Effective process times for aggregate modeling of manufacturing systems

Effective Process Times for Aggregate Modeling of Manufacturing Systems

A.A.A. Kock

A catalogue record is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-1306-2

Reproduction: Wöhrmann Print Service
Cover design: Sam Gatignon
Source cover: courtesy to www.activewin.com (picture of Pentium 4 0.13μ wafer)

This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Dutch Ministry of Economic Affairs.

Effective Process Times for Aggregate Modeling of Manufacturing Systems

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 30 juni 2008 om 16.00 uur

door

Adrianus Arnoldus Antoinetta Kock
geboren te Geleen

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. J.E. Rooda
en
prof.dr.ir. O.J. Boxma

Copromotor:
dr.ir. L.F.P. Etman

Preface

Professor Rooda often says 'research is like a marine oil spill, it keeps extending'. And now, as I am writing the last pages of my Ph.D. thesis, I understand exactly what he is saying. There are still some unresolved issues in the EPT that I would love to resolve. But I have been researching the EPT for several years now, and I feel it is time to move on, especially since the end of my PhD contract is rapidly approaching.

For their contribution to my thesis, I would like to thank:

The first promotor, Koos Rooda, for creating such a dynamic and interesting working-environment. Thank you Koos for the good advice you have given, and for the interest you often have shown.

The copromotor, Pascal Etman, for everything he has done for me. Pascal, thank you for all our discussions, your advices and your support when I needed it.

The second promotor, Onno Boxma, for the detailed comments he gave on the thesis and for the pleasant cooperation between the Stochastic Operations Research group and the Systems Engineering group.

Ivo Adan for the many discussions we had, for his assistance with in particular Chapter 6, and for his detailed comments on the mathematics in my thesis.

Marcel van Vuuren for four years of pleasant cooperation, for coauthoring two of the chapters and for his aide with several mathematical problems I encountered.

Erjen Lefeber for our many discussions, both in and out of the STW-project team, and the many relevant, detailed and difficult questions he asked.

Adam Wierman for the pleasant discussions and his important contribution to Chapter 6.

Marco Vijfvinkel, whose MSc thesis was the basis for Chapter 4.

Casper Veeger for his effort in the industrial case of Chapter 5, and Bart Lemmen for coauthoring Chapter 5.

Frank Nijsse and Freek Wullems for coauthoring Chapter 2.

Albert Hofkamp for keeping the χ-software running, and for his rapid fixes of the compiler.

Eric Blom, René Bouman and Maciej Lazurko for doing their MSc project and Roel Oomens his internship, all in the context of my research project.

Hervé Buclon, Joris van der Eerden, Joost van Herk, Johan Jacobs, Ton de Kok, Sven Weber, and Kees de Wit for their contribution to one of the chapters.

The STW-user committee, consisting of Martin Prins (ASML), Jan van Dorremalen and Simone Resing (CQM), Frans Brouwers and Edgar van Campen (NXP) and Frank Nijsse (Steelweld BV). The STW-guidance was provided by Margriet Jansz, Marijke de Jong and Corine Meuleman in consecutive order.

Professor van Steenhoven, professor Armbruster, professor Koole, and professor Mummolo for their contribution to the assessment of the thesis.

Mieke Lousberg for the interest she often showed, and for her help in filling out the many, many forms.

Finishing a PhD project is also about motivating yourself, and keeping going when times are tough. That is not possible without people who both love and encourage you. I want to thank

My parents Wiel and Agnes, my brother Rob, my sister Leontine and my girlfriend, Charlotte, for giving me the love and support I needed, and for being there for me.

Casper, Joost, Maarten, Michiel, Ricky, Roel, and Simon for the pleasant lunches and evenings we enjoyed (and for the entertainment we need to unwind).

Ad Kock

Sittard, May 2008

Summary

Effective Process Times for Aggregate Modeling of Manufacturing Systems

Modern manufacturing systems are becoming more complex. Analyzing the flow time and throughput performance may be quite involved. Often it is hard to predict the impact certain changes may have on the system behavior. Queueing models are helpful here.

Two classes of queueing models can be distinguished: analytical models and simulation models. Analytical models are fast to evaluate and need little input, yet they are not straightforward to develop and adhere to strict assumptions. Simulation models are more flexible and can be used to model any detail. However they are computationally expensive, and require a large amount of input data regarding the shop floor details.

This thesis proposes a method for model aggregation to reduce the number of details that has to be covered by either the analytical model or the simulation model. Through aggregation, a workstation is represented by a single effective process time distribution, which includes all the losses due to the outages such as setup, machine downs, or operator availability. Key to the methods presented in the thesis is that the aggregate process time distribution is measurable directly from shop floor data such as lot arrivals and lot departures at the workstation, without quantifying the contributing factors. This arrival and departure data may be obtained from the programmable logic controllers (PLCs) used in the control system of many manufacturing systems.

For the aggregation, we start from the concept of Effective Process Times (EPT). The EPT was introduced by Hopp and Spearman (1996, 2001) as the process time seen by a lot at a workstation from a logistical point of view. Jacobs, Etman, Van Campen, and Rooda (2001, 2003) showed that effective process times can be measured without quantifying the individual time losses. In this way, they were able to measure the process time coefficient of variation at several single-lot machine workstations in a semiconductor fab. This second moment of the process time distribution is needed in (analytical) queueing models of

manufacturing systems. Van Vuuren (2007) presents analytical queuing models that use the first two moments of the EPT workstation distributions as input for finitely buffered workstations (single- or multi-server) and assembly stations.

This thesis further develops the 'Effective Process Time' modeling framework for the performance analysis of manufacturing systems. It presents methods to measure EPT-realizations for finitely buffered workstations and assembly-stations. Sample path equations are used to compute the EPT-realizations from three events: lot arrival times, lot departure times, and process end times. The EPT-realizations are combined to form EPT-distributions from the mean, variance and possibly higher moments. Alternatively, distribution functions may be fitted to the measured EPT. The proposed EPT-method is tested in two industrial cases, one from the automotive industry and one from light bulb production. The EPT models provide accurate throughput and flow time approximations.

The thesis shows that the EPT concept may also be used to aggregate only part of the workstation. A model of a lithography track-scanner combination is presented in which the litho-cell itself is modeled in detail, but the influence of the environment is aggregated into a single delay distribution. Typically, for the inside of the litho cell, a lot of process data is available, whereas of the environment (the loading) less data is available. The developed models were applied on a simulation example, and an industrial case, using data obtained from the Crolles-2 wafer fab. The simulation test case showed that the model is accurate, and may be used to predict the effect of changes in the machine configuration. The industry case showed that an accurate flow time approximation could be obtained (with an error of 8% in the flow time approximation). The case revealed that a significant part of the flow time is due to the environment. Furthermore, the model was used to calculate a flow time-throughput curve.

Finally, the thesis presents an aggregation model for workstations with integrated processing machines. Equipment with integrated processing is commonly encountered in semiconductor manufacturing. They simultaneously process a flow of wafers of multiple lots. The proposed aggregate model is a simple $G/G/m$ queueing system but with the process times depending on the momentary number of customers in the system. Simulation experiments were conducted on four test cases (a sequential single server flow line, a short flow line with parallel servers, a case with four parallel single-server lines and a workstation with parallel servers). The third scenario (with four parallel lines) strongly resembles a workstation of litho cells. The results show that the proposed model gives accurate flow time approximations. The proposed model is far more accurate than the standard $G/G/m$ approximation that is typically used.

The research described in this thesis was carried out as part of the STW project EPT. The project is a collaboration of the Systems Engineering Group at the department of Mechanical Engineering and the Stochastic Operations Research Group at the department of Mathematics and Computer Science, both of the Eindhoven University of Technology.

Contents

Chapter 1

Introduction

Performance analysis of manufacturing systems is becoming increasingly important. The last decades, globalization has increased competition on the world wide market in nearly all industries. Customers demand better products, lower prices and shorter delivery times. Furthermore, the costs of materials and machines are increasing. For the production of goods at competitive prices, continuous improvement of the performance of manufacturing systems is required.

A manufacturing system can be defined as a collection of resources that converts raw material into a product. Well-known examples are car manufacturing and semiconductor wafer fabrication, which are among the largest and most cost-intensive manufacturing systems around the globe. The analysis and control of such large manufacturing systems is not straightforward. Therefore, from a logistical and managerial point of view, manufacturing systems are often analyzed at different levels. Rooda and Vervoort (2007) distinguish four levels, see Figure 1.1:

- At the network level, the manufacturing system is the factory (also referred to as plant, fabricator, or shortly fab). The elements of the system are areas and (groups of) machines. This level is also known as the factory level.
- At the sub-network level, the manufacturing system is an area of the factory with several machines or groups of machines (workstations). The elements of the system are individual machines. The sub-network level is also known as the area level.
- At the workstation level, the manufacturing system is a group of machines,

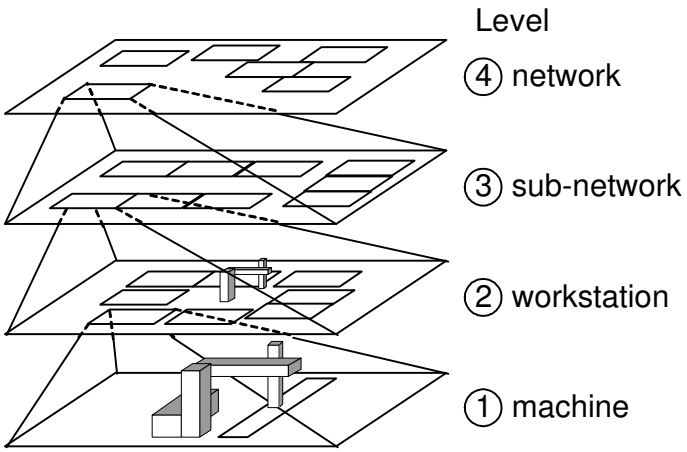


Figure 1.1: Abstraction levels in manufacturing systems (Rooda and Vervoort 2007)

that are typically scheduled as one entity.

- At the machine level, the manufacturing system is the individual machine (also referred to as equipment or tool). The elements of the system are components in the machine.

1.1 Performance analysis

Several tools and performance indicators are in use for the performance analysis of manufacturing systems. Two parameters that are often used are throughput δ (the number of lots processed per time unit) and mean flow time φ (the average time a lot spends in the system). Throughput δ as well as mean flow time φ are descriptive performance indicators, that is they quantify the performance of the system. They do not explain why the performance is the way it is, nor do they assist in finding solutions to improve the performance. For that purpose, other indicators are used.

A well-known indicator aiding performance improvement is the overall equipment effectiveness (OEE) (Nakajima 1988). The SEMI-E10 and SEMI-E79 norms (SEMI 2000, 2001) commonly used in the semiconductor industry are for instance based on the OEE. Recently a revision of the OEE, E, has been proposed by De Ron and Rooda (2005). The OEE quantifies mean time losses during processing. Losses are divided into availability losses, performance losses and quality losses. The OEE readily gives insight in the cause of undesired behavior at workstations. The OEE quantifies the production capacity losses, which relates to the *utilization* of the installed capacity. Note that the OEE does not quantify the *variability* in processing which also affects the manufacturing performance.

Workstation utilization and variability are the two basic parameters explaining the performance of a manufacturing system regarding throughput δ and mean

flow time φ. For a manufacturing system consisting of infinitely buffered workstations Equation (1.1), an approximate expression due to Sakasegawa (1977) and Whitt (1993), is insightful to explain the contribution of utilization and variability to the flow time performance (Hopp and Spearman 2001):

$$\varphi = \frac{c_a^2 + c_e^2}{2} \cdot \frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \cdot t_e + t_e. \tag{1.1}$$

Herein, c_a is the coefficient of variation in the inter-arrival times, c_e the coefficient of variation in the process time, m the number of parallel machines, or servers[†], in the workstation and u the utilization, i.e. the ratio between the mean process time t_e and the mean inter-arrival time t_a multiplied by m:

$$u = \frac{t_e}{m \cdot t_a}. \tag{1.2}$$

Note that t_e is the mean effective process time which includes all capacity losses due to the various outages such as machine breakdowns and setup time. Similarly, c_e is the coefficient of variation that results from the combination of the processing and the various outages. The t_e relates to the OEE (more specifically the E); for c_e no equivalent indicator is in use.

Once the performance of a system is analyzed, one may want to improve that performance. The performance metrics described above do not provide the possibility to *predict* the impact of changes in the system on system performance. Predicting the changes in system performance may be difficult due to the large number of processes and the interaction between processes in the manufacturing network. To understand the impact of changes in the system configuration, queueing models are used.

1.2 Models

For the performance prediction of manufacturing systems, typically discrete event simulation models (e.g. Kleijnen and Van Groenendaal (1992), Banks (1999), Law and Kelton (2000), Baines, Mason, and Siebers (2003), Fowler and Rose (2004)) or analytical queueing models (e.g. Dallery and Gershwin (1992), Buzacott and Shanthikumar (1993), Gershwin (1994), MacGregor Smith (2005), Shanthikumar, Ding, and Zhang (2007), Van Vuuren (2007)) are used.

In a simulation model, the relevant shop-floor realities may be included separately. As a result, the model does not necessarily need to conform to prespecified assumptions. However, since a distribution is typically required for each phenomenon that is modeled, large quantities of data are required to gather

[†]In this thesis, the words 'machine' and 'server' are used interchangeably.

Table 1.1: Properties of model-types (A: Analytical, S: Simulation)

Property	A	S
Assumptions	-	+
Amount of input data	+	-
Computational cost	+	-
Flexibility in application	-	+

the input for the simulation model. To an existing simulation model, new details can be added, thus simulation models are highly flexible. On the other hand, since each individual lot is tracked through the model, simulation models require a lot of computational effort. The simulation model is stochastic, so one needs to run multiple replications to obtain reliable results.

In an analytical model, often a Markov chain is used to represent the system behavior (Adan 2001). Markov chains with a limited number of states are computationally cheap to evaluate. The input of such a model typically consists of only mean process times and variances, hence little data is required. The model provides steady state output, hence no replications are required. However, to have computationally feasible Markov chains, the model has to adhere to restrictive assumptions (such as phase-type distributed process times). Furthermore, if the configuration of the system is changed, an entirely new Markov chain is required; adapting the model is not straightforward. In Table 1.1, the properties of both analytical models and simulation models are summarized.

Both model types have their own specific advantages and disadvantages. Analytical models are computationally fast, but it is difficult to include many shop-floor realities in the model. As a result, analytical queueing network models are little used in manufacturing industry. The gap between model assumptions and shop floor reality is often considered too large (Fowler and Rose 2004, Shanthikumar et al. 2007). If one would be able to aggregate the shop-floor realities and the processing into a single distribution for each workstation, and then be able to actually measure this aggregate distribution from simple shop-floor events such as lot arrivals and departures, then this may provide an opportunity to bridge this gap. Also for simulation models aggregation of shop-floor realities into a single workstation would be advantageous: a simulation model would require less input data, while the model becomes computationally cheaper since only one distribution per workstation is induced. The STW project "Effective process time" aims to provide such an aggregation method.

1.3 STW project on effective process time

The concept of effective process time (EPT) was first introduced by Hopp and Spearman (2001). They define the EPT as the time spent by a lot at a workstation from a logistical point of view. Thus, all time during which a lot claims

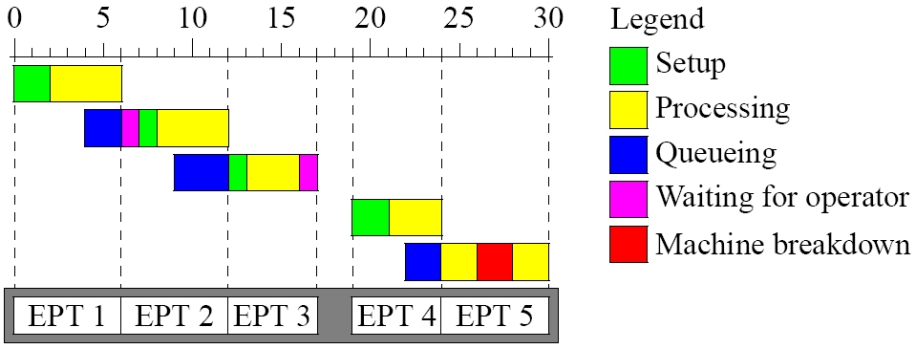


Figure 1.2: Concept of effective process time (picture from Coenen (2004))

machine capacity is included in the effective process time, as is illustrated in Figure 1.2. Hopp and Spearman show how the EPT of a workstation can be computed, given distribution parameters regarding the clean process time and preemptive and non-preemptive outages, as expressed in for instance the mean busy time between failures t_f, the mean time to repair t_r and setup Δ_u. Other outages are treated as either preemptive or non-preemptive outages. The notion of combining all individual influences on processing into a single distribution is also used in the context of sample path analysis (Chen and Chen 1990, Dallery and Gershwin 1992, Buzacott and Shanthikumar 1993, Rossetti and Clark 2003). However, in many practical cases, the outages may not all be quantifiable (Pierce 1994, McMullen and Frazier 1998, Hsieh 2002, Mendes, Ramos, Simaria, and Vilarinho 2005).

Jacobs et al. (2001, 2003) presented an algorithm to obtain effective process time distributions for infinitely buffered workstations from simple lot arrivals and departures. Their method does not require the quantification of the individual contributing factors. The motivation of their work was to arrive at a measurable metric for variability at a workstation (variance in processing), that can furthermore be used to build abstract but accurate aggregate models. They used closed form queueing equations, such as Equation (1.1) as well as simulation to predict the flow time. They feeded their EPT-based models with the first two moments of the effective process time distribution. Jacobs, Van Bakel, Etman, and Rooda (2006) extended their method to batch machines. Also several M.Sc. students contributed to these initial efforts: Van Bakel (2001), Rooney (2002), Wullems (2002) and Kock (2003). Wullems (2002) and Kock (2003) for instance started to work on the EPT for finitely instead of infinitely buffered workstations. Finitely buffered manufacturing lines are, among others, encountered in automotive manufacturing.

Following up on this initial work, the Systems Engineering group and the Stochastic Operations Research group, both of the Eindhoven University of Technology, initiated an STW project on the effective process time in 2004. The goal of the project was to develop an aggregate modeling methodology that enables one to build simple yet accurate models of manufacturing networks using operational

data such as arrival and departure events without the need to characterize all contributing disturbances and shop-floor realities. In the project two parts can be distinguished:

1. Development of the effective process time paradigm for aggregate modeling and parameter identification (carried out by the Systems Engineering group of the department of Mechanical Engineering). The results obtained are described in the present thesis.

2. Development of efficient queueing network approximations that fit into the EPT-based aggregate modeling framework (performed by the Stochastic Operations Research group of the department of Mathematics and Computer Science). Former STW-researcher Van Vuuren (2007) developed several new queueing network approximations for finitely buffered single- and multi-server flow lines, for assembly stations and for workstations with multiple arrival streams. The methods he developed are based on phase-type distributions decomposition, aggregation of states, matrix analytical methods and iterative numerical procedures. The distribution parameters are only the first two moments, for which the EPT mean and variance will be used.

1.4 EPT framework

A schematic overview of the EPT framework is presented in Figure 1.3. The box at the top represents the real-life manufacturing system from which shop-floor data is obtained. The box at the bottom represents the EPT-based aggregate model, either a simulation model or an analytical model. The oval boxes in between represent the EPT-algorithm and the distribution fitting procedure. The figure emphasizes that the STW-project aims at the development of aggregate models for which the parameters can be estimated from operational data at the factory floor. The consecutive steps in the EPT framework are explained further in detail.

First, based on the manufacturing system under investigation, one defines the structure of the EPT-based aggregate model. To keep the model intuitive and computationally cheap, the EPT-based model is kept as simple as possible, that is, shop-floor realities affecting processing behavior are aggregated in the EPT as much as possible.

For each workstation defined in the EPT-based model, event data, such as lot arrivals and departures, are gathered from the manufacturing system. This event data is used to compute the EPT-realizations per workstation. These EPT-realizations may then be translated into model-input, by selecting appropriate distributions and fitting the distribution parameters. The fitted distributions and parameters are then used in the EPT-based aggregate model. With the model,

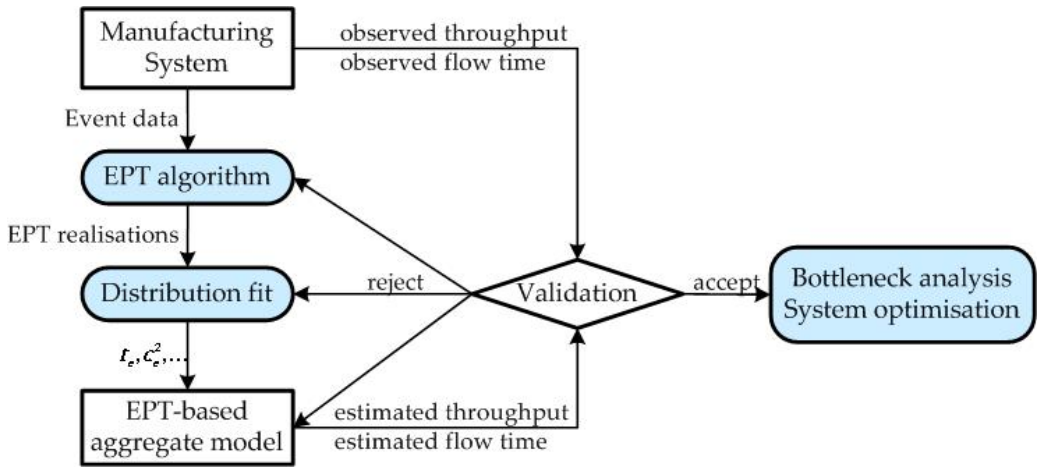


Figure 1.3: Schematic overview of the EPT framework

predictions for throughput, flow time behavior or other desired properties of the manufacturing system can be made.

The accuracy of the EPT-based aggregate model is evaluated by comparing the performance indicators estimated by the aggregate model to the performance indicators observed in the real manufacturing system. If the EPT-based model approximates the manufacturing system accurately enough, i.e. within a pre-specified error margin defined by the analyst, the EPT-based model is accepted. It can then be used for e.g. bottleneck analysis, or predicting the impact of changes in the system configuration or utilization. If the model is found not accurate enough, part of the aggregation process may be reconsidered. Possible solutions include: enhancing the level of modeling detail, acquiring more or more reliable data or refining the EPT-realizations.

1.5 Contribution and outline of the thesis

In this thesis, the effective process time framework is further developed. For finitely buffered flow lines, in Chapters 2 and 3 of the thesis, EPT-algorithms are presented that compute the first two moments of the process time distributions, required as input for the models developed by Van Vuuren (2007). For single server flow lines, it is shown that effective process times can be determined from three types of manufacturing events: lot arrivals, lot departures, and process finish times. For the multi-server case, it is shown that the single-server procedure can be used again by sorting the events per server on which the lots are processed, and by applying the single-server procedure for each server individually. The developed EPT-method is applied in two industrial cases, one from the automotive industry and one from light bulb production. The examples show that the EPT-based approximation models accurately approximate the flow time behavior of the system with approximation errors within a few percent. The cases

illustrate how the accuracy of the EPT-based model may be enhanced by explicitly modeling two product types, and by including the offset and the skewness as a third and fourth distribution parameter.

In Chapter 4, an EPT quantification method for assembly workstations in finitely buffered lines is proposed. The effective process time realization only starts running if all components of an assembly have arrived. Transport times are now explicitly modeled in the aggregate model. The new EPT-method for assembly stations is compared to treating the assembly station as an 'ordinary' finitely buffered workstation with the feeding component lines aggregated in the workstation process time distribution. We apply both alternatives in a case inspired by the automotive industry. The EPT-based aggregate simulation models are found to be accurate. Additionally, an EPT-based simulation model is compared to an EPT-based analytical queueing model such as developed by Van Vuuren (2007), showing that the models have comparable accuracies.

In semiconductor manufacturing, lithography is one of the main operations in the process flow. In the lithography area, litho cells are used. A litho cell consists of a track and scanner. The track is used for pre- and postprocessing of wafers, while the scanner is used to expose patterns onto the wafer. To this end, several process steps are carried out on the wafers in the track and the scanner. The litho cell can be viewed as a finitely buffered flow line. For the litho cell (track and scanner) this thesis presents a more detailed simulation model in Chapter 5. The model describes the processing behavior and outages of the track and scanner part of the litho cell in detail, while an EPT-like aggregation is used to describe the impact of the shop-floor on the performance of the litho cell. The proposed simulation model is tested on an industrial case. The model estimates the flow time of the considered litho cell with an error in the flow time approximation of 8% and in the throughput approximation of 2.6%.

Chapter 6 considers workstations consisting of integrated process type of machines. Recent developments in semiconductor wafer fabrication have shown a proliferation in the use of manufacturing tools with integrated process steps. An example of such an integrated process tool is the aforementioned track-scanner litho cell. Chapter 6 proposes a new aggregate model that is able to represent a multi-process step integrated manufacturing system: a $G/G/m$ approximation with process times depending on the level of work in progress (WIP, or number of customers in the system) is proposed. An accompanying EPT-algorithm to determine the EPT-realizations for the WIP-dependent $G/G/m$ model directly from operational factory data is presented. Four test scenarios show that the proposed aggregate model gives accurate flow time approximations at a utilization region around the training point (the utilization level at which the EPT-realizations were measured).

The current status of the STW research on effective process time is summarized in Table 1.2. In Table 1.2, N refers to the network level of a manufacturing system, while W refers to the workstation level and M to the machine level.

Table 1.2: Overview of the STW-project effective process time, categorised by level

Lvl	Topic	Reference
N	Queueing: finitely buffered line	Ch. 3, 4 of Van Vuuren (2007)
	Queueing: assembly line	Ch. 5 of Van Vuuren (2007)
	Queueing: multiple arrival streams	Ch. 6 of Van Vuuren (2007)
W	EPT: infinitely buffered workstations	Ch. 3 of Jacobs (2004)‡
	EPT: batch processing workstation	Ch. 4 of Jacobs (2004)‡
	EPT: finitely buffered workstation	Ch. 2 and 3 of this thesis
	EPT: assembly workstation	Ch. 4 of this thesis
	EPT: integrated manufacturing systems	Ch. 6 of this thesis
M	EPT: detailed litho cell model	Ch. 5 of this thesis

1.6 Guidelines for the reader

Chapters 2 to 6 are the research chapters of this thesis. Each research chapter is either accepted or submitted as a journal paper: Chapter 2 appeared as Kock, Wullems, Etman, Adan, Nijsse, and Rooda (2008c) and Chapter 3 appeared as Kock, Etman, and Rooda (2008a). Chapters 4 (Vijfvinkel, Kock, Etman, Van Vuuren, and Rooda 2007), 5 (Kock, Veeger, Etman, Lemmen and Rooda 2008d) and 6 (Kock, Etman, Rooda, Adan, Van Vuuren, and Wierman 2008b) are submitted as journal papers.

Note that each of these chapters is self-contained; after this introductory chapter, the reader may proceed with any of the chapters. As a consequence, the first two sections of each of the research chapter are alike to some extent. For each chapter, we have printed the abstract on the first page of the chapter.

‡The research by Jacobs was carried out as part of STW-project "ADOPT"

Chapter 2

Finitely buffered, single server flow lines

The present chapter extends the so-called effective process time (EPT) approach to single server flow lines with finite buffers and blocking. The power of the EPT-approach is that it quantifies variability in workstation process times without the need to identify each of the contributing disturbances, and that it directly provides an algorithm for the actual computation of EPTs. It is shown that EPT-realizations can be simply obtained from arrival and departure times of lots, by using sample path equations. The measured EPTs can be used for bottleneck analysis and for lumped parameter modeling. Simulation experiments show that for lumped parameter modeling of flow lines with finite buffers, in addition to the mean and variance, offset is also a relevant parameter of the process time distribution. A case from the automotive industry illustrates the approach.

This chapter originally appeared as:
Kock, Wullems, Etman, Adan, Nijsse, and Rooda. Performance Evaluation and Lumped Parameter Modeling of Single Server Flowlines subject to Blocking: an Effective Process Time Approach, Computers and Industrial Engineering 54 (4): 866-878. 2008
The original publication is available at DOI 10.1016/j.cie.2007.10.016: http://www.science-direct.com/

2.1 Introduction

Single server workstations with finite buffer sizes in a tandem flow line are an important class of manufacturing systems. Examples of such flow lines are semi-synchronous lines and assembly lines, as e.g. encountered in the automotive industry.

The performance of a flow line is commonly expressed in terms of throughput and flow time. Both performance indicators are influenced by blocking. The finite capacity of the buffers in the single server flow lines considered in this chapter introduces blocking in the line.

Blocking causes suspension of service to a lot (which implies loss of production capacity) since a finished lot cannot be sent on due to a saturated downstream buffer. Starvation refers to the situation where processing of the next lot is suspended due to an empty upstream buffer.

Variability in process times is the main reason that blocking and starvation occur. The variability of process times can be traced to several common sources. First, natural process times are variable due to differences in product types, machine states at product entry, operator behavior etcetera. Furthermore, disturbances such as setups, preventive maintenance, machine failures and absence of operators occur. These disturbances cause loss of production capacity effectively available at the workstation and increase the variability of process times, which in turn decreases the throughput. Subsequent workstations affect one another more prominently as the variability of process times increases. Variability of process times on workstation j can cause starvation on workstation $j+1$. Furthermore, in a flow line with finite buffers, variability of process times on workstation j can cause blocking on workstation $j-1$.

Obviously, for performance analysis of a finitely buffered flow line, an analysis tool that quantifies both the production losses and the level of variability of process times is required. A commonly applied performance analysis metric is the overall equipment efficiency, OEE. However, OEE can only be used for quantifying production losses. Therefore an alternative analysis tool will be used in this chapter.

Hopp and Spearman (2001) introduced this alternative concept to account for irregularities in process times of workstations. The alternative concept, effective process time (EPT), is defined as the total time seen by a lot at a workstation from a logistical point of view. Here, total time indicates the total time that the lot has effectively consumed production capacity of the workstation. EPT is based on the notion that, from a logistical perspective, a workstation does not care whether production capacity is claimed since the server is processing the lot or whether production capacity is claimed by other influences. These other influences are included in the EPT of the workstation.

Hopp and Spearman's notion of including processing disturbances in the effective process times is not new, see e.g. the work of Chen and Chen (1990), Dallery and Gershwin (1992), Buzacott and Shanthikumar (1993). The aforementioned authors all assume, or measure, distributions for the various processing disturbances and combine these into one single distribution. However, from industrial practice, it is often hard, if not impossible, to identify and quantify all individual disturbances, see e.g. the work of Pierce (1994), McMullen and Frazier (1998), Hsieh (2002), Mendes et al. (2005).

Starting from the concept of EPT, Jacobs et al. (2001) and Jacobs et al. (2003) presented a method to translate lot arrivals and lot departures at an infinitely buffered workstation into an EPT-distribution. The workstation process times and the disturbances from the factory floor are aggregated into a single distribution without the need to quantify the individual factors. In automated manufacturing environments, arrival and departure data are often available.

The obtained EPT-distributions can be used for performance analysis and optimization. Based on the characteristic parameters of the EPT-distributions, i.e. the mean effective process time t_e and the coefficient of variation c_e, a bottleneck analysis can be performed, after which an approximating model can be used to predict the changes in system performance. Two types of models may be distinguished: analytical queueing models and (discrete event) simulation models. Analytical queueing models are fast to evaluate, usually based on assumptions such as Markovian process times and Markovian times between failure and times to repair, see e.g. Chen and Chen (1990), Dallery and Gershwin (1992), Buzacott and Shanthikumar (1993), Gershwin (1994), Jeong and Kim (1999), Hopp, Spearman, Chayet, Donohue, and Gel (2002), Li, Alden, and Rabaey (2005), Diamantidis, Papadopoulos, and Heavey (2007), Van Vuuren (2007). Analytical models typically require the first two moments of the process times, for which t_e and c_e can be used. Alternatively, simulation models may be used (Banks 1999, Law and Kelton 2000). The EPT-distributions may be directly used as input to the simulation model, either by fitting an appropriate distribution function or by using the EPT-distribution as an empirical distribution.

This chapter aims to generalize the EPT-approach for application to single server flow lines subject to blocking. That is, the chapter considers finite buffers rather than infinite ones. Workstations can no longer be analyzed in isolation due to the dependencies introduced by blocking. Therefore, an EPT-algorithm for the blocking case is presented. Furthermore, the effect of the distribution shape on the accuracy of the EPT lumped parameter (ELP) model is investigated. Two theoretical examples and a case from automotive industry are used to illustrate the EPT-approach. Note that throughout the chapter, mainly the effects of blocking are discussed since starvation also occurs in infinitely buffered workstations.

