Decomposition of discrete-time open tandem queues with Poisson arrivals and general service times

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN
(Dr.-Ing.)

von der KIT-Fakultät für Maschinenbau des
Karlsruher Instituts für Technologie (KIT)

angenommene

DISSERTATION

von

M.Sc. Christoph Jacobi

Tag der mündlichen Prüfung:	14.11.2023
Hauptreferent:	Prof. Dr.-Ing. Kai Furmans
Korreferent:	Prof. Dr. Barış Tan

Kurzfassung

In der Grobplanungsphase vernetzter Logistik- und Produktionssysteme ist man häufig daran interessiert, mit geringem Berechnungsaufwand eine zufriedenstellende Approximation der Leistungskennzahlen des Systems zu bestimmen. Hierbei bietet die Modellierung mittels zeitdiskreter Methoden gegenüber der zeitkontinuierlichen Modellierung den Vorteil, dass die gesamte Wahrscheinlichkeitsverteilung der Leistungskenngrößen berechnet werden kann. Da Produktions- und Logistiksysteme in der Regel so konzipiert sind, dass sie die Leistung nicht im Durchschnitt, sondern mit einer bestimmten Wahrscheinlichkeit (z.B. 95%) zusichern, können zeitdiskrete Warteschlangenmodelle detailliertere Informationen über die Leistung des Systems (wie z.B. der Warte- oder Durchlaufzeit) liefern.

Für die Analyse vernetzter zeitdiskreter Bediensysteme sind Dekompositionsmethoden häufig der einzig praktikable und recheneffiziente Ansatz, um stationäre Leistungsmaße in den einzelnen Bediensystemen zu berechnen. Hierbei wird das Netzwerk in die einzelnen Knoten zerlegt und diese getrennt voneinander analysiert. Der Ansatz basiert auf der Annahme, dass der Punktprozess des Abgangsstroms stromaufwärts liegender Stationen durch einen Erneuerungsprozess approximiert werden kann, und so eine unabhängige Analyse der Bediensysteme möglich ist. Die Annahme der Unabhängigkeit ermöglicht zwar eine effiziente Berechnung, führt jedoch zu teilweise starken Approximationsfehlern in den berechneten Leistungskenngrößen.

Der Untersuchungsgegenstand dieser Arbeit sind offene zeitdiskrete Tandem-Netzwerke mit Poisson-verteilten Ankünften am stromaufwärts liegenden Bediensystem und generell verteilten Bedienzeiten. Das Netzwerk besteht folglich aus einem stromaufwärts liegenden $M/G/1$-Bediensystem und einem stromabwärts liegenden $G/G/1$-System. Diese Arbeit verfolgt drei Ziele, (1) die Defizite des Dekompositionsansatzes aufzuzeigen und dessen Approximationsgüte mittels statistischer Schätzmethoden zu bestimmen, (2) die Autokorrelation des Abgangsprozesses des $M/G/1$-Systems zu modellieren um die Ursache des Approximationsfehlers erklären zu können und (3) einen Dekompositionsansatz zu entwickeln, der die Abhängigkeit des Abgangsstroms berücksichtigt und so beliebig genaue Annäherungen der Leistungskenngrößen ermöglicht.

Im ersten Teil der Arbeit wird die Approximationsgüte des Dekompositionsverfahrens am stromabwärts liegenden $G/G/1$-Bediensystem mit Hilfe von Linearer Regression (Punktschätzung) und Quantilsregression (Intervallschätzung) bestimmt. Beide Schätzverfahren werden jeweils auf die relativen Fehler des Erwartungswerts und des 95%-Quantils der Wartezeit im Vergleich zu den simulierten Ergebnissen berechnet. Als signifikante Einflussfaktoren auf die Approximationsgüte werden die Auslastung des Systems und die Variabilität des Ankunftsstroms identifiziert.

Der zweite Teil der Arbeit fokussiert sich auf die Berechnung der Autokorrelation im Abgangsstroms des $M/G/1$-Bediensystems. Aufeinanderfolgende Zwischenabgangszeiten sind miteinander korreliert, da die Abgangszeit eines Kunden von dem Systemzustand abhängt, den der vorherige Kunde bei dessen Abgang zurückgelassen hat. Die Autokorrelation ist ursächlich für den Dekompositionsfehler, da die Ankunftszeiten am stromabwärts liegenden Bediensystem nicht unabhängig identisch verteilt sind.

Im dritten Teil der Arbeit wird ein neuer Dekompositionsansatz vorgestellt, der die Abhängigkeit im Abgangsstroms des $M/G/1$-Systems mittels eines semi-Markov Prozesses modelliert. Um eine explosionsartige Zunahme des Zustandsraums zu verhindern, wird ein Verfahren eingeführt, das den Zustandsraum der eingebetteten Markov-Kette beschränkt. Numerischen Auswertungen zeigen, dass die mit stark limitierten Zustandsraum erzielten Ergebnisse eine bessere Approximation bieten als der bisherige Dekompositionsansatz. Mit zunehmender Größe des Zustandsraums konvergieren die Leistungskennzahlen beliebig genau.

Abstract

During draft planning of interconnected logistics and production systems, decision makers are often interested in determining a satisfactory approximation of the system's key performance indicators with little computational effort. Employing discrete-time queueing models offers an advantage over continuous-time modelling in that the entire probability distribution of the performance parameters can be calculated. Since production and logistics systems are typically designed to guarantee performance not on average, but with a given probability (e.g. 95%), discrete-time queueing models can provide more detailed information about the system's performance, such as waiting or throughput time.

When it comes to analysing open queueing networks, decomposition methods often are the only feasible and computationally efficient approach to calculate steady-state performance measures in the individual discrete-time queues. The method decomposes the network into individual nodes, which are analysed in isolation. The approach is based on the assumption that a renewal process can approximate the point process of the departure stream of upstream stations, and thus an independent analysis of the queueing systems is possible. Although the assumption of independence is computationally attractive, performance results obtained with the renewal decomposition approach might be subject to severe approximation errors.

In this thesis, we study discrete-time open tandem networks with external Poisson arrivals and generally distributed service times. The external Poisson arrival stream makes the upstream queue of type $M/G/1$ while the downstream queue is of type $G/G/1$. This thesis pursues three goals, (1) to reveal the deficits of the renewal decomposition approach and to determine its approximation quality by means of statistical estimation methods, (2) to model the auto-correlation of the departure point process of the $M/G/1$-queue in order to explain the cause of the approximation error, and (3) to develop a decomposition approach that takes into account the dependence of the departure flow and thus allows for a converging accurate calculation of the performance parameters.

In the first part of the thesis, we analyse the approximation quality of the renewal decomposition method at the downstream queue based on linear regression (point estimation) and quantile regression (interval estimation). The dependent variables of the regression models are the relative error of the expected value and 95%-percentile of the waiting time compared to a simulation. Based on the ANOVA, we identify the system's utilisation and the arrival stream's variability as main drivers influencing the approximation quality.

The second part of the thesis focuses on the computation of the auto-correlation in the departure stream of the upstream $M/G/1$-queue. Consecutive inter-departure times are auto-correlated because the departure time of a customer depends on the system state left behind by the previous customer. Auto-correlation in the connecting stream is causal for the approximation error, as the arrival times at the downstream station are not independently identically distributed.

In the third part of the thesis, we present a novel decomposition approach that models the inter-dependency in the departure stream with a semi-Markov process. To prevent state-space explosion, we introduce a procedure to limit the state space of the embedded Markov chain. Our numerical results show that the performance measures obtained with a limited state space provide better approximations than the renewal decomposition approach. As the state space increases, the novel decomposition method converges arbitrarily accurate.

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Fördertechnik und Logistiksysteme (IFL) des Karlsruher Instituts für Technologie (KIT). Ich möchte allen, die mich auf dem Weg der Erstellung meiner Dissertation begleitet, unterstützt und gefördert haben, meinen herzlichen Dank aussprechen.

Kai Furmans danke ich für die Übernahme des Hauptreferats und für die Unterstützung und Förderung während meiner Zeit als wissenschaftlicher Mitarbeiter am IFL. Er hat mir in meiner Arbeit und meiner Forschung stets großes Vertrauen entgegengebracht und mir viele Freiheiten gewährt, von denen ich – nicht nur bei der Erstellung meiner Dissertation – sehr profitieren konnte. Barış Tan gilt mein Dank für die Übernahme des Korreferats und für seine wertvollen Anmerkungen zu meiner Arbeit.

Im Herbst 2022 habe ich einen dreimonatigen Forschungsaufenthalt an der Purdue University verbracht, während dessen wesentliche Teile dieser Arbeit entstanden sind. Ich danke George Shanthikumar, dass er mir diese wundervolle Erfahrung ermöglicht hat. Ich bin ihm ebenfalls dankbar für die fachlich anspruchsvollen Diskussionen, die wir in dieser Zeit zu meinem Forschungsgebiet geführt haben. Das Karlsruhe House of Young Scientists (KHYS) hat meinen Forschungsaufenthalt in den USA finanziell gefördert, wofür ich sehr dankbar bin.

Mein Dank gilt meinen aktiven und ehemaligen Kollegen am IFL, die eine angenehme Arbeitsatmosphäre geschaffen, und mich mit ihren fachlichen Diskussionen und Anmerkungen zu meiner Arbeit unterstützt haben. Besonders danken möchte ich Uta, dass sie sich so oft und lange Zeit genommen hat, intensiv mit mir über meine Arbeit zu diskutieren.

Meinen Eltern Steffi und Matthias danke ich herzlich für die bedingungslose Unterstützung und den Rückhalt, den ich während der Zeit meiner Dissertation erfahren habe. Susanne, Sven, Emil und Anton gebührt ein großer Dank dafür, dass sie mich jederzeit auf andere Gedanken bringen.

Mein letzter und größter Dank gilt Hannah, dafür, dass sie immer an mich glaubt, mir Mut zuspricht, Kraft gibt und immer für mich da ist.

Karlsruhe, Dezember 2023 Christoph Jacobi

Contents

Kurzfassung . . . . **i**

Abstract . . . . **iii**

Danksagung . . . . **v**

1 Introduction **1**

1.1 Research questions 3

1.2 Outline 5

2 Discrete-time stochastic modelling **7**

2.1 Discrete-time stochastic processes 8

2.2 Discrete-time point and renewal processes 9

2.3 Discrete-time Markov and semi-Markov processes 11

2.4 Advantages of discrete-time modelling 15

3 Literature review **19**

3.1 Diffusion approximation, mean value analysis, and exact methods 20

3.2 Decomposition methods 21

3.2.1 Transfer lines 22

3.2.2 Continuous-time open queueing networks 23

3.2.3 Discrete-time open queueing networks 27

3.3 Chapter conclusion 28

4 Decomposition error in discrete-time open tandem queues . **31**

4.1 Introduction 32

4.2 Theoretical background 34

4.3 Methodology . . . 35
4.3.1 Point and interval estimates . . . 38
4.3.2 Goodness of fit criteria and likelihood ratio tests . . . 39
4.3.3 Simulation model . . . 42
4.3.4 Design of experiments . . . 42
4.4 Results . . . 44
4.4.1 Study I: Expected value of waiting time . . . 44
4.4.2 Study II: 95% percentile of waiting time . . . 50
4.4.3 Performance of point and interval estimates . . . 50
4.5 Extensions: Bottlenecks and longer lines . . . 53
4.6 Main drivers of decomposition error . . . 55
4.7 Chapter conclusion . . . 57

5 On the output dynamics of the discrete-time M/G/1-queue . . . 59
5.1 Introduction and problem description . . . 60
5.2 Literature review . . . 62
5.3 Discrete-time Markov chain . . . 63
5.3.1 The stationary inter-departure time distribution . . . 67
5.3.2 The serial covariance of the departure process . . . 68
5.4 Numerical results . . . 71
5.5 The effect of auto-correlated arrivals on the analysis of downstream queues . . . 75
5.5.1 Waiting time with auto-correlated arrivals . . . 75
5.5.2 The effect of variability on the approximation quality of downstream queues . . . 77
5.6 Chapter conclusion . . . 78

6 A refined decomposition approach with converging accuracy for discrete-time open tandem queues . . . 81
6.1 Introduction and problem description . . . 82
6.2 Literature review . . . 84
6.3 The semi-Markov arrival decomposition approach . . . 85
6.3.1 The upstream $M/G/1$-queue with semi-Markov departures 86
6.3.2 State space limitation . . . 89
6.3.3 The downstream $SM/G/1$-queue . . . 93

6.4 Numerical results . . . 103
6.4.1 Verification . . . 103
6.4.2 Performance results . . . 105
6.4.3 Computational complexity . . . 108
6.5 Chapter conclusion . . . 110

7 Conclusion . . . 113
7.1 Summary . . . 113
7.2 Conclusion of the thesis . . . 117
7.3 Outlook . . . 117

A Regression analyses for decomposition error in tandem queues with bottlenecks . . . 119

B Proof for the state space reduction method . . . 125

C Continued performance results of the semi-Markov arrival decomposition approach . . . 131

Acronyms and symbols . . . 135

List of Figures . . . 139

List of Tables . . . 141

List of Publications . . . 143

Bibliography . . . 145

1 Introduction

In queueing theory, interconnected systems may be used to represent the stochastic behaviour of production systems and supply chains. For example, queuing network analysis allows us to compute the throughput time of a shipment or the number of buffer slots to be provided in front of a machine. Since production and logistics systems are typically designed to guarantee performance not on average, but with a given probability (e.g. 95%), logistics managers are often interested in the entire probability distribution of key performance indicators (such as waiting or throughput time). Having the entire probability distribution computed allows us to make more detailed statements about the performance of the system. For example, we can assess which percentage of orders are processed in 3h or less, or what promised throughput time will be met in 95% of the cases (Schleyer and Gue 2012). Applying discrete-time queueing models is appealing for the analysis of real manufacturing or logistics systems as the entire probability distributions of key performance indicators can be computed efficiently under very general assumptions. Particular research interest is given to the analysis of discrete-time queuing networks that may represent a supply chain or a production system, composed of several value-adding stages.

When it comes to analysing interconnected systems, a natural approach is to break the problem down into its component parts and study them in isolation. This *divide-and-conquer* approach has led to considerable breakthroughs in many scientific fields. For example, Gauss found that the sum over the first n integers can be broken down into the sum of smaller components, which are significantly easier to compute. Based on the observation that the sum of the first and the n-th element is the same as the sum of the second and the $(n-1)$-st element (and so on), he found the following formula:

$$1 + 2 + ... + n = \frac{n \cdot (n+1)}{2}. \tag{1.1}$$

In computer science, the quicksort (Hoare 1962) and the merge sort (invented by John von Neumann (Knuth 1998)) algorithms are famous examples of efficient sorting algorithms implementing the divide-and-conquer principle. In general, divide-and-conquer is an appealing design paradigm for computer algorithms as it allows for recurrence and parallelisation (Knuth 1998).

For the analysis of queueing networks, applying the divide-and-conquer concept seems naturally promising. The queueing network decomposition approach partitions the network into individual queues and analyses them in isolation. By being modular and scalable, this procedure is often the only feasible and computationally efficient approach to compute steady-state performance measures in a network of queues.

The major premise of the decomposition approach is that an independent analysis of the queues in the network is possible. By exploiting the memoryless property of the exponential distribution, Jackson (1957) demonstrated that the decomposition of a network of $M/M/1$-queues yields exact results. However, two simplifying assumptions must be made to transfer this idea to networks with generally distributed inter-arrival and service times (Govil and Fu 1999): First, we must assume that the individual queueing systems can be treated as statistically independent $GI/G/1$-queues. Second, we assume that a renewal process can approximate the point process which forms the input to each $GI/G/1$-queue. However, the renewal assumption of arrival point processes leads to the fact that the performance results computed with decomposition are approximate. While the existence of this approximation error is an undebated fact, its magnitude and influencing factors present an open research question in the scientific literature. It also remains of research interest to develop stochastic models that exploit point processes for analysing queueing networks with a comparable computational efficiency of the renewal decomposition approach.

In this thesis, we consider discrete-time open tandem networks with Poisson arrivals and general service times to analyse the approximation error of the renewal decomposition approach, investigate its causes, and introduce a refined decomposition approach that converges arbitrarily accurate. A tandem queuing network consists of an upstream and a downstream queue with infinite waiting rooms. The external Poisson arrival stream makes the upstream queue of type $M/G/1$ while the downstream queue is of type $G/G/1$. Since the upstream queue can be solved exactly, the research interest is in computing the performance measures of the downstream queue. As waiting time is a crucial key performance figure for the network's congestion and pace of order completion, the focus of this thesis is on the computation of the waiting time distribution in the downstream $G/G/1$-queue.

1.1 Research questions

Based on the pursued research contribution, we divide this thesis into three parts, each of which considers one of the following research topics.

Approximation quality of the renewal decomposition approach

Although it is widely accepted in the literature that the renewal decomposition approach yields approximate results, extensive studies on the approximation quality and their influencing factors are rare – especially in the discrete-time domain. In the first part of the thesis, we therefore present a comprehensive study on the approximation quality of the renewal decomposition approach in the discrete-time tandem queue and reveal the major influencing factors on the approximation quality. Our first research question is:

> **Which approximation quality should be expected when applying the renewal decomposition method in a discrete-time open tandem network and which factors impact the approximations?**

Output dynamics of the upstream queue

The renewal decomposition approach assumes that the departure point process of the upstream queue can be approximated by a renewal process. Therefore, our research interest in the second part is to analyse the output dynamics of the upstream queue. We investigate the self-dependent behaviour of the departure point process and discuss how the renewal assumption impacts the approximation of the performance measures at the downstream queue. The research question in the second part is:

> **How does the auto-correlated departure stream of the upstream queue impact the performance measures in the downstream queue?**

Refined decomposition approach for tandem queues

To overcome the renewal assumption, the third part of this thesis presents a refined decomposition approach for the discrete-time tandem queue that considers the interdependence among the connecting stream. The novelty of this decomposition method is that a semi-Markov process is used to model the connecting stream between the upstream and the downstream queueing system. However, stochastic models capturing the entire state space are computationally extensive. Therefore, we introduce a state space limitation method for the semi-Markov process that observes the state of the embedded Markov chain only if the number of customers in the queue does not exceed a given limit. This allows us to compute a satisfactory approximate solution with little computational effort. When better approximations are needed, the performance measures converge arbitrarily accurate with increasing computational expenditure. Our research question in the third part is:

> **How can a decomposition approach exploit auto-correlated arrival streams to compute satisfactory approximations with acceptable computational efficiency?**

1.2 Outline

The main body of this thesis is divided into three parts, each dedicated to answering one of the above research questions. All three parts are intended to be (or have already been) published as a stand-alone paper in the academic literature. Consequently, each main chapter is structured like a research paper and presents the motivation, related literature, stochastic models, numerical results, and conclusions bundled together. The three main chapters build on each other, and reference the previous results and conclusions to motivate each of the next research questions. In the following, we outline the structure of the thesis.

In Chapter 2, we briefly introduce the terms and methodologies of discrete-time stochastic modelling that will be applied throughout the thesis. Chapter 3 provides a literature overview on queuing network decomposition approaches in the continuous-time and the discrete-time domain.

Chapter 4 is taken from the paper "Point and interval estimation of decomposition error in discrete-time open tandem queues" published in *Operations Research Letters* (Jacobi and Furmans 2022b) and in the corresponding data article in *Data in Brief* (Jacobi and Furmans 2022c). In this chapter, we analyse the approximation quality of the discrete-time decomposition approach, compared to simulation, and with respect to the expected value and the 95th-percentile of waiting time. For both performance measures, we present regression models to compute forecasts of decomposition error. The ANOVA reveals major influencing factors on the approximation quality of the renewal decomposition approach.

Chapter 5 is based on the paper "On the output dynamics of the discrete-time $M/G/1$-queue" which is under review at *Annals of Operations Research* (Jacobi 2023b). In this chapter, we investigate the auto-correlation of the upstream queue's departure process to find situations in which the renewal assumption is not justified. We model the $M/G/1$-queue as a discrete-time Markov chain and compute the joint probability distribution of two departure instances. We present numerical results for the auto-correlation in the departure stream and discuss its effect on the analysis of downstream performance measures.

Chapter 6 is based on the working paper “A refined decomposition approach with converging accuracy for discrete-time open tandem queues with Poisson arrivals and general service times” (Jacobi and Shanthikumar 2023). This chapter presents the semi-Markov arrivals decomposition approach (SMAD), a refined decomposition approach, where the connecting stream between the upstream and the downstream station is described by a semi-Markov process. We model the departure process of the upstream $M/G/1$-queue as a semi-Markov process and introduce the state space limitation method of the embedded Markov chain. The downstream queue is a discrete-time $SM/G/1$-queue (semi-Markov arrival queue with general service times). We present numerical results to show that the approach produces reasonably accurate results when the state space is limited and converges arbitrarily accurate with increasing state space size.

Finally, Chapter 7 summarises the findings and contributions of this thesis and gives an outlook on future research perspectives.

2 Discrete-time stochastic modelling

Performance evaluation of stochastic systems necessitates the modelling of random variables and their state transitions over time. For example, it is of interest to know how many customers are in a queuing system at a given point in time or how the probability distribution of a random variable describing the time gaps between two arrival instances can be computed. Our objective is to develop stochastic models that capture the behaviour of these kind of random variables.

This chapter provides an overview of discrete-time stochastic modelling and introduces fundamental terms, definitions, and methodologies used throughout the thesis. However, we do not provide a comprehensive introduction into the topic; the interested reader is referred to Tran-Gia (1996) for the basics of discrete-time modelling, Daley and Vere-Jones (2008) for the theory of point processes, and Stewart (1994) and Limnios and Barbu (2008) for an introduction to discrete-time Markov and semi-Markov processes, respectively. A detailed introduction to queuing theory in the continuous-time domain can be found in the books written by Kleinrock (1975), Wolff (1989), and Buzacott and Shanthikumar (1992).

The remainder of this chapter is as follows. Section 2.1 introduces discrete-time modelling and defines stochastic processes. Section 2.2 defines discrete-time point and renewal processes and distinguish their properties. Section 2.3 introduces Markov and semi-Markov processes that will be used throughout the thesis to model the stochastic processes. Finally, Section 2.4 discusses the advantages of discrete-time modelling compared to stochastic models in the continuous-time domain.

2.1 Discrete-time stochastic processes

Discrete-time modelling means that the time axis is divided into time slots of equal length t_{inc}, and events (such as the arrival, start of service, or departure of a customer) are only observed at slot boundaries. Let N denote a random variable observed over time (e.g. the number of customers in a queuing system). The family $\mathcal{N} = \{N_t, t \in \Gamma\}$ of the discrete random variables $N_t \in \Xi$ describes a discrete-time stochastic process at time instance t (Tran-Gia 1996). The discrete state space $\Xi = \{0, 1, 2, ...\}$ may be countable (thus $\Xi = \mathbb{N}_0$), or bounded by some upper bound. The index set Γ denotes the set of all observation times of the stochastic system that are integer multiples of the slot parameter t_{inc} (Tran-Gia 1996).

Often, we are interested in the probability distribution of the time the stochastic process spends in state N_t. Let X denote the discrete random variable describing the time the random process $\mathcal{N}$ spends in N_t. Since the process is observed only at slot boundaries, this time is a multiple of the slot parameter t_{inc}, as well, and the probability for the discrete random variable X is described by

$$P(X = i \cdot t_{inc}) = x_i \quad \forall i \geq 0. \tag{2.1}$$

In the discrete-time stochastic models developed in this thesis, we describe the service, inter-arrival and inter-departure times by discrete random variables. For convenience, we do not include the slot parameter t_{inc} in the formulas developed in this thesis. In general, we assume $\Gamma = \mathbb{N}_0$, and thus, the stochastic process $\mathcal{N} = \{N_t, t = 0, 1, 2, ...\}$. However, in the stochastic models developed throughout this thesis, observing the system only at specific moments, for example, if a customer arrives at or departs from the system, might be convenient. These event points will be denoted as k, and we define the random processes $\mathcal{N} = \{N_k, k = 0, 1, 2, ...\}$ that observes the stochastic process only at event points. This consideration leads to the definition of discrete-time point processes, which will be discussed in the following.

2.2 Discrete-time point and renewal processes

An essential question for analysing stochastic systems is how the time gaps between consecutive events are distributed over time. For example, we are interested in modelling the inter-arrival or inter-departure times between two arrival / departure instances at a queueing system. We can think of arrivals or departures as discrete events that are randomly distributed along the time axis. The stochastic modelling of the set of these event points is called a (discrete-time) point process. In the following, we briefly define discrete-time point processes by introducing three different representations of the same point process. The definitions in this section are adapted from Whitt (1982), Tran-Gia (1996), and Daley and Vere-Jones (2008).

In general, we distinguish three representations of the same point process. A point process can be described by

- the random sequence $\{T_k, k = 1, 2, 3, ...\}$ of event points on a (positive) time axis,
- the time differences $\{X_k, k = 1, 2, 3, ...\}$ between two consecutive event points, and
- the associated counting process $\{N_t, t = 0, 1, 2, ...\}$ which counts the events at any discrete time point t.

A discrete-time point process is a sequence of event points in time that lie on the discrete-time axis. We consider point processes defined on the positive real line. The total number of points on the time axis is infinite, however, we assume that the number of points in any bounded interval is finite.

Let $T_k, k \geq 1$ denote the position of the k-th point, and $T_0 = 0$ (but with the notion that T_0 does not correspond to a point). We denote X_k the interval between the k-th and $(k-1)$-st points, such that $X_k = T_k - T_{k-1}$. Let the stochastic process $\mathcal{N} = \{N_t, t = 0, 1, 2, ...\}$ denote the associated counting process of the number of points in interval $[0, t]$, such that

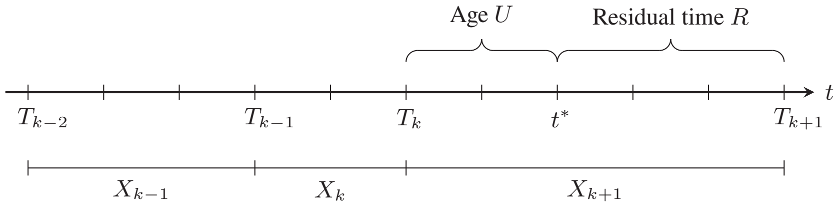

Figure 2.1: Point process on the discrete-time axis.

$$N_t = \max\left\{k \geq 1 : T_k \leq t\right\}, \quad t \geq 0. \tag{2.2}$$

Given $\{N_t\}$, we can construct the stochastic processes $\{T_k, k = 1, 2, 3, ...\}$ and $\{X_k, k = 1, 2, 3, ...\}$ by setting $T_0 = 0$, and compute

$$T_k = \min\left\{t \geq 0 : N_t \geq k\right\}, \quad k \geq 1, \tag{2.3}$$

and $X_k = T_k - T_{k-1}$. Therefore, the stochastic processes $\{N_t\}$, $\{T_k\}$ and $\{X_k\}$ are three different representations of the same point process.

Sometimes we are interested in observing the process not at an event point k, but at a time point that lies between two event points. Assume we observe the system at a random discrete observation point t^*. We call the time interval since the last event occurred the age of the process U and the time interval until the next event occurs the residual time R. Figure 2.1 shows the situation. A vital distinction when calculating the age and the residual time is whether we put the observation point t^* immediately before or immediately after the discrete time point. In the stochastic models developed in this thesis, we assume that the observation point lies immediately after the event.

Point processes are characterised by the fact that the random variables X_k in the interval sequence $\{X_k, k = 1, 2, 3, ...\}$ do not share the same probability distribution. We generally assume that each random variable X_k in the point process has a unique probability distribution. In contrast, we call a point process in which the time intervals of successive observation points are described by an i.i.d. random variable a renewal process. A renewal process is associated with the notion that the process resets after the event occurs and initiates again from the beginning. For example, replacing a defective component (the event) resets the process, and the lifetime distribution of the new component is the same as in the previous component. A renewal process is said to be memoryless if the time interval between two events has the same distribution as the recurrence time. It follows (Tran-Gia 1996) that the probability for event T_k to occur at a given time point is described by a Bernoulli process, the counting process $\{N_t\}$ is a Poisson process, and the time between two events is geometrically distributed.

2.3 Discrete-time Markov and semi-Markov processes

We model the state transitions in the discrete-time stochastic processes considered in this thesis using Markov processes (or Markov chains) and semi-Markov processes. The introduction provided in the following is based on the books by Stewart (1994) for Markov processes and Limnios and Barbu (2008) for semi-Markov processes. We do not intend to give a general introduction to the topic. Our aim is to provide the reader with a basic understanding of the modelling approaches deployed in the rest of this thesis.

We call a stochastic process $\{N_t, t = 0, 1, 2, ...\}$, $N_t \in \Xi$ a discrete-time Markov process or discrete-time Markov chain if the conditional transition probability function satisfies the Markov property

$$\begin{aligned} &P\big(N_{t+1} = n_{t+1} \,|\, N_t = n_t, N_{t-1} = n_{t-1}, ..., N_0 = n_0\big) \\ &= P\big(N_{t+1} = n_{t+1} \,|\, N_t = n_t\big). \end{aligned} \tag{2.4}$$

The Markov property implies that the state N_t contains all relevant information for the transition to the next state and this transition is independent of the transition

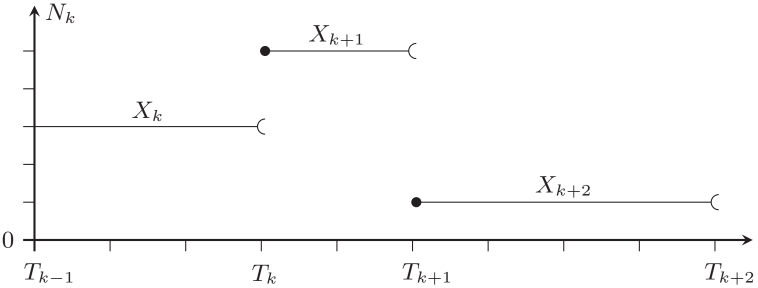

Figure 2.2: Sample path of a discrete-time semi-Markov process (Limnios and Barbu 2008).

path of the process in the past. We call this property memoryless, meaning that the past and the future of the stochastic system are conditionally independent. The conditional probability $P(N_{t+1} \mid N_t)$ is called single-step transition probability of the Markov chain. The Markov chain is said to be homogeneous, if the transition probability is independent of the observation point t, and we use the notation

$$p_{ij} = P\big(N_{t+1} = j \,\big|\, N_t = i\big) \quad i, j \in \Xi, t \in \mathbb{N}_0 \tag{2.5}$$

to describe the transition from state $N_t = i$ to state $N_{t+1} = j$. The transition probabilities form the entries of the transition probability matrix $\mathbf{P} = (p_{ij})_{i,j \in \Xi}$. Note that (depending on the state space size Ξ) matrix $\mathbf{P}$ might be of infinite size.

As already stated above, the memoryless property of the Markov process implies that the interval sequences X_k follow a geometric distribution. While this is intuitive for many applications, sometimes, we aim to relax the underlying assumption in order to allow arbitrarily distributed interval sequences in any state, while still having a "flexible" Markovian hypothesis. A process that has these properties is called a semi-Markov process.

Consider the coupled discrete-time stochastic process $\mathcal{Z} = \{(N_k, X_k), k = 1, 2, 3, ...\}$ which is composed of the stochastic processes $\{N_k, k = 1, 2, 3, ...\}$ and $\{X_k, k = 1, 2, 3, ...\}$. The stochastic process $\{N_k\}, N_k \in \Xi$ records the visited states at all time points k. Let $k = 1, 2, 3, ...$ denote all time points where the state in $\{N_k\}$ changes. As for the discrete-time point process, let $\{T_k, k = 1, 2, 3, ...\}$ denote the successive time points when the state changes in $\{N_k\}$, and $\{X_k\}$ the interval time the stochastic process spends in each state, such that $X_k = T_k - T_{k-1}$. By convention, we set $X_0 = T_0 = 0$. Figure 2.2 shows a sample path of such a stochastic process.

