM Journal on Scientific Computing*, 30(3):1490–1507.

Galliher, H. P. and Wheeler, R. C. (1958). Nonstationary Queuing Probabilities for Landing Congestion of Aircraft. *Operations Research*, 6(2):264–275.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.

Garavello, M. and Goatin, P. (2012). The Cauchy problem at a node with buffer. *Discrete and Continuous Dynamical Systems-Series A*, 32(6):1915–1938.

Gaury, E. G. A., Pierreval, H., and Kleijnen, J. P. C. (2000). An evolutionary approach to select a pull system among Kanban, Conwip and Hybrid. *Journal of Intelligent Manufacturing*, 11(2):157–167.

Gaver, D. P. (1969). Highway Delays Resulting from Flow-Stopping Incidents. *Journal of Applied Probability*, 6(1):137–153.

Gershwin, S. B. and Schick, I. C. (1983). Modeling and analysis of three-stage transfer lines with unreliable machines and finite buffers. *Operations Research*, 31(2):354–380.

Gershwin, S. B. and Schor, J. E. (2000). Efficient algorithms for buffer space allocation. *Annals of Operations Research*, 93(1-4):117–144.

Gillard, J. and Knight, V. (2014). Using Singular Spectrum Analysis to obtain staffing level requirements in emergency units. *Journal of the Operational Research Society*, 65(5):735–746.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434.

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733.

Giorno, V., Nobile, A. G., and Ricciardi, L. M. (1987). On Some Time-Non-Homogeneous Diffusion Approximations to Queueing Systems. *Advances in Applied Probability*, 19(4):974–994.

Glasserman, P. and Yao, D. D. (1996). Structured buffer-allocation problems. *Discrete Event Dynamic Systems: Theory and Applications*, 6(1):9–41.

Göttlich, S., Herty, M., and Klar, A. (2005). Network models for supply chains. *Communications in Mathematical Sciences*, 3(4):545–559.

Göttlich, S., Kühn, S., Schwarz, J. A., and Stolletz, R. (2016). Approximations of time-dependent unreliable flow lines with finite buffers. *Mathematical Methods of Operations Research*, (In Press):DOI: 10.1007/s00186–015–0529–6, 1–29.

Göttlich, S., Martin, S., and Sickenberger, T. (2011). Time-continuous production networks with random breakdowns. *Networks and Heterogeneous Media*, 6(4):695–714.

Grassmann, W. (1977a). Transient solutions in Markovian queues - An algorithm for finding them and determining their waiting-time distributions. *European Journal of Operational Research*, 1(6):396–402.

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

Green, L. and Kolesar, P. (1991). The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals. *Management Science*, 37(1):84–97.

Green, L., Kolesar, P., and Svoronos, A. (1991). Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research*, 39(3):502–511.

Green, L. V. and Kolesar, P. J. (1995). On the Accuracy of the Simple Peak Hour Approximation for Markovian Queues. *Management Science*, 41(8):1353–1370.

Green, L. V. and Kolesar, P. J. (1997). The Lagged PSA for Estimating Peak Congestion in Multiserver Markovian Queues with Periodic Arrival Rates. *Management Science*, 43(1):80–87.

Green, L. V. and Kolesar, P. J. (1998). A Note on Approximating Peak Congestion in $M_t/G/\infty$ Queues with Sinusoidal Arrivals. *Management Science*, 44(11):S137–S143.

Green, L. V., Kolesar, P. J., and 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., and 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., and Whitt, W. (2007). Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management*, 16(1):13–39.

Green, L. V. and Soares, J. (2007). Computing Time-Dependent Waiting Time Probabilities in $M(t)/M/s(t)$ Queuing Systems. *Manufacturing & Service Operations Management*, 9(1):54–61.

Green, L. V., Soares, J., Giglio, J. F., and Green, R. A. (2006). Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing. *Academic emergency medicine*, 13(1):61–68.

Griffiths, J. D., Holland, W., and Williams, J. E. (1991). Estimation of Queues at the Channel Tunnel. *Journal of the Operational Research Society*, 42(5):365–373.

Griffiths, J. D., Leonenko, G. M., and Williams, J. E. (2008). Time-Dependent Analysis of Non-Empty $M/E_k/1$ Queue. *Quality Technology & Quantitative Management*, 5(3):309–320.

Gross, D. and Miller, D. R. (1984). The Randomization Technique as a Modeling Tool and Solution Procedure for Transient Markov Processes. *Operations Research*, 32(2):343–361.

Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M. (2008). *Fundamentals of Queueing Theory*. Wiley, Hoboken, 4 edition.

Haller, M., Peikert, A., and Thoma, J. (2003). Cycle time management during production ramp-up. *Robotics and Computer-Integrated Manufacturing*, 19(1-2):183–188.

Hampshire, R. C., Jennings, O. B., and Massey, W. A. (2009). A time-varying call center design via Lagrangian mechanics. *Probability in the Engineering and Informational Sciences*, 23(2):231–259.

Hampshire, R. C. and Massey, W. A. (2010). Dynamic Optimization with Applications to Dynamic Rate Queues. In *Tutorials in Operations Research*, pages 208–247. INFORMS.

Han, M.-S. and Park, D.-J. (2002). Optimal buffer allocation of serial production lines with quality inspection machines. *Computers and Industrial Engineering*, 42(1):75–89.

Harrison, J. M. and Zeevi, A. (2005). A Method for Staffing Large Call Centers Based on Stochastic Fluid Models. *Manufacturing & Service Operations Management*, 7(1):20–36.

Hebert, J. E. and Dietz, D. C. (1997). Modeling and Analysis of an Airport Departure Process. *Journal of Aircraft*, 34(1):43–47.

Helber, S. (2001). Cash-flow-oriented buffer allocation in stochastic flow lines. *International Journal of Production Research*, 39(14):3061–3083.

Helber, S., Schimmelpfeng, K., Stolletz, R., and Lagershausen, S. (2011). Using linear programming to analyze and optimize stochastic flow lines. *Annals of Operations Research*, 182(1):193–211.

Holland, W. and Griffiths, J. D. (1999). A time-dependent approximation for the queue $M/M(1,s)/c$. *IMA Journal of Mathematics Applied in Business & Industry*, 10(3):213–223.