The chapter is organized as follows. In Section 2.2, an outline of the EPT-approach is presented. Subsequently, computation of EPT-realizations for single server workstations with finite buffers is considered in Section 2.3. EPT-based

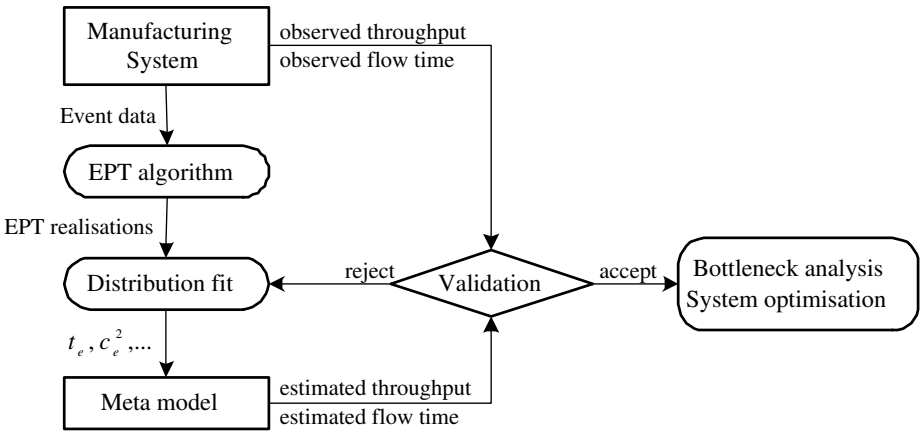


Figure 2.1: SCHEMATIC OVERVIEW OF THE EPT FRAMEWORK

lumped parameter modeling in the context of finitely buffered flow lines is discussed in Section 2.4. The concepts discussed throughout this chapter are illustrated using the aforementioned examples and case in Section 2.5 and Section 2.6. Finally, Section 2.7 concludes the chapter.

2.2 A framework for implementing EPT

The EPT-approach, based on the concept of Jacobs et al. (2003), consists of four stages, as visualized in Figure 2.1.

First, EPT-realizations are obtained from the discrete manufacturing system. An EPT-realization is defined by Jacobs et al. as: 'the time a lot was in process plus the time a lot (not necessarily the same lot) could have been in process'. EPT-realizations can be computed from event data, such as arrivals and departures of lots on workstations. The EPT-realizations are computed by means of an EPT-algorithm. The EPT or similar concepts (such as completion time) are used in sample path analyses of queuing systems. Sample path equations are typically used to determine lot departures from lot arrivals and the effective process time. The EPT-concept presented in this chapter uses the sample path equations differently, that is, effective process times are determined from arrival and departure data. The sample path equations are thus a means to obtain EPT-realizations from an operating production system. The operation time as defined by Rossetti and Clark (2003) is very similar to EPT; however, Rossetti and Clark do not use it to quantify the level of variability.

Next, the EPT-realizations are fitted to distributions. Here, distributions are fitted based on relevant workstation properties, such as the mean EPT t_e and the coefficient of variation c_e. Parameter t_e quantifies the mean effective capacity used for a lot by the workstation, c_e quantifies the effective variability.

Subsequently, a so-called EPT lumped parameter (ELP) model can be built using the fitted distributions, either an analytical queueing model or a (discrete event) simulation model. This ELP model can be used for performance prediction and optimization. The structure of the ELP model follows the original system in terms of the number of servers on each workstation, the buffer sizes of workstations, the flow of materials between workstations, etcetera. In this model, detailed modeling of shop-floor realities such as failures, repairs, setups, operators and lot sizes is avoided. The various sources of variability are aggregated into the EPT-distributions of the workstations. Jacobs et al. (2003) used the term 'meta model' rather than 'lumped parameter model'. However, the phrase 'meta model' may suggest that a simplified model is derived from another model. Since this is certainly not the case, the terminology 'lumped parameter model' is used in this chapter. Here, the lumped parameters refer to the distribution parameters of the EPT-distributions.

Before the ELP model is accepted, it is validated by comparing the throughput and flow time as estimated by the model to those observed in the actual system, since one is interested in how well the lumped parameter model describes the behavior of the actual system. If the estimated throughput and flow time are accurate enough, the ELP model and the EPT-distributions are accepted. If they are rejected, distribution fitting and model building are reconsidered. Possible changes include enhancing the level of detail of the model or using more parameters to fit more accurate distributions.

If the EPT-distributions and the ELP model are accepted, they can be used for performance analysis and optimization. A bottleneck analysis can be carried out based on the distribution parameters t_e and c_e of the various workstations. The effect of suggested improvements can be evaluated using the ELP model by accordingly adjusting the EPT-distribution parameters in the model.

Implementation of the EPT-approach provides several significant advantages. First, many shop-floor realities are included in the EPT-distributions and thus do not have to be included explicitly in the ELP model. Now, an ELP model can be obtained that is accurate, yet simple when compared to the detailed (simulation) models that are typically used. Second, since the processing disturbances are included in the EPT-distributions, directly obtained from industrial data, the EPT-parameters t_e and c_e readily give insight in the behavior of the flow line, allowing for straightforward bottleneck analysis.

2.3 Measuring EPT

The EPT was introduced by Hopp and Spearman (1996, 2001) to be used in analytical queuing models. Similar concepts, such as completion time, are used in sample path equations. In the literature describing such concepts, referred to in Section 2.1, the respective distributions are assumed to be known *a priori*.

However, it is not specified how these distributions should be estimated from industrial data.

Jacobs et al. (2003) presented a method to compute EPT-realizations for infinitely buffered, isolated workstations from industrial data. Their method does not assume the effective process time distributions *a priori*, but, in a way similar to using a sample path equation, determines these distributions. For a single machine workstation, the sample path equation is:

$$e_{i,j} = d_{i,j} - \max\left(a_{i,j}, d_{i-1,j}\right), \tag{2.1}$$

where $e_{i,j}$ denotes the EPT-realization of lot i on workstation j, $d_{i,j}$ is the departure of lot i from workstation j and $a_{i,j}$ is the arrival of lot i on workstation j. Here, we assume that lots do not overtake. From Equation (2.1), one sees that an EPT-realization encompasses all time during which the server could have been processing the lot. For the events holds that $a_{i,j} \leqslant d_{i,j}$. In case of timeless transport, $d_{i,j-1} = a_{i,j}$.

Algorithmic extensions have been presented for workstations with multiple parallel servers (Jacobs et al. 2003) and with batching (Jacobs et al. 2006). However, the algorithms are only applicable to workstations with an infinitely large buffer. This chapter studies finite buffers, which gives rise to blocking. Due to blocking, $e_{i,j}$ depends on events occurring on workstation $j+1$, rendering the previous algorithms inapplicable.

Considering finitely buffered workstations, the sample path equation for the departure of lots is given by (see page 184 of Buzacott and Shanthikumar (1993), or Adan and Van der Wal (1989)):

$$D_i^j = \max\left[\max\left\{D_i^{j-1}, D_{i-1}^j\right\} + S_i^j, D_{i-b_{j+1}}^{j+1}\right] \tag{2.2}$$

Herein, D_i^j is the i$^{\text{th}}$ departure from workstation j; the term $\max(D_i^{j-1}, D_{i-1}^j)$ represents the i^{th}time at which processing of the lot can start; S_i^j represents the completion time, b_{j+1} is the total capacity that can be held at workstation $j+1$, and $D_{i-b_{j+1}}^{j+1}$ is the time of the $i - b_{j+1}^{\text{th}}$ departure from workstation $j+1$, so that workstation $j+1$ has sufficient capacity to receive the i$^{\text{th}}$ lot. Substituting our notation into Equation (2.2), and assuming that lots do not overtake, we obtain

$$d_{i,j} = \max\left(\max\left\{d_{i-1,j}, d_{i,j-1}\right\} + e_{i,j}, d_{i-b_{j+1},j+1}\right), \tag{2.3}$$

or with $d_{i,j-1} = a_{i,j}$,

$$d_{i,j} = \max\left(\max\left\{d_{i-1,j}, a_{i,j}\right\} + e_{i,j}, d_{i-b_{j+1},j+1}\right). \tag{2.4}$$

Similar to Equation (2.1), processing starts if the lot has arrived and no other lot occupies the server (at $\max\left(d_{i-1,j}, a_{i,j}\right)$). Processing finishes $e_{i,j}$ time units

later. If processing of the lot is done, the lot could leave workstation j, provided that the receiving workstation has sufficient capacity. We call this the possible departure of lot i: $pd_{i,j}$. This gives

$$pd_{i,j} = \max\left(d_{i-1,j}, a_{i,j}\right) + e_{i,j}$$

$$d_{i,j} = \max\left(pd_{i,j}, d_{i-b_{j+1},j+1}\right)$$

The effective process time of lot i on workstation j is thus computed from:

$$e_{i,j} = pd_{i,j} - \max\left(d_{i-1,j}, a_{i,j}\right). \tag{2.5}$$

As can be seen, one should replace $d_{i,j}$ in Equation (2.1) by $pd_{i,j}$. Possible occurrences of blocking should not be included in the EPT-realization. They are a physical part of the finitely buffered flow line and will also appear in the EPT-based lumped parameter (ELP) model. Note that Equation (2.5) can also be used to compute the EPT for finitely buffered, single server workstations with overtaking. In that case, the i^{th} EPT-realization on workstation j, $e_{i,j}$, is computed from the arrival $a_{i,j}$ and the possible departure $pd_{i,j}$ of the i^{th} processed lot, and the actual departure $d_{i-1,j}$ of the previously processed lot.

2.4 Lumped parameter modeling

Distribution fitting is the second phase of the EPT-approach. The relevant distribution parameters are estimated based on the measured EPT-realizations and appropriate distribution functions are proposed.

Process time distributions based on the first two moments of the distribution are often used in models of manufacturing systems consisting of workstations with infinitely large buffers. The two-moment fits are supported by queuing theory, see e.g. Buzacott and Shanthikumar (1993), Sabuncuoglu, Erel, and de Kok (2002) and Curry, Peters, and Lee (2003).

For workstations in a flow line with finite buffer sizes, distribution fitting could be more complicated. Due to blocking, workstations are expected to affect one another more prominently. Therefore, extra information may be needed. Regardless, in queuing theoretical approaches, two-moment distribution fits are used for computational reasons. However, in case of simulation, the use of additional information, such as higher moments or the offset, may be reconsidered. A typical example thereof is presented by Kim and Alden (1997). They study constant natural process times with exponentially distributed times to failure and times to repair. In the EPT-approach, the sources of disturbances are lumped. In this respect, no assumptions regarding the distribution of the process times or disturbances are made. The necessity of additional distributional information in ELP models of finitely buffered flow lines will be studied here.

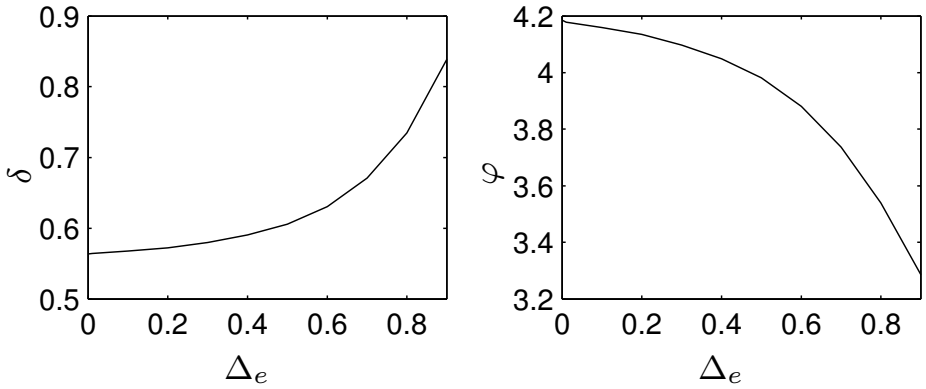


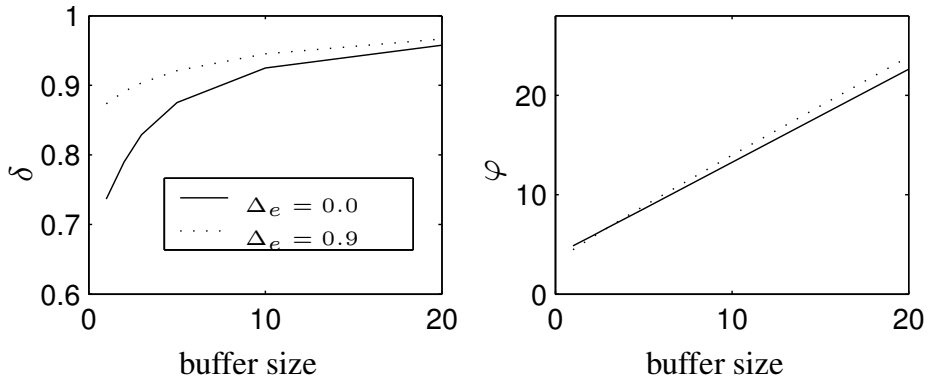
Figure 2.2: INFLUENCE OF THE OFFSET PARAMETER ON δ AND φ

Using simulation, the influence of the offset parameter is investigated. The offset parameter is chosen since, in practice, many operations require at least a minimum amount of time. The offset refers to the smallest possible value of a distribution. The simulation model is a flow line consisting of three unbuffered single server workstations in which lots do not overtake. The three workstations have process times distributed according to a shifted Gamma distribution. The distributional parameters are $t_e = 1.0$, $c_e = 1.0$ and offset Δ_e, which is varied from 0.0 to 0.9.

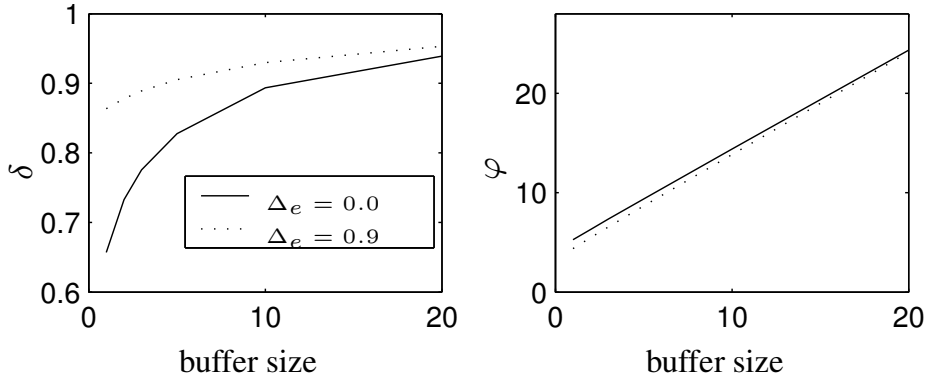
The corresponding simulation results are presented in Figure 2.2. The results show that for large offsets, significant differences in throughput (δ) and flow time (φ) are observed. Increasing Δ_e from 0.0 to 0.9 results in a throughput increase of 50% and a flow time decrease of 21% (see Figure 2.2).

The observed phenomenon can readily be explained by considering the nature of the offset. An offsetted distribution consists of a constant part, Δ_e, that is increased by a random variable with mean t_l and coefficient of variation c_l, where $t_l = t_e - \Delta_e$. Since the variance of the process time distribution does not change, it holds that $t_e^2 c_e^2 = t_l^2 c_l^2$. Now, if $t_l = 0.1 t_e$, $c_l^2 = 100 c_e^2$. Due to the large c_l, most process times will be small ($\gtrsim \Delta_e$), and sporadically a value greatly exceeding the average ($\gg t_e$) will occur. The sporadic large process time realization therefore causes massive amounts of blocking on preceding workstations and starvation on successive workstations. If $\Delta_e = 0.0$ however, all process times will be centered on t_e. Process times will thus often be larger than t_e, frequently causing some blocking and starvation on preceding or successive workstations.

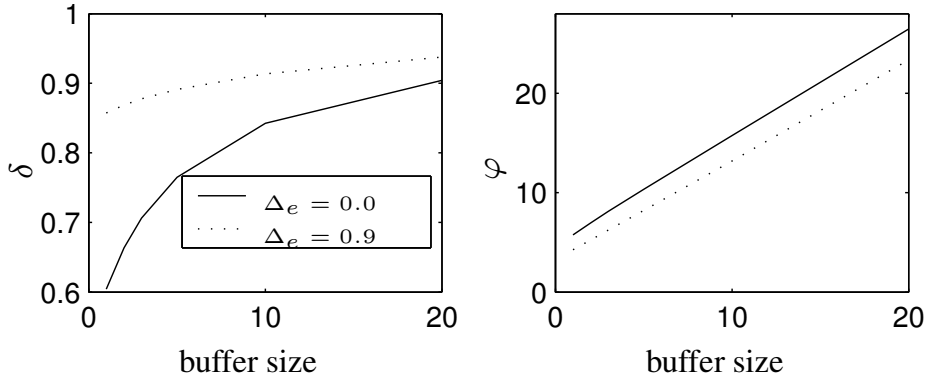
A new set of simulations is used to test the relevance of Δ_e. As stated above, one can expect the shape of the distribution to have more influence if the amount of blocking and starvation increases. This expectation is investigated using simulation. For a flow line consisting of three finitely buffered workstations with a single server, the buffer space between WS_0 and WS_1 and between WS_1 and WS_2 will be changed. In addition, the level of variability is changed. Process times on the workstations will have identical $t_e = 1.0$. However, $c_{e,i}$ (where i refers to the workstation number) is chosen at 1.0 at the first workstation, but is varied from



(a) $c^2_{e,1}$ = 1.0, $c^2_{e,2}$ = 0.5 and $c^2_{e,3}$= 0.5



(b) $c^2_{e,1}$ = 1.0, $c^2_{e,2}$ = 1.0 and $c^2_{e,3}$= 1.0



(c) $c^2_{e,1}$ = 1.0, $c^2_{e,2}$ = 2.0 and $c^2_{e,3}$= 2.0

Figure 2.3: INFLUENCE OF BUFFER SPACES ON δ AND φ

0.5 to 2.0 at the other two workstations. The throughput and flow time will be evaluated at offset levels of $\Delta_e = 0.0$ and $\Delta_e = 0.9$. The corresponding simulation results are depicted in Figure 2.3.

Several observations can be made from Figure 2.3. The first observation is that the mean throughput for $\Delta_e = 0.9$ approaches the throughput for $\Delta_e = 0.0$ as the buffer space increases.A second important observation, from comparing Figure 2.3(a) to Figure 2.3(c), is that the difference in mean throughput between $\Delta_e = 0.9$ and $\Delta_e = 0.0$ becomes larger as the squared coefficients of variation are increased.

A change in the squared coefficient of variation has more effect on performance for $\Delta_e = 0.0$ than for $\Delta_e = 0.9$. This observation can be explained by the fact that a flow line with $\Delta_e = 0.0$ is more likely to be blocked than a flow line with $\Delta_e = 0.9$. Since an increase in variability implies an increase in the amount of blocking, the flow line with $\Delta_e = 0.0$ is more heavily affected.

These results, and additional simulation results presented in Kock (2003), imply that, as the amount of blocking and starvation in the flow line increases (by reducing buffer space or by increasing the level of variability), the influence of higher order information of the distribution shape increases.

2.5 Examples

Two examples are presented to validate the computation of EPT-realizations and to illustrate the EPT-approach.

2.5.1 Example I

Consider a flow line consisting of five workstations labeled WS_i for $i = 0, \ldots, 4$. Each workstation has a buffer of size one and one server. The first workstation is never starved whereas the final workstation is never blocked. All workstations have exponentially distributed natural process times with mean $t_{0,i} = 1.0$ for all i. The servers are subject to operation dependent failures, with busy time between failures exponentially distributed with mean $t_{f,i} = 7.5$ for all i. Once a failure has occurred, the server is repaired. Repair times are exponentially distributed with mean $t_{r,i} = 2$ for all i. After the repair is finished, processing of the lot is continued for a period of time equal to the remaining process time. The flow line is represented using a detailed discrete event simulation model, explicitly modeling the failure and repair behavior. This model will be referred to as the 'original' model.

The first stage of the EPT-approach is carried out by applying Equation (2.5) to the arrival and departure events generated by the original model. This leads to a large set of EPT-realizations for each of the workstations. During the second stage of the approach, the realizations are translated into shifted Gamma distributions with mean $t_{e,i}$, squared coefficient of variation $c^2_{e,i}$, and offset $\Delta_{e,i}$ as presented in Table 2.1. The t_e and c^2_e values of the table are verified using Equations (2.6) and (2.8) as presented by Hopp and Spearman (2001). Herein, t_0 is the mean natural process time, c_0 is the corresponding coefficient of variation, c_r is the coefficient of variation of the times to repair, and A is the availability. Availability A of a workstation represents the fraction of time during which the server is able to perform. It is computed using (2.7). Equations (2.6) and (2.8) give values $t_{e,i} = 1.13$ and $c^2_{e,i} = 1.42$ for all i, which corresponds to the measured equivalents

Table 2.1: MEASURED EPT-PARAMETERS FOR EXAMPLE I OF CHAPTER 2

WS_i	$\Delta_{e,i}$	$t_{e,i}$	$c^2_{e,i}$
WS_0	0.00	1.13	1.41
WS_1	0.00	1.13	1.42
WS_2	0.00	1.13	1.42
WS_3	0.00	1.13	1.42
WS_4	0.00	1.13	1.42

in Table 2.1.

$$t_{\mathrm{e},i} = \frac{t_{0,i}}{A_i}, \tag{2.6}$$

$$A_i = \frac{t_{\mathrm{f},i}}{t_{\mathrm{f},i} + t_{\mathrm{r},i}}, \tag{2.7}$$

$$c^2_{\mathrm{e},i} = c^2_{0,i} + \left(1 + c^2_{\mathrm{r},i}\right) A_i \left(1 - A_i\right) \frac{t_{\mathrm{r},i}}{t_{0,i}}. \tag{2.8}$$

Since the natural process times are exponentially distributed, as are the failures and repairs, the effective process time distributions of the workstations do not have an offset, i.e. $\Delta_{\mathrm{e}} = 0.0$. As can be seen in Table 2.1, the estimated value of Δ_{e} is indeed 0.0.

The original simulation model has $\delta = 0.495 \pm 0.01\%$ and $\varphi = 14.15 \pm 0.01\%$. This implies that, with a probability of 95%, the range $(0.49495, 0.49505)$ contains the true value of δ and the range $(14.1486, 14.1514)$ contains the true value of φ.

During the third stage of the approach, the approximated distributions are used as input for a discrete event EPT-based lumped parameter (ELP) model. The structure of the ELP model follows the structure of the original system, i.e. five workstations consisting of one buffer space and one server. Servers have process times distributed according to the shifted Gamma distribution, with parameters according to Table 2.1. The ELP model approximates $\widetilde{\delta} = 0.491$ and $\widetilde{\varphi} = 14.26$, which means that the difference between the EPT approximation and the original situation is 0.81% in throughput and 0.77% in flow time. The error in the approximation is computed by:

$$\left| \frac{\delta - \widetilde{\delta}}{\delta} \right| \cdot 100\% \quad \text{and} \quad \left| \frac{\varphi - \widetilde{\varphi}}{\varphi} \right| \cdot 100\% \tag{2.9}$$

Note that Equation (2.9) is used in the remainder of this chapter to compute the error in approximations.

If both the original system and the ELP model do not contain buffer spaces, the original model gives performance measures $\delta = 0.399$ and $\varphi = 9.23$, whereas the approximation is $\widetilde{\delta} = 0.393$ and $\widetilde{\varphi} = 9.34$, giving an error of 1.5% for throughput and 1.2% for flow time. Increasing the number of buffer spaces on all workstations to 5 leads to $\delta = 0.656$ and $\varphi = 29.72$ for the original model compared to $\widetilde{\delta} =$

0.657 and $\widetilde{\varphi}$ = 29.76 for the EPT-based lumped parameter model. This is an error of 0.2% in throughput and 0.1% in flow time. Obviously, the error decreases as the number of buffer spaces in the line increases, which corresponds to the observations of Section 2.4.

2.5.2 Example II

Consider a flow line consisting of five workstations WS_i for $i = 0,\ldots,4$. Workstation WS_i has b_i buffer spaces and one single server, where $[b_0,b_1,b_2,b_3,b_4] = [0,2,1,2,1]$. The flow line produces two product types, pt_0 and pt_1 in the deterministic sequence $[pt_0,pt_1,pt_0,pt_1\ldots]$. The first workstation is never starved whereas the final workstation is never blocked. At WS_0, all products are processed with exponentially distributed natural process times with mean 1. At WS_1 and WS_3, natural process times for products of type pt_0 are distributed according to a shifted Gamma distribution with $\Delta_{0,0} = 0.6$, $t_{0,0} = 1.5$ and $c^2_{0,0} = 0.75$, whereas $\Delta_{0,1} = 0.2$, $t_{0,1} = 0.5$ and $c^2_{0,1} = 0.75$ on these stations for products of type pt_1. On workstations WS_2 and WS_4, products of type pt_0 are processed with natural process times according to a triangular distribution with $\Delta_{0,0} = 0.4$, $t_{0,0} = 0.5$ and maximum 0.6 and thus, $c^2_{0,0} = 6.67\times10^{-3}$; for pt_1 however $\Delta_{0,1} = 1.2$, $t_{0,1} = 1.5$ and maximum 1.8 giving $c^2_{0,1} = c^2_{0,0}$. On WS_i for $i = 1,2,3,4$, a constant setup time of 0.1 time units is required if the product type is changed. The servers are subject to operation dependent failures, with busy time between failures exponentially distributed with mean $t_{\mathrm{f},i} = 7.5$ for all i. Once a failure has occurred, the server is repaired. Repair times are exponentially distributed with mean $t_{\mathrm{r},i} = 2$ for all i. After the repair is finished, processing of the lot is resumed at the point where it was interrupted. Simulation results for the example have been obtained for 95% confidence levels with a relative width of 1% or less of the corresponding mean.

First, EPT-realizations are computed for each of the workstations by applying Equation (2.5) to the arrival events (a) and departure events (pd,d) obtained from the simulation model. Next, the realizations are translated into shifted Gamma distributions with mean $t_{\mathrm{e},i}$, squared coefficient of variation $c^2_{\mathrm{e},i}$, and offset $\Delta_{\mathrm{e},i}$ as presented in Table 2.2. The t_{e} and c^2_{e} values of Table 2.2 are verified using Equation (2.6). To properly apply these equations, the two natural process time distributions of a workstation are first translated into a general natural process time distribution. Let X denote the overall natural process time and X_0 and X_1 reflect the type-specific natural process times. Then

$$t_{0,i} = E\left[X_i\right], \quad \text{for } i = \{0,1\}, \tag{2.10}$$

$$c^2_{0,i} = \frac{E\left[X_i^2\right]}{\left(E\left[X_i\right]\right)^2} - 1, \tag{2.11}$$

$$E\left[X^2\right] = 0.01 + 0.1\left(t_{0,0}+t_{0,1}\right) + 0.5\left(t^2_{0,0}\left(c^2_{0,0}+1\right) + t^2_{0,1}\left(c^2_{0,1}+1\right)\right), \tag{2.12}$$

Table 2.2: MEASURED EPT-PARAMETERS FOR EXAMPLE II OF CHAPTER 2 WITH A SINGLE EPT-DISTRIBUTION

WS_i	$\Delta_{e,i}$	$t_{e,i}$	$c^2_{e,i}$
WS_0	0.00	1.27	1.66
WS_1	0.30	1.39	1.59
WS_2	0.50	1.39	0.82
WS_3	0.30	1.39	1.59
WS_4	0.50	1.39	0.82

$$t_0 = E[X] = 0.1 + \frac{t_{0,0} + t_{0,1}}{2}, \tag{2.13}$$

$$c_0^2 = \frac{E\left[X^2\right]}{(E[X])^2} - 1. \tag{2.14}$$

Equations (2.6), (2.13) and (2.14) yield $t_{e,0} = 1.27$, $c^2_{e,0} = 1.66$, $t_{e,1} = t_{e,3} = 1.39$, $c^2_{e,1} = c^2_{e,3} = 1.59$ and $t_{e,2} = t_{e,4} = 1.39$, $c^2_{e,1} = c^2_{e,3} = 0.82$. As can be seen in Table 2.2, the estimated EPT-parameters are correct. When considering the input distributions, one knows that $\Delta_{e,0} = 0.0$, $\Delta_{e,1} = \Delta_{e,3} = 0.3$ and $\Delta_{e,2} = \Delta_{e,4} = 0.50$, which also corresponds to the values presented in Table 2.2.

The observed flow line performance is $\delta = 0.462 \pm 0.01\%$ and $\varphi = 15.70 \pm 0.01\%$. This implies that, with a probability of 95%, the range (0.46195,0.46246) contains the true value of δ and the range (15.69843,15.70157) contains the true value of φ.

Next, shifted Gamma distributions with parameters as presented in Table 2.2 are used as input for an ELP model. The ELP model approximates $\widetilde{\delta} = 0.444$ and $\widetilde{\varphi} = 16.74$, which means that the difference between the EPT approximation and the original situation is 4.0% for throughput δ and 6.6% for flow time φ.

Part of these errors can be explained as follows. Firstly, the ELP model assumes identically and independently distributed (iid) process times on all workstations. In the case considered here, each lot is of a different type than the preceding one. Since $t_{i,0}$ differs from $t_{i,1}$ for $i = 1,2,3,4$, a correlation is expected between successive process times on a workstation. Due to the assumption of iid process times in the ELP model, these correlations between successive process times on a workstation are neglected. Secondly, in the ELP model, the process times of one lot on the successive workstations are assumed to be independent. In the original model however, process times for one lot on successive workstations are correlated due to the type-specific natural process times. The lumped parameter model again does not incorporate this correlation.

The error in the approximation can be reduced by fitting EPT-distributions for each product type per workstation. The new distributional properties are presented in Table 2.3. Comparing these values with Equations (2.13) and (2.14) again shows that the estimated values are correct. Inserting the distribution

Table 2.3: Measured EPT-parameters for example II of Chapter 2 with deterministic lot type sequence and product type specific EPT–distributions

	pt_0			pt_1		
WS_i	$\Delta_{e,i}$	$t_{e,i}$	$c^2_{e,i}$	$\Delta_{e,i}$	$t_{e,i}$	$c^2_{e,i}$
WS_0	0.00	1.27	1.66	0.00	1.27	1.66
WS_1	0.70	2.03	1.07	0.30	0.76	1.63
WS_2	0.50	0.76	1.11	1.30	2.03	0.42
WS_3	0.70	2.03	1.07	0.30	0.76	1.63
WS_4	0.50	0.76	1.11	1.30	2.03	0.42

properties of Table 2.2 into the lumped parameter model yields $\widetilde{\delta}$ = 0.460 and $\widetilde{\varphi}$ = 15.78, which is an error of 0.4% for throughput and 0.5% for flow time.

The latter procedure is repeated for different levels of buffering. If both the original system and the lumped parameter model contain no buffer spaces, the original model gives performance measures δ = 0.364 and φ = 10.16, whereas the approximation finds $\widetilde{\delta}$ = 0.358 and $\widetilde{\varphi}$ = 10.29, giving an error of 1.7% for throughput and 1.0% for flow time. Increasing the number of buffer spaces on all workstations to 5 leads to δ = 0.565 and φ = 25.06 for the original model compared to $\widetilde{\delta}$ = 0.565 and $\widetilde{\varphi}$ = 25.05 for the approximation. This is an error of less than 0.1% for both throughput and flow time. These results correspond to the observations of Section 2.4.

2.5.3 Implications

Two main observations can be derived from the examples presented here. First, the measured EPT-parameters comply with the analytically calculated parameters. Secondly, adding detail to the ELP model, by using product type specific EPT-distributions, results in more accurate approximations.

2.6 Industrial case

A case from an automotive manufacturing plant will be used to illustrate the practical applicability of the EPT-approach.

2.6.1 System description

Experimental data has been obtained from one of the clients of Steelweld B.V. This particular client produces two types of cars, called pt_0 and pt_1 in the remainder of this section. Focus is on a small semi-synchronous flow line within

the manufacturing plant. On this flow line, referred to as FL in the remainder of this section, lots are produced according to a constant product mix, i.e. $pt_0/(pt_0+pt_1) = 0.57$. The actual sequence of lots is determined by an overhead scheduler. Since the scheduler is not considered in this case, the stream of lots entering the system will have a random lot type sequence.

FL consists of a transport system and eleven workstations in tandem (i.e. sequential). The workstations are labeled WS_0 to WS_{10}. Here, WS_1 and WS_2 are manual workstations, served by one operator. Workstations WS_7 and WS_8 are single buffer positions. WS_0 and WS_5 are handling workstations. Workstation WS_{10} is used for (occasional) manual quality checks. All other workstations in the line are used for spotwelding.

2.6.2 First stage of the EPT-approach

The event data needed for the EPT analysis is obtained from the programmable logic controllers (PLCs) within FL. In their present configuration, only possible departures and actual arrivals can be measured using the PLCs; the actual departures thus would have to be reconstructed. However, since the workstations can contain at most one lot at a time, one knows that $a_{i,j}$ will always exceed $d_{i-1,j}$, hence $d_{i-1,j}$ is not required for determining EPT-realizations. However, $d_{i-1,j}$ should be known on the last workstation so that flow times can be computed for validation.

The actual arrival occurs only after transport from the sending workstation to the receiving workstation has ended. Therefore, if the logged actual arrival and possible departure are used, transport is excluded from the EPT-realization. However, the work cycle of these unbuffered workstations always begins with transport. Therefore, the actual arrival should be adapted so that the EPT-realization will include transport. Transport takes a fixed, known amount of time $\Delta_{\min}$, the value of which will not be reported here for reasons of confidentiality. By decreasing $a_{i,j}$ with $\Delta_{\min}$, transport is included in the EPT.