The stochastic process $\mathcal{Z} = \{(N_k, X_k), k = 1, 2, 3, ...\}$ is called a semi-Markov process if the conditional transition probability satisfies

$$\begin{aligned} &P\big(N_{k+1} = j, X_{k+1} = x \,\big|\, N_k, N_{k-1}, ..., N_0; T_k, T_{k-1}, ..., T_0\big) \\ &= P\big(N_{k+1} = j, X_{k+1} = x \,\big|\, N_k\big). \end{aligned} \tag{2.6}$$

The similarity between equation (2.6) and equation (2.4) immediately catches the eye. The property defined in equation (2.6) is what has been called earlier the flexible Markovian hypothesis. The expression of this property is that if we know the past visited states and interval times of the system, the next visited state and the associated interval time depend only on the present state. The Markov property therefore does not act on the discrete-time axis t, but on the time governed by the jump process $\{N_k\}$.

Obviously, the semi-Markov process $\{(N_k, T_k)\}$ describes a discrete-time point process, where the composed stochastic processes $\{N_k\}$ and $\{X_k\}$ are conditionally dependent. In the following, we describe how we can make use of this dependency to analyse the stochastic behaviour of the point process.

If equation (2.6) is independent of k, we call $\mathcal{Z} = \{(N_k, X_k), k = 1, 2, 3, ...\}$ (time) homogeneous. In this case, we can define the discrete-time semi-Markov kernel $\mathbf{Q} = (q_{ij}(x); i, j \in \Xi, x \in \mathbb{N})$ by

$$q_{ij}(x) = P\big(N_{k+1} = j, X_{k+1} = x \,\big|\, N_k = i\big). \tag{2.7}$$

Similar to the transition probability matrix in a Markov chain, the semi-Markov kernel $\mathbf{Q}$ is the essential quantity to define a semi-Markov process.

The stochastic process $\{N_k, k = 1, 2, 3, ...\}$ is called the embedded Markov chain of the semi-Markov process. If condition (2.6) holds true, $\{N_k\}$ is a homogeneous Markov chain with probability transition matrix $\mathbf{P} = (p_{ij})_{i,j\in\Xi}$, where the transition probabilities

$$p_{ij} = P\big(N_{k+1} = j \,\big|\, N_k = i\big) \quad i, j \in \Xi, k \in \mathbb{N} \tag{2.8}$$

can be computed by

$$p_{ij} = \sum_x q_{ij}(x) \quad i, j \in \Xi. \tag{2.9}$$

The interval time distributions of the semi-Markov chain is defined by the matrix $\mathbf{F} = (f_{ij}(x)_{i,j\in\Xi,x\in\mathbb{N}})$, where

$$f_{ij}(x) = P\big(X_{k+1} = x \,\big|\, N_k = i, N_{k+1} = j\big) \quad i, j \in \Xi, k \in \mathbb{N} \tag{2.10}$$

denotes the conditional probability for an interval time of x time units, given that the stochastic process transitions from state i to state j. Note the difference to the definition of the semi-Markov kernel.

For any states $i, j \in \Xi$ and non-negative integer x the condition

$$q_{ij}(x) = p_{ij} \cdot f_{ij}(x) \tag{2.11}$$

holds true. Given this condition, we can construct semi-Markov processes based on a homogeneous Markov chain whose state interval times are conditioned by a interval time distribution $\mathbf{F}$.

2.4 Advantages of discrete-time modelling

Discrete-time stochastic modelling – especially discrete-time queueing models – arose in the 1980s with an increasing research interest in data transmission in networks. Since data package transfer in networks such as ATM (Asynchronous Transfer Mode) is usually organised in time slots, discrete-time models are well suited to describe the stochastic behaviour of telecommunication networks. Essential contributions in this research field have been published by Ackroyd (1980), Bruneel and Kim (1993), and Hübner and Tran-Gia (1994).

The stochastic analysis of material handling and production systems also lends itself to discrete-time stochastic modelling. The advantages of discrete-time modelling for the analysis of production and material handling systems in terms of accuracy, level of detail, and efficiency have been introduced and discussed in detail by Schleyer (2007) and Epp (2018). In the following, we briefly summarise the main beneficial aspects of discrete-time modelling. Further, we add another aspect to the argument and examine the difficulties of continuous-time modelling for computing the probability distribution of performance measures.

Concerning the practical application of queueing theory for the design and analysis of production systems and material handling systems, it is of particular interest to calculate the entire probability distribution of performance measures with reasonable computational effort. This allows us to determine the 90%, 95%, or 99% percentile of the probability distribution and thus to make more detailed statements about the system's performance. The main advantage is that discrete-time modelling does not require any assumption to be made about the distribution of the inter-arrival and service time. For the $G/G/1$-queue, for example, we compute the entire waiting time distribution by using the method by Grassmann and Jain (1989) and the distribution of waiting customers with the method by Grassmann and Tavakoli (2019). In contrast, continuous-time models of the $G/G/1$-queue are usually limited to the computation of the first two moments (we discuss a relaxation of this limitation later). This has two consequences. First, using continuous-time models, we cannot report the system's performance beyond the expected value and its variability. This is a disadvantage if the system is to be designed to perform 90%, 95% or 99% of the time. Additionally, Schleyer (2007) demonstrated in numerical examples for the $G/G/1$-queue that the accuracy of the continuous-time models is flawed when computing performance values with input distributions having the same mean and variance but different skewness and kurtosis. Despite the higher level of detail in comparison to continuous-time models, discrete-time models preserve the advantage of being computationally efficient. Simulation models offering the same level of detail require considerable effort for implementation, verification, validation, and performing the experiments. Therefore, discrete-time models are well suited to analyse and report the long-term steady-state performance of stochastic systems and to conduct extensive "what-if" analyses.

In the past years, researchers developed stochastic models in the continuous-time domain that compute the entire probability distribution of performance measures, seeking to overcome the abovementioned limitation of reporting only the first two moments. The remarks in the following aim to discuss these approaches, identify their drawbacks and thus contribute another aspect to the advantage of discrete-time modelling.

In the continuous-time domain, the computation of stationary performance distributions is generally computationally expensive. Approaches for the stationary queue length distribution are based on the probability generating function given by the Pollaczek–Khinchine transform equation (Harchol-Balter 2013), and by the supplementary variable method (Cox 1955). Recently, Sherzer et al. (2022) presented a deep-learning method to fit the steady-state probability distribution of the number of customers in a continuous-time $M/PH/1$-queue. The approach receives the arrival rate and the first l moments of the Phase-type service time distribution as input and outputs the stationary queue length distribution.

However, deriving the steady-state distribution only on the basis of the first l moments may lead to substantially large errors. Even if the first l moments of a probability density function are appropriately defined, the asymptotic behaviour of the distribution may be unpredictable. The following example is borrowed from

Shanthikumar (2022) and shall demonstrate the problem. Assume that random variable X is a Phase-type distributed random variable that arbitrarily closely approximates the given first l moments of an (unknown) probability distribution. We further assume that the probability distribution is light-tailed, that is, X has an exponentially decaying complementary density function f_X (Gass 2013, p. 880), and f_X generates moments that are bounded upwards.

Supposing that the given l moments are sufficient to predict the probability distribution of X, the error in this prediction can be substantial. It can be shown that there exists a random variable X' that arbitrarily closely approximates the observed l moments, but the $(l+1)$-st moment is asymptotically infinite, $m_{l+1} \to \infty$. Let Y denote a random variable that has the probability density function f_Y of the form

$$f_Y(x) = x^{-\gamma} \quad \gamma > 0,\ 0 < x < \infty. \tag{2.12}$$

We construct the probability density function f' of random variable X' by the weighted sum of the distribution functions f_X and f_Y with weighting factor φ,

$$f'(x) = (1 - \varphi) \cdot f_X(x) + \varphi \cdot f_Y(x). \tag{2.13}$$

Let φ be arbitrarily close to zero, such that f' generates the observed l moments, as well. The moment m_l of function f' is computed by

$$\begin{aligned} m_l &= \int_{-\infty}^{\infty} (x-c)^l \cdot f'(x)dx \\ &= \int_{-\infty}^{\infty} (x-c)^l \cdot \Big((1-\varphi) \cdot f_X(x) + \varphi \cdot f_Y(x)\Big)dx \\ &= \int_{-\infty}^{\infty} (x-c)^l \cdot \Big((1-\varphi) \cdot f_X(x)\Big) + \varphi \cdot (x-c)^l \cdot x^{-\gamma}dx. \end{aligned} \tag{2.14}$$

In equation (2.14), the asymptotic behaviour of the term $(x-c)^l \cdot x^{-\gamma}$ is

$$(x-c)^l \cdot x^{-\gamma} \to \begin{cases} 0 & \gamma \leq l, \\ \infty & \gamma > l. \end{cases} \tag{2.15}$$

Consequently, for $m_1, m_2, ..., m_l$ the integral in equation (2.14) is properly defined, but asymptotically, the $(l+1)$-st moment is infinite, $m_{l+1} \to \infty$.

This example illustrates that the first l moments of a probability distribution do not necessarily uniquely define the distribution type. As a consequence, if we derive properties of a probability function (e.g. the 90%, 95% or 99% percentiles) to draw conclusions about the system's performance solely based on knowing the first l moments, the prediction error can be substantially large. As mentioned above, Schleyer (2007) identified this phenomenon (for the $G/G/1$-queue) using numerical experiments.

In conclusion, making service-oriented statements about the system's performance necessitates the computation of the distribution's tail. Since continuous-time models measure the first l moments of the distribution, deriving the steady-state behaviour of the queue may lead to substantially large errors. In contrast, discrete-time queueing models compute the entire probability distribution, regardless of the shape and tail of the input values. Discrete-time queueing models therefore are advantageous both compared to simulation models (in terms of computational complexity) and compared to continuous-time queueing models (in terms of accuracy).

3 Literature review

Performance analysis of queueing networks is a broad research field covering numerous publications from over 50 years. Bitran and Tirupati (1988) roughly divide the available methods into exact approaches, approximation methods, and simulations. As already stated, simulation is a powerful tool for performance evaluation. However, it requires considerable effort for validation, and performance evaluation is time-consuming. Exact results are obtained only under restrictive conditions. Thus, the majority of the approaches presented in the literature are approximate, and we can further distinguish (Bitran and Tirupati 1988):

- Diffusion approximation,
- Mean value analysis, and
- Decomposition methods.

Naturally, this chapter focuses on decomposition methods, both in the continuous- and the discrete-time domain. Our aim in this chapter is to present the approaches in the literature to approximate the point processes in the connecting inter-node streams of queueing networks and to compare the expected accuracy of the performance results. In Section 3.1 we start the literature review with a "look over the edge of the plate" and briefly introduce diffusion approximation, mean value analysis, and the product form solution for exact performance evaluation. Section 3.2 focuses on decomposition methods of transfer lines and open queueing in the continuous- and discrete-time domain. In Section 3.3, we briefly summarise the main conclusion from the literature review.

3.1 Diffusion approximation, mean value analysis, and exact methods

Diffusion approximation is an early attempt to consider queueing models with generally distributed service times. The approach is motivated by the heavy traffic limit theorem and relies on the assumption that the queues under consideration are almost never empty. Iglehart and Whitt (1970) demonstrate that under heavy traffic conditions, congestion measures at the second node of a queueing network are asymptotically the same as if the first queueing system was removed. This implies that the arrival process to the second node can be approximated by the arrival process at the first node. Diffusion approximation methods use these results to asymptotically approximate the point arrival processes as diffusion process (which is a continuous-path Markov process). The interested reader is referred to the papers published by Iglehart and Whitt (1970), Reiser and Kobayashi (1974), Gaver and Shedler (1973b,a) and the books written by Cox and Miller (1965) and Newell (1982).

Mean value analysis was initially introduced by Reiser and Lavenberg (1980) and is a recursive approach for analysing closed queueing networks. Sevcik and Mitrani (1981) develop the arrival theorem for this class of networks in case of exponentially distributed service times. The arrival theorem states that in an arrival instance, the distribution of customers seen (as an outside observer) by this arriving customer is the same as the steady-state customer distribution of the network with one less customer. This in an important finding as it enables the recursive computation of the customer distribution. From the perspective of material handling, deploying mean value analysis for performance evaluation of closed queueing network is still of research interest, for example, to determine the optimal fleet size and service availability of a fleet of automated guided vehicles (George and Xia 2011) or performance evaluation of automated material handling systems (Govind et al. 2010). We refer the interested reader to the work published by Dallery and Cao (1992), Onvural (1990), and Lagershausen (2013) for a more detailed study of this topic.

Jackson (1957) introduces an appealing approach to solve open queueing networks with external Poisson arrivals, exponentially distributed service times, and Markovian job transfer between the queueing systems. Jackson (1957) demonstrates that the steady-state probability distribution of the number of customers in the network can be computed exactly using a product-form solution. This approach generates exact results since $M/M/1$-queues with infinite waiting room generate a Poisson output process (Burke 1956), and further, random splitting and superposition of Markov processes again form a Markov process. Jackson's product-form solution is considered the first decomposition approach for open queueing networks, and sets a landmark for the development of other decomposition methods for the analysis of networks with generally distributed service times.

3.2 Decomposition methods

Since the assumption of Poisson arrivals and exponentially distributed service times in Jackson networks is restrictive for many applications, researchers developed methods to analyse queueing networks with generally distributed interarrival and service times. The common idea of the methods presented in the following is to decompose the original network of queues into a set of smaller subsystems which are easier to analyse. Each decomposition method involves three steps (Dallery and Gershwin 1992): First, the subsystems are appropriately characterised, second, a set of equations is derived to determine the unknown parameters in each subsystem, and third, an algorithm is developed to solve these equations. In the following, we first consider transfer lines which employ tandem queues as building blocks for the decomposition method, and then review decomposition methods for open queueing networks in the continuous- and the discrete-time domain.

3.2.1 Transfer lines

Transfer lines are an important sub-class of networks which arise e.g. in the context of manufacturing systems and chemical processes. In a transfer line, the material flow is sequential and the service stations are decoupled by buffers with finite capacity. Each station encounters a probability to fail, which may cause the upstream machine to block or the downstream machine to starve. With these characteristics, transfer lines clearly distinguish themselves from open queueing networks that have reliable queues with infinite buffers. The following descriptions shall give a brief overview over the analysis of transfer lines using decomposition. Extensive literature reviews that go into more detail are provided by Dallery and Gershwin (1992) and Papadopoulos and Heavey (1996).

Early investigations on the analysis of transfer lines have been carried out by Buzacott (1967) who investigates the effects of a given buffer capacity and the distribution of buffers on the line efficiency. Exact results for the analysis of transfer lines with two or three machines are presented by Gershwin and Schick (1983) for the continuous-time, and Gebennini and Grassi (2015) for the discrete-time domain. The analysis of longer lines relies on decomposition techniques. Gershwin (1987) approximates a line of k machines by a set of $k-1$ tandem systems with two machines and presents an iterative algorithm to solve the resulting set of equations. However, Dallery et al. (1988) point out that the proposed algorithm may fail to converge. Therefore, Dallery et al. (1988) replace the original set of equations by an equivalent one, which is again solved using an iterative procedure. Extensions to these approaches focus on the analysis of networks with splits and merges (cf. Gershwin (1991), Gershwin and Burman (2000)). Gershwin (1991) presents a decomposition method where the network is decomposed into two-machine lines with intermediate buffer. The machine parameters are chosen so that the behaviour of the material flow in the buffers of the two-machine lines closely matches that of the flow in the buffers of the original line. Gershwin and Burman (2000) present a generalisation of this work for inhomogeneous networks, where machines can operate with different speeds.

3.2.2 Continuous-time open queueing networks

In contrast to transfer lines, open queueing networks are characterised by a flexible network structure, which allows for example for the modelling of a dynamic job shop manufacturing system. We distinguish three basic steps in each decomposition method of open queueing network models (Bitran and Dasu 1992):

1. Characterisation of the arrival process at each station,
2. Analysis of the queue based on the characteristics of the arrival process,
3. Determination of the departure process.

Decomposition approaches for open queueing networks generally rely on two basic assumptions (Govil and Fu 1999): First, it is assumed that the nodes of the network (that is, the individual queueing systems) can be treated as being statistically independent. Second, it is assumed that the input to each queueing system is a renewal process characterised by the mean and the variance of inter-arrival time distribution of customers. In the continuous-time domain, this approach was first by applied by Kuehn (1979) with modifications presented by Shanthikumar and Buzacott (1981), Whitt (1983b), and Reiman (1990). In the discrete-time domain, Haßlinger and Rieger (1996) proposed a refinement of these so-called parametric decomposition which allows for the computation of the entire probability distributions of performance measures.

In the following, we briefly introduce the above-mentioned decomposition approaches, focusing (a) on the techniques to approximate the point arrival processes and (b) the achieved accuracy of the performance measures. However, it must be noted that there is a plethora of decomposition techniques to be found with various adaptions, for example, multiple customer classes (Bitran and Tirupati 1988, 1989, Whitt 1994). For a broader overview over the related literature, the interested reader is referred to the review articles written by Bitran and Tirupati (1988), Bitran and Dasu (1992), Govil and Fu (1999), Shanthikumar et al. (2007), and Worthington (2009).

The first parametric decomposition approach was introduced by Kuehn (1979) who presents an approximate analysis of open queuing networks with generally distributed inter-arrival and service times. The network is decomposed into single-station $GI/G/1$-queuing systems which are independently analysed. Kuehn (1979) uses conservation of flow to compute the mean arrival rates of each queuing system, the formula by Kraemer and Langenbach-Belz (1976) to compute mean waiting times, and the formula by Marshall (1968) to compute the variability of the departure process of each queuing system. Kuehn (1979) reports the method to yield generally an increasing accuracy under the conditions of

- low or heavy traffic,
- increasing randomness in the arrival and service processes (so that the network is close to a Markovian network),
- increasing network complexity,
- decreasing closedness of the network (that is, number of feedback loops).

This is due to the fact that under these conditions, the renewal assumption is usually better fulfilled. Kuehn (1979) presents numerical results of the approximations of the decomposition approach, compared to simulation. He finds relative errors in flow time (that is, the sum of waiting and service time at each queuing system) in the range of -30.01% and 69.35% in case of a two-node network with feedback loops. The high relative errors are observed for heavy traffic situations, where (in combination with the closedness of the network) the renewal assumption becomes critical. In another example of a network with 9 queuing systems, the flow times of decomposition approach are within the 95% confidence interval of the benchmark simulation.

Shanthikumar and Buzacott (1981) present a decomposition technique for dynamic job shop type open queuing networks with generally distributed service times. The renewal approximations of the arrival point processes relies on two types. The first type considered is general arrival processes which is represented by the mean flow rate and the coefficient of variation. Shanthikumar and Buzacott (1981) present approximation formulas to compute the scv-values for random splitting and superposition of the general renewal processes, as well as for the arrival and departure processes to and from the queues. The second type of arrival streams considered is Poisson arrival processes which is defined uniquely by the mean flow rate. Based on the definition of these arrival process approximations, Shanthikumar and Buzacott (1981) introduce a set of equations to compute the mean sojourn time for each node in the network in case of first-come-first-serve and shortest-processing-time-first service disciplines. The observed computational results for the mean sojourn times are reported to be inside the simulated 95% confidence interval for most cases considered. The approximation of the arrival processes by Poisson processes is reported to perform better for dynamic job shops (compared to flow jobs), which is an important finding as it suggests that the superposition of (numerous) renewal processes asymptotically forms a Poisson process.

Whitt (1983b) introduced the Queueing Network Analyzer (QNA), a software package for the analysis of open queuing networks with generally distributed inter-arrival and service times. As a generalisation of the product-form solution (Jackson 1957), the network is decomposed into $GI/G/m$-queuing systems. In the QNA, flow rates are obtained via traffic rate equations, just as with Markov models (Jackson 1957). To compute the variability measures for the internal flows, QNA employes two procedures presented by Whitt (1982) to approximate the point departure processes by renewal processes: The stationary interval method equates the moments of the renewal interval with the moments of the stationary interval in the point process to be approximated. The asymptotic method takes into account the dependence between successive intervals and determines the moments of the renewal interval by matching the asymptotic behaviour of the moments of the sums of successive intervals. Since neither the asymptotic method nor the stationary interval method yields promising results for a wide range of variability parameters, Whitt (1983b) introduces a hybrid procedure based on the work by Albin (1984a,b). In the QNA, the analysis of the queueing performance relies on the formula by Kraemer and Langenbach-Belz (1976) for values of the squared coefficient of variation in the arrival stream smaller than 1, and a corrected expression in cases with variability values greater 1. Finally, the variability of departure stream of the $GI/G/1$-queue is observed as approximation of the stationary interval method (Whitt 1984) and the formula by Marshall (1968).

The performance of the Queueing Network Analyzer is investigated by Whitt (1983a) with respect to several network structures (single $GI/G/1$-queue, superposition of arrival processes, a couple two-node network as investigated by Kuehn (1979), and several more complicated networks as suggested by various authors). The results of investigations of $GI/G/1$-queues suggest that the reliability of approximations decreases when variability in the arrival stream increases. In general, Whitt (1983a) concludes that the maximum relative error in the $GI/G/1$-queue might be approximated by $0.05 \cdot c_a^2$, that is, about 10% for $c_a^2 = 2.0$. Further, based on the insights from the heavy traffic bottleneck phenomenon, Whitt (1983a) generally suggests the quality of the approximations to improve with increasing utilisation (an assumption later disproved by Kim (2004)). In a comparison against the decomposition method by Kuehn (1979), the QNA yields about the same approximation quality.

Reiman (1990) presents two decomposition methods (the individual bottleneck decomposition and the sequential bottleneck decomposition) to analyse open queuing networks with generally distributed inter-arrival and service times. When applied to a Jackson network or a $M/G/1$-queue, the decomposition method yields exact results. Further, the approach is asymptotically exact in light traffic situations, and in networks with a single bottleneck station in heavy traffic. Reiman (1990) deploys the results from heavy traffic behaviour of queues to determine approximation formulae for mean waiting and sojourn times. He compares the performance of both decomposition approaches to the results presented by Kuehn (1979) and the QNA (Whitt 1983b). In the two-node feedback model introduced by Kuehn (1979), Reiman (1990) finds both decomposition approaches to perform mostly better than QNA (mean errors of sojourn time between -6.78% and 12.95%, compared to simulation). In a second example (which is two queues in series as investigated by Whitt (1983b)), QNA outperforms both approaches (-16.25% up to 6.12% errors in sojourn time, compared to simulation).

3.2.3 Discrete-time open queueing networks

Haßlinger and Rieger (1996) present a refinement of the parametric decomposition approach for the analysis of open queueing networks in the discrete-time domain. The discrete distribution of superpositions of renewal processes is reversibly obtained by the distribution of the minimum of the residual times of all superposed flows. A recursive method and a faster approach based on the z-transform for the computation of the stochastic split of a renewal process are presented. For the analysis of discrete $GI/G/1$-queues, the algorithm by Grassmann and Jain (1989) is applied to calculate the stationary waiting and idle time distributions. The stationary distribution of the number of customers is computed based on an polynomial factorisation approach (Haßlinger 1995). Haßlinger and Rieger (1996) compare the results obtained with their decomposition approach with the parametric decomposition approach by Kuehn (1979). Further, they compare the congestion measures of the $GI/G/1$-queues with the formulae by Kraemer and Langenbach-Belz (1976) and the QNA (Whitt 1983b). The results are reported to be in good agreement for various combinations of different arrival and service time distributions, except for deterministic service times.

Furmans (2004) presents a framework for stochastic finite elements to model material handling systems in the discrete-time domain. In this decomposition approach, the probability distribution of the superposition of independent flows is approximated by the minimum of the time to the next renewal period in each incoming stream. The probability distribution of the inter-arrival time after a stochastic split is obtained by the l-times iterative convolution of the inter-departure time distribution with itself, weighted with the probability that l consecutive customers are not routed to the respective direction. The waiting time distribution and the inter-departure time distribution of the resulting $GI/G/1$-queueing system are obtained with the algorithms presented by Grassmann and Jain (1989), and Jain and Grassmann (1988), respectively. For the calculation of the probability distribution of the number of customers in the queueing system, the algorithms presented by Furmans and Zillus (1996), or Grassmann and Tavakoli (2019) may be deployed.

3.3 Chapter conclusion

To conclude the literature review, it is crucial to emphasise that congestion measures obtained by decomposition techniques that follow the renewal assumption of the connecting streams are approximate. Below, we briefly discuss the implications for the analysis of open queuing networks and derive the research gap for this thesis.

In the continuous-time domain, the problem is two-fold. First, the assumption of independence among the queueing systems and the resulting approximation of point processes by renewal processes introduces approximations in the performance measures. While the computation of the first moment of the arrival streams is straightforward, approximations are needed to compute the second moment. Thus, several authors (Kuehn 1979, Buzacott and Shanthikumar 1993, Whitt 1983a,b) demonstrate that inter-arrival time variability parameters are significant for the approximation quality of the decomposition approach. The second problem arises from the use of approximate formulas for performance analysis (e.g. Kraemer and Langenbach-Belz (1976)) of the resulting $GI/G/1$-queues.

Decomposition methods in the discrete-time domain only face the problem of the renewal assumption since performance measures can be computed with great accuracy in the $GI/G/1$-queue. However, decomposition approaches in the discrete-time domain are rare, and the approximation quality of the approach still needs to be thoroughly investigated. Despite discussing the renewal assumption and its implications, Haßlinger and Rieger (1996) state that "further study is needed to construct [...] representations of non-renewal processes" in the discrete-time domain that enable the computation of exact results.

This thesis aims to close this research gap for the discrete-time tandem queue with Poisson arrivals. The contribution is three-fold: First, in Chapter 4, we quantify the approximation quality of the renewal decomposition approach and identify situations where the renewal assumption is critical. Based on the point and interval estimates developed in Chapter 4, we can predict the approximation error with great accuracy. Second, in Chapter 5, we investigate the output dynamics of the upstream $M/G/1$-queue, demonstrating that the departure stream is autocorrelated. We conclude that successive departures are dependent because the inter-departure time of one customer depends on the state in the system that the previous customer left behind. We discuss the implications for the approximation error of the renewal decomposition approach identified in Chapter 4. Third, in Chapter 6, we introduce a novel decomposition approach that is based on semi-Markov arrivals. Semi-Markov arrivals capture the state-dependent behaviour of the upstream departure process identified in Chapter 5. Therefore, the novel decomposition approach does not rely on the renewal assumption, which allows for a converging accurate queueing analysis in the downstream $G/G/1$-queue.

4 Point and interval estimation of decomposition error in discrete-time open tandem queues

This chapter is published under the title “Point and interval estimation of decomposition error in discrete-time open tandem queues” in *Operations Research Letters* (Jacobi and Furmans 2022b). The text in this chapter has been taken from the paper with minor changes. The regression models appear in the accompanied data article in *Data in Brief* (Jacobi and Furmans 2022c), the data is available in the *KITopen Repository* (Jacobi and Furmans 2022a).

The author of this thesis was responsible for the conceptualisation, methodology, software programming, validation, formal analysis, writing (original draft and review), and visualisation of the research presented in this chapter.

Chapter abstract

In this chapter, we analyse the approximation quality of the discrete-time decomposition approach, compared to simulation, and with respect to the expected value and the 95% percentile of waiting time. For both performance measures, we present OLS regression models to compute point estimates, and quantile regression models to compute interval estimates of decomposition error. The ANOVA

reveal major influencing factors on decomposition error while the regression models are demonstrated to provide accurate forecasts and precise confidence intervals for decomposition error.

4.1 Introduction

Queuing models are widely used for performance evaluation of production and logistics systems which are subject to the influence of randomness (Shanthikumar et al. 2007, Van Nieuwenhuyse and de Koster 2009, Wu et al. 2019, Yu and de Koster 2009, Lieckens and Vandaele 2012). When applying continuous-time queueing models, engineers calculate the first and second moment of performance indicators of interest (e.g. throughput, waiting time, and the number of customers in the queue) using the well-known formulas for $M/M/1$ and $M/G/1$ queues as well as approximation formulas for $G/G/1$-queues. Books that provide an overview of continuous-time queueing models are written by Buzacott and Shanthikumar (1993) and Wolff (1989).

However, production and logistics systems are typically designed to guarantee performance not on average, but with a given probability (e.g. 95%), which necessitates the calculation of the distribution of key performance indicators (such as waiting time) to know, for example, which percentage of orders are processed in 3h or less, or what promised throughput time will be met in 95% of the cases (Schleyer and Gue 2012). Applying discrete-time queueing models allows for the computation of the entire probability distributions of key performance indicators under very general assumptions. Discrete-time modelling means that events are only recorded at moments that are multiples of a constant time unit t_{inc}. Thus, the probability mass function of a discrete random variable x is denoted by

$$P(x = i \cdot t_{inc}) = x_i \quad \forall i = 0, ..., i_{max}. \tag{4.1}$$

Given the discrete random variables for the inter-arrival and service time, the probability distributions of performance measures can be computed, for example the waiting time (Grassmann and Jain 1989) or the inter-departure time (Jain and Grassmann 1988) distributions. A comprehensive introduction to discrete-time queueing models can be found in the books by Ackroyd (1980) and Bruneel and Kim (1993). The models have been successfully applied in various use cases related to logistics and production systems (Schleyer and Gue 2012, Epp et al. 2017, Schwarz and Epp 2016, Schleyer and Furmans 2007, Schleyer 2010).

The analysis of discrete-time open queueing networks relies on a decomposition approach. As in the continuous-time domain, the technique is known to yield approximate results in the case of non-Poisson arrivals and generally distributed service times. The drawback with approximations is that we cannot quantify the deviation of the performance measures calculated with a decomposition approach from their actual values. While the approximation quality of decomposition approaches has been studied in the literature for the continuous-time domain (see e.g. Suresh and Whitt (1990) and Kim et al. (2005)), decomposition error in the discrete-time domain has not yet been comprehensively examined. So far, no estimator is available to predict decomposition error for a given queueing network in the discrete-time domain.

In this chapter, we investigate discrete-time open tandem queues to analyse and forecast the approximation quality of the discrete-time decomposition technique, compared to simulation. We limit ourselves to the analysis of tandem queues with external Poisson arrivals that become non-renewal at the downstream queue with the aim to reveal fundamental dependencies regarding the approximation quality of the discrete-time decomposition approach.

The remainder of this chapter is organised as follows. In Section 4.2, we present the theoretical background of the discrete-time decomposition approach. Section 4.3 introduces the methodology to compute point and interval estimates for decomposition error, defines statistical tests to evaluate their accuracy, and introduces the design of experiments of this study. In Section 4.4, we show numerical results for decomposition error and present the regression models for point and interval estimation of decomposition error. In Section 4.5, we discuss extensions of the study presented. Section 4.6 discusses the main drivers on decomposition error. Section 4.7 concludes the chapter.

4.2 Theoretical background

Open queueing networks allow for the analysis of systems with infinite buffer capacity and generally distributed inter-arrival and service times. Generalisations of Jackson's product form solution (Jackson 1957, 1963) with respect to generally distributed inter-arrival and service times are proposed by Reiser and Kobayashi (1974) with modifications presented by Kuehn (1979), Shanthikumar and Buzacott (1981), Whitt (1983b), and Bitran and Tirupati (1988, 1989). Each decomposition approach relies on two basic assumptions (Govil and Fu 1999): First, it is assumed that the individual queueing systems can be treated as being statistically independent $GI/G/1$-queues. Second, it is assumed that the point process which forms the input to each $GI/G/1$-queue can be approximated by a renewal process. It is therefore important to emphasise that congestion measures obtained by decomposition techniques are approximate, since the assumption of independence among queueing systems does not properly account for the correlations of the arrival stream which have a significant effect on the performance measures (Kim et al. 2005).

Decomposition approaches for discrete-time open tandem queues are based on these conditions, as well. The arrival stream of a downstream queue is approximated as renewal process by the inter-departure time distribution of the upstream queue, which can be efficiently computed with the algorithm by Jain and Grassmann (1988). The waiting time distribution of the resulting $GI/G/1$-queue is obtained with the algorithm presented by Grassmann and Jain (1989). Further performance measures, such as the distribution of customers, can be computed with the approaches presented by Haßlinger (1995), and Grassmann and Tavakoli (2019).