Horonjeff, R. (1969). Analyses of Passenger and Baggage Flows in Airport Terminal Buildings. *Journal of Aircraft*, 6(5):446–451.

Iida, T. (2002). A non-stationary periodic review production-inventory model with uncertain production capacity and uncertain demand. *European Journal of Operational Research*, 140(3):670–683.

Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y., and Wu, X. (2007). A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary $M(t)/M/s(t)$ Queueing Systems with Exhaustive Discipline. *INFORMS Journal on Computing*, 19(2):201–214.

Ingolfsson, A., Campello, F., Wu, X., and Cabral, E. (2010). Combining integer programming and the randomization method to schedule employees. *European Journal of Operational Research*, 202(1):153–163.

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

Inman, R. R. (1999). Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. *Production and Operations Management*, 8(4):409–432.

Jacquillat, A. and Odoni, A. R. (2015). Endogenous control of service rates in stochastic and dynamic queuing models of airport congestion. *Transportation Research Part E: Logistics and Transportation Review*, 73:133–151.

Jagerman, D. L. (1975). Nonstationary Blocking in Telephone Traffic. *The Bell System Technical Journal*, 54(3):625–661.

Jaikumar, R. and Bohn, R. E. (1992). A dynamic approach to operations management: An alternative to static optimization. *International Journal of Production Economics*, 27(3):265–282.

Janic, M. (2005). Modelling Airport Congestion Charges. *Transportation Planning and Technology*, 28(1):1–26.

Janic, M. (2009). Modeling Airport Operations Affected by a Large-Scale Disruption. *Journal of Transportation Engineering*, 135(4):206–216.

Jennings, O. B., Mandelbaum, A., Massey, W. A., and Whitt, W. (1996). Server Staffing to Meet Time-Varying Demand. *Management Science*, 42(10):1383–1394.

Jennings, O. B. and Massey, W. A. (1997). A modified offered load approximation for nonstationary circuit switched networks. *Telecommunication Systems*, 7(1-3):229–251.

Jiménez, T. and 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.

Jung, M. and Lee, E. S. (1989a). A multi-echelon and multi-indenture repairable item queuing model during emergencies. *Mathematical and Computer Modelling*, 12(7):851–864.

Jung, M. and Lee, E. S. (1989b). Numerical Optimization of a Queueing System by Dynamic Programming. *Journal of Mathematical Analysis and Applications*, 141(1):84–93.

Jung, W. (1993). Recoverable inventory systems with time-varying demand. *Production and Inventory Management Journal*, 34(1):77–81.

Kahraman, A. and Gosavi, A. (2011). On the distribution of the number stranded in bulk-arrival, bulk-service queues of the $M/G/1$ form. *European Journal of Operational Research*, 212(2):352–360.

Kambo, N. S. and Bhalaik, H. S. (1979). A note on the nonhomogeneous $M/M/\infty$ queue. *OpSearch*, 16(2&3):103–106.

Keller, J. B. (1982). Time-Dependent Queues. *SIAM Review*, 24(4):401–412.

Khintchine, A. Y. (1969). *Mathematical methods in the theory of queueing*. Charles Griffin, London, 2nd edition.

Kim, J. W. and Ha, S. H. (2012). Advanced workforce management for effective customer services. *Quality & Quantity*, 46(6):1715–1726.

Kim, S. and Lee, H.-J. (2001). Allocation of buffer capacity to minimize average work-in-process. *Production Planning and Control*, 12(7):706–716.

Kimber, R. M. and Daly, P. N. (1986). Time-dependent queueing at road junctions: Observation and prediction. *Transportation Research Part B: Methodological*, 20(3):187–203.

Kimber, R. M. and Hollis, E. M. (1978). Peak-period traffic delays at road junctions and other bottlenecks. *Traffic Engineering & Control*, 19(10):442–446.

Kimber, R. M., Marlow, M., and Hollis, E. M. (1977). Flow/delay relationships for major/minor priority junctions. *Traffic Engineering & Control*, 18(11):516–519.

Kimura, T. (2004). Diffusion Models for Computer/Communication Systems. *Economic Journal of Hokkaido University*, 33:37–52.

Kirchner, C., Herty, M., Göttlich, S., and Klar, A. (2006). Optimal control for continuous supply network models. *Networks and Heterogenous Media*, 1(4):675–688.

Knessl, C. (2000). Exact and Asymptotic Solutions to a PDE That Arises in Time-Dependent Queues. *Advances in Applied Probability*, 32(1):256–283.

Knessl, C. and Yang, Y. (2001). Analysis of a Brownian particle moving in a time-dependent drift field. *Asymptotic Analysis*, 27(3-4):281–319.

Knessl, C. and Yang, Y. P. (2002). An Exact Solution for an $M(t)/M(t)/1$ Queue with Time-Dependent Arrivals and Service. *Queueing Systems*, 40(3):233–245.

Ko, Y. M. and Gautam, N. (2010). Transient analysis of queues for peer-based multimedia content delivery. *IIE Transactions*, 42(12):881–896.

Ko, Y. M. and Gautam, N. (2013). Critically Loaded Time-Varying Multiserver Queues: Computational Challenges and Approximations. *INFORMS Journal on Computing*, 25(2):285–301.

Kolesar, P. (1984). Stalking the Endangered CAT: A Queueing Analysis of Congestion at Automatic Teller Machines. *Interfaces*, 14(6):16–26.

Kolesar, P. J. and Green, L. V. (1998). Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management*, 7(3):282–293.

Kolesar, P. J., Rider, K. L., Crabill, T. B., and Walker, W. E. (1975). A Queuing-Linear Programming Approach to Scheduling Police Patrol Cars. *Operations Research*, 23(6):1045–1062.

Kolmogorov, A. (1931). Sur le problème d'attente. *Matematicheskii Sbornik*, 38(1-2):101–106 (in French).

Koole, G. and van der Sluis, E. (2003). Optimal shift scheduling with a global service level constraint. *IIE Transactions*, 35(11):1049–1055.

Koopman, B. O. (1972). Air-Terminal Queues under Time-Dependent Conditions. *Operations Research*, 20(6):1089–1114.

Kose, S. Y. and Kilincci, O. (2015). Hybrid approach for buffer allocation in open serial production lines. *Computers and Operations Research*, 60:67–78.