No data was available for WS_7 and WS_8. Therefore, WS_5 is the last workstation on which actual departures can be computed. Hence, workstations WS_6 and above will not be studied in the case.

Since not all gathered events are useable, the data must be filtered. First of all, a number of the events result in EPT-realizations that are unrealistically low or even negative if either possible or actual arrivals are registered too late. Furthermore, since the machines are reliable, large EPT-realizations due to failures and repairs only occur sporadically. Since only a few of these realizations occur within the considered time period, no reliable statistics concerning these high realizations can be obtained. EPT-realization $e_{i,j}$ is thus only used during the analysis if it satisfies Equation (2.15), hence machine failures are excluded.

Table 2.4: AUTOMOTIVE CASE: FITTED DISTRIBUTIONS

WS_i	$t_{e,i}$	$c^2_{e,i}$	$\Delta_{e,i}$	$t_{e,i}/\Delta_{e,i}$
WS_0	82.73	0.106	57.19	1.45
WS_1	76.78	1.259	27.61	2.78
WS_2	94.32	0.765	19.72	4.78
WS_3	116.61	0.149	90.71	1.29
WS_4	112.09	0.077	78.88	1.42
WS_5	130.82	0.021	114.37	1.14

$$\Delta_{\min} \leqslant e_{i,j} \leqslant \Delta_{\max} \tag{2.15}$$

2.6.3 Second stage of the EPT-approach

Distribution fitting, the second stage of the EPT-approach, is done by computing the values for Δ_e, t_e and c_e^2 per workstation from the obtained filtered EPT-realizations, as presented in Table 2.4. On workstations WS_1 and WS_2, the values for c_e^2 are high: these are manual workstations, where only one of the two product types is processed. Due to the manual labor, variance for the product that is processed is already high; however the c_e^2-value is increased even further since the second product type has very low process times on the workstation. The other workstations are robotic workstations, which explains the low values for c_e^2. The data in Table 2.4 have been slightly rescaled, in order to respect the confidentiality of the data. Based on this data, shifted Gamma distributions were fitted for all workstations.

2.6.4 Third stage of the EPT-approach

In the third stage, the shifted Gamma distributions with parameters as presented in Table 2.4 are used as input for an EPT-based aggregate model, a discrete event simulation model in this case. The structure of the model is identical to the structure of *FL*, i.e., six unbuffered single server workstations in a flow line.

A distribution capturing the starvation observed on the first workstation has been obtained from the data to model the starvation of the first workstation in the flow line. In order to obtain this starvation distribution, a filter similar to Equation (2.15) has been applied. The starvation distribution has properties $t_s = 63.63$, $c_s^2 = 2.564$ and $\Delta_s = 29.58$. If it is starving, the first workstation requests a lot from the generator. The generator sends a lot on to the first workstation after an appropriate period of starvation. Similarly, for the final workstation in the flow line, a distribution capturing the observed blocking is obtained. The parameters of this blocking distribution are $t_{b,5} = 15.10$, $c_{b,5}^2 = 8.04$ and $\Delta_{b,5} = 1.97$.

Table 2.5: AUTOMOTIVE CASE: FITTED DISTRIBUTIONS FOR TWO TYPES

	pt_0				pt_1			
WS_i	$t_{e,i}$	$c^2_{e,i}$	$\Delta_{e,i}$	$t_{e,i}/\Delta_{e,i}$	$t_{e,i}$	$c^2_{e,i}$	$\Delta_{e,i}$	$t_{e,i}/\Delta_{e,i}$
WS_0	86.01	0.139	59.16	1.45	78.42	0.041	57.19	1.37
WS_1	40.46	1.400	27.61	1.47	127.52	0.362	67.05	1.90
WS_2	138.68	0.157	86.76	1.60	38.08	0.308	19.72	1.93
WS_3	112.59	0.243	90.71	1.24	121.87	0.036	92.68	1.31
WS_4	105.67	0.021	78.88	1.34	121.89	0.119	110.43	1.10
WS_5	134.26	0.016	120.29	1.12	126.30	0.023	114.37	1.10

The true mean flow time φ of FL is determined by computing the individual flow times from the obtained data and deleting the unrealistic flow times. Flow time realizations are thus again filtered using a filter similar to Equation (2.15). Due to filtering, some EPT-realizations are discarded during data analysis. Consequently, the mean throughput cannot be computed as the amount of bodies produced during the measured time period. Instead, mean throughput δ will be computed by determining the mean interdeparture time of bodies on workstation WS_0.

The ELP model underestimates the throughput $\widetilde{\delta}$ by less than 1.0%, whereas the flow time $\widetilde{\varphi}$ is overestimated by 3.7% (simulation results presented in this section have a confidence level of 99% and a relative width of less than 0.1% of the mean). As can be seen, only a small error remains in the approximation. This error can partially be explained using the inter- and intra-correlations of workstations, as was presented in Section 2.5.2. To improve on this, type-specific EPT-distributions can be fitted, as presented in Table 2.5. The new distributions are used in the ELP model. The model now overestimates both $\widetilde{\delta}$ and $\widetilde{\varphi}$ by less than 1.0%. By adding more detail, the approximations have become more accurate.

2.6.5 Fourth stage of the EPT-approach

A bottleneck analysis is performed, after which the suggested improvements are simulated by accordingly changing the EPT-distributions. It is used to determine which workstations are the major restrictions on throughput and flow time. Workstations with high t_e or c^2_e are potential bottlenecks since they may cause starvation or blocking.

Using the information of Table 2.5, one can see that the values of t_e range from 38.08 to 138.68. Out of this range, acceptable values of t_e seem to lie between 100 and 125 s (although lower values are obviously desirable). Therefore, parameters $t_{e,1,1}$, $t_{e,2,0}$, $t_{e,5,0}$ and $t_{e,5,1}$ are reduced to 125.00 s. Here, the first index refers to the workstation number, the second index to the product type. Furthermore, Table 2.5 illustrates that for most situations, $c^2_e < 0.25$. Reduction of $c^2_{e,1,1}$ and

$c^2_{e,1,2}$ to 0.25 is assumed to be feasible, whereas it is assumed that $c^2_{e,1,0}$ can be reduced to 0.75. The suggested changes have been implemented in the ELP model. Implementation of these changes would, according to the ELP model, result in an increase of 3.5% in δ and a decrease of 4.0% in φ. The simulation study with the unscaled data predicted improvements of the same order of magnitude; which was further confirmed (for the throughput) during implementation on the factory floor; the flow time was not studied during implementation.

2.7 Conclusions

A new method for performance analysis and lumped parameter modeling of single server flow lines subject to blocking has been proposed. The method is based on the effective process time (EPT). In previous work, EPT has only been considered for infinitely buffered, isolated workstations. Here, a calculation method for EPT-realizations for single server flow lines subject to blocking has been presented and validated. The method translates event data (actual and possible arrivals and departures of lots) into EPT-realizations using sample path like equations.

The EPT of a lot is the time experienced by the lot on a workstation from a logistical perspective. It is implemented by means of an approach consisting of four stages, the so-called EPT-approach. In the first stage, EPT-realizations are gathered from industrial data. Next, the realizations are translated into distributions. Typically, distributions are fitted using the first two moments (t_e, c_e). Simulation results however show that for flow lines subject to blocking the offset Δ_e should be used as an additional distribution parameter. In the third stage, an ELP model can be built and validated. Finally, in the fourth stage, the flow line can be optimized.

The EPT-approach has been applied to a case study taken from automotive industry. The ELP model accurately estimated both throughput and flow time. Adding more detail to the ELP model (i.e. including product type-specific shifted Gamma distributions) further reduced errors to less than 1.0%. Based on the EPT-approach, changes in t_e and c^2_e were proposed to increase throughput and to decrease flow time. The presented industrial case shows that the concept can be applied in an industrial context. In the following chapters, we further extend the method to other types of workstations, including finitely buffered multi-server stations and assembly stations where material flows converge.

Chapter 3

Finitely buffered, multi-server flow lines

An effective process time (EPT) approach is proposed for aggregate model building of multi-server tandem queues with finite buffers. Effective process time distributions of the workstations in the flow line are measured without identifying the contributing factors. A sample path equation is used to compute the EPT-realizations from arrival and departure events of lots at the respective workstations. If the amount of blocking in the line is high, the goodness of the EPT-distribution fits determines the accuracy of the EPT-based aggregate model. Otherwise, an aggregate model based on just the first two moments of the EPT-distributions is sufficient to obtain accurate predictions. The approach is illustrated in an industrial case study using both simulation and analytical queueing approximations as aggregate models.

This chapter originally appeared as:
Kock, Etman, and Rooda. Effective Process Time for Multi-Server Flowlines with Finite Buffers. IIE Transactions 40 (3):177-186. 2008
The original publication is available at DOI 10.1080/07408170701488029: http://informaworld.com

3.1 Introduction

Multi-server tandem queues with finite buffers commonly occur in industrial practice. The performance of these lines is typically expressed in terms of throughput and flow time. Irregularities in processing play a key role in the throughput and flow time performance. The blocking of workstations may occur due to a limited buffer capacity.

The performance prediction of finitely buffered multi-server tandem queues is typically performed using discrete-event simulation models (e.g., Banks (1999), Law and Kelton (2000), Baines et al. (2003)) or queueing models. Simulation models are usually more accurate than queueing models since they can incorporate more shop floor effects. However, queueing models tend to be computationally far less expensive than simulation models. Both types of models have to be fed with appropriate data on processing, disturbances and other effects that occur on the shop floor. The methods reported in the literature either assume a distribution or measure individual influences on processing (Chen and Chen 1990, Dallery and Gershwin 1992, Buzacott and Shanthikumar 1993).

In industrial practice, it is often hard to identify and quantify all relevant shop floor details that contribute to the flow time performance of the workstations. Jacobs et al. (2001, 2003) present an algorithm to obtain effective process time distributions for infinitely buffered workstations from lot arrivals and departures. The advantage of their method is that it does not require the quantification of the individual contributing factors. The motivation of their work is to arrive at a measurable metric for variability at a workstation (variance in processing).

In this chapter we generalize this concept to build Effective Process Time (EPT)-based aggregate queueing models of finitely buffered, multi-server tandem flow lines. Using an aggregation based on the EPT paradigm (Hopp and Spearman 1996, 2001), we aim to arrive at simplified queueing models, either simulation or analytical, for which the aggregate process time distribution parameters can be obtained from shop floor event data, such as arrivals and departures.

The contribution of the chapter is two-fold. First, we show that a sample path equation can be used to compute EPT-realizations in multi-server workstations with blocking. Second, we investigate the influence of the shape of the EPT-distribution fit on the accuracy of the EPT-based aggregate queueing model. In particular we consider the offset (i.e., the smallest measured EPT-realization) as a third distribution parameter in a shifted gamma distribution in addition to the EPT mean and variance. The accuracy of both the mean flow time and the variance of the flow time prediction are considered.

The chapter is structured as follows. First we present our proposed aggregate modeling approach using the EPT, and discuss the applicability of the aggregation. Then, calculation of the EPT is presented. This is followed by several examples to experimentally investigate the role of the shape of the EPT-distribution

fit on the accuracy of the aggregate model prediction. Next, in an industrial case problem, the use of EPT-distributions in queueing models and simulation models is illustrated. Finally the main conclusions and some remarks on future work are offered.

3.2 Aggregate modeling using the EPT-approach

Queueing models are used in the prediction of flow line performance. Two well-known classes of models are discrete event simulation models and analytical queueing models.

A simulation model is a representation of the operation of an actual real-world system (Banks 1999) , in our case a manufacturing flow line. In a simulation model, various shop floor details may be modeled in detail. As an example we cite Baines et al. (2003) who included operator behavior in their model. Generally authors attempt to include the most important effects in their model so as to arrive at an accurate simulation model representation of the factory floor. A drawback is that running a simulation model to obtain statistically relevant outcomes may become computationally expensive. An additional difficulty is to obtain all the required data about the shop floor details for inclusion in the model. In practice, some of the data may be difficult to obtain.

Analytical queueing models are an interesting alternative to simulation models. One may distinguish between exact and approximate analytical models. Examples can be found in Dallery and Gershwin (1992), Buzacott and Shanthikumar (1993), MacGregor Smith (2005), Van Vuuren, Adan, and Resing-Sassen (2005) and Van Vuuren and Adan (2005b). Analytical models cannot give as detailed a description as simulation models since they generally contain restrictive assumptions on the details of shop floor behavior that may be included. However, if one can limit the number of states in the model then analytical queueing models are less computationally expensive to evaluate compared to a simulation model. In some cases even exact or explicit approximative expressions can be derived. Even though the number of parameters in analytical models is typically much smaller than in simulation models, feeding the model with appropriate data is then not trivial.

We aim at an aggregate modeling approach that enables one to obtain its parameters from simple events that are readily measurable from the shop floor such as lot arrivals and departures. For this we start by considering the EPT as the aggregate process time distribution.

3.2.1 Concept

The EPT aggregates the raw processing time and all the shop floor operations and disturbances that hamper the processing operation under study, into a single process time distribution. Examples of operations and disturbances are machine breakdowns, setup, rework, operator availability, lot size, metrology, tool change, etc. The combining of multiple phenomena into a single distribution is referred to as aggregation. The EPT concept was introduced by Hopp and Spearman (2001), although the concept of aggregation is of course not new. Hopp and Spearman defined the EPT of a lot as "the time spent by the lot on a workstation from a logistical point of view." They give explicit expressions to compute the mean EPT and the EPT coefficient of variation under various outages, either preemptive or non-preemptive. They use the EPT mean and EPT variance in explicit queueing approximation equations, such as Kingman's equation, to estimate and explain the mean flow time performance.

In many practical cases, the outages may not all be quantifiable. Nevertheless, aggregation approaches (such as the EPT) are appealing, in particular if the EPT can be measured without identifying the contributing factors. For workstations with infinite buffers, a method to actually do this was first proposed by Jacobs et al. (2001, 2003). From lot arrival and departure events they calculate an EPT-realization for each departing lot. By collecting consecutive EPT-realizations, a workstation EPT-distribution is obtained. All influences on processing at the workstation are then aggregated into the EPT-distribution.

This idea may be further generalized into an EPT-based aggregate modeling framework. Then the EPT is not only used as a performance metric that quantifies the effective workstation capacity (mean) and variability (variance), but also to build an aggregate simulation or analytical queueing model. So the idea is that the EPT is a measurable quantity on the factory floor and the aggregate queueing model can stay simple while being fed directly with parameter values obtained from the measured EPT-distributions. The basic approach we propose is as follows.

Step 1 Measure arrival and departure events at the workstations in the manufacturing system, and for multi-server workstations register which lot was processed on which machine.

Step 2 Translate the events into EPT-realizations, one for each departing lot.

Step 3 From the EPT-realizations, compute the mean and variance.

Step 4 Build an aggregate queueing model, either simulation or analytical, using the measured EPT means and variances of the workstations.

In this chapter we develop an EPT-based aggregate modeling approach for multi-server tandem flow lines subject to blocking. Blocking refers to the situation

where a lot cannot leave a machine when its processing on that machine is finished since the receiving buffer of the subsequent station is full. As a consequence, the server cannot commence processing a new lot. Blocking can have a large impact on throughput and flow time performance.

For the aggregate model building of flow lines with blocking we will in particular consider approximative analytical queueing methods such as those developed by Van Vuuren and Adan (2005b) and Van Vuuren (2007). These methods require as input for the workstations the mean and variance of the process time for which we will obviously use the EPT mean and variance. The authors demonstrated, using a range of test problems, the accuracy of their approximation compared to a simulation model representation. A clear advantage of such an analytical approximation is the speed of evaluation compared to running a simulation model.

In the following discussions we will use the following notations and definitions. The mean of the EPT-distribution is denoted as t_e. The ratio of m (the number of parallel machines in a workstation) to t_e quantifies the mean effective capacity available at the workstation. The ratio of the raw processing time t_0 and the mean effective process time t_e quantifies the capacity loss. The latter ratio relates to the industry metric OEE (see e.g., SEMI (2000)) and the revision E proposed by De Ron and Rooda (2005). The squared coefficient of variation of the EPT-distribution is denoted as c_e^2. Following Hopp and Spearman (2001) we refer to this as a quantification of the variability in processing. We call the model in which certain shop floor behaviors are not included explicitly but represented by an aggregate EPT-distribution, an EPT-based aggregate model or simply an EPT-based model. The structure of the EPT-based model (i.e., material flows, number of workstations, number of servers per workstation and number of buffer spaces) is identical to the original system (or detailed model of the original system). Finally, the queuing performance is expressed in throughput (δ (lots/hour)) and flow time (φ (hours)).

3.2.2 Considerations

For certain cases, shop floor behaviors may be aggregated without a significant loss of accuracy. For an $M/G/1$ workstation the mean flow time depends solely on the first two moments of the process time distribution. For a multi-server station with generally distributed arrivals ($G/G/n$) this remarkable property is approximately still valid, provided that the service times and arrivals are phase-type distributed (Adan 2001, Van Vuuren and Adan 2005a). Hence, the performance is predicted accurately as long as the first two moments of the process time distribution are known, regardless of the shape of the distribution function. This implies that, for workstations with infinite buffers, it is sufficient to fit a two-moment distribution (e.g., a gamma distribution) to the measured EPT-realizations.

For finitely buffered flow lines this shape independence property may no longer hold. As a consequence, the first two moments (mean and variance) may not be sufficient to obtain accurate predictions from the aggregate queueing model. In this case the EPT-distribution has to be described more accurately by using a higher-order distribution fit. For instance, in most manufacturing lines, processing at the workstations takes at least some minimum time. The shift or offset may be included as a third parameter in the distribution fit to account for this, e.g., using a shifted gamma or other type of distribution. In Section 3.4 we investigate in further detail the contribution of the offset to the mean flow time for flow lines subject to blocking.

Alternatively, one may decide to include one or more shop floor behaviors explicitly in the aggregate queueing model. For instance, if different lot types give rise to different processing characteristics, one can fit a separate (two-moment) distribution for each lot type. The lot type then becomes an integral part of the aggregate model. For a simulation aggregate model, this poses no additional difficulties. For an analytical aggregate model, new model equations may need to be derived to account for the shop-floor behavior that becomes part of the aggregate model (lot type in the example).

One may also want to leave out a certain shop floor behavior from the EPT-distribution and measure and model it separately. This happens when the time scales of events are different. For instance, when the machines are highly reliable, machine breakdowns occur only very infrequently. Thus, it may happen that for the measurement period under consideration, one may have produced thousands of lots (and thus have obtained the same number of EPT-realizations) while only a couple of machine breakdowns have occurred. If the breakdowns have a significant effect on the shape of the EPT-distribution, but only a few actual breakdown events occur, then no statistically reliable distribution parameter estimates can be obtained. Data on the breakdown behavior should then be collected separately on a different time scale, and be excluded from the EPT. Again, the breakdown then has to be modeled explicitly in the aggregate model. Note in this respect the analytical queueing approximations developed by Tolio, Gershwin, and Matta (2002).

Taking these considerations into account, the EPT-approach may be recast in the following manner:

Step 0 Define the structure of the model, and define which shop-floor realities or disturbances are to be modeled explicitly and thus are to be excluded from aggregation in the EPT.

Step 1 Measure arrival and departure events at the workstations in the manufacturing system; for multi-server workstations register which lot was processed on which machine; obtain data regarding the explicit realities.

Step 2 Translate the events into EPT-realizations, one for each departing lot.

Step 3 Fit for each workstation a suitable distribution to the measured EPT-realizations.

Step 4 Build an aggregate queueing model, either simulation or analytical, using the fitted EPT-distributions.

Step 5 If the EPT model is sufficiently accurate, stop. Otherwise, return to Step 4 and reconsider the distribution fitting or go back to Step 0 and reconsider the aggregation.

Preferably we start by building the simplest possible model, and refine this model when necessary. The accuracy of an EPT-based model may be validated by comparing the estimated throughput and flow time to the throughput of the actual system and the flow time of the lots in the actual system. We will mainly focus on mean throughput and mean flow time. Additional information, such as higher moments or the offset, may be also considered but, as we will show, the required quality of the EPT-distribution fit regarding the actual shape becomes more pronounced.

3.2.3 Application

Once a suitable EPT-based model is obtained, it can serve two main purposes.

First, the obtained EPT-parameters provide insight into the performance of the flow line. Parameter t_e details the average amount of time claimed by a lot at the workstations. The workstation that has the lowest effective capacity is the actual bottleneck. Parameter c_e^2 quantifies the amount of variability associated with the effective processing of lots. Workstations with a high value for c_e^2 may be a problem since they interrupt the steady flow of lots.

Second, the EPT-based model may be used to predict the effect of changes in the line configuration or in numerical optimization procedures. Accurate but quick to evaluate models are then a prerequisite. An analytical model has a great advantage in such cases when compared to a simulation queueing model.

3.3 EPT computation for finitely buffered workstations

Jacobs et al. (2001, 2003) compute EPT-distributions for infinitely buffered multi-server workstations in isolation. They present an EPT-algorithm that computes an EPT-realization for each departing lot. Their algorithm is based on the observation that as long as there are lots in the workstation then capacity is claimed. Each arriving lot starts a new capacity claim if the number of lots in the workstation is less than the number of installed servers. Each departing lot ends its

capacity claim. Thus, the number of ongoing capacity claims is equal to the minimum of the number of lots in the system and the number of servers. The method proposed by Jacobs et al. (2001, 2003) also incorporates time losses due to dispatching issues (assignment of lots to machines) in the EPT, for instance for the case where a server is available for processing but none of the lots waiting in the queue is ever processed on that particular machine. We will refer to this as a violation of the EPT non-idling assumption as we explain later in this section.

Workstations subject to blocking cannot be considered in isolation. We therefore follow a different approach to calculate their EPT values. We show that a simple sample path equation can be used to compute the EPT-realizations in a flow line subject to blocking. The key observation when blocking is present is that the EPT excludes time losses due to blocking. Blocking is excluded since it is due to the finite nature of the buffers. The EPT-based model will also have the same finite buffers, which means that the blocking phenomenon is already covered in the structure of the EPT-based aggregate model. For similar reasons, starvation of a workstation should not be included in the EPT.

3.3.1 EPT for finitely buffered, single server workstations

The EPT for a finitely buffered workstation is computed using three events: the possible departure $pd_{i,j}$ (the time epoch at which workstation j finishes processing the i^{th} lot and tries to send it on to the next workstation in the line); the actual departure $d_{i,j}$ (the time epoch at which the i^{th} lot physically leaves workstation j); and the actual arrival $a_{i,j}$ (the time epoch at which the i^{th} departing lot enters (the buffer of) workstation j). If no blocking occurs, $pd_{i,j} = d_{i,j}$ holds since the receiving workstation has sufficient capacity available to receive the lot. Note that, if transport is instantaneous, $d_{i,j}$ equals $a_{i,j+1}$.

An EPT-realization ends upon the possible departure of the respective lot. The EPT-realization begins at the time at which the workstation could have started processing the lot, that is either at the moment that the lot arrived in the buffer or the moment that the preceding lot was finished. Thus, the EPT-realization begins at $\max a_{i,j}, d_{i-1,j}$ and ends at $pd_{i,j}$. The EPT-realization due to the i^{th} lot departure from the workstation can then be computed from:

$$e_{i,j} = pd_{i,j} - \max\left\{a_{i,j}, d_{i-1,j}\right\}, \tag{3.1}$$

which is a reverse use of the sample path equation for finitely buffered, single-server workstations (Buzacott and Shanthikumar 1993, Adan and Van der Wal 1989); instead of computing departure events, we compute EPT-realizations.

3.3.2 EPT for finitely buffered, multi-server workstations

Calculation of EPTs for multi-server workstations subject to blocking can be done using the same equation. First, sort the processed lots by the machine they were processed on and then apply Equation (3.1) for each machine in the workstation.

This approach to calculating the EPT-realizations per machine assumes that lots waiting in the queue will be processed on the next available machine. This is often referred to as the non-idling assumption. Note that in our case the non-idling assumption has to be interpreted from the EPT point of view. From an EPT point of view the state of a machine that finishes processing a lot changes from busy to available. The machine is from an EPT point of view busy again when the next lot to be processed is present in the queue. Consider a workstation with two parallel machines, with two lots on the workstation. From an EPT point of view, one assumes that both machines process one lot. However, due to lot-dedication, it may be possible that in reality both lots are processed sequentially on the same machine. Then, the EPT non-idling assumption is violated.

The EPT non-idling assumption is violated when a machine becomes available and lots are present in the buffer but none of these lots are processed on the available machine. By applying Equation (3.1) to each machine separately this particular loss of capacity is not accounted for in the EPT and has to be accounted for separately. This case will not be further considered in this chapter.

Finally, if we have an infinitely buffered workstation instead of a finitely buffered one, $pd_{i,j}$ may be replaced by $d_{i,j}$ in Equation (3.1). When the EPT non-idling assumption is satisfied, then it can be shown that using Equation (3.1) is equivalent to the algorithm proposed by Jacobs et al. (2003)

3.4 Examples

In this section, the applicability of the EPT-method for finitely buffered, multiple-server flow lines is evaluated using several examples. First, we briefly illustrate that Equation (3.1) provides the correct EPT-parameters. Next, we show that EPT-based models for finitely buffered flow lines may require more input than just the first two moments of the EPT-distribution.We study this more extensively using the "offset" as the third distribution parameter. Finally, we show that the variance of the flow time distribution may also be approximated using the EPT-approach.

3.4.1 Validation of Equation (3.1)

Consider a two-workstation flow line. The first workstation, which consists of a single server, is never starved. The service time at the first workstation is exponentially distributed with mean process time $\lambda^{-1} = 1.00$ hours/lot. The second workstation, which is never blocked, consists of two (identical) parallel servers and a single buffer space. The process times are again exponentially distributed, with mean process time $\mu = 2.05$ hours/lot.

Following the EPT-approach, events are measured per lot per workstation. These events are the actual and possible departures, and the arrivals. The collected events are used as input data for Equation (3.1), with which EPT-realizations are computed. The gathered EPT-realizations are represented as gamma distributions. For the first workstation, the mean effective process time we measure is $t_{e,0}$ = 1 hour; whereas the squared coefficient of variation is $c^2_{e,0}$= 1. For the second workstation, parameters $t_{e,1} = 2.05$ hours and $c^2_{e,1} = 1$ are measured. These values correspond to the input given above.

3.4.2 Influence of the shape of the EPT-distribution

Consider a line consisting of three unbuffered workstations. The first workstation is never starved, the third workstation is never blocked. The first workstation contains one machine, the second and third workstation each contain two machines. The clean process time on the first workstation is triangularly distributed with minimum 0.9, maximum 1.1 and modus 1.0. On the second and third workstations, the process time is also triangularly distributed, but now with minimum 1.8, maximum 2.2 and modus 2.0.

On all machines, a setup is required after every tenth lot that has been processed. A setup is triangularly distributed with minimum 0.5, maximum 1.5 and mean 1.0. Machines are prone to failure. The busy time between failures is exponentially distributed on each machine with mean $t_f = 15.0$. After a failure, the machine is repaired and the repair time is exponentially distributed with mean $t_r = 3.0$. After a repair, processing of the lot is resumed from the point at which it was left. For this system, the simulated mean flow time is $\varphi = 7.111$. The 95% confidence interval of the simulation results presented in this section is less than 1% of the corresponding parameter.

From this system, EPT-realizations were obtained using Equation (3.1). The mean and variance of the distributions were $t_{e,0} = 1.292$, $c^2_{e,0} = 0.777$, $t_{e,1} = 2.490$, $c^2_{e,1} = 0.400$, and $t_{e,2} = 2.492$, $c^2_{e,2} = 0.405$ for the three workstations respectively. These values were inserted in an EPT-based model. The model approximates $\widetilde{\varphi} = 7.563$ hours. Hence, it overestimates the flow time by 6.4%.

From our measurements, we know that in the real system, the smallest EPTs

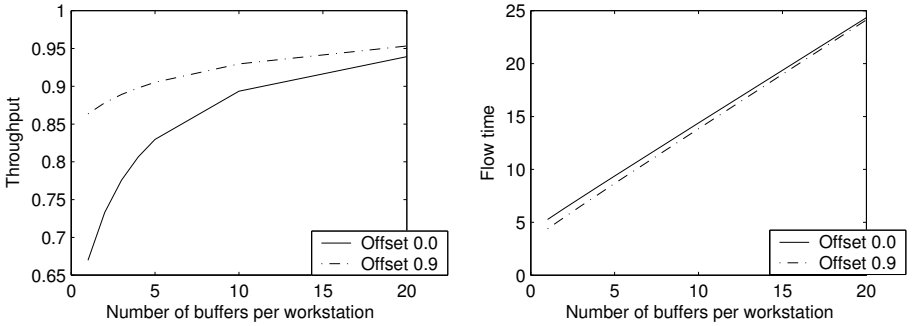


Figure 3.1: INFLUENCE OF BUFFER SIZE ON THROUGHPUT δ AND FLOW TIME φ FOR A THREE WORKSTATION FLOW LINE

measured at the workstations (referred to as offset) were respectively $\Delta_{e,0} = 0.9$, $\Delta_{e,1} = 1.8$ and $\Delta_{e,2} = 1.8$. However, this knowledge is not used in the EPT-based model. By fitting a shifted gamma distribution (Christensen 1989), this offset can be included in the EPT model. The estimated parameters of the shifted gamma distribution are $\Delta_{e,0} = 0.9, t_{e,0} = 1.292, c^2_{e,0} = 0.777, \Delta_{e,1} = 1.8, t_{e,1} = 2.490, c^2_{e,1} = 0.400$ and $\Delta_{e,2} = 1.8, t_{e,2} = 2.492, c^2_{e,2} = 0.405$. Then, the EPT-based model approximates $\widetilde{\varphi}$ = 7.223. Now, the mean flow time is only overestimated by 1.6%. Inclusion of the offset here improves the accuracy of the EPT model.

3.4.3 Relevance of the offset

In many practical cases, a minimum (positive) value for the process time distribution is present (processing requires at least a fixed minimum amount of time). As the previous example illustrates, for flow lines subject to blocking the shape of the EPT-distribution may need to be represented in more detail than obtained by just using the first two moments to obtain a sufficient prediction accuracy of the EPT-based model. In this subsection, we experimentally investigate the contribution of the offset. Our hypothesis is that the shape of the process time distribution (i.e., inclusion of the offset in this example) becomes increasingly important when the flow times on one workstation heavily affect the flow times on other workstations, i.e., when blocking occurs. The stronger the blocking effect, the stronger we expect the shape of the EPT-distribution fit to impact the accuracy of the EPT-based model.

First, consider a three-workstation flow line with one server per workstation. Process times are distributed with a shifted gamma distribution with a mean of one and a squared coefficient of variation of one. The offset (or shift) is taken at 0.0 and 0.9. In Figure 3.1, we see that the influence of the offset is reduced if the buffer size is increased for both the throughput and flow time. Increasing the buffer level corresponds to decreasing the amount of blocking. Hence, this observation confirms our hypothesis.

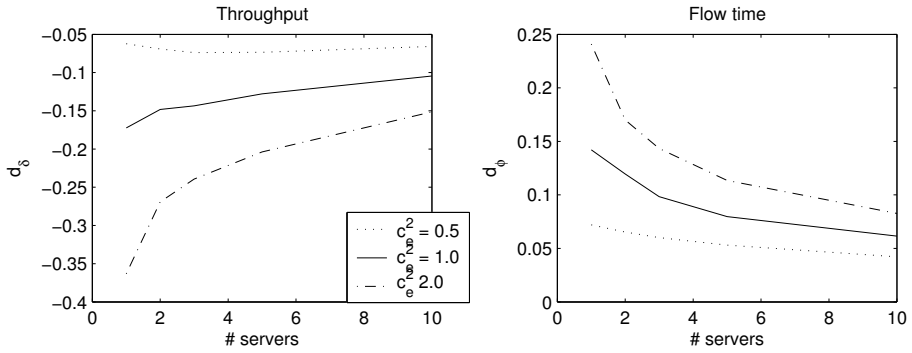


Figure 3.2: RELATIVE DIFFERENCE BETWEEN A TEN-STATION FLOW LINE WITH AND WITHOUT INCLUSION OF AN OFFSET OF 0.9 FOR VARIOUS LEVELS OF VARIABILITY. THE WORKSTATION PARAMETERS ARE $t_E = 1.0$ AND CAPACITY OF 1.

Next, consider a ten-workstation flow line with $n \in \{1,2,\ldots,10\}$ servers per workstation. Each workstation has one buffer space. Process times are distributed according to a shifted gamma distribution with a mean of one and a squared coefficient of variation of $c_e^2 \in \{0.5, 1.0, 2.0\}$ and offsets (shifts) of 0.0 and 0.9 respectively. The results are displayed in Figure 3.2, where $d_\delta = \frac{\delta_{\Delta e=0.0} - \delta_{\Delta e=0.9}}{\delta_{\Delta e=0.0}}$ and $d_\varphi = \frac{\varphi_{\Delta e=0.0} - \varphi_{\Delta e=0.9}}{\varphi_{\Delta e=0.0}}$ From this figure, we see that the influence of the offset becomes smaller when there are more parallel servers in the system. Including extra parallel servers leads to a reduction in blocking. Again, this observation confirms our hypothesis. A second observation from Figure 3.2 is that, if the level of variability in the line (i.e., c_e^2) is reduced, the relevance of the offset also becomes smaller. Reducing the variability implies that the level of blocking is also reduced. Hence, again our hypothesis is confirmed.