In an effort to investigate the approximation quality of the decomposition techniques, tandem lines have been studied extensively in the literature. Suresh and Whitt (1990) examine the impact of non-renewal processes on the approximation quality with different traffic intensities. Wu and McGinnis (2013) introduce the intrinsic ratio, a fundamental property of tandem queues that is based on the insight that some servers are directly affected by the external arrival process. Whitt (1995) suggests using a variability function (instead of a single parameter as in the QNA) for the arrival stream of the downstream queue, which is a function of the traffic intensity of the incoming queue. Sagron et al. (2015) extend this method to multi-class systems that address the scenario when the upstream server in a tandem queue experiences downtimes (e.g. set-up, maintenance, and repair), events that increase the station's departure variability, while causing starvation of a downstream bottleneck station. To achieve better computational efficiency, Sagron et al. (2017) approximate the between-class effect (the variability caused by interactions with other classes) in a queue with downtimes using a Regression-Based Variability Function (RBVF). RBVF receives the squared coefficient of variation of the arrival and service times, as well as the expected value of the service process as input and approximates the variability function using methods of linear regression.

4.3 Methodology

The object of investigation in this thesis is a tandem queue, that is, two discrete-time queueing systems are arranged one after the other (see Figure 4.1). The upstream queueing system is fed by an external arrival stream with arrival rate $1/E(A^U)$ of customers. If the service station is busy upon arrival of a customer, this customer waits in the waiting room for the service to begin. After being served at the upstream station with service rate $1/E(B^U)$, all customers enter the waiting area of the downstream queueing system. The size of the waiting area is infinite, meaning that all customers wait to be served with service rate $1/E(B^D)$ at the downstream station and to hereafter leave the tandem queue.

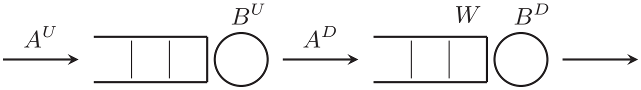

Figure 4.1: Parameters in the tandem queue.

We only consider steady-state systems where the utilisation parameters $\rho^U = E(B^U)/E(A^U)$ and $\rho^D = E(B^D)/E(A^D)$ are smaller than 1. Since the arrival process at the downstream queue is approximated as point process with inter-arrival time distribution A^D, only the downstream queueing system is prone to decomposition error. For the sake of clarity, Table 4.1 defines the system performance metrics of the tandem queue.

In our analyses, we assume that the random variables describing the service processes are described by discretised gamma distributions. Let X be a gamma-distributed random variable with shape parameter k and scale parameter θ. The probability density function of X is given by (Bijma et al. 2017)

$$f(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)}, \quad x,k,\theta > 0, \tag{4.2}$$

where $\Gamma(k)$ is the gamma function. We use the squared coefficient of variation (scv) as normalised measure of statistical dispersion to measure the process variability. Let $E(X)$ define the expected value of X, and $Var(X)$ its variance. The variability of X is defined as

$$scv(X) = Var(X)/E^2(X). \tag{4.3}$$

In order to generate gamma-distributed random variables X with predefined values for $E(X)$ and $scv(X)$, we use the well-known closed-form expressions for the shape and scale parameters of the gamma distribution,

$$\begin{aligned} E(X) &= k\theta, \\ Var(X) &= k\theta^2. \end{aligned} \tag{4.4}$$

Table 4.1: Performance metrics of the tandem queue.

A^U, A^D	Random variable describing the inter-arrival time of the external (downstream) arrival process
B^U, B^D	Random variable describing the service time at the upstream (downstream) queue
ρ^U, ρ^D	Utilisation of the upstream (downstream) queue
W	Random variable describing the waiting time of a customer at the downstream queue

Finally, we define σ_X^τ as the τ-percent percentile of the probability mass function (pmf) of random variable X.

In this chapter, we are interested in the error of the waiting time W at the downstream queue computed by the discrete-time decomposition approach, compared to discrete-event simulation. We conduct two distinct studies with different dependent variables. In Study I, let $\Delta(E)$ be the divergence of the expected value of waiting time

$$\Delta(E) = \frac{E_{Sim}(W) - E_{Queue}(W)}{E_{Sim}(W)}, \tag{4.5}$$

where $E_{Sim}(W)$ and $E_{Queue}(W)$ denote the expected value of waiting time, computed with the discrete-time queueing approach and simulation, respectively. In Study II, let $\Delta(\sigma)$ be the divergence of the 95% percentile of waiting time

$$\Delta(\sigma) = \frac{\sigma_{W,Sim}^{95} - \sigma_{W,Queue}^{95}}{\sigma_{W,Sim}^{95}}, \tag{4.6}$$

where $\sigma_{W,Queue}^{95}$ denotes the 95% percentile of waiting time, computed with the discrete-time queueing approach, and $\sigma_{W,Sim}^{95}$ the 95% percentile of waiting time, obtained with simulation.

In the following, we introduce the methodologies used for the computation of point and interval estimates of decomposition error and briefly describe the empirical evaluation criteria, the simulation model, and our design of experiments.

4.3.1 Point and interval estimates

We use Ordinary Least Square (OLS) multiple linear regression to compute point estimates, and quantile regression to compute interval estimates for decomposition error. The methodological background on OLS regression can be found e.g. in Sen et al. (1990). Quantile regression aims at the estimation of conditional quantile functions-models in which quantiles (percentiles) of the conditional distribution of the dependent variable are expressed as functions of observed covariates (Koenker and Bassett 1978, Koenker and Hallock 2001). Unlike OLS which is used to compute the conditional mean of the dependent variable, quantile regression can be used to explain the determinants of the dependent variable at any point of the pmf of the dependent variable.

The dependent variables of the regression models in Study I and Study II are $\Delta(E)$ and $\Delta(\sigma)$, respectively. In both studies, we consider the same sample of M observations for the estimation of decomposition error. To help simplify the notations introduced in the following, we do not differentiate between both studies, but instead set $y_m = \Delta(E)$ in Study I, and $y_m = \Delta(\sigma)$ in Study II for a given data point m. The observations include $\boldsymbol{y}$ and $\boldsymbol{X}$, where $\boldsymbol{y}$ denotes the M-vector of decomposition error, and $\boldsymbol{X}$ is the $(N \times K)$ design matrix of the independent variables, with $K-1$ dependent (explanatory) variables.

Point estimates for decomposition error are computed with the well known formula for multiple linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \tag{4.7}$$

where ε is the M-vector of the random error terms of the regression model. The estimates $\hat{\boldsymbol{\beta}}$ for problem (4.7) are found by minimising the sum of squares residuals

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^K} \sum_{n=1}^{N} \left(y_m - \boldsymbol{x}_n^{\mathsf{T}} \boldsymbol{\beta} \right)^2. \tag{4.8}$$

In contrast to OLS, quantile regression finds the estimates $\hat{\boldsymbol{\beta}}(\tau)$ for a given quantile $\tau \in (0,1)$ by minimising the weighted sum of the absolute deviations

$$\hat{\boldsymbol{\beta}}(\tau) = \min_{\boldsymbol{\beta}(\tau) \in \mathbb{R}^K} \sum_{n=1}^{N} \Big| y_m - \boldsymbol{x}_n^{\intercal} \boldsymbol{\beta}(\tau) \Big| \omega_n, \tag{4.9}$$

where the weight ω_n is defined as

$$\omega_n = \begin{cases} 2\tau & y_m - \boldsymbol{x_n}^{\intercal} \boldsymbol{\beta}(\tau) > 0, \\ 2 - 2\tau & \text{otherwise.} \end{cases} \tag{4.10}$$

The quantile regression estimates $\hat{\boldsymbol{\beta}}(\tau)$ in problem (4.9) can be computed very efficiently by linear programming methods. In this chapter, we use the modified version of Barrodale and Roberts algorithm (Koenker and d'Orey 1987, 1994) to calculate the quantile regression estimates.

We always consider the quantile regression models in pairs, so that they form the upper and lower endpoints of the 90%, 95% or 99% confidence interval (CI) of decomposition error, respectively. Consequently, we fit quantile regression models $Q(\tau)$ for the pairs of $\tau = .05$ and $\tau = .95$ for the 90% CI, $\tau = .025$ and $\tau = .975$ for the 95% CI, and $\tau = .005$ and $\tau = .995$ for the 99% CI.

4.3.2 Goodness of fit criteria and likelihood ratio tests

To evaluate the accuracy of the fitted OLS models, we are interested in the empirical distribution of the error term $\boldsymbol{\varepsilon}$ in problem (4.7). A preliminary evaluation of the data set shows that the Gauss-Markov conditions (Sen et al. 1990), and especially $E(\boldsymbol{\varepsilon}) = 0$, hold for our data set. Consequently, mean error measurements for the cumulated error terms of $\boldsymbol{\varepsilon}$ (such as *MSE* and *RMSE*) will be (nearly) zero and therefore not meaningful for interpretation. Instead, we evaluate the absolute values $|\varepsilon_n|, n \in N$ and denote $|\varepsilon_n|$ as *forecasting error (FE)* for observation m. To arrive at the determination of the accuracy of the OLS model, we compute the relative frequency distribution function of *FE* for all observations in ε. Interpreting the relative frequency distribution of *FE*, the higher the percentage of small values, the better the model fits the data and thus the higher the accuracy of the model.

The goodness of fit criterion of quantile regression is calculated with the algorithm by Koenker and Machado (1999). Analogous to the conventional R^2 statistic of OLS regression, we call it Pseudo R^2. Let $\hat{\boldsymbol{\beta}}(\tau)$ denote the minimiser of problem (4.9), and $\hat{V}(\tau)$ the error sum of the conditional quantile function. Further, let $\tilde{V}(\tau)$ denote the error sum of the corresponding conditional quantile function, that is restricted to only consider the intercept parameter of $\hat{\boldsymbol{\beta}}(\tau)$. Conventionally, the goodness of fit criterion is defined as

$$R^2_{Pseudo}(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau). \tag{4.11}$$

Note that Pseudo R^2 is not comparable to the standard coefficient of determination R^2 although it lies between 0 and 1. It is only useful for the comparison between quantile regression models since it is based on the weighted sum of absolute residuals, while R^2 is based on residual variance. Finally, it should be noted that Pseudo R^2 may be a skewed measure as it is not corrected by the degrees of freedom. However, a definition for the goodness of fit that follows the concept of Adjusted R^2 known from OLS regression is not available for quantile regression analyses.

We use likelihood ratio tests to test the overall significance of the OLS regression models (Sen et al. 1990). We are interested in testing whether all the independent variables have any effect on decomposition error and test the general linear hypothesis

$$H : \boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{\gamma} = \boldsymbol{0}, \tag{4.12}$$

where $\boldsymbol{C}$ is a $(M \times K)$ matrix of rank $M < K$ and $\boldsymbol{\gamma}$ is a M-vector. Note that hypothesis (4.12) allows us to test the overall significance of the OLS model, where

$$H : \beta_1 = 0, \beta_2 = 0, ..., \beta_{(K-1)} = 0, \tag{4.13}$$

as well as the significance of elected independent variables (so-called nested models), where

$$H : \beta_m = \gamma_m, \tag{4.14}$$

for arbitrary values of m and γ_m. Hypothesis (4.12) is rejected if

$$\frac{M^{-1}(\boldsymbol{Cb} - \boldsymbol{\gamma})^\intercal[\boldsymbol{C}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{C}^\intercal]^{-1}(\boldsymbol{Cb} - \boldsymbol{\gamma})}{s^2} \geq F_{M,N-K-1,\alpha}, \tag{4.15}$$

where $F_{M,N-K-1,\alpha}$ is the upper α-percent point of the F-distribution with $(M, N-K-1)$ degrees of freedom,

$$\begin{aligned} \boldsymbol{b} &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}, \text{and} \\ s^2 &= (N-K-1)^{-1}\boldsymbol{y}^\intercal[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal]\boldsymbol{y}. \end{aligned} \tag{4.16}$$

We report the test statistic (4.15) as well as the p-value of the hypothesis test, which is the probability of observing a value of F larger than the one observed under H with degrees of freedom $(M, N-K-1)$ and significance level α. Generally speaking, when the test statistic is large, and the p-value is small, we can safely reject H and conclude that the OLS model provides a better fit to the data than a model which contains no independent variables (hypothesis (4.13)) or the nested model (hypothesis (4.14)).

4.3.3 Simulation model

We use a discrete-event simulation model of a tandem queue to obtain the waiting time distribution at the downstream station. Each simulation run is composed of 50 replications with 10,000,000 simulated time steps each. In each simulation run, the first 100,000 time steps are discarded. The observed width of the 95%-CI of the expected value of waiting time is 0.0286, which is less than 0.5% of the average simulated waiting time. Therefore, the performance metrics obtained with the simulation model are – despite being prone to some variance – valid estimates for the waiting time.

4.3.4 Design of experiments

Each tandem queue is parameterised with rate and variability parameters of the external arrival stream and the service processes in both queueing systems. For the sake of conciseness, we limit ourselves to experiments where the arrival process at the first queue is Poisson, and the service times at both queues are gamma-distributed. Given its flexibility, the gamma distribution allows for the modelling of a wide range of dispersion and is therefore well suited to represent the stochastic behaviour of the service process. Further, it is well known that the exponential distribution is a special case of the gamma distribution when the scv-value equals 1. We first consider tandem queues where the utilisation parameters at the upstream and the downstream queue are equal. This allows us to define a generic utilisation parameter ρ for the tandem queue, $\rho = \rho^U = \rho^D$. A relaxation of this assumption will be discussed in Section 4.5.

Based on these conditions, we parameterise each tandem queue with four parameters, the external arrival rate, the service rate, and the variability parameters of both service processes. We use the algorithm described in Dupuy et al. (2015) to generate 1,166 data points in a four-dimensional space-filling latin hypercube design. The expected values of the external inter-arrival and the service times are independently randomly selected from the interval [1.0, 30.0]. The variability

Table 4.2: Summary statistics of the IVs and flow parameters in the training data set.

	Mean	STD	Min	Max
ρ	0.59	0.24	0.06	0.99
$scv(B^U)$	1.30	0.80	0.10	2.96
$scv(B^D)$	1.45	0.79	0.10	2.95
$scv(A^D)$	1.18	0.41	0.18	2.79
$E(B)$	12.32	6.35	1.23	29.00
$E(A^D)$	21.22	6.27	2.72	30.00

parameters of the service time distributions are independently randomly selected from the interval [0.1, 3.0]. We then use the closed-form expressions for the shape and scale parameters of the gamma distribution (cf. equation 4.2) to compute the probability distributions for the service and external arrival processes, leading to a final data set of 1,166 data points. Following the rule of thumb established by Jones et al. (1998) and investigated by Loeppky et al. (2009), this sample size is considered sufficient.

We define the utilisation of the tandem queue ρ, the variability parameters of both service processes $scv(B^U)$ and $scv(B^D)$, and the variability of the arrival process at the downstream queueing system $scv(A^D)$ as independent variables (IVs) of the regression models. We partition the data set into two subsets, the *training data set* which consists of 932 randomly chosen data points, and the *test data set* which consist of the remaining 234 data points. Table 4.2 provides the summarising statistics for the IVs in the training data set and the flow parameters for the tandem queue. Note that the expected values for the service processes at the upstream and the downstream queue are equal, and thus, we list $E(B)$ for both queues. We normalise the IVs of both subsets with the mean- and std-values listed in Table 4.2. The data sets are accessible in a repository (Jacobi and Furmans 2022a) and described in detail in the accompanied data article (Jacobi and Furmans 2022c).

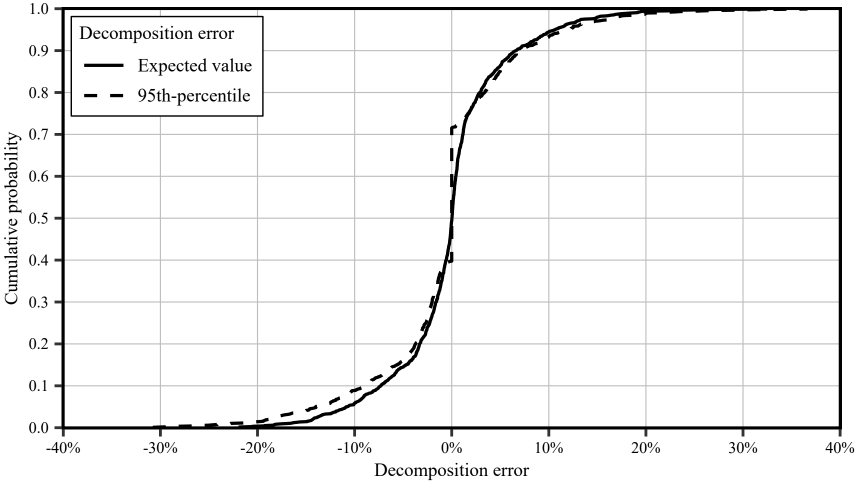

Figure 4.2: Empirical cumulative distribution functions of the decomposition error of waiting time regarding the expected value and 95% percentile.

4.4 Results

We first consider the distribution of decomposition error in the overall data set. The empirical cumulative distribution of decomposition error reveals that both, positive (meaning that discrete-time queueing theory underestimates the waiting time) and negative errors (overestimation of the waiting time) are found. We find the relative errors in the range of -21.9% and 32.5% (referring to Study I) and -30.8% and 36.7% (referring to Study II). The mean absolute values of decomposition error equal 3.93% and 4.51% regarding the expected value and the 95% percentile of waiting time, respectively (see Figure 4.2).

4.4.1 Study I: Expected value of waiting time

The OLS regression coefficients for Study I are presented in Table 4.3. Recall that in Study I, the dependent variable is $\Delta(E)$, cf. equation (4.5). The OLS regression analysis is found to be statistically significant ($F(10, 921) = 2123$, $p < .001$), explaining the majority of the variance of the relative error of the expected value of waiting time ($R^2_{Adj.} = 0.958$). The ANOVA reveals all direct and the majority of the interaction effects to be statistically significant. Since the non-significant coefficient is small, we did not find evidence for the regression model to perform significantly better without incorporating this interaction ($F(921, 922) = 1.234$, $p = .267$). We identify the service process variability at the upstream queueing system and the arrival process variability at the downstream queueing system, as well as the utilisation as major impact factors. Despite being statistically significant, the effect of the variability of the service process at the downstream queueing system is found to be a minor influencing factor.

The Pseudo R^2 of each quantile regression model is well above 0.8. All quantile regression equations show similar patterns of changes in coefficient values as the OLS regression. We find the majority of direct and interaction effects to be statistically significant. As in the OLS regression, the interaction effect between the service process variability (at the upstream queueing system) and the utilisation is found to be non-significant among each model. While the absolute sizes of the coefficients for most factors vary little across the equations, it should be noted that the weights of the service process variability at the upstream queueing system, and the arrival process variability at the downstream queueing system rise with increasing quantile.

Table 4.3: OLS and quantile regression estimates in Study I.

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
const.	0.0048***	-0.0068***	-0.0065***	-0.0039***	0.0228***	0.0309***	0.0361***
	0.0007	0.0012	0.0013	0.0011	0.0035	0.0034	0.0036
$scv(B^U)$	-0.0668***	-0.0460***	-0.0471***	-0.0551***	-0.0713***	-0.0770***	-0.0768***
	0.0017	0.0033	0.0046	0.0037	0.0084	0.0072	0.0074
$scv(B^D)$	-0.0039***	-0.0020***	-0.0011*	-0.0013**	-0.0081***	-0.0093***	-0.0076**
	0.0005	0.0006	0.0005	0.0005	0.0015	0.0015	0.0024
$scv(A^D)$	0.0591***	0.0437***	0.0424***	0.0530***	0.0585***	0.0676***	0.0731***
	0.0027	0.0056	0.0070	0.0058	0.0129	0.0111	0.0108
ρ	-0.0325***	-0.0306***	-0.0301***	-0.0311***	-0.0300***	-0.0291***	-0.0283***
	0.0009	0.0014	0.0017	0.0015	0.0042	0.0041	0.0047
$scv(B^U)$	-0.0048***	-0.0087***	-0.0081***	-0.0064***	-0.0057***	-0.0052***	-0.0043*
$\times\ scv(B^D)$	0.0008	0.0013	0.0016	0.0013	0.0013	0.0011	0.0019
$scv(B^U)$	0.0194***	0.0128***	0.0133***	0.0128***	0.0241***	0.0229***	0.0291***
$\times\ scv(A^D)$	0.0005	0.0008	0.0007	0.0007	0.0033	0.0031	0.0054
$scv(B^U)$	-0.0441***	-0.0293***	-0.0292***	-0.0366***	-0.0475***	-0.0488***	-0.0475***
$\times\ \rho$	0.0011	0.0027	0.0033	0.0028	0.0052	0.0046	0.0058

OLS and quantile regression estimates in Study I (continued).

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
$scv(B^D)$	0.0080***	0.0049***	0.0057***	0.0052***	0.0112***	0.0107***	0.0058
× $scv(A^D)$	0.0008	0.0014	0.0016	0.0012	0.0027	0.0025	0.0043
$scv(B^D)$	-0.0006	0.0008	0.0010	0.0017**	-0.0005	-0.0014	0.0006
× ρ	0.0005	0.0007	0.0008	0.0008	0.0018	0.0017	0.0030
$scv(A^D)$	-0.0405***	-0.0411***	-0.0396***	-0.0391***	-0.0407***	-0.0450***	-0.0518***
× ρ	0.0011	0.0023	0.0023	0.0019	0.0041	0.0043	0.0051
Adj. / Ps. R^2	0.958	0.917	0.902	0.895	0.829	0.843	0.872

Notes: Standardised regression coefficients with standard errors listed below. The standard errors of quantile regression estimates are based on 100 bootstrapping replications. The sample is training data set with sample size 932.

* $p < .1$

** $p < .05$

*** $p < .001$

Table 4.4: OLS and quantile regression estimates in Study II.

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
const.	0.0052***	-0.0169***	-0.0107***	-0.0093***	0.0289***	0.0349***	0.0838***
	0.0012	0.0023	0.0022	0.0020	0.0033	0.0064	0.0163
$scv(B^U)$	-0.0735***	-0.0671***	-0.0631***	-0.0547***	-0.0847***	-0.0931***	-0.1010***
	0.0028	0.0082	0.0089	0.0074	0.0109	0.0129	0.0146
$scv(B^D)$	-0.0046***	-0.0036	0.0002	0.0004	-0.0093***	-0.0120***	-0.0413***
	0.0008	0.0031	0.0021	0.0012	0.0019	0.0044	0.0097
$scv(A^D)$	0.0680***	0.0916***	0.0776***	0.0554***	0.0735***	0.0777***	0.0591***
	0.0046	0.0146	0.0163	0.0116	0.0169	0.0198	0.0236
ρ	-0.0381***	-0.0351***	-0.0360***	-0.0321***	-0.0439***	-0.0471***	-0.0636***
	0.0015	0.0029	0.0031	0.0022	0.0052	0.0071	0.0120
$scv(B^U)$	-0.0038***	-0.0134*	-0.0052**	-0.0066***	-0.0095**	-0.0035	0.0254**
$\times\ scv(B^D)$	0.0013	0.0062	0.0019	0.0014	0.0032	0.0069	0.0114
$scv(B^U)$	0.0189***	0.0078***	0.0091***	0.0113***	0.0241***	0.0322***	0.0263***
$\times\ scv(A^D)$	0.0008	0.0012	0.0015	0.0013	0.0030	0.0050	0.0073
$scv(B^U)$	-0.0476***	-0.0291***	-0.0378***	-0.0338***	-0.0649***	-0.0676***	-0.0342***
$\times\ \rho$	0.0019	0.0086	0.0067	0.0053	0.0072	0.0097	0.0128

OLS and quantile regression estimates in Study II (continued).

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
$scv(B^D)$	0.0092***	0.0089	0.0005	0.0049*	0.0148***	0.0090	-0.0046
$\times\ scv(A^D)$	0.0014	0.0063	0.0032	0.0019	0.0039	0.0058	0.0081
$scv(B^D)$	0.0008	0.0109***	0.0056*	0.0030*	0.0038	0.0065	0.0218***
$\times\ \rho$	0.0008	0.0028	0.0022	0.0015	0.0026	0.0045	0.0071
$scv(A^D)$	-0.0522***	-0.0740***	-0.0597***	-0.0520***	-0.0468***	-0.0476***	-0.0503***
$\times\ \rho$	0.0019	0.0061	0.0055	0.0029	0.0048	0.0060	0.0110
Adj. / Ps. R^2	0.920	0.809	0.798	0.807	0.698	0.681	0.635

Notes: Standardised regression coefficients with standard errors listed below. The standard errors of quantile regression estimates are based on 100 bootstrapping replications. The sample is training data set with sample size 932.

* $p < .1$

** $p < .05$

*** $p < .001$

4.4.2 Study II: 95% percentile of waiting time

The regression coefficients for Study II are presented in Table 4.4. In Study II, the dependent variable is $\Delta(\sigma)$, cf. equation (4.6). We find a statistically significant OLS regression equation ($F(10, 921) = 1064$, $p < .001$), which explains the majority of the variance ($R^2_{Adj.} = 0.920$) of decomposition error regarding the 95% percentile of waiting time. The impact patterns of the interaction effects are the same as in Study I. Again, we did not find evidence for the OLS estimate to better perform without incorporating the non-significant interaction effect between the service process variability and utilisation ($F(921, 922) = 0.917$, $p = .339$). Analogous to Study I, the service process variability (at the upstream queueing system), the arrival process variability (downstream queueing system), and the utilisation are found to be the major direct effects. Despite being statistically significant, the service process variability at the downstream queueing system is a minor impact factor.

The Pseudo R^2 of all quantile regression models is well above 0.6. Except for the service process variability at the downstream queueing system, which is non-significant for the models with $\tau \leqslant .05$, all direct effects are found to be statistically significant among each regression model. The majority of interaction coefficients is found to be significant or marginally significant. However, we did find non-significant coefficients among the interaction effect of the service process variability and the arrival process variability (both at the downstream queueing system), as well as in the $Q(.975)$ model. As in Study I, the absolute sizes of coefficients vary little for most factors across the equations. However, the weight of the utilisation increases by rising quantiles, while (in contrast to Study I) the weight of the arrival process variability decreases.

4.4.3 Performance of point and interval estimates

The accuracy of the point estimates is presented in Tables 4.5 and 4.6. For the majority of data points, we find an absolute error of the OLS predictions

Table 4.5: Performance of point estimates: Relative frequency distributions and means of forecasting error for training and test data in Study I.

FE	Train	Test	Test (a)	Test (b)
[0.000, 0.005]	40.5%	37.2%	41.9%	8.7%
(0.005, 0.010]	30.5%	30.8%	34.6%	13.0%
(0.010, 0.020]	20.5%	22.6%	18.4%	34.8%
(0.020, 0.050]	8.0%	9.0%	5.1%	39.1%
(0.050, ∞)	0.5%	0.4%	0.0%	4.4%
Mean	0.0087	0.0092	0.0073	0.0210

Notes: Subsets (a) and (b) denote the subsets of test data with absolute decomposition error smaller than 3% and above 10%. The sample sizes are 136 and 23.

of less than 1 percentage point from the simulated value. The mean absolute forecasting errors are less than 1 percentage point in Study I and only slightly above 1 percentage point in Study II. In both studies, this accuracy is achieved for the training and the test data set, which indicates that our OLS prediction approach is robust to overfitting.

Despite the minor mean errors, the results suggest that the accuracy of point estimates decreases when forecasting severe values of decomposition error. To investigate this effect, we examine the subsets of test data with minor decomposition errors, that is, all data points with absolute decomposition errors smaller than 3% (in the following referred to as subset (a)), and with severe decomposition errors, that is, all data points with absolute decomposition errors above 10% (subset (b)). The sample sizes of subsets (a) and (b) are 136 and 23 in Study I, and 127 and 33 in Study II, respectively. The relative frequency distributions of *FE* and its mean errors (cf. Tables 4.5 and 4.6) suggest that subset (a) is forcasted with significantly higher accuracy than the data points from subset (b) in both studies. Further, the share of data points that is forecasted with a *FE* greater than 0.05 is significantly higher in subset (b). However, it cannot be concluded that data points with severe absolute decomposition errors are frequently predicted with minor accuracy. In the test data from Study I, we find 96% of the data points

Table 4.6: Performance of point estimates: Relative frequency distributions and means of forecasting error for training and test data in Study II.

FE	Train	Test	Test (a)	Test (b)
[0.000, 0.005]	40.8%	30.8%	34.6%	11.8%
(0.005, 0.010]	36.5%	29.5%	33.1%	0.0%
(0.010, 0.020]	7.4%	23.9%	19.7%	35.3%
(0.020, 0.050]	12.7%	12.8%	11.8%	41.1%
(0.050, ∞)	2.6%	3.0%	0.8%	11.8%
Mean	0.0118	0.0117	0.0095	0.0260

Notes: Subsets (a) and (b) denote the subsets of test data with absolute decomposition error smaller than 3% and above 10%. The sample sizes are 127 and 33.

with an absolute decomposition greater than 10% to be forecasted with a *FE* less than 0.05 (in Study II the share is 89%).

Interval estimation compensates for this effect. By providing the 90%, 95%, and 99% confidence intervals, we evaluate the precision of the point estimates. Table 4.7 presents the performance of the interval estimates for Study I and Study II, listing the mean interval lengths and the actual shares of decomposition errors included in the respective confidence intervals. As expected, the average interval lengths increase with rising confidence in finding a data point in the corresponding interval. In both studies, the average interval lengths differ only marginally between training and test data which indicates that the approach of interval estimation is robust to over-fitting. In the training data set, the confidence intervals contain exactly the respective share of values they were determined for. These shares are only slightly undermined for the test data.

The interval estimates are designed to indicate uncertainty in the forecast of point estimates. The results are presented in Table 4.7. In subset (a), the precision of interval estimations increases, compared to the entire test data set. This is indicated by the narrower intervals, as well as the high shares of values that are included in the respective intervals (which is especially to be emphasised for

Table 4.7: Performance of interval estimates: Mean lengths and actual share of values for confidence intervals (CI) in Study I and Study II, based on quantile regression models.

	90% CI		95% CI		99% CI	
	Length	Share	Length	Share	Length	Share
Study I Train	0.0348	90.58%	0.0416	95.07%	0.0506	98.93%
Study I Test	0.0345	86.32%	0.0415	91.45%	0.0509	94.02%
Study I Test (a)	0.0271	87.50%	0.0318	93.38%	0.0374	94.85%
Study I Test (b)	0.0692	78.26%	0.0810	82.61%	0.1006	91.30%
Study II Train	0.0467	90.26%	0.0661	95.50%	0.1343	98.82%
Study II Test	0.0473	91.45%	0.0658	92.74%	0.1303	95.73%
Study II Test (a)	0.0432	92.91%	0.0594	96.85%	0.1442	99.21%
Study II Test (b)	0.0689	81.82%	0.1006	81.82%	0.1281	84.85%

Notes: Subsets (a) and (b) denote the subsets of test data with absolute decomposition error smaller than 3% and above 10%. The sample sizes are 136 and 23 (Study I), and 127 and 33 (Study II).

Study II). As discussed above, in subset (b), the forecast uncertainty of the point estimates increases, which is indicated by longer mean intervals and a smaller share of values contained in the intervals.

We conclude that minor decomposition errors are predicted with satisfactory point estimation accuracy and great precision. Predicting severe decomposition errors is subject to uncertainty: the absolute error of the point estimate might be considerable, which is indicated by large confidence intervals. By combining the methods, the regression models satisfy both, the aspect of an accurate point estimation forecast, as well as the quantification of its uncertainty.

4.5 Extensions: Bottlenecks and longer lines

The investigations of the heavy-traffic bottleneck phenomenon in open queuing systems (Suresh and Whitt 1990) suggest that the performance of bottleneck downstream queues is strongly related to the variability of the non-renewal arrival process variability, which impacts the approximation quality of decomposition methods. Therefore, we extend our analyses to tandem queues and longer lines with bottlenecks. As Suresh and Whitt (1990) mention, in a narrower sense, the bottleneck is the queue with the highest traffic intensity. However, increasing the traffic intensity of a queue by only a small amount may shift the bottleneck position. Therefore, it is intuitive to state that either of the queues is the bottleneck if it's utilisation is substantially greater than some ϵ, $|\rho^U - \rho^D| > \epsilon$.