Kuraya, K., Masuyama, H., and Kasahara, S. (2011). Load distribution performance of super-node based peer-to-peer communication networks: A nonstationary Markov chain approach. *Numerical Algebra, Control and Optimization*, 1(4):593–610.

Kuraya, K., Masuyama, H., Kasahara, S., and Takahashi, Y. (2009). Decentralized user information management systems for peer-to-peer communication networks: An approach by nonstationary peer-population process. *Ubiquitous Computing and Communication Journal*, CSNDSP 2008:1–8.

Kuwahara, M. (2007). A theory and implications on dynamic marginal cost. *Transportation Research Part A: Policy and Practice*, 41(7):627–643.

Kwan, S. K., Davis, M. M., and 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):265–275.

Lackman, R. A., Spragins, J. D., and Tipper, D. (1992). Scheduling real-time and non-real-time traffic under nonstationary conditions. *Annals of Operations Research*, 36(1):193–224.

Lapre, M. A., Mukherjee, A. S., and Van Wassenhove, L. N. (2000). Behind the Learning Curve: Linking Learning Activities to Waste Reduction. *Management Science*, 46(5):597–611.

Lau, H. C. and Song, H. (2008). Multi-echelon repairable item inventory system with limited repair capacity under nonstationary demands. *International Journal of Inventory Research*, 1(1):67–92.

Lee, H.-T., Chen, S.-K., and Chang, S. (2009). A meta-heuristic approach to buffer allocation in production line. *Journal of C.C.I.T*, 38(1):167–178.

Lee, S.-D. and Ho, S.-H. (2002). Buffer sizing in manufacturing production systems with complex routings. *International Journal of Computer Integrated Manufacturing*, 15(5):440–452.

Leese, E. L. and Boyd, D. W. (1966). Numerical methods of determining the transient behaviour of queues with variable arrival rates. *Journal of the Canadian Operational Research Society*, 4(1):1–13.

LeVeque, R. J. (1992). *Numerical Methods for Conservation Laws*. Birkhäuser Verlag, Basel, second edition.

Li, J. (2013). Continuous improvement at Toyota manufacturing plant: Applications of production systems engineering methods. *International Journal of Production Research*, 51(23-24):7235–7249.

Li, J. and Meerkov, S. M. (2009). *Production Systems Engineering*. Springer, New York.

Liberopoulos, G. and Tsarouhas, P. (2002). Systems analysis speeds up Chipita's food-processing line. *Interfaces*, 32(3):62–76.

Liu, L., Liu, X., and Yao, D. D. (2004). Analysis and Optimization of a Multistage Inventory-Queue System. *Management Science*, 50(3):365–380.

Liu, Y. and Wein, L. M. (2008). A Queueing Analysis to Determine How Many Additional Beds Are Needed for the Detention and Removal of Illegal Aliens. *Management Science*, 54(1):1–15.

Liu, Y. and Whitt, W. (2011). Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. *Queueing Systems*, 67(2):145–182.

Liu, Y. and Whitt, W. (2012a). Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. *Operations Research*, 60(6):1551–1564.

Liu, Y. and Whitt, W. (2012b). The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems*, 71(4):405–444.

Liu, Y. and Whitt, W. (2014). Many-server heavy-traffic limit for queues with time-varying parameters. *The Annals of Applied Probability*, 24(1):378–421.

Lovell, D. J., Vlachou, K., Rabbani, T., and Bayen, A. (2013). A diffusion approximation to a single airport queue. *Transportation Research Part C: Emerging Technologies*, 33:227–237.

Luchak, G. (1956). The Solution of the Single-Channel Queuing Equations Characterized by a Time-Dependent Poisson-Distributed Arrival Rate and a General Class of Holding Times. *Operations Research*, 4(6):711–732.

Luchak, G. (1957). The Distribution of the Time Required to Reduce to Some Preassigned Level a Single-Channel Queue Characterized by a Time-Dependent Poisson-Distributed Arrival Rate and a General Class of Holding Times. *Operations Research*, 5(2):205–209.

Lyubarskii, G. Y. (1982). Busy time of a nonstationary single-channel service system and related questions. *Automation and Remote Control*, 43(12):1537–1543.

Mandelbaum, A. and Massey, W. A. (1995). Strong Approximations for Time-Dependent Queues. *Mathematics of Operations Research*, 20(1):33–64.

Mandelbaum, A., Massey, W. A., and Reiman, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems*, 30(1-2):149–201.

Mandelbaum, A., Massey, W. A., Reiman, M. I., Stolyar, A., and Rider, B. (2002). Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials. *Telecommunication Systems*, 21(2-4):149–171.

Manohar, P., Ram, S. S., and Manjunath, D. (2009). Path Coverage by a Sensor Field: The Nonhomogeneous Case. *ACM Transactions on Sensor Networks*, 5(2):1–26.

Margolius, B. H. (1999). A sample path analysis of the $M_t/M_t/c$ queue. *Queueing Systems*, 31(1-2):59–93.

Margolius, B. H. (2005). Transient Solution to the Time-Dependent Multiserver Poisson Queue. *Journal of Applied Probability*, 42(3):766–777.

Margolius, B. H. (2007). Transient and periodic solution to the time-inhomogeneous quasi-birth death process. *Queueing Systems*, 56(3-4):183–194.

Margolius, B. H. (2008). The matrices R and G of matrix analytic methods and the time-inhomogeneous periodic Quasi-Birth-and-Death process. *Queueing Systems*, 60(1-2):131–151.

Massey, W. A. (1985). Asymptotic analysis of the time dependent $M/M/1$ queue. *Mathematics of Operations Research*, 10(2):305–327.

Massey, W. A. (2002). The Analysis of Queues with Time-Varying Rates for Telecommunication Models. *Telecommunication Systems*, 21(2-4):173–204.

Massey, W. A. and Pender, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75(2-4):243–277.

Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, 13(1-3):183–250.

Massey, W. A. and Whitt, W. (1997). Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems*, 25(1-4):157–172.

Massim, Y., Yalaoui, F., Amodeo, L., Chatelet, E., and Zeblah, A. (2010). Efficient combined immune-decomposition algorithm for optimal buffer allocation in production lines for throughput and profit maximization. *Computers and Operations Research*, 37(4):611–620.

Matta, A. (2008). Simulation optimization with mathematical programming representation of discrete event systems. In *Proceedings of the 2008 Winter Simulation Conference*, pages 1393–1400, Miami, FL, USA.

