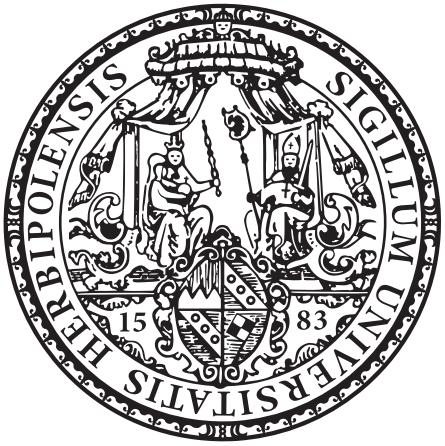
Julius-Maximilians-Universität Würzburg
Wirtschaftswissenschaftliche Fakultät



Capacity Planning and Control with Advanced Information

Inauguraldissertation
zur Erlangung des akademischen Grades
doctor rerum politicarum (Dr. rer. pol.)

vorgelegt von
M.Sc. Julian Frederick Kurz
geboren in Freiburg im Breisgau

Name und Anschrift:	Julian Frederick Kurz Salamanderplatz 10 70806 Kornwestheim
Erstgutachter:	Prof. Dr. Richard Pibernik
Zweitgutachter:	Prof. Dr. Christoph Flath
Datum der Einreichung:	28. Januar 2017

Acknowledgements

First and foremost, I would like to thank my doctoral advisor, Prof. Dr. Richard Pibernik. I am deeply grateful for his persistent support. He continuously challenged and enhanced my work and there are no words suitable to express how instrumental Prof. Pibernik has been in my development as a researcher. I enjoy looking back to our countless debates about research-related as well as personal subjects.

Secondly, I would like to thank Prof. Dr. Christoph Flath for serving as my second advisor and for his support during the development of this thesis.

I thank Prof. Jan van Mieghem for the opportunity to spend one quarter at the Kellogg School of Management. His lecture, our discussions and the contacts that emerged from this stay have been crucial for the success of my research journey. In addition, I would like to thank Prof. Itai Gurvich for our discussions regarding the properties of stochastic processes. My special thanks go to Prof. Kuang Xu for his feedback and support but especially, for providing the inspiration for my work on queueing with future information.

Additionally, I thank all colleagues at the Chair of Logistics and Quantitative Methods in Business Administration for the interesting discussions and legendary Wednesday nights. Fabian, thank you for your substantial support over the entire term of the PRACTICE project.

This dissertation was financially supported by the EU (FP7/ 2007-2013), grant agreement no. 609611 (PRACTICE). I thank McKinsey & Company for the financial support and for providing an exceptional work infrastructure.

Finally, I would like to thank my family and my girlfriend Johanna. I am deeply grateful for their continuing support of my professional and personal development. You always encouraged me to focus on my work, but also cared about my general well-being in a thoughtful and loving way.

Contents

Deutschsprachige Zusammenfassung (Summary in German Language)

Diese Dissertation besteht aus drei inhaltlich abgeschlossenen Teilen, die jedoch ein übergeordnetes Thema zur Grundlage haben: Wie können Daten über zukünftige Bedarfe zur Kapazitätsplanung und -steuerung genutzt werden? Im Rahmen von Industrie 4.0 werden zunehmend Daten erzeugt und für prädikative Analysen genutzt. Zum Beispiel werden Flugzeugtriebwerke mit Sensoren ausgestattet, die verschiedene Parameter in Echtzeit ermitteln und übertragen. In Kombination mit Flugplänen können diese Daten, unter Einsatz geeigneter Machine Learning Algorithmen, zur Vorhersage des Zeitpunkts der nächsten Wartung und des Wartungsbedarfs genutzt werden. In dieser Arbeit werden diese Vorhersagedaten zur optimalen Planung und Steuerung der Kapazität eines MRO (Maintenance, Repair and Overhaul) Dienstleisters genutzt.

Im ersten Artikel, "Capacity Planning for a Maintenance Service Provider with Advanced Information", Kapitel 2 beziehungsweise Kurz (2016), wird die aus mehreren Stationen bestehende Produktionsstätte des MRO Dienstleisters mit Hilfe eines Netzwerks aus GI/G/1 Warteschlagen beschrieben [25]. Durch Lösung eines Optimierungsproblems werden die Kapazitäten der einzelnen Stationen so ermittelt, dass Kapazitäts- und Strafkosten für eine zu lange Durchlaufzeit minimiert werden. Darüberhinaus wird untersucht, wie Vorhersagedaten bezüglich des Eintreffens und Wartungsaufwands zukünftiger Aufträge genutzt werden können, um die Gesamtkosten zu reduzieren.

Der Artikel "Flexible Capacity Management with Future Information", Kapitel 3 beziehungsweise Kurz und Pibernik (2016), nutzt Informationen hinsichtlich zukünftiger Wartungsbedarfe für die Steuerung einer flexiblen Kapazität [27]. Die Produktionsstätte des MRO Dienstleisters wird als M/M/1 Warteschlange beschrieben, die zwischen einer Basiskapazität und einer erhöhten Kapazität wechseln kann. Allesdings kann die hohe Kapazität nur einen definierten Zeitanteil genutzt werden. In dem Artikel werden Politiken entwickelt, welche die erwartete Warteschlangenläge minimieren, falls keine Informationen bezüglich des Eintreffens zukünftiger Aufträge verfügbar sind beziehungsweise alle Informationen in einem unendlich langen Zeitfenster vorliegen. Es zeigt sich, dass die erwartete Warteschlangenlänge drastisch reduziert werden kann, falls Informationen über zukünftige Bedarfe genutzt werden können.

Im dritten Artikel, "Queueing with Limited Future Information", Kapitel 4 oder Kurz (2016), wird neben der Steuerung einer flexiblen Kapazität auch die Zulassungskontrolle behandelt: Welche Aufträge sollten umgeleitet werden, zum Beispiel an einen Subdienstleister, falls ein definierter Anteil aller ankommenden Triebwerke nicht angenommen werden muss [26]? Es werden Politiken zur Steuerung der flexiblen Kapazität und für die Zulassungskontrolle entwickelt, die zukünftige Informationen in verschieden langen Zeitfenstern berücksichtigen: keine Informationen, endlich und unendlich lange Zeitfenster. Numerische Analysen zeigen, dass die Berücksichtigung von Informationen über die Zukunft im Vergleich zu reaktiven Politiken zu einer Verringerung der mittleren Warteschlangenlänge führt. Andererseits wird ersichtlich, dass die Nutzung von kürzeren Zeitfenstern unter bestimmten Umständen vorteilhaft sein kann.

Den inhaltlichen Rahmen dieser Dissertation bilden die Einleitung im folgenden Kapitel sowie ein Ausblick in Kapitel 5. Im Hauptteil nicht dargestellte Beweise werden in den Anhängen A bis C zusammengefasst.

1 Introduction

During the last years, digitization has gained significant importance and can be observed in almost any area, from business to academia, politics to private life. Especially enterprises have invested in technology to generate, store and analyze data with sophisticated algorithms, and this trend is almost certainly going to grow even faster in the future. The term analytics, as it is used today, comprises the collection and analysis of data, using it to make predictions and finally prescribe well-defined actions. A stylized analytics process is illustrated in Figure 1.1. However, as found by McKinsey & Company (2016)

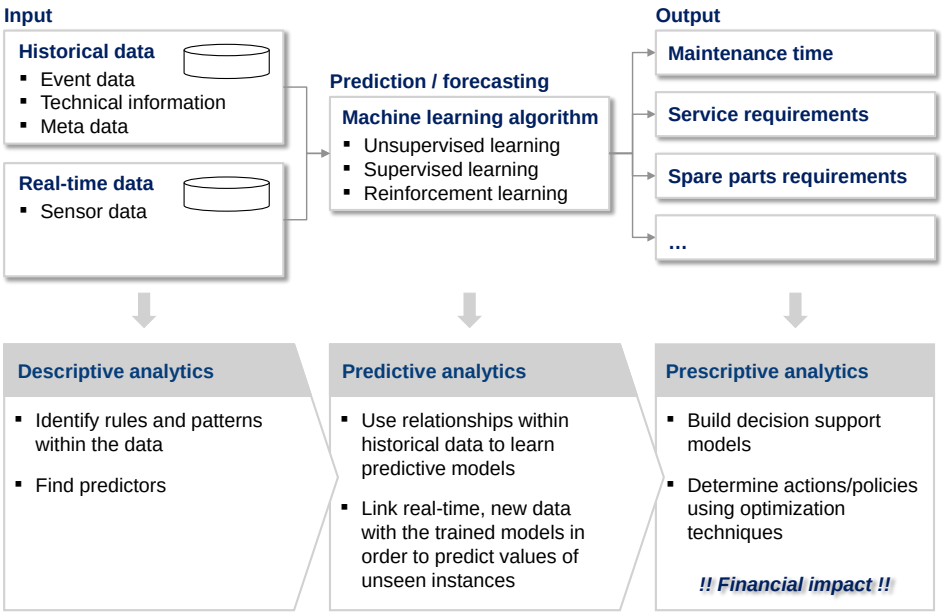


Figure 1.1: Example analytics process.

and others, more attention is dedicated to the data itself, its analysis and prediction of future events rather than prescriptions that have an actual impact

on economic performance [23, 29]. As an example: State-of-the-art aircraft engines are equipped with sensors that measure a variety of technical parameters and transmit the resulting data to central data warehouses periodically. Currently, a lot of effort is made towards developing algorithms to analyze the newly available data and, by using the engines' operation schedules, predict the engines' future conditions. Indisputably, this initiative improves flight safety and, on the other hand, the average time on-wing. However, the link towards using predictive information regarding the engines' future overhaul needs for the planning and control of the capacity of maintenance facilities is still missing.

Thus, the objectives of this dissertation are to (i) investigate how information about future jobs arriving to a service facility can be used for capacity planning and control and (ii) to analyze the benefits generated by using these information with respect to cost and waiting time. While trying to shed light on this superordinate subject, the thesis is composed of three independent parts that have been published as self-contained research articles. In the first article, a framework describing the usage of predictive information is developed and the resulting benefits with respect to capacity and lead time-related costs are investigated analytically and numerically. In the second and third article, we define policies that use information about future jobs to control a flexible capacity or divert jobs. Additionally, we investigate how predictive information impacts the queueing systems' performance, i.e., mean queue length or waiting time.

More specifically, the first article, "Capacity Planning for a Maintenance Service Provider with Advanced Information", Chapter 2 or Kurz (2016), solves the job shop-like capacity planning problem of an MRO service provider which overhauls aircraft engines of multiple, potentially competing customers [25]. First, the MRO's production network, which consists of eleven workstations for two engine types—from disassembly to final testing and certification—is modeled as a network of GI/G/1 queues. An algorithm based on the Queueing Network Analyzer developed by Ward Whitt in 1983 is used to compute the mean turnaround time per engine type, given service rate as well as service and interarrival time variability per workstation [44]. A total cost function

composed of capacity costs per work station and tardiness penalty costs for not meeting contractually specified mean lead times is defined. The capacity allocation problem minimizes this total cost function by choosing the optimal capacity per workstation. Depending on the parameters of the system, three ways to solve the optimization problem are distinguished: two cases where the solution can be obtained by solving a system of equations (based on the Karush-Kuhn-Tucker conditions for convex problems). For the third case, the optimization problem must be solved numerically using a subgradient method. The model is then used to investigate the implications of information about future jobs arriving to the system on total costs. A framework demonstrating the usage of advanced information is developed and, as a result, mean service requirements as well as service and interarrival time variabilities can be reduced. Analytical and numerical analyses imply that total costs can be reduced significantly if information about future jobs is available. Since the MRO service provider serves multiple competing customers, cryptographic methods such as Secure Multiparty Computation need to be used in order to prevent leakage of private data. Thus, this part of the chapter provides a decision-making tool for the investment case that needs to be considered when deciding whether to invest in the technology required or not.

The second article, "Flexible Capacity Management with Future Information", Chapter 3 or Kurz and Pibernik (2016), goes one step further [27].[1] We do not only consider the benefits of reduced service requirements or service and interarrival time variabilities, but investigate how information regarding the actual arrival times and service requirements of individual jobs (referred to as *future information*) can be used for capacity control. We model the service facility as an M/M/1 queue with a server that can switch between a base and a high capacity by activating the contingent capacity. However, we assume that the contingent capacity can only be used for a certain share of the time. The objective of the article is to develop capacity control policies minimizing the time-average queue length while taking different information into account. The reactive threshold-type policy, which only relies on the

[1] This part of the dissertation is coauthored by Richard Pibernik.

current queue length and not on future information, serves as a proxy for the performance of any kind of proactive policy. We develop a proactive policy that combines reactive (static threshold) and forward-looking elements. Basically, the forward-looking part of the policy is also based on a threshold. However, the threshold is dynamic and depends on the arrival times and service requirements of jobs that arrive at the system in the future. We are able to derive closed-form expressions for the time-average queue lengths for the cases of no or infinitely long lookahead window. Also, asymptotic optimality of the proactive policy is proved as the arrival rate approaches the time-average service rate. It is interesting to note that the mean queue length converges to a finite value in this case, while it diverges if the reactive policy is applied. Thus, the availability of future information leads to significantly reduced waiting times. We do not develop policies for the case of finite lookahead windows, but only provide some first analytical insights regarding their implications and effects.

Therefore, this problem is approached in the third article, "Queueing with Limited Future Information", Chapter 4 or Kurz (2016) [26]. We again model the service facility as an M/M/1 queue. Besides flexible capacity, we also consider diversion: Should a job be admitted to the queue or diverted? This model corresponds to the situation where the service provider is using a subcontractor to mitigate demand spikes. Similar as to the flexible capacity case, we assume that only a certain share of total arrivals can be diverted. This problem has already been solved for no and infinite lookahead windows by Spencer et al. (2014) and Xu and Chan (2016) [36, 49]. Thus, the objective of the article is to develop proactive capacity control and diversion policies minimizing the mean queue length given a lookahead window of finite length, i.e., limited future information. Depending on the rate of job arrivals to the system and the length of the lookahead window available, we characterize two distinct regimes. In the first regime, sufficient future information is available such that the proactive policies can be used as defined for a lookahead window of infinite length. The second regime corresponds to the case where the lookahead window is not long enough, i.e., we have insufficient future information. Therefore, we develop modified proactive capacity control and diversion

policies such that feasibility of the policies is always guaranteed. However, for both regimes, it is not possible to find analytic expressions for the mean queue length. Thus, a numerical analysis based on three conjectures is performed to shed light on the performance of the proactive policies given limited future information. The results suggest that the proactive and the modified proactive policies lead to lower mean queue lengths than their reactive counterparts. It is interesting to note that there exist parameter combinations where using less future information can actually be beneficial in terms of mean queue length.

An overview of the scientific contribution of this dissertation is presented in Table 1.1. For all three articles, queues have been used to solve capacity planning and control problems. It has been shown that information regarding future jobs has a positive effect on relevant performance indicators (total costs or mean queue length). Therefore, this dissertation provides examples how digitization and the "fourth industrial revolution" can improve businesses' operating and economic performance by prescribing actions on the basis of newly available data. Although the thesis is motivated from the point of view of an aircraft engine MRO (maintenance, repair and overhaul) facility, the policies and insights developed are not restricted to this application. There exist a variety of settings where advanced information can be used to plan and control the capacity of a production or service facility: general make-to-order / make-to-stock production with pre-orders, the allocation of server capacities for online applications or call centers. For call centers, e.g., advanced information can be generated and used for capacity planning if customers have to navigate through a menu indicating their inquiry before being connected to an agent.

Finally, summary and conclusion as well as avenues for future research are provided in Chapter 5. All proofs not stated in the main part of this dissertation and additional information regarding the numerical analyses are relegated to Appendices A to C.

	Capacity planning with adv. info. (Chapter 2)	**Capacity mgmt. with future info.** (Chapter 3)	**Queueing with limited future info.** (Chapter 4)
Analytical model	• Network of GI/G/1 queues with two job types • Optimization model: minimize capacity and tardiness penalty costs	• M/M/1 queue with flexible capacity • Policy minimizing mean queue length given no / infinite future information	• M/M/1 queue with diversion or flexible capacity • Policies minimizing mean queue length given limited future information
Methodological contribution	• Reformulation of the optimization model to solve it explicitly in two out of three cases • Investigation of the effects of advanced information on total costs	• First article considering future information for capacity control • Derivation of analytic expressions for the mean queue length	• First article investigating the effects of limited future information for diversion and capacity control • Development of new methodology to deal with insufficient future information
Conceptual findings	• Total costs can be reduced significantly if advanced information is available	• Future information in an infinite lookahead window leads to low mean queue length for all possible arrival rates • Proactive policy outperforms reactive policy	• Proactive policies with limited future information dominate reactive policies • Artificially truncating the lookahead window can be beneficial for certain parameter combinations

Table 1.1: Overview of scientific contribution.

2 Capacity Planning for a Maintenance Service Provider with Advanced Information

Analytical and iterative optimization techniques are employed to solve a job shop-like capacity planning problem for a maintenance service provider with contractually defined lead time requirements. The problem is motivated by a real-life case example, namely the overhaul of airline aircraft engines through an external service provider. The production network is modeled as a network of GI/G/1 queues, where the service rates are the decision variables and capacity costs and penalty costs for not meeting contractually defined lead times are minimized. Additionally, we analytically investigate the effects of collaborative maintenance management as a source of advanced information regarding future maintenance demand. More specifically, we consider the benefits of improved service rates and service and demand variabilities on production capacities and total costs. Numerical examples are provided to verify the proposed optimization procedure and illustrate the effects of collaborative maintenance management.[2]

2.1 Introduction and Overview

Today, the rise of Cyber-Physical Systems and the Internet of Things facilitates real-time condition monitoring and condition-based maintenance of technical equipment [50]. Combining real-time condition data with plans of

[2]This paper was published in the *European Journal of Operational Research* [25].

future equipment usage makes it possible to predict future condition and maintenance requirements. This is what we refer to as *advanced information*. Huge data streams are collected, stored and used for the optimization of maintenance schedules, production plans and supply chain operations. Sophisticated technical equipment such as Rolls-Royce's Trent aircraft engines are equipped with sensors that measure different parameters and send the data to data warehouses in real time [20]. For Trent engines, this data includes more than 20 parameters such as oil pressure, oil temperature, and vibration levels that can be used for condition-based maintenance scheduling and optimized planning of maintenance operations. However, data usage is challenging in a service provider or contract manufacturer setting, as the data is collected by the owner and oftentimes not shared with the supplier. This can be, e.g., due to privacy concerns or potential loss of competitive information. Especially if more than one customer is involved and information from different parties needs to be considered, a sound analysis of potential benefits is necessary in order to convince customers to participate and invest in a collaborative planning system.

In this paper, we analyze the benefits associated with advanced information for a maintenance service provider that can be obtained through collaborative maintenance management. More specifically, we investigate how advanced information influences the optimal capacity allocations in a complex maintenance and repair process, the associated capacity costs and the potential penalties for tardiness. Since advanced information oftentimes requires substantial investments, it is important to understand the relationship between advanced information and the costs of the system. The results of our analysis help supply chain partners to decide whether or not to invest in advanced information technologies and collaborative planning systems. The research described in this paper is part of a larger initiative whose objective is to investigate potential applications of secure cloud-based computation in supply chain management scenarios.[3] While we focus on cost-optimal capacity planning for a maintenance service provider setting in this document

[3]PRACTICE - Privacy-Preserving Computation in the Cloud, `http://practice-project.eu`

(highlighted in Figure 2.2), we also investigate the benefits of secure (data privacy-preserving) collaborative demand forecasting with decision tree learning, spare parts management with advanced information, secure collaborative arrival slot scheduling, and secure vendor managed inventory [51].

Our research is motivated by a practical use case, namely the overhaul of aircraft engines of multiple airlines through an MRO[4] service provider. Airlines ship their engines to the service provider when they need to be overhauled. The service provider then processes the engines and sends them back after completion. A typical MRO engine overhaul production system is illustrated in Figure 2.1. There are two sources of uncertainty in this system. One

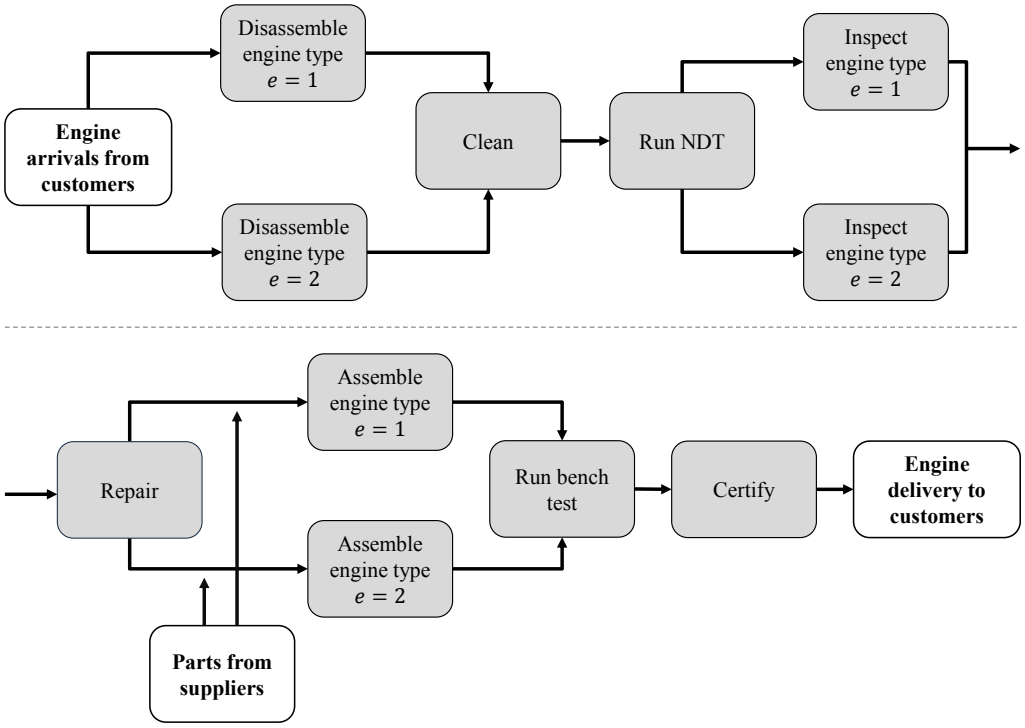


Figure 2.1: Engine overhaul production network (based ib [51]).

is that the engine arrival stream is a stochastic process where the interarrival time between subsequent engines is a random variable. Second, the service times at the work stations illustrated in Figure 2.1 are random variables and

[4]Maintenance, repair and overhaul

depend on the required service, on worker skill, and on spare parts availability.

In this paper we seek to determine the capacities of the work stations in the MRO production network such that total costs (capacity costs as well as penalty costs for not meeting contractually defined turnaround times) are minimized. Additionally, we investigate the benefits of advanced information that can be obtained from a closer collaboration between the service provider and its customers. Through information sharing and collaborative scheduling of engine arrivals, the variabilities of engine interarrival and service times can be reduced, and the mean time required for service at certain work stations can be shortened. Investigating the influence of these effects on optimal capacity is a complex problem, as these parameters are interlinked throughout all work stations in the network. Figure 2.2 illustrates a framework for collaborative maintenance management in the aerospace MRO business leading to the aforementioned improvements.

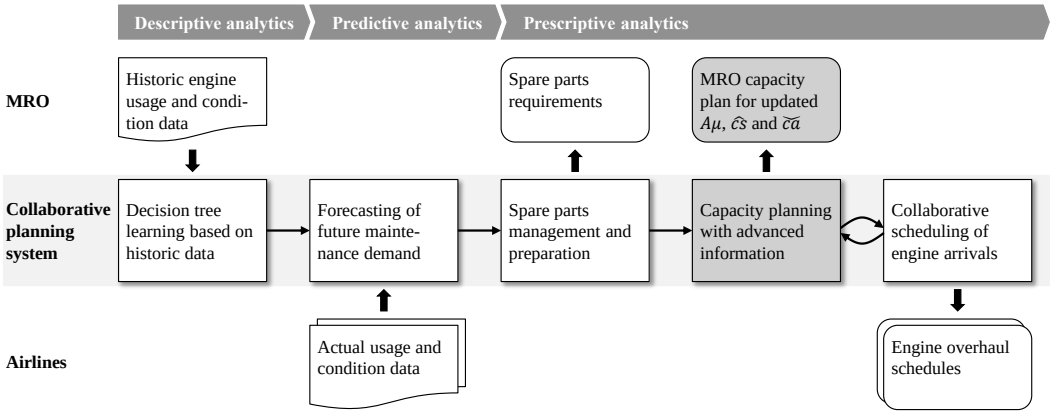


Figure 2.2: A framework for collaborative aircraft engine overhaul management.

While we focus on capacity planning (shaded in Figure 2.2), we briefly explain the basic ideas of the other process steps in the collaborative maintenance management framework.

Using historical and actual engine usage and condition data, it is possible to predict when an engine will arrive at the service provider, and in what condition (this is *advanced information*). If the service provider is able to use this data, spare parts can be ordered such that service levels are improved.

Additionally, the provider can plan for and prepare overhaul operations beforehand to optimize the activities carried out when the engine arrives. Both improvements contribute to a reduction of mean service time and service time variability. Once a forecast of future engine arrivals is available, actual engine arrivals can be scheduled such that their interarrival time variability is reduced.

These improvements are only possible through investments in a collaborative planning system. On the one hand, aircraft engines need to be able to report their usage and condition parameters (captured through sensors) to a central data warehouse in a timely fashion. As the data reported by the individual airlines is highly confidential, cryptographic methods are necessary to prevent competitive information from being leaked. The goal of this paper is to develop a tool that can estimate the benefits of different collaborative scenarios for the service provider and customer, so that they can decide whether or not to invest in the collaborative planning system.

The remainder of this paper is structured as follows. Section 2.2 provides a review of the relevant literature. In Section 2.3, we model the service provider's production network as a network of queues. We employ analytical and iterative optimization methods to find the optimal capacities per work station in the network. In Section 2.4 we analyze the effects of advanced information on the queueing network and total costs. Structural insights are supported by numerical analyses in Section 2.5. Section 2.6 provides conclusions and future outlooks. The proofs of propositions and the queueing network parameter algorithm can be found in the appendix.

2.2 Literature Review

This literature review focuses on existing research regarding queueing networks for production capacity planning. Although this research area dates back to the 1950s, many problems remain unsolved and the field is still subject to continuing research. Additionally, we provide an overview of the most important publications dealing with the effects of advanced information ob-

tained through supply chain collaboration. As this paper is motivated by an aerospace case, relevant literature regarding the overhaul of aircraft engines is also outlined.

We use a queueing network model to describe the service provider's production network. To date, most publications considering queueing networks for production planning have dealt with the fundamental tradeoff between capacity costs and lead time. They focus either on finding a cost-optimal capacity allocation subject to a maximum lead time constraint [8, 31] or, vice versa, finding a lead time-minimizing capacity allocation subject to a cost (or budget) constraint [9, 14]. In contrast, we combine both capacity and lead time penalty costs in the objective function in order to reflect the service provider setting.

As exact analytical results for queueing network performance metrics such as the sojourn (or lead) time can be computed only for a limited range of queueing networks, approximate expressions were developed for more general types of queueing networks (e.g., if the service and interarrival times at the nodes are not exponentially distributed). In this paper we use the Krämer-Langenbach-Belz approximation, published in 1976, to compute the number of customers in a single GI/G/1 queue [24]. In order to approximate sojourn times in a queueing network, Whitt proposed in 1983 to decompose the network and evaluate each node separately [44]. In his Queueing Network Analyzer (QNA) approach, Whitt developed iterative algorithms to describe the queueing parameters at node level resulting from the network structure. Adjustments and improvements to this approach are summarized in Bitran and Morabito (1996) [5]. Lately, Wu and McGinnis (2012, 2013) developed new methods to approximate cycle and waiting times in queueing networks based on the intrinsic ration, especially suited for networks in heavy traffic [46, 47].