We create a further data set containing 969 data points, following the procedure described in section 4.3.4, but with the relaxation that the expected values of service times are now independent. We choose $\epsilon = 0.1$ and find 403 data points where the downstream queue is the bottleneck. We use OLS and quantile regression to identify the major and minor effects on decomposition error in bottleneck queues. The coefficients of the regression analyses, where the dependent variables are $\Delta(E)$ (Study I), and $\Delta(\sigma)$ (Study II) are provided in the accompanied data article and in Appendix A. We find the previously identified major and minor effects on decomposition error to apply in this analysis, as well. However, the empirical distributions of the decomposition error show that the approximation quality of the decomposition approach depends significantly on which of the queues is the bottleneck. In the case of similar traffic intensities, we find mean absolute values of decomposition error to be 5.45% (6.51%) for the expected value (95% percentile) of waiting time, which is in line with the expectations of previous examinations. When the bottleneck is downstream, the mean absolute values of decomposition error regarding the expected value (95% percentile) of waiting time equal 4.87% (5.50%). In contrast, when the bottleneck is upstream, we find mean absolute values of decomposition error of 1.36% (1.46%) for the expected value (95% percentile) of waiting time.

Similar results are observed in longer lines. We investigate a set of lines with i queues in series, where i equals 3, 5, 7, and 9. For each line length i, we evaluate 250 data points. The utilisation parameters of the first $i - 1$ queues are equal, and the last queue in each case is the bottleneck. Table 4.8 shows the mean absolute decomposition errors for the expected value (Study I), and the 95%

Table 4.8: Absolute mean decomposition errors for Study I and Study II in longer lines.

	Length 3		Length 5		Length 7		Length 9	
Q.	Study I	Study II	Study I	Study II	Study I	Study II	Study I	Study II
1	0.23	0.39	0.24	0.26	0.24	0.22	0.17	0.16
2	2.43	2.43	2.16	2.32	2.43	2.55	2.18	2.39
3	8.94	10.94	2.56	2.38	2.61	2.39	2.72	2.67
4			2.50	2.45	3.12	3.05	2.82	2.83
5			9.68	12.53	3.04	3.22	2.86	3.11
6					3.24	3.36	3.49	3.28
7					9.24	10.67	3.54	4.10
8							3.48	3.03
9							10.91	12.91

Note: Values for decomposition error in percent.

percentile (Study II) of waiting time. It can be clearly seen that the last queues are prone to significant decomposition errors with 9.69% on average in Study I, and 11.67% on average in Study II. This is significantly more than the decomposition errors for the intermediate queues which are 2.82% on average in Study I, and 2.85% on average in Study II. The results confirm the long-range variability effect formulated by Suresh and Whitt (1990), that states that variability in the external arrival stream or the service times can have a dramatic effect on a downstream queue with a much higher traffic intensity.

4.6 Main drivers of decomposition error

From the analyses of decomposition techniques in the continuous-time domain, it is well known that utilisation and variability parameters for arrival and service processes are significant for the approximation quality of congestion measures. Based on the regression coefficients, we identify utilisation and arrival process

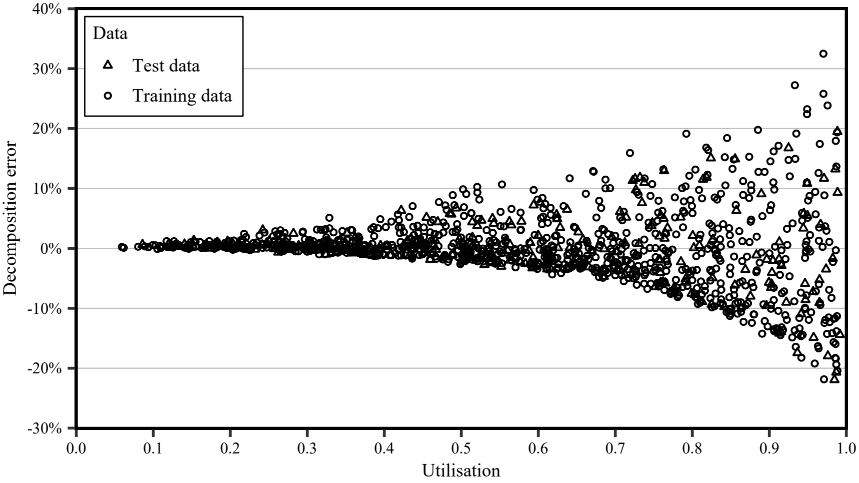

Figure 4.3: Utilisation is the main driver for decomposition error.

variability as major impact factors on decomposition error. Service process variability was revealed as a minor impact factor.

Utilisation is found to be the main driver for decomposition error (cf. Figure 4.3): In low-traffic queueing systems, the mean absolute decomposition error is significantly lower than the mean absolute errors in the entire data set. Severe absolute decomposition errors are only observed in heavy-traffic systems. In tandem queues with bottlenecks, we find the decomposition error to be significantly higher when the bottleneck is downstream. This leads to the conclusion that downstream bottlenecks are analysed with limited accuracy, which should be of particular interest since the performance evaluation of bottlenecks is obviously particularly critical. The arrival process variability determines the tendency (that is, overestimation or underestimation of the waiting time) of the decomposition technique. For scv-values of the arrival process at the downstream queue lower than 1.0, the decomposition approach underestimates waiting time. Overestimation of waiting time occurs for scv-values of the downstream arrival process greater than 1.0. Variability of the service process is a minor impact factor. This is indicated by the fact that when the arrival process at the downstream queue is Poisson, we did not find considerable decomposition errors, regardless of the utilisation of the queueing system nor the scv-value of the service process.

We conclude the discrete-time decomposition approach to analyse low traffic queueing systems with high accuracy. In heavy-traffic systems, the approximation quality depends on the arrival process variability. The analysis of queueing systems with highly volatile as well as deterministic arrival processes is prone to considerable decomposition errors. When the arrival process is Poisson, the decomposition approach yields high accuracy, regardless of the service process variability.

4.7 Chapter conclusion

In this chapter, we analyse the approximation quality of the renewal decomposition approach in the discrete-time tandem queue with Poisson arrivals. In our design of experiments, we combine variability parameters of the gamma-distributed service times in the interval $(0.1, 3.0)$ with the flow parameters to cover the utilisation in the interval $(0.3, 0.99)$. We compute the expected value (Study I) and the 95% percentile (Study II) of waiting time using the renewal decomposition approach and simulation, and define decomposition error as the relative error between both measures, respectively. We deploy the variability parameters of the service time distributions, the variability of the connecting stream, and utilisation as independent variables for the point and interval estimates of decomposition error. The point estimates are based on multiple linear regression, the interval estimates are based on quantile regression. Both estimation methods are applied for Study I and Study II, respectively. Using test data, we demonstrate that the regression models provide accurate forecasts and precise confidence intervals for decomposition error. Further, we use the ANOVA of the models to reveal major influencing factors on the renewal approximation quality: We find utilisation to be the main driver for decomposition error since in low-traffic queueing systems the mean absolute decomposition error is significantly lower than the mean absolute errors in the entire data set. Severe absolute decomposition errors are only observed in heavy-traffic systems. Finally, we find that the downstream arrival process variability determines whether the decomposition approach overestimates or underestimates the waiting time.

5 On the output dynamics of the discrete-time M/G/1-queue

This chapter is based on the paper "On the output dynamics of the discrete-time $M/G/1$-queue" which is under review at *Annals of Operations Research* (Jacobi 2023b). The software that has been used to compute the results presented here is available in the *KITopen Repository* (Jacobi 2023a).

The author of this thesis was responsible for the conceptualisation, methodology, software programming, validation, formal analysis, writing, and visualisation of the research presented in this chapter.

Chapter abstract

The departure process of the $M/G/1$-queue is a point process that results from the interaction of Poisson arrivals with the renewal service process. Consecutive departure instances are sequentially dependent as the inter-departure time of one customer depends on the state that the previous customer left behind in the system. However, decomposition methods often treat the departure point process as a renewal process, which causes approximation errors in the analysis of downstream queues. Therefore, in this chapter, we investigate the ϕ-lag auto-correlation of the departure process of the discrete-time $M/G/1$-queue to find situations in which the renewal assumption does not hold. To this end, we model the $M/G/1$-queue as a discrete-time Markov chain and derive the serial covariance function of the departure process for two inter-departure times that are $\phi > 0$ departure

instances apart. Numerical results show that auto-correlation is positive and a non-monotone function of the utilisation parameter. Short (long) inter-departure times are therefore likely to be followed by another short (long) inter-departure time. As this effect is not considered in an i.i.d. sampling of the inter-departure time distribution, the renewal decomposition approach is approximate for the analysis of downstream queues.

5.1 Introduction and problem description

We consider the $M/G/1$-queue in the discrete-time domain, that is, the time axis is divided into time slots of equal length t_{inc}, and events (such as the arrival, start of service, or departure of a customer) are only observed at slot boundaries. Consequently, service, inter-arrival, and inter-departure times are integer multiples of t_{inc}, and the probability for any discrete random variable Z is described by

$$P(Z = i \cdot t_{inc}) = z_i \quad \forall i \geq 0. \tag{5.1}$$

For convenience, we will refrain from including the slot parameter t_{inc} in the formula described in this chapter.

In the upstream $M/G/1$-queue, the service time B^U is an i.i.d. discrete random variable, and the arrival of customers is characterised by a Poisson process with mean λ. Let the random variable C describe the number of customers that arrive at a slot boundary with probability function

$$P(C = c) = \frac{e^{-\lambda} \cdot \lambda^c}{c!}. \tag{5.2}$$

Then, the inter-arrival time $A^U \sim Geo(q)$ is described by a geometric distribution with parameter $q = 1 - e^{-\lambda}$.

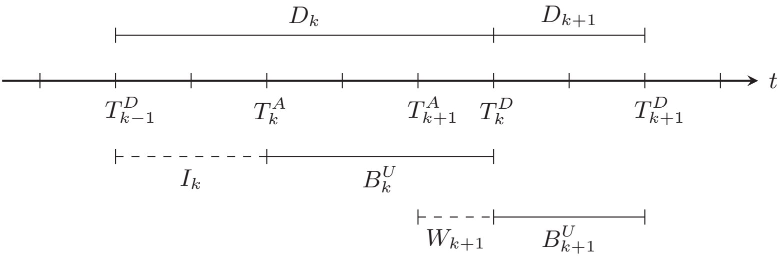

Figure 5.1: Discrete-time modelling of arrival and service processes in the $M/G/1$-queue.

The following example illustrates the cases that arise when computing the output dynamics of the $M/G/1$-queue. We observe the queue only in time instants immediately after the departure of a customer. For these instants, let the random variable N_k denote the number of customers left in the queue immediately after the departure of customer k. We assume that customer k arrives at the system at time T_k^A, and upon arrival, customer k waits $W_k \geq 0$ time units for the service to begin. After being served with service time $B_k^U > 0$, customer k leaves the queueing system at time instance $T_k^D > T_k^A$ (see Figure 5.1). The dynamics of this departure process are determined by two cases. First, we assume that after the departure of customer $k-1$, the system was starving, that is $N_{k-1} = 0$. The inter-departure time $D_k = T_k^D - T_{k-1}^D$ of customer k therefore equals the sum of the idle time $I_k = T_k^A - T_{k-1}^D$ and the customer's service time B_k^U. In the second case, we assume that the system was occupied by at least one customer after the departure of customer $k-1$. In this case, the inter-departure time D_k equals the service time of customer k. In summary, we determine

$$D_k = \begin{cases} I_k + B_k^U & N_{k-1} = 0, \\ B_k^U & N_{k-1} > 0. \end{cases} \tag{5.3}$$

Equation (5.3) shows that the interaction of the renewal arrival and the renewal service processes generates a non-renewal departure process. Indeed, consecutive departure instances are sequentially dependent since the inter-departure time D_k depends on the system state that the previous customer $k-1$ left behind.

It is well known that the departure processes of queuing systems are point processes (Whitt 1981, 1982). However, a point process is usually difficult to deploy for queueing analysis, so it is often approximated by a renewal process. For decomposition methods, the renewal assumption is convenient as it allows for an independent analysis of the queuing systems in a network of queues. Recently, it has been shown that the renewal assumption can result in significant approximation errors when using decomposition methods to compute performance measures in a line of downstream $GI/G/1$-queues (Jacobi and Furmans 2022b). This chapter analyses the sequential dependencies in the departure stream of the discrete-time $M/G/1$-queue and discusses the effects on the analysis of queues that receive this point process as input.

The remainder of this chapter is organised as follows. We review the related literature in Section 5.2. In Section 5.3, we introduce the embedded Markov chain model for the analysis of the $M/G/1$-queue and derive the computation of the ϕ-lag auto-correlation function based on the joint probability distribution of the departure stream. In Section 5.4, we present numerical results for the output dynamics of the discrete-time $M/G/1$-queue and discuss these results in Section 5.5. Section 5.6 concludes this chapter.

5.2 Literature review

The output process dynamics of stochastic systems have been studied in various research fields, e.g. to determine buffer size requirements at a downstream node or switch in an Asynchronous Transfer Mode (ATM) network (Mitchell et al. 1998, Lee et al. 2000) and to study production systems subject to blocking (Hendricks 1992, Hendricks and McClain 1993, Tan and Lagershausen 2017). The output dynamics of queuing models have been of research interest as well, e.g. the analysis of the covariance structure of the departure process for a $M/G/1$-queue with finite waiting space (King 1971) or the special case $M/D/1$-queue (Pack 1975). Reynolds (1975) and Daley (1976) present surveys that discuss the related theory. We detail the papers of particular interest to our study below.

Employing joint departure time distributions, Jenkins (1966a) studied the correlation structure of the departure process for queues with Poisson arrivals and Erlang service times. The results for auto-correlation of the departure stream were presented for lag 1 and 2. However, Jenkins (1966a) reported an extension of the method proposed to be “unwieldy beyond the second order”. In a sequel, Jenkins (1966b) showed how to exploit probability generating functions to relate the joint distribution of the numbers of customers left behind by two successive departing customers and the joint distribution of two customers arriving in two successive departure intervals. Daley (1968) studied the correlation structure of the departure sequence of several queueing systems in the continuous-time domain for, among others, the $M/G/1$-queue. This chapter is closely related to this work since our approach to the computation of the joint departure time distribution relies on the considerations made by Daley (1968).

In conclusion, the literature analysing the output dynamics of queues focuses on the continuous-time domain rather than discrete-time models, as in our case. Our contribution is to efficiently compute the auto-correlation of the departure process for lags $\phi \in [1, 9]$ and to use the output dynamics to explain the approximation quality of downstream performance measures using decomposition.

5.3 Discrete-time Markov chain

We introduce the discrete-time Markov chain for steady-state analysis of the $M/G/1$-queue. As in our introductory example, we observe the system state only at departure instances of a customer. Let N_k denote the state of the Markov chain immediately after the departure of customer k. The stochastic process $\{N_k, k = 1, 2, ...\}$ is an irreducible and aperiodic (see Neuts (1979a) for a proof) embedded discrete-time Markov chain with transition matrix $\mathbf{P}$,

$$\mathbf{P} = \begin{bmatrix} p_0 & p_1 & p_2 & p_3 & p_4 & \cdots \\ p_0 & p_1 & p_2 & p_3 & p_4 & \cdots \\ 0 & p_0 & p_1 & p_2 & p_3 & \cdots \\ 0 & 0 & p_0 & p_1 & p_2 & \cdots \\ 0 & 0 & 0 & p_0 & p_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{5.4}$$

where

$$p_i = \sum_b \frac{e^{-\lambda b} \cdot (\lambda b)^i}{i!} \cdot P(B = b) \quad \forall i \geq 0 \tag{5.5}$$

denotes the probability that during the service time b of a customer, $i \geq 0$ customers arrived at the queue. Recall that the system is only observed immediately after the departure of a customer. This means that in each state of the Markov chain, the queuing system has just completed a service period of length b. If $i > 1$ customers arrived during this service time b, the number of customers increases by $i - 1$. If $i = 1$ customer arrived during service time b, the state remains unchanged. If $i = 0$ customers arrived during service time b, the number of customers decreases by 1 (or – if the system was starving after the last departure – remains zero). Due to this behaviour, the number of customers in a state transition can reduce by a maximum of one. This results in the triangular structure of the transition matrix (5.4).

The stationary distribution π of the Markov chain can be computed if $\rho = \lambda E(B^U) < 1$. In this case, $\mathbf{P}$ is a stochastic matrix and has an invariant probability vector π that satisfies (Neuts 1979a)

$$\begin{aligned} \pi &= \mathbf{P}\pi, \\ \pi\mathbf{e} &= 1. \end{aligned} \tag{5.6}$$

However, the computation of the stationary distribution π is non-trivial since the state space of $\{N_k\}$ and thus the stochastic matrix $\mathbf{P}$ are infinite. Usually, one would truncate the state space at a sufficiently large number of customers, and compute the steady-state vector using the iterative Gauss-Seidel method. However, for large instances, this approach is time-consuming, even on today's powerful computers. Instead of computing the steady-state vector *iteratively*, we use the *recursive* formula that was introduced by Ramaswami (1988). We denote $\pi(r), r \geq 0$ the r-th entry of the steady-state vector. In the r-th recursion step, $\pi(r)$ is computed as follows:

$$\pi(r) = \pi(0) \cdot \Big(1 - \sum_{i=0}^{r-1} p_i\Big) + \sum_{n=1}^{r-1} \pi(n) \cdot \Big(1 - \sum_{i=0}^{r-n} p_i\Big) \quad \forall\, r > n \geq 0 \tag{5.7a}$$

$$\pi(0) = 1 - \sum_{r=1}^{\infty} \pi(r). \tag{5.7b}$$

In equation (5.7a), it can be seen that the computation of the r-th entry in the steady-state vector only relies on the previously computed steady-state probabilities $\pi(n), n < r$, and the transition probabilities p_i, which can be computed for any i using formula (5.5). The first entry of the steady-state vector $\pi(0)$ can be computed either as the inverse of the mean return time to state 0 in the Markov chain (Ramaswami 1980), or using equation (5.7b), once the entries of $\pi(r), r > 0$ have been computed in dependence of $\pi(0)$. We briefly explain the recursion method in the following.

Each recursion step r produces a reduced stochastic $(r + 1) \times (r + 1)$ matrix $\mathbf{P}_r$,

$$\mathbf{P}_r = \begin{bmatrix} p_0 & p_1 & p_2 & p_3 & p_4 & \cdots & p_{r-1} & \sum_{i=r} p_i \\ p_0 & p_1 & p_2 & p_3 & p_4 & \cdots & p_{r-1} & \sum_{i=r} p_i \\ 0 & p_0 & p_1 & p_2 & p_3 & \cdots & p_{r-2} & \sum_{i=r-1} p_i \\ 0 & 0 & p_0 & p_1 & p_2 & \cdots & p_{r-3} & \sum_{i=r-2} p_i \\ 0 & 0 & 0 & p_0 & p_1 & \cdots & p_{r-4} & \sum_{i=r-3} p_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & p_0 & \sum_{i=1} p_i \end{bmatrix}, \tag{5.8}$$

where the transition probabilities $p_i, i < r$ are identical to the entries in the original transition matrix $\mathbf{P}$. The transition probabilities in the last column are the row-sum probabilities $p_i, i \geq r$ in each row i in $\mathbf{P}$. Note these infinite sums can be re-written as follows:

$$\sum_{i=r-n}^{\infty} p_i = 1 - \sum_{i=0}^{r-1-n} p_i \quad \forall r > n \geq 0. \tag{5.9}$$

Given $\mathbf{P}_r$, we compute the entry of the steady-state vector $\pi(r), r > 0$ by finding the last equation from the problem $\pi = \mathbf{P}_r \pi$, i.e. the column sum of the last column in $\mathbf{P}_r$:

$$\begin{aligned}\pi(r) = \pi(0) \cdot \sum_{i=r} p_i \\ + \pi(1) \cdot \sum_{i=r} p_i \\ + \pi(2) \cdot \sum_{i=r-1} p_i \\ + \cdots \\ + \pi(r-1) \cdot \sum_{i=2} p_i \\ + \pi(r) \cdot \sum_{i=1} p_i.\end{aligned} \tag{5.10}$$

Using the relationship in equation (5.9), we can re-arrange equation (5.10) and find the recursive formula (5.7):

$$\begin{aligned}\pi(r) &= \pi(0) \cdot \sum_{i=r}^{\infty} p_i + \sum_{n=1}^{r-1} \pi(n) \cdot \sum_{i=r-n+1}^{\infty} p_i + \pi(r) \cdot (1 - p_0) \\ \Leftrightarrow p_0 \cdot \pi(r) &= \pi(0) \cdot \sum_{i=r}^{\infty} p_i + \sum_{n=1}^{r-1} \pi(n) \cdot \sum_{i=r-n+1}^{\infty} p_i \\ \Leftrightarrow \pi(r) &= \pi(0) \cdot \Big(1 - \sum_{i=0}^{r-1} p_i\Big) + \sum_{n=1}^{r-1} \pi(n) \cdot \Big(1 - \sum_{i=0}^{i-n} p_i\Big).\end{aligned} \tag{5.11}$$

5.3.1 The stationary inter-departure time distribution

Based on the steady-state vector π, we can compute the output dynamics of the $M/G/1$-queue. First, we focus on the stationary inter-departure time distribution. We assume that $\{N_k\}$ is in steady-state, and thus the departure process $\{D_k, k = 1, 2, ...\}$ is stationary. Let the discrete random variable D denote the inter-departure time. We obtain the probability distribution of D based on the considerations made in equation (5.3). Thus, the inter-departure time is equal to the service time, if at least one customer remains in the queue at a departure instant. If the queue starves after a departure instant, the inter-departure time is equal to the sum of the idle time and the service time. We compute the probability $P(D = d)$ as follows:

$$P(D = d) = \begin{cases} \sum_{i=1}^{\infty} \pi(i) \cdot P(B = d) & i > 0, \\ \sum_{b=1}^{d} \pi(0) \cdot P(A = d - b) \cdot P(B = b) & \text{otherwise.} \end{cases} \tag{5.12}$$

Given the fact that the departure process is stationary, the expected value $E(D) < \infty$ and variance $Var(D) > 0$ exist.

5.3.2 The serial covariance of the departure process

We compute the ϕ-lag auto-covariance function $\gamma(\phi)$ of the departure process for two random inter-departure time instances D_k and $D_{k+\phi}$. The auto-covariance function $\gamma(\phi)$ is given by

$$\begin{aligned} \gamma(\phi) &= E\big[(D_k - E(D)) \cdot (D_{k+\phi} - E(D))\big] \\ &= \sum_x \sum_y \big(x - E(D)\big) \cdot \big(y - E(D)\big) \cdot P(D_k = x, D_{k+\phi} = y). \end{aligned} \tag{5.13}$$

In order to find $\gamma(\phi)$, we have to compute the joint inter-departure time distribution $f^{(\phi)}(x, y) = P(D_k = x, D_{k+\phi} = y)$. This problem has been studied in the literature for the $M/G/1$-queue in the continuous-time domain (see Daley (1968), Theorems 6 and 7). The joint inter-departure time probability is composed of

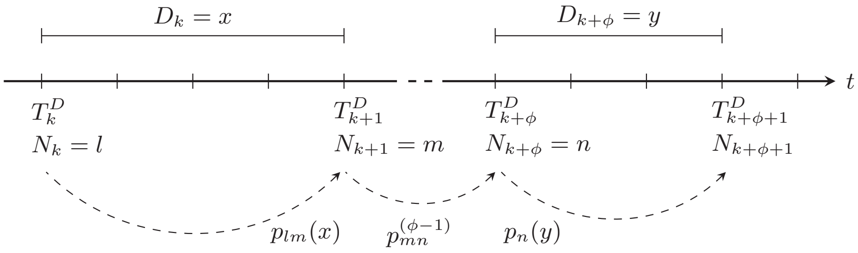

Figure 5.2: Discrete-time modelling of the joint inter-departure time distribution.

three components. Firstly, we compute the probability that the stochastic process transitions from some state l to state m, which produces the first inter-departure time of length x. Secondly, we compute the probability that the Markov chain transitions in $(\phi - 1)$ steps from state m to an arbitrary state n. Thirdly, we compute the probability that in state n, the queueing system outputs the second inter-departure time of length y. Figure 5.2 shows the approach.

We explain the components in detail in the following. First, we define for some states $l, m, n \geq 0$, the lag $\phi > 0$, and the inter-departure times $x, y > 0$:

$$p_{lm}(x) = P(N_{k+1} = m, D_k = x \,|\, N_k = l), \tag{5.14a}$$

$$p_{mn}^{(\phi)} = P(N_{k+\phi} = n \,|\, N_{k+1} = m), \tag{5.14b}$$

$$p_n(y) = P(D_{k+\phi} = y \,|\, N_{k+\phi} = n). \tag{5.14c}$$

The probability $p_{lm}(x)$ in equation (5.14a) is the probability that the stochastic process $\{N_k\}$ transitions from an arbitrary state l to state m, while the inter-departure time of this transition equals x. If the queuing system is not starving after the departure in state l, the probability $p_{lm}(x)$ equals the one-step transition probability from state l to state m, multiplied by the probability that the service time equals x. If the queuing system is starving (i.e., $l = 0$), the one-step transition probability to state m is p_m, and the inter-departure time x is equal to the sum of the service time and the idle time.

In summary, we compute $p_{lm}(x)$ as follows:

$$p_{lm}(x) = \begin{cases} p_{m-l+1} \cdot P(B=x) & m \geq l-1 \geq 0, \\ p_m \cdot \sum_{b=1}^{x} P(A=x-b) \cdot P(B=b) & l=0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.15}$$

The probability $p_{mn}^{(\phi)}$ in equation (5.14b) is the ϕ-step transition probability from an arbitrary state m to state n, which can be observed from the transition probability matrix $\mathbf{P}$. Note that $p_{mn}^{(0)} = \delta_{mn}$, the Kroneker delta (Daley 1968),

$$p_{mn}^{(0)} = \delta_{mn} := \begin{cases} 1 & m=n, \\ 0 & \text{otherwise.} \end{cases} \tag{5.16}$$

The probability $p_n(y)$ in equation (5.14c) is the dependent probability that the stochastic process $\{N_k\}$ is in state n, and the inter-departure time of the next departure instance equals y. Again, for $n > 0$, the inter-departure time y equals the service time, and for $n = 0$, the inter-departure time y is equal to the sum of the service time and the idle time. In summary, we compute $p_n(y)$ as follows:

$$p_n(y) = \begin{cases} P(B=y) & n>0, \\ \sum_{b=1}^{y} P(A=y-b) \cdot P(B=b) & n=0. \end{cases} \tag{5.17}$$

Based on the probabilities in equations (5.14), we find the joint probability

$$\begin{aligned} f_{lmn}^{(\phi)}(x,y) &= P(N_k = l, N_{k+1} = m, D_k = x, N_{k+\phi} = n, D_{k+\phi} = y) \\ &= \pi(l) \cdot p_{lm}(x) \cdot p_{mn}^{(\phi-1)} \cdot p_n(y), \end{aligned} \tag{5.18}$$

which leads to the joint inter-departure time distribution

$$f^{(\phi)}(x,y) = \sum_{l=0}^{\infty} \sum_{m=(l-1)^+}^{\infty} \sum_{n=(m-\phi+1)^+}^{\infty} f_{lmn}^{(\phi)}(x,y), \tag{5.19}$$

where the notation $(\cdot)^+$ is equal to the expression $\max\{\cdot\,, 0\}$.

5.4 Numerical results

In this section, we present numerical results for the output dynamics of several $M/G/1$-queues with different service time variability and utilisation parameters. For convenience, we report the results using the auto-correlation as a normalised measure of the auto-covariance. Based on $\gamma(\phi)$ derived in the last section, the auto-correlation function $r(\phi)$ is defined as

$$r(\phi) = \frac{\gamma(\phi)}{Var(D)}. \tag{5.20}$$

We consider three types of $M/G/1$-queues with different service time variability parameters. We use the squared coefficient of variation $scv(B^U)$ to measure the service time variability. In the first case, $scv(B^U) = 0$, that is, the queue is of type $M/D/1$. In the second and third case, the service time distributions are discretised gamma distributions with low ($scv(B^U) = 0.79$) and high ($scv(B^U) = 2.59$) variability. The expected values of the service time distributions are equal in both cases, $E(B^U) = 8.17$. We vary the utilisation parameters $\rho \in \{0.30, 0.45, 0.60, 0.80, 0.95\}$ of the queues by increasing the flow parameters $\lambda \in \{0.037, 0.055, 0.073, 0.098, 0.116\}$. In the $M/D/1$-queues, the service time equals 8 time units, and the flow parameters are adjusted accordingly to obtain the utilisation parameters above. In total, we consider 15 different $M/G/1$-queues, classified into three groups according to their service time variability and compute the inter-departure time auto-correlation for lags 1 to 9.

Figure 5.3 shows the effects of the service time variability and utilisation on the departure stream auto-correlation for lags 1 to 9. It can be seen that in all queues,

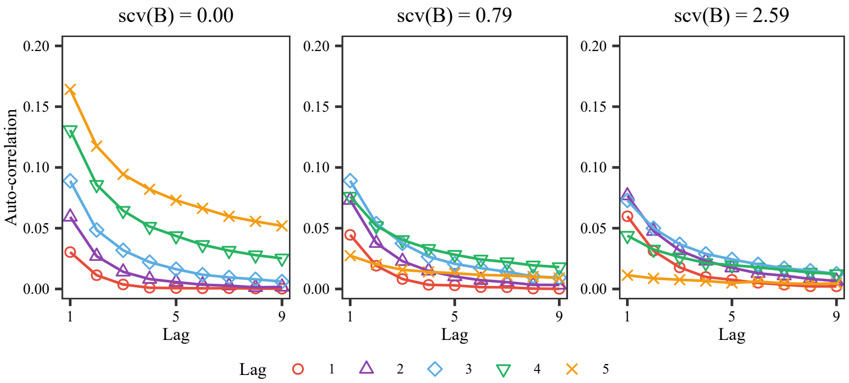

Figure 5.3: Auto-correlation for inter-departure times in $M/G/1$-queues with different service time variability and utilisation parameters.

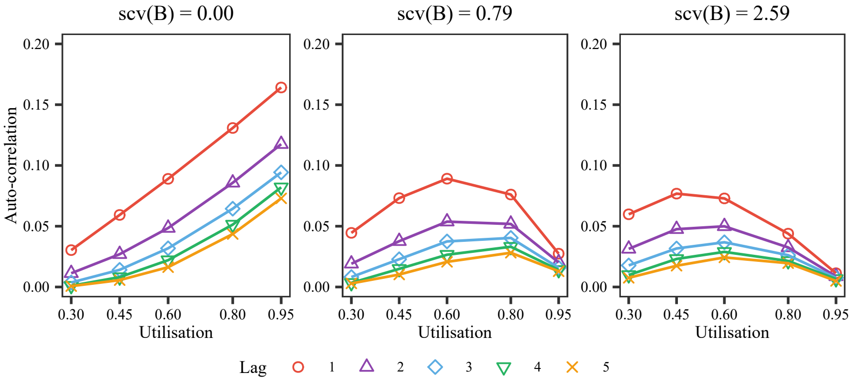

Figure 5.4: Auto-correlation for inter-departure times in $M/G/1$-queues in dependence of utilisation.

auto-correlation is positive and decreases with increasing lags. In the $M/D/1$-queues with high utilisation, auto-correlation is well above zero, even for large lags. In the low-traffic $M/D/1$-queues ($\rho \leq 0.5$), auto-correlation approaches zero as the lag increases. In these cases, auto-correlation is negligible for $\phi \geq 5$. In the second and third category of queues, where $scv(B^U) > 0$, auto-correlation is positive, but considerably smaller compared to the $M/D/1$-type queues. Again, auto-correlation decreases with increasing lags and approaches zero for most of the queues when $\phi \geq 5$. However, when the service time variability is greater than zero, it can be seen that auto-correlation is a non-monotone function of the utilisation parameter ρ.