From these experiments, we conclude that the offset only needs to be included in the EPT-distribution fit if the amount of blocking is high, that is, for few parallel servers, small buffer sizes and high levels of variability. Otherwise, an EPT-distribution that is fit with just the mean and variance is sufficient. This does not only hold for the offset but also for the distribution shape in general. The advantage is then that analytical queueing models based on the first two moments of the process time distribution, such as proposed by Van Vuuren and Adan (2005b) and Van Vuuren (2007), can be used.

3.4.4 Estimation of the variance of the flow time

Estimation of the variance of the flow time is relevant for instance in the context of customer reliability. In this example, we experimentally investigate the possibility of estimating the second moment of the flow time. Reconsider the three workstation example of Section 3.4.2, where the first workstation consists of one server, while the second and third workstation both have two servers. All three workstations are unbuffered. For that system,we obtained φ =7.111. The

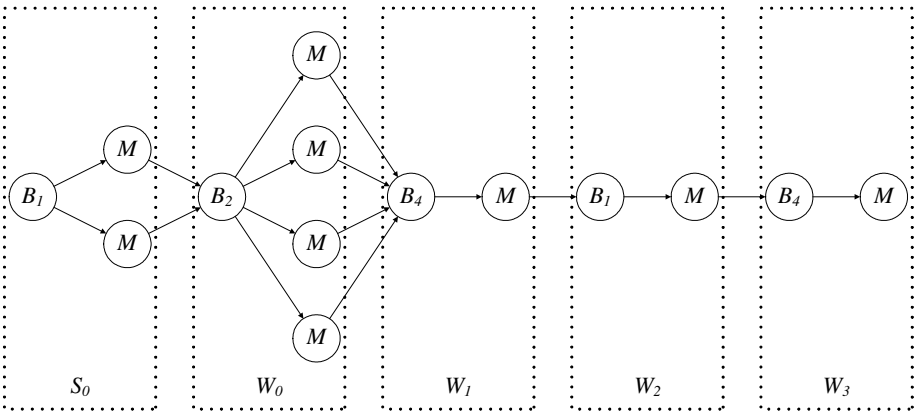


Figure 3.3: LAYOUT OF THE INDUSTRIAL CASE STUDY

variance of the flow time can also be measured: $S_\varphi^2 = 8.611$.

If we build an EPT-based model using solely t_e and c_e^2, then we approximate $\widetilde{\varphi}$ = 7.563 and $\widetilde{S}_\varphi^2 = 6.457$, which are respectively an overestimation of 6.4% and an underestimation of 25%. By explicitly including the offset in the EPT-based model using a shifted gamma distribution, we approximate $\widetilde{\varphi}$ = 7.223 and $\widetilde{S}_\varphi^2$= 8.120, an overestimation of 1.6% and an underestimation of 5.7% respectively.

Including more detail in the distribution fit further enhances the accuracy of the EPT-based model. Therefore, using the work of Osogami and Harchol-Balter (2006), we fit a shifted Erlang-Coxian distribution to the EPT of a machine. Then, we obtain $\widetilde{\varphi}$ = 7.118 and $\widetilde{S}_\varphi^2$ = 8.434, an overestimation of 0.1% and an underestimation of 2.1% respectively. We see that by describing the EPT-distribution in greater detail, the prediction accuracy of the EPT model increases. To accurately predict the variance in the flow time a more detailed distribution fit is required compared to predicting just the mean flow time. Refer to Blom (2007) for more results on this subject.

3.5 Industrial lamp socket case

The proposed method is tested on a second case inspired by industry practice. The industrial case considers a manufacturing line for lamp sockets (see Van Vuuren (2003)). The layout of the case is shown in Figure 3.3.

At supply station S_0, sheets of aluminum are die-cut into small cylinders. The rolls of aluminum sheet arriving at S_0 are large enough to safely assume that S_0 is never starved. The lots of cylinders are transported to W_0, where a screw thread is cut into the cylinders. Then, the lamp sockets are placed inside a glass oven (W_1), and a small amount of liquid glass is poured into the sockets. At W_2, the finishing bath, the socket is soaked in a solvent of either nickel or stain.

Table 3.1: LAMP-SOCKET CASE: PARAMETERS OF THE WORKSTATIONS

Station	m	b	μ_0 [lot/hr]	μ_b [fails/hr]	λ_0 [reps./hr]	p	λ_1 [reps./hr]
S_0	2	1	5.89	0.016	2	0.2	0.8
W_0	4	2	1.54	0.003	2	0.4	0.8
W_1	1	4	3.56	0.040	2	0.8	1
W_2	1	1	32.67	0.020	12	0.5	1
W_3	1	4	16.44	0.040	12	0.5	3

Finally, at W_3, lots are packed into boxes and stored for shipping. It is assumed that W_3 is never blocked.

Workstation S_0 has two parallel servers. At W_0, lots can be placed in a finite buffer of capacity two; the workstation has four parallel machines. W_1 has a finite buffer of capacity four, and one server. W_2 has a single server and a single buffer space; finally W_3 has a single server and four buffer spaces. Note that each single lot in this case corresponds to 6000 bulbs. The process times are approximately constant on the workstations, aside from the failure behavior. The time consumed by a lot on the workstation is thus accurately captured by the clean process time, the busy time between failures (exponentially distributed) and a description of the failure behavior.

In this chapter, failure behavior is assumed to consist of up to two exponentially distributed stages. First, when a machine experiences a breakdown an operator will check whether he or she can make an emergency repair, with rate λ_0. With probability p, the emergency repair suffices and the machine is fixed. With probability $1-p$, the repair is not sufficient and a professional mechanic has to be notified. This mechanic repairs the machine in the second stage with rate λ_1, and repairs the machine with a probability of one. The respective parameters for all workstations are presented in Table 3.1. In the table, b refers to the number of buffer spaces per workstation, m refers to the number of parallel machines, μ_0 is the inverse of the clean process time and μ_b is the inverse of the mean time to failure.

A detailed simulation model was built using the simulation modeling language $\chi - 0.8$ (Hofkamp and Rooda 2002). In the detailed model, workstations have clean process times modified by failures and repairs as quantified in Table 3.1. In this case, the detailed simulation model was treated as the real-life situation, from which the a, pd and d events were measured for each workstation. Using the EPT-algorithms presented in Section 3.3, the EPT-realizations for all workstations were gathered. These EPT-realizations were fitted into (shifted) gamma distributions. The obtained EPT-parameters are reported in Table 3.2. The following EPT-based aggregate models were built: a simulation model in which the offset is incorporated in the EPT-distribution fits (this model is referred to as EA-1); a simulation model in which the offset is included in the EPT-distribution fit at W_1,W_2 and W_3 (referred to as EA-2); a simulation model in which the EPT-

Table 3.2: LAMP-SOCKET CASE: EPT-PARAMETERS OF THE WORKSTATIONS

Workstation	t_e [hr]	c_e^2 [-]	Δ_e[hr]
S_0	0.1738	0.3572	0.1698
W_0	0.6518	0.0143	0.6494
W_1	0.2888	0.1497	0.2809
W_2	0.0310	0.7141	0.0306
W_3	0.0614	0.0988	0.0608

Table 3.3: LAMP-SOCKET CASE: ESTIMATED THROUGHPUT $\widetilde{\delta}$ AND FLOW TIME $\widetilde{\varphi}$

Parameter	Original	EA-1	EA-2	EA-3	EA-4
δ ($\widetilde{\delta}$)[lots/hr]	3.460	3.460	3.460	3.462	3.453
φ ($\widetilde{\varphi}$)[hr]	4.138	4.138	4.139	4.136	4.04

distribution fits have no offsets (i.e., all shifts in the shifted gamma distribution are set to zero) (called EA-3); and a queueing approximation model using the approach of Van Vuuren and Adan (2005b) (labeled EA-4).

Simulation results comparing the four EPT-based models to the detailed model are presented in Table 3.3. These results show that all models are very close to each other, since the amount of blocking and starvation of the bottleneck workstation (W_1) is low. The low level of blocking and starvation is reflected by the obtained throughput (δ = 3.460), which is nearly equal to the theoretical upper bound for the bottleneck ($\delta_{max} = (t_e/m)^{-1} = 0.2888^{-1} = 3.462$). This illustrates that in a (highly) unbalanced line, the level of blocking and starvation at the bottleneck workstation is decisive for the relevance of the offset.

This assertion was tested by changing the configuration of the line. First, the clean process times were changed to make the line more evenly balanced. Furthermore, in order to increase the variance in the line, the mean times between failure were decreased. The changes are given in Table 3.4, along with the new EPT-parameters. The new results of the four EPT models, compared to the original model, are presented in Table 3.5. The relevance of the offset has indeed increased. However, the influence is still reasonably small: for EA-3 the approximation error has grown to 14% for flow time and 4% for throughput. The queueing model (EA-4) tries to approximate the behavior of EA-3. The error present in the queueing approximation happens to cancel out the error induced by neglecting the offset. In other cases, the two errors may add up. Summarizing, the case study illustrates that, for moderate levels of variability and moderate levels of buffering, the shape of the distribution fit (in this case represented by the offset) is not very influential on the prediction of the flow line performance. The EPT-based aggregate models still provide accurate approximations.

The EPT-parameters of Table 3.2 can be used to perform a bottleneck analysis.Workstations with a low effective capacity $r_{e_j} = m_j/t_{e,j}$ (with m_j being the number of servers at workstation j) or high $c_{e,j}^2$ are potential bottlenecks. A

Table 3.4: Lamp-socket case: changed workstation parameters and resulting EPT-parameters

Station	μ_0 [hr]	μ_b [hr]	t_e [hr]	c_e^2 [-]	Δ_e [hr]
S_0	1.78	0.50	0.9836	1.1637	0.5618
W_0	0.89	0.10	1.2639	0.2196	1.1236
W_1	3.56	0.60	0.3990	1.1670	0.2809
W_2	3.56	0.30	0.3301	0.8500	0.2809
W_3	3.56	0.60	0.3230	0.2468	0.2809

Table 3.5: Lamp-socket case: estimated throughput $\widetilde{\delta}$ and flow time $\widetilde{\varphi}$ after changes

Parameter	Original	EA-1	EA-2	EA-3	EA-4
δ ($\widetilde{\delta}$) [lots/hr]	1.925	1.931	1.899	1.860	1.933
ft ($\widetilde{\varphi}$) [hr]	5.586	5.467	5.503	6.396	6.09

closer look at these bottleneck stations may reveal options for improvement. Before they are implemented on the shop floor, the effects of changes in t_e and c_e^2 can be predicted using the EPT-based aggregate model.

3.6 Conclusion and future work

Process time distributions play a key role in the throughput and flow time performance of a multi-server tandem queue subject to blocking. In industry practice, often only average production losses are quantified. In this chapter, an EPT-approach was proposed that enables one to measure aggregate process time distributions of workstations which incorporate outages that delay the processing without the need to quantify each of the contributing factors. The mean and variance of a measured EPT-distribution quantify the effective workstation capacity and variability, respectively, which can be used for bottleneck analysis. The measured EPT-distributions may also be fitted using a suitable distribution function for EPT-based aggregate model building. The EPT-based aggregate model can be either a simulation or an analytical queueing model with the advantage that it does not require the explicit modeling of the shop floor details that are covered by the EPT-distributions.

The EPT-distribution of a finitely buffered, multi-server workstation can be determined using three manufacturing events: (i) the arrival of a lot in (the buffer of) the workstation; (ii) the moment in time at which processing of the lot is finished; and (iii) the departure of the lot from the workstation. Using a simple sample path equation, these events can be translated into EPT-realizations.

For performance prediction using the EPT-based queueing model, often just the first two moments of the EPT workstation distributions suffice. Then, computationally cheap queueing models, such as those proposed by Van Vuuren and

Adan (2005b) and Van Vuuren (2007), can be used with the measured EPT mean and variance as input parameters. However, if blocking plays a major role in the system, then the shape of the EPT-distribution needs to be represented more accurately. This happens when buffer sizes are small or zero, variability is high and only few (or just one) parallel servers are present at a workstation.We have illustrated this in examples using the offset as a "third" distribution parameter, representing a minimum positive process time. We also showed that the EPT-distribution shape needs to be represented in greater detail if an accurate prediction of, for instance, the variance of the flow time is desired.

The EPT-based models presented in this chapter assume that the EPT non-idling assumption holds. This implies that, from an EPT point of view, a server is not idle if an unprocessed lot is in the buffer. This assumption may be violated when one machine has a long breakdown and the other machine(s) in the workstation take over. Jacobs et al. (2003) proposed a method to cope with such a situation for infinitely buffered multi-server workstations.

The method developed in this chapter is potentially very interesting for performance analysis of asynchronous assembly lines, as for instance encountered in automotive industry. Assembly of various components into an assembled part occurs at various stages of production. The next chapter investigates the EPT of an assembly machine, and the role of transport therein.

Chapter 4

Assembly lines

In many manufacturing systems, assembly is used to merge components. Multiple lines feed an assembly workstation that combines various components into a part. We propose an approach for performance measurement and prediction of finitely buffered assembly lines, with particular focus on the assembly station. The proposed method is based on the effective process time (EPT). We aggregate the various types of disturbances on the shop-floor into workstation EPT-distributions. For this, we contribute a model that builds EPT-distributions for each separate workstation, both in the main line and in the feeding lines. Equations are presented to compute the EPT-realizations at the assembly station. Two examples show that accurate, yet simple approximation models can be built of assembly lines using the proposed method. The proposed EPT-method also provides new opportunities to derive analytical queueing approximations for assembly lines.

This chapter is submitted as:
Vijfvinkel, Kock, Etman, Van Vuuren, and Rooda. Performance measurement and prediction of finitely buffered asynchronous assembly lines: an effective process time approach. *submitted* 2008

4.1 Introduction

Assembly lines play a major role in the manufacturing of various types of products. Assembly lines are used to merge components, either into a new, bigger component or an end-product. Thus, in an assembly line, material flows converge. Assembly lines can be categorized as either synchronous lines or asynchronous lines. In a synchronous line, products are transported from all workstations at the same time; the bottleneck workstation determines the speed of the other workstations. In an asynchronous line, transport of products between workstations need not occur at the same time for all workstations. In this chapter, we consider performance analysis of asynchronous, finitely buffered assembly flow lines.

Throughput and flow time are two measures that quantify the overall performance of a manufacturing line. Throughput is the number of lots that are processed per unit of time; flow time refers to the mean time spent by a lot (or, a product) in the manufacturing line. Due to the finite buffercapacity in the line, capacity losses due to e.g. machine downs, setup, or rework, and the corresponding variability may cause blocking in the system. Blocking implies that a workstation cannot start processing a new part, the finished (old) part remains on the workstation since the buffer of the downstream workstation is full. This results in throughput loss in the overall system, and consequently also in an increase in the average flow time. Therefore, one has to carefully monitor capacity losses and variability at workstations in the manufacturing line, and take appropriate actions where necessary.

In several industries, it is nowadays good practice to quantify capacity losses through the overall equipment effectiveness (OEE) (SEMI 2000). The OEE is a performance measure that quantifies average production losses and splits them into availability losses, performance losses and quality losses. The OEE is highly suited to identify the various types of disturbances that cause capacity losses. However, the amount of variability in processing that results from the disturbances is not quantified.

Aside from performance measurement, performance prediction and performance optimization through the use of models is typically desired. Queueing models can be used to predict the effect of changes in the configuration of the manufacturing line on its performance. For asynchronous assembly lines subject to blocking, one may distinguish discrete-event simulation (see e.g. Banks (1999), Law and Kelton (2000)) and analytical queueing approximation models, see e.g. Chen and Chen (1990), Dallery and Gershwin (1992), Buzacott and Shanthikumar (1993), Gershwin (1994), Kim and Alden (1997), Li (2004), Li et al. (2005), Diamantidis et al. (2007), and Van Vuuren (2007). Queueing approximation approaches are often based on aggregate model descriptions. One of the main problems is how to provide the appropriate distribution data, in which the various shop-floor realities are included. Typically, restrictive assumptions are made.

On the other hand, simulation approaches allow incorporation of the various shop-floor realities, see e.g. Pierce (1994), McMullen and Frazier (1998), Banks (1999), Law and Kelton (2000), Hsieh (2002) and Mendes et al. (2005). However, these realities need to be properly quantified. In large-scale manufacturing systems, it is difficult to keep such data up to date, and often it is not feasible at all. Shanthikumar et al. (2007) surveyed data collection as one of the challenges in queueing modeling of complex manufacturing systems. In addition, simulation models are computationally expensive compared to queueing approximations.

In this chapter, we follow an aggregate modeling approach that requires the parameters in the aggregate model to be obtained from simple, measurable events from the shop-floor, such as arrivals and departures of lots at workstations. We refer to this approach as the Effective Process Time (EPT) approach. This approach combines both the data collection and the performance prediction, and aims to arrive at simple, fast and accurate models, either simulation or analytical queueing models.

Jacobs et al. (2001, 2003, 2006) considered the EPT of isolated workstations with an infinite buffer. Starting from the EPT concept of Hopp and Spearman (2001), they proposed algorithms to measure the effective process time at a workstation without identifying the individual contributing phenomena. They considered both single-lot and batching-type machines. Chapters 2 and 3 extended this work to finitely buffered manufacturing flow lines with one or multiple single-lot server(s) per workstation. Their research, in a way, starts from a reverse approach of the sample-path analysis.

By measuring simple events from the shop-floor, Chapters 2 and 3 obtained effective process time distributions. The measured EPT-distributions provide a great deal of information on the workstations investigated and the behavior of the manufacturing flow line. First, the mean effective process time t_e indicates the average time that a workstation effectively requires to process a lot (which relates to the mean capacity), and thus can be used in a bottleneck analysis. Second, the coefficient of variation of the distribution, c_e^2, quantifies the effective level of variability in the workstation. The first two moments of the effective process time distribution can be perfectly used in analytical queueing approximations. This may yield efficient and accurate approximations for finitely buffered manufacturing flow lines. But the EPT-based aggregate model may also be a simulation model, which becomes very easy to develop.

This chapter extends the EPT-approach to asynchronous manufacturing flow lines subject to blocking that include assembly workstations where several material flows merge into one. First, the effective process time approach is discussed in further detail. Next, a procedure to obtain EPT-realizations from assembly workstations with transport systems is proposed. The simulation results that are presented for two examples show that the EPT-based aggregate model is an accurate approximation of the original detailed assembly line.

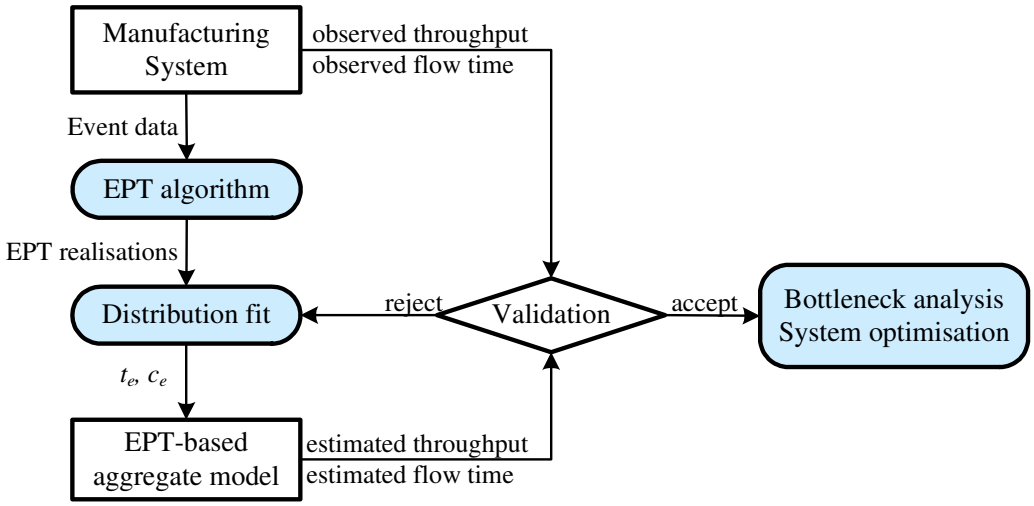


Figure 4.1: Schematic representation of the EPT–approach

4.2 The effective process time

Hopp and Spearman (2001) define the effective process time (EPT) as "the time spent by a lot on a workstation from a logistical point of view". The EPT includes all disturbances on processing into a single distribution. The notion of combining all individual influences on processing into a single distribution is also used in the context of sample path analysis (Chen and Chen 1990, Dallery and Gershwin 1992, Buzacott and Shanthikumar 1993). Here, it is referred to as completion time, processing time or service time. EPT can be measured from a manufacturing line using the EPT-approach.

The ideas and assumptions underlying the EPT-approach are described in detail in Chapter 2. Here we summarize the approach by means of Figure 4.1. Four steps are distinguished. The first step in the EPT-approach is to measure EPT-realizations from an existing, operating manufacturing system. An EPT-realization represents the time a lot consumed capacity from the respective workstation. EPT-realizations are obtained from event data, such as arrivals and departures of lots on workstations. For manufacturing flow lines subject to blocking, Chapters 2 and 3 show that sample path equations can be used to compute the EPT-realizations. So, instead of computing departures from arrivals and effective process times with the sample path analysis, the equations are used reversely to reconstruct the effective process times from the departure- and arrival-events measured in the production system.

The second step is to fit the EPT-realizations to distributions. Here, distributions are fitted based on relevant workstation properties. As discussed above, the relevant parameters are usually the mean EPT t_e and the squared coefficient of variation c_e^2. Chapters 2 and 3 showed that the offset (i.e., the smallest measured effective process time of a lot) Δ_e may, for finitely buffered, single-server workstations, also be relevant for distribution fitting. They compared simulation

results for flow lines with and without an offset. They observed differences in throughput of up to 50%, in particular if buffer sizes are small, or even zero, just a single server is present in the workstations, and variability of the workstations is high.

In the third step, an EPT-based aggregate model is built from the measured EPT-distributions. The model may be a simulation model or an analytical queueing model. The EPT-distributions are measured directly from the operational manufacturing system, without quantifying the individual disturbances on the shop-floor. The EPT-based model is typically used for performance analysis of the current configuration of the manufacturing system. The structure of the EPT-based aggregate model follows the original system to a large extent (e.g. the number of servers at each workstation, the buffer sizes of workstations and the flow of materials between workstations). In the aggregate model, detailed modeling of shop-floor realities such as failures, repairs, setups, operators and lot sizes is avoided. The complex workstation behavior, the shop-floor realities included, is described in the EPT-distributions of the workstations.

The fourth step is to validate the EPT-based aggregate model by comparing the throughput and flow time as estimated by the model to the throughput and flow time observed in the actual system. If the estimated throughput and flow time are found to be accurate enough, the aggregate model and the EPT-distributions are accepted. If not, distribution fitting and aggregate model building are reconsidered. Possible changes include enhancing the level of detail of the model (e.g. excluding specific shop-floor realities from the EPT-realization) or using more parameters to fit more accurate EPT-distributions.

The measured EPT-distributions and the corresponding EPT-based aggregate model can be used for performance analysis and optimization of the current configuration of the operational manufacturing system. A bottleneck analysis can be carried out based on the EPT-distribution parameters t_e and c_e^2 of the various workstations. Stations with a high t_e and c_e^2 value may hamper overall line performance. The effect of suggested improvements can be predicted using the EPT-based aggregate model by accordingly adjusting the EPT-distribution parameters in the model.

The EPT-approach provides the following benefits. First, many shop-floor realities do not have to be quantified individually, but are included directly in the measured EPT-distributions. The idea is that the aggregate model is sufficiently accurate, yet simple when compared to the detailed models that are typically used, and that the EPT-parameters can be measured easily from the operating system. Second, since the workstation behavior and shop-floor realities are included in the EPT-distributions and are directly obtained from industrial data, the EPT-parameters t_e and c_e^2 readily give insight in the behavior of the manufacturing system, allowing for straightforward bottleneck analysis, even without building an aggregate model.

In previous work, the concept of EPT has been applied to several types of workstations. Jacobs et al. (2001, 2003, 2006) compute EPT-distributions for infinitely buffered workstations in isolation. They compute the EPT-distribution of a workstation from lot-arrivals and lot-departures; this data is usually available in automated manufacturing environments. Workstations consist of one or more (parallel) single-lot servers (Jacobs et al. 2003) or batch servers (with one or more recipes) (Jacobs et al. 2006). Chapters 2 and 3 considered finitely buffered workstations in a manufacturing flow line. In their work, workstations have either one or multiple single-lot servers.

4.3 EPT for finitely buffered assembly workstations

Here, we develop the concept of the EPT for application to assembly workstations. We consider an assembly workstation with two or more merging material streams (components). For each component, there is a finite buffer between the end of the component line and the assembly machine. We consider assembly workstations that operate in push mode. This means that, if the corresponding buffer in the assembly workstation has sufficient space available, a component is transported into this buffer as soon as the component line finishes it. The assembly process starts as soon as all component types are available. Figure 4.2 shows an example of an assembly workstation, including the last workstation of the component lines and the transport systems. Note that in Figure 4.2, no distinction is made between a main product stream, which supplies the most important component, and sub-component streams, which implies that all component lines have equal importance. In some cases, a main component line may be distinguished.

For an assembly line, the EPT may be identified in two different ways. The first approach analyses the assembly station with explicit inclusion of all feeding component lines. Thus, it builds EPT-distributions for each separate workstation, both in the main line and in the feeding lines. If one of the component lines is a main component line, or if one of the component lines is expected to dominate the performance of the assembly line, it may not be worthwhile to include all component lines in the model. The second approach is then to consider only workstations in the main component line and to aggregate the feeding component lines into the assembly station. Both alternatives are discussed in the next subsections.

4.3.1 Isolation of assembly workstation

By analyzing all workstations in the component lines and the assembly workstation separately, the behavior of the component lines can be isolated from the assembly workstation. The EPT of the assembly workstation then only includes

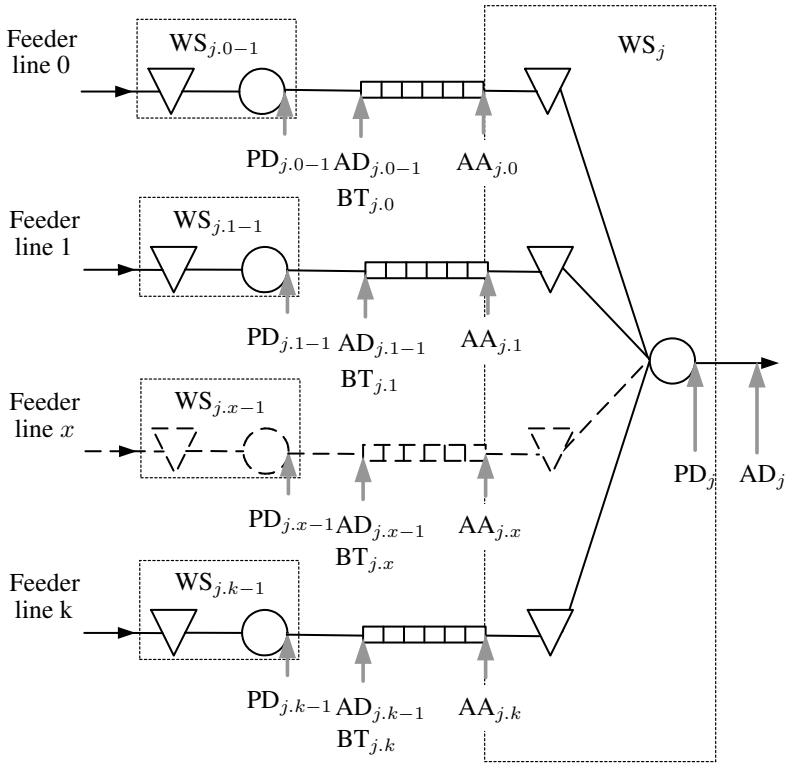


Figure 4.2: Example of an assembly line with finite buffers and transport systems

the processing time, disturbances and other shop-floor realities that relate to processing at the assembly station. The time that a component has to wait for other components is excluded from the EPT.

To compute the EPT of the assembly workstation and the sub-component lines, we need three types of events: arrivals, possible departures and actual departures. The events for component x are depicted in Figure 4.2. Then, we need: $a_{i,j}$, $pd_{i,j}$ and $d_{i,j}$. Let $a_{i,j.x}$ denote the arrival of component x of assembly i in the component buffer of the assembly workstation. Furthermore, let $d_{i,j}$ be the actual departure of assembly i from assembly workstation j. Event $pd_{i,j}$ refers to the possible departure of assembly i from assembly workstation j; implying that the processing of assembly i has finished (the assembly is ready to be sent on; blocking at the downstream workstation may temporarily prevent this).

If the assembly workstation has a buffer, transport towards the assembly workstation cannot be included in the EPT. Transport between a component line and the assembly workstation begins when the component departs from the previous workstation ($d_{i,j.x-1}$). Transport ends when the component has arrived at the assembly workstation ($a_{i,j.x}$). Once transport of one of the components has ended, it is possible that not all components are available. Then, these components have to wait for processing until all components have arrived. The EPT commences when all component types are available and the assembly machine is idle, and ends when processing has finished. The sample path equations to compute the

EPT-realization $e_{i,j}$ for lot i on workstation j and transport time $tt_{i,j.x}$ for the corresponding component x are:

$$e_{i,j} = pd_{i,j} - \max\left\{\max_{x\in 0,\ldots,k}\left(a_{i,j.x}\right), d_{i-1,j}\right\} \tag{4.1}$$

$$tt_{i,j.x} = a_{i,j.x} - d_{i,j.x-1} \tag{4.2}$$

where we assume that lots do not overtake. Note that if transport is instantaneous, only Equation (4.1) is required to compute EPT-realizations. However, even if the assembly workstation has no buffers, transport towards the assembly workstation cannot be included in the EPT. In this chapter, Equations (4.1) and (4.2) will be referred to as Algorithm 1AMB-t (1 Assembly Machine with finite Buffers and Transport).

4.3.2 EPT of the main component line

In some cases, one may not want to model all individual component lines. If a main component line can be defined, or if one component line is expected to dominate the overall behavior, one may want to model the assembly station as a simple station, that includes the behavior of the feeding stations. In that case, if the main component has arrived at the assembly workstation, the time spent waiting on sub-components is included in its EPT. Note that this waiting time depends on the behavior of both the main component line and the sub-component lines. If the sub-component lines are slow, and the main component line is fast, then much waiting time will be included, which results in a high EPT. Conversely, if the main line is slow and the feeder lines are fast, little waiting time will be included, which results in a lower EPT. This makes the EPT-distribution of the assembly station depend on the parameter settings at other stations.

For the EPT of the assembly workstation, of all arrivals we only use the arrival of the main component (say, component 0). The EPT is computed similar to the other stations in the line by means of the sample path equation for a workstation consisting of a finite buffer and one single-lot machine Chapter 2. For the assembly station we get, assuming that lots do not overtake:

$$e_{i,j} = pd_{i,j} - \max\left(a_{i,j.0}, d_{i-1,j}\right). \tag{4.3}$$

If a workstation has a buffer, transport towards the workstation cannot be included in the EPT. In that case, transport time is computed using the following equation:

$$tt_{i,j.0} = a_{i,j.0} - d_{i,j.0-1}. \tag{4.4}$$

A special case occurs if the workstation is unbuffered. Transport can then be included in the EPT, since the EPT-realization then always starts at $d_{i,j.0-1}$ and

ends at $pd_{i,j}$:

$$e_{i,j} = pd_{i,j} - d_{i,j.0-1}. \tag{4.5}$$

For a bottleneck analysis, the EPT-parameters are easily interpretable. For the assembly workstation, high EPT-parameters (compared to the other workstations in the main stream) may be caused by bad performance of the workstation itself, or by bad performance of at least one sub-component line.

A disadvantage of this method is that if one of the workstations in the feeding component lines is the bottleneck, the exact cause cannot be identified, since its contribution is aggregated in the EPT of the assembly workstation. Another drawback is that the resulting EPT-based model cannot be used to predict the effect of changes in the line configuration on the performance. This is because the EPT of the assembly workstation depends on the behavior of the main component line and thus is only valid for this specific system configuration for which the measurements have been carried out.

4.4 Assembly workstation test example

This simulation example illustrates the use of EPT-Algorithm 1AMB-t. The example demonstrates the EPT-based aggregate model for varying numbers of component workstations. The example shows that the transport time and EPT-realizations measured by Algorithm 1AMB-t are correct. Furthermore, the example shows that the EPT-based aggregate models are accurate approximations of the system that is analyzed.

Consider an assembly line consisting of one assembly workstation, that is fed by a number of (parallel) component workstations, ranging from 1 to 11. The component workstations consist of one single-lot machine. The assembly workstation consists of one assembly machine and a finite buffer for every component type (possibly with size zero). All machines in the line process lots (on average) equally fast, thus the line is balanced. We also obtained results for an unbalanced line. Since these results correspond to the balanced case, we omit them here. The component workstations are never starved, while the assembly workstation is never blocked. The clean process times are exponentially distributed with mean $t_0 = 1.0$. A machine fails after an exponentially distributed busy time with mean $t_f = 15.0$. Upon failure, a machine is repaired after an exponentially distributed repair time with mean $t_r = 2.0$. After a failure is repaired, the remaining process time is completed. For this example, with known disturbances and clean process times, the EPT-parameters can be derived analytically, which gives mean effective process time $t_e = 1.1333$, squared coefficient of variation $c_e^2 = 1.4152$ and offset $\Delta_e = 0.0$.