Matta, A., Pezzoni, M., and Semeraro, Q. (2012). A kriging-based algorithm to optimize production systems approximated by analytical models. *Journal of Intelligent Manufacturing*, 23(3):587–597.

McCalla, C. and Whitt, W. (2002). A Time-Dependent Queueing-Network Model to Describe the Life-Cycle Dynamics of Private-Line Telecommunication Services. *Telecommunication Systems*, 19(1):9–38.

Meerkov, S. M. and Zhang, L. (2008). Transient behavior of serial production lines with Bernoulli machines. *IIE Transactions*, 40(3):297–312.

Mejía-Téllez, J. and Worthington, D. (1994). Practical methods for queue length behaviour for bulk service queues of the form $M/G^{0,C}/1$ and $M(t)/G^{0,C}/1$. *European Journal of Operational Research*, 73(1):103–113.

Minh, D. L. (1978). The Discrete-Time Single-Server Queue with Time-Inhomogeneous Compound Poisson Input and General Service Time Distribution. *Journal of Applied Probability*, 15(3):590–601.

Mok, S. K. and Shanthikumar, J. G. (1987). A transient queueing model for Business Office with standby servers. *European Journal of Operational Research*, 28(2):158–174.

Monden, Y. (1983). *Toyota production system: practical approach to production management*. Engineering & Management Press, Atlanta.

Moore, S. C. (1975). Approximating the Behavior of Nonstationary Single-Server Queues. *Operations Research*, 23(5):1011–1032.

Mourani, I., Hennequin, S., and Xie, X. (2007). Failure models and throughput rate of transfer lines. *International Journal of Production Research*, 45(8):1835–1859.

Nahas, N., Ait-Kadi, D., and Nourelfath, M. (2006). A new approach for buffer allocation in unreliable production lines. *International Journal of Production Economics*, 103(2):873–881.

Nasr, W. W. and Taaffe, M. R. (2013). Fitting the $Ph_t/M_t/s/c$ Time-Dependent Departure Process for Use in Tandem Queueing Networks. *INFORMS Journal on Computing*, 25(4):758–773.

Nelson, B. L. and Taaffe, M. R. (2004). The $Ph_t/Ph_t/\infty$ Queueing System: Part I-The Single Node. *INFORMS Journal on Computing*, 16(3):266–274.

Newell, G. F. (1966). The $M/G/\infty$ Queue. *SIAM Journal on Applied Mathematics*, 14(1):86–88.

Newell, G. F. (1968a). Queues with Time-Dependent Arrival Rates I - The Transition through Saturation. *Journal of Applied Probability*, 5(2):436–451.

Newell, G. F. (1968b). Queues with Time-Dependent Arrival Rates: II - The Maximum Queue and the Return to Equilibrium. *Journal of Applied Probability*, 5(3):579–590.

Newell, G. F. (1968c). Queues with Time-Dependent Arrival Rates III - A Mild Rush Hour. *Journal of Applied Probability*, 5(3):591–606.

Newell, G. F. (1971). *Applications of Queueing Theory*. Chapman and Hall Ltd., London.

Newell, G. F. (1979). Airport Capacity and Delays. *Transportation Science*, 13(3):201–241.

Ng, A. H., Bernedixen, J., and Syberfeldt, A. (2012). A comparative study of production control mechanisms using simulation-based multi-objective optimisation. *International Journal of Production Research*, 50(2):359–377.

Nozari, A. (1985). Control of Entry to a Nonstationary Queuing System. *Naval Research Logistics Quarterly*, 32(2):275–286.

Omosigho, S. E. and Worthington, D. J. (1985). The single server queue with inhomogeneous arrival rate and discrete service time distribution. *European Journal of Operational Research*, 22(3):397–407.

Omosigho, S. E. and Worthington, D. J. (1988). An approximation of known accuracy for single server queues with inhomogeneous arrival rate and continuous service time distribution. *European Journal of Operational Research*, 33(3):304–313.

Ong, K. L. and Taaffe, M. R. (1988). Approximating nonstationary $Ph(t)/Ph(t)/1/c$ queueing systems. *Mathematics and Computers in Simulation*, 30(5):441–452.

Palm, C. (1943). *Intensitätsschwankungen im Fernsprechverkehr - Untersuchungen über die Darstellung auf Fernsprechverkehrsprobleme anwendbarer stochastischer Prozesse*. Ericsson Technics, Stockholm.

Pang, G. and Whitt, W. (2012a). Infinite-server queues with batch arrivals and dependent service times. *Probability in the Engineering and Informational Sciences*, 26(2):197–220.

Pang, G. and Whitt, W. (2012b). The Impact of Dependent Service Times on Large-Scale Service Systems. *Manufacturing & Service Operations Management*, 14(2):262–278.

Papadopoulos, H. T. and Vidalis, M. I. (2001a). A heuristic algorithm for the buffer allocation in unreliable unbalanced production lines. *Computers and Industrial Engineering*, 41(3):261–277.

Papadopoulos, H. T. and Vidalis, M. I. (2001b). Minimizing WIP inventory in reliable production lines. *International Journal of Production Economics*, 70(2):185–197.

Parlar, M. (1984). Optimal dynamic service rate control in time dependent $M/M/S/N$ queues. *International Journal of Systems Science*, 15(1):107–118.

Parthasarathy, P. R. and Sudhesh, R. (2006). An exact solution for an $M/M/1$ queue with piecewise-constant rates. *The Mathematical Scientist*, 31(1):48–52.

Paullin, R. L. and Horonjeff, R. (1969). Sizing of departure lounges in airport buildings. *Transportation Engineering Journal of ASCE*, 95(2):267–277.

Pender, J. (2014a). A Poisson-Charlier approximation for nonstationary queues. *Operations Research Letters*, 42(4):293–298.

Pender, J. (2014b). Gram charlier expansion for time varying multiserver queues with abandonment. *SIAM Journal on Applied Mathematics*, 74(4):1238–1265.

Powell, W. B. and Simão, H. P. (1986). Numerical simulation of transient bulk queues with general vehicle dispatching strategies. *Transportation Research Part B: Methodological*, 20(6):477–490.

Purdue, P. (1974a). Stochastic theory of compartments. *Bulletin of Mathematical Biology*, 36(3):305–309.