In 1989, Bitran and Tirupati developed tradeoff curves between work-in-process (WIP), lead time, and capacity (represented by the service rates at the nodes) [8]. They formulated a convex program minimizing total cost of capacity subject to a total lead time constraint for a general open network of queues with Markovian or deterministic routing. Bretthauer and Côté

(1996) developed nonlinear programming models for time-dependent capacity planning in manufacturing systems [14]. They present two models, one to minimize capacity costs under a WIP constraint and one to minimize lead times subject to a capacity cost constraint. In 2002, Hopp et. al developed an Optimized Queueing Network (OQNet) capacity planning tool to support the design and reconfiguration of queueing network-like semiconductor production facilities [21]. Using a heuristic algorithm, they minimize facility costs such that specific volume and lead time targets are satisfied, while these performance metrics are determined using the general open queueing network approximations. More discussion regarding tradeoff curves between WIP, lead time and capacity in this context, also using QNA approximations and convex/MIP models for capacity planning of semiconductor manufacturing networks, can be found in the review of Bitran and Morabito (1999) [6]. Da Silva and Morabito (2009) determine how to optimally allocate capacity in a job shop-like queueing network of a metallurgical plant [31]. They use approximate parametric decomposition models to compute performance metrics such as WIP and lead time and apply optimization models minimizing capacity costs. Finally, Morabito et al. (2014) study network routing decisions for multicommodity flows. They provide and compare different performance approximation approaches for generalized open queueing networks [30].

To the best of our knowledge, there have been no studies to date that incorporate both capacity costs and lead time penalties in the objective function. This means that in existing publications, capacity costs or lead time are constraints that must be fulfilled; in our setting, however, contractually defined lead times can be exceeded if marginal capacity costs are more expensive than marginal penalty costs. On the one hand, this best reflects the actual business case motivating our research, but on the other hand it also allows us to explore the effects of advanced information on total costs (capacity and penalty costs). Additionally, whereas most other studies use heuristic algorithms to solve the optimization problems, we are (depending on the problem parameters) partially able to find explicit solutions.

The effects of advanced information have been explored in a variety of settings. In the most general of terms, advanced information can be used

to reduce uncertainty. In return, this lowers capacity or inventory requirements. For supply chain collaboration through information sharing, we refer the reader to Poler et al. (2008) [33]. Collaborative scheduling of demand, taking into account customer demand and supplier capacity information is discussed, for example, in Chen and Hall (2007) [16]. In the MRO context, collaborative planning leads to reduced variability of service and interarrival times and finally to reduced costs. So far, this has not yet been explicitly studied for service provider production networks.

Interesting publications regarding the overhaul operations of aircraft engines include, for example, Stranjak et. al (2008) [39]. They present an agent-based simulator to predict and schedule aircraft engine overhauls, motivated by competitiveness considerations in a highly complex and dynamic business environment. Additionally, Rolls-Royce recently published an online article in which they explain how they use big data gathered by the sensors of the Trent engines and advanced analytics to achieve cost-efficient engine maintenance scheduling for their customers [20]. Reményi and Staudacher (2014) present a simulation-based approach that is used to identify scheduling rules for aircraft engine maintenance carried out by a MRO for multiple airlines [34]. For more information regarding aircraft engine usage and condition parameters, we refer the reader to Batalha (2012) [18]. We add to this stream of research and literature by Taigel (2015) [43].

2.3 Optimized Queueing Network Model

We need to explicitly model the complex nature of the maintenance production system to determine optimal capacities per work station and to highlight how advanced information impacts both capacities and costs. Therefore, we model the system as a queuing network where the service rates of the individual work stations are chosen as optimization variables. The basic capacity planning model for maintenance services is developed in Section 2.3.1. In Section 2.3.2, we describe how to approximate sojourn times in the queueing network. Finally, analytical and iterative solution methods for the optimiza-

tion problem are detailed in Section 2.3.3.

2.3.1 Queueing Network Model for Maintenance Services

Within the service or production process, products (e.g., aircraft engines) pass through successive work stations under an FCFS[5] policy. In practice, service providers or contract manufacturers and their customers often have contracts specifying maximum lead times (also referred to as sojourn or turnaround times) for service or production processes. Over a finite planning horizon, the actual mean lead time is computed and compared to the contractually defined maximum lead time per product family. If this time is exceeded, the service provider incurs a penalty which increases with increasing mean lead time. The objective of the capacity allocation model developed here is to determine the capacity per work station such that total costs are minimized. It is worth noting again that total costs are comprised of both capacity costs (associated with the service rate) and penalty costs (incurred for not meeting contractually defined maximum lead times). Before developing the mathematical model to optimally determine production capacities, we summarize the notation used throughout the next sections.

The production network consists of work stations $j \in \mathcal{J} = \{1, \ldots, J\}$, processing product families $e \in \mathcal{E} = \{1, \ldots, E\}$. Each product family e follows a predetermined acyclic path $\mathcal{J}_e \subseteq \mathcal{J}$ through the network.

The decision variables of the optimization model are defined as μ_j, representing the capacity or mean service rate at each work station $j \in \mathcal{J}$. The associated per unit capacity costs are denoted by c_j. Let $S_j(\mu_j)$ denote the service rate-dependent mean sojourn times at station j and let S_e^T denote the contractually defined maximum mean total sojourn time for product family e after which a penalty is incurred.[6] The tardiness penalty cost per time unit for product family e is denoted by γ_e and is incurred for positive differences of the actual total sojourn time minus the contractually defined maximum sojourn time.

[5]First come, first served

[6]We assume the same penalty costs and parameters S_e^T per product family for all customers.

As we will see in Section 2.3.2, we need some additional parameters in order to compute the sojourn times $S_j(\mu_j)$. We define λ_j as the mean arrival rate of jobs at work station j (sum of the individual arrival rates of all product families served at work station j). ca_j denotes the squared coefficient of variation (ratio of the variance to the squared mean of a random variable) of product interarrival times at work station j. Service time variability at the work stations is captured as the squared coefficient of variation of service times cs_j. Finally, $L_j(\mu_j)$ denotes the approximate expected number of products at work station j given service rate μ_j.

The cost-optimal capacity allocation problem (CAP) can be written as follows:

$$\text{minimize} \quad \mathcal{C}(\mu) = c^\top \mu + \sum_{e=1}^{E} \gamma_e \Big[\sum_{j \in \mathcal{J}_e} S_j(\mu_j) - S_e^T \Big]^+, \tag{CAP}$$

where $[\cdot]^+ = \max\{0, \cdot\}$ and the domain of $\mathcal{C}$ is the open convex set

$$\mathbf{dom}\ \mathcal{C} = \{\mu \mid \mu_j > \lambda_j,\ j = 1, \ldots, J\} \subset \mathbb{R}^J_{++}.$$

The first term of the objective function represents the total direct capacity costs increasing linearly with capacity. The second term accounts for total penalty costs for all product families.[7] The $[\cdot]^+$ operator ensures that a penalty is only incurred for a product family e if

$$\sum_{j \in \mathcal{J}_e} S_j(\mu_j) > S_e^T.$$

The total costs function is developed to conform to the actual costs incurred by the maintenance service provider of aircraft engines introduced in Section 2.1. The service provider can allocate capacity such that contractually defined lead times are not met as long as marginal capacity cost reductions exceed increased marginal penalty costs for a given product family.

The domain restriction $\mu_j > \lambda_j,\ \forall j \in \mathcal{J}$, assures that the network of

[7] We do not consider fixed cost and material cost as they are not influenced by capacity changes in the presented setting.

queues is stable. This means that the sojourn times $S_j(t)$ (or queue lengths $L_j^q(t)$) at the work stations do not grow towards infinity as time approaches infinity, $\lim_{t\to\infty} L_j^q(t)/t = 0,\ \forall j \in \mathcal{J}$. Service rates are assumed to be continuous variables, which is quite common in the context of production planning with queueing networks; see e.g. Bitran and Tirupati (1989) and the references therein [8].

Since the service rate-dependent mean sojourn times $S_j(\mu_j)$ determine the penalty costs, they are the key performance measures in the production network. However, computing the sojourn times in a queueing network is not trivial due to the interdependencies of the individual nodes. Therefore, we subsequently describe an approximation method based on various queueing network parameters.

2.3.2 Sojourn Time Approximation for Queueing Networks

The individual work stations can be modeled as queues with independent and generally distributed interarrival times, generally distributed services times, one server, and infinite waiting room (abbreviated as GI/G/1/∞). This type of queue allows us to account for all possible kinds of probability distribution functions for the interarrival and service times. We do not consider M/M/1 queues first, but use GI/G/1 queues directly as this research is motivated by the MRO case where interarrival and service times are generally not exponentially distributed. One could assume that interarrival times may, without scheduling, be modeled as exponentially distributed random variables for each product family, resulting in a squared coefficient of variation $ca = \mathrm{Var}[X]/\mathrm{E}[X]^2 = \lambda^2/\lambda^2 = 1$. On the other hand, if we imply scheduled arrivals as described in the introduction, interarrival time variability decreases and interarrival times can become close to constant, $ca \approx 0$.

We can use parametric decomposition to approximate the production system as a network of J independent GI/G/1 queues, where each product family e follows a predetermined acyclic path comprising work stations $\mathcal{J}_e$ through the network. We assume that the union of all product family-specific paths

is equal to the set of work stations in the network, $\bigcup_{e\in\mathcal{E}} \mathcal{J}_e = \mathcal{J}$.[8]

The production network of the aerospace case described in Section 2.1 and displayed in Figure 2.1 for two engine types, $\mathcal{E} = \{1, 2\}$, can be schematically illustrated as shown in Figure 2.3.

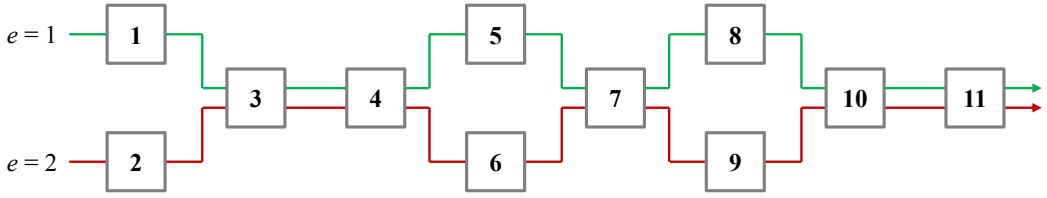


Figure 2.3: Schematic illustration of example production network.

The Krämer-Langenbach-Belz approximation for GI/G/1 queues [24] approximates the steady-state mean number of products L at a given node with a service rate μ as

$$L(\mu) = \frac{\lambda}{\mu} + \frac{(ca + cs)\lambda^2}{2} \frac{1}{\mu(\mu - \lambda)} g(\mu) \tag{2.1}$$

where

$$g(\mu) = \begin{cases} \exp\left\{\frac{-2(1-ca)(\mu-\lambda)}{3\lambda(ca+cs)}\right\} & \text{if } ca \leq 1 \\ 1 & \text{otherwise.} \end{cases} \tag{2.2}$$

With Little's law, $L = \lambda S$, which also applies to GI/G/1 queues, and assuming for all nodes that $ca, cs \in [0, 1]$, the KLB approximation (2.1) for the mean number of products in the system can also be used to approximate the mean sojourn time (sum of expected waiting and service time per customer) at a work station j,

$$S_j(\mu_j) = \frac{1}{\mu_j} + \frac{(ca_j + cs_j)\lambda_j}{2} \frac{1}{\mu_j(\mu_j - \lambda_j)} \exp\left\{\frac{-2(1 - ca_j)(\mu_j - \lambda_j)}{3\lambda_j(ca_j + cs_j)}\right\}, \tag{2.3}$$

[8] For details regarding the decomposition approximation, the reader is referred to Shanthikumar and Buzacott [35], Whitt [44], Bitran and Tirupati [8] and Negri da Silva and Morabito [31].

for all $j \in \mathcal{J}$. Equation (2.3) provides an approximation of the mean sojourn times $S_j(\mu_j)$ for each work station in the production network. We can assume to know the arrival processes to the system and the service time variabilities cs_j for all work stations. However, although we assume independence of the queues at the individual nodes, it is still necessary to determine the arrival rates λ_j and their variability parameters ca_j for all work stations (except the ones where the engines arrive), depending on the network structure. These parameters are approximated prior to the optimization based on techniques proposed by Bitran and Morabito 1996 [5]. While the arrival rates can be found quite easily from the subsets $\mathcal{J}_e$, the interarrival time variability at a work station is driven by the interarrival time and the service time variability of the previously visited work station. The algorithms to determine the queueing network parameters can be found in Appendix A.3.

We follow the approach proposed by several authors, e.g., by Bitran and Tirupati 1989 [8], and use the following assumption.

Assumption 2.1. *The effect of a change in service rate μ_j on the interdeparture time variability cd_j and therefore on the interarrival time variability $ca_i,\ i > j$, can be neglected.*

Therefore, λ_j and ca_j are computed for all work stations $j \in \mathcal{J}$ once for a reasonable guess of $\mu \in \mathbf{dom}\ \mathcal{C}$ prior to the optimization routine. $S_j(\cdot)$ is treated as a univariate function of μ_j. During optimization, all other parameters are assumed to remain constant. Assumption 2.1 generally holds as long as the service rates estimate is close to the optimal service rates. However, if the service rates are hard to estimate for a certain problem, the shortcomings of the assumption can be circumvented by iteratively solving the optimization problem, taking the optimal service rates as initializing service rates for the queueing network parameter algorithm and solving the optimization problem again with the updated parameters.

2.3.3 Solution Procedure

With the definition of the approximate mean sojourn times we can derive structural insights regarding the capacity allocation problem defined in Sec-

tion 2.3.1.

Proposition 2.1. *With Assumption 2.1 and* $ca + cs > 0$*, the capacity allocation problem* (CAP) *is convex in* $\mathbf{dom}\ \mathcal{C}$ *and has a unique optimal solution* $\mu^\star \in \mathbf{dom}\ \mathcal{C}$*.*

Proof. See Appendix A.1. □

Due to the $[\cdot]^+$ operator in the total cost function, $\mathcal{C}$ is generally not differentiable in $\mathbf{dom}\ \mathcal{C}$. Therefore, a point $\mu^\star \in \mathbf{dom}\ \mathcal{C}$ is a minimizer of $\mathcal{C}$ if and only if $0 \in \partial\mathcal{C}(\mu^\star)$, where $\partial\mathcal{C}(\mu)$ denotes the subdifferential (or the set of subgradients) of $\mathcal{C}$ at point $\mu \in \mathbf{dom}\ \mathcal{C}$ [10]. In order to construct $\partial\mathcal{C}(\mu)$, we first note that the subdifferential of the pointwise maximum of differentiable convex functions $f_1, \ldots, f_m$ is given as the convex hull of the union of the gradients of the active functions,

$$\partial f(x) = \partial \max_{i=1,\ldots,m} f_i(x) = \mathbf{Co} \cup \{\nabla f_i(x) \mid f_i(x) = f(x)\}.$$

For the $[\cdot]^+$ operator as defined in the capacity allocation problem,

$$f_e(\mu) = \Big[\sum_{j \in \mathcal{J}_e} S_j(\mu_j) - S_e^T \Big]^+,$$

the subdifferential is therefore given by

$$\partial f_e(\mu) = \begin{cases} 0 & \text{if } \sum_{j\in\mathcal{J}_e} S_j(\mu_j) < S_e^T \\ \nabla \sum_{j\in\mathcal{J}_e} S_j(\mu_j) & \text{if } \sum_{j\in\mathcal{J}_e} S_j(\mu_j) > S_e^T \\ \Big[0, \nabla \sum_{j\in\mathcal{J}_e} S_j(\mu_j)\Big] & \text{if } \sum_{j\in\mathcal{J}_e} S_j(\mu_j) = S_e^T. \end{cases} \tag{2.4}$$

Using this we can write the subdifferential of $\mathcal{C}$ for $\mu \in \mathbf{dom}\ \mathcal{C}$ as

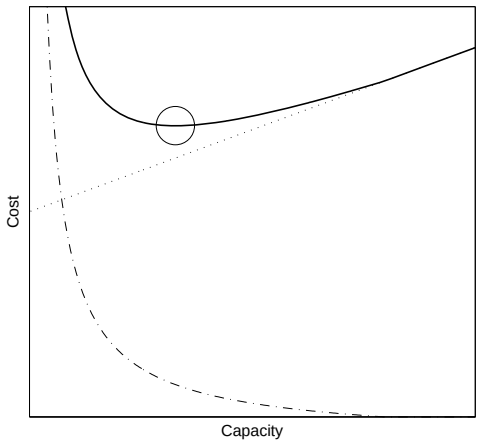
$$\partial\mathcal{C}(\mu) = c^\top + \sum_{e\in\mathcal{E}_>} \gamma_e \nabla \sum_{j\in\mathcal{J}_e} S_j(\mu_j) + \sum_{e\in\mathcal{E}_=} \gamma_e \Big[0, \nabla \sum_{j\in\mathcal{J}_e} S_j(\mu_j)\Big], \tag{2.5}$$

with $\mathcal{E}_>, \mathcal{E}_=, \mathcal{E}_< \subseteq \mathcal{E}$ denoting the disjoint subsets of product families where

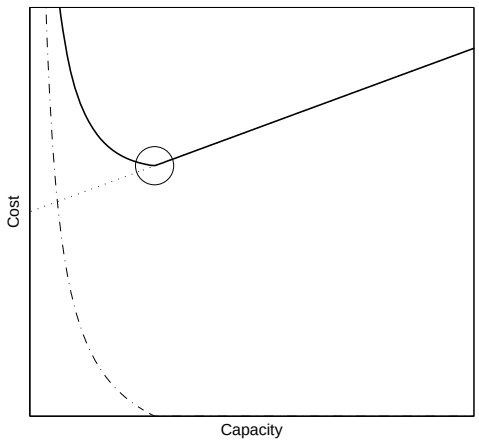
$$\sum_{j\in\mathcal{J}_e} S_j(\mu_j) > S_e^T, \qquad \sum_{j\in\mathcal{J}_e} S_j(\mu_j) = S_e^T \qquad \text{and} \qquad \sum_{j\in\mathcal{J}_e} S_j(\mu_j) < S_e^T,$$

respectively. It is easy to see that $\mathcal{E}_> \cup \mathcal{E}_= \neq \emptyset$ for the minimizer $\mu^\star$, since (2.5) reduces to $\partial \mathcal{C}(\mu) = c^\top \neq 0$ otherwise. Furthermore, since $\partial \mathcal{C}(\mu) \in \mathbb{R}^J$, there must be enough product families with mean total sojourn time greater or equal to the contractually defined maximum mean total sojourn time such that all work stations are encompassed in their paths, $\bigcup_{e \in \mathcal{E}_> \cup \mathcal{E}_=} \mathcal{J}_e = \mathcal{J}$.

Figure 2.4 illustrates the total cost function as a one-dimensional curve for different values of S_e^T. The optimal point is indicated by the circle. Panel 2.4a represents the cost curves for subsets $\mathcal{E}_>$, panel 2.4b the cost curves for subsets $\mathcal{E}_=$.



(a) Low S_e^T: Mean lead time is higher than S_e^T for the optimal capacity, a penalty is incurred. The curve is differentiable at the minimum.



(b) High S_e^T: Mean lead time equals S_e^T for the optimal capacity, no penalty is incurred. The curve has a kink at the minimum.

Figure 2.4: One-dimensional illustration of total cost function $\mathcal{C}(\mu)$ (—), composed of capacity costs $(\cdots)$ and expected penalty costs $(\cdot\ -)$.

As we are not able to solve $0 \in \partial \mathcal{C}(\mu^\star)$ directly, we distinguish two limiting cases that can be solved directly and a general case for any other solution depending on the differentiability of (CAP). The two limiting cases are defined such that all relevant penalty terms are either differentiable (corresponding to panel 2.4a) or not differentiable (corresponding to panel 2.4b). The general case applies if the penalty cost term is differentiable for only a part of the product families. Due to the complex structure of the problem, it is a priori not possible to determine which case applies for a given set of parameters.

Therefore, we solve the optimization problem for all three cases using different methodologies and determine the "true" optimal solution from their respective outcomes. In formal terms, the three cases can be defined as follows.

- *Case 1*: For given parameters S_e^T, the optimal solution $\mu^\star \in \mathbf{dom}\ \mathcal{C}$ is such that $\mathcal{E}_> \cup \mathcal{E}_< = \mathcal{E}$ and $\mathcal{E}_= = \emptyset$, i.e., $\mathcal{C}$ is differentiable in $\mathbf{dom}\ \mathcal{C}$.
- *Case 2*: For given parameters S_e^T, the optimal solution $\mu^\star \in \mathbf{dom}\ \mathcal{C}$ is such that $\mathcal{E}_> \cup \mathcal{E}_< = \emptyset$ and $\mathcal{E}_= = \mathcal{E}$, i.e., there exists no path $\mathcal{J}_e,\ e \in \mathcal{E}$, where the penalty cost term is differentiable in $\mathbf{dom}\ \mathcal{C}$.
- *Case 3*: For given parameters S_e^T, the optimal solution $\mu^\star \in \mathbf{dom}\ \mathcal{C}$ is such that $\mathcal{E}_> \cup \mathcal{E}_< \neq \emptyset$ and $\mathcal{E}_= \neq \emptyset$. This means that there exist paths $\mathcal{J}_e,\ e \in \mathcal{E}_> \cup \mathcal{E}_<$ where the penalty cost terms are differentiable and paths $\mathcal{J}_e,\ e \in \mathcal{E}_=$ where the penalty cost terms are not differentiable in $\mathbf{dom}\ \mathcal{C}$.

For *Case 1* and *Case 2*, the solution of the optimization problem (CAP) can be found explicitly from the problem formulation as shown in Proposition 2.2 and Proposition 2.3, respectively. For the general case we use an iterative optimization technique.

Case 1

In the following proposition we show that we can explicitly determine $\mu^{\star 1} \in \mathbf{dom}\ \mathcal{C}$ the first case directly from the problem statement.

Proposition 2.2. *With Assumption 2.1, $ca + cs > 0$ and if the contractually defined maximum mean total sojourn times S_e^T are such that the total costs function $\mathcal{C}$ is differentiable at the Case 1-optimal solution $\mu^{\star 1} \in \mathbf{dom}\ \mathcal{C}$, then $\mu^{\star 1}$ can be found by solving at most $J(E^2 - 1)$ equations.*

Proof. See Appendix A.1. □

For each node in the queueing network, the equation to be solved is explicitly given by

$$c_j = -\sum_{e \in \mathcal{E}_>^j} \gamma_e \frac{\partial S_j(\mu_j^{\star 1})}{\partial \mu_j}, \qquad \forall j \in \mathcal{J}, \tag{2.6}$$

where $\mathcal{E}_>^j \subseteq \mathcal{E}_>$ is the set of product families where node j is contained in all paths, $j \in \mathcal{J}_e,\ \forall e \in \mathcal{E}_>^j$ (see equation (A.13) in Appendix A.2 for the explicit derivative $\partial_{\mu_j} S_j(\mu_j)$). As an example, for the network depicted in Figure 2.3, we only need to solve J equations as there is only one possible outcome of $\mathcal{E}_>$, namely $\mathcal{E}_> = \mathcal{E}$, due to the condition $\bigcup_{e\in\mathcal{E}_>} \mathcal{J}_e = \mathcal{J}$.

Case 2

In the following proposition we show that the solution of the second case $\mu^{\star 2} \in \mathbf{dom}\ \mathcal{C}$ can also be determined directly by a reformulation of the problem and exploitation of duality properties.

Proposition 2.3. *With Assumption 2.1, $ca + cs > 0$ and if the contractually defined maximum mean total sojourn times S_e^T are such that there exists no path $\mathcal{J}_e$, $e \in \mathcal{E}$, where the penalty cost term is differentiable at the Case 2-optimal solution $\mu^{\star 2} \in \mathbf{dom}\ \mathcal{C}$, then $\mu^{\star 2}$ can be found by solving a system of $J + E$ equations.*

Proof. See Appendix A.1. □

Explicitly, the system of equations to be solved is given by

$$\begin{aligned} 0 &= c_j + \sum_{e\in\mathcal{E}^j} \nu_i^\star \frac{\partial S_j(\mu_j^{\star 2})}{\partial \mu_j} && \forall j \in \mathcal{J} \\ 0 &= \sum_{j\in\mathcal{J}_e} S_j(\mu_j^{\star 2}) - S_e^T && \forall e \in \mathcal{E}, \end{aligned}$$

with KKT multipliers $\nu_i^\star$ as explained in the proof of the proposition given in Appendix A.1.

As the negative slope of S_j increases with decreasing μ_j, it is obvious that *Case 2* becomes more likely to be the right way to solve (CAP) (as compared to *Case 1*) for increasing S_e^T. In other words, if the contractually defined maximum mean total sojourn times are relatively high, then we define capacity such that actual and contractually defined sojourn times coincide (incremental penalty cost increases for capacity reduction are higher than

incremental capacity cost savings).

Case 3

If $\mu^\star \in \mathbf{dom}\ \mathcal{C}$ is such that $\mathcal{E}_> \neq \emptyset$ and $\mathcal{E}_= \neq \emptyset$ for given parameters S_e^T, there are paths for which the penalty cost term is differentiable and paths where this is not the case. Thus, we need to solve the problem with an iterative method for nonlinear and nondifferentiable optimization, e.g., the subgradient method.

The subgradient method is applicable to unconstrained optimization problems. Therefore, in order to determine the solution of (CAP), we define the extended-value extension $\breve{\mathcal{C}}$ with $\mathbf{dom}\ \breve{\mathcal{C}} = \mathbb{R}^J$ of the total cost function as $\breve{\mathcal{C}}(\mu) = \infty,\ \mu \notin \mathbf{dom}\ \mathcal{C}$ and $\breve{\mathcal{C}}(\mu) = \mathcal{C}(\mu),\ \mu \in \mathbf{dom}\ \mathcal{C}$. Thus, we can minimize $\breve{\mathcal{C}}$ with the subgradient method which deploys the iteration

$$\mu^{(k+1)} = \mu^{(k)} - \alpha_k g^{(k)}. \tag{2.7}$$

The k-th iterate $\mu^{(k)}$ and any subgradient $g^{(k)} \in \partial\breve{\mathcal{C}}(\mu^{(k)})$ are used to determine the $(k+1)$-th iteration, where $\alpha_k > 0$ denotes the k-th step size. Although Equation (2.7) looks like the ordinary gradient method for differentiable functions, it is different since it is not a descent method, meaning that $\breve{\mathcal{C}}(\mu^{(k+1)}) > \breve{\mathcal{C}}(\mu^{(k)})$ can happen. In conclusion, after each iteration we set

$$\breve{\mathcal{C}}_{\text{best}}^{(k)} = \min\{\breve{\mathcal{C}}_{\text{best}}^{(k-1)}, \breve{\mathcal{C}}(\mu^{(k)})\}, \tag{2.8}$$

and $i_{\text{best}}^{(k)} = k$ if $\breve{\mathcal{C}}(\mu^{(k)}) = \breve{\mathcal{C}}_{\text{best}}^{(k-1)}$, i.e., if $\mu^{(k)}$ is the best point found so far. The gradient $g^{(k)}$ can for example be computed as

$$g^{(k)} = c^\top + \sum_{e\in\mathcal{E}_>} \gamma_e \nabla \sum_{j\in\mathcal{J}_e} S_j(\mu_j^{(k)}) + \sum_{e\in\mathcal{E}_=} r\gamma_e \nabla \sum_{j\in\mathcal{J}_e} S_j(\mu_j^{(k)}) \in \partial\breve{\mathcal{C}}(\mu^{(k)}),$$

where $r \in [0,1]$ is a random number to reflect the convex hull property in (2.5).