Transport times between the component workstations and the assembly workstation are triangularly distributed with minimum 0.2, mean 0.3 and maximum 0.4

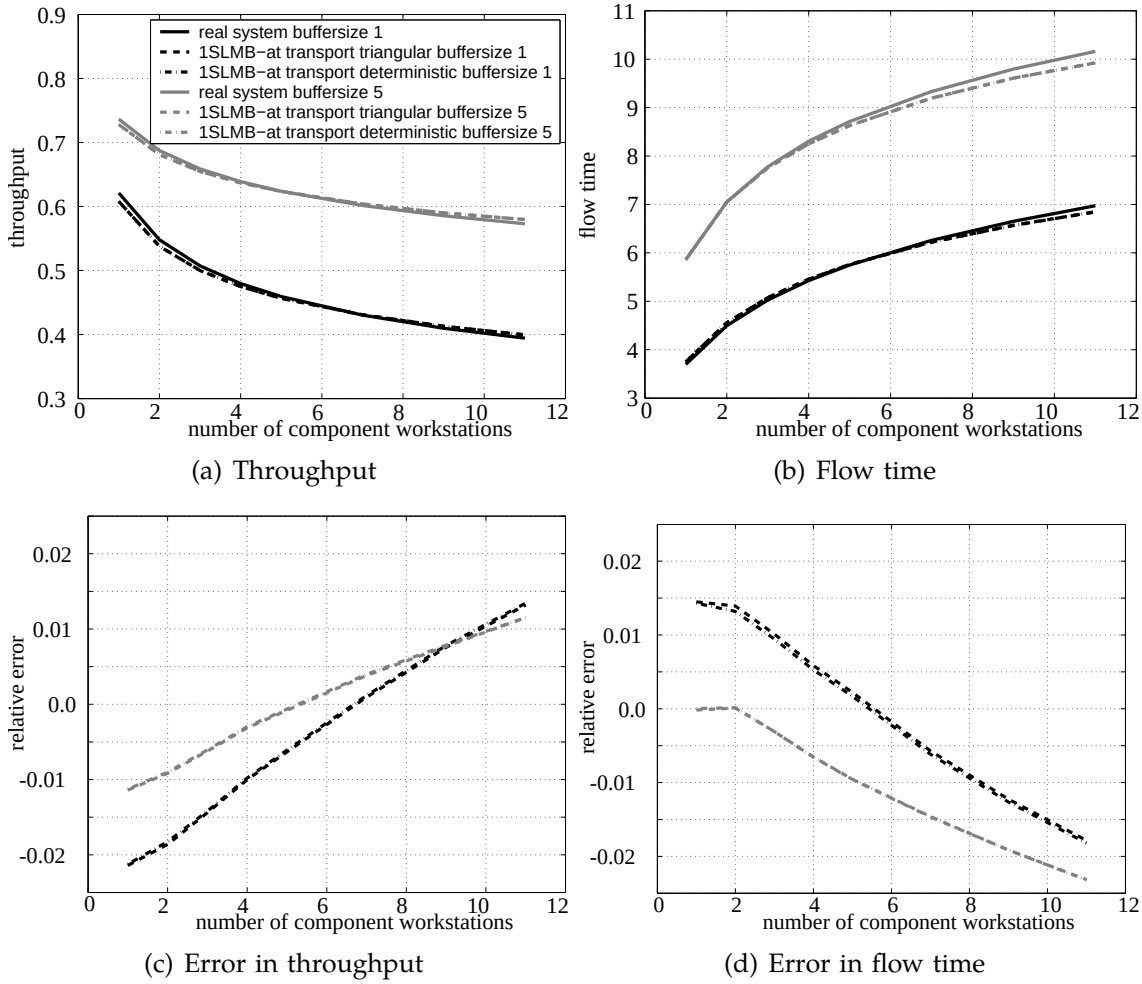
which leads to mean transport time $t_t = 0.3$, transport time squared coefficient of variation $c_t^2 = 0.01852$ and minimum transport time $\Delta_t = 0.2$. The choice of these triangular distributions for transport reflect that in reality transport between workstations is only lowly variable, much less than the variability in processing. This is observed in many assembly lines, e.g. in car manufacturing.

The EPT-parameters are measured from the operational manufacturing system, in this test example the simulation model described above. First, the required events are measured in the real system (represented by a detailed simulation model) and converted into EPT- and transport time realizations using Algorithm 1AMB-t. From the collected realizations, t_e, c_e^2, Δ_e, t_t, c_t^2 and Δ_t can be estimated. The measured values correspond to the analytical ones (error$< 0.1\%$). Hence, Algorithm 1AMB-t measures the correct EPT- and transport time realizations.

Next, EPT-distributions are fitted to the TT- and EPT-realizations and EPT-based aggregate models are derived. The EPT-realizations for processing are fitted into shifted gamma distributions. The EPT-based aggregate models have the same system structure as the real system, but the real machine behavior is replaced by the EPT-distribution. For this example, two types of models were built. In the first type, triangular distributions are fitted to the TT-realizations. In the second type, transport is modeled deterministically by using t_t as deterministic transport time.

In Figure 4.3, the throughput and flow time predicted by the two EPT-based aggregate model types are compared with the throughput and flow time of the real system. The main observation from this figure is that the EPT-based aggregate models are very accurate. The error in approximation of flow time and throughput is less than 2%. Next, one can see that modeling the transport deterministically is only slightly less accurate than modeling transport with a distribution, for the case in which transport has a low variability. The insight gained here, is that transport may be modeled as a constant in many practical cases.

The simulation results presented here have been obtained for simulation run lengths of 230.000 lots, the first 30.000 of which constituted the transient phase. A simulation experiment consists of at least 6 simulation runs. Using these simulation runs, 95% confidence intervals, based on the student-t distribution, on the throughput and flow time are computed. Extra simulation runs are performed until the relative width of these confidence intervals is smaller than 0.1% of respectively the throughput and flow time. The simulations were conducted in the specification language χ-0.8 Hofkamp and Rooda (2002).



(a) Throughput (b) Flow time

(c) Error in throughput (d) Error in flow time

Figure 4.3: Prediction accuracy of EPT-based models for an assembly station with feeding component workstations

4.5 Assembly line case problem

Next, we present a case problem of an assembly line in a car manufacturing environment. The case is inspired on an industrial case by VDL Steelweld b.v. of a part of an automotive plant, previously described in Chapter 2. Here, two assembly workstations are fed with components from a component line. Components are transported to the main line on a conveyor belt. Unfortunately, as of yet, we could not extract the necessary events from the PLCs of the conveyor belt to measure the EPT-realizations for the two assembly stations in the main line. To nevertheless demonstrate the potential of the proposed method, we replace the real operating line by a simulation model of the line, where we explicitly include machine failures and repair as shop-floor realities that we want to aggregate in the EPT models. The input chosen for this reality is inspired on the knowledge we have from the VDL Steelweld simulation case, however numbers are modified to respect the confidentiality of the data.

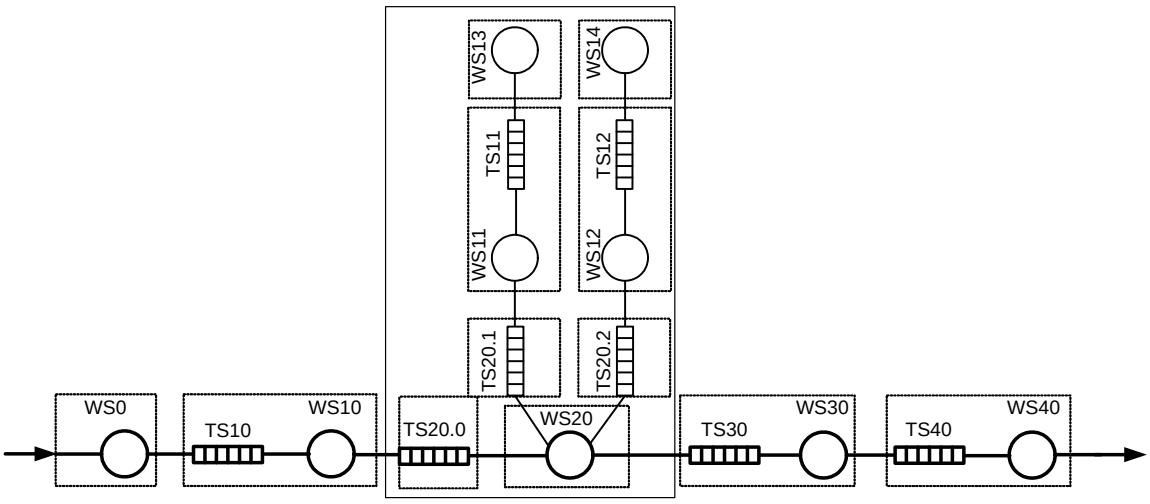


Figure 4.4: Nine-workstation assembly line

The EPT-approach is used to build EPT-based aggregate models of the (simulation) industrial reality. The case illustrates that these models are accurate approximations of the assembly line. In addition, the case shows that the EPT-parameters and EPT-based aggregate models can be used to improve system performance.

Case description

The assembly line consists of one main component line, that is fed by two sub-component lines. A main component and two sub-components are assembled in assembly workstation WS_{20} and continue as a single part. The line comprises nine workstations and eight transport systems of which workstations WS_{00}, WS_{10}, WS_{20}, WS_{30}, WS_{40} and transport systems TS_{10}, $TS_{20.0}$, TS_{30} and TS_{40} form the main component line. Main components enter the line at WS_0, subcomponents enter at WS_{13} and WS_{14}, assembled products leave the line at WS_{40}. The line is considered in isolation, which means that WS_0, WS_{13} and WS_{14} are never starved and WS_{40} is never blocked. The assembly line is visualized in Figure 4.4.

A workstation consists of one machine and has no buffer. Machines have clean process times which are gamma distributed with mean t_0 and squared coefficient of variation c_0^2. Machines fail after a certain busy time which is exponentially distributed with mean t_f. Upon failure, a machine is repaired after an exponentially distributed repair time with mean t_r. The machine parameters can be found in Table 4.1. The assembly line operates in push mode. The assembly process starts as soon as a main component and two subcomponents are available at the assembly machine.

Components and assembled products are transferred between workstations via transport systems. Transport can only take place if the receiving workstation is

Table 4.1: Workstation parameters

Workstation	t_0	c_0^2	t_f	t_r	Availability
WS_{00}	1.2	0.7	38.0	2.0	95.0%
WS_{10}	0.5	0.1	60.0	2.0	96.8%
WS_{11}	0.5	0.1	60.0	1.0	98.4%
WS_{12}	0.5	0.1	60.0	1.0	98.4%
WS_{13}	1.5	1.0	18.0	2.0	90.0%
WS_{14}	1.5	1.0	31.5	3.5	90.0%
WS_{20}	1.1	0.7	45.0	5.0	90.0%
WS_{30}	0.5	0.1	60.0	2.0	96.8%
WS_{40}	1.2	0.7	38.0	2.0	95.0%

Table 4.2: Transport system parameters

Transport system	t_{min}	$\bar{t}$	t_{max}
TS_{10}	0.20	0.25	0.30
TS_{11}	0.08	0.10	0.12
TS_{12}	0.08	0.10	0.12
$TS_{20.0}$	0.20	0.25	0.30
$TS_{20.1}$	0.08	0.10	0.12
$TS_{20.2}$	0.08	0.10	0.12
TS_{30}	0.20	0.25	0.30
TS_{40}	0.20	0.25	0.30

ready to accept the component or assembled product (i.e. it has enough storage capacity). Transport times are triangularly distributed with minimum t_{min}, mean $\bar{t}$ and maximum t_{max}. The transport parameters can be found in Table 4.2.

The performance of the assembly line is determined by means of a discrete event simulation model. The model gives $\delta = 0.3773$, $\varphi_0 = 10.21$, $\varphi_1 = 9.486$ and $\varphi_2 = 9.485$, where δ denotes the throughput and φ_0, φ_1 and φ_2 denote the flow times of respectively the main component and two sub-components. These results are obtained with an accuracy such that the relative width of the 95% confidence interval is at most 0.1% of the corresponding mean.

EPT-based analysis of the main component line

First, the main component line only is investigated by using the approach presented in Section 4.3.2. Therefore, in the analysis, the behavior of WS_{11}, WS_{12}, WS_{13} and WS_{14} is included in the EPT of WS_{20}. As noted in Section 4.3.2, since the workstations are unbuffered, transport towards a workstation can be included in the EPT of the corresponding workstation. As a result, the only events to be measured are the transport start times and the process finish times for the workstations in the main component line. For these workstations, the EPT-realizations are computed by Equation (4.5). The measured EPT-parameters are reported in

Table 4.3: EPT-parameters of the main component line using Equation (4.5)

Workstation	t_e	c_e^2	Δ_e
WS_{00}	1.2635	0.8560	0.0000
WS_{10}	0.7666	0.1582	0.2751
WS_{20}	2.2143	1.0691	0.2045
WS_{30}	0.7669	0.1614	0.2958
WS_{40}	1.5137	0.6011	0.2028

Table 4.4: EPT-parameters of the sub-component lines

Workstation	t_e	c_e^2	Δ_e
WS_{11}	0.6084	0.1162	0.1542
WS_{12}	0.6083	0.1145	0.1606
WS_{13}	1.6674	1.2401	0.0000
WS_{14}	1.6656	1.4179	0.0000

Table 4.3. In Figure 4.4, each part of the structure of the aggregate model is indicated by a box surrounding it.

The EPT-parameters in Table 4.3 show that in the present configuration, the combination of the assembly workstation with the component lines is the bottleneck. However, since the behavior of the component lines is aggregated into the EPT of the assembly workstation, one does not know whether the bottleneck is the assembly station or one of the workstations in the component lines. Applying solely the EPT measurement method as presented in Chapter 2 apparently does not suffice. In order to find the exact bottlenecks, the behavior of the sub-component lines must be investigated.

EPT-based aggregate model: main and sub-component lines

The behavior of the assembly workstation and sub-component lines is investigated in greater detail by using the approach presented in Section 4.3.1. The behavior of the assembly workstation is further refined to the individual workstations of the sub-component lines and the assembly workstation itself. For the assembly workstation, transport is separated from the EPT, using Algorithm 1AMB-t. The EPT-parameters that are estimated from these EPT-realizations, are presented in Table 4.4 and Table 4.5.The EPT-parameters obtained on workstations WS_{00}, WS_{10}, WS_{30} and WS_{40} do not change. In the EPT-based aggregate model, the transport and machine behavior are modeled by a shifted gamma distribution, based on the data from Table 4.5. Combined with the results of Table 4.3, the EPT-based aggregate model gives $\widetilde{\delta} = 0.3680$, $\widetilde{\varphi}_0 = 10.41$, $\widetilde{\varphi}_1 = 9.702$ and $\widetilde{\varphi}_2 = 9.646$ which leads to errors of respectively -2.5%, $+2.0\%$, $+2.3\%$ and $+1.7\%$. Since the model is accurate, we expect that it can be used to predict the effect of modifications of the assembly line.

Table 4.5: EPT- and TT-parameters for assembly workstation WS_{20} according to the approach presented in Section 4.3.1

Workstation	t_e	c_e^2	Δ_e
WS_{20}	1.2221	1.5176	0.0000
Transport system	t_t	c_t^2	Δ_t
$TS_{20.0}$	0.2500	$6.67\ 10^{-3}$	0.2000
$TS_{20.1}$	0.1000	$6.67\ 10^{-3}$	0.0800
$TS_{20.2}$	0.1000	$6.67\ 10^{-3}$	0.0800

The measured EPT-parameters in Tables 4.3, 4.4 and 4.5 indicate which workstations restrict line performance most. The table shows that workstations WS_{20}, WS_{13}, WS_{14} and WS_{40} have the largest t_e. Note that the ratio t_0/t_e reflects the capacity loss (recall the t_0 values of Table 4.1). The largest values of c_e^2 were observed on WS_{00}, WS_{13}, WS_{14} and WS_{20}. Thus, improvements in these workstations are likely to have the greatest impact on line performance. In future work, one may consider actual optimization tools here, such as a sensitivity analysis.

On a hypothetical basis, we assume that the sum of the t_e values can be reduced by 0.5 (none of the individual t_e values may be increased though). Similarly, the sum of the c_e^2 values may be decreased by 0.25. This means that the sums of t_e and c_e^2 are both reduced by 5%. The aim is to reduce the largest t_e and c_e^2 as much as possible. This means that the largest t_e and c_e^2 become 1.4489 and 1.2040 respectively, and that the performance of WS_{13}, WS_{14} and WS_{40} needs to be improved.

The suggested modifications are implemented in the EPT-based aggregate simulation model and their effects on throughput and flow time are predicted. The modification predicts $\widetilde{\delta} = 0.3982$, $\widetilde{\varphi}_0 = 9.657$, $\widetilde{\varphi}_1 = 9.289$ and $\widetilde{\varphi}_2 = 9.293$. This would lead to improvements of respectively 8.2%, 7.2%, 4.3% and 3.7%.

To implement the suggested modifications in the original assembly line, the modified EPT-parameters are translated into real workstation parameters. In practice, this step is not trivial because the individual contributors to capacity consumption and variability are not known. In this test case, the translation is simple because the EPT-parameters can be computed analytically. Table 4.6 shows the implementation of the proposed line configuration that was chosen by the authors. Implementing the modifications gives $\delta = 0.4096$, $\varphi_0 = 9.446$, $\varphi_1 = 9.075$, $\varphi_2 = 9.115$, which leads to performance improvements of 8.6%, 7.5%, 4.3% and 3.9%. These results are closely matched by the predictions of the aggregate model. Hence, for this case, the EPT-based aggregate model accurately predicted the effect of modifications in the assembly line.

Table 4.6: New workstation parameters after modifications

Workstation	t_0	c_0^2	t_f	t_r	Availability
WS_{00}	1.2	0.7	38.0	2.0	95.0%
WS_{10}	0.5	0.1	60.0	2.0	96.8%
WS_{11}	0.5	0.1	60.0	1.0	98.4%
WS_{12}	0.5	0.1	60.0	1.0	98.4%
WS_{13}	**1.304**	**0.928**	18.0	2.0	90.0%
WS_{14}	**1.323**	**0.844**	31.5	**3.0**	**91.3**%
WS_{20}	1.1	0.7	45.0	5.0	90.0%
WS_{30}	0.5	0.1	60.0	2.0	96.8%
WS_{40}	**1.139**	**0.695**	38.0	2.0	95.0%

4.6 Conclusions and recommendations

In this chapter, a new method to analyse and predict the performance of assembly lines is proposed. The method is based on the effective process time (EPT), see Hopp and Spearman (2001), Jacobs et al. (2003) and Chapters 2 and 3. The chapter addresses assembly workstations that are subject to blocking. A new EPT-algorithm is derived for assembly workstations with finite buffers, where transport of components to buffers requires time. This EPT-algorithm computes EPT-realizations from easily measurable events on the shop-floor (i.e. product arrival, completion and departure times). EPT-realizations are used to measure workstation performance (indicated by EPT-parameters) and to model workstation performance (by deriving an EPT-based aggregate model). The EPT-realizations are thus measured directly from workstation data, rather than first quantifying all disturbances affecting processing and translating these disturbances into EPT-distributions.

In this chapter, we contribute a method to measure EPT-realizations for assembly systems. The method isolates the behavior of all component lines that feed the assembly workstation. If all components have arrived, the EPT-realization of the assembly commences. In this way, the EPT-realization of the assembly station is independent of the performance of the component lines. The obtained realizations are used to build an EPT-based aggregate model. The EPT-realizations are determined using Algorithm 1AMB-t, developed in this chapter

As an alternative, we study the approach to aggregate the behavior of component lines in the EPT of the assembly station. The EPT of the assembly system can then be obtained be obtained using the same equations as for the other stations, as presented in Chapter 2. This alternative may be attractive if a main component line can be identified. The measured EPT-parameters clearly show the biggest constraint in the main component flow line. However, the behavior of the feeding component sub-lines now can not be distinguished from the assembly workstation. This means the measured EPTs can be used for bottleneck analysis but not for building an EPT-based aggregate model.

The proposed approach is illustrated in two examples. The first example considers an assembly workstation, fed by multiple component workstations. By using EPT-distributions as a simple representation of the real machine behavior, two types of EPT-based aggregate models have been built. In the first model, transport is modeled according to a triangular distribution, in the second model transport is modeled deterministically. The following conclusions stem from this example. First, the proposed Algorithm 1AMB-t measures the correct EPT-realizations. Second, the EPT-based aggregate models are accurate approximations of the real system. Furthermore, the example shows that, if transport has a low variability, it can be represented by a constant value without great loss of accuracy.

The second example studies an assembly line in the context of an automotive plant. The assembly line consists of one unbuffered assembly workstation and eight ordinary unbuffered workstations with transport systems in-between. Using the proposed method, an EPT-based aggregate model was built. This model approximates both flow time and throughput within 2.5% of their actual values. Based on the EPT-parameters measured by Algorithm 1AMB-t, improvements are suggested for mean EPT t_e and squared coefficient of variation of the EPT c_e^2. It is assumed that the summed value of all t_e's as well as the summed values of c_e^2 can be reduced by 5%. According to the EPT-based aggregate model, the proposed changes will yield a throughput increase of 8.2%. The proposed changes in EPT-parameters are translated into a new line configuration for the assembly line. This new configuration shows an increase in throughput of 8.6%. The EPT-based aggregate model thus accurately predicts the effect on line performance of the proposed changes.

In this chapter, we use discrete-event simulation to evaluate the proposed EPT-based aggregate models and to predict the effect of line configuration changes on line performance. Discrete event simulation models, however, are computationally expensive. A very interesting alternative are analytical queueing approximations for finitely buffered flow lines as for instance developed in Van Vuuren et al. (2005) and the PhD thesis by Van Vuuren (2007). The thesis by Vuuren includes queueing approximations for assembly stations. Such queueing approximations are often computationally much cheaper. It is recommended to investigate the use and applicability of such queueing approximations for assembly lines to obtain computationally cheap and accurate EPT-based aggregate models of assembly lines.

Chapter 5

Lumped Parameter Modeling of the Litho Cell

Lithography is often the bottleneck in a wafer fab. Utilization is typically high, resulting in high WIP levels and large cycle times. To optimize performance, one has to keep capacity losses as well as variability in processing low. Often, insight can be gained from analytical $G/G/m$ queueing models, or from simulation models. The applicability of $G/G/m$ models for lithography stations is limited since they assume that just one lot is processed at a time, while most litho cells process more than one lot at a time. On the other hand, the simulation models that are typically developed incorporate various shop floor details, the quantification of which may be hard and time-consuming.

In this paper, a lumped parameter model is proposed for the litho cell. The model consists of two parts: a detailed representation of the processing inside the track and scanner, and an aggregate representation of the factory floor feeding the loadport. The track-scanner is modeled as a tandem flow line with blocking. The shop floor is represented by a delay distribution that incorporates all contributions outside the machine. Simulation results for both a theoretical example and an industrial case show that the proposed model provides a reasonably simple, yet accurate approximation of the litho cell.

This paper is submitted as:
Kock, Veeger, Etman, Lemmen and Rooda. Lumped parameter modeling of the litho cell. *submitted* 2008

The sections of the paper previously appeared as Kock, Etman, and Rooda (2006), Kock, Veeger, Etman, Lemmen, and Rooda (2007)

5.1 Introduction

The litho cell is an expensive piece of equipment in a wafer fab. A litho cell consists of a track and scanner. The track is used for pre- and postprocessing of wafers, while the scanner is used to expose patterns onto the wafer. The litho cell is often the (designed) bottleneck. Furthermore, the litho cell plays a central role in wafer fabrication. Therefore, optimal configuration and operation of a litho cell is highly desirable.

To facilitate continuous improvement, several performance indicators are in use including throughput δ, mean time between failures t_f, mean time to repair t_r and mean cycle time φ. Another commonly used performance measure is the OEE which quantifies capacity losses at the workstation (Nakajima 1988, SEMI 2000). The OEE is the product of six equipment capacity losses grouped into three categories: availability, efficiency and quality. The OEE classification of capacity losses directly relates to utilization. The OEE does neither cover the contribution of variability in processing to the flow time, nor the loss of throughput due to blocking inside the machine.

Simulation models are also helpful in optimizing the performance. Nayani and Mollaghasemi (1998), Arisha and Young (2004), Mummolo, Mossa, and Digiesi (2004) developed simulation models, with explicit modeling of contributing factors such as machine downs, repairs, operating rules, setups, reticle changes, maintenance, tool changes, operator availability and operator skill. In practice, it may be hard to identify all elements contributing to the processing behavior of a litho cell.

In this paper, we look at the problem of litho cell performance analysis from the viewpoint of availability of data. We observe that on the factory floor, much data is available on the nominal processing behavior of the litho cell (that is, clean process times, recipes, number of wafers in a lot, etcetera). Also, the down and repair times of the machine are known accurately. However, there is much less data available about the impact of external factors on the litho cell such as operator behavior, dispatching rules, maintenance, setups, reticle changes etcetera. Identifying all external factors that affect the performance of the litho cell is often not feasible. The simulation models described by Pierce and Drevna (1992), Nayani and Mollaghasemi (1998), Arisha and Young (2004) include only the internal factors. External factors were not considered, whereas they may contribute significantly to the flow line performance.

We propose a litho cell model based on the above observations. We divide the behavior of the litho cell into two parts: the processing inside the litho cell (which is already well known on the factory floor) and the influences due to the environment (which are hard to quantify). The discrete-event simulation model consists of these two parts.

The inside part describes the processing and availability behavior of the litho cell

in detail. Since this behavior is known on the factory floor, this part is modeled "as is". We use a serial flow line in which the wafers visit the various process steps in the track and scanner.

The environmental part models the influence of the factory floor on the litho cell. This part consists of many disturbances and factors, some of which may be difficult to measure or identify. Therefore, we represent this part using an aggregate distribution which lumps the contributing factors into one single distribution in such a way that the aggregated distribution can be measured from basic events on the factory floor, such as lot arrival times and times when the loadport becomes available to receive a new lot. For this, we use a method similar to the effective process time method described by Jacobs et al. (2001, 2003, 2006) and Chapters 2 to 4.

The paper is organized as follows. In Section 5.2, the litho cell is described in further detail. Then, the effective process time concept is discussed in the context of litho cells in Section 5.3. Next, in Section 5.4, the aggregate model of the litho cell is presented. In Section 5.5, a simulation testcase is described. Finally, in Section 5.6, we present an industrial case study from the Crolles2 wafer fabrication plant. In Section 5.7, the main conclusions are presented.

5.2 Litho cell

A litho cell projects patterns on wafers using a reticle. Typically, 1 to 25 wafers are combined in a lot. A lot is taken by an operator or the AMHS (automated material handling system) from the buffer and clicked onto one of the four loadports (in the remainder of this paper, we will refer to the four parallel loadports of the litho cell as the loadport), after which the wafers are sequentially loaded onto the machine. After loading, the wafer surface is cleaned, coated and baked. Next, the wafer is aligned by the litho cell so that the machine knows the exact position. The desired pattern, presented on a reticle, is exposed onto the coating. Finally, the exposed wafer is developed and hard-baked. In between these process steps, cooling and heating plates may be used to ensure that a wafer has the correct temperature. Typically, process steps are not all equally fast. However, parallel units are provided to ensure that each process step has more or less the same capacity.

The logistics of the waferstream inside the litho cell are relatively straightforward. Wafers are never allowed to overtake one another inside the cell. If a process step contains multiple parallel units, wafers leave it in a FIFO-way. Transport of wafers in the litho cell can be either synchronous or asynchronous. If transport is synchronous, all wafers are transported to the next process step when the slowest wafer has finished processing. If transport is asynchronous, a wafer will be transported to the next process step as soon as the current step is completed and the next step is free.

Not all lots are processed equally fast in the litho cell. This variability is caused by several reasons: some are related to the lot, some to the machine, and others to causes outside the litho cell. Significant factors that influence the required processing time of a lot are the numbers of wafers within a lot, the recipe (i.e., which process steps in the litho cell are required, and the process time per step) and the failure behavior of the litho cell. Data on these three factors can usually be obtained in most semi-conductor fabs.

The influence of other factors is not so easily available. Examples are the availability of operators, the necessity of machine setups, the presence of reticles, and -closely related- the time required to retrieve a reticle from a reticle stocker, or the operator behavior in general. The large number of influences complicate building of a simulation model of the litho cell. Aggregation of contributing factors in a single distribution may be helpful.

5.3 Effective process time concept

The concept of using a distribution that lumps processing, failures and other disturbances together has been discussed in literature (Dallery and Gershwin 1992, Buzacott and Shanthikumar 1993). Hopp and Spearman (2001) use for instance aggregate process time distributions in their factory physics book. They argue that, from a logistical point of view, a workstation does not care whether the delay is caused by setup, down-time, or another reason as long as lots are waiting in the buffer. Therefore, they use the term effective process time in their presentation of queueing relations. Jacobs et al. (2003) propose a method to actually 'measure' EPT-realizations from operational data, such as arrivals and departures of lots at a workstation. Jacobs et al. refer to the EPT as the time during which the lot has consumed production capacity of the workstation. The EPT-distribution is thus obtained without quantifying the individual contributions. This work was extended by Jacobs et al. (2006) and Chapters 2 to 4. So far, the EPT has been considered for infinitely buffered multiple (single-lot or batch) machine workstations and for finitely buffered serial flow lines subject to blocking. Blocking refers to the situation where process capacity of a machine is lost since a finished product can not be sent away due to a full downstream buffer.

The most simple case to compute EPTs is the single server workstation with an infinite buffer (Jacobs et al. 2003). From an EPT point of view, machine capacity is claimed if at least one lot is present in the workstation buffer. The EPT-realization of the i^{th} arriving lot starts upon arrival in an empty workstation or upon departure of the previously processed lot. The EPT-realization lasts until the lot departs from the workstation. Hence, with a_i the i^{th} arrival of a lot in the buffer, d_i the corresponding departure time, and d_{i-1} the departure of the previously processed lot, the i^{th} EPT-realization e_i can then be expressed as

Equation (5.1):

$$e_i = d_i - \max\{a_i, d_{i-1}\}. \tag{5.1}$$

Litho cells are more complicated to analyze from an EPT point of view. A litho cell may contain up to four lots at the same time, however only one lot at a time can start processing. Van der Eerden, Saenger, Walbrick, Niesing, and Schuurhuis (2006) and Lazurko (2005) present a method to apply the EPT to litho cells. They model the litho cell as a single-lot station with an additional delay distribution to account for the multi-lot conveyor-like processing. Due to the aggregation, their model cannot be used for predicting the effect of changes in the litho cell configuration, including changes in throughput rate.

We consider here a different sort of model, with which the effect of changes in the litho cell configuration can be analyzed. The inside of the litho cell is modeled explicitly. The environment is aggregated using an EPT-like method.

5.4 Proposed litho cell model

Next, we explain our aggregate model of the litho cell in further detail. The model assumes that the process times at the various process steps in the machine are available in a database, and that the failure behavior of the litho cell is recorded, expressed by the mean time between failures t_f and the mean time to repair t_r SEMI (2001). The model further assumes that time losses due to the operational environment occur, but that quantification of each of these losses is either not feasible or is rather elaborate.

An aggregate model consisting of two parts is proposed. The first part models the environmental factors between the arrival buffer and the litho cell, such as transport from the stocker to the litho cell, setup, dispatching or preventive maintenance, are lumped into a delay distribution. This explicitly does not include the behavior of the entire fab. The second part models in detail the processing behavior inside the machine. Figure 5.1 visualizes the model.

Model description

The model is presented in terms of communicating parallel processes. The model may be implemented using any discrete-event simulation environment. Each circle in Figure 5.1 represents a process. The arrows represent the communications.

Process S is a start process, which generates new lot arrivals according to a specific distribution. The start process emulates the lot arrivals in the stocker from the automated material handling system (AMHS). The lots are stored in the stocker, in our model the infinitely large buffer B.