Purdue, P. (1974b). Stochastic theory of compartments: One and two compartment systems. *Bulletin of Mathematical Biology*, 36(5-6):577–587.

Ramakrishnan, C. S. (1980). A note on the $M/D/\infty$ queue. *OpSearch*, 17(2&3):118.

Rider, K. L. (1976). A Simple Approximation to the Average Queue Size in the Time-Dependent $M/M/1$ Queue. *Journal of the ACM*, 23(2):361–367.

Ridley, A. D., Massey, W., and Fu, M. (2004). Fluid Approximation of a Priority Call Center With Time-Varying Arrivals. *Telecommunications Review*, 15:69–77.

Rosenlund, S. I. (1976). Busy Periods in Time-Dependent $M/G/1$ Queues. *Advances in Applied Probability*, 8(1):195–208.

Rothkopf, M. H. and Johnston, R. G. (1982). Routine Analysis of Periodic Queues. *IIE Transactions*, 14(3):214–218.

Rothkopf, M. H. and Oren, S. S. (1979). A Closure Approximation for the Nonstationary $M/M/s$ Queue. *Management Science*, 25(6):522–534.

Sabuncuoglu, I., Erel, E., and Gocgun, Y. (2006). Analysis of serial production lines: Characterisation study and a new heuristic procedure for optimal buffer allocation. *International Journal of Production Research*, 44(13):2499–2523.

Savsar, M. (2006). Buffer allocation in serial production lines with preventive and corrective maintenance operations. *Kuwait Journal of Science and Engineering*, 33(2):253–266.

Schneider, H. (1981). Effect of service-levels on order-points or order-levels in inventory models. *International Journal of Production Research*, 19(6):615–631.

Schwarz, J. A., Selinka, G., and Stolletz, R. (2016). Performance analysis of time-dependent queueing systems: survey and classification. *Omega,*, (In Press):DOI: 10.1016/j.omega.2015.10.013, 1–20.

Seki, Y. and Hoshino, N. (1999). Transient behavior of a single-stage kanban system based on the queueing model Yoichi. *International Journal of Production Economics*, 60-61:369–374.

Selinka, G., Franz, A., and Stolletz, R. (2016). Time-dependent Performance Approximation of Truck Handling Operations at an Air Cargo Terminal. *Computers & Operations Research*, 65:164–173.

Shanbhag, D. N. (1966). On Infinite Server Queues with Batch Arrivals. *Journal of Applied Probability*, 3(1):274–279.

Shang, K. H. (2012). Single-Stage Approximations for Optimal Policies in Serial Inventory Systems with Nonstationary Demand. *Manufacturing & Service Operations Management*, 14(3):414–422.

Sharma, O. P. and Gupta, U. C. (1983). $M/M/\infty$ Queues in series with non-homogeneous inputs. *Mathematische Operationsforschung und Statistik. Series Optimization*, 14(3):445–453.

Shi, C. and Gershwin, S. B. (2009). An efficient buffer design algorithm for production line profit maximization. *International Journal of Production Economics*, 122(2):725–740.

Shi, C. and Gershwin, S. B. (2014). A segmentation approach for solving buffer allocation problems in large production systems. *International Journal of Production Research*, (In Press):1–21.

Shi, L. and Men, S. (2003). Optimal buffer allocation in production lines. *IIE Transactions*, 35(1):1–10.

Singer, M. and Donoso, P. (2008). Assessing an ambulance service with queuing theory. *Computers & Operations Research*, 35(8):2549–2560.

Smith, J. M. and Cruz, F. R. B. (2005). The buffer allocation problem for general finite buffer queueing networks. *IIE Transactions*, 37(4):343–365.

So, K. C. (1997). Optimal buffer allocation strategy for minimizing work-in-process inventory in unpaced production lines. *IIE Transactions*, 29(1):81–88.

Spearman, M. L. (1992). Customer Service in Pull Production Systems. *Operations Research*, 40(5):948–958.

Stadje, W. (1990). A note on the simple queue with variable intensities and two servers. *Operations Research Letters*, 9(1):45–49.

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

Stolletz, R. (2008a). Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research*, 190(2):478–493.

Stolletz, R. (2008b). Non-stationary delay analysis of runway systems. *OR Spectrum*, 30(1):191–213.

Stolletz, R. (2011). Analysis of passenger queues at airport terminals. *Research in Transportation Business & Management*, 1(1):144–149.

Stolletz, R. and Lagershausen, S. (2013). Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions. *International Journal of Production Research*, 51(5):1366–1378.

Swaroop, P., Zou, B., Ball, M. O., and Hansen, M. (2012). Do more US airports need slot controls? A welfare based approach to determine slot levels. *Transportation Research Part B: Methodological*, 46(9):1239–1259.

Sze, D. Y. (1984). A Queueing Model for Telephone Operator Staffing. *Operations Research*, 32(2):229–249.

Taaffe, M. R. and Clark, G. M. (1988). Approximating Nonstationary Two-Priority Non-Preemptive Queueing Systems. *Naval Research Logistics*, 35(1):125–145.

Taaffe, M. R. and Ong, K. L. (1987). Approximating nonstationary $Ph(t)/M(t)/s/c$ queueing systems. *Annals of Operations Research*, 8(1):103–116.

Takahashi, K. (2003). Comparing reactive Kanban systems. *International Journal of Production Research*, 41(18):4317–4337.

Takahashi, K., Morikawa, K., and Nakamura, N. (2004). Reactive JIT ordering system for changes in the mean and variance of demand. *International Journal of Production Economics*, 92(2):181–196.

Takahashi, K. and Nakamura, N. (2002). Decentralized reactive Kanban system. *European Journal of Operational Research*, 139(2):262–276.

Tan, B. (2015). Mathematical programming representations of the dynamics of continuous-flow production systems. *IIE Transactions*, 47(2):173–189.

Tan, X., Knessl, C., and Yang, Y. P. (2013). On finite capacity queues with time dependent arrival rates. *Stochastic Processes and their Applications*, 123(6):2175–2227.

Tarabia, A. M. K. (2000). Transient Analysis of $M/M/1/N$ Queue - An Alternative Approach. *Tamkang Journal of Science and Engineering*, 3(4):263–266.