We have now introduced ways to compute the solution of the optimization problem in each of the three possible constellations depending on the input

parameters. To determine the "true" optimal solution of the problem, we finally choose $\mu^\star = \arg\min_\mu\{\mathcal{C}(\mu^{\star 1}), \mathcal{C}(\mu^{\star 2}), \mathcal{C}(\mu^{i^{(k)}_{\text{best}}})\}$.

2.4 Benefits of Advanced Information

In the previous section we developed a queueing network model for a maintenance service provider and a corresponding solution procedure that allows us to determine the optimal capacities in the production system. This constitutes the basis for an in-depth evaluation of the benefits associated with advanced information that can be obtained through collaborative maintenance management of the MRO service provider with his customers. More specifically, we want to determine the cost benefits of collaborative demand forecasting, spare parts management, and collaborative engine arrival scheduling as introduced in Section 2.1. As explained in the subsequent subsections, collaborative maintenance management may result in improved service rates and reduced service and interarrival time variabilities.[9]

For all three effects, we want to investigate the behavior of total costs for increasing improvement. That is, how does the slope of total costs evolve with the improvements? Do the benefits obtained through improvements at multiple work stations add up? What are the numerical effects on mean turnaround time? Can we find bounds on capacities and cost improvements in order to obtain good estimates of the minimum expectable benefits, even without solving the optimization problem? Are there interdependencies between the two types of variability reduction and which effects yields the higher benefits? These guiding managerial questions are answered as far as analytically possible for improved mean service rates and reduced service time variabilities in Sections 2.4.1 and 2.4.2, respectively. The effect of reduced interarrival time variabilities are explained in more detail in Section 2.4.2. A numerical study

[9]It is interesting to note that sole forecast information does not lead to a cost improvement. This means that knowledge about future arrivals has no value as long as the information is not used for service preparation, spare parts management, or scheduling overhaul time slots. Without further processing of the information, the service rate is not improved and service and interarrival time variabilities are not reduced (e.g., although we know future arrivals, they are still Poissonian from a queueing network point of view).

validating and enhancing the theoretical results is conducted in Section 2.5.

2.4.1 Mean Service Rate Improvement

In this section we want to determine the effects of improved service rates on the production network. We provide managerial insights regarding the benefits of capacity and penalty costs that can be expected as well as some structural properties regarding the updated optimal capacity allocation.

Until now the service provider did not receive any information regarding future engine arrivals and their condition. In the collaborative maintenance management scenario, customers grant access to historic and real-time engine usage and condition data that can be used as follows: Let us assume that there exists for each engine a set of usage parameters $x_u \in \mathbb{R}^{X_u}$ (e.g., traveled distance or number of takeoffs since last overhaul) and a set of condition parameters $x_c \in \mathbb{R}^{X_c}$ (e.g., oil temperature, oil pressure, or vibration levels). Using historic usage and condition data from overhauled engines and the associated overhaul activities, we can use machine learning techniques to forecast both the overhaul time and, based on the predicted condition parameters, the service that needs to be executed. Therefore, we can use a regression model and hard-time usage thresholds to forecast the overhaul time T and the related usage parameters x_u^T given the actual data collected until time t, $(T, x_u^T) = f(t, x_u^t)$. Additionally, let $P \in [0,1]^Z$ be the set of probabilities that part $z \in \{1, \ldots, Z\}$ needs to be replaced. With probabilistic decision tree learning we can forecast the probabilities p_z that spare part z will be needed for replacement (accordingly, the probability of regular overhaul and repair is given by $\bar{p}_z = 1 - p_z$) once the engine is delivered for overhaul given the actual and predicted usage and condition parameters, $P^T = g(x_u^T, x_c^T)$. With this information, it is possible to optimize spare parts management and to prepare the repair processes (e.g., guarantee that skilled personnel is available). This may reduce the time needed for inspection, repair and assembly by a factor ξ_j and hence increase the service rates at the affected work stations of the network. For more information we refer the reader to Taigel 2015 [43] and Zilli et al. 2015 [51].

The queueing network model defined in Section 2.3 now allows us to investigate the effects of mean service rate improvement on optimal capacity and total costs. We assume that the service rate improvement can be quantified by the service provider ex ante. We define a matrix $A \in \mathbb{R}_+^{J \times J}$ to account for the the service rate improvement. The elements on the main diagonal of A are given by $\xi_j \geq 1,\ \forall j \in \mathcal{J}$, all other elements are zero. The factor ξ_j represents the relative mean service rate improvement at work station j through collaborative forecasting of maintenance demand and preparation ($\xi_j = 1$ means no improvement, $\xi_j > 1$ indicates an improvement at the respective work station).

A myopic choice for the updated capacity vector would be $\widetilde{\mu} = A^{-1}\mu^\star$, i.e., dividing the original optimal capacities by the improvement factor. This operation would yield the optimal updated capacity if the production system consisted of a single work station and no penalty costs were incurred. However, due to the network structure (an improvement at one node can induce a change in the optimal updated capacities at preceding and subsequent work stations) and the nonlinear penalty cost term in our optimization model, we will later see that the myopic choice does not yield the optimal solution. Nevertheless, as penalty costs would remain constant when choosing $\widetilde{\mu} = A^{-1}\mu^\star$, i.e., $S_j(\xi_j \xi_j^{-1} \mu_j) = S_j(\mu_j)$,

$$\underline{\Delta \mathcal{C}} = c^\top \mu^\star - c^\top A^{-1} \mu^\star = \sum_{j \in \mathcal{J}} c_j \mu_j^\star (1 - 1/\xi_j) \tag{2.9}$$

provides a lower bound on total costs improvement.

In order to obtain the economically optimal capacity allocation, we define an updated capacity allocation problem (UCAP) which is a simple variation of the original (CAP) problem, thus enabling us to investigate the effects of mean service rate improvements on all work stations in the network, capacity, and penalty costs.

$$\text{minimize} \quad \widetilde{\mathcal{C}}(\mu) = c^\top \mu + \sum_{e \in \mathcal{E}} \gamma_e \Big[\sum_{j \in \mathcal{J}_e} S_j(\xi_j \mu_j) - S_e^T \Big]^+, \tag{UCAP}$$

with $\mathbf{dom}\ \widetilde{\mathcal{C}} = \{\mu \mid \xi_j \mu_j > \lambda_j,\ j = 1, \ldots, J\}$. Even without a formal proof it is straightforward that total costs will be reduced for any service rate improvement.

Proposition 2.4. *If there is a mean service rate improvement at any node, i.e.,* $\exists\, j \in \mathcal{J},\ \xi_j > 1$*, total costs will be reduced. Capacity costs are decreasing in the improvement,* $c^\top \widetilde{\mu} < c^\top \mu^\star$*, and penalty costs are weakly decreasing in the improvement,*

$$\sum_{e \in \mathcal{E}} \gamma_e \Big[\sum_{j \in \mathcal{J}_e} S_j(\xi_j \widetilde{\mu}_j) - S_e^T \Big]^+ \leq \sum_{e \in \mathcal{E}} \gamma_e \Big[\sum_{j \in \mathcal{J}_e} S_j(\mu_j^\star) - S_e^T \Big]^+ .$$

Proof. See Appendix A.2. □

It is interesting to note that total sojourn times are only reduced if a penalty was incurred for the original optimization problem's solution. Additionally, since penalty costs are weakly decreasing with improved service rates, a product family (or engine type) originally incurring a penalty can "loose" the penalty if the improvement is substantial.

As mentioned before, the updated optimal capacities do not correspond to the original capacities divided by the improvement, and an improvement at one work station can lead to a shift of the optimal capacity at a different node in the network. Therefore, the following proposition provides some structural insights regarding the updated optimal capacities.

Proposition 2.5. *For optimization problems where a penalty is incurred for all product families, capacities at work station without improvement will remain constant,* $\widetilde{\mu}_j = \mu_j^\star$*. For optimization problems where no penalty is incurred for any product families, capacities at work stations without improvement will be reduced,* $\widetilde{\mu}_j < \mu_j^\star$*. In any case, updated optimal capacities at work stations with improvement will be in the interval* $\widetilde{\mu}_j / \mu_j^\star \in (1/\xi_j, 1)$*.*

Proof. See Appendix A.2. □

From the proposition we can deduce some interesting results. If the solution of the original optimization problem is such that a penalty is incurred for all

product families, we can conclude that capacities remain constant at work stations without improvement, $\widetilde{\mu}_j = \mu_j^\star,\ \forall j$ where $\xi_j = 1$. This is due to the independence of the nodes when solving the *Case 1* optimization problem. As soon as at least one product family does not incur a penalty, the result is generally no longer valid. For product families without penalty, increases and reductions in service rate at the work stations will balance each other out such that mean approximate total sojourn times still coincide with contractually defined maximum mean sojourn times. Mean service rates will increase at work stations with improvement and decrease at work stations without improvement, although this may be counterintuitive at first sight. Independent of the penalty cost term, updated optimal capacities will always range between the myopic updated capacity choice and the optimal capacity of the original (CAP).

In Section 2.5.3 we conduct a numerical analysis of the structural properties of service rate improvements in order to provide further insights regarding our guiding managerial questions.

2.4.2 Service and Interarrival Time Variability Reduction

In this section we explore the effects of reduced interarrival and service time variabilities on the production network. We not only compare the impact of both types of variability reductions on total costs, but also provide some structural insights regarding the propagation of the effects through the network.

As outlined in the previous section, optimized spare parts management (i.e., higher service levels) and preparation of service (i.e., required material and skilled personnel are available) lead not only to increased service rates, but also to reductions in service time variabilities at the respective work stations. Despite potential correlations between improvements of mean service rates and service time variabilities, we can still investigate the sole effects of variability reduction on queueing network performance as we model the GI/G/1 work stations with the Krämer-Langenbach-Belz equation where service rate and variability are independent variables.

The scheduling of engine arrivals was another process step in the collaborative maintenance management scenario introduced in Section 2.1. Assume that all time slots of type e engine arrivals are labeled, $.., T_{e,n-1}, T_{e,n}, T_{e,n+1}, ..,$ and the n-th engine just arrived. Without scheduling, the engine arrival process is a stochastic process with high variance. For example, it could be modeled as a Poisson process where the time between two arrivals $\delta T_e = T_{e,n} - T_{e,n-1}$ is described as an exponentially distributed random variable with mean λ_e, $\mathbb{P}[\delta T_e \leq t] = 1 - \exp\{-\lambda_e t\}$. The squared coefficient of variation for exponentially distributed interarrival times is given by $ca_e = \lambda_e^2/\lambda_e^2 = 1$. Now assume we can forecast the $(n+1)$-th engine's arrival slot and determine a discrete set of arrival slot options around the predicted slot resulting from the aircraft-specific flight plan. Then we can schedule the arrivals such that the squared coefficient of variation of the interarrival times vanishes, i.e., engines arrive with constant interarrival times. It is unlikely that this process could be managed on the customers' side, as the service provider is likely to have multiple customers with the same engine types, meaning that successive engines could stem from different customers. For more information we refer the reader to Zilli et al. 2015 [51].

Proposition 2.6. *For a single work station j with constant service rate, the following statements hold.*

i) The sojourn time $S_j(\cdot)$ is strictly increasing in cs_j and ca_j.

ii) Ceteris paribus, the cost reduction through improved interarrival time variability exceeds the cost reduction through improved service time variability.

Proof. See Appendix A.2. □

Statement i) tells us that sojourn times are reduced if interarrival or service time variabilities are reduced. Therefore, it is obvious that total costs will be reduced by any improvement of service or interarrival time variability.

While the impact of a reduction in ca_j and cs_j on sojourn times and total costs is intuitive, an investigation of the strength of the two effects from a

queueing network perspective yields some interesting insights. As stated in *ii*), when we consider one work station in isolation, the cost reduction through reduced interarrival time variability exceeds the cost reduction through reduced service time variability considering one work station in isolation.

On the other hand, if we consider two subsequent work stations j and $i = j + 1$, assuming there is no superposition or splitting between the two work stations, we can derive which properties drive the impact of service and interarrival time variabilities from a queueing network perspective. If we take the derivatives of the interdeparture time variability equation $cd_j = ca_i = \rho_j^2 cs_j + (1 - \rho_j^2) ca_j$ with $\rho_j = \lambda_j / \mu_j$ (see Appendix A.3 for further reference) with respect to cs_j and ca_j, $\partial ca_i / \partial cs_j = \rho_j^2$ and $\partial ca_i / \partial ca_j = 1 - \rho_j^2$, we find that the effect of reduced service time variability at work station j on work station i is larger than the effect of reduced interarrival time variability for $\rho_j > \sqrt{0.5}$. The opposite holds if $\rho_j < \sqrt{0.5}$. Therefore, the traffic intensity ρ_j is an important measure that determines the propagation of variability reduction through the network. Additionally, we observe that the benefits of reduced interarrival time variabilities increase for lower service time variabilities, meaning that maintenance demand forecasting and optimized planning foster the benefits of scheduling.

Due to the interdependence of work stations in a queueing network, reductions of service or interarrival time variabilities will always have an impact on the interarrival time variabilities of all successive work stations in the production process. On the other hand, a reduction of the external interarrival time variabilities has no influence on the service time variabilities at the nodes. Because of the complexity of the queueing network performance parameter algorithm detailed in Appendix A.3, in the subsequent section we numerically illustrate how an improvement of the variabilities affects the capacities and costs of the optimization problem. Subsection 2.5.4 contains a numerical analysis of the structural properties of variability reductions in order to shed some light on the managerial questions defined at the beginning of this section.

2.5 Numerical Analysis

The following numerical analysis sheds light on the actual numerical behavior of the models introduced above. As we already derived some structural insights, numerical examples validate those insights and provide a more detailed evaluation of the actual performance and expectable benefits.

The experimental design of the numerical collaborative maintenance management scenario analysis is explained in the subsequent section. In Section 2.5.2 we define the reference case of our numerical analysis, including a comparison of the different solution methods introduced in Section 2.3. Numerical analyses regarding the benefits of the collaborative maintenance management scenario are provided in Section 2.5.3.

2.5.1 Experimental Design

After a more analytical treatment of the effects obtained through advanced information in Section 2.4, we now conduct a series of factorial analyses in order to answer some basic managerial questions.

For all three effects, we investigate the actual numerical behavior of total costs for increasing improvement. We address managerial questions such as: How does the slope of total costs evolve with the improvements? Do the benefits obtained through improvements at multiple work stations add up? What are the numerical effects on mean turnaround time? For improved service rates, we are also interested in the total costs benefit in comparison to the lower bound of cost improvement derived in Section 2.4.1. We also investigate how capacities at the work stations change with improved service rates, and check this against the structural insights provided in Proposition 2.5. Finally, we conduct a factorial analysis regarding the effects of mutual dependence of service and interarrival time variability improvements. Figure 2.5 depicts the input and output parameters of our analysis.

Mean service rates are improved through advanced demand information and increased spare parts service levels. If we consider the network shown in Figure 2.1, the NDT (Non-Destructive Test) and inspection process steps are

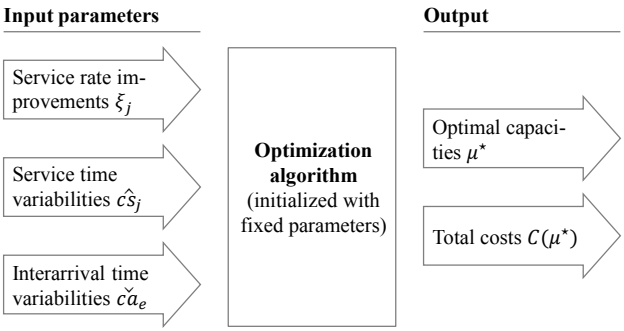


Figure 2.5: Input and output parameters.

accelerated, since more information is known in advance through continuous sharing of real time sensor data. Optimized spare parts management and service preparation lead to accelerated sojourn times of the repair and assembly process steps. Therefore, we assume an improvement of the mean service rates at work stations $j \in \{4, \ldots, 9\}$ in the range of $\xi_j = [1, 1.5]$, whereas $\xi_j = 1,\ \forall j \in \mathcal{J} \setminus \{4, \ldots, 9\}$.

Service time variabilities are also reduced due to optimized spare parts management and service preparation. Therefore, we investigate the benefits at the corresponding repair and assembly work stations $j \in \{7, 8, 9\}$ with $\hat{cs}_j = [cs_j/2, cs_j]$.

Finally, we assume collaborative scheduling of engine arrivals such that the interarrival time variabilities are minimized. As the improvement range we select $\check{ca}_e = [0.1, 1],\ \forall e \in \{1, 2\}$. This means that we start with Poissonian demand and end up with close to constant interarrival times.

In the following section we define the reference case used throughout the numerical experiments conducted in Sections 2.5.3 and 2.5.4.

2.5.2 Reference Case

In this section we define the queueing network parameters used for the numerical experiments. Additionally, we investigate the performance of the three proposed solution methods for four different sets of sojourn time requirements, one of which is chosen as the reference case for the factorial analyses conducted

in the following section.

According to the production network of the service provider setting depicted in Figure 2.1, we use a queueing network with $J = 11$, i.e., 11 work stations and two paths through the network, $\mathcal{J}_1 = \{1, 3, 4, 5, 7, 8, 10, 11\}$ and $\mathcal{J}_2 = \{2, 3, 4, 6, 7, 9, 10, 11\}$ for engine types $e \in \{1, 2\}$, respectively. The external arrival rates for the two engine types are given by $\lambda_1 = 5$ and $\lambda_2 = 8$. The squared coefficients of variation of the external arrivals are given by $ca_1 = ca_2 = 1$, as we assume Poisson processes without arrival scheduling. As cost parameters we choose $c_j = 1,\ \forall j \in \mathcal{J}$, and $\gamma_1 = 20$, $\gamma_2 = 22$.

The queueing network parameters of the individual nodes resulting from the computations described in Appendix A.3 are summarized in Table 2.1. The values for ca_j were obtained using an initial guess of $\mu_j^\star = \lambda_j + 2$.

Node	1	2	3	4	5	6	7	8	9	10	11
λ_j	5	8	13	13	5	8	13	5	8	13	13
cs_j	0.044	0.044	0.044	0.117	0.064	0.064	0.753	0.283	0.283	0.028	0.028
ca_j	1.000	1.000	0.470	0.150	0.262	0.211	0.187	0.359	0.454	0.254	0.084

Table 2.1: Queueing network parameters used for the numerical analysis.

We start our numerical analysis by comparing the three solution approaches defined in Section 2.3.3. We evaluate four scenarios of the production network with different contractually defined maximum mean sojourn times in order to evaluate the performance of the solution methods depending on the applying case. The numerical solutions are provided in Table 2.2.

Scenario	***Case 1***	***Case 2***	***Case 3***
a) $S_1^T = 0.7,\ S_2^T = 0.7$	$\boldsymbol{\mathcal{C}(\mu^{\star 1}) = 150.3}$	$\mathcal{C}(\mu^{\star 2}) = 160.9$	$\breve{\mathcal{C}}^{(k)}_{\text{best}} = 150.3$
b) $S_1^T = 1.4,\ S_2^T = 1.4$	$\mathcal{C}(\mu^{\star 1}) = 135.7$	$\boldsymbol{\mathcal{C}(\mu^{\star 2}) = 125.5}$	$\breve{\mathcal{C}}^{(k)}_{\text{best}} = 125.6$
c) $S_1^T = 0.7,\ S_2^T = 1.4$	$\mathcal{C}(\mu^{\star 1}) = 143.4$	$\mathcal{C}(\mu^{\star 2}) = 149.5$	$\boldsymbol{\breve{\mathcal{C}}^{(k)}_{\text{best}} = 138.4}$
d) $S_1^T = 1.4,\ S_2^T = 0.7$	$\mathcal{C}(\mu^{\star 1}) = 142.7$	$\mathcal{C}(\mu^{\star 2}) = 146.8$	$\boldsymbol{\breve{\mathcal{C}}^{(k)}_{\text{best}} = 138.4}$

Table 2.2: Solution of the capacity allocation problem with different maximum mean total sojourn times and methods (optimal solution highlighted in bold).

We see that *Case 1* applies for low S_e^T (where the objective function

is differentiable), *Case 2* applies for high S_e^T (where the penalty cost term is not differentiable on all paths), and the subgradient method produces optimal results for partially low, partially high S_e^T (where some paths are differentiable, some not). As observable for the first scenario with $S_1^T = S_2^T = 0.7$, the *Case 1*-optimal solution and the solution of the subgradient method (with diminishing step size rule) coincide. This is due to the fact the algorithm is guaranteed to converge to the optimal value for differentiable functions, i.e., $\lim_{k\to\infty} f(x^{(k)}) = f^\star$ [11].

Since the maximum mean total sojourn times S_e^T are not considered in *Case 1*, the actual *Case 1*-optimal mean total lead times are constant throughout all scenarios. Given the scenarios in Table 2.2, the mean total lead times are given by $\sum_{j\in\mathcal{J}_1} S_j(\mu_j^{\star 1}) = 1.08 > S_1^T$ and $\sum_{j\in\mathcal{J}_2} S_j(\mu_j^{\star 1}) = 1.02 > S_2^T$. Knowing these values, we instantly see which case applies for which numerical values of S_e^T. If $S_1^T \leq 1.08$ and $S_2^T \leq 1.02$, *Case 1* applies. If $S_1^T > 1.08$ and $S_2^T > 1.02$, *Case 2* applies. For all other cases where one contractually defined maximum mean total sojourn time is smaller and one larger than for the *Case 1*-optimal solution, *Case 3* applies. The scenarios and the respective cases are illustrated in Figure 2.6.

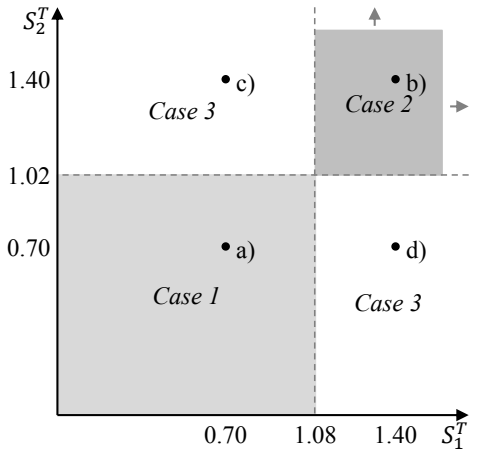


Figure 2.6: Parameter map with case boundaries and scenarios.

For all paths where $S_e^T > \sum_{j\in\mathcal{J}_e} S_j(\mu_j^{\star 1})$, i.e., where the global solution is not the *Case 1*-optimal solution, the optimal solution $\mu^\star$ will always be such that $\sum_{j\in\mathcal{J}_e} S_j(\mu_j^\star) = S_e^T$. This means that capacities are chosen such that

capacity costs are minimal under the constraint that contractually defined and actual mean total sojourn times coincide.

In order to investigate the effects of collaborative maintenance management on total costs of aircraft engine overhaul services, we choose option a) in Table 2.2 with $S_1^T = S_2^T = 0.7$ as reference scenario. Without improved spare parts management and preparation, i.e., $\xi_j = 1$ and $\hat{cs}_j = cs_j$, $\forall j \in \mathcal{J}$, and without collaborative scheduling of engine arrivals, i.e., the arrival process is assumed to be Poissonian with $ca_e = 1$, $\forall e \in \mathcal{E}$, total costs are given by 150.3. For the solution of the updated optimization problem we use the subgradient method-based procedure defined for *Case 3* without any differentiability requirements.

2.5.3 Mean Service Rate Improvement

In this section we investigate the effects of mean service rate improvement on total costs. Figure 2.7 shows total costs for increasing service rate improvement factors ξ_j. The lowest thick line in the interval $\widetilde{\mathcal{C}} \in [126.9, 150.3]$ is

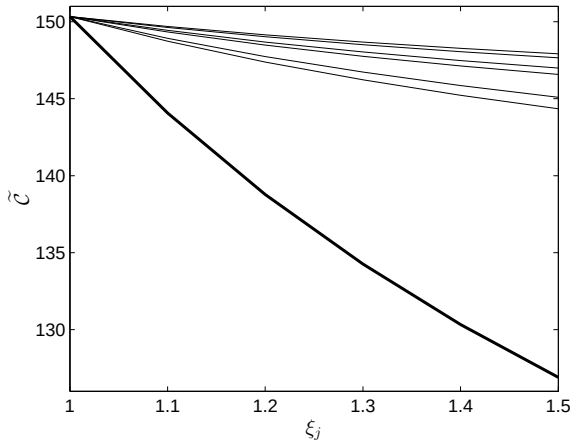


Figure 2.7: Numerical study for improved service rates $\xi_j \mu_j$.

generated by increasing all factors for work stations $j = \{4, \ldots, 9\}$ from 1.0 to 1.5 simultaneously. The six upper thin lines represent total costs evolution when only increasing the improvement factors for single nodes. Note that the

lowest two of the thin lines correspond to work stations 4 and 7, which are the only work stations in the subset which serve both engine types.

If we compare optimal service rates $\mu^\star$ and $\widetilde{\mu}$ for $\xi_j = 1,\ \forall j \in \mathcal{J}$ and $\xi_j = 1.5,\ \forall j \in \{4, \ldots, 9\}$, respectively, we obtain

$$\widetilde{\mu}./\mu^\star = [1.00 \ \ 1.00 \ \ 1.00 \ \ 0.69 \ \ 0.72 \ \ 0.70 \ \ 0.70 \ \ 0.72 \ \ 0.71 \ \ 1.00 \ \ 1.00]^\top$$

where $\widetilde{\mu}./\mu^\star$ is the element-wise division. This verifies the structural insight proposed in Section 2.4.1 for product families incurring a penalty: capacities remain constant at work stations j without improvement, $\widetilde{\mu}_j = \mu_j^\star$ if $\xi_j = 1$, whereas updated optimal service rates are in the interval $\widetilde{\mu}_j/\mu_j^\star \in (1/1.5, 1) = (0.\bar{6}, 1)$ for work stations j where an improvement $\xi_j = 1.5$ is imposed.

Total costs reduction for maximum improvements at all work stations is given by $\Delta\mathcal{C} = \mathcal{C}(\mu^\star) - \widetilde{\mathcal{C}}(\widetilde{\mu}) = 150.3 - 126.9 = 23.4$. If we compute the lower bound (2.9) derived in Section 2.4.1, we obtain $\underline{\Delta\mathcal{C}} = 22.8$. Therefore, we can conclude that the lower bound is tight and provides an accurate initial estimation of total costs reduction through service rate improvement.

2.5.4 Service and Interarrival Time Variability Reduction

Figure 2.8 shows total costs for improved service time variabilities. Again,

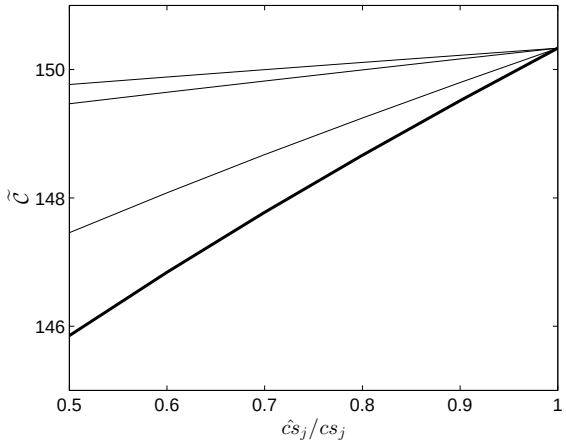


Figure 2.8: Numerical study for improved service time variability $\hat{cs}_j$.

the thick line in the interval $\widetilde{\mathcal{C}} \in [145.9, 150.3]$ corresponds to the improvement of $\hat{cs}_j$ from cs to $cs_j/2$ simultaneously for all work stations $j \in \{7, 8, 9\}$ and the thin lines to the improvement at single work stations. The largest improvement (i.e., the lowest thin line) is found at work station 7, since this work station is visited by both product families and exhibits the largest service time variability without improvement as presented in Table 2.1. Node 7 models the assembling work station, where spare parts delivered from external suppliers are needed. Through collaborative forecasting of spare parts demand, the spare parts service levels can be increased, which explains the imposed reduced service time variability at the work station.