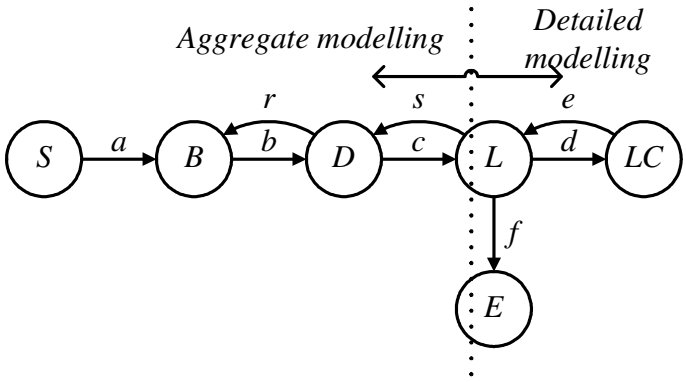


Figure 5.1: Aggregate model of the litho cell

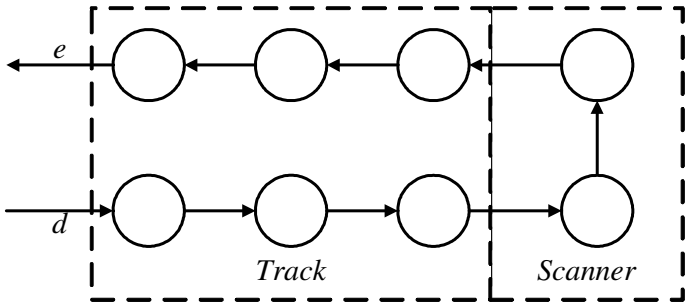


Figure 5.2: Wafer-flow inside litho cell *LC*

After a lot is requested from the stocker, some time elapses before the lot is actually clicked onto one of the four loadports of the litho cell. This elapsed time, caused by the operational environment of the litho cell, is modeled by delay process *D*. Once a lot is requested by the loadport process, delay process *D* waits until at least one lot is present in buffer *B* (stocker). Then, the process delays the lot according to a random variable with an appropriate distribution (chosen by the modeler) with a mean of t_d and coefficient of variation of c_d. After the lot is delayed, the delay process forwards the lot to the loadport process.

Loadport process *L* models the four parallel loadports that are physically present. The process is modeled as a four place (FIFO) buffer. *L* splits the lots into wafers, which are sent wafer by wafer to litho cell *LC*. After they are processed, the wafers are collected again into the same lot by the loadport. In the model, finished lots are assumed to leave loadport process *L* immediately. Any delay in this respect that is observed in practice is included in the delay of subsequent lots.

Litho cell process *LC* represents the wafer-flow inside the litho cell. *LC* is modeled as a single-server serial flow line with blocking, see Figure 5.2. Each process step is modeled individually using a deterministic process time (available from the process database). The busy time between failures and time to repair of the litho cell as a whole are modeled with an appropriate distribution with means t_f and t_r.

Measurement of the delays

The parameters in delay process D are estimated using a method similar to the EPT. This is done using three events. The time at which a lot arrives in buffer B is called the arrival. The i^{th} arrival time of a lot in B is denoted a_i. The i^{th} moment at which loadport L requests a new lot (i.e., the machine is willing to receive a new lot since there are less than four lots on the loadport) is r_i. A new lot is only requested after a slot on the loadport becomes available. Taking the finished lot from the loadport is done by an operator or by the automated material handling system. Delays caused by a finished lot being left on the loadport are thus comparable to delays caused by lots not being moved from the stocker to the loadport. Therefore, we combine these delays into a single distribution: we assume that finished lots are removed immediately. Finally, the i^{th} moment that a lot has actually arrived on loadport L is called l_i.

The delay stops after a lot has arrived on loadport L. The delay starts if i) a lot has arrived in the buffer (so current time $\geq a_i$) *and* ii) the litho cell is waiting for a new lot (i.e., current time $\geq r_i$). Hence, the delay experienced in the buffer for the i^{th} lot processed on the litho cell, which is denoted Δ_{i}, is quantified by Equation (5.2):

$$\Delta_{\text{i}} = l_i - \max(a_i, r_i) \tag{5.2}$$

Once sufficient realizations are obtained, an appropriate distribution can be fitted. From this distribution, we determine the mean delay time t_{d} and the coefficient of variation in the delay c_{d}.

5.5 Simulation example problem of a litho cell

The EPT-based aggregate model is illustrated using a simulation test setup. For the 'real life system', a simulation model is used on which the proposed approach is tested.

Real life model

The accuracy of the aggregate model is investigated using a detailed simulation model as reference. This facilitates investigation of the accuracy of the aggregate model for various utilization levels. An important advantage is the reproducibility of the experimental setup. Process parameters are chosen rather arbitrarily in the simulation model. A case from industrial practice is presented in the next section. The detailed simulation model is used to collect delay-realizations using Equation (5.2). From the obtained realizations, parameters t_{d} and c_{d} are measured. These are used in the proposed aggregate model.

The structure of the real life model corresponds to Figure 5.1. Start S, loadport L, litho cell LC and exit E are the same as in the EPT-based aggregate model. The aggregate part is replaced by a detailed model of setups and operator behavior.

Once the loadport requests a lot, in the real life model, an operator has to take the lot from the stocker, place it on the loadport and perform several actions. The time associated with this is gamma-distributed with mean 200 seconds and CV 0.05. With a probability of 5%, the operator needs to collect a reticle for a lot that is not readily available. In that case, the additional reticle collection time required by the operator is $500 + X$ seconds, where X is exponentially distributed with mean 500 seconds. After the operator has put a new lot on the loadport, he leaves the machine unattended with a probability of 10%. The time during which the operator is unavailable is exponentially distributed with mean 1800 seconds. For easy implementation of the discrete event simulation, in the real life model, the setup and operator are implemented as part of buffer B. Note that the aggregate model aggregates these three disturbances into a single delay distribution.

Start process S produces a new lot with exponential inter-arrival times, with mean t_a. A lot consists of either 15, 20 or 25 wafers (with an equal probability for either possibility). The process time of the wafers on a process step is either 30, 60 or 90 seconds inside the litho-cell (which is multiplied by capacity of the process step). Process S sends the generated lots to the buffer.

The track-scanner part of the machine consists of 8 process steps with capacity for 2, 5, 4, 1, 2, 1, 4 and 6 wafers respectively. The busy time between failures is exponentially distributed with mean 10000 seconds. Upon failure of the machine, the machine has to be repaired. Repair time is exponentially distributed with mean 3000 seconds. After a machine failure, there is a probability of 50% that processing of the wafers is extended with the required repair time. Otherwise, the wafers are immediately taken from the machine and all (partly) processed lots are removed from the loadport. In practice, partly processed lots are sent away for strip-and-clean, after which they return to be reprocessed. Here, it is assumed that lots that are reprocessed are already accounted for in the start process S. Transport of wafers is asynchronous.

Results

The simulation models have been built using the χ-1.0 language (Van Beek, Man, Reniers, Rooda, and Schiffelers 2006). The simulation results shown below have been obtained for a confidence level of 99%. The confidence interval of parameter X_i is given by $0.98X_i \leqslant X_i \leqslant 1.02X_i$.

For the case we considered, the maximal obtainable throughput is $\delta_{\max} = 2.9$ [lots/hr]. The real life model has been compared to the aggregate model for throughput ratios of $\delta/\delta_{\max}$ of 0.65, 0.75, 0.85 and 0.95. Simulation results are

Table 5.1: Lumped parameters and cycle time estimates

δ/δ_{max}	t_d [s]	c_d^2	φ_R [hr]	φ_L [hr]	e
0.65	342.6	3.26	1.4278	1.4270	-0.1%
0.75	349.7	3.33	1.8409	1.8368	-0.2%
0.85	355.7	3.38	2.7595	2.7244	-1.3%
0.95	361.5	3.42	6.7272	6.5208	-3.1%

Table 5.2: cycle time estimates for delay distribution measured at $\delta/\delta_{MAX} = 0.65$

δ/δ_{max}	φ_R [hr]	φ_L [hr]	e
0.65	1.4278	1.4270	-0.1%
0.75	1.8409	1.8182	-1.2%
0.85	2.7595	2.6557	-3.8%
0.95	6.7272	5.9137	-12.1%

reported in Tables 5.1, 5.2 and 5.3. In Table 5.1, we trained the aggregate model at a certain throughput ratio and used it to estimate the cycle time at the same throughput ratio. φ_R refers to the mean cycle time of the real life model, while φ_L refers to the mean cycle time of the aggregate model. The error of the aggregate model, e, is defined as

$$e = \left| \frac{\varphi_R - \varphi_L}{\varphi_R} \cdot 100\% \right|.$$

Table 5.1 shows that the aggregate model closely approximates the real life model, with observed differences in cycle times φ_L and φ_R of 3.1% or less. Furthermore, the values of t_d and c_d show only a small correlation with the utilization level (changing about 5% if the throughput-ratio is increased from 0.65 to 0.95). Our aggregate model assumes that the values of t_d and c_d are throughput-ratio independent. Thus, for this case, the aggregate model can be used as a predictive model at other throughput levels. This observation is reconfirmed by the results of Tables 5.2 and 5.3. In Table 5.2, we use the delay distributions measured at $\delta/\delta_{max} = 0.65$ to predict the cycle time performance for $\delta/\delta_{max} \in \{0.65, 0.75, 0.85, 0.95\}$. In Table 5.3, we use the delay distributions measured at $\delta/\delta_{max} = 0.95$ to predict the cycle time performance again for $\delta/\delta_{max} \in \{0.65, 0.75, 0.85, 0.95\}$. The results in the tables show that the cycle time estimations are good, even if the difference in throughput ratio is large.

Table 5.3: cycle time estimates for delay distribution measured at $\delta/\delta_{MAX} = 0.95$

δ/δ_{max}	φ_R [hr]	φ_L [hr]	e
0.65	1.4278	1.4553	1.9%
0.75	1.8409	1.8568	0.9%
0.85	2.7595	2.7509	-0.3%
0.95	6.7272	6.5433	-2.7%

5.6 Semiconductor manufacturing case

The proposed aggregate modeling approach is illustrated on an operational litho cell at the Crolles2 wafer fab. First, we shortly introduce the Crolles2 wafer fab, where we measured the data for the case. Next, we give an overview of the data we collected and the issues that came up during the processing of the data. The model is validated and then used to compare the cycle time contribution of the environment of the litho cell to the cycle time contribution of the track and scanner itself. Furthermore, a cycle time-throughput curve is generated.

Crolles2 wafer fab

In 2001, the Crolles2 Alliance was formed by Philips Semiconductors (now NXP), STMicroelectronics and Motorola (now Freescale). The alliance built a new 300mm production facility in Crolles, France. Since its start-up, Crolles2 evolved from a mixed R&D-PilotLine to a combined Production-R&D facility in 2007, when we collected the data for our case. About 80% of all processed lots was sold for commercial end-user products, whereas the remaining 20% of the lots was used for research and development programs, as well as engineering or process improvement.

Crolles2 can be characterized as a mid-volume multi-process multi-product Logic fab in which both high volume products as well as small series and prototype products are produced. Standard production lots contain 25 wafers. Lots are processed in several so-called areas: lithography, implant, etch, thermal treatment, metal, dielectrics, chemical mechanical polishing, wet processing, and metrology.

Case data

For the case, we used four independent sources of information. The manufacturing execution software (MES) was used to collect data on the arrivals of lots in the buffer, the moment at which lots enter the loadport and the moment at which lots depart from the machine. Second, the fault detection software (FDC) was used to extract data from the litho cell itself. The FDC data consisted of process-start and process-end data per module in the machine, per recipe. Thirdly, Brooks XSITE^{tm} was used to collect data concerning the uptime-behavior of the litho cell. Finally, we obtained a file with info on recipes, routing and track information from the litho team at the area.

From the obtained data, we extracted the following information:

1. The arrival process of lots in the buffer; the series-sizes in which lots are processed.

2. The recipes that were processed on the machine. The frequency with which a recipe is processed.
3. The delays, such as required for the delay-process described in the previous section.
4. The uptime behavior of the machine (expressed in the mean time between failures t_f and the mean time to repair t_r).
5. The process time per module per recipe.
6. The number of parallel modules.

While we extracted this information from the data, we encountered the following issues:

- **Hold lots:** Some lots are clicked onto the loadports, but are taken off for some reason. In such cases, the lot needs to be cleaned, reworked, or some other action is required. These lots leave the buffer, and return later. In between, the lot is "on hold". Since they are gone for some time (which is not registered as an arrival or departure in the MES), the WIP present in the buffer is estimated incorrectly when a lot is on hold. Hence, the delay realizations cannot correctly be reconstructed with Equation (5.2). Therefore, delays measured while a hold lot was present are not taken into account.
- **Unscheduled machine downs:** If the machine fails while processing lots, the machine may need to be flushed. Lots are removed from the machine and are often sent to wet processing areas. While the machine is down, no new lots are put on the loadports. As a result, delays during a failure are always large. Furthermore, the delay will last the entire repair-period. Therefore, delays measured during a down are not taken into account. From a model point of view, when the machine is down, no lots will be requested from the loadport. Moreover, no new lots are assigned to the litho cell while it is down. Thus, the arrival process is heavily affected by the down behavior of the machine. Therefore, interarrival times that include down time are not taken into account. In return, in the approximation model, the generator is put on hold if the machine is down.

  The times between failures and times to repair are modeled using exponential distributions. One cannot effectively predict when a failure will happen: if one could predict the occurrence of a failure, the responsible component would be replaced during preventive maintenance. Therefore, it is safe to assume that times to repair and times between failure are unpredictable. This corresponds to an exponential distribution.
- **Delay times:** From the measured delays, we could see that the expected delay strongly depends on the state of the loadports. If there are three lots

on the loadports, delays appeared to be much lower than when the loadports are empty. Combining these different delays into a single distribution leads to a wrong estimation of the utilization, and an overestimation of the variability of the delay process. Therefore, the delay process is split into four: a delay for empty loadports, with one lot on the loadports, with two lots or with three lots on the loadports.

In the model, we use a gamma distribution to model the delay process. We use a gamma distribution since it gives only non-negative samples and since nearly any desired (positive) combination of mean and variance can be implemented with a gamma distribution.

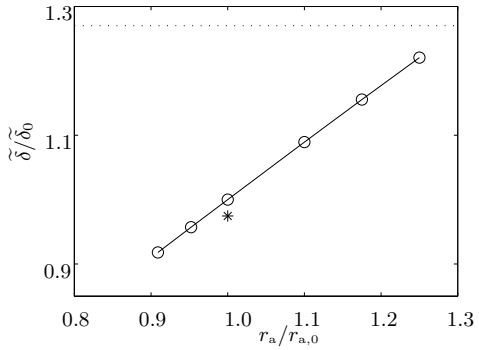
- **Process times within the scanner:** During data collection, we were not able to obtain a complete data set regarding the processing inside the scanner. From our data, and from knowledge on the factory floor, we extracted that the scanner contains five sequential processes, where the actual scan is the bottleneck of the track and scanner. For the bottleneck, we estimated the process time mean and variance based on the interarrival times on the first process step succeeding the scanner.

Model Validation

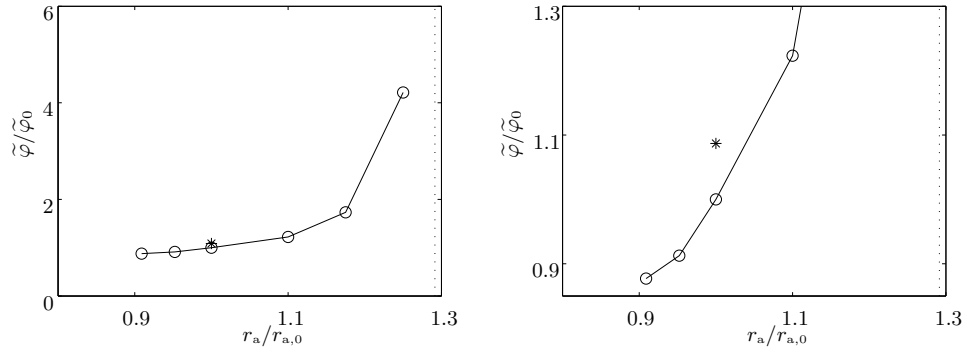
The model is validated using measured throughput and cycle time data. Let $\widetilde{\delta}_0$ and $\widetilde{\varphi}_0$ be the throughput and cycle time estimated by our model. Let δ_0 and φ_0 be the throughput and cycle time observed in our data set. We observed that $\widetilde{\delta}_0 = 1.027\delta_0$ and $\widetilde{\varphi}_0 = 0.919\varphi_0$, i.e. a cycle time error of 8.1%. This error may be due to the following. First, as stated above, hold lots are excluded from the cycle time behavior in the data-set. However, the hold lots do have an impact on the cycle time of the other lots. This effect is included in the data-set, but neglected in the simulation model. Second, upon a machine down, in some cases, the machine is flushed, in other cases, it is not flushed. Although this is included in the model, it may be that the conditions determining wether the litho cell is flushed or not in reality differ from the way they were implemented in the model. Finally, the cycle time behavior of the litho cell strongly depends on the process times on the scanner. Increasing the mean process times by 1% leads to a cycle time estimation error of 8.5% (this change is not significant), while doubling the squared coefficient of variation (which is then still < 0.1) leads to a cycle time underestimation of 1.8% (the error in prediction reduces).

Application of the model

The aggregate model can be used for several purposes. The model can for instance be used to investigate the contribution of the environment of the litho cell to the total cycle time. The delay represents the contribution of the dispatching



(a) Throughput



(b) cycle time

Figure 5.3: cycle time, throughput curves

policy, setup policies and operator behavior on the cycle time of the litho cell. After we removed the delay, we observed that the throughput is approximately the same ($\widetilde{\delta}_{\text{no delay}} = 0.99\widetilde{\delta}_0$, hence within simulation accuracy), while the mean cycle time was reduced by 66% ($\widetilde{\varphi}_{\text{no delay}} = 0.335\widetilde{\varphi}_0$). By removing the delay, we essentially lowered the capacity claimed by lots from the system. In other words, the utilization was lowered. Lowering the utilization leads to a lower cycle time.

By applying different inter-arrival times to the model, a cycle time curve may be estimated. In this curve, we vary the arrival rate between two machine downs. We looked at the effect hereof on throughput and cycle time (expressed in the change in throughput or cycle time relative to the working point found at the validation, i.e. $\widetilde{\delta}/\widetilde{\delta}_0$ and $\widetilde{\varphi}/\widetilde{\varphi}_0$), see e.g. Figure 5.3. Recall that r_a refers to the arrival rate of lots *when the litho cell is not down*. As a result, if the arrival rate is increased, the throughput increases as well, but with a slope of less than 1, as can be seen in Figure 5.3(a): increasing the arrival rate increases the likelihood of a down. The cycle time-throughput curve (left-hand side of Figure 5.3(b)) clearly shows that the cycle time increases nonlinearly with utilization and nicely resembles the well known cycle time curves from the factory physics queueing equation (Hopp and Spearman 2001). The righthand-side of Figure 5.3(b) zooms in at the cycle time curve around the training point. In Figure 5.3, the original training point (respectively $\delta_0/\widetilde{\delta}_0$ and $\varphi_0/\widetilde{\varphi}_0$) is represented by an asterisk.

The model may be used to investigate the effect of changes in the litho cell configuration. As an example, one may consider the effect of changing the internal buffer. Several other applications are possible. The model can be used to investigate what happens if an additional recipe is processed on the litho cell. Also, one may investigate what happens if the product-mix is changed.

5.7 Conclusions and recommendations

In this paper, an aggregate model that predicts cycle time and throughput of an individual litho cell is proposed. This model consists of two parts. In the detailed (litho cell) part, the logistics are modeled as a serial flow line using the process times and system failure characteristics that were obtained from the machine logs. In the aggregate part, external factors which cause lots to be delayed from loading onto the machine were lumped into a single distribution. This distribution was obtained from three events: lot arrivals in the buffer a_i, lot requests from the loadports r_i and lot arrivals on the loadport l_i. Two relevant aggregate distribution parameters were computed: mean t_d and coefficient of variation c_d.

The proposed model has been applied in a simulation test example and an industrial case. In the test example, the proposed approximative model is compared to a complete detailed simulation model of the litho cell, that represents the real life situation. The simulation results show that the aggregate model gives an accurate representation of the real life situation, estimating cycle time and throughput within 3.5% for the case considered. Furthermore, the performance of the model appears to be almost independent of utilization and thus can be used as a means to predict performance changes for the litho cell.

The simulation results of the industrial case show that the aggregate model is a good representation of the real life situation for the case considered: the observed error in the cycle time approximation is about 8%. The model can be used to investigate changes in the configuration of the litho cell. The model also clearly quantifies the contribution of the factory floor to the cycle time performance of the litho cell. This contribution appears to be significant. Most simulation studies only consider the track-scanner module and disregard the 'outside' of the litho cell.

Chapter 6

Aggregate modeling of multi-processing workstations

In this chapter, an aggregate model for manufacturing systems consisting of flow lines with finite buffers and parallel servers is proposed. The proposed model is a multi-server station with process times depending on the work in process (WIP). An algorithm is developed to measure the WIP-dependent process times directly from industrial data such as arrival times at and departure times from the manufacturing system. Simulation results show that the aggregate model accurately predicts the mean flow time.

This chapter is submitted as:
Kock, Etman, Rooda, Adan, Van Vuuren, and Wierman. Aggregate modeling of multi-processing workstations. *submitted* 2008

6.1 Introduction

In semiconductor manufacturing, there is a trend of proliferation of integrated processing (Wood 1996). These integrated processing tools allow multiple wafers of one or more lots to be processed simultaneously. Multiple processes or process steps are contained within a single tool. The logistics inside such integrated tools are often flow line alike. For example, integrated lithography cells allow wafers of up to four lots to be pipelined through a sequence of several processes, including resist coat, expose, and develop. In addition, vacuum processors are integrated around standardized frames that include wafer handlers and load-locks. Other examples of integrated processing tools are wet-benches (lots traverse through a sequence of chemical baths), metal deposit tools (several surface treatment and metal-alloy deposition processes are combined in a single tool) and ion-implant (ion implant consists of two sequential steps: loading and ion emanation onto the wafers).

Due to the sequence of processes that is carried out in an integrated processing tool, the mean flow time φ and throughput δ in the tool increases as the work in process, WIP, increases. The presence of such tools on the factory floor complicates the performance analysis.

For the performance analysis of semiconductor manufacturing there are two categories of models in common use: (discrete-event) simulation models and analytical models. Simulation models allow the inclusion of various details of the processes. However, every detail requires data to be collected and adds to the computational expense of the simulation model. Arisha and Young (2004), Nayani and Mollaghasemi (1998), Pierce and Drevna (1992) develop simulation models of integrated processing tools, with explicit modeling of, e.g., machine downs, repairs, operating rules, setups, maintenance, operator availability and operator skill. The cluster tool model described in Pierce and Drevna uses over 1100 variables and parameters and 500 distributions.

Analytical models, on the other hand, are usually computationally cheap to evaluate and require little input data, such as the mean and variance of process times. However, they adhere to restrictive assumptions, such as, e.g., phase-type distributed process times (Asmussen 2003). An appealing approach to estimate the performance of complex manufacturing systems is to represent (part of) the system by a so-called flow equivalent server (FES) (Norton 1926): an exponential single-server station with service rates depending on the WIP. Indeed, under restrictive assumptions, the aggregate system behavior can be described exactly by a FES, i.e, it is possible to replace part of a queueing network (representing the manufacturing system) by a single-server station without affecting the behavior of the rest of the network (Chandy, Herzog, and Woo 1975, Boucherie 1998). Exact FES models were originally derived for balanced, closed queueing networks with exponential process times. Later, extensions were proposed for special networks with Coxian process times and constant process times (Stewart

and Zeiszler 1980, Thomasian and Nadji 1981, Rhee 2006). However, the assumptions required for an (exact) FES model are too prohibitive to be of practical use in the present context of integrated tools.

In this chapter, we propose an aggregate model that, similar to the FES, replaces the integrated tool by a single- or multi-server station with WIP-dependent processing times. However, unlike the FES, we do not make a priori assumptions regarding process time distributions. Key to our approach is that the process time distributions can be obtained directly from arrival and departure events from the factory floor. The advantage is clear: we do not need to quantify all shop-floor realities individually. To estimate the parameters of the process time distributions we adopt the "Effective Process Time" (EPT) paradigm (Hopp and Spearman 1996, 2001, Jacobs et al. 2001, 2003).

The system, studied in this chapter, is an open network with finite buffers and no feedback; in particular, the configuration is flow-line alike, motivated by the integrated processing tools used in semiconductor manufacturing, which is common for the logistics inside integrated tools. The accuracy of the mean flow time predicted by the aggregate model is investigated for several configurations, ranging from a flow line with twelve sequential servers to a station with twelve parallel servers. Simulation results convincingly demonstrate that the proposed aggregate model yields accurate predictions. Hence, the conclusion is that the modeling framework of multi-server stations with WIP-dependent process times combined with the EPT paradigm provides an effective and powerful tool for the performance evaluation of integrated tools.

The outline of this chapter is as follows: we first present an overview of the effective process time paradigm in Section 6.2. In Section 6.3, we explain the main concept of the aggregate model. We introduce the algorithm to translate arrival and departure data into EPT-realizations in Section 6.3.3. The algorithm is tested on a set of examples in Section 6.4. Finally, in Section 6.5 we present our main findings and the discussion.

6.2 Previous work using the EPT paradigm

The phrase effective process time was originally introduced by Hopp and Spearman (1996, 2001). They define the EPT as 'the time seen by a lot at a workstation from a logistical point of view'. The EPT aggregates the raw processing time and all shop-floor realities and disturbances on processing at a workstation into a single process time distribution. The inclusion of multiple phenomena into a single distribution is referred to as aggregation. Hopp and Spearman give explicit expressions to compute the mean EPT and the EPT coefficient of variation from the raw processing time and the various outages, either preemptive (setup-alike) or non-preemptive (breakdown-alike). They use the EPT mean and variance in closed form approximations for $G/G/m$ queues to explain and estimate the mean

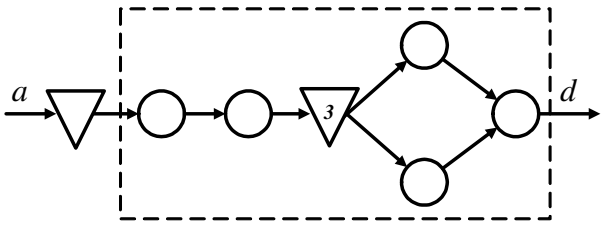


Figure 6.1: Tandem flow line with finite buffers; circles indicate process steps, triangles buffers, a lot arrivals and d lot departures

flow time performance.

In many practical cases, outages may not all be quantifiable. Jacobs et al. (2001, 2003) show that the EPT can be measured without the need to identify and quantify all contributing shop-floor realities. For workstations with ample buffer space and that process a single lot at a time, they present an algorithm to calculate EPT-realizations directly from lot arrival and departure events. The obtained empirical distribution can then be used to fit a parameterized EPT-distribution.

This idea can be generalized into an EPT-based modeling framework, as explained by Kock et al. (2008a). Event collection, EPT calculation, distribution fitting and aggregate modeling are presented as an integrated framework. The EPT is not only used as a performance metric quantifying capacity (mean) and variability (variance), but also to build simulation or analytical models fed by parameter values obtained from empirical EPT-distributions.

EPT-algorithms to compute EPT-realizations from arrival and departure events were proposed by Jacobs et al. (2001, 2003, 2006), Kock et al. (2008c,a), Vijfvinkel et al. (2007), for infinitely buffered 'single lot' workstations, finitely buffered 'single lot' workstations, assembly workstations and batch workstations. These references focus on discrete-event simulation models. Analytical models may be used as an alternative. Closed form expressions for (mean) performance measures of $G/G/m$ queues can be used for infinitely buffered multi-server workstations. For finitely buffered flow lines and assembly lines, queueing approximations as discussed by Dallery and Gershwin (1992), Van Vuuren et al. (2005), Van Vuuren (2007) may be used.

6.3 An aggregate multi-server station

In the present chapter, we consider flow lines consisting of multi-server workstations with finite buffers. Specifically, we assume that, on arrival, lots are put into an infinite buffer to wait until processing starts, and once in process, lots do not recirculate. An example is visualized in Figure 6.1.

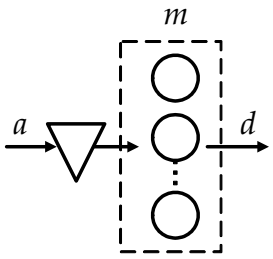


Figure 6.2: Structure of the proposed aggregate model

6.3.1 Model concept

The idea is to aggregate the entire flow line into a multi-server station with FIFO dispatching and WIP-dependent process times; see Figure 6.2. The number of servers, denoted by m, is an important user-defined parameter. Initially, one may expect that the choice of m will be related to the structure of the flow line, i.e., the "degree of parallel processing"; this relation will be investigated in Section 6.4. The process time of a lot depends on the WIP present in the system just before the start of processing. The dependence on the WIP reflects that, in the real system, the mean flow time and throughput depend on the number of lots in the system. Clearly, the real system is *not* a m-server station; hence, the challenge is to subtract the required WIP-dependent process times from the arrival and departure events in the real system. This is explained in the following two sections.

6.3.2 EPT measurement

The input to the calculation of EPT-realizations consists of a chronological list of events obtained from the shop-floor. Each event is defined by the lot *id*, the event type *ev* (arrival in the infinite buffer of the flow line, denoted 'A', or departure from the flow line, denoted 'D') and the time of occurrence of the event τ. Then, by acting as if the event list has been produced by an m-server station, we are able to retrieve the EPT-realizations. Since the process times in the multi-server station are WIP-dependent, we introduce bucket b for each WIP-level b, $1 \leqslant b < \infty$. An EPT-realization is assigned to bucket b if b lots are present at the start of the EPT-realization. Thus, each bucket collects EPT-realizations corresponding to a certain WIP, and at the end of the event list, provides an empirical EPT-distribution. Since the EPT-distributions are expected to converge as b tends to infinity, we can limit the number of buckets by N, say, where bucket N contains all process times registered with a WIP $\geq N$.

Most likely, the real system and the m-server station do not perfectly match. Hence, it may happen that, when lot *id* departs at time τ, it has not yet started processing in the m-server station; this is readily seen to happen when a $G/G/2$

is aggregated into a $G/G/1$, since overtaking takes place in the first, but not in the second system. This inconsistency will be solved as follows. We pick one of the lots in process at time τ, say lot *jd* that started processing at time t when the WIP was b; the pick rule(s) will be specified in the next section. Then we "interchange" the departure times of lot *id* and *jd*; so lot *jd* leaves at time τ, having received an EPT of $\tau - t$ time units for WIP b, after which lot *id* immediately enters service and remains so until the "old" departure time of lot *jd*.

In the next section we describe the algorithm to calculate EPT-realizations in more detail.

6.3.3 EPT-algorithm

The EPT-algorithm is depicted in Figure 6.3. It uses the following variables: n represents the current WIP, list *rs* stores (id, τ, n) containing the start times of the lots that are in process (according to the m-server station). List *ws* contains the *id* of each lot in the system that has not yet started processing (again, according to the m-server station). The algorithm uses the functions `append`, `get`, `remove`, `head`, `tail` and `find` operating on the lists *rs* and *ws*. Function `append` adds an element to the end of the list, `get` reads the element with lot *id* from the list. Function `remove` removes the element with *id* from the list. Function `head` takes the first element in the list and function `tail` takes all elements except the first. Finally, `find` picks one specific element from the list according to a user-defined rule, to be discussed later.

The EPT-algorithm distinguishes five cases:

($a1$) A lot arrives when $n < m$ lots are present. Capacity is available: lot start with *id*, time τ and WIP-level n is added to *rs*.

($a2$) A lot arrives when $n \geqslant m$ lots are present. All m servers are busy, thus the lot is stored in the buffer *ws*.

($d1$) A lot departs, $n < m$ lots remain behind. Bucket b and start time t of the departing lot are retrieved from *rs*, after which the lot is removed.

($d2$) A lot departs, $n \geqslant m$ lots remain behind and *id* of the departing lot is known in *rs*: bucket b and start time t of the lot are retrieved from *rs* after which *id* is removed from *rs*; the first lot waiting in *ws* is added as new lot start to *rs* with time τ and WIP-level n.

($d3$) A lot departs, $n \geqslant m$ lots remain behind, and *id* of the departing lot is not known in *rs*. So lot *id* departs, while it has not started processing according to the m-server station. Then, using function `find`, we select an alternative lot that has started already, *jd*. We compute the EPT-realization using the

```
n:= 0; rs:=[]; ws := []
loop
    read id, ev, τ
    if ev = 'A' then
        n := n + 1
        if n ⩽ m then                                   (a1)
            rs:= append(rs, (id, τ, n))
        elseif n ⩾ m then                               (a2)
            ws:= append(ws, id)
        endif
    elseif ev = 'D' then
        n := n − 1
        if n < m then                                   (d1)
            (t, b):= get(rs, id)
            rs:= remove(rs, id)
        elseif n ⩾ m and id ∈ rs then                   (d2)
            (t, b):= get(rs, id)
            rs:= remove(rs, id)
            jd:= head(ws); ws:= tail(ws)
            rs:= append(rs, (jd, τ, n))
        elseif n ⩾ m and id ∉ rs then                   (d3)
            (jd, t, b):= find(rs,rule)
            rs:= remove(rs, jd)
            rs:= append(rs, (jd, τ, n))
            ws:= remove(ws, id)
        endif
        write τ − t, b
    endif
endloop
```

Figure 6.3: EPT-ALGORITHM

start time of *jd*. Then, lot *jd* is restarted and lot *id* is removed from buffer *ws*.

Note that in $(d3)$ lot *id* immediately departs and lot *jd* (re)starts service, instead of the other way around; the reason is that, although the lot identity is not relevant for the EPT-realization, we should be able to connect the right lot to the departure of lot *jd* after time τ.