This effect becomes clearly visible in Figure 5.4, where we plot the same data as before, with utilisation now on the x-axis. Note that due to the fact that for higher lags, auto-correlation is close to zero, we only plot the data for $\phi \leq 5$. In Figure 5.4, the increase of auto-correlation in the $M/D/1$-type queues for increasing utilisation can be clearly seen. In the $M/G/1$-queues, auto-correlation increases for all lags when utilisation is below 50 percent, reaches a peak, and decreases as utilisation approaches 95 percent.

In summary, we have obtained two major effects in our numerical results. First, we found that the inter-departure time auto-correlation is positive and decreases, as the lag increases. Second, we found that in $M/D/1$-type queues, auto-correlation monotonously increases for increasing utilisation while in $M/G/1$-type queues, auto-correlation is a non-monotone function of the utilisation. In the following, we discuss these findings.

The positive values of auto-correlation can be explained using the relationship identified in equation (5.3). In equation (5.3), we distinguished two types of inter-departure times, service times and the sum of idle and service times. We roughly categorise them as short (i.e. service times) and long (i.e. service plus idle times) inter-departure times. Since the inter-departure time of customer k depends on the state that customer $k-1$ left behind in the system, short inter-departure times are likely to be observed in bulks. Consider for example a situation where $n > 0$ customers wait in the queue. In this case, (at least) the next n inter-departure times must be service times, i.e. short inter-departure times. Equivalently, long inter-departure times are only observed after the queue has starved. As it can be seen in the transition probability matrix $\mathbf{P}$, it is more likely for the queue to starve again after a long inter-departure time, than a large number of customers arriving during the service time. Therefore, a short (long) inter-departure time is likely to be followed by another short (long) inter-departure time, resulting in a positive departure stream auto-correlation.

To explain the dependence of the inter-departure time auto-correlation from utilisation, we first consider the extreme cases: Assume the utilisation of the $M/G/1$-queue to be zero, which means that the output stream of the queue is equal to the input stream. In contrast, when the utilisation is equal to 1, the output stream is equal to the service process. In both cases, the output stream is renewal and thus auto-correlation is zero. When the queue is in steady-state, as described above, the output is sequentially dependent, which results in alternating bulked sequences of short and long inter-departure times. Given a low utilisation of the queue, it is unlikely that a large number of customers waits in the queue. Thus, the output of the queue is mainly determined by the arrival process, and observing a sequence of short inter-departure times is unlikely. As a consequence, the auto-correlation is rather low. However, when the queue is moderately utilised, sequences of short and long inter-departure times become equally likely. In contrast to an i.i.d. sampling of the inter-departure time distribution, however, the inter-departure times are observed in alternating bulks of short and long times. As a consequence, the inter-departure time auto-correlation reaches a peak for medium utilisation. When the utilisation approaches 1, for $M/G/1$-type queues, the same explanation applies as for low-utilised queues. When the queue is in heavy-traffic, the output is mainly determined by the service process, which is renewal for $M/G/1$-type queues. Thus, auto-correlation is low. In contrast, in $M/D/1$-type queues, the service time is deterministic, and thus, inter-departure times are closely resembling to themselves when the queue is in heavy-traffic. Therefore, auto-correlation ever-increases for $M/D/1$-type queues, as the utilisation increases.

5.5 The effect of auto-correlated arrivals on the analysis of downstream queues

Consider a tandem network of two discrete-time queues, where the upstream queueing system is a $M/G/1$-queue that feeds a downstream $G/G/1$-queue. To analyse this tandem queue, we usually deploy a decomposition approach that approximates the output point process of the $M/G/1$-queue as renewal process and thus treats the downstream queue as $GI/G/1$-queue with inter-arrival time distribution D. Jacobi and Furmans (2022b) showed that in this type of tandem queue, the analysis of performance measures in the downstream $GI/G/1$-queue can be subject to considerable approximation errors. Jacobi and Furmans (2022b) identified two major drivers for the approximation quality of the renewal decomposition approach. First, they found that severe approximation errors are only observed in heavy-traffic queues and second, they found that the decomposition approach over- or underestimates waiting time depending on the variability of the connecting stream. In the following, we aim to explain these results based on the findings of this chapter. We first derive the waiting time computation of a downstream $G/G/1$-queue with auto-correlated arrivals, and compare it to the waiting time computed with a renewal arrival stream. Then, we discuss the effect of variability on the approximation quality of the renewal decomposition approach.

5.5.1 Waiting time with auto-correlated arrivals

Consider the waiting time of a downstream $GI/G/1$-queue that is fed with the i.i.d. distributed inter-departure time distribution D from the upstream $M/G/1$-queue. Let B^D denote the i.i.d. random variable of the general service time distribution, $B^D \sim G$, and W the corresponding waiting time of the downstream queue, respectively. Consider the waiting time W_{k+1} of customer $k+1$ arriving at the downstream queue. Let B_{k+1}^D denote the service time of this customer. If the system is not empty upon arrival of customer $k+1$, the waiting time is determined by the sum of the waiting time of the previous customer k and the service time of customer k, minus the time gap D_{k+1} between the arrival instances of customers $k+1$ and k. If the system is empty upon arrival, customer $k+1$ does not have to wait. In summary, W_{k+1} is determined by

$$\begin{aligned} W_{k+1} &= \max\left\{W_k + B_k^D - D_{k+1}; 0\right\} \\ &= \max\left\{\max\left\{W_{k-1} + B_{k-1}^D - D_k; 0\right\} + B_k^D - D_{k+1}; 0\right\} \\ &= \max\left\{W_{k-1} + B_{k-1}^D + B_k^D - (D_k + D_{k+1}); B_k^D - D_{k+1}; 0\right\}. \end{aligned} \tag{5.21}$$

In equation (5.21), the waiting time equals zero, if customer $k+1$ finds an empty system upon arrival. If customer $k+1$ finds one customer in the system upon arrival, the waiting time is equal to the residual service time of customer k, that is, $B_k^D - D_{k+1}$. If customer $k+1$ finds more than one customer in the system upon arrival, the waiting process has not renewed, and the waiting time is determined by the serial dependencies of the inter-arrival times D_k and D_{k+1}. In this case, W_{k+1} is computed by

$$W_{k+1} = W_{k-1} + B_{k-1}^D + B_k^D - (D_k + D_{k+1}). \tag{5.22}$$

Equation (5.22) shows the dependency of the waiting time from the serial inter-arrival times D_k and D_{k+1} in the case that customer $k+1$ has to wait upon arrival. Applying the decomposition approach for the computation of the waiting time distribution, we assume that the arrival stream is renewal and sample the inter-arrival times D_k and D_{k+1} from the i.i.d. random variable D. However, the i.i.d. sampling of D does not account for the sequential dependence of the inter-departure times identified above. Consider again the example, where a number of $n > 0$ customers waits in the $M/G/1$-queue at a given observation point. In this situation, the next n inter-departure times must be sampled from the service time distribution. Therefore, the arrival stream at the downstream queue is determined by the n-fold convolution of the service time distribution. In contrast, the analysis of the $GI/G/1$-queue using the inter-arrival time distribution D does not account for the number of customers in the queue. Thus, for the situation described above, the renewal decomposition approach determines the arrival stream as n-fold convolution of the inter-arrival time distribution D. As a consequence, the results for the downstream waiting time computed with the renewal decomposition approach are flawed.

5.5.2 The effect of variability on the approximation quality of downstream queues

In a line of queues, the variability of the connecting stream has a major influence on whether the renewal decomposition approach over- or underestimates waiting time in the downstream $GI/G/1$-queue (Jacobi and Furmans 2022b): When the arrival stream variability is smaller than 1, the decomposition approach underestimates waiting time in the downstream $GI/G/1$-queue (that is, the true waiting time is longer). When the arrival stream variability is greater than 1, the decomposition approach overestimates waiting time in the downstream $GI/G/1$-queue (that is, the true waiting time is shorter).

To explain this finding, we consider again the example described above, where a number $n > 0$ customer waits in the upstream $M/G/1$-queue. The $M/G/1$-queue is assumed to be in heavy-traffic, so the variability of the departure stream is mainly determined by the variability of the service process. First, we assume that the service time variability at the upstream $M/G/1$-queue is small, $scv(B^U) < 1$. Since the variability of the external Poisson arrival process is equal to 1, we know that the departure stream variability must be smaller than 1, as well. However, as the departure stream is still influenced by the external arrival process, we can safely assume that $scv(B^U) < scv(D) < 1$. As explained in equation (5.3), the inter-departure times of the sequence of n customers departing from the $M/G/1$-queue are solely determined by the service process. Since $scv(B^U) < scv(D)$, the sequence of n service times arriving at the downstream $G/G/1$-queue is more likely to contain short times than an i.i.d. sampling of the inter-departure time distribution D. The stream of n customers arriving at the downstream $G/G/1$-queue therefore (temporarily) brings more work to the downstream queue, thus leading to a higher congestion as if the inter-departure times where i.i.d. sampled from the inter-departure time distribution D. As a consequence, the waiting time in the downstream $G/G/1$-queue is longer than the waiting time in the corresponding $GI/G/1$-queue – The renewal decomposition approach underestimates waiting time.

Now consider the opposite case, where a number of $n > 0$ customers waits in the $M/G/1$-queue whose service time variability is greater than 1, $scv(B^U) > 1$. The $M/G/1$-queue is again assumed to be in heavy-traffic, so that the variability of the departure stream is mainly determined by the service process, and therefore $scv(D) > 1$. However, as the external Poisson arrivals also influences the departure process, we conclude that $scv(B^U) > scv(D) > 1$. Therefore, a sequence of n customers arriving at the downstream $G/G/1$-queue is more likely to contain longer times than an i.i.d. sampling of the inter-departure time distribution D. Thus, the sequence of n service times (temporarily) brings less work to the downstream $G/G/1$-queue, compared to an i.i.d. sampling of the inter-departure time distribution D. As a consequence, the waiting time in the downstream $G/G/1$-queue is shorter than the waiting time in the corresponding $GI/G/1$-queue – The renewal decomposition approach overestimates waiting time.

5.6 Chapter conclusion

In this chapter, we consider the output dynamics of the discrete-time $M/G/1$-queue to analyse the auto-correlation of the inter-departure times. Consecutive inter-departure times are sequentially dependent because the inter-departure time D_k of customer k depends on the system state that customer $k-1$ left behind. Consequently, the overlay of the renewal arrival and the renewal service processes generates non-renewal departures. To compute the ϕ-lag auto-correlation in the departure stream, we model the $M/G/1$-queue as a discrete-time Markov chain.

Based on the stationary distribution, we derive formulas to compute the joint inter-departure time distribution of two inter-departure times that are $\phi > 0$ instances apart. We present numerical results for the auto-correlation in $M/D/1$- and $M/G/1$-queues for several lags and utilisation parameters. The numerical results show that auto-correlation is positive and monotonically decreasing for increasing lags ϕ and a non-monotone function of the utilisation.

Based on these results – and taking into account the findings from the previous chapter – we explain the reasons for the approximation errors of the renewal decomposition approach when analysing downstream queues. For steady-state systems, we identify the dependency of the waiting time from the serial dependency of inter-arrival times at the downstream queue. The positive auto-correlation in the departure stream affect the flow factor and variability compared to an i.i.d. sampling of inter-departure times since a long (short) inter-departure time is likely to be followed by another long (short) inter-departure time. Consequently, results computed with the renewal decomposition approach are flawed and – depending on the departure stream variability – overestimate or underestimate the waiting time.

6 A refined decomposition approach with converging accuracy for discrete-time open tandem queues with Poisson arrivals and general service times

This chapter is based on a working paper entitled “A refined decomposition approach with converging accuracy for discrete-time open tandem queues with Poisson arrivals and general service times” (Jacobi and Shanthikumar 2023). The software that has been used to compute the results presented in this chapter is available in the *KITopen Repository* (Jacobi 2023a).

The author of this thesis was responsible for the conceptualisation, methodology, software programming, validation, formal analysis, writing, and visualisation of the research presented in this chapter.

Chapter abstract

Decomposition often is the only feasible and computationally efficient approach to compute steady-state performance measures for queueing networks. However, performance results may be subject to severe approximation errors as decomposition methods usually assume that the connecting streams can be approximated by

renewal processes. In this chapter, we study the discrete-time tandem queue with external Poisson arrivals and generally distributed service times. To overcome the renewal assumption, we present the semi-Markov arrivals decomposition approach (SMAD), a refined decomposition approach, where the connecting stream between the upstream and the downstream station is described by a semi-Markov process. Using this modelling approach, the auto-correlation of the upstream inter-departure times is preserved for downstream queuing analysis. To avoid state space explosion, we demonstrate how to limit the state space of the embedded Markov chain in the upstream queue to an upper bound κ. Limiting the state space is computationally efficient, but leads to approximations as the departure point process depends on the state of the Markov chain. We present numerical results for several state space limits to demonstrate that the approach produces reasonably accurate results when the state space limit is tight, and converges arbitrarily accurate with increasing state space size.

6.1 Introduction and problem description

Closed-form solutions to analyse the discrete-time tandem queue with Poisson arrivals and general service times usually require great computational effort. Decomposition is often the only feasible and computationally efficient approach to compute steady-state performance measures (such as the probability distribution of the number of customers and waiting time) in the queues. This approach partitions the network into individual queuing systems and analyses them in isolation. It is based on the assumption that the output stream of the upstream $M/G/1$-queue – which is fed into the downstream $GI/G/1$-queue – can be approximated by a renewal process. However, it is well known that the departure process is a point process that is generally difficult to deploy for queueing system analysis (Whitt 1981, 1982). Recently, it has been shown that the renewal assumption of the departure point process may result in severe approximation errors when computing performance measures in downstream queues (Jacobi and Furmans 2022b).

The renewal assumption ignores the fact that the inter-departure times of the upstream $M/G/1$-queue are auto-correlated (Jacobi 2023b). Inter-departure times are sequentially dependent because the inter-departure time of customer k depends on the state that the previous customer $k-1$ left behind in the queueing system (Jacobi 2023b): If the system starves after the departure of customer $k-1$, the next inter-departure time is the sum of the queue's idle time and the service time of the next service period. If the upstream queue did not starve after the departure of customer $k-1$, the next inter-departure time is equal to the service time of customer k. The i.i.d. sampling of the inter-departure time distribution, however, does not account for this effect. Therefore, performance results obtained with the renewal decomposition approach are approximate.

To overcome the renewal assumption for the analysis of tandem queues with Poisson arrivals and general service times, this chapter presents the semi-Markov decomposition approach (SMAD). The novelty of this decomposition method is that a semi-Markov process (SMP) is used to model the connecting stream between the upstream $M/G/1$- and the downstream $G/G/1$-queue. SMAD does not rely on the renewal assumption of the inter-departure time distribution because the SMP allows to compute the conditional probability distribution of inter-departure times based on the system's state in the upstream $M/G/1$-queue. For downstream queueing analysis, we deploy the discrete-time $SM/G/1$-queue (semi-Markov arrival queue with general service times) which was introduced by Rieger and Haßlinger (1994). To avoid state space explosion and computational inefficiency, we demonstrate how to limit the state space of the embedded Markov chain in the SMP to an upper bound κ. Our numerical results show that the approach produces reasonably accurate results when the state space is limited and converges arbitrarily accurate with increasing state space size.

The remainder of this chapter is organised as follows. In Section 6.2, we present the related literature. Section 6.3 introduces the stochastic models for the upstream and downstream queues and describes how to reduce the computational complexity of the decomposition method by limiting the state space of the SMP. Section 6.4 presents numerical results and compares SMAD with the renewal decomposition approach and simulation. Section 6.5 concludes the chapter.

6.2 Literature review

Approaches to solve queuing systems with auto-correlated arrivals have been worked on for decades in both the continuous-time (Lucantoni and Neuts 1994, Ferng and Chang 2001b, Lee et al. 2003, Shioda 2003, Kim et al. 2008) and the discrete-time domains. In the continuous-time domain, starting with the work from Neuts (1979b) on so-called versatile processes (later called Neuts-flows (Ramaswami 1980)), researchers developed queueing models with Markovian Arrival Processes (MAP), Markov Modulated Poisson Process (MMPP), and their generalisation, the Batch Markovian Arrival Processes (BMAP) (Vishnevskii and Dudin 2017). A definition of the BMAP and its stochastic properties, as well as a literature review on their applications in the continuous-time domain, has been written by Vishnevskii and Dudin (2017).

BMAP-flows have been deployed to analyse tandem queues in the continuous-time domain and helped overcome the drawbacks of the renewal assumption. Lian and Liu (2008) study the tandem queue with MAP inputs and exponential service times and computed the joint queue length distribution for both servers. Gómez-Corral (2002) considers a tandem queue with external MAP flows and exploited the phase-type distribution of the upstream service time to model the output process as MAP. Heindl and Telek (2002) present a similar approach for large open networks. Ferng and Chang (2001a) propose a moment matching scheme to emulate the tagged output process as a MMPP that is fed into the intermediate queues. Heindl (2001) considers a decomposition approach with external MMPP, where the internal traffic processes are described as semi-Markov processes which are subsequently converted to a MMPP. Heindl (2003) later extends this framework to allow the splitting of SMPs and superposition of MMPPs for general queuing network analysis.

In the discrete-time domain, the analysis of queues with auto-correlated arrivals focuses on developing models with semi-Markov arrivals. While early approaches have been limited to special cases (Arjas 1972, de Smit 1986), Rieger and

Haßlinger (1994) present a Markov-chain model for the discrete-time $SM/G/1$-queue and derived the distribution for the number of customers in the queue. We deploy this model in our decomposition approach and thus present it in detail in Section 6.3. An algorithm for the stationary distributions of the waiting and idle time for the discrete-time $SM/G/1$-queue is presented by Haßlinger (2000). This approach performs a Wiener–Hopf factorisation and is an extension of the efficient algorithm for the waiting time distribution of the discrete-time $GI/G/1$-queue by Grassmann and Jain (1989).

Haßlinger and Rieger (1996) present a decomposition approach for general discrete-time open queuing networks. The authors outlined the renewal assumption of the inter-node flows, developed methods for splitting and superposition of discrete-time renewal processes, and present an approach for the computation of the distribution of the number of customers in the $GI/G/1$-queues. Further, Haßlinger and Rieger (1996) discuss the generalisation of the decomposition approach to auto-correlated arrivals using semi-Markov processes for the interconnection of queues. However, they conclude that “further study is needed to construct optimum SMP representations of non-renewal processes, regarding state space limitations for a tractable analysis.”

The decomposition approach presented in this chapter addresses this research gap. We model the output process of the discrete-time $M/G/1$-queue as a semi-Markov process and introduce a straightforward method to limit the state space of the embedded Markov chain in order to increase the computational efficiency of the decomposition method.

6.3 The semi-Markov arrival decomposition approach

In this section, we introduce the semi-Markov arrival decomposition approach (SMAD). We present the stochastic model for the upstream $M/G/1$-queue that generates a semi-Markov departure stream in Section 6.3.1. In Section 6.3.2, we show how to limit the state space of the embedded Markov chain to reduce the computational complexity in the downstream queue. Finally, in Section 6.3.3, we introduce the downstream $SM/G/1$-queueing model that is fed with the (limited state space) semi-Markov departure process.

6.3.1 The upstream $M/G/1$-queue with semi-Markov departures

We model the upstream $M/G/1$-queue as a discrete-time semi-Markov process. We extend the investigations on the output dynamics of the $M/G/1$-queue from Chapter 5 to connect the state transitions of the embedded Markov chain with the computation of the state-dependent inter-departure times. As described in Section 2.3, "if we know the past visited states and interval times of the system, the next visited state and the associated interval time depend only on the present state". Since we have investigated the embedded Markov chain model of the $M/G/1$-queue in detail in Chapter 5, our focus here is on the conditional probability function to compute the state-dependent inter-departure times. From this, we derive the computation of the semi-Markov kernel which eventually forms the input to the downstream queue.

Let the stochastic process $\mathcal{Z}^U = \{(N_k^U, D_k), k = 1, 2, ...\}$ denote a semi-Markov process where $N_k^U \in \mathbb{N}_0$ is the number of customers in the $M/G/1$-queue immediately after the departure instance of customer k, and $D_k \in \mathbb{N}$ is the inter-departure time between customers k and $k+1$. Let the probability function

$$f(t \,|\, i) = P(D = t \,|\, N^U = i) \tag{6.1}$$

denote the conditional probability that the inter-departure time is equal to t, given that the embedded Markov chain of the semi-Markov process Z_k is in state $N_k^U = i$. The probability function $f(t \,|\, i)$ is equal to the service time, if the system is not empty immediately after the departure instance (that is, $i > 0$), and equal to the sum of the remaining inter-arrival time and the service time, if the system is starving after departure instance k (that is, $i = 0$):

$$f(t \,|\, i) = \begin{cases} \sum_b P(A = t - b) \cdot P(B^U = b) & i = 0, \\ P(B^U = t) & i > 0. \end{cases} \tag{6.2}$$

The semi-Markov kernel $\mathbf{Q} = \{q_{ij}(t); i, j \in \mathbb{N}_0, t \in \mathbb{N}\}$ is defined based on these considerations, where

$$q_{ij}(t) = P(N^U_{k+1} = j, D_{k+1} = t \,|\, N^U_k = i) \tag{6.3}$$

denotes the conditional probability that the number of customers in the system transitions from i to j with inter-departure time t. For all $j \in \mathbb{N}_0$, the kernel entries are defined as

$$q_{ij}(t) = \begin{cases} \displaystyle\sum_b \frac{e^{-\lambda b}(\lambda b)^j}{j!} \cdot P(A = t - b) \cdot P(B^U = b) & i = 0, \\ \displaystyle\frac{e^{-\lambda t}(\lambda t)^{j-i+1}}{(j-i+1)!} \cdot P(B^U = t) & j + 1 \geq i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.4}$$

Recall that the state of the stochastic process is only observed immediately after a departure instance. In equation (6.4), we therefore distinguish the same cases as for the definition of the conditional probability function $f(t \,|\, i)$. If the system was starving after the departure instance ($i = 0$), the departure time t is the sum of the idle time of the system and the service time of the next customer. Assume that this service time is equal to b. Thus, the probability that the system is idle for $t - b$ time units is computed by the inter-arrival time distribution $P(A = t - b)$. After this customer arrived at the system, j additional customers may arrive during the service period of this customer. The probability is computed with the well-known formula for the Poisson arrival process.

In the second case considered in equation (6.4), the system is not starving after the departure instance. Therefore, the next customer can immediately start the service, and the next inter-departure time is equal to the service time of this customer. During the service period, more customers may arrive at the system. In order to observe j customers at the next departure instance, $j - i + 1$ customers have arrived during the service period. In case that zero customers arrived during the service time, $j = i - 1$ customers remain in the system at the next departure instance.

Given the semi-Markov kernel $\mathbf{Q}$, we compute the transition matrix $\mathbf{P} = (p_{ij})_{i,j \in \mathbb{N}_0}$ of the embedded Markov chain $\mathcal{N}^U = \{N_k^U, k = 1, 2, ...\}$ by

$$p_{ij} = \sum_{t=0}^{\infty} q_{ij}(t) \quad \forall i, j \in \mathbb{N}_0, \tag{6.5}$$

and equivalently, the equation

$$q_{ij}(t) = p_{ij} \cdot f(t \,|\, i) \tag{6.6}$$

holds. Equation (6.6) is the central result of the considerations made in this section. With this expression, we can connect the computation of state-dependent inter-departure times (cf. equation (6.2)) with the transition probability matrix $\mathbf{P}$. Since the computation of the transition probabilities is straightforward (also compare equation (5.5) from the previous chapter), the SMP is clearly defined, and the computation of the semi-Markov kernel entries is straightforward, as well. However, from equation (6.6), it can be seen that the semi-Markov kernel is infinite, because the transition probability matrix $\mathbf{P}$ is infinite. Thus, we recall the computation of the stationary distribution below, and derive the state space limitation method from this in the following section.

Let π denote the stationary distribution of the embedded Markov chain $\mathcal{N}^U = \{N_k^U, k = 1, 2, ...\}$, such that

$$\begin{aligned} \pi &= \mathbf{P}\pi, \\ \pi\mathbf{e} &= 1. \end{aligned} \tag{6.7}$$

The state space of the discrete-time Markov chain $\mathcal{N}^U$ and the stochastic matrix $\mathbf{P}$ are infinite and thus, the computation of π using the Gauss-Seidel method is non-trivial. However, the stationary distribution can be efficiently computed using the recursion method introduced by Ramaswami (1988). This procedure was introduced in detail in Section 5.3.

6.3.2 State space limitation

To prevent state-space explosion during performance analysis in the downstream queue, we limit the state space of the embedded Markov chain $\mathcal{N}^U$ to an upper bound $\kappa > 0$. Intuitively, the upstream queue's departure stream is determined by whether or not the queue has starved after a departure instant (cf. equation (6.2)). In the most restricted case, $\kappa = 1$, the state space Ω of the embedded Markov chain is limited to two states only, $\Omega = \{0, \kappa\}$, representing an empty system and a system occupied by any number of customers. As the upper bound κ increases, the state space of the embedded Markov chain tracks an increasing number of customers $\Omega = \{0, 1, 2, ..., \kappa\}$, where the state κ again represents all states where at least κ customers are in the system, $N^U \geq \kappa$.

On the one hand, an SMP with a large state limit models the queue's departure behaviour with increasing accuracy. This is since κ successive inter-departure times can be tracked. Consider again the example described in Chapter 5 where a number of n customers waits in the upstream $M/G/1$-queue, thus producing an output sequence of n service times. In this situation, the SMP can only accurately track the sequence of n service times if $\kappa \geq n$. On the other hand, the SMP forms the input for the downstream queue and thus, limiting the state space increases the

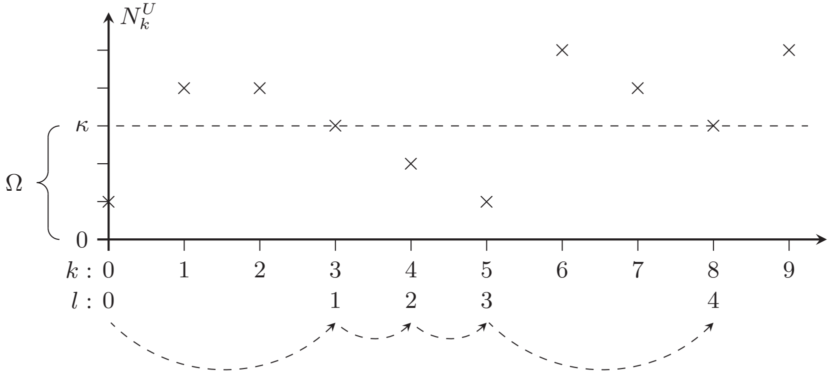

Figure 6.1: State space limitation method in the upstream $M/G/1$-queue.

computational efficiency of the decomposition method. However, approximations must be found for the stochastic behaviour of the limited transition probability matrix and therefore, the performance results obtained with a limited state space are approximate. In the following, we describe the state space limitation method.

Let $\mathcal{N}'$ denote the limited Markov chain with state space $\Omega = \{0, 1, 2, ..., \kappa\}$. The discrete-time Markov chain $\mathcal{N}'$ observes the stochastic process $\mathcal{N}^U$ only when it is in one of the states in Ω. Let T_l denote the observation points of the associated point process $\mathcal{N}'$, such that

$$\begin{aligned} T_1 &= \inf\big\{k : N_k^U \in \Omega \,\big|\, N_0^U \in \Omega\big\}, \\ T_l &= \inf\big\{k : k > T_{l-1}, N_k^U \in \Omega\big\}, \quad l = 2, 3, 4, ... \end{aligned} \tag{6.8}$$

Without loss of generality, we assume $T_0 = 0$ and define the discrete-time Markov chain $\mathcal{N}' = \{N_{T_k}^U, k = 0, 1, 2, ...\}$.

Figure 6.1 shows the number of customers N_k^U over time and visualises the state space limitation approach. In Figure 6.1, each observation point k corresponds to a departure instant of a customer from the $M/G/1$-queue. The change of the number of customers over time follows the same dynamics as described in

Chapter 5: During a service period, any number $j \geq 0$ of customers may arrive at the system, according to the Poisson arrival process. Since we only observe the state of the system immediately after the departure of a customer, the number of customers N^U can only reduce step-wise (as seen e.g. for the sequence $(2, 3, 4, 5)$ of observation points k). In observation point $k = 0$, the number of customers is smaller than κ, and therefore $N_0^U \in \Omega$. During the following service period, more customers arrive at the system. In the next departure instances at $k = 1$ and $k = 2$, we find $(\kappa + 1)$ customers in the system, and therefore $N_1^U, N_2^U \notin \Omega$. At instance $k = 3$, the number of customers is $N_3^U = \kappa$, thus $N_3^U \in \Omega$. As described above, the limited Markov chain $\mathcal{N}'$ observes the state of the original Markov chain $\mathcal{N}^U$ only if the number of customers lies in Ω. Consequently, the first two observation points T_l of the limited Markov chain are $T_0 = 0$ and $T_1 = 3$. For the observation points $k = 4$ and $k = 5$, $N_k^U \in \Omega$, and thus $T_2 = 4$ and $T_3 = 5$. In point $k = 6$, the number of customers exceeds the state limit κ again and declines back into the boundary for $k = 8$. Therefore, the next observation point of the limited Markov chain is at $T_4 = 8$.

Since $\kappa > 0$, the stochastic process $\mathcal{N}'$ has a transition probability matrix $\mathbf{P}' = (\hat{p}_{ij})_{i,j \in \Omega}$ of size $(\kappa{+}1) \times (\kappa{+}1)$. As we have demonstrated above, the stochastic behaviour of the limited Markov chain $\mathcal{N}'$ is equal to the stochastic behaviour of the original Markov chain when the system is in a state $N^U < \kappa$. However, approximations must be found for the transition probabilities from and to state κ. Based on the invariant probability vector π of the original Markov chain $\mathcal{N}^U$, we define $\mathbf{P}'$ as follows:

$$\mathbf{P}' = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0,\kappa-1} & \hat{p}_{0,\kappa} \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1,\kappa-1} & \hat{p}_{1,\kappa} \\ 0 & p_{21} & p_{22} & \cdots & p_{2,\kappa-1} & \hat{p}_{2,\kappa} \\ 0 & 0 & p_{32} & \cdots & p_{3,\kappa-1} & \hat{p}_{3,\kappa} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{\kappa-1,\kappa-1} & \hat{p}_{\kappa-1,\kappa} \\ 0 & 0 & 0 & \cdots & \hat{p}_{\kappa,\kappa-1} & \hat{p}_{\kappa,\kappa} \end{bmatrix}. \tag{6.9}$$

In the stochastic matrix $\mathbf{P}'$, the probabilities p_{ij} are identical to those in the matrix $\mathbf{P}$. Approximations must be found for the transition probabilities in the last row and the last column. The values $\hat{p}_{i,\kappa}$ in the last column of matrix $\mathbf{P}'$ are the row sum probabilities of all $p_{ij}, j \geq \kappa$ in each row $i < \kappa$,

$$\hat{p}_{i,\kappa} = \sum_{j=\kappa}^{\infty} p_{ij} = 1 - \sum_{j=0}^{\kappa-1} p_{ij} \quad \forall i = 0, 1, ..., \kappa - 1. \tag{6.10}$$

As in the original matrix $\mathbf{P}$, the first $(\kappa - 1)$ entries in row κ equal zero. The value $\hat{p}_{\kappa,\kappa-1}$ in line κ is computed by solving

$$\hat{p}_{\kappa,\kappa-1} = \pi(\kappa) \cdot p_{\kappa,\kappa-1} \cdot \left(\sum_{j=\kappa}^{\infty} \pi(j) \right)^{-1}, \tag{6.11}$$

where $\pi(j)$ is the j-th entry of the stationary distribution vector π. Finally, we compute the probability $\hat{p}_{\kappa,\kappa}$ by

$$\hat{p}_{\kappa,\kappa} = 1 - p_{\kappa,\kappa-1}. \tag{6.12}$$

Given the stochastic matrix $\mathbf{P}'$, we compute the invariant probability vector $\hat{\pi}$ that satisfies

$$\begin{aligned} \hat{\pi} &= \mathbf{P}'\hat{\pi}, \\ \hat{\pi}\mathbf{e} &= 1, \end{aligned} \tag{6.13}$$

and approximates the stationary distribution π of the stochastic process $\mathcal{N}^U$. A proof that $\hat{\pi}(i) = \pi(i)$ for all $i = 0, 1, ..., \kappa - 1$ can be found in Appendix B.