Finally, Figure 2.9 displays the effects of reduced interarrival time variabilities. The thick line corresponding to a simultaneous improvement of

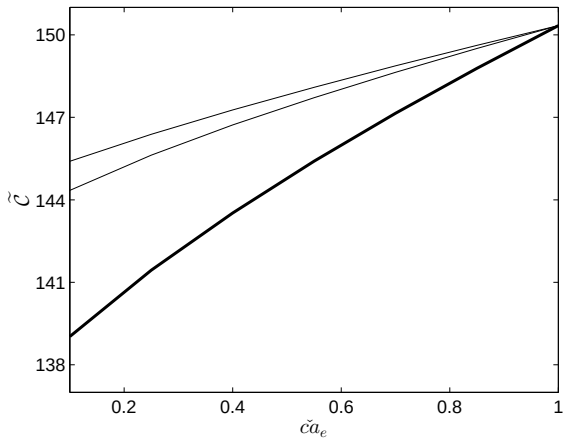


Figure 2.9: Numerical study for improved interarrival time variability $\check{c}a_e$.

$\check{c}a_e$, $\forall e \in 1, 2$, from 1 to 0.1 shows the highest cost reduction, ranging from 139.0 to 150.3. When considering single engine types, the cost reduction through improved $\check{c}a_2$ is superior (the lower of the two thin lines) compared to the cost reduction through improved $\check{c}a_1$ as mean arrival rate and unit penalty costs in the numerical example are comparatively higher, $\lambda_2 = 8 > \lambda_1 = 5$ and $\gamma_2 = 22 > \gamma_1 = 20$, respectively.

It is interesting to note that total cost reductions do not add up for improved service and interarrival time variabilities. However, they do add up for

improved service rates due to the independence of the nodes when solving the differentiable optimization problem where penalties are incurred on all paths. For improved service and interarrival time variabilities, the effect of reducing variabilities at multiple work stations simultaneously exceeds the sum of the effects when only reducing the variability at one work station (in our analysis by approximately 4% for service and interarrival time variabilities).

2.5.5 Summary of Numerical Analysis

In order to investigate whether the three effects of advanced information sum up, different combinations of specific improvements from the ranges defined in Section 2.5.1 are summarized in Table 2.3. These are: mean service rate improvements $\xi_j = 1.5,\ \forall j \in \{4, \dots, 9\}$ in column $A\mu$, service time variability reductions $\hat{cs}_j = cs_j/2,\ \forall j \in \{7, 8, 9\}$ in column $\hat{cs}$, and interarrival time variability reductions with $\check{ca}_e = 0.1,\ \forall e \in \{1, 2\}$ in column $\check{ca}$.

$A\mu$	$\hat{cs}$	$\check{ca}$	$\sum_{j\in\mathcal{J}_1} S_j$	$\sum_{j\in\mathcal{J}_2} S_j$	$\widetilde{\mathcal{C}}(\tilde{\mu})$	Cost reduction
-	-	-	1.08	1.02	150.3	0%
•	-	-	1.00	0.94	126.9	16%
-	•	-	1.05	0.98	145.9	3%
-	-	•	1.01	0.92	139.0	8%
•	•	-	0.98	0.91	123.2	18%
•	-	•	0.94	0.85	116.3	23%
-	•	•	0.98	0.88	134.4	11%
•	•	•	0.92	0.82	112.5	25%

Table 2.3: Numerical analysis of combinations of improvements.

As already shown in Figure 2.7, an improvement in mean service rates has the highest impact on total costs. As expected, with Proposition 2.6 in mind, the collaborative scheduling of engine arrivals exhibits the second-highest cost reduction. The reduction of service time variabilities has the smallest effect on total costs. On the other hand, in many cases this improvement will happen automatically when mean service rates are improved, and therefore requires

little additional implementation costs. In the columns displaying the sums of the approximate mean sojourn times on paths $\mathcal{J}_1$ and $\mathcal{J}_2$ we see that not only total costs, but also mean turnaround times are reduced as stated in Proposition 2.5.

Finally, it can be observed from the column showing optimal total costs $\widetilde{\mathcal{C}}(\widetilde{\mu})$ that the benefits of the three effects of advanced information do not add up. For any combination of mean service rate improvement with another effect, the sum of the individual improvements exceeds the improvement when solving the optimization problem for the combination. In contrast, when solving the optimization problem for reduced service and reduced interarrival time variabilities, the improvement is larger than the sum of the individual benefits (15.9 versus 15.7, respectively). This supports our statement in Section 2.4.2 that maintenance demand forecasting and optimized planning foster the benefits of scheduling.

2.6 Conclusion and Discussion

This paper solved the problem of determining the cost-minimizing capacity in an acyclic production network for a service provider with contractually defined lead time requirements and associated penalty costs. The production network was described as a network of GI/G/1 queues, and we were able to find analytical and iterative methods to solve the optimization problem through classification of different cases distinguished by the differentiability of the objective function. For differentiable objective functions we were able to solve the optimization problem with the first-order necessary condition. If the objective function is not differentiable on all paths products take through the network, the problem can be solved with the Karush-Kuhn-Tucker conditions of an equivalent reformulation of the optimization problem. Finally, we developed a general near-optimal solution algorithm based on the subgradient method capable of solving the problem with any properties.

Additionally, we considered a collaborative maintenance management scenario in the aerospace business. In this scenario, the service provider and cus-

tomers jointly employ advanced information collected by sensors measuring aircraft engine parameters in order to further minimize maintenance costs. As the associated collaborative planning systems are an investment for all parties involved, we showed how to compute the expected benefits under improvement assumptions such as improved service rates and reduced service and interarrival time variabilities. Numerical analyses illustrated the findings, allowed us to derive further structural properties, and supported investment decisions regarding a collaborative planning system.

With the information displayed in Table 2.3 and knowledge regarding implementation costs of the collaborative planning system, the service provider and his customers can now decide whether to consider the investment or discard the opportunity. Of course, to do so, a benefit sharing scheme must be developed to incentivize the individual players to participate: whereas capacity and penalty costs are reduced for the service provider, a fair share of the benefits would need to be transferred to the customers through a price reduction. The proposed methods allow the parties to establish a fair benefit sharing rule (depending on the implementation costs per party) and to quickly assess the cost-effectiveness of the collaborative planning system. Customers also benefit from reduced mean turnaround times, which translates into a reduction of the total number of engines needed for flight operations.

We provided some preliminary ideas regarding collaborative demand forecasting, spare parts management and engine arrival slot scheduling. We continue to work on more formal models for collaborative maintenance management. Although this paper was developed specifically for the aircraft engine MRO case, the models and insights apply to any similar service provider setting.

3 Flexible Capacity Management with Future Information

We consider a maintenance service provider that overhauls aircraft engines in a central service facility. The service provider can choose between a low and a high capacity. However, the high capacity can only be used for a certain share of the time. We model the service facility as a queue with adaptable service rate and develop capacity control policies minimizing the time-average queue length. The reactive threshold-type policy only considers the current queue length and we show that the time-average queue length diverges as the load of the system increases. State-of-the-art aircraft engines are equipped with sensors permitting prediction of future maintenance needs. Thus, the solely forward-looking policy is based on predicted arrival times and maintenance requirements of engines arriving to the system in the future. We show that the time-average queue length converges to a finite value as the load increases. As the reactive policy outperforms the solely forward-looking policy for low arrival rates, we combine both to a proactive policy. While we assume that future information is available within an infinite lookahead window when stating the policy, we derive conditions under which a proactive capacity control policy can be employed and how the policy has to be modified if the lookahead window is finite.[10]

[10]This paper has been submitted for publication [27]. It is coauthored by Richard Pibernik.

3.1 Introduction and Outline

Variable, but predictable. The idea of Spencer et al. (2014) is that future arrivals to a queue are variable, but that they can be predicted [36]. Information regarding the arrival times and the service requirements of jobs arriving to a queue in the future is referred to as *future information*. In this paper, we investigate how to optimally use future information for the control of a flexible capacity.

Our analysis is motivated by the capacity management problem faced by a provider of maintenance, repair and overhaul services (MRO) for aircraft engines. The MRO can switch between a low (base) and a high (base plus contingent) capacity. Both the base and the contingent capacity have been determined based on historical arrival rates of engines and their historical service requirements. Furthermore, the contingent capacity can only be employed for a predefined share of time, such that the time-average capacity (or service rate) is at most one. Engine arrivals, on the other hand, occur at a rate of less or equal than one. Until now, the decision to employ the contingent capacity has been taken in an ad-hoc manner, mainly based on the current number of engines awaiting service and subjective judgments of the operations managers. Future information in the aforementioned sense has not been available for capacity control. This, however, is about to change. Sophisticated technical equipment such as Rolls-Royce's Trent aircraft engines are equipped with sensors that measure different parameters and send the data to data warehouses in real-time. For Trent engines, this data includes more than 20 parameters such as oil pressure, oil temperature and vibration levels that—in combination with usage data of the airlines—can be used to accurately predict future maintenance needs, i.e., time of maintenance and service requirements [25]. Thus, in future, engine arrivals will be *variable, but predictable*. For the MRO, the question arises how to best utilize this future information to effectively employ its flexible capacity. More specifically, the task is to develop an optimal capacity control policy for *variable, but predictable* job arrivals and to assess its performance relative to a *reactive policy*, which is currently applied. This is the objective of the research presented in

this paper.

Following Spencer et al. (2014), we assume that (the amount of) future information can be characterized by a lookahead window for which the company has perfect information about the jobs arriving to the system. Currently, the company has a lookahead window of length zero, that is, the company has no future information. Job arrivals are random and the decision to activate the contingent capacity is based on the current queue length. A lookahead window of infinite length would imply that the company knew all future jobs arriving to the system when determining whether or not to employ the contingent capacity. At a first glance it may appear rather simple to determine the optimal (sequence of) decisions with respect to the contingent capacity if all future jobs are known. Determining an optimal capacity control policy and the resulting time-average queue length in advance, however, is not trivial. Although all future orders will be known at some point in time t_0, their exact realizations are not known when implementing a capacity control policy (before t_0). Thus, the task is to devise an optimal capacity control policy for uncertain future arrivals of orders given that at some point in time their realizations will be revealed. This reflects the notion of *variable, but predictable*.

In this paper, we develop and study two capacity control policies for the case of an infinite lookahead window. For a capacity decision at time t_0, the first policy (*solely forward-looking policy*) only utilizes future information about engine arrivals. Based thereupon, we develop a second policy (called the *proactive policy*) which utilizes both future information about engine arrivals and the current queue length at time t_0. We compare the performance of these two policies with the performance of the reactive policy which only uses information about the current queue length—the reactive policy serves as a proxy for the MRO's current practice. We assume that the MRO intends to minimize the time-average waiting time of the customers for the completion of their orders. Since it can be shown that the time-average waiting time is directly linked to the time-average queue length, we use the latter as our performance criterion to assess the different policies. As detailed in Section 3.4, we observe that the time-average queue length diverges if no future informa-

tion is available as the arrival rate approaches the time-average service rate. On the other hand, if we assume to know future information until infinity, we find that the time-average queue length converges as the load of the system approaches one. However, as soon as the lookahead window is smaller than infinity, the time-average queue length diverges as the arrival rate approaches the time-average service rate.

Studying the extremes with respect to the availability of future information (lookahead window of zero versus an infinite lookahead window) brings a number of advantages: it allows us to i) develop analytically tractable properties of the different capacity control policies, ii) carry out performance comparisons and iii) shed light on the value of future information. Of course, from a practical perspective, a finite lookahead window (the company only knows order arrivals over a shorter period of time, e.g., the next four weeks) is more realistic and relevant. Naturally, the performance of a capacity control policy that uses future information will strongly depend on the length of the lookahead window. In this paper, we first focus on analytically tractable operating modes, i.e., the reactive mode and the proactive mode with infinite lookahead window. Based thereupon, we derive conditions under which a proactive policy can be used for the case of a finite lookahead window. Also, we provide insights how the policy could be modified if these conditions are not met. Thus, our results establish a basis for the further development of optimal capacity control policies with limited future information.

The remainder of this paper is organized as follows: We provide a literature review in Section 3.2. In Section 3.3, we define the basic queueing model and some stochastic primitives. The reactive threshold-based policy is developed in Section 3.4 and the solely forward-looking policy with infinite lookahead in Section 3.5. Both policies are combined to a proactive policy with infinite lookahead in Section 3.6. First insights regarding the effects of limited future information are derived in Section 3.7. Finally, Section 3.8 provides concluding remarks and opportunities for further research. All proofs not stated in the main part of the document are relegated to the appendix.

3.2 Related Literature

Recently, the topic of queuing with future information has received increasing attention. Spencer et al. (2014) consider information about future job arrivals to determine optimal admission control policies for an M/M/1 queue in the overload regime [36]. They show that a finite time-average queue length can be achieved if future information is available in an infinite lookahead window, also as the load of the system approaches one. This paper is the one most closely related to our research. Xu and Chan (2016) use future information to reduce waiting times in an emergency department via diversion [49]. They combine an online and an offline policy to optimally divert patients for all arrival rates. Additionally, they perform a numerical analysis and argue that future information is also valuable if the information regarding future jobs arriving to the system is noisy. Xu (2015) investigates the amount of future information needed in order for the time-average queue length to converge as the load of the system approaches one [48]. He finds that the amount of future information needed increases with increasing arrival rate and a finite time-average queue length for a system with a load approaching one can only be achieved with an infinite lookahead window. Finally, Zhang (2014) develops models to proactively serve jobs that have not yet arrived to the system but can be observed in a finite prediction window. The focus of his work is on M/M/1 and GI/G/1 queues in light traffic.

Queuing with adaptable service rates and without future information is a problem that has been studied extensively. Bekker et al. (2008) develop various models for M/M/1 and M/G/1 queues in which the server can work at two different speeds, depending on the number of customers in the system or the workload [3]. Bekker et al. (2011) study two models [4]: First, a model that continuously adapts the service rate based on the waiting time of the first customer in line. Secondly, a model with a primary server that is supplemented by a secondary server if the waiting time of the first customer in line exceeds a certain threshold. They derive the steady-state waiting time distributions for both models. Lee et al. (2006) develop an (m, M) control rule for M/M/1 and M/G/1 queues [28]. When the workload exceeds M, a

higher service rate is used. When the workload falls below $m < M$, the system switches back to the lower service rate. Their approach avoids high switching rates.

Flexible queuing architectures are also considered for make-to-stock and make-to-order capacity management: Buyukkaramikli et al. (2013) model a production system that operates under a lead time performance constraint as M/M/1 queue with periodically adjusted service rate [15]. They find the optimal capacity levels, the capacity control policy and compute the optimal period lengths minimizing capacity costs while obeying a lead time service level. Allon and van Mieghem (2010) develop a GI/G/1 inventory shortfall model using dual-drift reflected Brownian motion to determine an optimal base-surge capacity control policy (China versus Mexico sourcing–low cost and slow versus high cost and fast) [1]. Similarly, Bradley (2004) develops a capacity control policy for an inventory shortfall queue modeling in-house production and a subcontractor, also based on reflected Brownian motion [13].

Finally, there is a stream of research considering (capacity planning for) MRO service providers. Kurz (2015), for example, models a maintenance service facility as a network of GI/G/1 queues and determines optimal capacities with the objective of minimizing the sum of capacity costs and penalty costs for not meeting contractually defined turnaround times [25]. In addition, Kurz (2015) investigates the effects of advanced information on service requirements and service and interarrival time variability.

Our work contributes to the research in queuing theory that addresses the problem of managing a flexible capacity. While our approach is inspired by Spencer et al. (2014) who introduced the notion of *variable, but predictable* job arrivals, we are, to the best of our knowledge, the first to develop flexible capacity control policies incorporating future information.

3.3 Setup and Problem Definition

We model the facility of the service provider as an M/M/1 queue with adjustable service rate. With $r \in (0, 1)$ and $p > r$, define the low and high service

rates as $\mu_1 = 1 - r$ and $\mu_2 = \mu_1 + p = 1 - r + p$, respectively. The arrival rate is given as $\lambda \in (1 - r, 1]$, i.e., we are considering the overload regime.[11] The M/M/1 queue with flexible capacity is illustrated in Figure 3.1.

Figure 3.1: Illustration of the flexible capacity model.

Let the capacity control policy be denoted as $\{\pi(t) : t \in \mathbb{R}_+\}$, $\pi(t) \in \{0, 1\}$. Then,

$$\mu(t) = \begin{cases} 1 - r, & \text{if } \pi(t) = 0, \\ 1 - r + p, & \text{if } \pi(t) = 1, \end{cases}$$

and we can define the time share the high capacity is active as

$$\mathbb{E}\pi = \lim_{T \to \infty} \frac{1}{T} \int_0^T \pi(t) \, \mathrm{d}t \in [0, 1].$$

In many practical applications, there exists an upper bound on the time the contingent capacity can be used. For example, overtime is limited by labor legislation or, if the capacity is sourced externally, amount and expected usage are specified contractually. Thus, for the remainder of the paper, we will focus on capacity control policies obeying the following capacity constraint.

Definition 3.1. *A capacity control policy π is called feasible, if the time share the contingent capacity is active $\mathbb{E}\pi$ does not exceed the ratio r/p. Π denotes the family of all feasible capacity control policies.*[12]

[11] Taking into account future information is especially interesting in the overload regime. Although future information could also be used if the arrival rate was below the base capacity, benefits would be limited as the time-average queue length of a system with low load tends to be quite small anyways.

[12] We use this specific upper bound for simplicity and readability, however, without loss of generality. All results can easily be transferred to cases where the upper bound is different from r/p, as long as an upper bound is provided.

With this definition, the time-average service rate can be at most 1,

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mu(t)\,\mathrm{d}t = (1-\mathbb{E}\pi)(1-r)+\mathbb{E}\pi(1-r+p)$$
$$\leq \frac{p-r}{p}(1-r)+\frac{r}{p}(1-r+p)=1.$$

Clearly, any policy must be such that $\mathbb{E}\pi \to r/p$ as $\lambda \to 1$, which is the most critical and therefore interesting regime. However, the performance, in terms of the time-average queue length, of the flexible capacity management policy can vary depending on when the contingent capacity is activated. If we, for example, deploy the contingent capacity when the system is empty or when the queue is already very long, performance may be worse than for a static model with $\mu = 1$. On the other hand, if we deploy the contingent capacity during a time with a very large number of arrivals, we can minimize the time-average queue length.

When developing capacity control policies, we will always try to minimize the time-average queue length while obeying the feasibility constraint. Let $\{Q[n] : n \in \mathbb{Z}_+\}$, $Q[n] \in \mathbb{Z}_+$, be a discrete-time queue length process of the M/M/1 queue with flexible capacity.

Definition 3.2. *Given a capacity control policy π,*

$$\mathcal{Q}(r,p,\lambda,\pi) = \limsup_{N\to\infty}\mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}Q[n]\right)$$

defines the time-average queue length, $\mathcal{Q}(r,p,\lambda,\pi) \in \mathbb{R}_+$.

We try to minimize the time-average queue length because, especially in the MRO business, turnaround times are extremely critical and long waiting times may lead to high penalties, see Kurz (2016) for more details [25]. The time-average waiting time is directly linked to the time-average queue length by Little's law. Thus, the company that motivates our research project has installed a costly contingent capacity. Now, we want to determine the "best usage" of this contingent capacity to minimize the mean waiting time of the customers' jobs.

Following Spencer et al. (2014), we define future information based on a lookahead window: If we are at time t_0, a lookahead window of length $w \in \mathbb{R}_+$ implies that we know the exact arrival times and service requirements (time needed to service the job) of all jobs arriving within the time interval $[t_0, t_0 + w]$. Depending on the length of the lookahead window, we distinguish four different operating modes and thus policies:

i) $w = 0$, no future information is known—capacity decisions are made based on the current queue length (*reactive policy*).

ii) $w = \infty$, we have perfect information regarding all jobs arriving to the system in the future:

   a) Capacity decisions are made only considering information regarding future jobs arriving to the system (*solely forward-looking policy*).

   b) Capacity decisions are made based on the current queue length and information regarding future job arrivals (*proactive policy*).

iii) $0 < w < \infty$, future information is available within a finite lookahead window—capacity decisions are made based on current queue length and limited future information (*proactive policy with limited future information*).

The research questions we try to answer in this paper can be summarized as follows: What is the feasible capacity control policy minimizing the time-average queue length of an M/M/1 with two capacity levels and a lookahead window of length w? What is the resulting time-average queue length and how does it behave depending on the arrival rate?

In the remainder of this paper we will not assign a specific size (service time) to each job, but use a service token model to capture the randomness in the service times of different jobs. In a service token model, jobs' service times are induced by the randomness of the speed of the server as an exogenous process. This allows us to compute the time-average queue length without considering the underlying workload process. In the service token model, jobs wait in a queue with infinite waiting room. Once a service token is generated

by the server, it is consumed by the first job in line which then leaves the system. Thus, the queue length coincides with the number of jobs in the system, as illustrated in Figure 3.2. Additionally, the resulting queue length

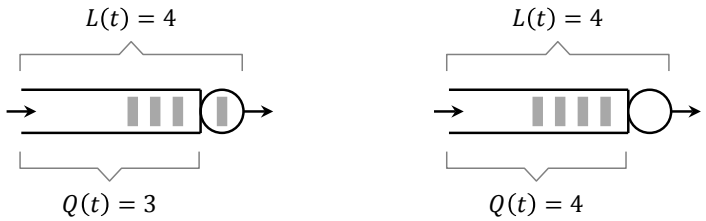


Figure 3.2: Illustration of the service token model: The figure on the left hand side displays a standard M/M/1 queue, where one job is currently served at the server. Thus, as this job is not accounted for as "in the queue", the number of jobs in the queue is given as $Q(t) = L(t) - 1$, where $L(t) \in \mathbb{Z}_+$ denotes the number of jobs in the system. The right hand side represents the equivalent service token-based M/M/1 queue. Here, all jobs are waiting in the queue until a service token has been produced. Thus, $L(t) = Q(t)$.

process is insensitive to the order in which jobs are completed.[13] Xu and Chan (2016) show via simulation that policies developed based on a service token model perform well and the resulting time-average queue length serves as a good approximation for settings with job-specific service times [49].

3.4 Reactive Capacity Control

We first consider the setting in which the company does not have information about future jobs and the decision to deploy the contingent capacity is only based on the current queue length. In the following, we develop an optimal reactive capacity control policy and derive an expression for the corresponding time-average queue length. The reactive policy will serve as a benchmark to evaluate capacity control policies that incorporate future information.

If the current queue length is the only information available, the company will naturally employ a threshold-type capacity control policy to decide when

[13] When having information about future arrivals and job-specific service times, the time-average queue length could be further reduced by using scheduling rules, e.g., the *shortest remaining processing time* discipline.

to switch to the high capacity and back.

Definition 3.3. *We call π_R^K a K-threshold policy if the contingent capacity is activated if and only if the queue length at time t is larger than K.*

It is well established that a threshold policy is the optimal stationary policy when controlling a contingent capacity that can be added to a base capacity. Crabill (1972) shows that a threshold policy is optimal if total costs, costs depending on the number of jobs in an M/M/1 queue with infinite waiting room and capacity costs, are being minimized [17]. However, an expression for the optimal threshold value is not provided. By adjusting the cost factors, this problem can easily be transferred to a setting where we do not try to minimize total costs but the time-average queue length subject to the capacity constraint stated in Definition 3.1. Thus, a K-threshold policy is optimal, $\mathcal{Q}^\star_{\Pi^0}(r,p,\lambda) = \inf_{\pi\in\Pi^0} \mathcal{Q}(r,p,\lambda,\pi) = \mathcal{Q}(r,p,\lambda,\pi_R^K)$, where Π^w denotes the family of reactive policies with lookahead window length w. The following theorem characterizes the optimal feasible threshold policy.

Theorem 3.1. *Fix $r \in (0,1)$ and $p > r$. Let*

$$K(r,p,\lambda) = \left\lceil \log_{\frac{\lambda}{1-r}} \mathcal{K}(r,p,\lambda) \frac{1}{1-\lambda} \right\rceil, \tag{3.1}$$

with $\mathcal{K}(r,p,\lambda) = \frac{r(r-1)(\lambda+r-p-1)}{p\lambda}$. Then, $\pi_R^{K(r,p,\lambda)}$ is the optimal threshold policy and feasible for all $\lambda \in (1-r,1]$.

Proof. See Appendix B.1. □

The proof of the theorem is based on the continuous-time Markov chain illustrated in Figure 3.3. An example of a queue length process of an M/M/1 queue in reactive mode is displayed in Figure 3.4. It can be observed that the high capacity is used at all times t where $Q(t) > K(r,p,\lambda)$.

Given the optimal threshold value, we can derive an expression for the time-average queue length and investigate its behavior as $\lambda \to 1$.

Theorem 3.2. *Given $r \in (0,1)$, $p > r$ and $K(r,p,\lambda)$, the time-average queue*

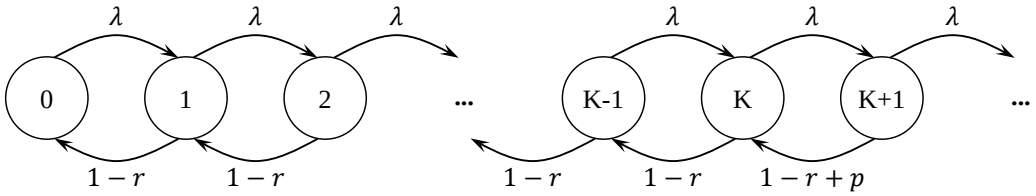


Figure 3.3: Flow diagram of an M/M/1 queue with reactive capacity control.

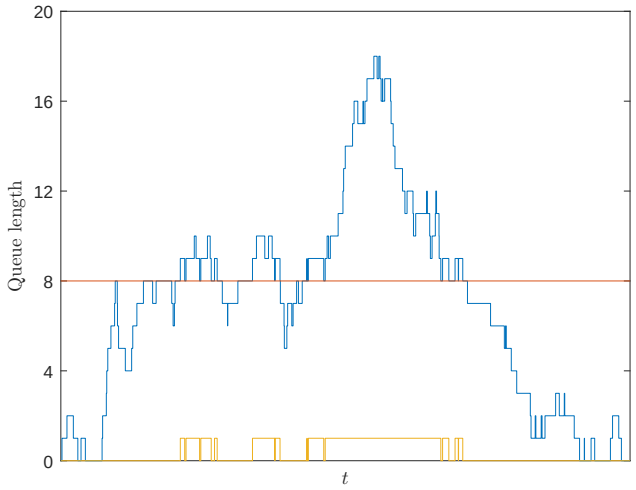


Figure 3.4: Simulated queue length process (blue) with $r = 0.2$, $p = 0.4$ and $\lambda = 0.98$ resulting in $K(r, p, \lambda) = 8$ (red). The yellow line corresponds to the reactive capacity control policy π_R^K.

length can be computed as

$$\mathcal{Q}(r, p, \lambda, \pi_R^K) = \left[\frac{(K-1)\rho_1^{K+1} - K\rho_1^K + \rho_1}{(\rho_1 - 1)^2} + \frac{\rho_2 + K(1-\rho_2)}{(1-\rho_2)^2} \rho_1^K \right] \nu(0|\pi_R^K). \tag{3.2}$$

As $\lambda \to 1$,

$$\mathcal{Q}(r, p, \lambda, \pi_R^{K(r,p,\lambda)}) \sim \mathcal{O}\left(\log_{\frac{1}{1-r}} \frac{1}{1-\lambda} \right), \tag{3.3}$$

i.e., the time-average queue length diverges.