Tardif, V. and Maaseidvaag, L. (2001). An adaptive approach to controlling kanban systems. *European Journal of Operational Research*, 132(2):411–424.

Tempelmeier, H. (2003). Practical considerations in the optimization of flow production systems. *International Journal of Production Research*, 41(1):149–170.

Terwiesch, C. and Bohn, R. E. (2001). Learning and process improvement during production ramp-up. *International Journal of Production Economics*, 70(1):1–19.

Thakur, A. K. and Rescigno, A. (1978). On the stochastic theory of compartments: III. General time-dependent reversible systems. *Bulletin of Mathematical Biology*, 40(2):237–246.

Thakur, A. K., Rescigno, A., and Schafer, D. E. (1972). On the stochastic theory of compartments: I. A single-compartment system. *The Bulletin of Mathematical Biophysics*, 34(1):53–63.

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

Tipper, D. and Sundareshan, M. K. (1990). Numerical Methods for Modeling Computer Networks Under Nonstationary Conditions. *IEEE Journal on Selected Areas in Communications*, 8(9):1682–1695.

Tošić, V. (1992). A review of airport passenger terminal operations analysis and modelling. *Transportation Research Part A: Policy and Practice*, 26(1):3–26.

Tripathi, S. K. and Duda, A. (1986). Time-dependent analysis of queueing systems. *INFOR*, 24(3):199–220.

Upton, R. A. and Tripathi, S. K. (1982). An Approximate Transient Analysis of the $M(t)/M/1$ Queue. *Performance Evaluation*, 2(2):118–132.

Van As, H. R. (1986). Transient Analysis of Markovian Queueing Systems and Its Application to Congestion-Control Modeling. *IEEE Journal on Selected Areas in Communications*, 4(6):891–904.

Van de Coevering, M. C. T. (1995). Computing transient performance measures for the $M/M/1$ queue. *OR Spektrum*, 17:19–22.

Van Dijk, N. M. (1992). Uniformization for nonhomogeneous Markov chains. *Operations Research Letters*, 12(5):283–291.

Vanberkel, P. T., Boucherie, R. J., Hans, E. W., and Hurink, J. L. (2014). Optimizing the strategic patient mix combining queueing theory and dynamic programming. *Computers & Operations Research*, 43:271–279.

Vandergraft, J. S. (1983). A Fluid Flow Model of Networks of Queues. *Management Science*, 29(10):1198–1208.

Viti, F. and van Zuylen, H. J. (2009). The Dynamics and the Uncertainty of Queues at Fixed and Actuated Controls: A Probabilistic Approach. *Journal of Intelligent Transportation Systems*, 13(1):39–51.

Viti, F. and van Zuylen, H. J. (2010). Probabilistic models for queues at fixed control signals. *Transportation Research Part B: Methodological*, 44(1):120–135.

Vouros, G. A. and Papadopoulos, H. T. (1998). Buffer allocation in unreliable production lines using a knowledge based system. *Computers and Operations Research*, 25(12):1055–1067.

Wall, A. D. and Worthington, D. J. (2007). Time-dependent analysis of virtual waiting time behaviour in discrete time queues. *European Journal of Operational Research*, 178(2):482–499.

Wang, W.-P., Tipper, D., and Banerjee, S. (1996). A Simple Approximation for Modeling Nonstationary Queues. In *Proceedings of IEEE INFOCOM'96. Conference on Computer Communications*, pages 255–262.

Weiss, S., Schwarz, J. A., and Stolletz, R. (2015). Buffer Allocation Problems for stochastic flow lines with unreliable machines. In *Proceedings of the 10th Conference on Stochastic Models of Manufacturing and Service Operations*, pages 271–277, Volos, Greece.

Weiss, S. and Stolletz, R. (2015). Buffer allocation in stochastic flow lines via sample-based optimization with initial bounds. *OR Spectrum*, 37(4):869–902.

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.

Whitt, W. (1999). Using different response-time requirements to smooth time-varying demand for service. *Operations Research Letters*, 24(1-2):1–10.

Whitt, W. (2007). What You Should Know About Queueing Models to Set Staffing Requirements in Service Systems. *Naval Research Logistics*, 54(5):476–484.

Whitt, W. (2013). OM Forum - Offered Load Analysis for Staffing. *Manufacturing & Service Operations Management*, 15(2):166–169.

Wirasinghe, S. C. and Bandara, S. (1990). Airport gate position estimation for minimum total costs -Approximate closed form solution. *Transportation Research Part B: Methodological*, 24(4):287–297.

Wirasinghe, S. C. and Shehata, M. (1988). Departure lounge sizing and optimal seating capacity for a given aircraft/flight mix - (i) single gate. (ii) several gates. *Transportation Planning and Technology*, 13(1):57–71.

Worthington, D. and Wall, A. (1999). Using the discrete time modelling approach to evaluate the time-dependent behaviour of queueing systems. *Journal of the Operational Research Society*, 50(8):777–788.

Wragg, A. (1963). The solution of an infinite set of differential-difference equations occurring in polymerization and queueing problems. *Mathematical Proceedings of the Cambridge Philosophical Society*, 59(01):117–124.

Wu, K. (2014). Classification of queueing models for a workstation with interruptions: a review. *International Journal of Production Research*, 52(3):902–917.

Wu, K., McGinnis, L., and Zwart, B. (2011). Queueing models for a single machine subject to multiple types of interruptions. *IIE Transactions*, 43(10):753–759.

Xanthopoulos, A. S. and Koulouriotis, D. E. (2014). Multi-objective optimization of production control mechanisms for multi-stage serial manufacturing-inventory systems. *The International Journal of Advanced Manufacturing Technology*, 74(9-12):1507–1519.

Xu, K., Tipper, D., Qian, Y., Krishnamurthy, P., and Tipmongkonsilp, S. (2014). Time-Varying Performance Analysis of Multihop Wireless Networks with CBR Traffic. *IEEE Transactions on Vehicular Technology*, 63(7):3397–3409.

Yang, Y. and Knessl, C. (1997). Asymptotic analysis of the $M/G/1$ queue with a time-dependent arrival rate. *Queueing Systems*, 26(1-2):23–68.

Yang, Z.-Z., Chen, G., and Song, D.-P. (2013). Integrating truck arrival management into tactical operation planning at container terminals. *Polish Maritime Research*, 20(Special Issue):32–46.