Based on this observation, we conclude that the stochastic process $\mathcal{N}'$ and its limited transition probability matrix $\mathbf{P}'$ appropriately approximate the stochastic behaviour of the original Markov chain $\mathcal{N}^U$. Given the limited transition matrix $\mathbf{P}'$, we can efficiently approximate the kernel entries of the original semi-Markov process defined in equation (6.6) by solving

$$\hat{q}_{ij}(t) = \hat{p}_{ij} \cdot f(t \,|\, i). \tag{6.14}$$

Recall the importance of equation (6.6) defined above. The kernel entries $\hat{q}_{ij}(t)$ efficiently approximate the stochastic behaviour of the original SMP on the limited state space Ω. However, as stated above, tight state space limits lead to approximations. Consider again the example $\kappa = 1$, where the limited transition matrix $\mathbf{P}'$ is a (2×2) matrix. In this case, the semi-Markov process only distinguishes if the system is starving or non-starving after the departure of a customer. However, longer sequences of state-dependent inter-departure times cannot be tracked. This is due to the fact that the embedded Markov chain $\mathcal{N}'$ may always transition from state κ to state 0, regardless of the number of customers in the queue. In contrast, the number of customers in the original stochastic process $\mathcal{N}$ can only decrease step-wise with each transition. As a consequence, the approximation of the SMP improves, as the state limit κ increases.

6.3.3 The downstream $SM/G/1$-queue

The discrete-time $SM/G/1$-queue has been studied in the literature by Rieger and Haßlinger (1994), who derive an analytical solution for the stationary distribution, the distribution of the number of customers, and the idle time distribution. We will briefly introduce the embedded Markov chain model as described by Rieger and Haßlinger (1994) and derive the computation of the waiting time distribution based on the stationary distribution.

Let the stochastic process $\mathcal{N}^D = \{(N_k^D, r_k, N_k^U), k = 1, 2, ...\}$ denote the states of the $SM/G/1$-queue at all instants k when a customer arrives at or departs from the system. The state of the stochastic process is given by $Z_k = (N_k^D, r_k, N_k^U) \in \mathbb{N} \times \{1 - R, ..., R - 1\} \times \{0, ..., \kappa\}$. Note that the parameters N_k^U and κ stem from the upstream $M/G/1$-queue and will be used to model the arrival process at the $SM/G/1$-queue.

The embedded stochastic process $\mathcal{N}^D$ forms a homogeneous Markov chain (see Rieger and Haßlinger (1994) for a proof). The transition from state k to state $k+1$ is caused by an arrival, a departure, or a simultaneous event of both. Accordingly, the random variable N_k^D is incremented, decremented, or remains unchanged, and the residual time r_k until the next event is updated. The semi-Markov arrival process is state-dependent and is determined by the state N_k^U of the governing chain. The decomposition method proposed here computes the state-dependent inter-arrival times according to the approximation formula (6.14).

The state components are denoted as follows:

N_k^D The number of customers in the $SM/G/1$-queue at instant k,

$r_k > 0$ A customer departs at instant k, and the residual arrival time is r_k time units,

$r_k < 0$ A customer arrives at instant k, and the residual departure time is $|r_k|$ time units,

$r_k = 0$ Simultaneous arrival and departure of a customer,

N_k^U The state of the governing chain at instant k.

6.3.3.1 Transition probabilities

The following descriptions are based on the considerations made by Rieger and Haßlinger (1994). Before introducing the transition probabilities, we define the probability function

$$u_{ij}(r) = P(B^D - D = r) = \sum_{t=\{1,1-r\}^+}^{\{R,R-r\}^-} \hat{q}_{ij}(t) \cdot P(B^D = r + t), \qquad (6.15)$$

which denotes the likelihood that a simultaneous arrival and departure event leaves a residual time $r \in \{1 - R, ..., R - 1\}$ in the system. Note that the probability function $u_{ij}(r)$ considers state-dependent arrivals $\hat{q}_{ij}(k)$ from the (limited) semi-Markov arrival process (cf. equation (6.14)). The function will be exploited to compute the transition probabilities presented in the following.

As stated above, the transition from state k to state $k+1$ is caused by the arrival of a customer, the departure of a customer, or the simultaneous arrival and departure of a customer. As a consequence, we observe the state of the Markov chain only if the service process renews or the arrival process generates a new point. For the computation of the transition probabilities, we distinguish four transition types,

1. a departure event that leaves a non-empty system behind,
2. an arrival event,
3. a simultaneous arrival and departure event, and
4. a departure event that leaves a starving system behind.

We use the probability functions $u_{ij}(r)$ (cf. equation (6.15)), $\hat{q}_{ij}(t)$ (cf. equation (6.14)), and $P(B^D)$ to compute the transition probabilities. It will be shown that the first three transition types defined above are similar in the sense that they apply these probability functions to different cases. In the fourth transition type, the $SM/G/1$-queue starves, which requires additional considerations.

Type 1: Departure event that leaves a non-empty system behind

The first transition type considers the departure event of a customer that leaves a non-empty system behind. Let $N \geq 1$ and $r \geq 1$. Let the state of the stochastic process before the transition be described by $Z_k = (N+1, r+l, i)$. Thus, in instant k, we have $N_k^D = N+1$, $r_k = (r+l)$, and $N_k^U = i$. Equivalently, let the state of the stochastic process after the transition be described by $Z_{k+1} = (N, r, j)$, so that $N_{k+1}^D = N$, $r_{k+1} = r$, and $N_{k+1}^U = j$. As we consider the departure of a customer, it can be seen that the customer count in the $SM/G/1$-queue decreases by 1, $N_{k+1}^D = N_k^D - 1$.

The transition probability from state k to state $k+1$ depends on one of three cases, the departure, arrival, or simultaneous event of both in the previous state k. In the first case, $r_k = (r+l) > 0$, which means that in state k, a customer departs from the system and thus, the service process renews. Since in state $k+1$, another customer departs, the transition from state k to state $k+1$ is a service period of length l. Thus, the transition probability is equal to $P(B^D = l)$. Note that in this transition, the state of the SMP must not change, therefore $i = j$, and the residual arrival time is updated to $r_{k+1} = r$.

In the second case, $r_k = (r+l) < 0$, which means that in state k, a customer arrives at the system. Since in state $k+1$ a customer departs from the system, the residual service time must be smaller than the next inter-arrival time. In state $k+1$, the residual arrival time remaining in the system is $r_{k+1} = r$, and thus, the inter-arrival time in state k is $|\, l \,|$ time units. Note that $l < 0$ since $r > 0$, and therefore, the transition probability is equal to $\hat{q}_{ij}(-l)$.

In the third case, $r_k = (r+l) = 0$, which means that in state k, one customer arrived at the system, as simultaneously another customer departed from the system. Therefore, the service process renews while simultaneously, the arrival process generates a new arrival event in state k. Since in state $k+1$, a customer departs from the system, the system has completed an entire service period, and the renewed service time in k must be smaller than the inter-arrival time. Therefore, we use equation (6.15) to compute the probability that the residual arrival time

$r_{k+1} = r$ remains in the system. Note that in equation (6.15), $r < 0$ for $B^D < D$, and therefore we compute $u_{ij}(-r)$.

In summary, the transition probability for a departure event that leaves a non-empty system behind is computed by

$$P\big(Z_{k+1} \,\big|\, Z_k\big) = \begin{cases} P(B^D = l) & (r+l) > 0,\ i = j, \\ \hat{q}_{ij}(-l) & (r+l) < 0, \\ u_{ij}(-r) & (r+l) = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.16}$$

Type 2: Arrival event

The second transition type considers an arrival instant of a customer. Let $N \geq 2$ and $r \geq 1$. Let the state of the stochastic process before the transition be described by $Z_k = (N-1, -r-l, i)$. Thus, in instant k, we have $N_k^D = N-1$, $r_k = (-r-l)$, and $N_k^U = i$. Equivalently, let the state of the stochastic process after the transition be described by $Z_{k+1} = (N, -r, j)$, so that $N_{k+1}^D = N$, $r_{k+1} = -r$, and $N_{k+1}^U = j$. As we consider the arrival of a customer, it can be seen that the customer count in the $SM/G/1$-queue increases by 1, $N_{k+1}^D = N_k^D + 1$.

As before, the transition probability from state k to state $k+1$ depends on one of three cases, the departure, arrival, or simultaneous event of both in the previous state k. In the first case, $r_k = -(r+l) > 0$, which means that in state k, a customer departs from the system and thus, the service process renews. Note that $l < 0$ and $|\,l\,| > r$. The residual service time remaining after the arrival instance in state $k+1$ is $r_{k+1} = -r$. Therefore, the service period has a total length of $r + |\,l\,|$ time units, and thus the transition probability is equal to $P(B^D = r + |\,l\,|)$. Note that in this transition, the state of the SMP must not change, and thus $i = j$.

In the second case, $r_k = -(r+l) < 0$, which means that in state k, a customer arrives at the system. Since in state $k+1$, another customer arrives, the transition from state k to state $k+1$ is an inter-arrival period of length $l > 0$. Thus, the transition probability is equal to $\hat{q}_{ij}(l)$. In state $k+1$, the residual service time is updated to $r_{k+1} = -r$.

In the third case, $r_k = -(r+l) = 0$, which means that in state k, one customer arrived at the system, as simultaneously another customer departed from the system. Therefore, the service process renews while simultaneously, the arrival process generates a new arrival event in state k. Since in state $k+1$, a customer arrives at the system, the inter-arrival time in state k must be smaller than the service time. Therefore, we use equation (6.15) to compute the probability that the residual service time $r_{k+1} = -r$ remains in the system. Note that in equation (6.15), $r > 0$ for $B^D > D$, and therefore we compute $u_{ij}(r)$.

In summary, the transition probability for an arrival event is computed by

$$P\big(Z_{k+1} \,\big|\, Z_k\big) = \begin{cases} P(B^D = r + |\,l\,|) & -(r+l) > 0,\ i = j, \\ \hat{q}_{ij}(l) & -(r+l) < 0, \\ u_{ij}(r) & -(r+l) = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.17}$$

Type 3: Simultaneous arrival and departure event

The third transition type considers the simultaneous arrival and departure of a customer. Let $N \geq 2$. Let the state of the stochastic process before the transition be described by $Z_k = (N, r, i)$. Thus, in instant k, we have $N_k^D = N$, $r_k = r$, and $N_k^U = i$. Equivalently, let the state of the stochastic process after the transition be described by $Z_{k+1} = (N, 0, j)$, so that $N_{k+1}^D = N$, $r_{k+1} = 0$, and $N_{k+1}^U = j$. As we consider the simultaneous arrival and departure of a customer, it can be seen that the customer count in the $SM/G/1$-queue does not change, $N_{k+1}^D = N_k^D = N$.

As before, the transition probability from state k to state $k+1$ depends on one of three cases, the departure, arrival, or simultaneous event of both in the previous state k. In the first case, $r > 0$, which means that in state k, a customer departs from the system. As in state $k+1$, another customer departs, the system has completed an entire service period of length r. Therefore, the transition probability is equal to $P(B^D = r)$. Note that during the service process, the state of the SMP arrival process must not change, and therefore $i = j$.

In the second case, $r < 0$, which means that in state k, a customer arrives at the system. As in state $k + 1$, another customer arrives, the time between both states is equal to the inter-arrival time r. Therefore, the transition probability is equal to $\hat{q}_{ij}(r)$.

In the third case, $r = 0$, which means that in state k, one customer arrived at the system, as simultaneously another customer departed from the system. Therefore in both states, the service process renews while simultaneously, the arrival process generates a new point. Thus, the transition probability is equal to the likelihood that a simultaneous event in state k leaves a residual time of zero in the system in state $k + 1$, $u_{ij}(0)$.

In summary, the transition probability for a simultaneous arrival and departure event is computed by

$$P\big(Z_{k+1} \,\big|\, Z_k\big) = \begin{cases} P(B^D = r) & r > 0,\ i = j, \\ \hat{q}_{ij}(r) & r < 0, \\ u_{ij}(0) & r = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.18}$$

Type 4: Departure event that leaves a starving system behind

The fourth transition type considers state transitions where the systems starves after the departure event. In the state transitions defined above, we assured that the number of customers in the system is always greater than zero. However, the system can only starve if the number of customers in the queue in a departure event is equal to one and the residual time r is non-negative. To increase computational efficiency, the Markov chain immediately transitions to the next arrival event, where the customer count is again one. Therefore, we do not explicitly model the idle phase of the $SM/G/1$-queue, and thus, the customer count N_k^D is always greater than zero.

Let the state of the stochastic process before the transition be described by $Z_k^D = (1, r, i)$. Thus, in instant k, we have $N_k^D = 1$, $r_k = r$, and $N_k^U = i$. Equivalently, let the state of the stochastic process after the transition be described by $Z_{k+1}^D = (1, 0, j)$, so that $N_{k+1}^D = 1$, $r_{k+1} = 0$, and $N_{k+1}^U = j$. As described above, the customer count remains unchanged, $N_k^D = N_{k+1}^D = 1$, but with the notion that the queue starves in between the state transition. Therefore, in state $k + 1$, the service process renews while simultaneously the arrival process generates a new point, and thus $r_{k+1} = 0$.

The transition probability from state k to state $k + 1$ depends on two cases, the departure of a customer, or simultaneous arrival and departure of a customer in state k. In the first case, $r > 0$, which means that in state k, a customer departs from the system, and the service process of the last remaining customer begins. The system starves only if the service time elapses before the next customer arrives. Since the residual arrival time is $r_k = r$, the transition probability is equal to the sum of all service time probabilities, where the service time is smaller than or equal to r (see first case in equation (6.19)). Note that if the service time equals r, the $SM/G/1$-queue starves for zero time units. As before, the state of the arrival SMP must not change, thus $i = j$.

In the second case, $r = 0$, which means that in state k, the last customer in the queue departs from the system, and simultaneously, another customer arrives. Since the customer count in state k equals one, the service process of the new customer immediately begins. As described above, the system starves only if this service time elapses before the next customer arrives. Therefore, the transition probability is the cumulative probability that the simultaneous arrival and departure event in state k leaves a negative residual time (that is $B^D < D$) in the system (see second case in equation (6.19)).

Note that states Z_k where $r < 0$ cannot exist because the arrival of a customer immediately causes a transition to state $Z = (1, 0, j)$. In summary, the transition probability for a departure event with starvation is defined as

$$P\big(Z_{k+1}\,\big|\,Z_k\big) = \begin{cases} \sum_{b=1}^{r} P(B^D = b) & r > 0,\ i = j, \\ \sum_{a=0}^{R-1} u_{ij}(-a) & r = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.19}$$

6.3.3.2 The waiting time distribution

We consider the distribution of waiting time at the arrival instants of customers. Let the probability $\pi(Z)$ denote the stationary probability of state $Z = (N^D, r, N^U)$. Since we only consider the system at arrival instants (that is, $\{Z = (N^D, r, N^U)\,|\,r \leq 0\}$), a normalisation constant θ is required, such that

$$\frac{1}{\theta} \sum_{N^U=0}^{\kappa} \Big(\pi\big(1, 0, N^U\big) + \sum_{N^D=2}^{\infty} \sum_{r=0}^{R-1} \pi\big(N^D, -r, N^U\big)\Big) \overset{!}{=} 1. \tag{6.20}$$

Let $P(W = w)$ denote the probability that an arriving customer has to wait $w \geq 0$ time units upon arrival. The probability that an arriving customer does not have to wait is equal to the probability that a customer enters an empty system (that is, $N^D = 1$),

$$P(W = 0) = \frac{1}{\theta} \sum_{N^U=0}^{\kappa} \pi\big(1, 0, N^U\big). \tag{6.21}$$

If the arriving customer finds another customer in the system upon arrival (that is, $N^D = 2$), the arriving customer has to wait the residual service time of that customer,

$$P(W = |r|) = \begin{cases} \dfrac{1}{\theta} \displaystyle\sum_{N^U=0}^{\kappa} \pi\big(2, -r, N^U\big) & r < 0, \\ \dfrac{1}{\theta} \displaystyle\sum_{N^U=0}^{\kappa} \pi\big(2, 0, N^U\big) \cdot P(B = r) & r = 0. \end{cases} \tag{6.22}$$

In equation (6.22), we distinguish two cases. First, upon arrival, a customer is already in service, and the residual service time is $|r|$ time units. In the second case, a departure instant coincides with the arrival of the customer. Thus, the service process renews in the arrival instant and the arriving customer has to wait the entire service time.

If the arriving customer finds more than one customer already present in the system ($N^D > 2$), he has to wait the residual service time of the customer in service and the sum of the service times of all customers waiting ahead in the queue. The probability that l consecutive service operations require n time units is denoted with the l-fold convolution of the service time vector, $b_n^{l\otimes}$. The waiting time probability is computed as follows:

$$P(W = |r| + n) = \begin{cases} \dfrac{1}{\theta} \displaystyle\sum_{N^U=0}^{\kappa} \pi\big(N^D, -r, N^U\big) \cdot b_n^{(N^D-2)\otimes} & r < 0, \\ \dfrac{1}{\theta} \displaystyle\sum_{N^U=0}^{\kappa} \pi\big(N^D, 0, N^U\big) \cdot b_n^{(N^D-1)\otimes} & r = 0. \end{cases} \tag{6.23}$$

In equation (6.23), we again distinguish the cases of a sole arrival event and a simultaneous arrival and departure event. In the first case, we obtain the residual service time $|r|$ from the state and compute the $(N^D - 2)$-fold convolution of the service time distribution. In the second case, the service process renews at the arrival instant, and thus, we compute the residual service time as $(N^D - 1)$-fold convolution of the service time distribution.

6.4 Numerical results

In this section, we verify the semi-Markov decomposition approach and present numerical results for the performance and computational efficiency of SMAD. We compute the waiting time in the downstream $SM/G/1$-queue using SMAD for the set of state limits $\kappa \in \{1, 5, 10, 20, 50\}$ and compare these results with the discrete-time renewal decomposition approach (DTQA), and simulation.

It is crucial for the comparison of these methods to carefully define the input parameters. Recall that the discrete-time queuing methods receive the inter-arrival time distribution as input, whereas SMAD receives the flow rate λ as input. As stated above (cf. equation (5.1)), the geometric inter-arrival time distribution is defined by the flow rate λ, and the state space of the Markov chain is infinite. However, the vector of the inter-arrival times that is input to DTQA must be of finite length. Therefore, we have to truncate the inter-arrival time vector. We set the truncation epsilon to 10^{-10}, and compute the corrected flow rate $\tilde{\lambda} = 1/E(\tilde{A})$, where $E(\tilde{A})$ is the expected value of the truncated inter-arrival time vector.

6.4.1 Verification

We consider a tandem queue where the service time distributions are equal at the upstream and the downstream station, $P(B = 15) = P(B = 16) = 0.5$, and the arrival stream is defined by $\lambda = 0.0613$. The utilisation of the tandem queue is $\rho = 0.950$. We compute the probability distribution of waiting time with five state space limits κ and compare the results to the waiting time distributions obtained with the renewal decomposition approach and simulation. Table 6.1 shows the first ten positions of the waiting time distributions, their expected values, and 95% percentiles.

The probability distributions in Table 6.1 indicate that the results of the novel decomposition approach improve with increasing accuracy parameter κ. As κ increases, the probabilities of finding the system empty, the expected waiting times, and the 95% percentile of the distributions converge towards the respective

Table 6.1: Distributions of waiting time (truncated) in the downstream queue, obtained with the renewal decomposition approach (DTQA), the semi-Markov arrival decomposition approach (SMAD) for five state limits κ, and simulation.

		SMAD(κ)					
W	DTQA	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 50$	Sim
0	0.3420	0.2976	0.2368	0.2260	0.2229	0.2219	0.2219
1	0.2259	0.2013	0.1537	0.1431	0.1399	0.1391	0.1387
2	0.1492	0.1434	0.1150	0.1057	0.1025	0.1019	0.1014
3	0.0984	0.1022	0.0904	0.0827	0.0798	0.0791	0.0788
4	0.0648	0.0728	0.0727	0.0669	0.0642	0.0636	0.0635
5	0.0425	0.0520	0.0590	0.0553	0.0529	0.0523	0.0522
6	0.0278	0.0371	0.0482	0.0462	0.0441	0.0436	0.0436
7	0.0181	0.0265	0.0396	0.0390	0.0373	0.0368	0.0368
8	0.0117	0.0190	0.0326	0.0332	0.0318	0.0314	0.0312
9	0.0075	0.0136	0.0270	0.0285	0.0275	0.0272	0.0268
$\sim$	...	...	...	...	...	...	...
EV	1.88	2.46	4.14	5.03	5.64	5.90	5.96
Q-95	6	8	14	18	21	22	22

Note: The probability distributions are truncated after 10 positions. The .999-percentile of the simulated probability distribution is 64.

simulation results. For $\kappa = 50$, we do not find evidence for the waiting time distribution to be significantly different from the simulated distribution. We performed a Chi-Square Goodness-of-Fit Test and found a significant relationship between both distributions $\big(\chi^2(11; 612{,}682) = 5.69, p = .893\big)$.

In terms of accuracy, the novel decomposition approach outperforms the renewal decomposition technique. Figure 6.2 plots the first ten positions of the waiting time distributions obtained with SMAD(50) alongside the results obtained with the renewal decomposition approach. It visualises the difference between both approaches, which Table 6.1 specifies.

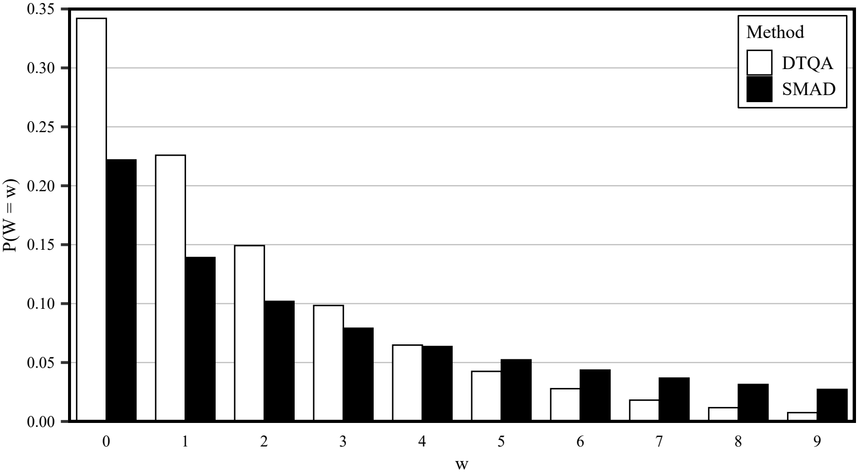

Figure 6.2: Distributions of waiting time (truncated) for the $SM/G/1$-queue ($\kappa = 50$, black bars) and the corresponding $GI/G/1$-queue (white bars).

Besides the apparent approximation error found for the renewal decomposition approach, the semi-Markov arrival decomposition approach notably outperforms the renewal decomposition technique not only for $\kappa = 50$, but for any state limit κ. This is an important finding as it suggests that the state space limitation method yields better approximations than the renewal assumption, even for tight state limits.

6.4.2 Performance results

We compute the waiting time distributions for a design of experiments where the service times are gamma-distributed. Given its flexibility, the gamma distribution allows for the modelling of a wide range of dispersion and is therefore well suited to represent the stochastic behaviour of the service process. Furthermore, the gamma distribution is well-defined by its shape and scale, which translates to the expected value and variance (see definitions in Chapter 4). We define two gamma distributions that share the same expected value, $E(B) = 10$.

Table 6.2: Expected values of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with the renewal decomposition approach (DTQA), the semi-Markov arrival decomposition approach (SMAD) for five state space limits κ, and simulation.

			SMAD(κ)					
ρ	scv	DTQA	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 50$	Sim
0.30	L / L	2.71	**2.76**	**2.74**	**2.74**	**2.75**	**2.74**	2.76 ± 0.04
0.45	L / L	4.96	5.01	5.03	**5.04**	**5.04**	**5.04**	5.12 ± 0.08
0.60	L / L	8.84	**9.51**	**9.50**	**9.31**	**9.31**	**9.31**	9.38 ± 0.15
0.80	L / L	21.76	21.61	**25.14**	**25.54**	**25.44**	**25.44**	24.96 ± 0.59
0.90	L / L	41.91	49.25	**57.22**	**56.13**	**54.37**	(n.a.)	55.41 ± 1.97
0.30	H / L	**3.84**	**3.93**	**3.87**	**3.86**	**3.86**	**3.86**	3.81 ± 0.13
0.45	H / L	7.30	7.68	7.57	**7.47**	**7.48**	**7.48**	7.40 ± 0.09
0.60	H / L	**13.64**	14.95	14.69	**13.86**	**13.80**	**13.79**	13.53 ± 0.46
0.80	H / L	**35.86**	32.54	**36.20**	**35.62**	**37.22**	**36.77**	36.50 ± 1.49
0.90	H / L	67.65	75.87	**81.76**	**81.37**	**81.24**	(n.a.)	82.39 ± 3.48

Notes: The service time variability parameters are encoded as L = Low ($scv(B) = 0.50$) and H = High ($scv(B) = 1.51$) at the upstream / downstream station, respectively. Simulation results show the 95% confidence interval, bold numbers lie within the confidence interval.

The first service time distribution has a low variability, $scv(B) = 0.50$, the second one has high variability, $scv(B) = 1.51$. We define five flow parameters $\lambda \in \{0.030, 0.044, 0.058, 0.077, 0.086\}$ to observe the tandem queue for five utilisation parameters, $\rho \in \{0.30, 0.45, 0.60, 0.80, 0.90\}$. We consider two tandem queue configurations. In the first configuration, the service time distribution at both queues has a low variability (L / L), in the second configuration, the downstream queue is unchanged and the upstream queue has a high variability service time (H / L). The performance results for the variability configurations (L / H) and (H / H) are presented in Appendix C.

Tables 6.2 shows the results for the expected values of waiting time at the downstream queue obtained with DTQA, SMAD(κ), and simulation. The utilisation parameters are varied as described above, the service time variability parameter configurations are (L / L) and (H / L). The 95% percentiles of waiting time are listed in the corresponding Table 6.3. The simulation results show the upper and lower bound of the 95% confidence interval. Bold numbers in both tables indicate that this value lies inside this confidence interval.

Comparing the results obtained with DTQA and simulation, we can identify the decomposition error defined and discussed in Chapter 4. Most of the expected waiting times computed with DTQA are outside of the confidence intervals. We again find that utilisation is the driving factor for decomposition error. For tandem queues with utilisation smaller than or equal to 80%, the decomposition error is smaller than 6%. We identify three cases where the decomposition error is greater than 10%. In the first configuration (L / L), decomposition error for the queues with utilisation $\rho = 0.8$ and $\rho = 0.9$ is 13% and 24%, respectively. In the second configuration (H / L), the decomposition error for the queue with utilisation $\rho = 0.9$ is 18%.

The results computed with SMAD converge towards the simulation results as the state space limit κ increases. In the first configuration of the design of experiments (L / L), the waiting times are already inside the confidence intervals for $\kappa = 1$ in the queues where $\rho = 0.30$ and $\rho = 0.60$. The expected waiting times for queue $\rho = 0.45$ are barely outside the confidence interval, however, the 95% percentiles are inside. For $\kappa = 10$ and $\kappa = 20$, all values lie inside their respective confidence intervals. We can see that the waiting times computed with SMAD only slightly differ for $\kappa = 20$ and $\kappa = 50$. This can especially be seen in the 95% percentiles of waiting times which indicates the robustness of the waiting time vectors computed with SMAD.

In summary, the waiting times computed with SMAD are approximate for small κ (that is, a tight state space limits). However, in our numerical examples, we only found few cases where SMAD(1) performs worse than DTQA. In general, when DTQA produces satisfactory approximations, the waiting times computed with SMAD(1) are similar to those computed with DTQA. However, in cases where DTQA produces low accuracy, SMAD(1) is a better approximation. Based on the performance results presented, we conclude that SMAD computes reasonably

Table 6.3: 95% percentiles of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with the renewal decomposition approach (DTQA), the semi-Markov arrival decomposition approach (SMAD) for five state space limits κ, and simulation.

			SMAD(κ)					
ρ	scv	DTQA	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 50$	Sim
0.30	L / L	**17**	**17**	**17**	**17**	**17**	**17**	17 ± 0
0.45	L / L	**26**	**26**	**26**	**25**	**25**	**25**	26 ± 1
0.60	L / L	37	**40**	**40**	**39**	**39**	**39**	40 ± 1
0.80	L / L	76	78	92	95	94	94	89 ± 2
0.90	L / L	219	**184**	**187**	**182**	**176**	(n.a.)	184 ± 11
0.30	H / L	**21**	22	22	22	22	22	20 ± 1
0.45	H / L	**33**	36	35	**34**	**34**	**34**	33 ± 1
0.60	H / L	**53**	57	56	**53**	**53**	**53**	52 ± 2
0.80	H / L	**118**	111	**126**	**125**	130	127	121 ± 5
0.90	H / L	202	237	**254**	**254**	**254**	(n.a.)	256 ± 12

Notes: The service time variability parameters are encoded as L = Low ($scv(B) = 0.50$) and H = High ($scv(B) = 1.51$) at the upstream / downstream station, respectively. Simulation results show the 95% confidence interval, bold numbers lie within the confidence interval.

accurate results of waiting time for $\kappa = 10$, if the tandem queue is low-traffic, and $\kappa = 20$ for heavy-traffic queues. In Tables 6.2 and 6.3 it can be seen that the results for $\kappa = 50$ are missing for $\rho = 0.90$. In these cases, the state space became too large to compute the stationary distribution of the Markov chain.

6.4.3 Computational complexity

After evaluating the performance of SMAD, we investigate its computational complexity. It is intuitive that the computational complexity increases with increasing state space limit κ. To explore this effect, we collected the computer times for the computation of the performance results presented above. The results have been computed on an Intel Core i7-8550U 64-bit machine, with 16 GB of RAM and

Table 6.4: Computer times in seconds for the renewal decomposition approach (DTQA) and the semi-Markov arrival decomposition approach (SMAD) for five state space limits κ.

		SMAD(κ)				
ρ	DTQA	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 50$
0.30	1	181	485	854	1,678	4,143
0.45	1	167	403	654	1,246	3,293
0.60	4	215	459	750	1,426	3,799
0.80	31	187	453	758	1,335	4,336
0.90	102	248	555	989	2,100	(n.a.)

2.8 gigahertz processor. Table 6.4 shows the computer times in seconds for the renewal decomposition approach (DTQA), and SMAD for five state space limits.

Table 6.4 shows that DTQA is the most efficient decomposition approach. For most queues, the computer time is well below one minute. On average, DTQA requires 38 seconds to compute the waiting time distribution. However, it should be noted that the computational complexity of DTQA increases with growing utilisation of the system. In the high-traffic queue ($\rho = 0.9$), the computational time required by DTQA is 1:42 minutes. This is due to the fact that the waiting time vector grows exponentially in high-traffic queues. Table 6.4 shows that the computational complexity of SMAD increases exponentially with growing state space limit κ. For $\kappa = 1$, the approach requires an average computer time of 3 minutes. For $\kappa = 10$, the average compute time is 14 minutes, for $\kappa = 20$ it is 26 minutes, and for $\kappa = 50$, it is 65 minutes. In contrast to DTQA, however, the computational complexity of SMAD only slightly increases with increasing utilisation of the queue.

As seen above, the computational efficiency of SMAD can be drastically increased by limiting the state space with small values of κ. However, as with any Markov chain, the computational complexity of SMAD depends on the entire state space, not just one component. Since the state space is three-dimensional, $(N^D, r, N^U) \in \mathbb{N} \times \{1 - R, ..., R - 1\} \times \{0, ..., \kappa\}$, appropriate upper and lower bounds must be defined for the state dimensions N^D and r. Therefore, two additional factors affect computational efficiency. First, the number of customers in the downstream queue N^D increases as the utilisation increases. As defined above, the maximum number of customers in the queue is unlimited. Thus, an upper bound $\hat{N}^D_{max}$ must be found in order to compute the stationary distribution. The upper bound $\hat{N}^D_{max}$ must be increased when analysing high-traffic queues, compared to the analyses of queues with low traffic intensity. Second, the maximum residual time R for the next event is influenced by the length of the probability vectors describing the inter-arrival and service times. For efficient implementation, it is useful to define an upper ($\hat{R}_U$) and a lower bound ($\hat{R}_L$) for R. The upper bound $\hat{R}_U$ is set to the maximum inter-arrival time and the lower bound $\hat{R}_L$ is set to the maximum service time. Note that $\hat{R}_L < 0$, thus, the maximum service time is multiplied by (-1).