For function `find` in case $(d3)$, we propose three rules: 1) random lot, 2) lot with the shortest elapsed process time, 3) lot with the longest elapsed process time. The rationale behind rule 2 is that the lot might be a fast mover, and therefore, we assign the smallest possible process time; the rationale behind rule 3 is opposite. Clearly, for $m = 1$, the pick rules are identical, since then there is only one lot to pick. The impact of the choice of the pick rule on the performance predictions

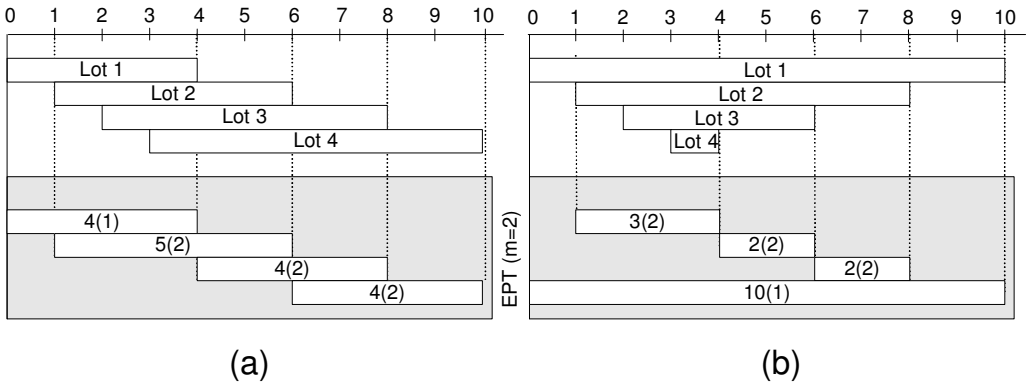


Figure 6.4: EXAMPLE GANTT-CHARTS, (A) WITHOUT OVERTAKING, (B) WITH OVERTAKING, USING RULE 2

will be investigated in Section 6.4.

In case $(d1)$, $(d2)$ and $(d3)$, the EPT-realization is printed as $\tau - t$ with bucket b.

6.3.4 Gantt-chart examples

Figure 6.4 shows Gantt-charts for two manufacturing systems; Figure 6.4(a) corresponds to a system without overtaking, and Figure 6.4(b) to a system with overtaking. The bottom part of the Gantt-charts shows the EPT-realizations computed by the EPT-algorithm, with $m = N = 2$; EPTs are labeled $t(b)$, where t is the duration of the EPT and b the bucket. Note that case $(d3)$ is invoked twice in Figure 6.4(b), but not in Figure 6.4(b).

6.4 Model validation

By means of discrete-event simulation we will test the aggregate model in four scenarios depicted in Figure 6.5; all simulation results are generated using the $\chi - 0.8$ software (Hofkamp and Rooda 2002).

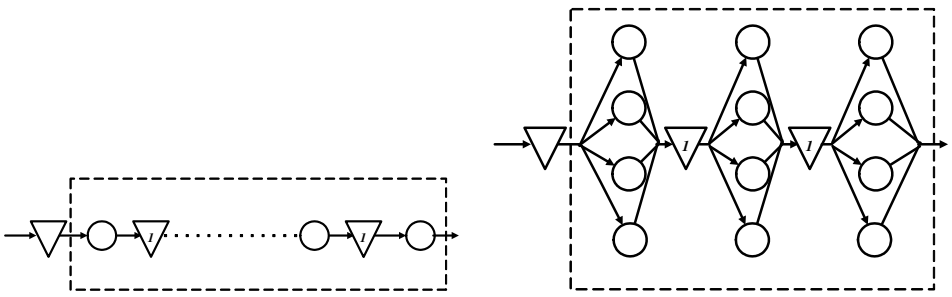
In each example, the arrival process is Poisson with rate δ and the process times on workstations are gamma-distributed with mean 1.0 and squared coefficient of variation $c^2 \in \{0.1, 1.0, 2.0\}$. Mean flow time predictions in the real system are based on simulation runs of 2.000.000 lots. The utilization of the system is defined as the ratio of the throughput δ and the maximum attainable throughput $\delta_{\max}$, which is determined in one simulation run of 100.000 lots using unlimited supply of lots. For each scenario EPT-realizations are measured using the EPT-algorithm (Figure 6.3) in a simulation run of 2.000.000 lots at a given utilization level, the so-called training level. For scenario I, the training level is $\delta/\delta_{\max} \in \{0.6, 0.9\}$ while for scenarios II, III and IV, we take $\delta/\delta_{\max} = 0.8$. On

the empirical EPT-distributions, we fit Gamma distributions matching the mean t_e and coefficient of variation c_e^2. Then mean flow times are predicted by the multi-server station with WIP-dependent Gamma-distributed process times at utilization levels $0.3 \leqslant \delta/\delta_{max} \leqslant 0.95$; at each utilization level the mean flow time prediction is based on five runs of 10.000.000 lots.

6.4.1 Scenario I: Twelve sequential single server workstations

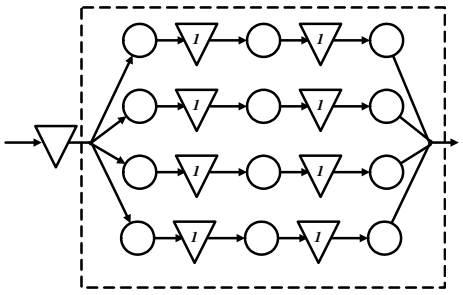
The system consists of a flow line of twelve sequential single-server workstations, see Figure 6.5(a). Each workstation has one buffer space. For this system, we have $\delta_{max} = \{0.875, 0.553, 0.440\}$ [lots/hour] for $c^2 = \{0.1, 1.0, 2.0\}$.

In Figures 6.6 we present EPT-realizations measured for $\delta/\delta_{max} = 0.9$, $c^2 = 1.0$ and $m = 1$. The x-axis in Figure 6.6(a) is the WIP (or bucket), whereas the y-z planes represent histograms of the EPT-realizations. Clearly, the bulk of the EPT-realizations is in buckets ranging from 1 to 40, with a peak near 20. The empirical probability distribution function (PDF) is plotted in Figure 6.6(b). From bucket 30, say, onwards, the distributions do not significantly change; buckets 40 or higher hardly contain any realization explaining the noisy behavior. Hence, it makes sense to aggregate all realizations in buckets ≥ 30 into bucket $N = 30$.

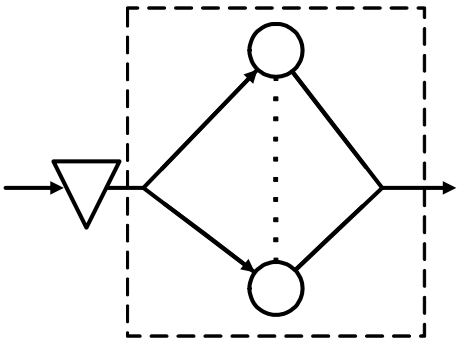


(a) SCENARIO 1: 12 SEQUENTIAL SINGLE-SERVER WORKSTATIONS

(b) SCENARIO 2: 3 WORKSTATIONS WITH 4 SERVERS



(c) SCENARIO 3: 4 PARALLEL LINES



(d) SCENARIO 4: 12 PARALLEL SERVERS

Figure 6.5: TEST SCENARIOS FOR ALGORITHM OF FIGURE 6.3

Figure 6.7 plots the mean EPT t_e and squared coefficient of variation (SCV) c_e^2 as a function of the WIP-level. Clearly, these plots depend on the squared coefficient of variation c^2 of the processing times in the real system.

The monotonic behavior of t_e as a function of WIP-level is as expected: the higher the WIP in the flow line, the faster lots will leave the line. Also the behavior of c_e^2 may be explained: initially, at low WIP, c_e^2 tends to increase, due to the (random) distribution of the WIP in the flow line, and eventually, c_e^2 will converge to a value close to c^2. We would like to point out that monotonicity properties of t_e and c_e^2, as observed in Figure 6.7, may be exploited in an analytical model to accomplish, e.g., state space reduction.

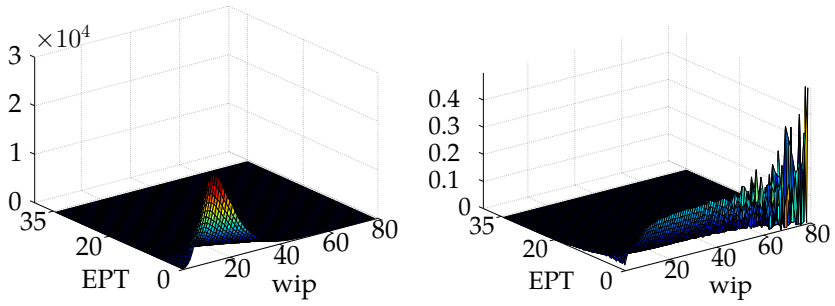
Figures 6.8(a) and 6.8(b) show, for various values of m, mean flow time predictions of the aggregate model trained at utilization level $\delta/\delta_{max} = 0.6$ and 0.9, respectively; the EPT-realizations are obtained by employing pick rule 1. The figure shows that, for $m = 1$, the best prediction is obtained at the training level (as expected). For $m = 1$, mean flow time predictions are also listed in Tables 6.1 and 6.2. From the results we can conclude that mean flow times at low utilization levels are more accurately predicted by the aggregate model trained at $\delta/\delta_{max} = 0.6$ than the one trained at $\delta/\delta_{max} = 0.9$, whereas the reverse is true for high utilizations. Further, the predictions seem to be more accurate for smaller values of c^2.

A naive approach is to approximate the flow line by an $M/G/1$ queue; in the present context, this means that the flow line is aggregated into a 1-server station with $N = 1$, i.e., all EPT-realizations are assigned to one bucket. This approach would produce poor approximations, since it completely fails to take into account the increased efficiency of the integrated processing tool for larger WIP-levels.

Table 6.1: SCENARIO I: MEAN FLOW TIME PREDICTION ($m = 1$, TRAINED AT $\delta/\delta_{MAX} = 0.6$)

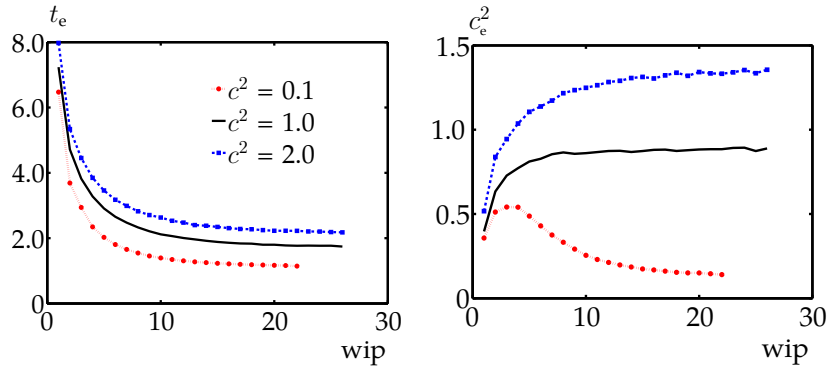
	$c^2 = 0.1$		$c^2 = 1.0$		$c^2 = 2.0$	
$\frac{\delta}{\delta_{max}}$	Approx.	Real	Approx.	Real	Approx.	Real
0.3	12.02	12.85	14.11	14.67	15.40	15.77
0.5	13.39	13.79	17.16	17.18	19.44	19.29
0.6	14.49	14.49	18.86	18.85	21.62	21.62
0.7	16.23	15.51	20.94	21.06	24.19	24.69
0.85	21.75	18.58	25.22	27.26	29.19	33.54
0.95	69.62	27.07	30.36	48.94	33.38	67.10

The aggregate 1-server station may be slightly refined by exploiting the following observation. There are two possibilities to start processing at WIP-level 1: either a lot arrives in an empty flow line, or the previous departure left behind a single lot. The mean EPT of a lot entering an empty flow line is 12, whereas the mean EPT of a single lot left behind is clearly less (in fact, 6 according to simulation). Thus, splitting bucket 1 in two buckets may improve the predictions. Figure 6.9 shows mean flow time predictions for $c^2 = 1.0$ with training level $\delta/\delta_{max} = 0.6$.



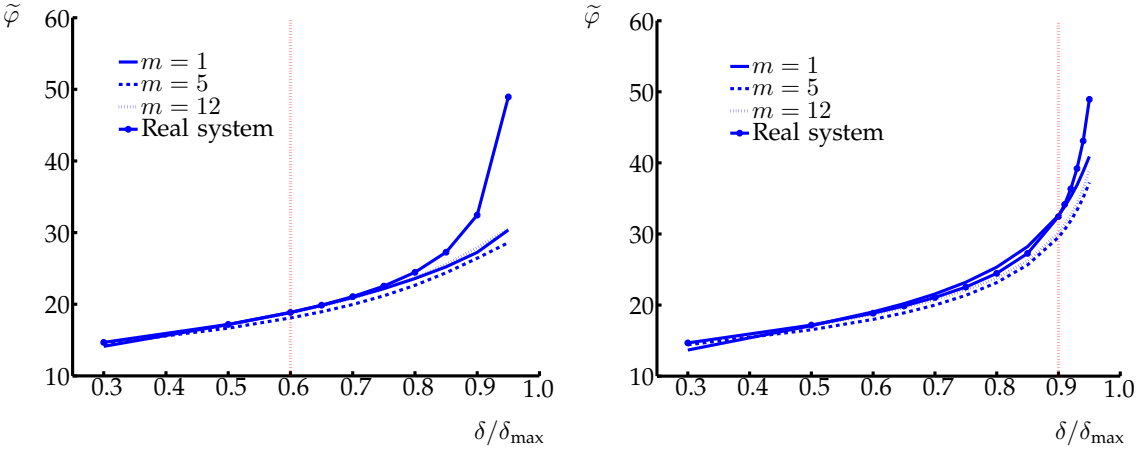
(a) HISTOGRAM (b) EMPIRICAL PDF

Figure 6.6: SCENARIO I: EPT-REALIZATIONS ($\delta/\delta_{MAX} = 0.9$, $c^2 = 1.0$ AND $m = 1$)



(a) t_E PER BUCKET (b) c_E^2 PER BUCKET

Figure 6.7: SCENARIO I: MEAN AND SCV OF EPT ($\delta/\delta_{MAX} = 0.9$, $c^2 = 1.0$ AND $m = 1$)



(a) TRAINED AT $\delta/\delta_{MAX} = 0.6$ (b) TRAINED AT $\delta/\delta_{MAX} = 0.9$

Figure 6.8: SCENARIO I: FLOW TIME PREDICTION ($c^2 = 1.0$, RULE 1)

Since the prediction only slightly improves for low δ/δ_{max}, we will not further pursue the option of splitting of buckets.

Next we investigate sensitivity with respect to the number of EPT-measurements. Figure 6.10 shows that, if the number of EPT-realizations is drastically reduced from 2.000.000 to 15.000 lots, the mean flow time predictions are still accurate.

Table 6.2: SCENARIO I: MEAN FLOW TIME $\widetilde{\varphi}$ ESTIMATION FOR $m = 1$ IF THE MODEL IS TRAINED AT $\delta/\delta_{\text{MAX}} = 0.9$

	$c^2 = 0.1$		$c^2 = 1.0$		$c^2 = 2.0$	
$\frac{\delta}{\delta_{\max}}$	Approx.	Real	Approx.	Real	Approx.	Real
0.3	11.35	12.85	13.66	14.67	15.34	15.77
0.6	13.10	14.49	19.04	18.85	22.82	21.62
0.85	18.13	18.58	28.22	27.26	35.32	33.54
0.9	21.08	20.95	32.60	32.45	41.26	41.18
0.92	22.46	22.55	35.27	36.35	44.84	47.10
0.95	26.14	27.07	40.91	48.94	52.42	67.10

This suggests that it is not necessary to collect an "enormous" amount of data, which is convenient from a practical point of view.

Finally we consider an unbalanced flow line: the processing speed of server 6 is is slowed down by a factor 1.5, and thus it becomes the bottleneck station. Mean flow time predictions for utilization levels from 0.3 until 0.95 are depicted in Figure 6.11. For $m = 1$, the predictions are even slightly more accurate than in the balanced case.

6.4.2 Scenario II: Three workstations, four parallel servers each

The first workstation in the three station flow line of Figure 6.5(b) has an infinite buffer, the other two have one buffer place. The maximum obtainable throughput is $\delta_{\max} = \{3.666, 3.174, 2.989\}$ [lots/hour] for $c^2 = \{0.1, 1.0, 2.0\}$. The training level is $\delta/\delta_{\max} = 0.8$.

In Figures 6.12 we show t_e and c_e^2 as a function of the WIP-level, for $m = 1$ and $m = 4$. As expected, the shape of the t_e and c_e^2 curves depend on the choice of m; in particular, the limiting value of t_e for $m = 4$ is (roughly) four times the limiting value for $m = 1$.

Figure 6.13 presents mean flow time predictions in the range of $0.3 \leqslant \delta/\delta_{\max} \leqslant 0.95$. It shows that the predictions for $m = 12$ are accurate at low utilizations, but underestimate the mean flow time at high utilizations; a possible explanation is that the 12-server station allows for more overtaking than in the real system. The predictions for $m = 1$ and $m = 4$ are very accurate in the utilization range $0.6 \leqslant \delta/\delta_{\max} \leqslant 0.9$. In this case, one might initially guess that $m = 4$ would be the best choice, since it properly reflects the "degree of parallel processing"; but, surprisingly, the predictions for $m = 1$ are of the same quality.

Table 6.3 gives additional results for $m = 4$, demonstrating the effect of the pick rule. The estimates for the three rules are fairly close, but seem to be ordered: rule 3 gives the lowest prediction, rule 2 the highest and rule 1 is in between.

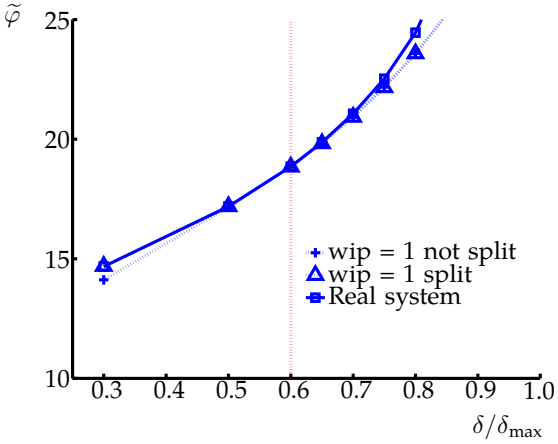


Figure 6.9: SCENARIO I: EFFECT OF SPLITTING BUCKET 1 ($c^2 = 1.0$, $m = 1$, TRAINED AT $\delta/\delta_{MAX} = 0.6$)

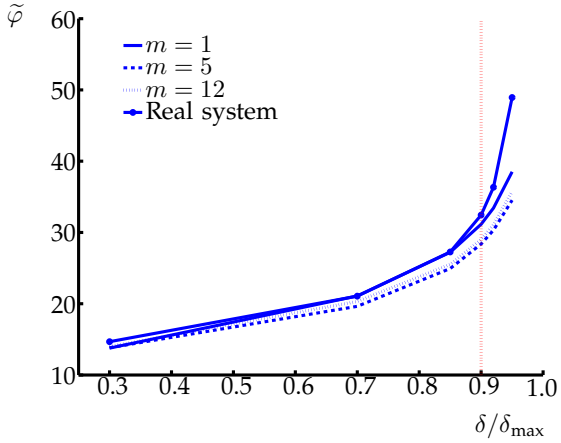


Figure 6.10: SCENARIO I: FLOW TIME PREDICTION, 15.000 LOTS ($c^2 = 1.0$, RULE 1, TRAINED AT $\delta/\delta_{MAX} = 0.9$)

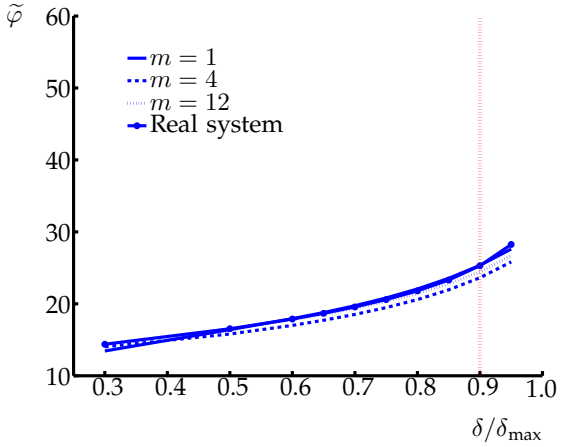
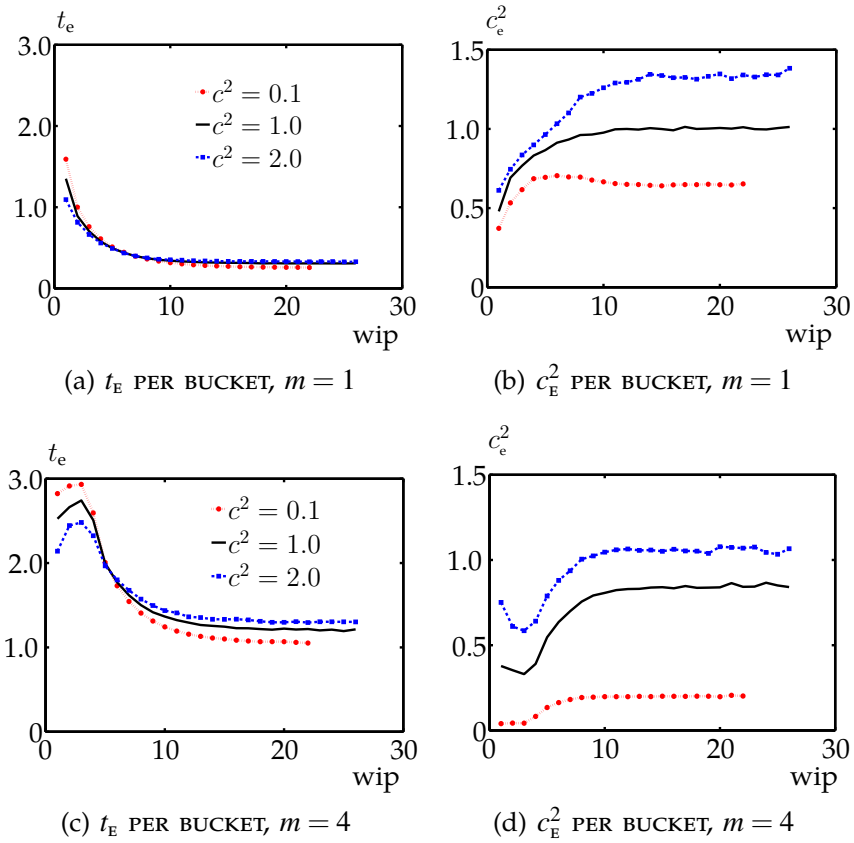


Figure 6.11: SCENARIO I: FLOW TIME PREDICTION, UNBALANCE 1.5 ($c^2 = 1.0$, RULE 1, TRAINED AT $\delta/\delta_{MAX} = 0.9$)

This ordering is also reflected in the c_e^2curves in Figure 6.14, which seems to be a direct consequence of the pick rule.

Table 6.3: SCENARIO II: MEAN FLOW TIME PREDICTION ($m = 4$, TRAINED AT $\delta/\delta_{MAX} = 0.8$)

	$c^2 = 0.1$				$c^2 = 2.0$			
$\frac{\delta}{\delta_{max}}$	rule 1	rule 2	rule 3	Real	rule 1	rule 2	rule 3	Real
0.3	2.90	2.90	2.89	3.02	2.50	2.85	2.29	3.03
0.6	3.08	3.08	3.08	3.18	3.12	3.36	3.00	3.33
0.7	3.23	3.23	3.23	3.32	3.51	3.73	3.38	3.62
0.8	3.50	3.50	3.50	3.58	4.12	4.39	3.95	4.18
0.9	4.14	4.15	4.13	4.29	5.46	5.92	5.09	5.68
0.95	4.91	4.94	4.89	5.63	7.28	7.84	6.57	8.04



(a) t_E PER BUCKET, $m = 1$ (b) c_E^2 PER BUCKET, $m = 1$

(c) t_E PER BUCKET, $m = 4$ (d) c_E^2 PER BUCKET, $m = 4$

Figure 6.12: SCENARIO II: EFFECTIVE PROCESS TIMES PER BUCKET ($\delta/\delta_{MAX} = 0.8$, $c^2 = 1.0$, RULE 1)

6.4.3 Scenario III: Four parallel lines of three sequential, single server workstations

We now consider a system of four parallel single-server flow lines, with three workstations per line, see Figure 6.5(c). Each workstation has one buffer space, except for the first stations in the lines sharing an infinite buffer. For this system, the maximum obtainable throughput is $\delta_{max} = \{3.659, 2.691, 2.319\}$ [lots/hour] for $c^2 = \{0.1, 1.0, 2.0\}$. The training level is $\delta/\delta_{max} = 0.8$.

Figure 6.15 shows the mean flow time prediction for $0.3 \leqslant \delta/\delta_{max} \leqslant 0.95$; additional results for $m = 4$ and each of the pick rules are displayed in Table 6.4. The results for scenario III are comparable to ones for scenario II. Note, however, at high utilizations the prediction errors in scenario III are larger than in scenario II (cf. Figure 6.15 and Figure 6.13). Apparently, in scenario II, the aggregate model more accurately captures interaction between lots.

Finally, we note that the picture of mean flow times, obtained by slowing down one of the four lines by a factor 1.5, is similar to Figure 6.11.

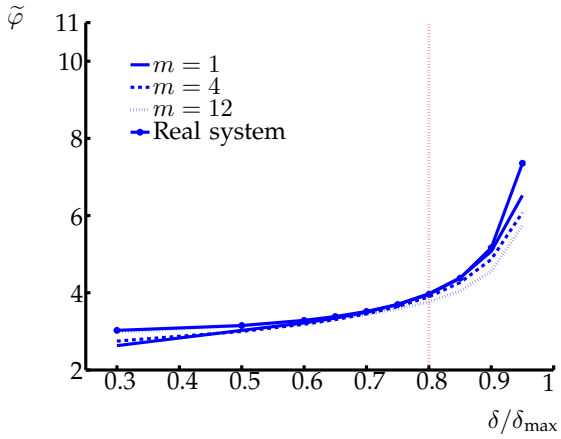
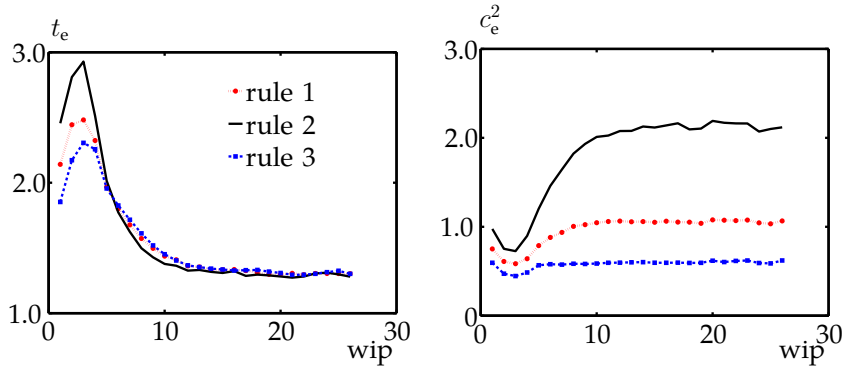


Figure 6.13: SCENARIO II: FLOW TIME PREDICTION ($c^2 = 1.0$, RULE 1, TRAINED AT $\delta/\delta_{\max} = 0.8$)



(a) t_E PER BUCKET, $m = 2$ (b) c_E^2 PER BUCKET, $m = 2$

Figure 6.14: SCENARIO II: EFFECTIVE PROCESS TIMES PER PICK RULE ($\delta/\delta_{\max} = 0.8$, $c^2 = 2.0$, $m = 2$)

Table 6.4: SCENARIO III: MEAN FLOW TIME PREDICTION ($m = 4$, TRAINED AT $\delta/\delta_{\max} = 0.8$)

	$c^2 = 0.1$				$c^2 = 2.0$			
$\frac{\delta}{\delta_{\max}}$	rule 1	rule 2	rule 3	Real	rule 1	rule 2	rule 3	Real
0.3	3.58	3.60	3.55	3.84	4.20	4.67	3.97	5.23
0.6	4.14	4.15	4.13	4.27	5.57	5.83	5.46	5.96
0.7	4.37	4.38	4.36	4.42	6.11	6.34	6.00	6.28
0.8	4.72	4.74	4.72	4.70	6.83	7.11	6.66	6.92
0.9	5.51	5.54	5.49	5.44	8.16	8.66	7.81	8.86
0.95	6.62	6.66	6.59	6.84	9.60	10.39	9.09	12.75

6.4.4 Scenario IV: One workstation with twelve parallel servers

To conclude, we consider a single workstation with twelve parallel servers, see Figure 6.5(d). For this system, the maximum obtainable throughput is $\delta_{\max} = \{12, 12, 12\}$ [lots/hour] for $c^2 = \{0.1, 1.0, 2.0\}$. The training level is again set at $\delta/\delta_{\max} = 0.8$.

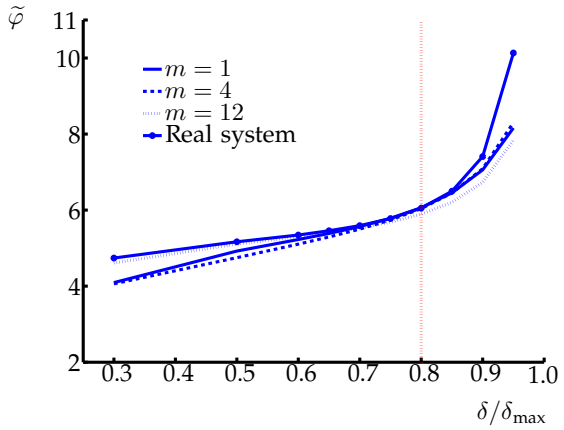
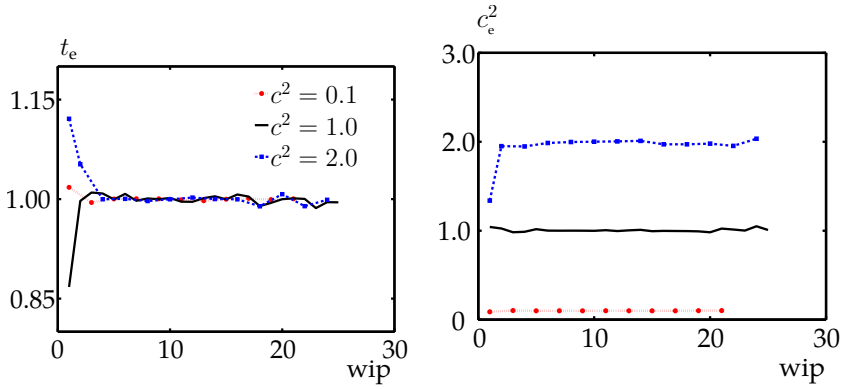


Figure 6.15: SCENARIO III: FLOW TIME PREDICTION ($c^2 = 1.0$, RULE 1, TRAINED AT $\delta/\delta_{\max} = 0.8$)



(a) t_E PER BUCKET, $m = 12$ (b) c_E^2 PER BUCKET, $m = 12$

Figure 6.16: SCENARIO IV: EFFECTIVE PROCESS TIMES PER BUCKET ($\delta/\delta_{\max} = 0.8$)

Figure 6.16 shows t_e and c_e^2 as a function of the WIP-level for $m = 12$. Clearly, the measurements in buckets smaller than 6 or larger than 15 experience noise (due to few observations): one would expect flat curves here.

Figure 6.17 shows mean flow time predictions for $0.3 \leqslant \delta/\delta_{\max} \leqslant 0.95$. The figure also depicts the standard $M/G/12$ approximation, i.e., $m = 12$ and $N = 1$. Obviously, now this "naive" approximation is very accurate, and the $M/G/12$ with "WIP-dependent" process times is almost as accurate. Further, the predictions for $m = 1$ are less accurate at low utilization and the ones for $m = 20$ are less accurate at high utilization.