Yin, G. and Zhang, H. (2002). Countable-State-Space Markov Chains with Two Time Scales and Applications to Queueing Systems. *Advances in Applied Probability*, 34(3):662–688.

Yom-Tov, G. B. and Mandelbaum, A. (2014). Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing. *Manufacturing & Service Operations Management*, 16(2):283–299.

Zhang, J. and Coyle, E. J. (1991). The Transient Solution of Time-Dependent $M/M/1$ Queues. *IEEE Transactions on Information Theory*, 37(6):1690–1696.

Zhang, L., Wang, C., Arinez, J., and Biller, S. (2013). Transient analysis of Bernoulli serial lines: performance evaluation and system-theoretic properties. *IIE Transactions*, 45(5):528–543.

Zhang, Z. G. (2009). Performance Analysis of a Queue with Congestion-Based Staffing Policy. *Management Science*, 55(2):240–251.

Appendix A

A flow line as described in Section 6.3.1 with $M = 1$, Poisson demand, exponentially distributed processing times, and a constant buffer capacity under steady-state conditions can be modeled as a birth and death process with discrete and infinite state space $\{0, 1, 2, ..., n, ...\}$. A state n is defined as the number of unused buffer spaces, plus 1 if station M is not blocked, plus the number of backlogged customer orders. A transition to a greater state occurs with rate λ_c and a transition to a smaller state with rate μ_M. The steady-state probability P_n of state n is given by

$$P_n = (1 - \rho)\rho^n, \qquad \forall n \geq 0, \tag{A.1}$$

with $\rho = \lambda_c/\mu_M$ and stability condition $\rho < 1$ (Gross et al., 2008, p.59). Thus, in the following we focus on the relevant range of $0 < \rho < 1$.

Proof of Theorem 6.4.1:
The expected WIP is given by

$$\begin{aligned}
\mathrm{E}[W(B_M)] &= (1 - P_0) + \sum_{n=0}^{B_M}((B_M + 1) - n) \cdot P_n \\
&= (1 - (1 - \rho)) + \sum_{n=0}^{B_M}((B_M + 1) - n)(1 - \rho)\rho^n \\
&= \rho + (1 - \rho)((B_M + 1)\sum_{n=0}^{B_M}\rho^n - \sum_{n=0}^{B_M} n\rho^n) \\
&= \rho + \frac{\rho(B_M + 2) - \rho^{B_M+2} - B_M - 1}{(\rho - 1)}.
\end{aligned} \tag{A.2}$$

Note that if the machine is not blocked, it is processing a workpiece. This part of the WIP is covered by the first summand of (A.2). The above simplifications are based on closed form representations of geometric series.

We find the derivative with respect to B_M assuming B_M to be continuous. However, the obtained properties also hold for B_M taking only positive integer values. The first derivative with respect to B_M is given by

$$\frac{\partial \mathrm{E}[W(B_M)]}{\partial B_M} = \frac{\rho - 1 - \ln(\rho)\rho^{B_M+2}}{\rho - 1} = 1 + \frac{\ln(\rho)\rho^{B_M+2}}{1-\rho}. \tag{A.3}$$

For (A.2) to be strictly increasing it is sufficient to show

$$\frac{\ln(\rho)\rho^{B_M+2}}{1-\rho} > -1 \Leftrightarrow \ln(\rho)\rho^{B_M+2} > \rho - 1 \Leftrightarrow \ln(\rho)\rho^{B_M+2} - \rho > -1. \tag{A.4}$$

The left hand side of Inequality (A.4) is increasing in B_M. Hence, we consider the case of $B_M = 0$. For this case we analyze the left hand side by building the following first and second derivatives with respect to ρ

$$\frac{\partial}{\partial \rho} \ln(\rho)\rho^2 - \rho = \rho + 2\rho\ln(\rho) - 1, \tag{A.5}$$

$$\frac{\partial^2}{\partial \rho^2} \ln(\rho)\rho^2 - \rho = 2\ln(\rho) + 3. \tag{A.6}$$

The necessary and sufficient conditions for a global minimum are fulfilled by $\rho = 1$ with value -1. It follows that (A.3)> 0, for $0 < \rho < 1$.

The second derivative of $\mathrm{E}[W(B_M)]$ is always positive for $0 < \rho < 1$

$$\frac{\partial^2 \mathrm{E}[W(B_M)]}{\partial B_M^2} = \frac{\ln^2(\rho)\rho^{B_M+2}}{1-\rho} > 0. \tag{A.7}$$

Hence, $\mathrm{E}[W(B_M)]$ is strictly convex and increasing in B_M.□

Note that $\mathrm{E}[W(B_M)]$ converges for large B_M to a linear increase with gradient 1 as

$$\lim_{B_M \to \infty} \frac{\partial \mathrm{E}[W(B_M)]}{\partial B_M} = \lim_{B_M \to \infty} 1 + \frac{\ln(\rho)\rho^{B_M+2}}{1-\rho} = 1, \tag{A.8}$$

and

$$\lim_{B_M \to \infty} \frac{\partial^2 \mathrm{E}[W(B_M)]}{\partial B_M^2} = \lim_{B_M \to \infty} \frac{\ln^2(\rho)\rho^{B_M+2}}{1-\rho} = 0. \tag{A.9}$$

Proof of Theorem 6.4.2:
Tardif and Maaseidvaag (2001) use intuitive arguments and Little's law to show that the backlog decreases in the number of Kanbans. Here, we provide a closed-form solution and use it to rigourously establish first and second order properties. From the steady-state distribution we obtain

$$\begin{aligned}
\mathrm{E}[W^-(B_M)] &= \sum_{n=B_M+1}^{\infty} (n-(B_M+1))P_n \\
&= (1-\rho)\left(\sum_{n=B_M+1}^{\infty} n\rho^n - (B_M+1)\sum_{n=B_M+1}^{\infty}\rho^n\right) \\
&= (1-\rho)\left(\frac{\rho^{B_M+1}(-(B_M+1)\rho+\rho+(B_M+1))}{(1-\rho)^2}\right. \\
&\left. - \frac{(B_M+1)\rho^{B_M+1}}{(1-\rho)}\right) \qquad \text{(A.10)} \\
&= \frac{\rho^{B_M+2}}{(1-\rho)}. \qquad \text{(A.11)}
\end{aligned}$$