Proof. See Appendix B.1. □

It is intuitive that the time-average queue length diverges at the same order

as the threshold level. The queue length process always fluctuates around K since the associated random walk has a positive drift at all times $\{t \in \mathbb{R}_+ : Q(t) \leq K\}$ and a negative drift for all other times $\{t \in \mathbb{R}_+ : Q(t) > K\}$. Thus, if we have no information about the future, the optimal time-average queue length diverges at rate $\mathcal{O}(\log_{1/(1-r)} \frac{1}{1-\lambda})$, as $\lambda \to 1$.

For $\lambda \in (1-r, 1]$, the time-average queue length, the optimal threshold and the time share the high capacity is active are illustrated in Figure 3.5 for $r = 0.2$ and $p = 0.4$. We can observe that the time share the high capacity

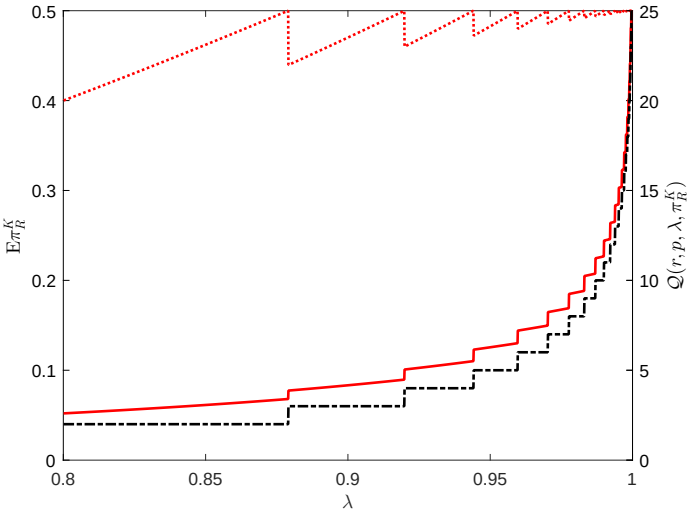


Figure 3.5: Performance of the reactive policy: time share the high capacity is active (dashed line, left y-axis), time-average queue length (solid line, right y-axis) and optimal threshold level (dashed line, right y-axis) versus arrival rate.

is active, $\mathbb{E}\pi_R^K = \mathbb{P}[Q > K(r,p,\lambda)]$, increases with λ until it reaches r/p (0.5 in the this example). In order for the policy to remain feasible, the threshold then jumps to next higher positive integer and $\mathbb{P}[Q > K(r,p,\lambda)]$ to a lower value in $(r/p(1-r), r/p]$. The reactive policy performs very well if λ is low. As λ increases to 1, the optimal threshold and the time-average queue length grow exponentially.

Finally, we conclude by providing some insights regarding the effects of choosing different values for $p > r$. While the time-average service rate is

always less or equal than one, the policy's performance depends on the choice of p (in proportion to r).

Proposition 3.1. *Fix $r \in (0,1)$. For any $\lambda \in (1-r, 1]$, the worst time-average queue length, $\sup_{p>r} \mathcal{Q}(r,p,\lambda,\pi_R^K)$, is obtained at $p \downarrow r$. If $p \downarrow r$, the time-average queue length corresponds to the time-average number of jobs in an M/M/1 queue with a static service rate $\mu = 1$. Thus, any reactive K-threshold policy outperforms the static service rate model.*

Proof. See Appendix B.1. □

This is illustrated in Figure 3.6. The time-average number of jobs in the

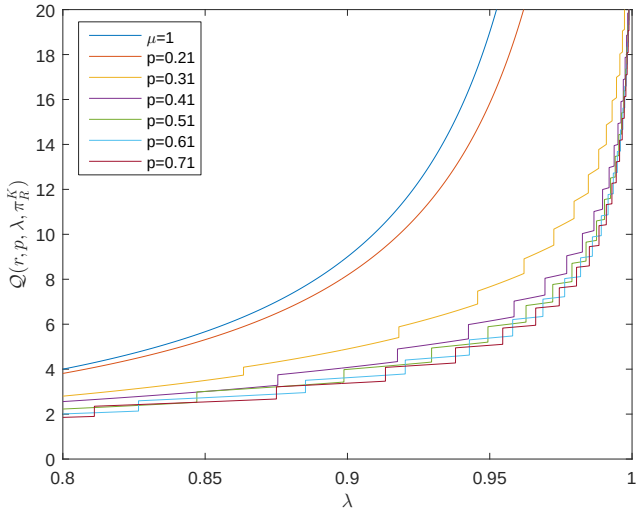


Figure 3.6: Time-average queue length for static service rate and different values of p versus arrival rate.

system of an M/M/1 queue with static service rate and $\rho < 1$ is given as $\mathbb{E}(L_{\mathrm{M/M/1}}) = \rho/(1-\rho)$. We can observe that $\mathcal{Q}(r,p,\lambda,\pi_R^K)$ converges to $\mathbb{E}(L_{\mathrm{M/M/1}})$ for small p, i.e., $\mathbb{E}(L_{\mathrm{M/M/1}})$ provides an upper bound for the time-average queue length of an M/M/1 queue with flexible capacity. Intuitively, if the contingent capacity is in the order of r, it has to be applied for a large fraction of time, $\mathbb{E}\pi_R^K \to 1$ as $p \downarrow r$, and therefore the flexible capacity management model approaches the static model with $\mu = 1$. On the

other hand, if p increases, the time-average queue length $Q(r, p, \lambda, \pi_R^K)$ decreases for all λ up to a certain level. We see that for $p = \{0.51, 0.61, 0.71\}$, $\min\{Q(r, 0.51, \lambda, \pi_R^K), Q(r, 0.61, \lambda, \pi_R^K), Q(r, 0.71, \lambda, \pi_R^K)\}$ is alternating between the three contingent capacity levels over λ. Naturally, an M/M/1 queue with $\mu = 1 - r + p$ provides a lower bound for the time-average queue length of a system with flexible capacity.

In conclusion, we find that the time-average queue length diverges in the reactive mode as $\lambda \to 1$. However the results of Spencer et al. (2014) and Xu and Chan (2016) for diversion suggest that there exists an optimal policy taking into account future information such that the time-average queue length converges to a (relatively small) constant [36, 49]. This problem is subject of the subsequent section.

3.5 Solely Forward-Looking Capacity Control

In this section, we assume to know future information regarding arrival and service times of the jobs within an infinite lookahead window, $w = \infty$. The decision to activate the contingent capacity at time t_0 is made based on information about all future arrivals, i.e., the realizations of the arrival times and service requirements in the time interval $[t_0, \infty)$.

After introducing some technical details, we develop the optimal policy for this setting, prove its feasibility and derive an expression for the time-average queue length. Based thereupon, we prove asymptotic optimality of the policy as $\lambda \to 1$. We then investigate further properties of the policy and the resulting time-average queue length and provide numerical results to illustrate and validate our results.

3.5.1 Solely Forward-Looking Policy

Job arrivals to the system occur according to a Poisson process $\{A(t) : t \in \mathbb{R}_+\}$, $A(t) \in \mathbb{Z}_+$, with arrival rate $\lim_{t\to\infty} A(t)/t = \lambda \in (1-r, 1]$. When developing the solely forward-looking policy (with infinite lookahead window), we assume to have perfect knowledge regarding the realization of this stochastic

process. Define the service token generation process of the base capacity server as a Poisson process with rate $(1-r)$ as $\{S_1(t) : t \in \mathbb{R}_+\}$, $S_1(t) \in \mathbb{Z}_+$. We also assume to have perfect knowledge regarding the realization of this stochastic process. The number of jobs in the system $\{X_0(t) : t \in \mathbb{R}_+\}$, $X_0(t) \in \mathbb{Z}$, is therefore given as

$$X_0(t) = A(t) - S_1(t).$$

Since $(1 - r) < \lambda$, X_0 is transient. Note that X_0, also referred to as *doubly-infinite queue*, can have negative values (if more service tokens are produced than consumed), because the service token process is independent of the state of the system. Given any doubly infinite queue X, the corresponding queue length process $\{Q(t) : t \in \mathbb{R}_+\}$, $Q(t) \in \mathbb{Z}_+$, is defined as the reflected version of X, where $\{Y(t) : t \in \mathbb{R}_+\}$, $Y(t) \in \mathbb{Z}_+$, denotes the reflection map or regulator,

$$Q(t) = X(t) + Y(t) = X(t) + \sup_{0 \le s \le t} [-X(s)]^+.$$

Here, $[\cdot]^+ = \max\{\cdot, 0\}$. Thus, the *initial queue length process*, which is the queue length process if only the base capacity is applied, is given as $Q_0(t) = A(t) - S_1(t) + Y_0(t)$.

Future information means that, although interarrival and service times are random variables, we know their realizations in the future. This is illustrated in Figure 3.7. If we are now at time t_0, no future information ($w = 0$) means that we only know the initial queue length process from the past (upper chart). Future information ($w > 0$) implies that we know the times when jobs arrive and their service requirements (lower chart). Thus, we can compute the future initial queue length process in the interval $[t_0, t_0 + w]$ and we will focus on effectively using this information in the remainder of the section.

The embedded discrete-time process of $Q(t)$ is given by $\{Q[n] : n \in \mathbb{Z}_+\}$, where $T_n \in \mathbb{R}_+$ is the corresponding time of the n-th event in $Q(t)$, i.e., a job arrival or a job departure after generation of a service token. The values of $Q[n]$ are well defined as the sample paths of Poisson processes are right-continuous with left limits (RCLL) almost surely. Therefore, $Q[n] = Q(T_n+) = Q(T_n)$, where $T_n+ = \lim_{\epsilon \downarrow 0} T_n + \epsilon$, and $Q[n-1] = Q(T_n-) =$

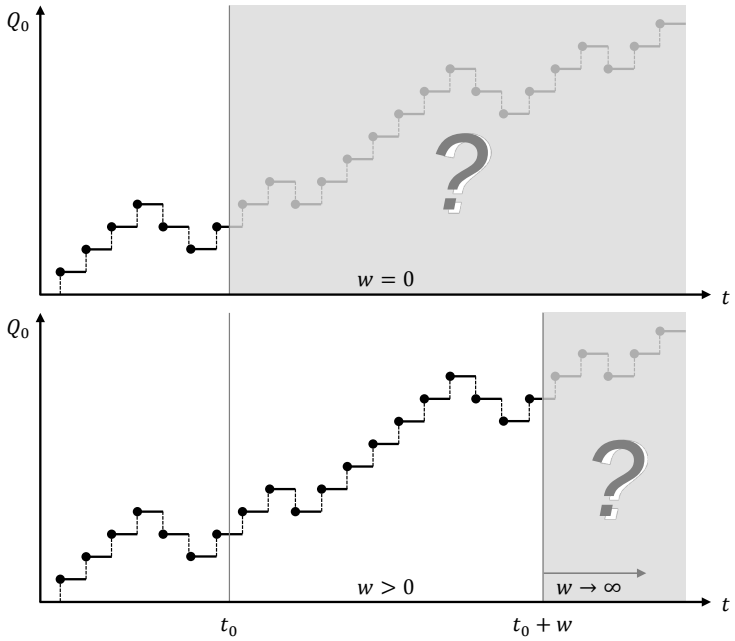


Figure 3.7: Illustration of future information with the initial queue length process Q_0.

$Q(T_{n-1})$.

With the solely forward-looking policy, we assume that the server generates service tokens according to a non-stationary Poisson process with rates $(1 - r)$ or $(1 - r + p)$ if switched to high capacity. The service tokens are either consumed instantaneously when the system is not empty or wasted if the system is empty. Define $Q_2(t)$, $X_2(t)$ and $Y_2(t)$ as the corresponding processes after applying the feasible solely forward-looking policy with infinite lookahead window π_F^∞.

Lemma 3.1. *With infinite lookahead window, there exists a feasible solely forward-looking policy such that $Y_2(t) = Y_0(t)$, for all $t \in \mathbb{R}_+$, resulting in a time-average queue length which is finite almost surely as $\lambda \to 1$.*

Proof. Since $X_0(t)$ is a transient random walk with positive drift on $\mathbb{Z}$, its all-time minimum $-M = -\max_{t\in\mathbb{R}_+} Y_0(t)$, where $Y_0(t)$ denotes the reflection map of $X_0(t)$, is geometrically distributed with parameter $(1 - r)/\lambda < 1$ and thus finite almost surely. $Y_0(t)$ accounts for the total number of service tokens

wasted until time t and is increasing. Therefore, if we consider the family of solely forward-looking policies such that for the reflection map of the resulting process $Y_2(t) = Y_0(t) \leq M,\ \forall t \in \mathbb{R}_+$, i.e., no service tokens are wasted once the all-time minimum $-M$ of $X_0(t)$ has been reached, it follows instantly from the functional strong law of large numbers (FSLLN) that there must exist a feasible solely forward-looking policy with $(1-r) + \mathbb{E}\pi_F^\infty p \geq \lambda$ such that the resulting time-average queue length is finite almost surely.

More specifically, there exists a policy such that if $(1-r) + \mathbb{E}\pi_F^\infty p = \lambda$ and $Y_2(t) = Y_0(t) \leq M,\ \forall t \in \mathbb{R}_+$, $X_2(t)$ has a drift of zero. The expected difference between a stochastic process with a drift of zero and a finite lower bound $-M$ is finite. This concludes the proof. □

In order to define a feasible solely forward-looking policy meeting the requirements stated in Lemma 3.1, we need some further definitions. Given a lookahead window size of $w \in \mathbb{R}_+$, define $W(n) \in \mathbb{Z}_+$ as the window size with respect to the discrete-time initial queue length process $Q_0[n]$,

$$W(n) = \sup\{k \in \mathbb{Z}_+ : T_{n+k} \leq T_n + w\},$$

where $T_n \in \mathbb{R}_+$ denotes the time of the n-th event in $Q_0(t)$, i.e., all events in $A \cup S_1$.[14] For $x \in \mathbb{Z}_+$, define the set of indices

$$U(Q, n, x) = \inf\{j \in \{1, \ldots, x\} : Q[n+j] = Q[n] - 1\},$$

yielding the first index $n + j \in \{n+1, \ldots, n+x\}$ for which the process $Q[n+j]$ drops below $Q[n]$. Based on Spencer et al. (2014), we define the *no-job-left-behind* (NOB) arrivals to the system with base capacity $(1-r)$ as

$$\Psi^w = \{n \in \Phi(Q_0) : U(Q_0, n, W(n)) = \infty\},$$

where $\Phi(Q) = \{n \in \mathbb{Z}_+ : Q[n] > Q[n-1]\}$ describes the locations of all arrivals in the process $Q[n]$ and $\bar{\Phi}(Q) = \mathbb{Z}_+ \setminus \Phi(Q)$ defines all departures [36].

[14] Although we do not consider finite lookahead windows in first part of this section, we introduce the notation now for completeness. More details regarding finite lookahead windows will be given in Section 3.7.

Define the NOB arrival process as

$$A^w_\Psi(t) = |\{n \in \Psi^w : T_n \leq t\}|,$$

where $|\cdot|$ denotes the cardinality. $A^w_\Psi(t)$ counts the number of NOB arrivals until time t, is RCLL and increasing and $\bar{A}^w_\Psi(t)$ denotes all non-NOB arrivals, $\bar{A}^w_\Psi(t) = A(t) - A^w_\Psi(t)$.

The following lemma will be needed in order to obtain the time-average queue length of the flexible capacity management model.

Lemma 3.2. *Fix $r \in (0,1)$, $p = \infty$ and $\lambda \in (1-r, 1]$. With infinite lookahead window and if we switch to the high capacity at all NOB arrival instants,*

$$\pi^\infty_{\mathrm{NOB}}(t) = \begin{cases} 1, & \text{if } t \in \{T_n\}_{n \in \Psi^\infty} \\ 0, & \text{otherwise,} \end{cases}$$

the time-average queue length is given as

$$\mathcal{Q}(r, \infty, \lambda, \pi^\infty_{\mathrm{NOB}}) = \frac{1-r}{\lambda - (1-r)}. \tag{3.4}$$

Proof. The proof is based on Proposition 2 in Spencer et al. (2014) [36]. They consider future information in a setting where the server has capacity $(1-r)$, arrivals occur at rate $\lambda \in (1-r, 1]$ and a fraction r of all arrivals can be diverted. Due to the service token model, switching on the contingent capacity $p = \infty$ at a NOB arrival instant and thus instantaneously producing a service token that is consumed by the first job in line corresponds to a diversion of a NOB job in Spencer's model. Therefore, as they derive that the time-average queue length is given as the time-average number of jobs in an M/M/1 queue with $\rho = \frac{1-r}{\lambda} < 1$, given in Equation (3.4), the same holds for the flexible capacity management model with $p = \infty$ and $\pi^\infty_{\mathrm{NOB}}$ as defined above. □

Denote the queue length process obtained for the parameters defined in Lemma

3.2 as Q_1. It is given as

$$Q_1(t) = Y_0(t) + \bar{A}^{\infty}_{\Psi}(t) - S_1(t). \tag{3.5}$$

Furthermore, the queue length process Q_2 with finite contingent capacity $p < \infty$ can be computed as

$$Q_2(t) = Y_0(t) + A(t) - S_1(t) - S_2(\varpi^{\infty}_F(t)), \tag{3.6}$$

where

$$\varpi^{w}_F(t) = \int_0^t \pi^{w}_F(s)\,\mathrm{d}s.$$

$S_2(\varpi^{\infty}_F(t))$ denotes the process counting the additional service tokens produced at the contingent capacity rate p when the contingent capacity is active, i.e., $\pi^{\infty}_F(t) = 1$. Finally, we can define the solely forward-looking policy.

Theorem 3.3. *Fix $r \in (0,1)$ and $p > r$. Given a queue length process Q_1, define the solely forward-looking capacity control policy with infinite lookahead window π^{∞}_F such that the contingent capacity is activated if and only if $Q_2(t) > Q_1(t)$,*

$$\pi^{\infty}_F(t) = \begin{cases} 1, & \text{if } Q_2(t) > Q_1(t), \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

The solely forward-looking policy π^{∞}_F is feasible for all $\lambda \in (1-r, 1]$. More precisely,

$$\mathbb{E}\pi^{\infty}_F = \frac{\lambda - 1 + r}{p} \leq \frac{r}{p}.$$

Before providing the proof of the theorem, we establish some intuition regarding the queue length processes Q_1 and Q_2 as well as the solely forward-looking policy π^{∞}_F.

First, we consider the definition of the queue length process with $p = \infty$ (3.5), which corresponds to a system with constant service rate and diversion as described in Lemma 3.2. The non-decreasing process Y_0 accounts for any

wasted service tokens until the first NOB arrival and thus remains constant once the all-time minimum of the initial doubly-infinity queue X_0 has been reached, as stated in Lemma 3.1. $\bar{A}^{\infty}_{\Psi}$ corresponds to the Poisson arrival process without NOB arrivals. The NOB arrivals, which are all arrivals after which the doubly-infinity queue X_0 will not be smaller than its current value again, are (virtually) deleted, thus leading to recurrence of the resulting queue length process. The NOB arrival process A^{∞}_{Ψ} increases with step size one and can be thought of as the drift of X_0. Also, the NOB arrivals describe the earliest arrivals such that, if they are deleted, no (more) service tokens are wasted, i.e., if considering a system with job-specific service time, the server is never idle. Spencer et al. (2014) use this result by comparing the queue length process to one obtained using a greedy deletion rule to argue that the NOB policy $\pi^{\infty}_{\mathrm{NOB}}$ is asymptotically optimal as $\lambda \to 1$ [36].

Q_1 serves as a lower bound for any queue length process obtained with $p < \infty$, i.e., we want to find a feasible solely forward-looking policy such that the difference between Q_2 and Q_1 is minimal. This is exactly what we obtain with the definition of Q_2 in combination with π^{∞}_{F}, (3.6) and (3.7). Y_0 is included in the definition of Q_2 for the same reason as in Q_1. But in contrast, in Q_2 we cannot delete arrivals, but need to produce a service token (at finite generation rate) for each arrival. Thus, arrivals occur according to the original Poisson process A. Once the low capacity is active, i.e., $\pi^{\infty}_{F}(t) = 0$, service tokens are produced according to S_1. S_2 is only increasing when $\pi^{\infty}_{F}(t) = 1$ and constant otherwise, $S_2(\varpi^{\infty}_{F}(t))$.

The solely forward-looking policy is such that the server switches on the contingent capacity as soon as a NOB arrival occurs (at NOB arrivals instants $Q_2(t)$ must exceed $Q_1(t)$, as the arrival is not deleted) and only switches back to low capacity once all NOB arrivals have been accounted for. Figure 3.8 illustrates the three different queue length processes. A simulated version of the processes is shown in Figure 3.9.

Proof of Theorem 3.3: The proof is based on the amount of extra arrivals that need to be served by the contingent capacity compared to the queue length process obtained by deleting NOB arrivals. We can rewrite the queue length

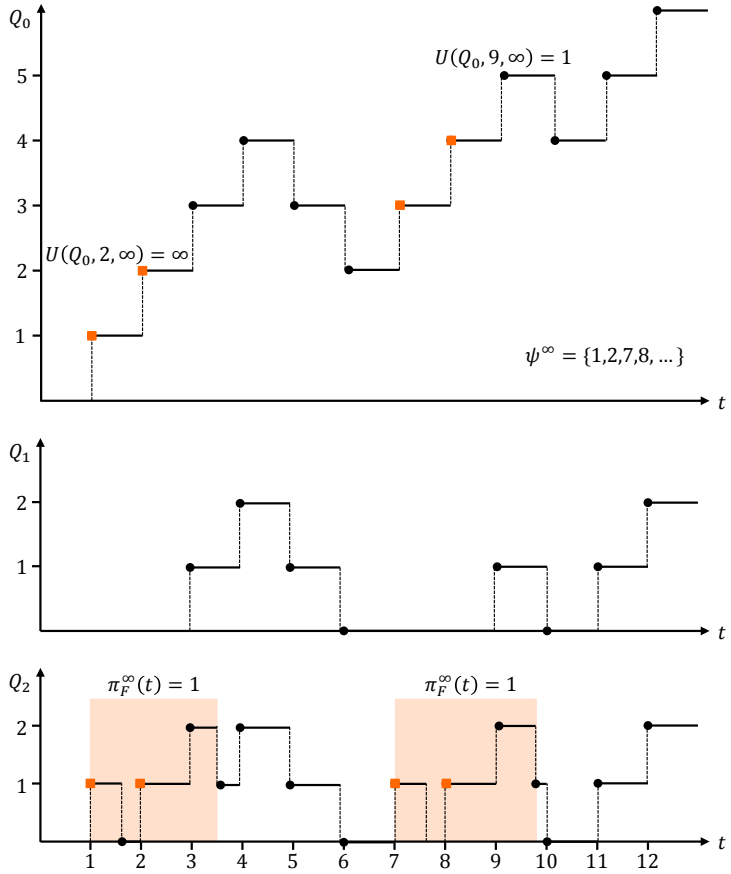


Figure 3.8: Illustration of the solely-forward looking policy: The upper process illustrates the initial queue length process Q_0. The little orange boxes correspond to the NOB arrivals, i.e., all arrivals for which the queue length process is not lower again. The process in the middle corresponds to the process using the π^∞_{NOB} policy for $p = \infty$. NOB arrivals are virtually deleted. Finally, the lower process is the queue length process when applying the solely forward-looking policy π^∞_F. NOB arrivals are not deleted, but compensated for until Q_2 coincides with Q_1.

process with solely forward-looking capacity control as

$$Q_2(t) = Q_1(t) + [Q_2(t) - Q_1(t)] = Q_1(t) + \delta Q(t).$$

By using the definitions (3.5) and (3.6) of the processes Q_1 and Q_2 we find

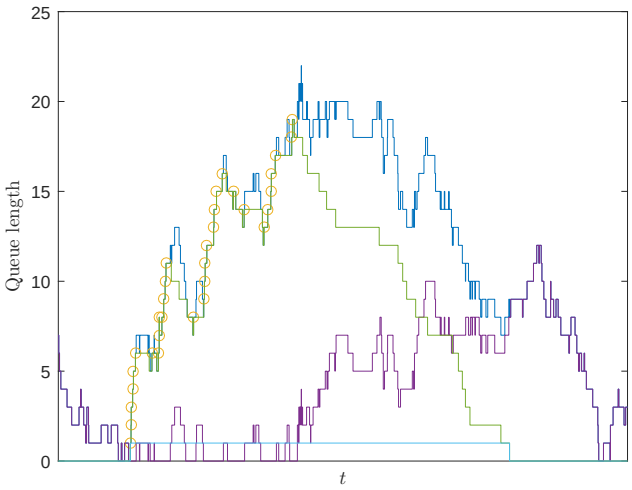


Figure 3.9: Snapshot of a queue with solely forward-looking flexible capacity control, where $r = 0.2$, $p = 0.4$ and $\lambda = 1$. Displayed are the processes Q_1 (purple), Q_2 (blue) and δQ (green) and the solely forward-looking policy π_F^∞ (turquoise). The yellow circles indicate NOB arrivals.

that

$$\begin{aligned}\delta Q(t) &= Y_0(t) - Y_0(t) + A(t) - \bar{A}_\Psi^\infty(t) - S_1(t) + S_1(t) - S_2(\varpi_F^\infty(t)) \\ &= A_\Psi^\infty(t) - S_2(\varpi_F^\infty(t)).\end{aligned}$$

Therefore, the time share that the high capacity is active corresponds to the probability that $\delta Q > 0$. From the FSLLN we can deduce the following corollary.