In Scenario II we already touched the issue of selecting the pick rule; see Table 6.3, demonstrating that the effect of the pick rule on the mean flow time prediction is limited. However, this choice may be relevant in situations where the rule is often invoked. For example, this is expected to happen if the 12-server station is aggregated as a 2-server station; the predicted mean flow time, as a function of $\delta/\delta_{\max}$, is depicted in Figure 6.18, and indeed, the accuracy now strongly depends on the pick rule. In all examples, however, it appeared that rule 1, i.e., the random rule, performed well and thus, this rule seems to be a safe

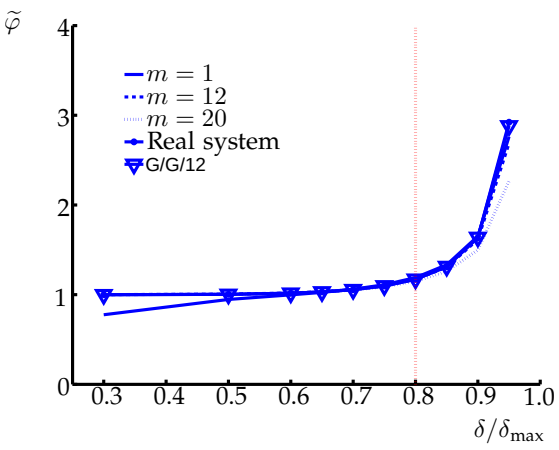


Figure 6.17: SCENARIO IV: FLOW TIME PREDICTION ($c^2 = 1.0$, RULE 1, TRAINED AT $\delta/\delta_{\max} = 0.8$)

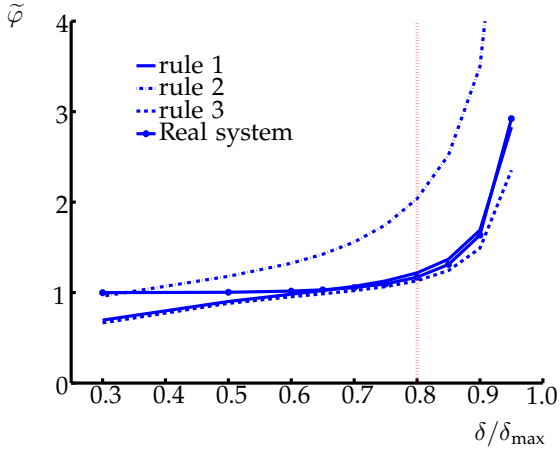


Figure 6.18: SCENARIO IV: FLOW TIME PREDICTION ($c^2 = 1.0$, $m = 2$, MEASURED AT $\delta/\delta_{\max} = 0.8$)

choice. Moreover, the numerical experiments in this chapter convincingly show that the aggregate model with $m = 1$ always produces accurate mean flow time predictions, and in this case, the pick rule is irrelevant.

Finally, we consider an unbalanced case by slowing down the processing speed of six of the twelve servers by a factor 1.5, while keeping $c^2 = 1.0$ for all processing times. Evaluating mean flow time predictions for $m \in \{1, 2, 4, 12\}$ over the range $0.3 \leqslant \delta/\delta_{\max} \leqslant 0.95$ leads to similar results as shown in Figure 6.17. However, in this case, the standard $M/G/12$ approximation is inaccurate: it sometimes overestimates the mean flow time by more than 10%, while the $M/G/12$ approximation with WIP-dependent process times remains accurate.

6.5 Conclusions and discussion

In this chapter, we propose an aggregate m-server model with WIP-dependent process times. The process times are computed from lot arrivals at and lot departures from the system that is aggregated. An advantage is that these events can be directly measured from the factory floor. An algorithm is presented to calculate the WIP-dependent effective process time realizations.

The accuracy of the mean flow time prediction has been investigated in four scenarios, ranging from a flow line to a single workstation with parallel servers. The results show that predictions are accurate, but the quality depends on the choice of m, and to a lesser degree, on the pick rule; surprisingly, the choice $m = 1$ appears to be good across all scenarios. The feature of WIP-dependent process times appears to be crucial: the quality of mean flow time predictions by multi-server stations with WIP-*in*dependent process times is usually poor. The overall conclusion is that the aggregate 1-server station always performs well (and, in

this case, the choice of the pick rule is not relevant). The simulation study in this chapter is restricted to flow lines consisting of multi-server workstations with finite buffers; we expect, however, that the scope of this approach goes (far) beyond this class of manufacturing systems.

The aggregate model has been developed keeping integrated processing equipment in mind. A follow-up chapter by Veeger, Etman, van Herk, and Rooda (2008) demonstrates how the present methodology can be applied to workstations with integrated processing tools in a semiconductor manufacturing environment, where commonly used $G/G/m$ approximations perform unsatisfactorily.

Chapter 7

Conclusions and Recommendations

The thesis presents four contributions in the framework of the Effective Process Time for performance analysis of discrete event manufacturing systems. This chapter reviews the main conclusions and gives an outlook towards possible future research.

7.1 Conclusions

7.1.1 Finitely buffered workstations

In Chapters 2 and 3, an EPT-approach for flow lines consisting of finitely buffered, (multi-)server workstations is developed. It is shown that, for such systems, the EPT of the workstations can be determined from three manufacturing events: (i) the arrival of a lot in (the buffer of) the workstation; (ii) the moment in time at which processing of the lot is finished; and (iii) the departure of the lot from the workstation. These events can be translated into EPT-realizations with a sample path equation: the EPT-realization starts when a lot is present and the server is idle (departure of the previously processed lot has occurred), and the EPT-realization ends at the finish of processing.

For infinitely buffered workstations, usually the first two moments of the measured EPT workstation distributions suffice. In Chapters 2 and 3, we show that, if blocking plays a major role in the system, then the shape of the EPT-distribution needs to be represented more accurately. This happens when buffer sizes are

small or zero, variability is high, and only few (or just one) parallel servers are present at a workstation. From simulation test cases, it is concluded that using the offset as a third distribution parameter leads to increased accuracy of the approximations.

7.1.2 Assembly workstations

Chapter 4 addresses assembly workstations that are subject to blocking. The chapter contributes a method to measure EPT-realizations for assembly systems. The method isolates the behavior of all component lines that feed the assembly workstation. In this way, the EPT-realization of the assembly station is not affected by the performance of the component lines. As an alternative, the methods of Chapters 2 and 3 may be used by aggregating the behavior of the component lines in the EPT of the assembly station. This alternative may be attractive if one of the components to be assembled can be seen as main component or if one of the component lines dominates the behavior of the whole assembly line. However, such an aggregation of the complete line prohibits the use of the EPT-model for performance prediction. The measured EPT-distributions can only be used for performance quantification. For performance prediction, the newly developed EPT for assembly workstations should be used.

The proposed approach is illustrated in two examples: a theoretical example and an industrial case. The first example illustrates the case that transport can be modeled as a constant. The second example illustrates that the EPT-based model accurately predicts the effect on line performance of changes that are made in the line configuration.

7.1.3 Litho cell model

In Chapter 5 an aggregate model that predicts flow time and throughput of an individual litho cell is proposed. This model consists of two parts. In the detailed (litho cell) part, the logistics inside the track and scanner are modeled as a serial flow line using the process times and system failure characteristics that were known from the machine logs. In the aggregate part, external factors which cause lots to be delayed from loading onto the machine were lumped into a single delay distribution in a way similar to calculating EPTs.

The proposed model has been applied in an industrial case. The simulation results of the industrial case show that the aggregate model is an accurate representation of the real life situation for the case considered, the flow time is underestimated by 8%. It is shown that 34% of the flow time is due to the operation of the machine. Also, a flow time-throughput curve is plotted for the litho cell. With the model we can investigate changes in the configuration of the litho

cell, such as the size of the internal buffer, the capacity of the bottleneck process step in the litho cell, or the product mix.

7.1.4 Integrated manufacturing systems

Chapter 6 proposes an aggregate $G/G/m$ model with WIP-dependent process times to model integrated manufacturing systems. Process times are sampled from buckets, where the bucket in this thesis corresponds to the WIP present in the system at the start of processing of a lot according to the aggregate model.

In four test scenarios, it is shown that the aggregate model provides accurate flow time predictions in a region around the training point $\delta/\delta_{\max}$ at which the EPT-realizations were determined. The accuracy of the model depends on the choice of the number of machines m in the aggregate model and the pick rule that is used in the EPT-algorithm of Figure 6.3: in case a lot leaves the system while, from the perspective of the aggregate model the lot is still in process. Then, the EPT-algorithm picks the EPT-start time of one of the lots currently in process. We considered: (1) pick random, (2) pick the lot with the shortest elapsed process time, (3) pick the lot with the longest elapsed process time. Flow times approximated with pick rule 2 are always larger than those approximated with pick rule 1, while pick rule 3 always estimates the smallest flow times.

In this thesis, the $G/G/m$ model with WIP-dependent process times is a simulation model. From a mathematical perspective, it is interesting to investigate whether an analytical counterpart within the new model class can be developed.

7.2 Recommendations

7.2.1 $G/G/m$ station with WIP-dependent process times

In this thesis, the $G/G/m$ station with WIP-dependent process times has only been tested for finitely buffered flow lines without feedback. We expect the aggregate model of Chapter 6 to be more widely applicable. The concept can probably also be used to approximate infinitely buffered flow lines, reentrant flow lines and job-shop like systems. It is recommended to investigate this opportunity. A second topic is to investigate whether the aggregate modeling method can be extended such that also the flow time distribution can be predicted. The current method considers only the mean flow time. For customer reliability, also the distribution of the flow time is of importance. For single-lot machine workstations, Blom (2007) considered flow time prediction using the EPT-approach. A topic related to this is the modeling of the dispatching rule. On the factory floor, dispatching rules define the sequence of lots processed on the machine at

the workstation. One may want to explicitly model the dispatching rule in the aggregate model.

7.2.2 Networks

Currently, the EPT studies flow lines, in particular flow lines with finite buffers. Recall that the goal of the STW project is to arrive at simple yet accurate models of manufacturing networks, both simulation models and analytical models. Bierbooms (2008) has proposed a first method to implement EPT in open networks of queues.

The new aggregate modeling approach proposed in Chapter 6 may be used to simplify the models of entire networks. As an example, one may choose to build a (detailed) model for the bottleneck workstations and use the aggregate approach of Chapter 6 to model the remaining part(s) of the network. Alternatively, the entire manufacturing system may be aggregated, the approach proposed in Chapter 6 may be used to model the system at different levels of abstraction.

7.2.3 Optimization

Optimization using EPT-based aggregate models is an interesting next step to investigate. Two types of EPT-based models may be used: simulation models and analytical models. The simulation model makes the optimization computationally expensive, whereas the analytical model may not be able to cover particular distribution details such as offset or product-type dependent parameters. A promising approach is to use an optimization method where analytical models and simulation models are used together. Such a hybrid optimization method was investigated by Vijfvinkel (2005). It is recommended to continue this research.

References

I. Adan. *Lecture notes Stochastical Process Design*. 2001. [cited at p. 4, 33]

I.J.B.F. Adan and J. van der Wal. *Monotonicity of the throughput in single server production and assembly networks with respect to the buffer sizes*, pages 345–356. 1989. [cited at p. 16, 36]

A. Arisha and P. Young. Intelligent simulation-based lot scheduling of photolithography toolsets in a wafer fabrication facility. In *2004 Winter Simulation Conference*, pages 1935–1942, 2004. [cited at p. 66, 80]

S. Asmussen. *Applied Probability and Queues*. Springer, New York, 2nd edition, 2003. [cited at p. 80]

T. Baines, S. Mason, and P.O. Siebers. Humans: the missing link in manufacturing simulation? *Simulation modeling: practice and theory*, 12:515–526, 2003. [cited at p. 3, 30, 31]

P.P. van Bakel. Effective process times for batch machines. Master's thesis, Eindhoven University of Technology, Department of Mechanical Engineering, Systems Engineering Group, 2001. SE 420284. [cited at p. 5]

J. Banks. Introduction to simulation. In P.A. Farrington, H.N. Nembhard, D.T. Sturrock, and G.W. Evans, editors, *Proceedings of the 1998 Winter Simulation Conference*, pages 7–13, 1999. [cited at p. 3, 13, 30, 31, 48]

D.A. van Beek, K.L. Man, M.A. Reniers, J.E. Rooda, and R.R.H. Schiffelers. Syntax and consistent equation semantics of hybrid chi. *Journal of Logic and Algebraic Programming*, 68:129–210, 2006. [cited at p. 72]

R. Bierbooms. Analysis of open networks of queues. Master's thesis, Tilburg University, 2008. [cited at p. 100]

E.L.W. Blom. EPT and flowtime distributions. Master's thesis, Eindhoven University of Technology, Department of Mechanical Engineering, Systems Engineering Group, December 2007. SE 420513. [cited at p. 41, 99]

R.J. Boucherie. Norton's equivalent for queueing networks comprised of quasi-reversible components linked by state-dependent routing. *Performance Evaluation*, 32:83–99, 1998. [cited at p. 80]

J.A. Buzacott and J.G. Shanthikumar. *Stochastic Models of Manufacturing Systems*. Prentice Hall, Englewood Cliffs, New Jersey, 1st edition, 1993. [cited at p. 3, 5, 13, 16, 17, 30, 31, 36, 48, 50, 68]

K.M. Chandy, U. Herzog, and L. Woo. Parametric analysis of queuing networks. *IBM Journal of Research and Development*, 19:36–42, 1975. [cited at p. 80]

L. Chen and C-L. Chen. A fast simulation approach for tandem queueing series. In O. Balci, R.P. Sadowski, and R.E. Nance, editors, *Proceedings of the 1990 Winter Simulation Conference*, pages 539–546, 1990. [cited at p. 5, 13, 30, 48, 50]

R. Christensen. *Data Distributions: A Statistical Handbook*. Entropy Limited, Lincoln, Massachusetts, second edition, 1989. ISBN 0-938-87651-1. [cited at p. 38]

S.A. Coenen. Analysis of the retrofit manufacuring system using measured data. Internship SE-420407, Eindhoven University of Technology, Systems Engineering Group, 2004. [cited at p. 5]

G.L. Curry, B.A. Peters, and M. Lee. Queueing network model for a class of material-handling systems. *International Journal of Production Research*, 41:3901–3920, 2003. [cited at p. 17]

Y. Dallery and S.B. Gershwin. Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems: Theory and Applications*, 12: 3–94, 1992. [cited at p. 3, 5, 13, 30, 31, 48, 50, 68, 82]

A.C. Diamantidis, C.T. Papadopoulos, and C. Heavey. Approximate analysis of serial flow lines with multiple parallel-machine stations. *IIE Transactions*, 39 (4):361–375, 2007. [cited at p. 13, 48]

J. van der Eerden, T. Saenger, W. Walbrick, H. Niesing, and R. Schuurhuis. Litho area cycle time reduction in an advanced semiconductor manufacturing line. In *ASMC 2006*, Boston, 2006. [cited at p. 68]

J.W. Fowler and O. Rose. Grand challenges in modeling and simulation of complex manufacturing systems. *Simulation*, 80:469–476, 2004. [cited at p. 3, 4]

S.B. Gershwin. *Manufacturing Systems Engineering*. Englewood Cliffs, NJ: Prentice-Hall, 1994. [cited at p. 3, 13, 48]

A.T. Hofkamp and J.E. Rooda. χ *Reference manual*. Systems Engineering Group, Eindhoven University of Technology, November 2002. URL:http://se.wtb.tue.nl/. [cited at p. 42, 57, 86]

W.J. Hopp and M.L. Spearman. *Factory Physics: Foundations of Manufacturing Management*. London: Irwin McGraw-Hill, 1st edition, 1996. [cited at p. vii, 15, 30, 81, 109]

W.J. Hopp and M.L. Spearman. *Factory Physics: Foundations of Manufacturing Management*. London: Irwin McGraw-Hill, 2nd edition, 2001. ISBN 0-256-24795-1. [cited at p. vii, 3, 4, 12, 15, 20, 30, 32, 33, 49, 62, 68, 77, 81, 109]

W.J. Hopp, M.L. Spearman, S. Chayet, K. Donohue, and E.S. Gel. Using an optimized queueing network model to support wafer fab design. *IIE Transactions*, 34:119–130, 2002. [cited at p. 13]

S.-J. Hsieh. Hybrid analytic and simulation models for assembly line design and production planning. *Simulation modeling: practice and theory*, 10(1):87–108, 2002. [cited at p. 5, 13, 48]

J.H. Jacobs. *Performace quantification and simulation optimization of manufacturing flow lines*. PhD thesis, Eindhoven University of Technology, department of Mechanical Engineeering, 2004. [cited at p. 8, 9]

J.H. Jacobs, L.F.P. Etman, E.J.J. van Campen, and J.E. Rooda. Quantifying operational time variability: the missing parameter for cycle time reduction. In *2001 IEEE/SEMI Advanced semiconductor manufacturing conference*, pages 1–10, 2001. [cited at p. vii, 5, 13, 30, 32, 35, 49, 51, 67, 81, 82, 109]

J.H. Jacobs, L.F.P. Etman, E.J.J. van Campen, and J.E. Rooda. Characterization of operational time variability using effective process time. *IEEE Transactions on Semiconductor Manufacturing*, 16:511–520, 2003. [cited at p. vii, 5, 13, 14, 15, 16, 30, 32, 35, 37, 44, 49, 51, 52, 62, 67, 68, 81, 82, 109]

J.H. Jacobs, P.P. van Bakel, L.F.P. Etman, and J.E. Rooda. Quantifying variability of batching equipment using effective process times. *IEEE Transactions on Semiconductor Manufacturing*, 19(2):269–275, 2006. [cited at p. 5, 16, 49, 51, 52, 67, 68, 82]

K.C. Jeong and Y.D. Kim. An approximation method for performance analysis of assembly/disassembly systems with parallel-machine stations. *IIE Transactions*, 31:391–394, 1999. [cited at p. 13]

D.S. Kim and J.M. Alden. Estimating the distribution and variance of time to produce a fixed lot size given deterministic processing times and random downtimes. *International Journal of Production Research*, 35:3405–3414, 1997. [cited at p. 17, 48]

J. Kleijnen and W. van Groenendaal. *Simulation: a statistical perspective*. John Wiley and Sons, Chichester, 1st edition, 1992. ISBN 0 471 93055 5. [cited at p. 3]

A.A.A. Kock. Performance evaluation and simulation meta modeling of single server flow lines subject to blocking: an effective process time approach. Master's thesis, Eindhoven University of Technology, Department of Mechanical Engineering, Systems Engineering Group, December 2003. SE 420367-A. [cited at p. 5, 19]

A.A.A. Kock, L.F.P. Etman, and J.E. Rooda. Lumped parameter modeling of the litho cell. In A. Dolgui, G. Morel, and C.E. Pereira, editors, *INCOM06*, volume 2, pages 709–714, St. Etienne, France, 2006. [cited at p. 65]

A.A.A. Kock, C.P.L. Veeger, L.F.P. Etman, B. Lemmen, and J.E. Rooda. Cycle time and throughput performance analysis of a litho cell using an aggregate modeling approach. In *ASMC 2007*, Stresa, Italy, 2007b. [cited at p. 65]

A.A.A. Kock, L.F.P. Etman, and J.E. Rooda. Effective process time for multi-server flowlines with finite buffers. *IIE Transactions*, 40(3):177–186, 2008a. [cited at p. 8, 29, 82]

A.A.A. Kock, L.F.P. Etman, J.E. Rooda, I.J.B.F. Adan, M. van Vuuren, and A. Wierman. Aggregate modeling of integrated manufacturing systems. *submitted*, 2008b. [cited at p. 8, 79]

A.A.A. Kock, F.J.J. Wullems, L.F.P. Etman, I.J.B.F. Adan, F. Nijsse, and J.E. Rooda. Performance evaluation and lumped parameter modeling of single server flowlines subject to blocking: an effective process time approach. *Computers and Industrial Engineering*, 54:866–878, 2008c. [cited at p. 8, 11, 82]

A.A.A. Kock, C.P.L. Veeger, L.F.P. Etman, B. Lemmen and J.E. Rooda. Lumped parameter modeling of the litho cell. *submitted*, 2008d. [cited at p. 8, 65]

A.M. Law and W.D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill Higher Education, Boston, 3rd edition, 2000. [cited at p. 3, 13, 30, 48]

M. Lazurko. A χ model of the litho-cell as a basis for testing multiple lot machine EPT algorithms. Master's thesis, Eindhoven University of Technology, Department of Mechanical Engineering, Systems Engineering Group, June 2005. SE 420443. [cited at p. 68]

J. Li. Throughput analysis in automotive paint shops: a case study. *IEEE Transactions on Automation Science and Engineering*, 1(1):90–98, 2004. [cited at p. 48]

J. Li, J.M. Alden, and J.R. Rabaey. Approximating feeder line reliability statistics with partial data collection in assembly systems. *Computers and Industrial Engineering*, 48(2):181–203, 2005. [cited at p. 13, 48]

J. MacGregor Smith. $M/G/c/k$ performance models. In *Fifth International Conference on "Analysis of Manufacturing Systems-Production Managament"*, pages 177–184, 2005. [cited at p. 3, 31]

P.R. McMullen and G.V. Frazier. Using simulation and data envelopment analysis to compare assembly line balancing solutions. *Journal of productivity analysis*, 11(2):149–168, 1998. [cited at p. 5, 13, 48]

A.R. Mendes, A.L. Ramos, A.S. Simaria, and P.M. Vilarinho. Combining heuristic procedures and simulation models for balancing a pc camera assembly line. *Computers and Industrial Engineering*, 49(3):413–431, 2005. [cited at p. 5, 13, 48]

G. Mummolo, G. Mossa, and S. Digiesi. Learning and tiredness phenomena in manual operations performed in lean automated manufacturing systems: a reference model. In *Proceedings of the international IMS (Intelligent Manufacturing Systems) Forum 2004, Cernobbio (CO), Italy*, 2004. [cited at p. 66]

S. Nakajima. *Introduction to TPM: Total Productive Maintenance*, volume 1. Cambridge Productivity Press, Cambridge, Massachusetts, 1988. ISBN 0-915299-23-2. [cited at p. 2, 66]

M. Nayani and M. Mollaghasemi. Validation and verification of the simulation model of a photolithography process in semiconductor manufacturing. In *1998 Winter Simulation Conference*, pages 1017–1022, 1998. [cited at p. 66, 80]

E.L. Norton. Design of finite networks for uniform frequency characteristic. taken from http://www.ece.rice.edu/~dhj/norton/ (last visited 11-12-2007), 1926. [cited at p. 80]

T. Osogami and M. Harchol-Balter. Closed form solutions for mapping general distributions to quasi-minimal PH distributions. *Performance Evaluation*, 63(6): 524–552, 2006. [cited at p. 40]

N.G. Pierce. Golden nuggets of amhs modeling and design for semiconductor wafer fabrication. In *Advanced Semiconductor Manufacturing Conference and Workshop, ASMC, 1994*, pages 200–204. IEEE/SEMI, 1994. [cited at p. 5, 13, 48]

N.G. Pierce and M.J. Drevna. Development of generic simulation models to evaluate wafer fabrication cluster tools. In *Advanced Semiconductor Manufacturing Conference and Workshop, ASMC, 1992*, pages 874–878. IEEE/SEMI, 1992. [cited at p. 66, 80]

Y. Rhee. Some notes on the reduction of network dimensionality in nested open queueing networks. *European Journal of Operational Research*, 174:124–131, 2006. [cited at p. 80]

A.J. de Ron and J.E. Rooda. Equipment effectiveness: OEE revisited. *IEEE Transactions on Semiconductor Manufacturing*, 18(1):190–196, 2005. [cited at p. 2, 33]

J.E. Rooda and J. Vervoort. *Analysis of Manufacturing Systems using χ 1.0*. Eindhoven University of Technology, source: http://se.wtb.tue.nl/education/mis, 2007. [cited at p. 1, 2]

M. Rooney. Effective process time characterization for workstations with unequal machines, exceptional first lots and idling. Master's thesis, Eindhoven University of Technology, Department of Mechanical Engineering, Systems Engineering Group, July 2002. SE 420306. [cited at p. 5]

M.D. Rossetti and G.M. Clark. Estimating operation times from machine center arrival and departure events. *Computers and Industrial Engineering*, 33:493–514, 2003. [cited at p. 5, 14]

I. Sabuncuoglu, E. Erel, and A.G. de Kok. Analysis of assembly systems for interdeparture time variability and throughput. *Computers and Industrial Engineering*, 34(1):23–40, 2002. [cited at p. 17]

H. Sakasegawa. An approximation formula $l_q = \alpha\beta^\rho/(1-\rho)$. *Annals for the Institute for Statistics Mathematics*, 29:67–75, 1977. [cited at p. 3]

SEMI. Standard for definition and measurement of equipment productivity. Technical Report SEMI E79-0200, Sematech, 2000. Originally published in 1999. [cited at p. 2, 33, 48, 66]

SEMI. Specification for definition and measurement of equipment reliability, availability and maintainability (ram). Technical Report SEMI E10-0301, Sematech, 2001. Originally published in 1986. [cited at p. 2, 69]

J.G. Shanthikumar, S. Ding, and M.T. Zhang. Queueing theory for semiconductor manufacturing systems: a survey and open problems. *IEEE Transactions on Automation Science and Engineering*, 4(4):513–522, 2007. [cited at p. 3, 4, 49]

W.J. Stewart and G.A. Zeiszler. On the existence of composite flow equivalent markovian servers. *ACM Sigmetrics Performance Evaluation Review*, 9(2):105–116, 1980. [cited at p. 80]

A. Thomasian and B. Nadji. Aggregation of stations in queueing network models of multiprogrammed computers. *ACM Sigmetrics Performance Evaluation Review*, 10(3):86–104, 1981. [cited at p. 80]

T. Tolio, S.B. Gershwin, and A. Matta. Analysis of two-machine lines with multiple failure modes. *IIE Transactions*, 34(1):51–62, 2002. [cited at p. 34]

C.P.L. Veeger, L.F.P. Etman, J. van Herk, and J.E. Rooda. Generating cycle time-throughput curves using EPT-based aggregate modeling. In *2008 IEEE/SEMI Advanced Semiconductor Manufacturing Concerence (ASMC)*, Boston, 2008. [cited at p. 96]

M. Vijfvinkel. The effective process time for model-based optimization of assembly lines with finite buffers. Master's thesis, Eindhoven University of Technology, Department of Mechanical Engineering, Systems Engineering Group, December 2005. SE 420449. [cited at p. 100]

M. Vijfvinkel, A.A.A. Kock, L.F.P. Etman, M. van Vuuren, and J.E. Rooda. Performance measurement and prediction of finitely buffered asynchronous assembly lines: an effective process time approach. *submitted*, 2007. [cited at p. 8, 47, 82]

M. van Vuuren. Performance analysis of multi-server tandem queues with finite buffers. Master's thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, 2003. [cited at p. 41]

M. van Vuuren. *Performance Analysis of Manufacturing Systems: Queueing Approximations and Algorithms*. PhD thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, 2007. [cited at p. vii, 3, 6, 7, 8, 9, 13, 33, 40, 44, 48, 63, 82, 110]

M. van Vuuren and I.J.B.F. Adan. Performance analysis of assembly systems. In *Proceedings of the Markov Anniversary Meeting 2006*, pages 89–200, 2005a. [cited at p. 33]

M. van Vuuren and I.J.B.F. Adan. Performance analysis of tandem queues with small buffers. In *Fifth International Conference on "Analysis of Manufacturing Systems-Production Managament"*, pages 127–135, 2005b. [cited at p. 31, 33, 40, 42, 44]

M. van Vuuren, I.J.B.F. Adan, and S.A.E. Resing-Sassen. Performance analysis of multi-server tandem queues with finite buffers and blocking. *OR Spektrum*, 27:315–339, 2005. [cited at p. 31, 63, 82]

W. Whitt. Approximations for the $GI/G/m$ queue. *Production and Operations Management*, 2(2):114–161, 1993. [cited at p. 3]

S.C. Wood. Simple performance models for integrated processing tools. *IEEE Transactions on Semiconductor Manufacturing*, 9:320–328, 1996. [cited at p. 80]

F.J.J. Wullems. Data collection for simulation of flow lines with blocking; the role of machine failure in throughput loss. Master's thesis, Eindhoven University of Technology, Department of Mechanical Engineering, Systems Engineering Group, November 2002. SE 420316. [cited at p. 5]

Samenvatting

Moderne productiesystemen worden steeds ingewikkelder. Het is vaak moeilijk om de invloed van veranderingen op de verblijftijd- en doorzetprestatie te voorspellen. Wachtrijmodellen kunnen hierbij behulpzaam zijn.

Twee soorten wachtrijmodellen worden onderscheiden: analytische modellen en simulatiemodellen. Analytische modellen kunnen snel worden doorgerekend en behoeven weinig invoer; het ontwerpen ervan behoeft echter specialistische kennis, waarbij strikte aannamen worden gemaakt. Simulatiemodellen zijn flexibeler en kunnen gebruikt worden om ieder gewenst detail te modelleren; het doorrekenen van zulk een model kost echter veel rekentijd, en er is veel data nodig om de details van de fabrieksvloer te beschrijven.

Dit proefschrift stelt een methode van aggregeren voor om het aantal details in analytische modellen of simulatiemodellen te verkleinen. Door aggregatie wordt een werkstation weergegeven middels één effectieve-procestijdverdeling, die tijdverliezen als gevolg van omstellen, machinefalen en beschikbaarheid van operators omvat. Een fundamenteel aspect van de voorgestelde methode is dat de geaggregeerde procestijdverdeling rechtstreeks gemeten kan worden uit data van de fabrieksvloer, zoals aankomsttijden en vertrektijden van halfproducten bij een werkstation. De individuele tijdsverliesfactoren hoeven hierbij niet gekwantificeerd te worden. Deze aankomst- en vertrekdata kan verkregen worden uit de programmeerbare logica controllers (PLCs), die veel gebruikt worden in de machinebesturing van productiesystemen.

In het proefschrift wordt de voorgestelde methode van aggregeren aangeduid met de Effectieve-ProcesTijd (EPT). De term effective-procestijd werd geïntroduceerd door Hopp and Spearman (1996, 2001) als 'de procestijd gezien door een lot op een workstation vanuit een logistiek oogpunt'. Jacobs et al. (2001, 2003) hebben aangetoond dat de effectieve-procestijd gemeten kan worden zonder de individuele tijdsverliezen te kennen. Door EPTs te meten hebben zij de variatiecoëfficiënt van enkele single-lot machines in een halfgeleider-fabriek

weten te bepalen. Dit tweede moment van de procestijdverdeling is nodig voor (analytische) wachtrijmodellen van productiesystemen. Van Vuuren (2007) presenteert analytische wachtrijmodellen voor eindig gebufferde productielijnen en assemblage-stations waarin de eerste twee momenten van de EPT-verdelingen worden gebruikt.

Het EPT-modelleerraamwerk voor de prestatie-analyse van productiesystemen wordt in dit proefschrift verder uitgebreid. Het proefschrift presenteert methodes voor het meten van de EPT voor eindig gebufferde werkstations en assemblage-stations. EPT-realisaties worden berekend met sample-pad vergelijkingen op basis van drie tijdstippen: de aankomsttijd van lots, de vertrektijd van lots en het tijdstip waarop het bewerken van een lot ten einde komt. Het gemiddelde, de variantie, en mogelijk hogere momenten van de gemeten EPT-realisaties kunnen gebruikt worden als parameters van de EPT-verdeling. Daarnaast kan een distributiefunctie gefit worden op de verzamelde realisaties. De voorgestelde EPT-methode is getoetst in twee cases vanuit de industrie: uit de automobielindustrie en uit de gloeilampenindustrie. De EPT-modellen geven nauwkeurige benaderingen voor zowel de verblijftijd als de doorzet.

Het proefschrift laat zien dat het EPT-concept ook gebruikt kan worden om slechts een deel van het werkstation te aggregeren. Een model van een lithografie machine wordt gepresenteerd, waarbij de litho-cel in detail wordt gemodelleerd terwijl de machine omgeving geaggregeerd gemodelleerd wordt. Meestal is over het inwendige van de litho-cel veel proces-data beschikbaar, terwijl over de omgeving (het laden) weinig bekend is. Het ontwikkelde model wordt geïllustreerd middels een simulatie case en een industriële case. Beide cases laten zien dat de aggregaat modellen nauwkeurige verblijftijdvoorspellingen opleveren, en dat ze gebruikt kunnen worden om het effect van veranderingen in de machine-configuratie te voorspellen. In de industrie-case blijkt dat tweederde van de verblijftijd wordt veroorzaakt door de omgeving van de machine, terwijl slechts eenderde wordt veroorzaakt door de litho-cel zelf. Daarnaast is met het model een verblijftijd-doorzet curve berekend.

Tenslotte wordt in het proefschrift een aggregatiemodel voorgesteld voor stations bestaande uit machines die een stroom van lots tegelijkertijd in process kunnen hebben. Zulke machines worden veel gebruikt in de halfgeleiderindustrie. Het voorgestelde aggregaatmodel is een $G/G/m$-type wachtrij model, waarvan de procestijden afhangen van het aantal klanten in het systeem. Op vier verschillende productielijn-scenarios worden simulatie-experimenten uitgevoerd. Volgens de simulatie-resultaten biedt het voorgestelde model nauwkeurige verblijftijdsbenaderingen. Het voorgestelde model is nauwkeuriger dan de normale $G/G/m$ benaderingen met WIP-onafhankelijke procestijden.

Het onderzoek, beschreven in dit proefschrift, werd uitgevoerd als deel van het STW project EPT. Het project is een samenwerking van de Systems Engineering group (faculteit Werktuigbouwkunde) en de Stochastic Operations Research group (faculteit Wiskunde en Informatica).

Curriculum vitae

1980	Born in Geleen, the Netherlands on March 24
1992-1998	Secondary School, Atheneum at Serviam Scholengemeeschap, Sittard
1998-2002	Bachelor of Science in Mechanical Engineering, Eindhoven University of Technology, Eindhoven; awarded 'Great appreciation'
2002-2003	Master of Science in Mechanical Engineering, Eindhoven University of Technology, Eindhoven; awarded 'Cum Laude', awarded 'Corus jong talentprijs voor de materiaalkunde en werktuigbouwkunde 2004' and nominated for the Mignot award 2003-2004 (elected as best Master's thesis of the department of Mechanical Engineering)
2004-now	PhD on Effective Process Time, Eindhoven University of Technology, Eindhoven