Again, we build the derivative with respect to B_M while assuming B_M to be continuous. For $0<\rho<1$ we find

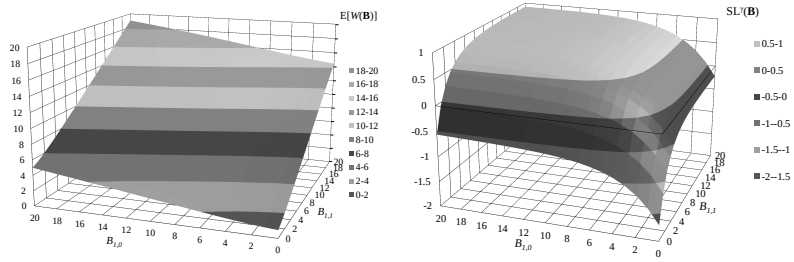
$$\frac{\partial \mathrm{E}[W^-(B_M)]}{\partial B_M} = \frac{\ln(\rho)\rho^{B_M+2}}{1-\rho} < 0, \qquad \text{(A.12)}$$

$$\frac{\partial^2 \mathrm{E}[W^-(B_M)]}{\partial B_M^2} = \frac{\ln^2(\rho)\rho^{B_M+2}}{1-\rho} > 0. \qquad \text{(A.13)}$$

Hence, $\mathrm{E}[W^-(B_M)]$ is strictly decreasing convex. In steady state the γ-service level is given by $SL^{\gamma}(B_M) = 1 - \frac{\mathrm{E}[W^-(B_M)]}{\lambda}$. From the result for the expected backlog it follows that the γ-service level is strictly increasing and concave in B_M.□

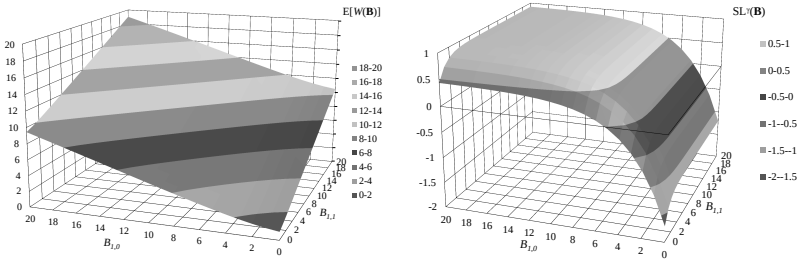
Appendix B

Results for the expected average WIP and γ-service level for a system with $T = 1920$, $M = 1$, exponentially distributed processing times with rate $\mu_1(t) = 2/3,\ t \in [0; 960)$, $\mu_1(t) = 1,\ t \in [960; 1920)$, Poisson demand process with rate $\lambda_c(t) = 0.5,\ t \in [0, 1920]$, $I = 1$ and varying t_1^*.



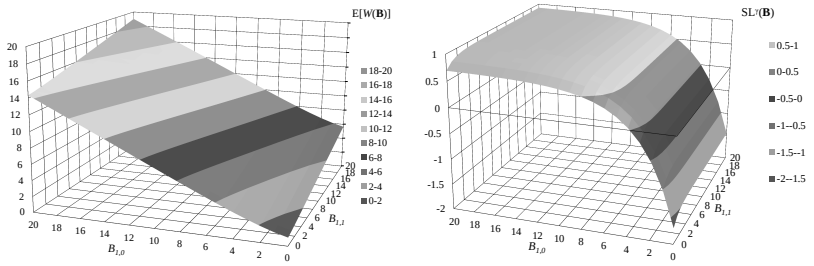
(a) Expected average WIP (b) γ-service level

Figure B.1: Expected average WIP and γ-service level for $t_i^* = 480$



(a) Expected average WIP (b) γ-service level

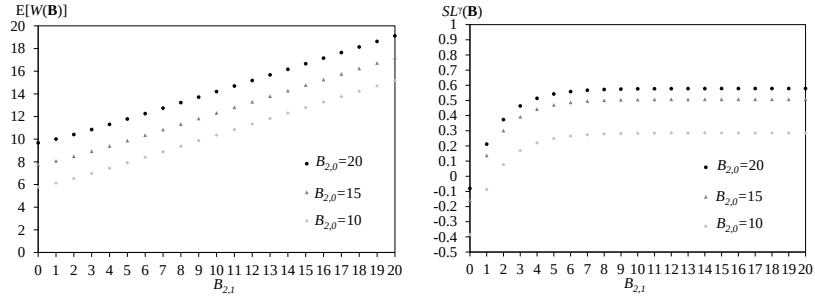
Figure B.2: Expected average WIP and γ-service level for $t_i^* = 960$



(a) Expected average WIP (b) γ-service level

Figure B.3: Expected average WIP and γ-service level for $t_i^* = 1440$

Results for the expected average WIP and γ-service level for a system with $T = 1920, M = 2$, a change in the processing distribution of $m = 1$ from an exponential distribution with rate $\mu_1(t) = 2/3, t \in [0, 960)$, to an Erlang-k distribution with $\mu_1(t) = 1, t \in [960, 1920]$ and $cv_1^2(t) = 0.5, t \in [960, 1920]$. Exponential processing distribution of station $m = 2$ with rate $\mu_m(t) = 1, t \in [0, 1920], m > 1$, Poisson demand process with rate $\lambda_c(t) = 0.5,\ t \in [0, 1920], I = 1$, and $t_1^* = 960$.



(a) Expected average WIP (b) γ-service level

Figure B.4: Expected average WIP and γ-service level for $M = 2$, $I = 1$, $t_i^* = 960$, $B_{1,0} = B_{1,1} = 1$

Curriculum Vitae

Justus Arne Schwarz

Personal information

Place of birth:	Hamburg, Germany
Nationality:	German

Professional experience

03/2012 - 12/2015	Research Assistant, Chair of Production Management, University of Mannheim, Germany

Education

03/2012 - 12/2015	Doctoral candidate, Business School, University of Mannheim, Germany
10/2006 - 12/2011	Industrial Engineering and Management (Diploma), Karlsruhe Institute of Technology, Germany
09/2009 - 07/2010	Engineering and Technology Management (Master of Science), Portland State University, USA
06/2006	Abitur, Gymnasium Oberalster, Hamburg, Germany