Corollary 3.1. *The time-average rate of NOB job arrivals is given as*

$$\lim_{t\to\infty} \frac{A_\Psi^\infty(t)}{t} = \lim_{t\to\infty} \frac{A(t) - S_1(t)}{t} = \lambda - (1 - r).$$

If the contingent capacity was always active, additional service tokens would be produced according to a Poisson process with rate p, and we know from

the FSLLN that

$$\lim_{t\to\infty} \frac{S_2(\mathbf{1}\{t \in \mathbb{R}_+\})}{t} = p.$$

But, as the number of additional service tokens produced until time t by the contingent capacity cannot exceed the number of NOB arrivals that have occurred until time t, since otherwise $Q_2(t) < Q_1(t)$ ($\delta Q(t) < 0$), and in the long run both numbers are equal, it follows that

$$\begin{aligned} \lambda - (1-r) = \lim_{t\to\infty} \frac{S_2(\varpi_F^\infty(t))}{t} &= \mathbb{E}\pi_F^\infty p \\ \implies \mathbb{E}\pi_F^\infty &= \frac{\lambda - 1 + r}{p}. \end{aligned}$$

This concludes the proof of the theorem. □

From this theorem, we observe that the time share the high capacity is active is linearly increasing in $\lambda \in (1-r, 1]$ and $\lim_{\lambda\downarrow(1-r)} \mathbb{E}\pi_F^\infty = 0$. This implies that if the arrival rate approaches its lower bound, the contingent capacity is not used any longer, and the system corresponds to an M/M/1 queue with arrival rate $\lambda = 1 - r$ and static service rate $\mu = 1 - r$. Therefore, as it can be found when analyzing the analytic expression for the time-average queue length derived subsequently, the queue length process should become transient as $\lambda \downarrow (1-r)$, $\lim_{\lambda\downarrow(1-r)} \mathcal{Q}(r, p, \lambda, \pi_F^\infty) = \infty$. Having the definitions (3.5) and (3.6) of Q_1 and Q_2 in mind, we can assume that the time-average queue length of the queue length process with solely forward-looking capacity control Q_2 will be of the following form:

$$\mathcal{Q}(r, p, \lambda, \pi_F^\infty) = \mathcal{Q}(r, \infty, \lambda, \pi_{\text{NOB}}^\infty) + x, \tag{3.8}$$

where $x \in \mathbb{R}_+$. Formally, x is given as

$$x = \limsup_{N\to\infty} \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} \delta Q[n]\right),$$

where $\delta Q[n]$ is the embedded discrete-time process of

$$\delta Q(t) = A^{\infty}_{\Psi}(t) - S_2(\varpi^{\infty}_F(t)).$$

$\delta Q(t)$ corresponds to the queue length process of an IPP/M/1 queue with service rate p and time-average arrival rate $\lambda-(1-r)$, as given in Lemma 3.1, where IPP stands for interrupted Poisson process.[15] The time-average queue length of an IPP/M/1 queue can only be computed numerically using matrix-analytic methods, see Ibe (2013), Section 12.5 [22]. However, as we already have some insights regarding the limiting behavior of all three components in equation (3.8), we can approximate x by a simple expression. First, assume that $\lambda \uparrow 1$ and $p \downarrow r$, i.e., $\mu_2 \downarrow 1$. Then, the high capacity will always be active and the system corresponds to an M/M/1 queue with an arrival and service rate of 1. Therefore, $\mathcal{Q}(r, p, \lambda, \pi^{\infty}_F) \to \infty$, and, since $\mathcal{Q}(r, \infty, \lambda, \pi^{\infty}_{\mathrm{NOB}}) = (1 - r)/r$, we can conclude that $x \to \infty$. Next, consider the case where $\lambda \in (1-r, 1]$ and $p \uparrow \infty$. This corresponds to the M/M/1 system with diversion described in Lemma 3.2, and since $\mathcal{Q}(r, p, \lambda, \pi^{\infty}_F) = \mathcal{Q}(r, \infty, \lambda, \pi^{\infty}_{\mathrm{NOB}}) = (1-r)/r$ it follows that $x = 0$. The M/M/1 system with diversion provides a lower bound for any system with finite contingent capacity. If p is larger than (and not to close to) r and $\lambda \downarrow (1 - r)$, we know that $\mathcal{Q}(r, p, \lambda, \pi^{\infty}_F) \to \infty$. Since also $\mathcal{Q}(r, \infty, \lambda, \pi^{\infty}_{\mathrm{NOB}}) \to \infty$, x can be any number in $\mathbb{R}_+ \cup \{\infty\}$. Finally, if, for the same system, $\lambda \uparrow 1$ we know from Lemma 3.1 that $\mathcal{Q}(r, p, \lambda, \pi^{\infty}_F) \in \mathbb{R}_+ \setminus \{\infty\}$, and since $\mathcal{Q}(r, \infty, \lambda, \pi^{\infty}_{\mathrm{NOB}}) = (1 - r)/r$ it follows that $x \in \mathbb{R}_+ \setminus \{\infty\}$.

We know that the queue length process Q_2 is larger than 0 when the high capacity is active. The expected number of jobs present in a non-empty M/M/1 queue with arrival rate β, service rate $\gamma > \beta$ and load $\rho = \beta/\gamma < 1$

[15] An IPP is a Poisson process with an ON state where the rate is given as $\lambda^{\mathrm{ON}}_{\mathrm{IPP}} = \lambda$ and an OFF state where $\lambda^{\mathrm{OFF}}_{\mathrm{IPP}} = 0$. The time spent in the ON state is exponentially distributed with mean $1/(1-r)$ (idling period of an M/M/1 queue with arrival rate $(1-r)$). The time spent in the OFF state has the same pdf as the busy-period distribution of an M/M/1 queue which involves a Bessel function of the first kind and has mean $1/(\lambda - 1 + r)$, see Asmussen (2008), Section III.8c, Corollary 8.7 [2].

can be computed as

$$\begin{aligned}\mathbb{E}(L|L>0) &= \frac{\mathbb{E}(\mathbf{1}\{L>0\}L)}{\mathbb{P}(L>0)} = \frac{\sum_{i=1}^{\infty} i(1-\rho)\rho^i}{\rho} \\ &= \frac{(1-\rho)}{\rho}\sum_{i=0}^{\infty} i\rho^i = \frac{(1-\rho)}{\rho}\frac{\rho}{(1-\rho)^2} \\ &= \frac{1}{1-\rho}.\end{aligned}$$

Therefore, the busy-period queue length of a service token-based M/M/1 system with arrival rate $\alpha = \lambda$ and service rate $\beta = 1 - r + p$ is given as

$$\mathbb{E}(L|L>0) = \frac{\beta}{\beta-\alpha} = \frac{1-r+p}{1-r+p-\lambda}.$$

By Theorem 3.3 we know that time share where the queue length process $Q_2(t) > Q_1(t)$ is given as $\mathbb{E}\pi_F^\infty = (\lambda - 1 + r)/p$. Thus, for all $\{t \in \mathbb{R}_+ : \pi_F^\infty(t) = 1\}$, i.e., where $Q_2(t) > Q_1(t)$, the expected distance between $Q_2(t)$ and $Q_1(t)$ can be approximated as the expected number of jobs in the system in a busy-period of an M/M/1 queue with arrival rate λ and service rate $1 - r + p$,

$$x \sim \mathbb{E}\pi_F^\infty \cdot \mathbb{E}(L|L>0) = \frac{\lambda-1+r}{p}\frac{1-r+p}{1-r+p-\lambda}. \tag{3.9}$$

Finally, we obtain the following: For $r \in (0,1)$ and $p > r$, the resulting time-average queue length for the feasible proactive capacity control policy π_F^∞ can be approximated as

$$\mathcal{Q}(r,p,\lambda,\pi_F^\infty) \sim \frac{1-r}{\lambda-1+r} + \frac{(\lambda-1+r)(1-r+p)}{p(1-r+p-\lambda)}.$$

As $\lambda \to 1$, the time-average queue length converges to a finite value almost surely,

$$\lim_{\lambda\to 1} \mathcal{Q}(r,p,\lambda,\pi_F^\infty) \sim \frac{1-r}{r} + \frac{r(1-r+p)}{p(p-r)}.$$

Finally, it remains to be shown that the solely forward-looking policy is asymp-

totically optimal for heavily loaded systems.

Theorem 3.4. *Fix $r \in (0,1)$ and $p > r$. The solely forward-looking policy π_F^∞ is asymptotically optimal as $\lambda \to 1$,*

$$\lim_{\lambda \to 1} \mathcal{Q}(r, p, \lambda, \pi_F^\infty) = \lim_{\lambda \to 1} \inf_{\pi \in \Pi^\infty} \mathcal{Q}(r, p, \lambda, \pi),$$

where Π^∞ denotes the family of all feasible policies with infinite lookahead window.

Proof. Spencer et al. (2014) show that the NOB policy $\pi_{\mathrm{NOB}}^\infty$ for $p = \infty$ is optimal by comparing the resulting queue length process when applying the policy to the one obtained with a greedy algorithm minimizing the area under Q_1 [36]. Therefore, we only need to show that there exists no policy such that the time-average excursion above Q_1, i.e., the difference between Q_2 and Q_1, is smaller than when applying the solely forward-looking policy π_F^∞.

Lemma 3.3. *As $\lambda \to 1$, any feasible policy $\pi \in \Pi^\infty$ for which $\mathcal{Q}(r, p, \lambda, \pi) < \infty$ almost surely must be such that $S_2(\varpi(t)) \leq A_\Psi^\infty(t)$ for all $t \in \mathbb{R}_+$.*

Proof. See Appendix B.2. □

The result presented in this lemma shows that, as $\lambda \to 1$, any feasible policy π for which the time-average queue length converges to a finite value must be such that no service tokens are wasted after the all-time minimum of Q_0 was reached. Therefore, $A_\Psi^\infty(t)$ is an upper bound of $S_2(\varpi(t))$. When additionally considering the solely forward-looking policy π_F^∞, we find the following lemma.

Lemma 3.4. *As $\lambda \to 1$, $S_2(\varpi(t)) \leq S_2(\varpi_F^\infty(t)) \leq A_\Psi^\infty(t)$ for all $t \in \mathbb{R}_+$.*

Proof. See Appendix B.2. □

The lemma tells us, for $\lambda \to 1$, that there exists no feasible policy π such that the number of service tokens produced by the contingent capacity until time t is higher than if the solely forward-looking policy π_F^∞ was used. Therefore, if we denote $\delta Q^\dagger(t) = A_\Psi^\infty(t) - S_2(\varpi_F^\infty(t))$ and $\delta Q^\ddagger(t) = A_\Psi^\infty(t) - S_2(\varpi(t))$, it

follows that

$$\begin{aligned}&\delta Q^{\dagger}(t) \leq \delta Q^{\ddagger}(t),\ \forall t \in \mathbb{R}_{+},\\ &\implies \limsup_{N\to\infty} \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}\delta Q^{\dagger}[n]\right) \leq \limsup_{N\to\infty} \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}\delta Q^{\ddagger}[n]\right).\end{aligned}$$

Thus, the time-average excursion above Q_1 is minimal for the solely forward-looking policy π_F^{∞} which proves the theorem. □

More insights regarding the properties of the solely forward-looking policy and a simulation study to verify the analytical results are presented in the next section.

3.5.2 Properties and Numerical Insights

In the previous section we derived an exact expression for the time share the contingent capacity is active and an approximation for the time-average queue length. In this section, we provide numerical insights regarding the properties of the solely forward-looking policy with infinite lookahead window.[16] We first simulate the system for a fixed base capacity $(1-r)$, a contingent capacity p and a variable arrival rate λ. Then, we fix $r \in (0,1)$ and $\lambda = 1$ and vary the contingent capacity $p > r$.[17]

To simulate the performance of the solely forward-looking policy, we fix $r = 0.2$ and $p = 0.4$, i.e., the contingent capacity can be used 50% of the time. Figure 3.10 displays the time-average queue length versus the arrival rate. The time-average queue length diverges as $\lambda \downarrow (1-r)$ which comes from the NOB part of the approximation, $\lim_{\lambda\downarrow(1-r)} \mathcal{Q}(r,\infty,\lambda,\pi_{\mathrm{NOB}}^{\infty}) \to \infty$. We see that the accuracy of the approximation decreases in this range, which makes sense since in this limit the queue behaves like an M/M/1 queue with arrival and service rate of $(1-r)$ and load $\lim_{\lambda\downarrow(1-r)} \rho = \frac{1-r}{\lambda} = 1$. For $\lambda > 0.85$, there is practically no difference between the approximation and the results

[16] For the sake of conciseness, we restrict the comparison of analytical results and simulation to this section, as the expression for time-average queue length for the solely forward-looking policy is an approximation and all other expressions are exact analytical results.
[17] The simulation was performed with MATLAB.

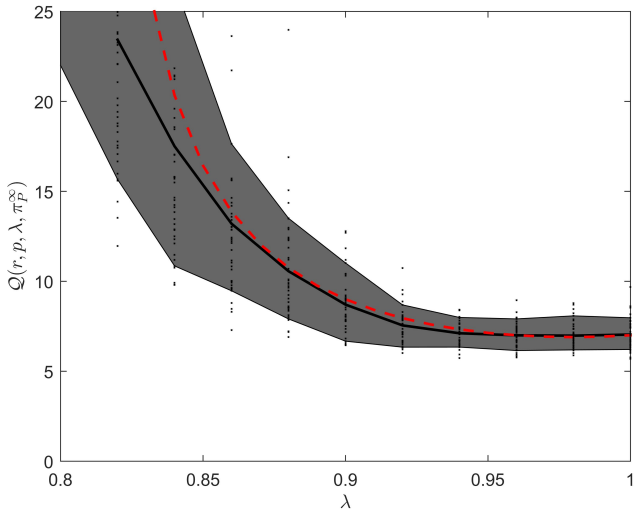


Figure 3.10: Time-average queue length for solely forward-looking policy with $r = 0.2$ and $p = 0.4$: the dashed line corresponds to the analytical solution, the solid line to the mean of 40 simulation runs per value of λ (with $N = 5,000$ events each) and the gray area to the 80% confidence interval of the simulation.

of the simulation. As we are specifically interested in the limit $\lambda \to 1$, we can conclude that the approximation performs very well if we fix both capacity levels and vary the arrival rate. Figure 3.11 displays the corresponding results for the time share the contingent capacity is active versus the arrival rate. We see that the analytical results and the results of the simulation coincide for all $\lambda \in (1 - r, 1]$.

Next, we fix $r = 0.2$ and $\lambda = 1$ and compare the analytical and simulated results for varying $p > r$. Figure 3.12 illustrates the time-average queue length versus the contingent capacity. We can observe that the approximation and the simulated mean time-average queue length coincide for all $p > r$. Thus, the approximation performs very well for fixed r and λ and varying p. As $p \downarrow r$, the time-average queue length diverges because we obtain an M/M/1 queue with arrival and service rate of 1. On the other hand, the queuing system with diversion ($p = \infty$) provides a lower bound to the solely forward-looking flexible capacity management policy. Finally, the corresponding results for the time share the contingent capacity is active are displayed in Figure 3.13.

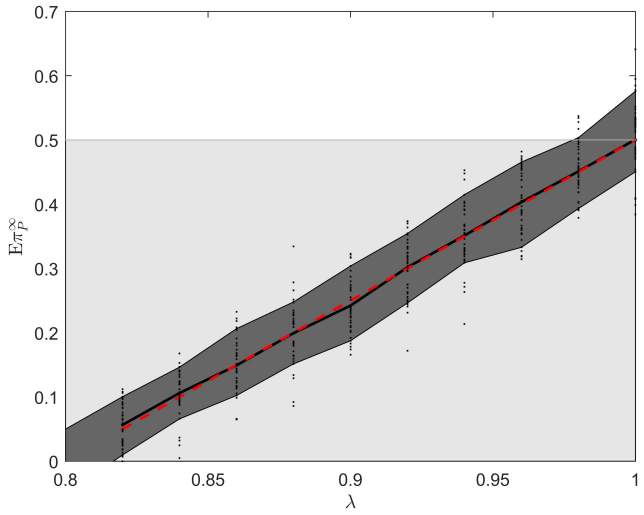


Figure 3.11: According $\mathbb{E}\pi_F^\infty$ for simulation displayed in Figure 3.10, where the black line corresponds the the (mean) simulated result and the dashed red line corresponds to the analytical result. The dark gray area corresponds to the 80% confidence interval and the light gray region indicates the feasible region.

We see that, again, the analytical result coincides with the mean of the simulated result. In conclusion, we observe that the approximate expression for the time-average queue length and the exact analytical result for the time share the contingent capacity is active are very accurate with respect to the corresponding mean results of the simulation.

In the remainder of this section we will compare the performance of the reactive and the solely forward-looking flexible capacity management policies. Therefore, the analytical results are plotted in Figure 3.14 versus the arrival rate. The first and most important difference we can observe is that the time-average queue length converges for the solely forward-looking policy as $\lambda \to 1$, while it diverges for the reactive policy. For lower arrival rates, however, the reactive policy outperforms the solely forward-looking policy. This may seem counter-intuitive at first; it makes sense, however, when considering the trajectories of the time share the contingent capacity is active. The solely forward-looking policy is strictly more efficient than the reactive policy: while

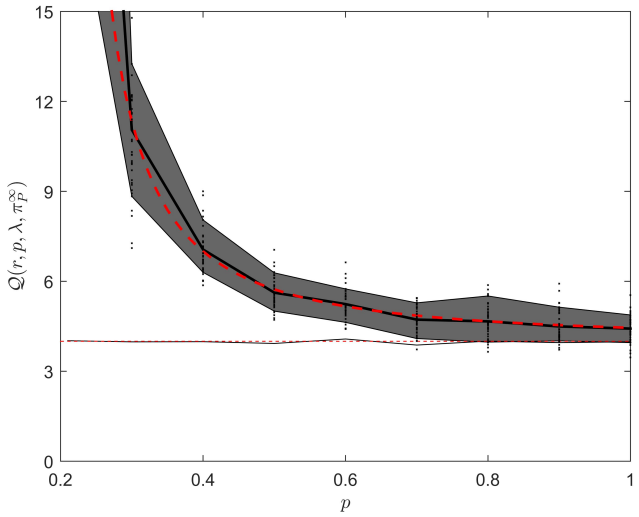


Figure 3.12: Time-average queue length for solely forward-looking policy with $r = 0.2$ and $\lambda = 1$: the dashed line corresponds to the analytical solution, the solid line to the mean of 40 simulation runs per value for p (with $N = 5,000$ events each) and the gray area to the 80% confidence interval of the simulation. The horizontal dashed line is the lower bound obtained by taking $p = \infty$ (diversion). The horizontal black line is the corresponding simulated result.

service tokens are wasted when using the reactive policy, no service tokens are wasted when applying the solely forward-looking policy. We can observe that for all $\lambda \in (1 - r, 1]$, $\mathbb{E}\pi_R^K \geq \mathbb{E}\pi_F^\infty$. The solely forward-looking policy is constructed such that it is asymptotically optimal as $\lambda \to 1$, but it does not perform well for low arrival rates. Thus, in the next section, we will merge the reactive and the solely forward-looking policy to obtain a hybrid policy. This proactive policy will combine the advantages of both policies.

Finally, it is interesting to notice that the time-average queue length is not monotonically decreasing in λ, i.e., the exists a distinct $\lambda < 1$ where $\mathcal{Q}(r, p, \lambda, \pi_F^\infty)$ is minimal. When considering a queueing system with diversion, i.e., $p = \infty$, the time-average queue length is strictly monotonically decreasing in the arrival rate. But since we need to add an additional term in the approximation for the queueing system with finite contingent capacity, which is strictly monotonically increasing in λ, the minimum is not necessarily always

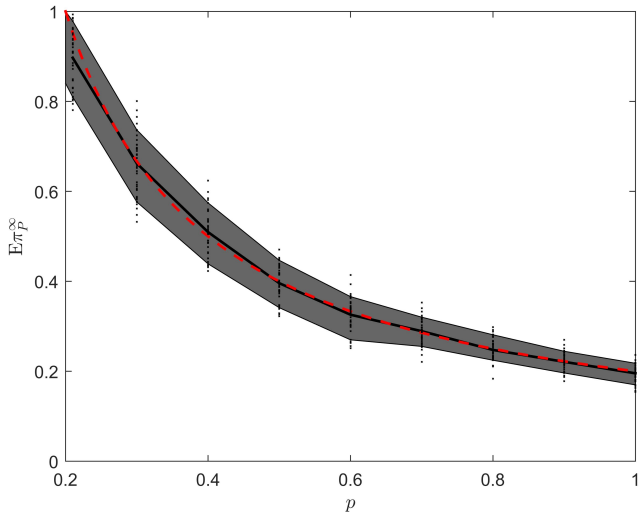


Figure 3.13: According $\mathbb{E}\pi_F^\infty$ for simulation displayed in Figure 3.12, where the black line corresponds the the (mean) simulated result and the dashed red line corresponds to the analytical result. The dark gray area corresponds to the 80% confidence interval.

obtained at $\lambda = 1$. Yet, the time-average queue length for the solely forward-looking policy with finite contingent capacity is convex in λ.

We could also use traditional optimization tools, such as a greedy algorithm, to determine when to use the contingent capacity (if we restrict the lookahead window to some large finite value). We, however, have provided a general policy and show that it is asymptotically optimal as $\lambda \to 1$, independent of the exact realization of the stochastic processes. Also, the policy enables us to find an expression for the time-average queue length, which is generally not possible if optimization is used. Additionally, our approach needs less computational time than solving an optimization model. Spencer et al. (2014) show that their policy, which determines which job should be diverted based on future information, yields the same result as applying a greedy algorithm [36]. The same holds for our problem: As $\lambda \to 1$, our policy minimizes the queue length at any point in time while, on the other hand, ensuring that no service tokens are wasted. Using the policy results in the lowest possible realization of a queue length process such that no service tokens are

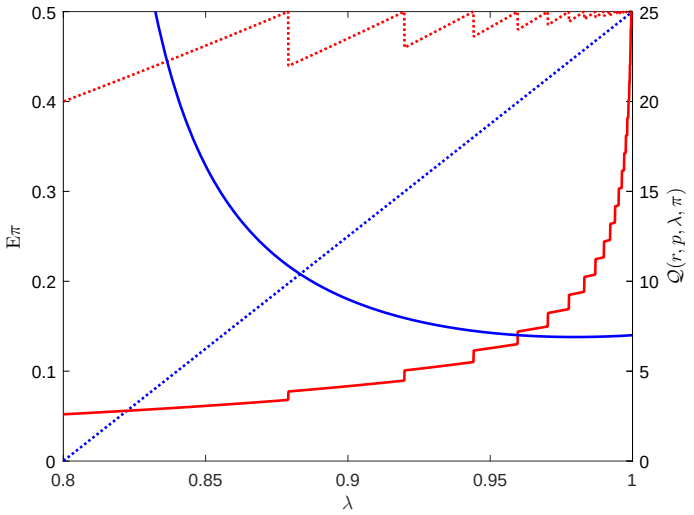


Figure 3.14: Performance of the solely forward-looking versus the reactive policy for $r = 0.2$ and $p = 0.4$. The red lines correspond to the reactive and the blue lines to the solely forward-looking policy. dashed lines illustrated the time share the high capacity is active (left y-axis). Solid lines correspond to the time-average queue length (right y-axis).

wasted. Thus, as $\lambda \to 1$, the result obtained by using the policy coincides with the result obtained through optimization.[18] Finally, as argued in Section 3.7, we can transfer the concepts developed in this section to settings with finite lookahead windows, where optimization cannot yield the best results as only a local minimum for the time interval considered can be achieved.

3.6 Proactive Capacity Control

Because the solely forward-looking policy is only asymptotically optimal as $\lambda \to 1$, we propose a hybrid policy, referred to as *proactive policy*, combining the reactive and the solely forward-looking policy such that the time-average queue length is minimized for all $\lambda \in (1-r, 1]$. The idea of the proactive policy is that the contingent capacity is activated if $Q_2(t) > Q_1(t)$ (solely forward-

[18]The argument only holds for $\lambda \to 1$, as the solely forward-looking policy is only asymptotically optimal.

looking part), but also if $Q_2(t) \leq Q_1(t)$ and $Q_2(t) > \widetilde{K}(r,p,\lambda)$ (reactive part), where $\widetilde{K} \in \mathbb{Z}_+$ denotes the proactive threshold that remains constant over time. In this section, we still assume that the lookahead window is infinite, $w = \infty$. The feasible proactive capacity control policy is defined as follows.

Theorem 3.5. *Fix $r \in (0,1)$, $p > r$ and let*

$$\widetilde{K}(r,p,\lambda) = \left\lceil \log_{\frac{1-r}{\lambda}} \frac{\lambda(\lambda-1)(\lambda+p+r-1)}{pr(r-1)} \right\rceil.$$

Then, the proactive capacity control policy

$$\pi_P^\infty(t) = \begin{cases} 1, & \text{if } Q_2(t) > \min\{Q_1(t), \widetilde{K}(r,p,\lambda)\}, \\ 0, & \text{otherwise,} \end{cases}$$

is feasible for all $\lambda \in (1-r, 1]$. More precisely, the time share the high capacity is active is given as

$$\mathbb{E}\pi_P^\infty = \frac{\lambda - 1 + r + p}{p} - \frac{\widetilde{\rho}_1^{\widetilde{K}+1} - 1}{\widetilde{\rho}_1 - 1} \widetilde{\nu}(0|\pi_R^{\widetilde{K}}) \leq \frac{r}{p},$$

with $\widetilde{\rho}_1$ and $\widetilde{\nu}(0|\pi_R^{\widetilde{K}})$ as defined below.

Proof. The contingent capacity is always active if $Q_2(t) > Q_1(t)$. Thus, we only have to look at the remaining part. The time share the contingent capacity is active in the proactive mode can be stated as

$$\mathbb{E}\pi_P^\infty = \mathbb{E}\pi_F^\infty + \mathbb{P}(Q_1 > \widetilde{K}).$$

In order to remain feasible and with $\mathbb{E}\pi_F^\infty$ as known from the previous section, we can conclude that the proactive threshold must be such that

$$\mathbb{P}(Q_1 > \widetilde{K}) \leq \frac{r}{p} - \frac{\lambda - 1 + r}{p} = \frac{1-\lambda}{p}.$$

Therefore, similar as in the proof of Theorem 3.1, we can compute the thresh-

old as

$$\widetilde{K}(r,p,\lambda) = \min\left\{n \in \mathbb{Z}_+ : \mathbb{P}(Q_1 > n) \leq \frac{1-\lambda}{p}\right\}.$$

From Lemma 3.2 we know that the queue length process Q_1 can be described as the process keeping track of the number of jobs in the system of an M/M/1 queue with arrival rate $(1-r)$ and service rate λ. When introducing the proactive threshold at which the contingent capacity is activated, the service rate will change to $(\lambda+p)$ if $Q_1(t) > \widetilde{K}$. Therefore, with $\widetilde{\rho}_1 = (1-r)/\lambda$ and $\widetilde{\rho}_2 = (1-r)/(\lambda+p)$, we can solve

$$\mathbb{P}(Q_1 > \hat{\widetilde{K}}) = 1 - \frac{\widetilde{\rho}_1^{\hat{\widetilde{K}}+1} - 1}{\widetilde{\rho}_1 - 1}\widetilde{\nu}(0|\pi_R^{\hat{\widetilde{K}}}) = \frac{1-\lambda}{p} \tag{3.10}$$

with

$$\widetilde{\nu}(0|\pi_R^{\hat{\widetilde{K}}}) = \left[\frac{\widetilde{\rho}_1^{\hat{\widetilde{K}}} - 1}{\widetilde{\rho}_1 - 1} + \frac{\widetilde{\rho}_1^{\hat{\widetilde{K}}}}{1-\widetilde{\rho}_2}\right]^{-1}$$

directly and obtain the result stated for the proactive threshold in the theorem with $\widetilde{K}(r,p,\lambda) = \lceil \hat{\widetilde{K}} \rceil$. The time share the contingent capacity is active for the proactive mode, $\mathbb{E}\pi_P^\infty$, follows directly from equation (3.10) and Theorem 3.3 and thus the proof is concluded. □

With Theorem 3.2 and the approximation (3.9), the time-average of the queue length process resulting from proactive capacity control with infinite lookahead window can be determined as

$$\mathcal{Q}(r,p,\lambda,\pi_P^\infty) = \left[\frac{(\widetilde{K}-1)\widetilde{\rho}_1^{\widetilde{K}+1} - \widetilde{K}\widetilde{\rho}_1^{\widetilde{K}} + \widetilde{\rho}_1}{(\widetilde{\rho}_1-1)^2} + \frac{\widetilde{\rho}_2 + \widetilde{K}(1-\widetilde{\rho}_2)}{(1-\widetilde{\rho}_2)^2}\widetilde{\rho}_1^{\widetilde{K}}\right]\widetilde{\nu}(0|\pi_R^{\widetilde{K}}) + \frac{(\lambda-1+r)(1-r+p)}{p(1-r+p-\lambda)}.$$

Thus, we can observe that $\lim_{\lambda\downarrow(1-r)} \mathcal{Q}(r,p,\lambda,\pi_P^\infty) = \lim_{\lambda\downarrow(1-r)} \mathcal{Q}(r,p,\lambda,\pi_R^K)$ and $\lim_{\lambda\uparrow 1} \mathcal{Q}(r,p,\lambda,\pi_P^\infty) = \lim_{\lambda\uparrow 1} \mathcal{Q}(r,p,\lambda,\pi_F^\infty)$.