To assess the complexity class of the algorithm, we first assume that the number of customers in the upstream $M/G/1$-queue is unbounded, as well. Therefore, equivalently to the upper bound $\hat{N}^D_{max}$, we define an upper bound $\hat{N}^U_{max}$ for the number of customers in the upstream queue. Let $m = \max\{\hat{N}^D_{max}, (\hat{R}_U + |\hat{R}_L|), \hat{N}^U_{max}\}$. Since the state space of the embedded Markov chain is three-dimensional, the computational complexity class of the $SM/G/1$-queue is $\mathcal{O}(m^3)$. However, due to the state space limitation method, the state component describing the number of customers in the upstream queue is bounded upwards by κ. With κ being a constant value, the computational complexity is directly proportional to κ and therefore, the algorithmic complexity reduces to $\mathcal{O}(m^2)$.

6.5 Chapter conclusion

Decomposition approaches for open queueing networks approximate the interconnecting streams as renewal processes. While this assumption allows for computationally efficient models, performance results obtained at downstream queues might be prone to considerable approximation errors. To overcome this problem, we propose SMAD, a decomposition method for tandem queues with Poisson arrivals and general service times. The novelty of this decomposition method is that a semi-Markov process (SMP) is used to model the connecting stream between the upstream $M/G/1$- and the downstream $G/G/1$-queue. Thus, SMAD captures the state-dependent inter-departure times in the upstream $M/G/1$-queue departure process.

To prevent state space explosion in the Markov chain of the downstream $SM/G/1$-queue, we introduce a state space limitation method for the embedded Markov chain of the SMP. The limited Markov chain observes the SMP only if the number of customers in the $M/G/1$-queue is smaller than or equal to the upper bound κ. We demonstrate that the stationary distribution of the limited embedded Markov chain is equal to the original stationary distribution, $\hat{\pi}(i) = \pi(i)$ for all $i = 0, 1, ..., \kappa - 1$. Therefore, tight space limits are computationally attractive, but lead to approximations in the computation of the state-dependent departure stream.

Our numerical results show that the waiting times in the downstream queue obtained with SMAD are reasonably accurate when the state space is limited and converge arbitrarily accurate with increasing state space size. Compared to the renewal decomposition approach (DTQA), SMAD computes better approximations than DTQA when the state space limit is tight (that is, κ is small). This is an important finding as it suggests that the state space limitation method yields better approximations than the renewal assumption, even for tight state limits. Increasing the state limit κ generally improves the quality of the results. In contrast to DTQA (which is prone to decomposition error), the waiting times computed with SMAD lie inside the confidence interval of the simulation model for reasonably large state limits. In conclusion, the semi-Markov arrival decomposition approach allows us to determine a satisfactory approximate solution that requires little computational effort. When better approximations are needed, SMAD converges arbitrarily accurate with increasing computational expenditure.

7 Conclusion

This chapter summarises the major research contributions of the thesis and gives an outlook on further research directions.

7.1 Summary

Discrete-time queueing models are well suited to compute key performance indicators of a stochastic system with little computational effort. For example, discrete-time queuing models allow to compute the throughput time of a shipment or the number of buffer slots to be provided in front of a machine. Applying discrete-time queueing models offers an advantage over continuous-time modelling in that the entire probability distribution of the performance parameters can be calculated. Thus, we can determine the performance of the system in more detail since we can report not only the averages but the 95%- or 99%-percentiles of the probability distribution.

To compute steady-state performance measures in the queueing systems of an open queueing network, decomposition methods are often the only feasible and computationally efficient approach. A decomposition approach partitions the network into individual queues and analyses them in isolation. The approach is based on the assumption that a renewal process can approximate the point process of the departure stream of upstream stations, and thus an independent analysis of the queueing systems is possible. Although the assumption of independence allows for highly efficient computation, performance results may be subject severe approximation errors.

This thesis considers discrete-time open tandem networks with Poisson arrivals and general service times to analyse the approximation error of the renewal decomposition approach, investigate its origins, and introduce a refined decomposition approach that converges arbitrarily accurate.

Approximation quality of the renewal decomposition approach

In the first part of the thesis, we conduct a simulation study to analyse the approximation quality of the renewal decomposition approach. In our design of experiments, the service times are gamma-distributed with variability parameters in the interval $(0.1, 3.0)$. We define the expected service times and the flow parameter of the external Poisson arrival process to cover low, medium and high utilisation in the tandem queue. We compute the expected value (Study I) and the 95th-percentile (Study II) of waiting time in the downstream queue using the renewal decomposition approach and simulation, and define decomposition error as the relative error between both measures, respectively. We find the relative errors in the range of -21.9% and 32.5% (referring to Study I) and -30.8% and 36.7% (referring to Study II). The mean absolute values of decomposition error equal 3.93% (4.51%) in Study I (Study II).

We deploy the variability parameters of the service time distributions, the variability of the connecting stream, and utilisation as independent variables for the point and interval estimates of decomposition error. The point estimates are based on multiple linear regression, the interval estimates are based on quantile regression. Both estimation methods are applied for Study I and Study II, respectively. Using test data, we demonstrate that the regression models provide accurate forecasts and precise confidence intervals for decomposition error. Further, we use the ANOVA of the models to reveal major influencing factors on the renewal approximation quality: We find utilisation to be the main driver for decomposition error since in low-traffic queueing systems the mean absolute decomposition error is significantly lower than the mean absolute errors in the entire data set. Severe absolute decomposition errors are only observed in heavy-traffic systems. Finally, we find that the downstream arrival process variability determines whether the decomposition approach overestimates or underestimates the waiting time.

Output dynamics of the upstream queue

The second part of the thesis focuses on the analysis of the output dynamics of the upstream $M/G/1$-queue. It is well known that the departure stream is a point process which is approximated as a renewal process in the decomposition approach. By observing the state of the queueing system only in departure instances, we show that consecutive departure times are sequentially dependent since the inter-departure time D_k of customer k depends on the system state that customer $k-1$ left behind. Consequently, the overlay of the renewal arrival and the renewal service processes generates non-renewal departures. To compute the ϕ-lag auto-correlation in the departure stream, we model the $M/G/1$-queue as a discrete-time Markov chain. Based on the stationary distribution, we derive formulas to compute the joint inter-departure time distribution of two departure times that are ϕ instances apart. We present numerical results for the auto-correlation in $M/D/1$- and $M/G/1$-queues for several lags and utilisation parameters. The numerical results show that auto-correlation is positive and monotonically decreases for increasing lags ϕ and is a non-monotone function of the utilisation parameter.

Based on these results – and taking into account the findings from the previous chapter – we can explain the reasons for the approximation errors in the renewal decomposition approach. Since auto-correlation is non-monotone for the utilisation parameter, we conclude that downstream performance measures can be computed exactly for the extreme cases $\rho = 0$ and $\rho = 1$. In these cases, the downstream arrival process is renewal since it is only determined by the external arrival process (for the case $\rho = 0$) or the renewal upstream service process (for the case $\rho = 1$). This finding is consistent to the heavy-traffic bottleneck phenomenon described in the literature. For steady-state systems, $\rho \in (0, 1)$, we can identify the dependency of the waiting time from the serial dependency of inter-arrival times at the downstream queue. The positive auto-correlation in the departure stream affect the flow factor and variability compared to an i.i.d. sampling of inter-departure times since a long (short) inter-departure time is likely to be followed by another long (short) inter-departure time. Consequently, results computed with the renewal decomposition approach are flawed and – depending on the arrival stream variability – overestimate or underestimate the waiting time.

Refined decomposition approach for tandem queues

In the third part of the thesis, we present the semi-Markov decomposition approach (SMAD) for the discrete-time tandem queue. The novelty of this approach is that we do not assume the connecting stream to be renewal, but model the output process of the upstream $M/G/1$-queue as a semi-Markov process (SMP). The SMP captures the sequential dependencies of the departure times described in the previous chapter. For the downstream system, we deploy a discrete-time $SM/G/1$-queue, which receives the SMP as input and thus considers the state-dependent departure times from the upstream queue for queueing analysis.

The SMP is composed of an embedded discrete-time Markov chain and a conditional probability function that describes the state-dependent inter-departure time. The departure time of the upstream queue is equal to the service time if the system is not empty after the departure instance, and equal to the sum of the residual arrival time and the service time if the system starves after the departure instance. To avoid state space explosion in the downstream queue, we limit the state space of the embedded Markov chain model to an upper bound κ. We demonstrate how to compute the corresponding limited transition probability matrix and proof that the stationary distribution of the limited embedded Markov chain is equal to the original stationary distribution, $\hat{\pi}(i) = \pi(i)$ for all $i = 0, 1, ..., \kappa - 1$.

The state space limit κ introduces approximations to the performance results, but increases the computational efficiency of the decomposition approach. We present numerical results where we compare the waiting time computed with SMAD for several state limits κ with the renewal decomposition approach and simulation. SMAD computes better approximations than the renewal decomposition approach, even when the state space limit is tight (that is, κ is small). This is an important finding as it suggests that the state space limitation method yields better approximations than the renewal assumption. Increasing the state limit κ generally improves the quality of the results. In conclusion, SMAD produces reasonably accurate results when the state space is limited. For increasing state limit κ, we conclude that the performance results computed with SMAD are arbitrarily accurate.

7.2 Conclusion of the thesis

To conclude the thesis, computing accurate performance measures for queueing networks with little computational effort is crucial. The computational results presented in this thesis demonstrate that the renewal decomposition approach is the most efficient one. However, the performance measures may be prone to considerable approximation errors in high-traffic queues with highly fluctuating downstream arrivals. To overcome the problem of reporting potentially flawed data, we introduce point and interval estimates to predict decomposition error efficiently. These statistical estimates allow a highly accurate forecast of the approximation quality of the decomposition technique only based on the input parameters of the queues.

With SMAD, we present a suitable alternative decomposition approach for the tandem queue in case the approximation quality achieved with the renewal decomposition is found to be unsatisfactory. Our numerical results show that the accuracy of SMAD outperforms the renewal decomposition approach, even if the state space is limited, and converges arbitrarily accurate with increasing state space limit. Increasing the accuracy accomplished with SMAD, however, requires substantial computational effort. Therefore, the renewal decomposition approach is not redundant and instead should be applied whenever decomposition error is estimated to be reasonably small. In conclusion, SMAD introduces a trade-off as to whether the parameter κ should be set small to increase computational efficiency or large so that converging accurate results are obtained.

7.3 Outlook

Based on the research questions answered in this thesis and the conclusions drawn, we identify future research perspectives in three areas.

Firstly, as discussed in detail in the thesis, SMAD is at a disadvantage compared to the renewal decomposition approach regarding computational efficiency. Since the approach is based on a Markov chain with a three-dimensional state space, the computational complexity grows exponentially as the state space limit increases. Therefore, we identify the need to study alternative stochastic models for the downstream queue. Haßlinger (2000) has presented a computational method for the waiting time distribution in the $SM/G/1$-queue based on a Wiener-Hopf factorisation. This approach has already led to solid efficiency gains for the computational efficiency for the waiting time distribution of the $GI/G/1$-queue (Grassmann and Jain 1989). The open research question is whether the state space limitation method presented in this thesis is compatible with Haßlinger's model.

Secondly, additional research is needed to transfer the approximation estimators of the renewal decomposition approach to general network types. It might be valuable for practitioners to be alert if the expected approximation quality is low. For tandem queues, this thesis identified the conditions under which the renewal decomposition approach produces low accuracy. Since we found reasonable explanations based on the analysis of auto-correlation in the connecting stream, we expect to transfer these findings to general network types, as well. However, it is interesting to investigate the approximation errors behind splits and superpositions since Poisson flows conditionally approximate thinned and superposed flows very well. Furthermore, the estimation methods based on regression models presented here require substantial effort for training and validation, which is no longer feasible for general network types. Stochastic ordering is an alternative approach to develop valid estimates for the approximation quality for the renewal decomposition approach (Shaked and Shanthikumar 2007).

Thirdly, research is needed to develop a general framework for decomposing discrete-time queueing networks with general service times and semi-Markov interconnecting streams. The results presented in this thesis demonstrate that considering sequential dependencies in the connecting streams is essential for better approximations. Haßlinger and Rieger (1996) discuss a framework for general discrete-time network types, including splitting and superposition of auto-correlated streams. This approach, however, is unfeasible due to state space explosion. Therefore, there is a need to transfer the state space limitation method presented in this thesis to the discrete-time $SM/G/1$-queue.

A Regression analyses for decomposition error in tandem queues with bottlenecks

The regression tables presented here appear in Data in Brief (Jacobi and Furmans 2022c).

Table A.1 shows the regression coefficients for the analysis of decomposition error for tandem queues with downstream bottlenecks, where the dependent (explanatory) variable is $\Delta(E)$, see equation (4.5).

Table A.2 shows the regression coefficients for the analysis of decomposition error for tandem queues with downstream bottlenecks, where the dependent (explanatory) variable is $\Delta(\sigma)$, see equation (4.6).

Note that, in contrast to the regression analyses presented in Chapter 4 (Tables 4.3 and 4.4), the regression analyses presented here differentiate the impact of the upstream (ρ_u) and downstream (ρ_d) utilisation parameter, respectively.

Table A.1: OLS and quantile regression estimates in Study I with downstream bottlenecks.

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
const.	-0.0704***	-0.2019***	-0.1530***	-0.1332***	-0.0426**	-0.0184	-0.0184
	0.0072	0.0492	0.0359	0.0266	0.0130	0.0176	0.0211
$scv(B_u)$	0.1314***	0.1222	0.2307**	0.2125*	0.1324**	0.0767	0.1116
	0.0233	0.1615	0.0989	0.0720	0.0411	0.0542	0.0708
$scv(B_d)$	0.0082***	-0.0210	0.0079	0.0099*	0.0059	0.0057	0.0052
	0.0015	0.0230	0.0116	0.0051	0.0036	0.0041	0.0044
$scv(A_d)$	-0.3605***	-0.4273	-0.5899**	-0.5556***	-0.3532***	-0.2373**	-0.2985**
	0.0473	0.3194	0.1963	0.1492	0.0847	0.1126	0.1434
ρ_u	0.1000***	0.1741*	0.1725**	0.1558**	0.0964***	0.0680**	0.0848**
	0.0145	0.0961	0.0619	0.0477	0.0249	0.0321	0.0407
ρ_d	-0.0255***	-0.0738***	-0.0442***	-0.0361***	-0.0090**	-0.0098**	-0.0036
	0.0020	0.0121	0.0094	0.0055	0.0040	0.0048	0.0049
$scv(B_u)$	0.0019	-0.0319*	-0.0087	-0.0071**	0.0054**	0.0075**	0.0056*
$\times\ scv(B_d)$	0.0022	0.0182	0.0127	0.0035	0.0020	0.0026	0.0032
$scv(B_u)$	0.0097***	0.0071	0.0006	0.0009	0.0259***	0.0261***	0.0242***
$\times\ scv(A_d)$	0.0018	0.0087	0.0074	0.0049	0.0064	0.0062	0.0063
$scv(B_u)$	0.1006***	0.1671	0.1721**	0.1668**	0.0914**	0.0486	0.0786
$\times\ \rho_u$	0.0189	0.1234	0.0694	0.0546	0.0324	0.0420	0.0551
$scv(B_u)$	0.0045*	-0.0202	0.0058	0.0004	0.0071***	0.0059**	0.0049*
$\times\ \rho_d$	0.0020	0.0208	0.0138	0.0056	0.0018	0.0021	0.0025

OLS and quantile regression estimates in Study I with bottlenecks (continued).

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
$scv(B_d)$	0.0142***	0.0329**	0.0229**	0.0204**	0.0028	0.0005	0.0037
$\times\ scv(A_d)$	0.0024	0.0159	0.0103	0.0067	0.0060	0.0066	0.0067
$scv(B_d)$	-0.0058**	0.0313	0.0026	0.0009	-0.0061*	-0.0057	-0.0064
$\times\ \rho_u$	0.0018	0.0192	0.0111	0.0044	0.0033	0.0036	0.0043
$scv(B_d)$	0.0091***	-0.0107	0.0020	0.0039	0.0082***	0.0064***	0.0059***
$\times\ \rho_d$	0.0017	0.0147	0.0079	0.0030	0.0014	0.0014	0.0017
$scv(A_d)$	0.1022***	0.0953	0.1661**	0.1548***	0.0931***	0.0675**	0.0743**
$\times\ \rho_u$	0.0108	0.0784	0.0518	0.0383	0.0202	0.0265	0.0314
$scv(A_d)$	-0.0712***	-0.0356	-0.0700***	-0.0647***	-0.0643***	-0.0673***	-0.0584***
$\times\ \rho_d$	0.0044	0.0281	0.0192	0.0133	0.0109	0.0115	0.0120
ρ_u	0.0008	-0.0080	-0.0038	-0.0048*	0.0097**	0.0107**	0.0117**
$\times\ \rho_d$	0.0023	0.0053	0.0042	0.0028	0.0036	0.0034	0.0042
Adj. / Ps. R^2	0.8329	0.6982	0.6870	0.7168	0.7146	0.7592	0.8241

Notes: Standardised regression coefficients with standard errors listed below. The standard errors of quantile regression estimates are based on 100 bootstrapping replications. The sample size 534.

* $p < .1$

** $p < .05$

*** $p < .001$

Table A.2: OLS and quantile regression estimates in Study II with downstream bottlenecks.

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
const.	-0.0831***	-0.2828***	-0.2213***	-0.1591***	-0.0492**	-0.0275	-0.0156
	0.0087	0.0695	0.0560	0.0414	0.0149	0.0173	0.0186
$scv(B_u)$	0.1549***	0.2868	0.3651**	0.2204*	0.1474**	0.0941*	0.0841
	0.0281	0.2243	0.1577	0.1132	0.0479	0.0549	0.0580
$scv(B_d)$	0.0101***	-0.0162	0.0099	0.0097	0.0033	0.0041	0.0056
	0.0018	0.0250	0.0137	0.0074	0.0041	0.0039	0.0043
$scv(A_d)$	-0.4181***	-0.7720*	-0.8768**	-0.5957**	-0.3892***	-0.2792**	-0.2494**
	0.0569	0.4493	0.3167	0.2368	0.1006	0.1145	0.1197
ρ_u	0.1184***	0.2926**	0.2665**	0.1663**	0.1078***	0.0805**	0.0773**
	0.0175	0.1410	0.1032	0.0751	0.0301	0.0333	0.0343
ρ_d	-0.0322***	-0.0818***	-0.0535***	-0.0457***	-0.0117**	-0.0146**	-0.0165**
	0.0025	0.0130	0.0107	0.0084	0.0043	0.0044	0.0055
$scv(B_u)$	0.0014	-0.0181	-0.0086	-0.0085	0.0068**	0.0060**	0.0075**
$\times\ scv(B_d)$	0.0026	0.0182	0.0152	0.0057	0.0025	0.0028	0.0032
$scv(B_u)$	0.0068**	0.0046	0.0004	-0.0015	0.0225***	0.0218**	0.0215**
$\times\ scv(A_d)$	0.0022	0.0115	0.0086	0.0068	0.0065	0.0067	0.0071
$scv(B_u)$	0.1202***	0.3086*	0.2814**	0.1779**	0.1013**	0.0594	0.0535
$\times\ \rho_u$	0.0227	0.1856	0.1211	0.0827	0.0368	0.0421	0.0441
$scv(B_u)$	0.0053*	-0.0228	0.0085	-0.0025	0.0064**	0.0077***	0.0074**
$\times\ \rho_d$	0.0025	0.0236	0.0185	0.0091	0.0020	0.0020	0.0025

OLS and quantile regression estimates in Study II with bottlenecks (continued).

	OLS	Q(.005)	Q(.025)	Q(.050)	Q(.950)	Q(.975)	Q(.995)
$scv(B_d)$	0.0197***	0.0287*	0.0254**	0.0273***	0.0069	0.0071	0.0064
$\times\ scv(A_d)$	0.0029	0.0166	0.0114	0.0070	0.0061	0.0061	0.0082
$scv(B_d)$	-0.0054*	0.0334	0.0054	-0.0008	-0.0083**	-0.0074**	-0.0055
$\times\ \rho_u$	0.0022	0.0243	0.0150	0.0068	0.0027	0.0031	0.0039
$scv(B_d)$	0.0087***	-0.0120	-0.0010	0.0041	0.0060***	0.0052***	0.0056**
$\times\ \rho_d$	0.0021	0.0168	0.0097	0.0045	0.0016	0.0016	0.0018
$scv(A_d)$	0.1162***	0.1625	0.2289**	0.1711**	0.1048***	0.0827**	0.0728**
$\times\ \rho_u$	0.0130	0.0987	0.0789	0.0646	0.0256	0.0278	0.0295
$scv(A_d)$	-0.0830***	-0.0385	-0.0854**	-0.0841***	-0.0675***	-0.0740***	-0.0761***
$\times\ \rho_d$	0.0052	0.0354	0.0290	0.0209	0.0096	0.0091	0.0111
ρ_u	-0.0001	-0.0077	-0.0047	-0.0022	0.0091**	0.0082**	0.0074*
$\times\ \rho_d$	0.0028	0.0075	0.0061	0.0050	0.0030	0.0031	0.0038
Adj. / Ps. R^2	0.8263	0.6366	0.6020	0.6260	0.7030	0.7529	0.8218

Notes: Standardised regression coefficients with standard errors listed below. The standard errors of quantile regression estimates are based on 100 bootstrapping replications. The sample size 534.

* $p < .1$

** $p < .05$

*** $p < .001$

B Proof for the state space reduction method

In the following, we present a proof for $\hat{\pi}(i) = \pi(i)$ for all $i = 0, 1, ..., \kappa - 1$ in the state space reduction method. We first consider the case $\kappa = 1$, that is, the reduction of the stochastic matrix $\mathbf{P}$ to a 2×2 matrix $\mathbf{P}'$, and then the case $\kappa > 1$ for larger matrices $\mathbf{P}'$.

Case 1: $\kappa = 1$

Consider the reduction of the stochastic matrix $\mathbf{P}$ to a 2×2 matrix $\mathbf{P}'$, such that

$$\mathbf{P}' = \begin{bmatrix} p_{00} & 1 - p_{00} \\ \hat{p}_{10} & 1 - \hat{p}_{10} \end{bmatrix}, \tag{B.1}$$

where

$$\begin{aligned} \hat{p}_{10} &= \pi(1) \cdot p_{10} \cdot \left(\sum_{j=1}^{\infty} \pi(j) \right)^{-1} \\ &= \pi(1) \cdot p_{10} \cdot (1 - \pi(0))^{-1}. \end{aligned} \tag{B.2}$$

Suppose $\mathbf{P}'$ has an invariant probability vector $\hat{\pi}$ that satisfies

$$\begin{aligned} \hat{\pi} &= \mathbf{P}' \hat{\pi}, \\ \hat{\pi} \mathbf{e} &= 1. \end{aligned} \tag{B.3}$$

It follows that

$$\begin{aligned}\hat{\pi}(0) &= \hat{\pi}(0) \cdot p_{00} + \hat{\pi}(1) \cdot \hat{p}_{10} \\ &= \hat{\pi}(0) \cdot p_{00} + \hat{\pi}(1) \cdot \pi(1) \cdot p_{10} \cdot (1 - \pi(0))^{-1}.\end{aligned} \tag{B.4}$$

Since in the original problem $\pi = \mathbf{P}\pi$,

$$\begin{aligned} & \pi(0) = \pi(0) \cdot p_{00} + \pi(1) \cdot p_{10} \\ \Leftrightarrow \quad & \pi(1) = \frac{\pi(0) \cdot (1 - p_{00})}{p_{10}}\end{aligned} \tag{B.5}$$

holds, and therefore equation (B.4) can be re-written as

$$\begin{aligned} & \hat{\pi}(0) = \hat{\pi}(0) \cdot p_{00} + \hat{\pi}(1) \cdot \pi(0) \cdot (1 - p_{00}) \cdot (1 - \pi(0))^{-1} \\ \Leftrightarrow \quad & \hat{\pi}(0) = \hat{\pi}(0) \cdot p_{00} + \frac{1 - \hat{\pi}(0)}{1 - \pi(0)} \cdot \pi(0) \cdot (1 - p_{00}) \\ \Leftrightarrow \quad & \hat{\pi}(0) \cdot (1 - p_{00}) = \frac{1 - \hat{\pi}(0)}{1 - \pi(0)} \cdot \pi(0) \cdot (1 - p_{00}) \\ \Leftrightarrow \quad & \frac{\hat{\pi}(0) \cdot (1 - p_{00})}{1 - \hat{\pi}(0)} = \frac{\pi(0) \cdot (1 - p_{00})}{1 - \pi(0)} \\ \Leftrightarrow \quad & \hat{\pi}(0) = \pi(0) \qquad \square \end{aligned} \tag{B.6}$$

Case 2: $\kappa > 1$

We will first show that $\hat{\pi}(i) = \pi(i)$ for all $i = 0, 1, 2, ..., \kappa - 2$, and then proof $\hat{\pi}(\kappa - 1) = \pi(\kappa - 1)$.

Consider the reduction of the stochastic matrix $\mathbf{P}$ to a $(\kappa + 1) \times (\kappa + 1)$ matrix $\mathbf{P}'$, where $\kappa > 1$, such that

$$\mathbf{P}' = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0,\kappa-1} & \hat{p}_{0,n} \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1,\kappa-1} & \hat{p}_{1,n} \\ 0 & p_{21} & p_{22} & \cdots & p_{2,\kappa-1} & \hat{p}_{2,n} \\ 0 & 0 & p_{32} & \cdots & p_{3,\kappa-1} & \hat{p}_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{\kappa-1,\kappa-1} & \hat{p}_{\kappa-1,n} \\ 0 & 0 & 0 & \cdots & \hat{p}_{\kappa,\kappa-1} & \hat{p}_{\kappa,\kappa} \end{bmatrix}, \tag{B.7}$$

where

$$\begin{aligned} \hat{p}_{\kappa,\kappa-1} &= \pi(\kappa) \cdot p_{\kappa,\kappa-1} \cdot \left(\sum_{j=n}^{\infty} \pi(j) \right)^{-1} \\ &= \pi(\kappa) \cdot p_{\kappa,\kappa-1} \cdot \left(1 - \sum_{j=0}^{\kappa-1} \pi(j) \right)^{-1}. \end{aligned} \tag{B.8}$$

We first show that $\hat{\pi}(i) = \pi(i)$ for $i = 0, 1, ..., \kappa - 2$. Suppose $\mathbf{P}'$ has an invariant probability vector $\hat{\pi}$ that satisfies

$$\begin{aligned} \hat{\pi} &= \mathbf{P}'\hat{\pi}, \\ \hat{\pi}\mathbf{e} &= 1. \end{aligned} \tag{B.9}$$

It follows that

$$\begin{aligned} & \hat{\pi}(0) = \hat{\pi}(0) \cdot p_{00} + \hat{\pi}(1) \cdot p_{10} \\ \Leftrightarrow \quad & p_{00} = \frac{\hat{\pi}(0) - \hat{\pi}(1) \cdot p_{10}}{\hat{\pi}(0)}. \end{aligned} \tag{B.10}$$

Since

$$\hat{\pi}(1) = \frac{\hat{\pi}(0) \cdot (1 - p_{00})}{p_{10}}, \tag{B.11}$$

we find

$$p_{00} = \frac{\hat{\pi}(0) - \hat{\pi}(0) \cdot (1 - p_{10})}{\hat{\pi}(0)}. \tag{B.12}$$

Equations (B.10), (B.11), and (B.12) also hold for $\mathbf{P}$ and π, therefore

$$\begin{aligned} \frac{\hat{\pi}(0) - \hat{\pi}(0) \cdot (1 - p_{10})}{\hat{\pi}(0)} &= \frac{\pi(0) - \pi(0) \cdot (1 - p_{10})}{\pi(0)} \\ \Leftrightarrow \quad \hat{\pi}(0) &= \pi(0). \end{aligned} \tag{B.13}$$

Starting with equation (B.11), we exploit the triangular form of the matrix $\mathbf{P}'$ to iteratively solve the linear equilibrium equations for $\pi(i)$ for all $i = 1, 2, ..., \kappa - 2$, and come to the conclusion that

$$\hat{\pi}(i) = \pi(i) \quad \forall i = 1, 2, ..., \kappa - 2. \tag{B.14}$$

In the following, we proof $\hat{\pi}(\kappa - 1) = \pi(\kappa - 1)$. Due to proposition (B.3),

$$\begin{aligned} \hat{\pi}(\kappa - 1) &= \sum_{i=0}^{\kappa-1} \hat{\pi}(i) \cdot p_{i,\kappa-1} + \hat{\pi}_n \cdot \hat{p}_{\kappa,\kappa-1} \\ &= \sum_{i=0}^{\kappa-2} \hat{\pi}(i) \cdot p_{i,\kappa-1} + \hat{\pi}(\kappa - 1) \cdot p_{\kappa-1,\kappa-1} + \hat{\pi}_n \cdot \hat{p}_{\kappa,\kappa-1}. \end{aligned} \tag{B.15}$$

From finding (B.14) it follows

$$\sum_{i=0}^{\kappa-2} \hat{\pi}(i) \cdot p_{i,\kappa-1} = \sum_{i=0}^{\kappa-2} \pi(i) \cdot p_{i,\kappa-1} \tag{B.16}$$

and for reasons of simplicity, we substitute

$$\begin{aligned}
\mathcal{X} &= \sum_{i=0}^{\kappa-2} \hat{\pi}(i) \cdot p_{i,\kappa-1} = \sum_{i=0}^{\kappa-2} \pi(i) \cdot p_{i,\kappa-1} \\
\mathcal{Y} &= 1 - p_{\kappa-1,\kappa-1}.
\end{aligned} \tag{B.17}$$

Equation (B.15) can be re-written as:

$$\begin{aligned}
\hat{\pi}(\kappa-1) &= \mathcal{X} + \hat{\pi}(\kappa-1) \cdot p_{\kappa-1,\kappa-1} + \hat{\pi}_n \cdot \hat{p}_{\kappa,\kappa-1} \\
&= \mathcal{X} + \hat{\pi}(\kappa-1) \cdot p_{\kappa-1,\kappa-1} + \hat{\pi}_n \cdot \pi(\kappa) \cdot p_{\kappa,\kappa-1} \cdot \left(1 - \sum_{j=0}^{\kappa-1} \pi(j)\right)^{-1} \\
\Leftrightarrow \quad \pi(\kappa) &= \Big(\hat{\pi}(\kappa-1) \cdot \big(1 - p_{\kappa-1,\kappa-1}\big) - \mathcal{X}\Big) \cdot \left(1 - \sum_{j=0}^{\kappa-1} \pi(j)\right) \cdot \Big(\hat{\pi}_n \cdot p_{\kappa,\kappa-1}\Big)^{-1} \\
&= \Big(\hat{\pi}(\kappa-1) \cdot \mathcal{Y} - \mathcal{X}\Big) \cdot \left(1 - \sum_{j=0}^{\kappa-1} \pi(j)\right) \cdot \left(\left(1 - \sum_{j=0}^{\kappa-1} \hat{\pi}_j\right) \cdot p_{\kappa,\kappa-1}\right)^{-1} \\
&= \Big(\hat{\pi}(\kappa-1) \cdot \mathcal{Y} - \mathcal{X}\Big) \cdot \Big(1 - \big(\mathcal{X} + \pi(\kappa-1)\big)\Big) \\
&\qquad \cdot \Big(\big(1 - \big(\mathcal{X} + \hat{\pi}(\kappa-1)\big)\big) \cdot p_{\kappa,\kappa-1}\Big)^{-1}
\end{aligned} \tag{B.18}$$

In the original problem $\pi = \mathbf{P}\pi$, and therefore

$$\begin{aligned}
\pi(\kappa-1) &= \sum_{i=0}^{n} \pi(i) \cdot p_{i,\kappa-1} \\
&= \sum_{i=0}^{\kappa-2} \pi(i) \cdot p_{i,\kappa-1} + \pi(\kappa-1) \cdot p_{\kappa-1,\kappa-1} + \pi(\kappa) \cdot p_{\kappa,\kappa-1} \\
&= \mathcal{X} + \pi(\kappa-1) \cdot p_{\kappa-1,\kappa-1} + \pi(\kappa) \cdot p_{\kappa,\kappa-1}
\end{aligned} \tag{B.19}$$

holds and can be re-written as

$$\pi(\kappa) = \Big(\pi(\kappa-1)\cdot\big(1-p_{\kappa-1,\kappa-1}\big)-\mathcal{X}\Big)\cdot\big(p_{\kappa,\kappa-1}\big)^{-1}. \tag{B.20}$$

Therefore we can state (B.18) = (B.20), and it follows:

$$\begin{aligned}
\frac{\big(\hat{\pi}(\kappa-1)\cdot\mathcal{Y}-\mathcal{X}\big)\cdot\big(1-\big(\mathcal{X}+\pi(\kappa-1)\big)\big)}{\big(1-\big(\mathcal{X}+\hat{\pi}(\kappa-1)\big)\big)\cdot p_{\kappa,\kappa-1}} &= \frac{\pi(\kappa-1)\cdot\mathcal{Y}-\mathcal{X}}{p_{\kappa,\kappa-1}}\\
\Leftrightarrow \quad \frac{\hat{\pi}(\kappa-1)\cdot\mathcal{Y}-\mathcal{X}}{\big(1-\big(\mathcal{X}+\hat{\pi}(\kappa-1)\big)\big)\cdot p_{\kappa,\kappa-1}} &= \frac{\pi(\kappa-1)\cdot\mathcal{Y}-\mathcal{X}}{\big(1-\big(\mathcal{X}+\pi(\kappa-1)\big)\cdot p_{\kappa,\kappa-1}}\\
\Leftrightarrow \quad \frac{\hat{\pi}(\kappa-1)\cdot\mathcal{Y}-\mathcal{X}}{1-\big(\mathcal{X}+\hat{\pi}(\kappa-1)\big)} &= \frac{\pi(\kappa-1)\cdot\mathcal{Y}-\mathcal{X}}{1-\big(\mathcal{X}+\pi(\kappa-1)\big)}\\
\Leftrightarrow \quad \hat{\pi}(\kappa-1) &= \pi(\kappa-1) \qquad \square
\end{aligned} \tag{B.21}$$

C Continued performance results of the semi-Markov arrival decomposition approach

In the following, we present continued performance results for the semi-Markov arrival decomposition approach for the variability configurations (L / H) and (H / H). Note that, in contrast to the results presented in Section 6.4.2, the maximum state space limit is $\kappa = 30$, and not $\kappa = 50$.

Table C.1 shows the expected values of waiting time in the downstream queue obtained with DTQA, SMAD, and simulation. Continuing the results presented in Table 6.2, in this table the variability parameter at the downstream queue is high ($scv(B^D) = 1.51$).

Table C.2 shows the .95 percentile of waiting time in the downstream queue obtained with DTQA, SMAD, and simulation. Continuing the results presented in Table 6.3, in this table the variability parameter at the downstream queue is high ($scv(B^D) = 1.51$).

Table C.1: Expected values of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with the renewal decomposition approach (DTQA), the semi-Markov arrival decomposition approach (SMAD) for five state space limits κ, and simulation (cont. Table 6.2).

			SMAD(κ)					
ρ	scv	DTQA	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 30$	Sim
0.30	L / H	**4.93**	**5.05**	**5.01**	**5.01**	**5.01**	**5.01**	5.02 ± 0.10
0.45	L / H	9.22	**9.57**	**9.50**	**9.47**	**9.47**	**9.46**	9.55 ± 0.23
0.60	L / H	16.46	**17.19**	17.90	**17.81**	**17.79**	**17.78**	17.51 ± 0.35
0.80	L / H	40.84	46.72	**50.73**	**49.60**	**49.57**	**49.49**	49.93 ± 1.12
0.90	L / H	77.57	99.84	109.76	**113.91**	**115.05**	**114.49**	113.47 ± 5.31
0.30	H / H	**5.93**	6.10	**6.06**	**6.04**	**6.04**	**6.04**	5.95 ± 0.10
0.45	H / H	**11.43**	**11.74**	**11.66**	**11.52**	**11.51**	**11.51**	11.48 ± 0.28
0.60	H / H	**23.38**	22.39	24.07	**23.87**	**23.74**	**23.72**	21.26 ± 0.47
0.80	H / H	54.53	53.55	**58.91**	**58.76**	**56.77**	**56.27**	57.66 ± 1.71
0.90	H / H	96.47	102.06	114.41	**119.79**	**122.71**	(n.a.)	122.55 ± 5.94

Notes: The service time variability parameters are encoded as L = Low ($scv(B) = 0.50$) and H = High ($scv(B) = 1.51$) at the upstream / downstream station, respectively. Simulation results show the 95% confidence interval, bold numbers lie within the confidence interval.

Table C.2: 95% percentiles of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with the renewal decomposition approach (DTQA), the semi-Markov arrival decomposition approach (SMAD) for five state space limits κ, and simulation (cont. Table 6.3).

			SMAD(κ)					
ρ	scv	DTQA	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 30$	Sim
0.30	L / H	**31**	**32**	**32**	**32**	**32**	**32**	31 ± 1
0.45	L / H	**48**	**49**	**49**	**49**	**49**	**49**	49 ± 1
0.60	L / H	**72**	**75**	79	**78**	**78**	**78**	76 ± 2
0.80	L / H	145	165	**177**	**176**	**176**	**175**	169 ± 6
0.90	L / H	245	321	352	**366**	**371**	**370**	392 ± 27
0.30	H / H	**35**	**36**	**36**	**36**	**36**	**36**	35 ± 1
0.45	H / H	**55**	57	**56**	**56**	**56**	**56**	55 ± 1
0.60	H / H	93	92	100	99	**98**	**98**	96 ± 2
0.80	H / H	184	186	205	206	**198**	**197**	197 ± 6
0.90	H / H	294	333	**373**	**395**	**408**	(n.a.)	388 ± 22

Notes: The service time variability parameters are encoded as L = Low ($scv(B) = 0.50$) and H = High ($scv(B) = 1.51$) at the upstream / downstream station, respectively. Simulation results show the 95% confidence interval, bold numbers lie within the confidence interval.

Acronyms and symbols

Acronyms

DTQA	Discrete-time queueing analysis
BMAP	Batch Markovian Arrival Process
MMPP	Markov Modulated Poisson Process
MAP	Markovian Arrival Process
OLS	Ordinary Least Square
SMAD	Semi-Markov Arrival Decomposition Approach
SMP	Semi-Markov Process

Discrete random variables

A	Inter-arrival time
A_k	Inter-arrival time of customer k
A^U/A^D	Inter-arrival time at the upstream / downstream system
B	Service time
B_k	Service time of customer k
B^U/B^D	Service time at the upstream / downstream system

C	Number of customers arriving at a slot boundary with Poisson probability distribution
D	Inter-departure time at the upstream system
D_k	Inter-departure time of customer k at the upstream system
N_t	Discrete random variable of the family $\mathcal{N}$ at time instance t
N_k	Discrete random variable of the family $\mathcal{N}$ at observation point k
N^U/N^D	Number of customers in the upstream / downstream queue
W	Waiting time at the downstream queue
W_k	Waiting time of customer k at the downstream queue
R	Residual time until the next event occurs in a discrete-time point process
R_L / R_H	Lower and upper bound of the residual time until the next event occurs in SMAD
T_k	Position of the k-th point in a discrete-time point process
U	Time since the last event occurred in a discrete-time point process
X_k	Interval sequence the associated stochastic process $\mathcal{N}$ spends in state N_k

Sets

Ξ	Discrete state space of a discrete-time stochastic process
Γ	Set of all observation points of a discrete-time stochastic process
Ω	Reduced state space of the discrete-time Markov chain describing the number of customers in the discrete-time $M/G/1$-queue at departure instances

Families of random variables

$\mathcal{N}$	Stochastic process, family of discrete random variables
$\mathcal{N}^U/\mathcal{N}^D$	Discrete-time Markov chain describing the number of customers in the upstream / downstream queue
$\mathcal{Z}$	SMP, family of tuples of discrete random variables
$\mathcal{Z}^U$	SMP describing the number of customers and departure times in the upstream queue

Matrices

$\mathbf{F}$	Interval time distribution of a discrete-time SMP
$\mathbf{P}$	Probability transition matrix of a discrete-time Markov chain
$\mathbf{P}_r$	Reduced probability transition matrix of a discrete-time Markov chain in recursion step r
$\mathbf{P}'$	Approximated probability transition matrix of a discrete-time Markov chain
$\mathbf{Q}$	Semi-Markov kernel of a SMP

Queuing parameters

λ	Flow rate of the external Poisson process
ρ	Utilisation of the tandem queue
ρ^U/ρ^D	Utilisation of the upstream / downstream system

Stationary probabilities

π	Vector of the stationary distribution of a discrete-time Markov chain
$\pi(i)$	Stationary probability of state i in a discrete-time Markov chain with one-dimensional state space
$\pi(Z)$	Stationary probability of state Z in a discrete-time Markov chain with higher-dimensional state space

Constants and iterators

τ	Quantile of the quantile regression method
ϕ	Lag of auto-correlation in the upstream system's departure stream
κ	State space limit in SMAD
θ	Normalisation constant in SMAD for the computation of the waiting time distribution

Measures of decomposition error

$\Delta(E)$	Decomposition error with respect to the expected value of waiting time
$\Delta(\sigma)$	Decomposition error with respect to the 95th-percentile of waiting time

List of Figures

2.1 Point process on the discrete-time axis 10
2.2 Sample path of a discrete-time semi-Markov process 12

4.1 Parameters in the tandem queue 36
4.2 Empirical cumulative distribution functions of the decomposition error of waiting time 44
4.3 Utilisation is the main driver for decomposition error 56

5.1 Discrete-time modelling of arrival and service processes in the $M/G/1$-queue 61
5.2 Discrete-time modelling of the joint inter-departure time distribution . 69
5.3 Auto-correlation for inter-departure times in $M/G/1$-queues with different service time variability and utilisation parameters . . . 72
5.4 Auto-correlation for inter-departure times in $M/G/1$-queues in dependence of utilisation 72

6.1 State space limitation method in the upstream $M/G/1$-queue 90
6.2 Distributions of waiting time (truncated) for the $SM/G/1$-queue and the corresponding $GI/G/1$-queue 105

List of Tables

4.1 Performance metrics of the tandem queue 37
4.2 Summary statistics of the IVs and flow parameters 43
4.3 OLS and quantile regression estimates in Study I 46
4.4 OLS and quantile regression estimates in Study II 48
4.5 Performance of point estimates: Relative frequency distributions and means of forecasting error for training and test data in Study I . . 51
4.6 Performance of point estimates: Relative frequency distributions and means of forecasting error for training and test data in Study II . . 52
4.7 Performance of interval estimates 53
4.8 Absolute mean decomposition errors in longer lines 55

6.1 Distributions of waiting time (truncated) in the downstream queue, obtained with DTQA, SMAD, and simulation 104
6.2 Expected values of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with DTQA, SMAD(κ), and simulation 106
6.3 95% percentiles of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with DTQA, SMAD(κ), and simulation 108
6.4 Computer times for DTQA and SMAD 109

A.1 OLS and quantile regression estimates in Study I with downstream bottlenecks 120
A.2 OLS and quantile regression estimates in Study II with downstream bottlenecks 122

C.1 Expected values of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with DTQA, SMAD, and simulation (cont.) . . 132

C.2 95% percentiles of waiting time in the downstream queue for several combinations of utilisation and service time variability parameters, obtained with DTQA, SMAD, and simulation (cont.) . . 133

List of Publications

Jacobi, C. (2023a). Discrete-time semi-Markov models for the analysis of dependent departure and arrival streams. *KITopen Repository*.

Jacobi, C. (2023b). On the output dynamics of the discrete-time M/G/1-queue. *Under review at Annals of Operations Research*.

Jacobi, C. and K. Furmans (2022a). Data sets for the analysis of decomposition error in discrete-time open tandem queues. *KITopen Repository*.

Jacobi, C. and K. Furmans (2022b). Point and interval estimation of decomposition error in discrete-time open tandem queues. *Operations Research Letters 50*(5), p. 529–535.

Jacobi, C. and K. Furmans (2022c). Regression analyses of the data sets for the analysis of decomposition error in discrete-time open tandem queues. *Data in Brief 45*.

Jacobi, C. and J. G. Shanthikumar (2023). A refined decomposition approach with converging accuracy for discrete-time open tandem queues with Poisson arrivals and general service times. *Unpublished working paper*.

Bibliography

Ackroyd, M. (1980). Computing the waiting time distribution for the G/G/1 queue by signal processing methods. *IEEE Transactions on Communications 28*(1), p. 52–58.

Albin, S. (1984a). Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Operations Research 32*(5), p. 1133–1162.

Albin, S. (1984b). *Approximating Queues with Superposition Arrival Processes*. Dissertation, Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, United States.

Arjas, E. (1972). On a fundamental identity in the theory of semi-Markov processes. *Advances in Applied Probability 4*(2), p. 258–270.

Bijma, F., M. Jonker and A. W. van der Vaart (2017). *Introduction to mathematical statistics*. Amsterdam: Amsterdam University Press.

Bitran, G. and S. Dasu (1992). A review of open queueing network models of manufacturing systems. *Queueing Systems 12*, p. 95–133.

Bitran, G. and D. Tirupati (1988). Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Science 34*(1), p. 75–100.

Bitran, G. and D. Tirupati (1989). Capacity planning in manufacturing networks with discrete options. *Annals of Operations Research 17*(1), p. 119–135.

Bruneel, H. and B. G. Kim (1993). *Discrete time models for communication systems including ATM*. Kluwer international series in engineering and computer science. Boston: Kluwer.

Burke, P. (1956). The output of a queueing system. *Operations Research 4*(6), p. 699–704.

Buzacott, J. A. (1967). Automatic Transfer Lines with Buffer Stocks. *International Journal of Production Research 5*(3), p. 183–200.

Buzacott, J. A. and J. G. Shanthikumar (1992). Design of manufacturing systems using queueing models. *Queueing Systems 12*(1), p. 135–213.

Buzacott, J. A. and J. G. Shanthikumar (1993). *Stochastic Models of Manufacturing Systems*. Prentice Hall International series in industrial and systems engineering. Upper Saddle River, NJ: Prentice Hall.

Cox, D. R. (1955). The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Mathematical Proceedings of the Cambridge Philosophical Society 51*(3), p. 433–441.

Cox, D. R. and H. D. Miller (1965). *The theory of stochastic processes*. London: Methuen & Co. Ltd.

Daley, D. J. (1968). The correlation structure of the output process of some single server queueing systems. *The Annals of Mathematical Statistics 39*(3), p. 1007–1019.

Daley, D. J. (1976). Queueing output processes. *Advances in Applied Probability 8*(2), p. 395–415.

Daley, D. J. and D. Vere-Jones (2008). *An Introduction to the Theory of Point Processes: Volume II General Theory and Structure* (2. ed.). Probability and Its Applications, A Series of the Applied Probability. New York: Springer.

Dallery, Y. and X.-R. Cao (1992). Operational analysis of stochastic closed queueing networks. *Performance Evaluation 14*(1), p. 43–61.

Dallery, Y., R. David and X.-L. Xie (1988). An efficient algorithm for analysis of transfer lines with unreliable machines and finite buffers. *IIE Transactions 20*(3), p. 280–283.

Dallery, Y. and S. Gershwin (1992). Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems 12*, p. 3–94.

de Smit, J. (1986). The single server semi-Markov queue. *Stochastic Processes and their Applications 22*(1), p. 37–50.

Dupuy, D., C. Helbert and J. Franco (2015). Dice-Design and Dice-Eval: Two R packages for design and analysis of computer experiments. *Journal of Statistical Software 65*(11).

Epp, M. (2018). *Performance Evaluation of Shuttle-based Storage and Retrieval Systems Using Discrete-time Queueing Network models*. Dissertation, Karlsruhe Institute of Technology, Institute for Material Handling and Logistics, Karlsruhe, Germany.

Epp, M., S. Wiedemann and K. Furmans (2017). A discrete-time queueing network approach to performance evaluation of autonomous vehicle storage and retrieval systems. *International Journal of Production Research 55*(4), p. 960–978.

Ferng, H.-W. and J.-F. Chang (2001a). Connection-wise end-to-end performance analysis of queueing networks with MMPP inputs. *Performance Evaluation 43*(1), p. 39–62.

Ferng, H.-W. and J.-F. Chang (2001b). Departure processes of BMAP/G/1 queues. *Queueing Systems 39*, p. 109–135.

Furmans, K. (2004). A framework of stochastic finite elements for models of material handling systems. In: *8th International Material Handling Research Colloquium*, Graz, Austria.

Furmans, K. and A. Zillus (1996). Modeling independent production buffers in discrete time queueing networks. In: *Proceedings of Rensselaer's Fifth*

International Conference Computer Integrated Manufacturing and Automation Technology, Grenoble, France, p. 275–280.

Gass, S. I. (2013). *Encyclopedia of Operations Research and Management Science*. Boston, MA: Springer.

Gaver, D. P. and G. S. Shedler (1973a). Approximate models for processor utilization in multiprogrammed computer systems. *SIAM Journal on Computing 2*(3), p. 183–192.

Gaver, D. P. and G. S. Shedler (1973b). Processor utilization in multiprogramming systems via diffusion approximations. *Operations Research 21*(2), p. 569–576.

Gebennini, E. and A. Grassi (2015). Discrete-time model for two-machine one-buffer transfer lines with buffer bypass and two capacity levels. *IIE Transactions 47*(7), p. 715–727.

George, D. K. and C. H. Xia (2011). Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research 211*(1), p. 198–207.

Gershwin, S. B. (1987). An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Operations Research 35*(2), p. 291–305.

Gershwin, S. B. (1991). Assembly / disassembly systems: An efficient decomposition algorithm for tree-structured networks. *IIE Transactions 23*(4), p. 302–314.

Gershwin, S. B. and M. H. Burman (2000). A decomposition method for analyzing inhomogeneous assembly / disassembly systems. *Annals of Operations Research 93*, p. 91–115.

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

Gómez-Corral, A. (2002). A tandem queue with blocking and Markovian arrival process. *Queueing Systems 41*, p. 343–370.

Govil, M. K. and M. C. Fu (1999). Queueing theory in manufacturing: A survey. *Journal of Manufacturing Systems 18*(3), p. 214–240.

Govind, N., T. M. Roeder and L. W. Schruben (2010). A simulation-based closed queueing network approximation of semiconductor automated material handling systems. *IEEE Transactions on Semiconductor Manufacturing 24*(1), p. 5–13.

Grassmann, W. K. and J. L. Jain (1989). Numerical solutions of the waiting time distribution and idle time distribution of the arithmetic GI/G/1 queue. *Operations Research 37*(1), p. 141–150.

Grassmann, W. K. and J. Tavakoli (2019). The distribution of the line length in a discrete time GI/G/1 queue. *Performance Evaluation 131*, p. 43–53.

Harchol-Balter, M. (2013). *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge: Cambridge University Press.

Haßlinger, G. (1995). A polynomial factorization approach to the discrete time GI/G/1/(N) queue size distribution. *Performance Evaluation 23*(3), p. 217–240.

Haßlinger, G. (2000). Waiting time, busy periods and output models of a server analyzed via Wiener–Hopf factorization. *Performance Evaluation 40*(1), p. 3–26.

Haßlinger, G. and E. S. Rieger (1996). Analysis of open discrete time queueing networks: A refined decomposition approach. *Journal of the Operational Research Society 47*, p. 640–653.

Heindl, A. (2001). Decomposition of general tandem queueing networks with MMPP input. *Performance Evaluation 44*(1), p. 5–23.

Heindl, A. (2003). Decomposition of general queueing networks with MMPP inputs and customer losses. *Performance Evaluation 51*(2-4), p. 117–136.

Heindl, A. and M. Telek (2002). Output models of MAP/PH/1(/K) queues for an efficient network decomposition. *Performance Evaluation 49*(1), p. 321–339.

Hendricks, K. B. (1992). The output processes of serial production lines of exponential machines with finite buffers. *Operations Research 40*(6), p. 1139–1147.

Hendricks, K. B. and J. O. McClain (1993). The output process of serial production lines of general machines with finite buffers. *Management Science 39*(10), p. 1194–1201.

Hoare, C. A. R. (1962). Quicksort. *The Computer Journal 5*(1), p. 10–16.

Hübner, F. and P. Tran-Gia (1994). Discrete-time analysis of cell spacing in ATM systems. *Telecommunication Systems 3*(3), p. 379–395.

Iglehart, D. L. and W. Whitt (1970). Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. *Advances in Applied Probability 2*(2), p. 355–369.

Jackson, J. R. (1957). Networks of waiting lines. *Operations Research 5*(4), p. 518–521.

Jackson, J. R. (1963). Jobshop-like queueing systems. *Management Science 10*(1), p. 131–142.

Jain, J. L. and W. K. Grassmann (1988). Numerical solution for the departure process from the GI/G/1 queue. *Computers & Operations Research 15*(3), p. 293 – 296.

Jenkins, J. H. (1966a). On the correlation structure of the departure process of the M/E/1 queue. *Journal of the Royal Statistical Society: Series B (Methodological) 28*(2), p. 336–344.

Jenkins, J. H. (1966b). Stationary joint distributions arising in the analysis of the M/G/1 queue by the method of the imbedded Markov chain. *Journal of Applied Probability 3*(2), p. 512–520.

Jones, D., M. Schonlau and W. Welch (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization 13*, p. 455–492.

Kim, C., V. I. Klimenok and D. S. Orlovsky (2008). The BMAP/PH/N retrial queue with Markovian flow of breakdowns. *European Journal of Operational Research 189*(3), p. 1057–1072.

Kim, S. (2004). The heavy-traffic bottleneck phenomenon under splitting and superposition. *European Journal of Operational Research 157*(3), p. 736 – 745.

Kim, S., R. Muralidharan and C. A. O'Cinneide (2005). Taking account of correlations between streams in queueing network approximations. *Queueing Systems 49*(3), p. 261 – 281.

King, R. A. (1971). The covariance structure of the departure process from M/G/1 queues with finite waiting lines. *Journal of the Royal Statistical Society: Series B (Methodological) 33*(3), p. 401–405.

Kleinrock, L. (1975). *Queueing systems, Volume 1: Theory*. A Wiley-Interscience publication. New York: Wiley.

Knuth, D. E. (1998). *The art of computer programming* (2. ed.), Volume 3: Sorting and searching. Reading, MA: Addison-Wesley.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), p. 33–50.

Koenker, R. and V. d'Orey (1987). Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 36*(3), p. 383 – 393.

Koenker, R. and V. d'Orey (1994). Remark AS R92: A remark on algorithm AS 229: Computing dual regression quantiles and regression rank scores. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 43*(2), p. 410–414.

Koenker, R. and K. F. Hallock (2001). Quantile Regressions. *The Journal of Economic Perspectives 4*(15), p. 143–156.

Koenker, R. and J. A. Machado (1999). Goodness of fit and related inference for quantile regression. *Journal of the American Statistical Association 94*(448), p. 1296–1310.

Kraemer, W. and M. Langenbach-Belz (1976). Approximate formulae for the delay in the queuing system GI/G/1. In: *8th International Teletraffic Congress*.

Kuehn, P. (1979). Approximate analysis of general queuing networks by decomposition. *IEEE Transactions on Communications 27*(1), p. 113–126.

Lagershausen, S. (2013). *Performance Analysis of Closed Queueing Networks*. Lecture Notes in Economics and Mathematical Systems. Berlin, Heidelberg: Springer.

Lee, H. W., N. I. Park and J. Jeon (2003). A new approach to the queue length and waiting time of BMAP/G/1 queues. *Computers & Operations Research 30*(13), p. 2021–2045.

Lee, Y., A. van de Liefvoort and V. Wallace (2000). Modeling correlated traffic with a generalized IPP. *Performance Evaluation 40*(1), p. 99–114.

Lian, Z. and L. Liu (2008). A tandem network with MAP inputs. *Operations Research Letters 36*(2), p. 189–195.

Lieckens, K. and N. Vandaele (2012). Multi-level reverse logistics network design under uncertainty. *International Journal of Production Research 50*(1), p. 23–40.

Limnios, N. and V. S. Barbu (2008). *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their use in Reliability and DNA Analysis*. Lecture Notes in Statistics. New York, NY: Springer.

Loeppky, J. L., J. Sacks and W. J. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics 51*(4), p. 366–376.

Lucantoni, D. M. and M. F. Neuts (1994). Some steady-state distributions for the MAP/SM/1 queue. *Stochastic Models 10*(3), p. 575–598.

Marshall, K. (1968). Some inequalities in queuing. *Operations Research 16*(3), p. 651–668.

Mitchell, K., A. van de Liefvoort and J. Place (1998). Correlation properties of the token leaky bucket departure process. *Computer Communications 21*(11), p. 1010–1019.

Neuts, M. F. (1979a). Queues solvable without Rouché's theorem. *Operations Research 27*(4), p. 676–781.

Neuts, M. F. (1979b). A versatile Markovian point process. *Journal of Applied Probability 16*(4), p. 764–779.

Newell, G. F. (1982). *Applications of Queueing Theory*. Monographs on Statistics and Applied Probability. London: Chapman and Hall.

Onvural, R. O. (1990). Survey of closed queueing networks with blocking. *ACM Computing Surveys (CSUR) 22*(2), p. 83–121.

Pack, C. D. (1975). The output of an M/D/1 queue. *Operations Research 23*(4), p. 750–760.

Papadopoulos, H. and C. Heavey (1996). Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. *European Journal of Operational Research 92*(1), p. 1 – 27.

Ramaswami, V. (1980). The N/G/1 queue and its detailed analysis. *Advances in Applied Probability 12*(1), p. 222–261.

Ramaswami, V. (1988). A stable recursion for the steady state vector in Markov chains of M/G/1 type. *Communications in Statistics. Stochastic Models 4*(1), p. 183–188.

Reiman, M. I. (1990). Asymptotically exact decomposition approximations for open queueing networks. *Operations Research Letters 9*(6), p. 363 – 370.

Reiser, M. and H. Kobayashi (1974). Accuracy of the diffusion approximation for some queuing systems. *IBM Journal of Research and Development 18*(2), p. 110–124.

Reiser, M. and S. S. Lavenberg (1980). Mean-value analysis of closed multichain queuing networks. *Journal of the ACM (JACM) 27*(2), p. 313–322.

Reynolds, J. F. (1975). The covariance structure of queues and related processes – A survey of recent work. *Advances in Applied Probability 7*(2), p. 383–415.

Rieger, E. S. and G. Haßlinger (1994). An analytical solution for the discrete time single server system with semi-markovian arrivals. *Queueing Systems 18*, p. 69–105.

Sagron, R., D. Grosbard, G. Rabinowitz and I. Tirkel (2015). Approximation of single-class queueing networks with downtime-induced traffic variability. *International Journal of Production Research 53*(13), p. 3871–3887.

Sagron, R., G. Rabinowitz and I. Tirkel (2017). Approximating class-departure variability in tandem queues with downtime events: Regression-based variability function. *Computers & Operations Research 88*, p. 161–174.

Schleyer, M. (2007). *Discrete Time Analysis of Batch Processes in Material Flow Systems*. Dissertation, Karlsruhe Institute of Technology, Institute for Material Handling and Logistics, Karlsruhe.

Schleyer, M. (2010). An analytical method for the calculation of the number of units at the arrival instant in a discrete time G/G/1-queueing system with batch arrivals. *OR Spectrum 34*(1), p. 293–310.

Schleyer, M. and K. Furmans (2007). An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals. *OR Spectrum 29*(4), p. 745–763.

Schleyer, M. and K. Gue (2012). Throughput time distribution analysis for a one-block warehouse. *Transportation Research Part E: Logistics and Transportation Review 48*(3), p. 652 – 666.

Schwarz, J. A. and M. Epp (2016). Performance evaluation of a transportation-type bulk queue with generally distributed inter-arrival times. *International Journal of Production Research 54*(20), p. 6251–6264.

Sen, A. K., M. S. Srivastava and M. S. Srivastava (1990). *Regression Analysis: Theory, Methods, and Applications*. New York, NY: Springer.

Sevcik, K. C. and I. Mitrani (1981). The distribution of queuing network states at input and output instants. *Journal of the ACM (JACM) 28*(2), p. 358–371.

Shaked, M. and J. G. Shanthikumar (2007). *Stochastic Orders*. New York: Springer.

Shanthikumar, J. G. (2022). Lecture notes in the doctoral classes of the Operations Management Department of Krannert Graduate School of Management. West-Lafayette, IN: Purdue University.

Shanthikumar, J. G. and J. A. Buzacott (1981). Open queueing network models of dynamic job shops. *International Journal of Production Research 19*(3), p. 255–266.

Shanthikumar, J. G., S. Ding and M. T. Zhang (2007). Queueing theory for semiconductor manufacturing systems: A survey and open problems. *IEEE Transactions on Automation Science and Engineering 4*(4), p. 513–522.

Sherzer, E., A. Senderovich, O. Baron and D. Krass (2022). Can machines solve general queueing systems? *arXiv preprint arXiv:2202.01729*.

Shioda, S. (2003). Departure process of the MAP/SM/1 queue. *Queueing Systems 44*, p. 31–50.

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

Suresh, S. and W. Whitt (1990). The heavy-traffic bottleneck phenomenon in open queueing networks. *Operations Research Letters 9*(6), p. 355 – 362.

Tan, B. and S. Lagershausen (2017). On the output dynamics of production systems subject to blocking. *IISE Transactions 49*(3), p. 268–284.

Tran-Gia, P. (1996). *Analytische Leistungsbewertung verteilter Systeme: Eine Einführung*. Berlin: Springer.

Van Nieuwenhuyse, I. and R. B. de Koster (2009). Evaluating order throughput time in 2-block warehouses with time window batching. *International Journal of Production Economics 121*(2), p. 654 – 664.

Vishnevskii, V. M. and A. N. Dudin (2017). Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. *Automation and Remote Control 78*, p. 1361–1403.

Whitt, W. (1981). Approximating a point process by a renewal process: The view through a queue, an indirect approach. *Management Science 27*(6), p. 619–636.

Whitt, W. (1982). Approximating a point process by a renewal process, I: Two Basic Methods. *Operations Research 30*(1), p. 125–147.

Whitt, W. (1983a). Performance of the Queueing Network Analyzer. *The Bell System Technical Journal 62*(9), p. 2817–2843.

Whitt, W. (1983b). The Queueing Network Analyzer. *The Bell System Technical Journal 62*(9), p. 2779–2815.

Whitt, W. (1984). Approximations for departure processes and queues in series. *Naval Research Logistics Quarterly 31*(4), p. 499–521.

Whitt, W. (1994). Towards better multi-class parametric-decomposition approximations for open queueing networks. *Annals of Operations Research 48*(3), p. 221–248.

Whitt, W. (1995). Variability functions for parametric-decomposition approximations of queueing networks. *Management Science 41*(10), p. 1704–1715.

Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice Hall.

Worthington, D. (2009). Reflections on queue modelling from the last 50 years. *Journal of the Operational Research Society 60*, p. 83–92.

Wu, G., X. Xu, Y. Y. Gong, R. B. de Koster and B. Zou (2019). Optimal design and planning for compact automated parking systems. *European Journal of Operational Research 273*(3), p. 948 – 967.

Wu, K. and L. McGinnis (2013). Interpolation approximations for queues in series. *IIE Transactions 45*(3), p. 273–290.

Yu, M. and R. B. de Koster (2009). The impact of order batching and picking area zoning on order picking system performance. *European Journal of Operational Research 198*(2), p. 480 – 490.