An illustrative example of the processes involved in the proactive mode is displayed in Figure 3.15. We can observe that the activation of the contingent capacity is triggered by NOB arrivals and if the queue length process Q_2

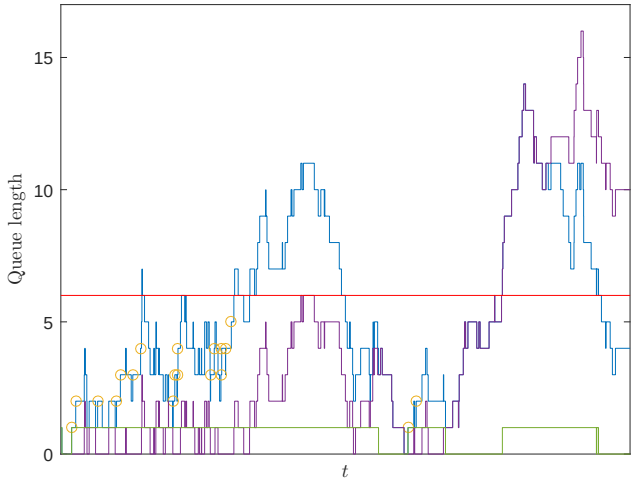


Figure 3.15: Simulation of the queue length process with proactive control with $r = 0.2$, $p = 0.4$ and $\lambda = 0.95$. The proactive threshold was computed as $\widetilde{K}(0.2, 0.4, 0.95) = 6$ (red line). The blue line corresponds to Q_2, the purple one to Q_1 and the yellow circles to NOB arrivals. We see that $\pi_P^\infty(t) = 1$ (green line) if $Q_2(t) > Q_1(t)$ (left part) or $Q_2(t) > \widetilde{K}(r, p, \lambda)$ (right part).

.

exceeds the threshold level $\widetilde{K}$.

Proposition 3.2. *For all* $\lambda \in (1 - r, 1]$,

$$\mathcal{Q}(r, p, \lambda, \pi_P^\infty) \leq \mathcal{Q}(r, p, \lambda, \pi_F^\infty),$$

i.e., the proactive policy dominates the asymptotically optimal policy.

Proof. See Appendix B.3. □

When comparing the time-average queue length resulting from the proactive policy to the one obtained with the reactive policy, it can be shown that, for

all $\lambda \in (1-r, 1]$,

$$\left[\frac{(\widetilde{\widehat{K}}-1)\widetilde{\rho}_1^{\widetilde{\widehat{K}}+1} - \widetilde{\widehat{K}}\widetilde{\rho}_1^{\widetilde{\widehat{K}}} + \widetilde{\rho}_1}{(\widetilde{\rho}_1-1)^2} + \frac{\widetilde{\rho}_2 + \widetilde{\widehat{K}}(1-\widetilde{\rho}_2)}{(1-\widetilde{\rho}_2)^2}\widetilde{\rho}_1^{\widetilde{\widehat{K}}}\right]\widetilde{\nu}(0|\pi_R^{\widetilde{\widehat{K}}})$$
$$+ \frac{(\lambda-1+r)(1-r+p)}{p(1-r+p-\lambda)}$$
$$\leq \left[\frac{(\hat{K}-1)\rho_1^{\hat{K}+1} - \hat{K}\rho_1^{\hat{K}} + \rho_1}{(\rho_1-1)^2} + \frac{\rho_2 + \hat{K}(1-\rho_2)}{(1-\rho_2)^2}\rho_1^{\hat{K}}\right]\nu(0|\pi_R^{\hat{K}}).$$

Note that it is not always true that $\mathcal{Q}(r,p,\lambda,\pi_P^\infty) \leq \mathcal{Q}(r,p,\lambda,\pi_R^K)$, $\forall\lambda \in (1-r,1]$, due to the ceiling function in the definitions of $K(r,p,\lambda) = \lceil\hat{K}\rceil$ and $\widetilde{K}(r,p,\lambda) = \lceil\widetilde{\widehat{K}}\rceil$. Thus, there exist triples $(\bar{r},\bar{p},\bar{\lambda})$, where $\bar{r} \in (0,1)$, $\bar{p} > \bar{r}$ and $\bar{\lambda} \in (1-r,1]$, such that $\mathcal{Q}(\bar{r},\bar{p},\bar{\lambda},\pi_P^\infty) > \mathcal{Q}(\bar{r},\bar{p},\bar{\lambda},\pi_R^K)$. However, the overshoot of the time-average queue length obtained using the proactive policy will always be small and occur for low arrival rates. Also, since the proactive policy is more efficient than the reactive policy, meaning that less service tokens are wasted on average, at arrival rates where these overshoots occur we find that $\mathbb{E}\pi_P^\infty(\bar{r},\bar{p},\bar{\lambda}) < \mathbb{E}\pi_R^K(\bar{r},\bar{p},\bar{\lambda})$. Finally, for high arrival rates the proactive policy necessarily always dominates the reactive policy,

$$\lim_{\lambda\to 1}\mathcal{Q}(r,p,\lambda,\pi_P^\infty) < \lim_{\lambda\to 1}\mathcal{Q}(r,p,\lambda,\pi_R^K),$$

for all $r \in (0,1)$ and $p > r$.

Finally, Figure 3.16 illustrates the performance of all three policies. We find that, with $r = 0.2$ and $p = 0.4$, the proactive policy dominates the reactive and the solely forward-looking policy for all $\lambda \in (0.8, 1]$.

3.7 Finite Lookahead Windows

So far we assumed to know all future information, $w = \infty$, which is unlikely the case with practical applications. Nevertheless, we can learn from the policies defined for infinite lookahead and modify them to work well in situations with a finite lookahead, $w < \infty$. However, in this case, we need to be careful in

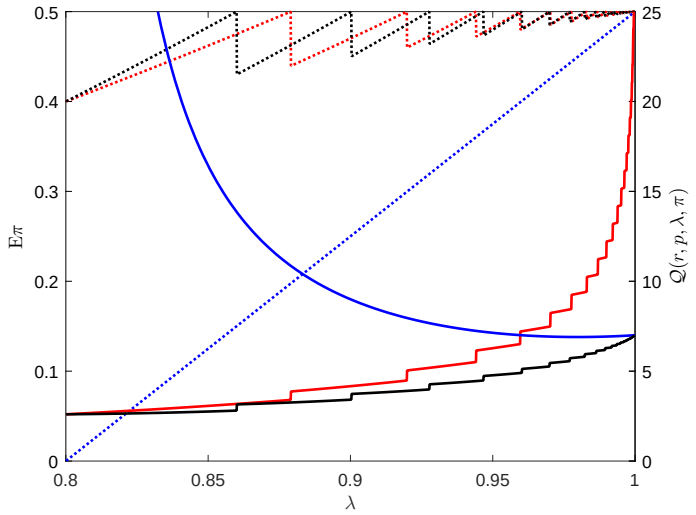


Figure 3.16: Performance of the proactive policy versus the reactive and the solely forward-looking policy for $r = 0.2$ and $p = 0.4$. The think black, red and blue lines correspond to the time-average queue lengths when applying π_R^K, π_F^∞ and π_P^∞, respectively (right y-axis). The dashed lines represent the corresponding time shares the high capacity is active (left y-axis).

order to ensure feasibility of the policy (while achieving a finite time-average queue length). In this section, we first provide insights regarding the effects of a finite lookahead window. Based thereupon, we define a set of parameters for which the proactive policy developed for infinite lookahead windows is also feasible in case of finite lookahead windows. Finally, we provide some first insights regarding potential modifications of π_P^w to ensure feasibility for all $\lambda \in (1-r, 1)$ given limited future information.

When using the proactive policy as defined in Section 3.6 for finite lookahead windows, the number of arrivals that will be characterized as NOB arrivals will exceed the number of NOB arrivals identified with infinite lookahead window, i.e., with $w < \infty$,

$$\Psi^\infty \subseteq \Psi^w \implies |\{n \in \Psi^\infty : T_n \leq t\}| \leq |\{n \in \Psi^w : T_n \leq t\}|, \ \forall t \in \mathbb{R}_+.$$

All arrivals that are characterized as NOB arrivals using an infinite lookahead window are also characterized as NOB arrivals if only a finite lookahead win-

dow is available (*real NOB* arrivals). However, some additional arrivals will be characterized as *virtual NOB* arrivals if the lookahead window is finite. Thus, if we apply the proactive policy, the time-average number of NOB arrivals will be larger than r as $\lambda \to 1$,

$$\lim_{\lambda\to 1}\lim_{t\to\infty}\frac{A^{\infty}_{\Psi}(t)}{t} = r < \lim_{\lambda\to 1}\lim_{t\to\infty}\frac{A^{w}_{\Psi}(t)}{t} = r + \chi(w)$$
$$\implies \mathbb{E}\pi^{w}_{P} > \frac{r}{p},$$

where $\chi(w) > 0$ if $w < \infty$. This implies that the proactive policy is not feasible if $w < \infty$ as $\lambda \to 1$.

Proposition 3.3. *The time-average number of arrivals characterized as NOB arrivals given a finite lookahead window of length $w < \infty$ can be computed as*

$$\lim_{t\to\infty}\frac{A^{w}_{\Psi}(t)}{t} = \lambda - (1-r)F_{1-r,\lambda}(w), \tag{3.11}$$

where

$$F_{1-r,\lambda}(w) = \int_{0}^{w}\frac{\rho_1^{-1/2}}{x}e^{-(\lambda+1-r)x}I_1\left(2x\sqrt{\lambda(1-r)}\right)\,\mathrm{d}x$$

denotes the cumulative distribution function of the busy-period of an M/M/1 queue with arrival rate λ and service rate $(1-r)$.[19]

Proof. See Appendix B.4. □

From this proposition we learn that the number of virtual NOB arrivals can be computed as $\chi(w) = (1-r)[1 - F_{1-r,\lambda}(w)]$ and we find the following insights regarding its trajectory: As $\lambda \to 1$, $\lim_{w\uparrow\infty}\chi(w) = 0$, i.e., there are no virtual NOB arrivals if the lookahead window is infinite, and $\lim_{w\downarrow 0}\chi(w) = 1 - r$, i.e., all arrivals are characterized as NOB arrivals if we apply the proactive policy and have no information regarding the future. Now, define (w_λ, λ_w) as the *feasibility double*. Given an arrival rate $\lambda < 1$, it is possible to determine the minimum length of the lookahead window w_λ required for the proactive

[19] $I_1(\cdot)$ denotes the modified Bessel function of the first kind of order one.

policy to remain feasible,

$$w_\lambda = F^{-1}_{1-r,\lambda}\left(\frac{\lambda - r}{1 - r}\right), \tag{3.12}$$

i.e., if $w \geq w_\lambda$, the policy π^w_P is feasible. We find that, as also shown by Xu (2015), as $\lambda \to 1$, there exists no feasible policy yielding a finite time-average queue length if $w < \infty$ [48].[20] Equivalently, given a lookahead window of length w, we can compute the maximum possible arrival rate λ_w such that π^w_P is feasible for all $\lambda \leq \lambda_w$ by solving

$$\lambda_w - (1 - r)F_{1-r,\lambda_w}(w) = r \tag{3.13}$$

Note that equations (3.12) and (3.13) can only be solved numerically.

Finally, we want to provide some first ideas regarding potential modifications of the proactive policy for the case where feasibility is not given, i.e., $\lambda > \lambda^w$ or $w < w_\lambda$. The contingent capacity is either activated at an arrival instant if we know, by relying on limited future information, that it is a critical arrival, or if the current queue length exceeds a static threshold level. However, "critical" arrivals, which were all arrivals we characterized as NOB in the case of an infinite lookahead window, need to be defined differently than in Section 3.5. We know that a policy is optimal if the rate of NOB job arrivals does not exceed $\lambda - (1 - r)$ (Corollary 3.1). Thus, we impose a second condition on NOB job arrivals to ensure that this rate is not exceeded given limited future information: *myopic NOB* arrivals can be defined as

$$\Psi^w = \left\{ n \in \Phi(Q_0) : \inf_{s \in [0,w]} Q_0(T_n + s) \geq Q_0(T_n) \wedge Q_0(T_n + w) \geq Q_0(T_n) + J \right\}.$$

This means that an arrival, event $n \in \Phi(Q_0)$, will be characterized as myopic NOB arrival if the initial queue length process we can compute for the time interval $[T_n, T_n + w]$ is always larger than its current value (same condition as for NOB arrivals with $w = \infty$) and if the initial queue length process at

[20] While Xu's results are obtained considering an M/M/1 queue with future information for diversion decisions, the results also apply for flexible capacity management.

time $T_n + w$, i.e., at the end of our lookahead window, is at least $J \in \mathbb{Z}_+$ jobs larger than its current value. Thus, by appropriately choosing J, we can ensure that the rate of myopic NOB job arrivals does not exceed $\lambda - (1 - r)$.

With this we have shown that the proactive policy can also be used for finite lookahead windows if certain conditions are met. Additionally, having defined a proactive policy for an infinite lookahead windows enables us to derive a modified proactive policy given limited future information. However, this policy, its properties and the resulting time-average queue length need to be evaluated numerically and we leave this problem for future research.

3.8 Conclusion and Outlook

In this paper, we focused on deriving flexible capacity control policies taking into account different information. The reactive policy only considers the current queue length, while the solely forward-looking policy only takes future information (in an infinite lookahead window) into account. Both policies are (asymptotically) optimal with respect to the families of feasible policies given no and full future information, respectively. We then combined both policies to what we termed a proactive capacity control policy: a policy that activates the contingent capacity depending on the current queue length and future information.

While, when developing the proactive policy, we assumed to know future information within an infinite lookahead window, we provided some first insights regarding finite lookahead windows. We showed under which conditions the proactive policy remains feasible and how it can be modified if these conditions are not met. We see a detailed analysis of this problem as an opportunity for future research.

In practice, if predictions regarding future job arrivals to a system are available, those predictions may be noisy. Therefore, it could be interesting to investigate the robustness of the capacity control policies developed in this paper with respect to observational noise. Xu and Chan (2016) show numerically that future information can still be valuable, also if it is noisy

(to a certain degree) for an M/M/1 queue with diversion [49]. In light of these results, it would be interesting to also analyze the robustness of flexible capacity control policies.

Finally, our results are not only relevant for the development of optimal control policies for a flexible capacity. Presumably, the optimal size of the base and contingent capacities will change when (more) future information is available. As our expression for the time-average queue length of the proactive policy is a function of r and p, the capacity sizing problem can be solved by minimizing a cost function and taking $r \in (0, 1)$ and $p > r$ as optimization variables. We consider this a promising avenue for future research.

4 Queueing with Limited Future Information

Predictive analytics foster accurate forecasting of the arrival times and service requirements of customers or jobs arriving to a queue. For example, state-of-the-art aircraft engines are equipped with sensors that transmit data in real-time and thus allow prediction of the time an engine arrives at the maintenance facility and the service requirements. In Kurz and Pibernik (2016), we develop flexible capacity control policies taking into account future information for a service provider who overhauls aircraft engines [27]. Spencer et al. (2014) and Xu and Chan (2016) derive diversion policies taking into account future information [36, 49]. However, so far all policies are tailored for the case of infinite future information, i.e., with the assumption that future information is available from now until infinity. In this paper, we investigate policies for diversion and capacity control if future information is only available within a finite lookahead window. First, we define reactive policies depending on the length of the available lookahead window. Then, we perform a numerical analysis to investigate the performance of the proactive policies with respect to the resulting mean queue length. We find that future information is valuable, also if only available within a short lookahead window. In comparison to their reactive, static threshold-based counterparts, the proactive policies lead to a reduction of the mean queue length by up to 71 % for diversion and 52 % for flexible capacity. Contrary to expectations, our analysis shows that using less future information than available can be beneficial in certain circumstances.[21]

[21] This paper has been submitted for publication [26].

4.1 Introduction and Literature Review

Predictive analytics are on the rise in all domains where data is available. The range of applications covers the prediction of individuals' behavior to cyber-physical systems. While a lot of research is focused on developing (machine learning) algorithms to extract predictive information from data, the actual usage of this newly available information is oftentimes not well-defined. Businesses capitalize only around 30 % of the financial potential of big data and advanced analytics, as found by the McKinsey Global Institute in December 2016 [23]. In this paper, we try to bridge the gap between the availability of predictive information and its employment, here for the control of queueing systems. If information regarding arrival and service times of jobs arriving at a service facility in the future, referred to *future information*, is available, we can forecast the workload. Thus, proactive actions can be defined to avoid demand peaks and therefore prevent the buildup of long queues. There exist two fundamental operating modes to cope with demand peaks:

a) Diversion–should a job be diverted or admitted to the queue, given that only a certain number of jobs can be diverted?

b) Flexible capacity–when should a contingent capacity be activated, given it can only be used for a certain share of the time?

Diversion can for example be realized by using a subcontractor. A flexible capacity can be achieved by employing temporary workers or using overtime. Thus, we will develop policies for diversion and capacity control taking future information into account.[22]

This research project was originally motivated by the overhaul of aircraft engines. State-of-the-art engines such as Rolls Royce's Trent engines are equipped with sensors that continuously measure and transmit a variety of parameters. This data can be analyzed to determine the engines' current condition and, when taking future flight plans into account, predict its future condition. Therefore, it is possible to estimate the point in time where the

[22] We assume that either diversion or flexible capacity can be employed, not both at the same time.

engine needs maintenance and the services and spare parts required. Nieto et al. (2015) develop an algorithm combining support vector machines and particle swarm optimization to predict the reliability and remaining useful lifetime of aircraft engines [32]. Sun et al. (2012) use a state space model in combination with a sequential Monte Carlo method to determine a time-to-failure distribution based on predicted degradation [40]. This newly available data can now be used by the maintenance, repair and overhaul (MRO) service provider to coordinate operations. Kurz (2016) provides more information regarding aircraft condition data and MRO operations [25]. In the paper, an optimization problem is solved to cost-optimally allocate production capacities. Additionally, it is investigated how aircraft engine condition information can be used to decrease total costs.

In general, forecasting the workload of queueing systems is a research topic emerging from various applications. Gans et al. (2015) use parametric forecasts (with updates) to estimate the arrival rate of incoming calls for a call center [19]. Sun et al. (2009) and Boyle et al. (2012) develop models to predict arrivals of patients to emergency departments [12, 41]. They show that it is not only possible to forecast mean arrival rates for future time intervals, but also make accurate predictions of the actual arrival counts for optimal resource planning. Xu and Chan (2016) go one step further and consider the actual arrival times on a single patient basis [49]. They develop admission control policies based on information regarding future patient arrivals with the objective to minimize the mean queue length. This paper advances their approach.

The paper of Xu and Chan (2016) is based on Spencer et al. (2014) [36]. They were the first to introduce the notion of *variable, but predictable* arrivals to a queueing system. The authors use future information to develop diversion policies minimizing the mean queue length. However, they only provide structural insights for the case where future information is available within an infinite lookahead window. That is, they assume that they know the exact arrival times and the speed of the server from now until infinity. Naturally, reactive policies are used to benchmark the performance of diversion policies considering future information. There exists a body of work on Markov

queueing admission control, where the decision maker makes dynamic admission / diversion decisions in order to optimize certain performance objectives. Examples for online admission policies can be found in Stidham (1985, 2002) and the references therein [37, 38].

Additionally, we also consider the control of a flexible capacity, i.e., the service provider has a base capacity and can activate a contingent capacity if required. Again, reactive or online capacity control is a well-established problem, see for example Bekker et al. (2011) or Tadj and Choudhury (2005) and the references therein [4, 42]. The first paper considering future information for the proactive control of a flexible capacity is Kurz and Pibernik (2016) [27]. The authors develop policies controlling a contingent capacity such that the mean queue length is minimized given no future information or a lookahead window of infinite length. Also, they provide some first insights regarding finite lookahead windows and our analysis advances these insights.

Subsequently, we investigate the implications of finite lookahead windows on proactive diversion and capacity control policies originally developed for infinite lookahead windows as, in practice, future information until infinity is never available. The setup of the diversion model, the capacity control model and the problem definition are provided in the next section. We develop diversion and capacity control policies depending on the length of the lookahead window available in Section 4.3. We analytically distinguish two regimes: sufficient future information, where the proactive policies can be used without modification, and insufficient future information, where the policies need to be modified. We are able to analytically derive the modified policies. However, the mean queue lengths given limited future information cannot be determined analytically. Thus, a numerical analysis performed to investigate the properties of theÂ´ policies is summarized in Section 4.4. For diversion, we find that the mean queue length can be reduced by up to 71 % if sufficient and 59 % is insufficient future information is available (compared to the reactive policy). With flexible capacity the improvements are 52 % and 32 %, respectively. Furthermore, we find that using less future information can lead to a lower mean queue length than using an infinite lookahead window. Finally, conclusion and opportunities for future research are provided in Section

4.5. All proofs are relegated to Appendix C.1.

4.2 Setup and Problem Definition

We model the facility of the service provider as an M/M/1 queue with admission control or adjustable service rate. With $r \in (0, 1/2)$, define the base capacity or service rate of the system as $1 - r \in (1/2, 1)$. Jobs arrive at the system at rate $\lambda \in (1-r, 1)$, i.e., we are considering the overload regime.[23] Denote the discrete-time diversion policies as $\{\pi_D[n] : n \in \mathbb{Z}_+\}$, $\pi_D[n] \in \{0, 1\}$, and n as the n-th event, i.e., a job arrival or departure. A job arrival n is diverted if and only if $\pi_D[n] = 1$ and admitted to the queue otherwise.

Definition 4.1 (feasible diversion policies)**.** *A diversion policy π_D is called feasible if and only if the rate of diverted jobs does not exceed r.*

This definition restricts the number of diversions that can be made. If, e.g., the service provider uses a subcontractor to avoid demand spikes, the mean number of jobs that can be diverted to the subcontractor are usually contractually specified. Thus, the total service rate of the system is less or equal than one.

The flexible capacity is modeled as a contingent capacity $p > r$ that can be added to the base capacity if needed, resulting in a high service rate $\mu_2 = 1-r+p$. Define the continuous-time capacity control policies as $\{\pi_C(t) : t \in \mathbb{R}_+\}$, $\pi_C(t) \in \{0, 1\}$. Then, the contingent capacity is active at time t if and only if $\pi_C(t) = 1$ and inactive otherwise.

Definition 4.2 (feasible capacity control policies)**.** *A capacity control policy π_C is called feasible if and only if the time share the contingent capacity is active does not exceed r/p.*

This definition restricts the time share the contingent capacity can be used as it is often the case with practical applications. Overtime is limited by labor legislation and the amount and expected usage of external contingent

[23] The benefits of future information are particularly substantial in the overload regime.

capacities are contractually specified. Again, the total service rate of the system is less or equal than one. Both feasibility definitions coincide in the sense that the number of jobs not served by the base capacity is less or equal than r. Figure 4.1 illustrates the M/M/1 queue with diversion and flexible capacity.

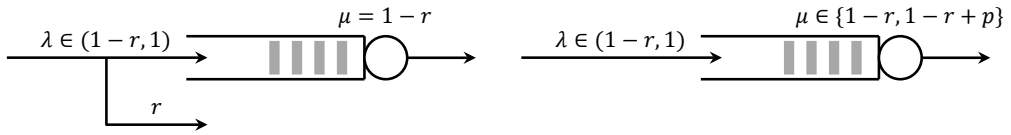


Figure 4.1: Illustration of the queue with diversion (left) and the queue with flexible capacity (right).

Let $w \in \mathbb{R}_+$ be the length of the lookahead window available. Future information within a lookahead window of length w means that we know all arrivals and their respective service requirements in the time interval $[t_0, t_0 + w]$, if we are currently at time t_0. We distinguish three operating modes:

i) No future information is available, $w = 0$: Diversion and capacity decisions are made based on the current queue length (*reactive policies*).

ii) We have perfect information regarding all jobs arriving to the system in the future, $w = \infty$: Decisions are made based on the current queue length and information regarding future job arrivals (*proactive policies with infinite lookahead window*).

iii) Future information is available within a finite lookahead window, $0 < w < \infty$: Decisions are made based on current queue length and limited future information (*proactive policies with limited future information*).

Xu and Chan (2016) and Kurz and Pibernik (2016) focus on operating modes i) and ii). In this paper, we focus on mode iii), proactive control with limited future information. When developing diversion and capacity control policies, we aim at minimizing the mean queue length while obeying the feasibility constraints. Thus, the main research questions of this paper can be stated as follows: Given a finite lookahead window, how can future information be

used to minimize the mean queue length of an M/M/1 queue with diversion / flexible capacity, while obeying the corresponding feasibility constraint? And how significant are the benefits with respect to mean queue length compared to traditional reactive policies?

We use a service token model to derive the capacity control policies. In a service token model, the randomness of the jobs' service times are transferred to the server. The server produces service tokens and the time between the generation of two tokens is exponentially distributed with mean $1/\mu$. As soon as one token has been produced, it is consumed by the first job in line (infinite waiting room) which therefore leaves the system. Thus, the queue length coincides with the number of jobs in the system, as illustrated in Figure 4.2. The service token model allows us to compute the time-average queue length

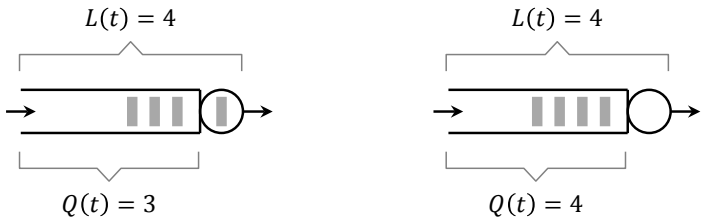


Figure 4.2: The left hand side illustrates a normal M/M/1 queue, where one job is currently being served at the server. This job is traditionally not accounted for as "in the queue" and the queue length $Q(t) \in \mathbb{Z}_+$ is one less than the number of jobs in the system $L(t) \in \mathbb{Z}_+$, $Q(t) = L(t) - 1$. The equivalent service token-based M/M/1 queue is illustrated on the right hand side. In the service token model, all jobs are waiting in the queue until the next service token has been produced, i.e., $L(t) = Q(t)$.

without considering the underlying workload process. For more information, the reader is referred to Xu and Chan (2016) and Kurz and Pibernik (2016) [27, 49].

4.3 Diversion and Capacity Control

We characterize reactive and proactive diversion and capacity control policies in the first subsection. We then investigate the effects of finite lookahead

windows on these policies and define parameters for which the policies remain feasible. Thereafter, we define modified policies if the feasibility parameters are not met.

4.3.1 Reactive and Proactive Policies

We will first introduce reactive and proactive diversion policies, advancing the policies derived by Xu and Chan (2016) [49]. We will then move on to reactive and proactive capacity control policies as developed by Kurz and Pibernik (2016) [27]. The policies are characterized as $\pi_i^{j,w}$, $i \in \{D, C\}$, $j \in \{L, K\}$, where the subscript D characterizes a diversion policy, C a capacity control policy, L and K represent the associated optimal threshold levels of a diversion or capacity control policy and w defines the length of the lookahead window. Furthermore, $Q_i^{j,w}$ characterizes the queue length process obtained when applying policy $\pi_i^{j,w}$.

For any proactive policy, the reactive counterpart serves as a benchmark for its performance. That means, we want to show that the mean queue length can be reduced by considering future information. The reactive diversion policy $\pi_D^{L,0}$ is a threshold policy, i.e., each arriving job that increases the queue length such that it reaches the threshold level will be diverted.

Definition 4.3 (reactive diversion policy)**.** $\pi_D^{L,0}[n] = 1$*, an arriving job* n *is diverted, if and only if* $Q_D^{L,0}[n] = L(r, \lambda, 0)$*, and* $\pi_D^{L,0}[n] = 0$ *otherwise.*

Thus, in order to achieve feasibility of the policy, we need to compute the largest threshold level such that the mean number of diverted jobs does not exceed r. According to Definition 4.1, a diversion policy is feasible if

$$\mathbb{E}\pi_D = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \pi_D[n] \leq \frac{r}{\lambda + (1-r)}.$$

The term $\lambda + (1 - r)$ is needed since we count all events, arrivals as well as departures from the queue. With this, the diversion rate is given as $d = \mathbb{E}\pi_D[\lambda + (1 - r)] \leq r$, i.e., the policy is feasible. As found by Xu and Chan (2016), the optimal threshold level such that the policy is feasible for all

$\lambda \in (1-r, 1)$ is given as

$$L(r, \lambda, 0) = \log_{\frac{\lambda}{1-r}} \frac{r}{1-\lambda} - 1.$$

Throughout this section, we will assume that threshold levels are integer to avoid the excessive use of floor and ceiling functions. With the threshold, the resulting mean queue length can be computed as

$$\mathbb{E}(Q_D^{L,0}) = \frac{r}{\lambda - (1-r)} L(r, \lambda, 0) + \frac{\lambda(1-\lambda) - r(1-r)}{(1-\lambda-r)^2}.$$

The drawback of this threshold-type policy is that diversion are only made when the queue length is large. The threshold level and therefore the mean queue length are increasing in the arrival rate and as $\lambda \to 1$, the mean queue length diverges. However, this changes if we take future information into account. Before stating the proactive diversion policy, we introduce some notation used throughout the remainder of this section.

Let $\{A(t) : t \in \mathbb{R}_+\}$, $A(t) \in \mathbb{Z}_+$, be the Poisson process with rate $\lambda \in (1-r, 1)$ representing the arrivals to the queue. Let $\{S_1(t) : t \in \mathbb{R}_+\}$, $S_1(t) \in \mathbb{Z}_+$, be the Poisson process representing the number of service tokens generated if only the base capacity is active, i.e., $\mu_1 = 1 - r$. Now, let $\{X_0(t) : t \in \mathbb{R}_+\}$, $X_0(t) \in \mathbb{Z}$, be the difference of the two Poisson processes,

$$X_0(t) = A(t) - S_1(t),$$

which is referred to *doubly-infinite queue*. Note that we know the past realization of these processes and, with future information, know their future realization within the time interval $[t, t+w]$. As $X_0(t)$ can have negative values but the number of jobs in a queue cannot be smaller than zero, the queue length process is computed as the corresponding reflected process,

$$\begin{aligned} Q_0(t) &= X_0(t) + \max_{s \in [0,t]} [-X_0(s)]^+ = X_0(t) + Y_0(t) \\ &= A(t) - S_1(t) + Y_0(t). \end{aligned}$$

$\{Q_0(t) : t \in \mathbb{R}_+\}$, $Q_0(t) \in \mathbb{Z}_+$ is referred to as *initial queue length process*. Let $\{Q_0[n] : n \in \mathbb{Z}_+\}$, $Q_0[n] \in \mathbb{Z}_+$ be the corresponding discrete-time version of $Q_0(t)$. Let T_n denote the time of the n-th event of Q_0 and let $\Phi(Q_0) = \{n \in \mathbb{Z}_+ : Q_0[n] > Q_0[n-1]\}$ be all events that are arrivals.

Definition 4.4 (w-critical arrivals)**.** *An arrival $n \in \Phi(Q_0)$ for which*

$$\min_{s\in[0,w]} Q_0(T_n + s) \geq Q_0(T_n)$$

is called a w-critical arrival.

With this definition, the set of w-critical arrivals is given as

$$\Psi^w = \left\{ n \in \Phi(Q_0) : \min_{s\in[0,w]} Q_0(T_n + s) \geq Q_0(T_n) \right\},$$

and the number of w-critical arrivals until time t can be computed as

$$A^w_\Psi(t) = |\{n \in \Psi^w : T_n \leq t\}|.$$

Equivalently, define $\bar{A}^w_\Psi(t) = A(t) - A^w_\Psi(t)$ as all non-w-critical arrivals. The proactive diversion policy proposed by Xu and Chan (2016) can be characterized as follows.

Definition 4.5 (proactive diversion policy)**.** *For sufficiently large w, an arriving job n is diverted, $\pi_D^{L,w}[n] = 1$, if and only if*

i) $Q_D^{L,w}[n] = L(r, \lambda, w)$, or

ii) the arrival is w-critical, $n \in \Psi^w$,

and $\pi_D^{L,w}[n] = 0$ otherwise.

We will come back to what "sufficiently large w" means in the next section. For now, we assume that the condition holds.

The continuous-time queue length process with the diversion policy stated in Definition 4.5 and with

$$\mathcal{N}(t) = \max\{n \in \mathbb{Z}_+ : T_n \leq t\}$$

can be computed as

$$\begin{aligned} Q_D^{L,w}(t) &= A(t) - S_1(t) - \sum_{n=1}^{\mathcal{N}(t)} \pi_D^{L,w}(T_n) + Y_D^{L,w}(t) \\ &= A(t) - \underbrace{\sum_{n=1}^{\mathcal{N}(t)} \mathbb{I}\{n \in \Psi^w\}}_{=A_\Psi^w(t)} - S_1(t) - \underbrace{\sum_{n=1}^{\mathcal{N}(t)} \mathbb{I}\{Q_D^{L,w}(T_n) = L\}}_{=S_D^L(t)} + Y_D^{L,w}(t) \\ &= \bar{A}_\Psi^w(t) - S_1(t) - S_D^L(t) + Y_D^{L,w}(t). \end{aligned}$$

For $\pi_D^{L,w}$ to be feasible we need

$$\lim_{t\to\infty} \frac{A_\Psi^w(t) + S_D^L(t)}{t} \leq r.$$

As shown by Kurz and Pibernik (2016),

$$\lim_{t\to\infty} \frac{A_\Psi^w(t)}{t} = \lambda - (1-r)F_{1-r,\lambda}(w), \tag{4.1}$$

with

$$F_{1-r,\lambda}(w) = \sqrt{\frac{1-r}{\lambda}} \int_0^w \frac{e^{-(\lambda+1-r)x}}{x} I_1\left(2x\sqrt{\lambda(1-r)}\right) \mathrm{d}x$$

being the cumulative distribution function of the busy-period of an M/M/1 queue with arrival rate $1-r$ and service rate λ.[24] Thus, we need to determine $L(r, \lambda, w)$ such that

$$\lim_{t\to\infty} \frac{S_D^L(t)}{t} \leq r + (1-r)F_{1-r,\lambda}(w) - \lambda. \tag{4.2}$$

As $F_{1-r,\lambda}(w)$ can only be computed numerically for all $w < \infty$, $L(r, \lambda, w)$ also needs to be computed numerically. The same holds for the mean queue length. However, for $w = \infty$, the rate of ∞-critical arrivals is given as

$$\lim_{t\to\infty} \frac{A_\Psi^\infty(t)}{t} = \lambda - (1-r),$$

[24] $I_1(\cdot)$ denotes the modified Bessel function of the first kind of order one, $I_1(y) = \sum_{i=0}^{\infty} \frac{(y/2)^{2i+1}}{k!(k+1)!}$.

which corresponds to the drift of the initial queue length process Q_0. Thus, the right hand side of inequality (4.2) simplifies to $1-\lambda$ and Xu and Chan (2016) find that the threshold can be determined as

$$L(r,\lambda,\infty) = \log_{\frac{\lambda}{1-r}} \frac{r}{1-\lambda} - 1,$$

which coincides with the optimal threshold of the reactive policy, $L(r,\lambda,\infty) = L(r,\lambda,0)$. For the resulting mean queue length we obtain

$$\mathbb{E}(Q_D^{L,\infty}) = \frac{1-\lambda}{1-\lambda-r} L(r,\lambda,\infty) - \frac{\lambda(1-\lambda)-r(1-r)}{(1-\lambda-r)^2}.$$

Figure 4.3 illustrates the diversion policies for $w=0$ and $w=\infty$. We observe

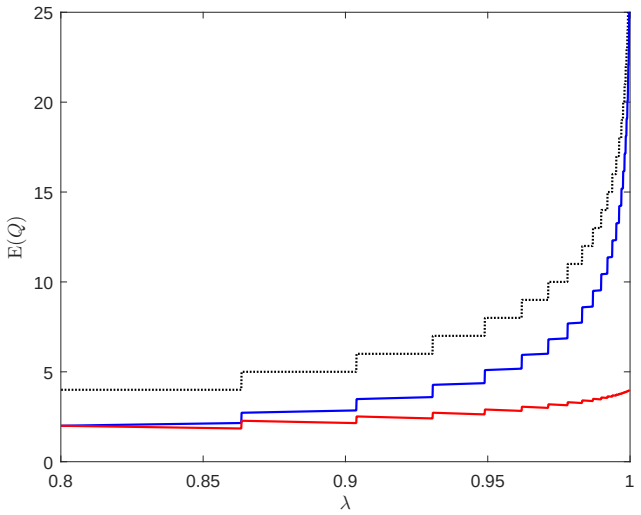


Figure 4.3: Performance of reactive and proactive diversion policies with $r = 0.2$. The blue line corresponds to the mean queue length of the reactive diversion policy, $\mathbb{E}(Q_D^{L,0})$, the red line to the proactive diversion policy with infinite lookahead, $\mathbb{E}(Q_D^{L,\infty})$, and the dashed black line to the threshold level $L(r,\lambda,0) = L(r,\lambda,\infty)$.

that the mean queue length diverges if no future information is available while it converges for an infinite lookahead window. Also, it can be shown that $\mathbb{E}(Q_D^{L,\infty}) \leq \mathbb{E}(Q_D^{L,0})$, for all $\lambda \in (1-r,1)$.

Similar as for diversion, the reactive capacity control policy is a threshold-type policy.

Definition 4.6 (reactive capacity control policy)**.** *The contingent capacity is active at time t if and only if $Q_C^{K,0}(t) > K(r, p, \lambda, 0)$.*

This means that additionally service tokens are produced at rate p when the queue length exceeds the threshold level. According to Definition 4.2, the following inequality must hold for all feasible capacity control policies:

$$\mathbb{E}\pi_C = \lim_{T\to\infty} \frac{1}{T} \int_0^T \pi_C(t)\,\mathrm{d}t \leq \frac{r}{p}.$$

As found by Kurz and Pibernik (2016), the optimal threshold level such that the policy is feasible for all $\lambda \in (1-r, 1)$ is given as

$$K(r, p, \lambda, 0) = \log_{\frac{\lambda}{1-r}} \frac{r(r-1)(\lambda + r - p - 1)}{p\lambda(1-\lambda)}.$$

The resulting mean queue length can be computed as

$$\mathbb{E}(Q_C^{K,0}) = \frac{1}{\lambda - (1-r)} \left[rK(r, p, \lambda, 0) - \lambda - \frac{r(1-r+p)}{\lambda - (1-r+p)} \right].$$

Again, we can observe that the threshold and the mean queue length diverge as $\lambda \to 1$.

Assuming future information is available, the proactive capacity control policy developed in Kurz and Pibernik (2016) is defined as follows [27].

Definition 4.7 (proactive capacity control policy)**.** *For sufficiently large w, the proactive capacity control policy is such that $\pi_C^{K,w}(t) = 1$ if and only if*

i) $Q_C^{K,w}(t) > K(r, p, \lambda, w)$*, or*

ii) $Q_C^{K,w}(t) > Q_D^{\infty,w}(t)$*,*

and $\pi_C^{K,w}(t) = 0$ otherwise.

Note that

$$Q_D^{\infty,w}(t) = \bar{A}_\Psi^w(t) - S_1(t) + Y_D^{\infty,w}(t),$$

i.e., diversions are only dependent on the forward-looking part of the policy and the diversion rate is given as the rate of w-critical arrivals, equation (4.1). The resulting queue length process when applying the capacity control policy $\pi_C^{K,w}$ is given as

$$Q_C^{K,w}(t) = A(t) - S_1(t) - S_2[\varpi_C^{K,w}(t)] + Y_C^{K,w}(t),$$

where

$$\varpi_C^{K,w}(t) = \int_0^t \pi_C^{K,w}(s)\,\mathrm{d}s$$

accounts for the total time the contingent capacity has been active until t. Feasibility requires

$$\mathbb{E}\pi_C^{K,w} = \lim_{t\to\infty} \frac{\varpi_C^{K,w}(t)}{t} \leq \frac{r}{p} \Longleftrightarrow \lim_{t\to\infty} \frac{S_2[\varpi_C^{K,w}(t)]}{t} \leq r.$$

The number of jobs that will be served by the contingent capacity due to the overshoot $Q_C^{K,w}(t) > Q_D^{\infty,w}(t)$ is again given by equation (4.1). Therefore, we again need to determine the threshold level by considering the remaining rate of service tokens that can be generated by the contingent capacity. However, the optimal threshold value can only be computed analytically for $w = \infty$,

$$K(r,p,\lambda,\infty) = \log_{\frac{1-r}{\lambda}} \frac{\lambda(\lambda-1)(\lambda+p+r-1)}{pr(r-1)}.$$

Note that, unlike for diversion, the threshold levels for reactive and proactive control do not coincide (although the difference is very small). The resulting mean queue length is given as

$$\mathbb{E}(Q_C^{K,\infty}) = \frac{1}{\lambda-(1-r)}\left[(\lambda-1)K(r,p,\lambda,\infty) + \lambda + 1 - 2r + \frac{(r-1)(p+r)}{\lambda+p+r-1}\right] + \frac{[\lambda-(1-r)](1-r+p)}{p(1-r+p-\lambda)}.$$

As shown by Spencer et al. (2014), the queue length process $Q_D^{\infty,\infty}$ can be interpreted as the process keeping track of the number of jobs in the system of an M/M/1 queue with arrival rate $1-r$ and service rate λ [36]. Thus, $Q_D^{\infty,\infty}$ and $Q_D^{\infty,w}$ are recurrent random walks as $Q_D^{\infty,w}(t) \leq Q_D^{\infty,\infty}(t)$.

Figure 4.4 illustrates the mean queue length for the reactive policy and the proactive policy with infinite lookahead window. We see that $\mathbb{E}(Q_C^{K,\infty}) \leq$

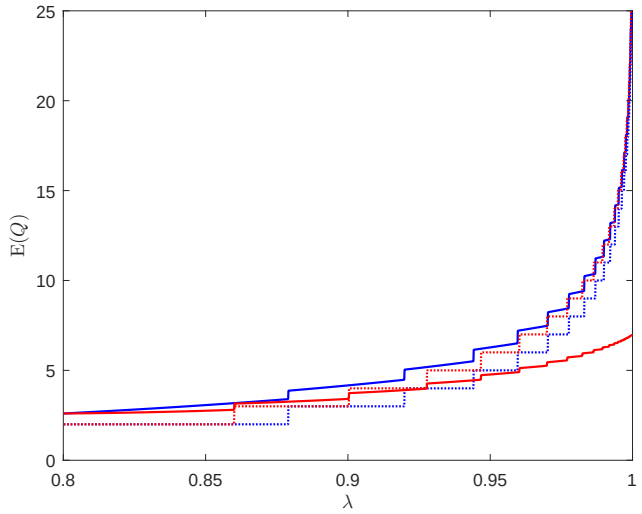


Figure 4.4: Performance of reactive and proactive capacity control policies for $r = 0.2$ and $p = 0.4$. The blue line corresponds to the mean queue length obtained with reactive capacity control, $\mathbb{E}(Q_C^{K,0})$, the red line to the proactive policy with infinite lookahead window, $\mathbb{E}(Q_C^{K,\infty})$. The dashed blue and red lines are the corresponding threshold levels $K(r,p,\lambda,0)$ and $K(r,p,\lambda,\infty)$, respectively.

$\mathbb{E}(Q_C^{K,0})$ for all $\lambda \in (1-r,1)$. Also, while the mean queue length diverges for the reactive policy, the mean queue length remains finite for the proactive policy also as $\lambda \to 1$. However, this will change if future information is only available within a finite lookahead window, as it is always the case with real-life applications.

4.3.2 Feasibility Considerations and Sufficient Future Information

In this section, we investigate the implications of limited future information, $w < \infty$, on proactive diversion and capacity control policies. We define a lower bound for the lookahead window length given a specific arrival rate such that the policies remain feasible for an appropriate choice of the threshold levels.

The proactive diversion and capacity control policies rely on w-critical arrivals. When applying the proactive policies as defined in the previous section with a lookahead window of finite length, more arrivals will be characterized as w-critical arrivals than with infinite lookahead window. Thus, for high arrival rates, the policies will become infeasible as more jobs will be diverted or more service tokens will be produced by the contingent capacity as possible for feasibility. In conclusion, given a specific arrival rate, there exists a minimum lookahead window length $\underline{w}(\lambda)$ such that $\pi_D^{\infty,\underline{w}(\lambda)}$ and $\pi_C^{\infty,\underline{w}(\lambda)}$ remain feasible.[25]

Proposition 4.1. *Given an arrival rate $\lambda \in (1-r, 1)$, the proactive diversion and capacity control policies are feasible for appropriate choices of $L(r, \lambda, w)$ and $K(r, p, \lambda, w)$ if and only if*

$$w \geq \underline{w}(\lambda) = F_{1-r,\lambda}^{-1}\left(\frac{\lambda - r}{1 - r}\right).$$

This is referred to as sufficient future information.

Proof. See Appendix C.1. □

Given the expression stated in the proposition, we can also compute the maximum possible arrival rate given a lookahead window of length w. We provide a numerical example of for the boundary of sufficient future information in Figure 4.5. Note that, as $\lambda \downarrow 0.8$, there exists a lowest possible lookahead window length given as $\lim_{\lambda \downarrow 0.8} \underline{w}(\lambda) = 6.25$.[26] The minimum lookahead window

[25] If the threshold levels are set to infinity, all diversions and service tokens produced by the contingent capacity will be due to w-critical arrivals.

[26] If we would set $w = 0$, all arrivals would be characterized as w-critical.

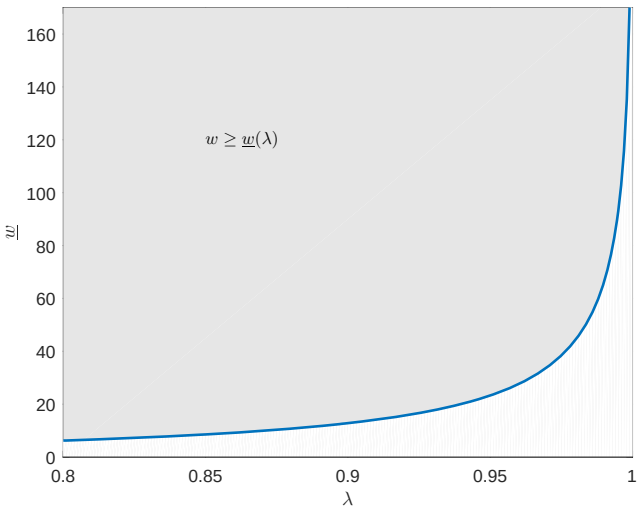


Figure 4.5: Illustration of lookahead window lengths required for sufficient future information for $r = 0.2$ and $p = 0.4$. The gray area represents the feasible region.

length increases exponentially in the arrival rate and as $\lambda \to 1$, $\underline{w}(\lambda) \to \infty$.

For finite lookahead windows, $w < \infty$,

$$\Psi^{\infty} \subseteq \Psi^{w} \implies (\Psi^{\infty} \cap \Psi^{w}) = \Psi^{\infty}.$$

Consequently, all arrivals characterized as ∞-critical are also characterized as w-critical. As $Q_D^{L,\infty}$ and $Q_C^{K,\infty}$ are recurrent random walks, the queue length processes $Q_D^{L,w}$ and $Q_C^{K,w}$ must also be recurrent.

Naturally, the threshold levels $L(r, \lambda, w)$ and $K(r, p, \lambda, w)$ need to be adapted depending on the lookahead window and as $w \downarrow \underline{w}(\lambda)$, $L \to \infty$, $K \to \infty$. However, the threshold level cannot be determined analytically as the result for the time-average queue length derived for the proactive control with infinite lookahead is not valid anymore. Thus, L and K need to be determined using simulation.

If sufficient information is available, i.e., the lookahead window is of length $w > \underline{w}(\lambda)$, we can still decide how much future information we actually want to use. This means, we can use any amount of future information in $[\underline{w}(\lambda), w]$.

We will investigate the effects of the choice of the lookahead window length used for proactive diversion and capacity control in Section 4.4.

4.3.3 Modified Policies for Insufficient Future Information

Assume that, given an arrival rate $\lambda \in (1-r,1)$, we do not have sufficient future information. A lookahead window of length $w < \underline{w}(\lambda)$ is referred to as *insufficient* or *limited future information*. We have to modify the policies developed in Section 4.3.1 and the modified policies and queue length processes will be denoted as $\widetilde{\pi}_D^{L,w}$ and $\widetilde{\pi}_C^{K,w}$, $\widetilde{Q}_D^{L,w}$ and $\widetilde{Q}_C^{K,w}$, respectively.

Definition 4.8 (myopic w-critical arrivals)**.** *With $J \in \mathbb{Z}_+$, an arrival $n \in \Phi(Q_0)$ for which*

$$\widetilde{\Psi}^w = \left\{ n \in \Phi(Q_0) : \min_{s\in[0,w]} Q_0(T_n+s) \geq Q_0(T_n) \wedge Q_0(T_n+w) \geq Q_0(T_n)+J \right\},$$

is called a myopic w-critical arrival.

This means that an arrival will be characterized as myopic w-critical, if the known future trajectory of the initial queue length process $Q_0(T_n+s)$ is always larger or equal than the current value $Q_0(T_n)$ (as for w-critical arrivals, Definition 4.4) and, additionally, the queue length at time T_n+w is at least J jobs above the value at time T_n. Subsequently, J will be referred to as *future distance*. Figure 4.6 illustrates the concept of myopic w-critical arrivals.

Let $\widetilde{A}_\Psi^w(t) = |\{n \in \widetilde{\Psi}^w : T_n \leq t\}$ be the process counting the number of myopic w-critical job arrivals until time t.

Proposition 4.2. *Given $r \in (0, 1/2)$, $\lambda \in (1-r,1)$ and $w < \underline{w}(\lambda)$, the future distance $J(r,\lambda,w)$ such that $\lim_{t\to\infty} \widetilde{A}_\Psi^w(t)/t \leq \Lambda \in [0,r]$ can be computed as*

$$J(r,\lambda,w,\Lambda) = \min\left\{ z \in \mathbb{Z}_+ : a(z) \leq \Lambda bc \right\}, \tag{4.3}$$

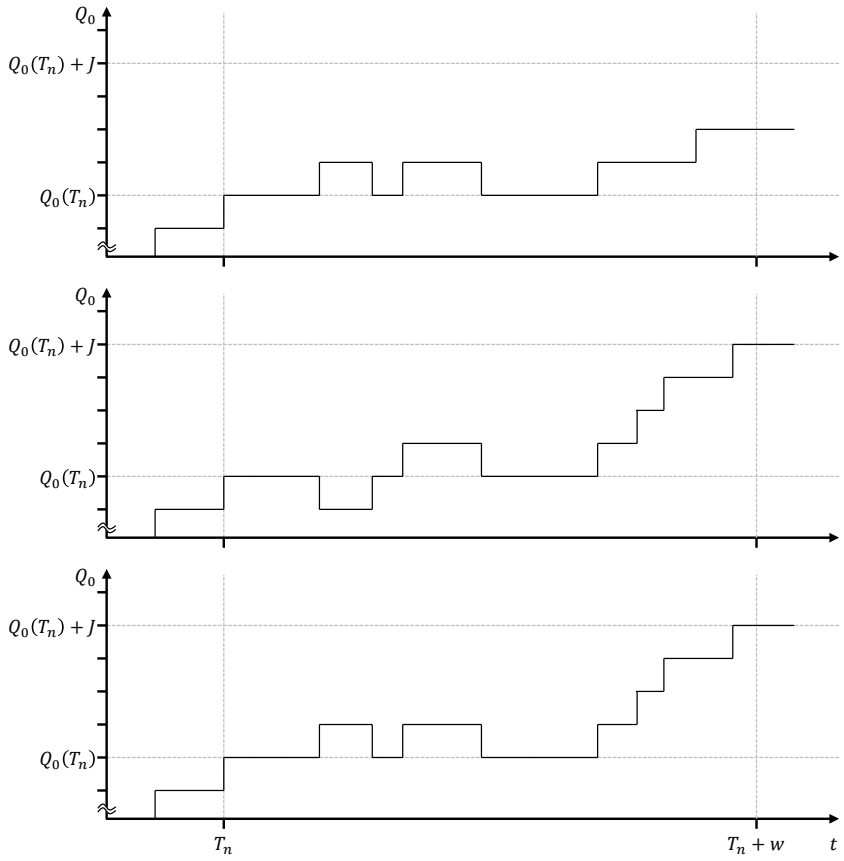


Figure 4.6: In the upper two charts, the arrival n would not be myopic w-critical, $n \notin \widetilde{\Psi}^w$, because $Q_0(T_n + w) < Q_0(T_n) + J$ in the first chart and $\min_{0\le s\le w} Q_0(T_n + s) < Q_0(T_n)$ in the second chart. In the lower chart, both criteria are met, thus $n \in \widetilde{\Psi}^w$.

with

$$
\begin{aligned}
a(z) &= \sum_{k=z}^{\infty} \left(\frac{\lambda}{1-r}\right)^{k/2} I_k\left(2w\sqrt{\lambda(1-r)}\right), \\
b &= \sum_{k=0}^{\infty} \left(\frac{\lambda}{1-r}\right)^{k/2} I_k\left(2w\sqrt{\lambda(1-r)}\right), \\
c &= [\lambda - (1-r)F_{1-r,\lambda}(w)]^{-1},
\end{aligned}
$$

and $I_k(\cdot)$ being the modified Bessel function of the first kind of order k.[27]

[27]The modified Bessel function of the first kind of order k, where $\Gamma(x)$ is the gamma

Proof. See Appendix C.1. □

With this proposition, we can compute the future distance such that the rate of myopic w-critical arrivals is given as $\Lambda \in [0, r]$. However, equation (4.3) can only be solved numerically. If we set $\Lambda = r$, the rate of w-critical arrivals will be the same as when using $\underline{w}(\lambda)$ in the previous section,

$$\lim_{t\to\infty} \frac{\widetilde{A}^{w}_{\Psi}(t)}{t} = \lim_{t\to\infty} \frac{A^{\underline{w}(\lambda)}_{\Psi}(t)}{t} = r.$$

As illustrated in Figure 4.7, we find $J = 0$ for all $w \geq \underline{w}(\lambda)$ (white area

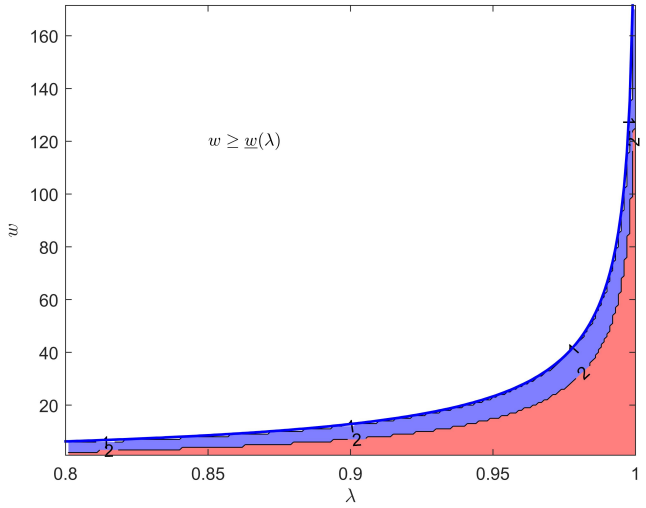


Figure 4.7: Contour plot of $J(r, \lambda, w, \Lambda)$ for $r = 0.2$ and $\Lambda = r$, different lookahead windows $w \in \{0\!:\!10\!:\!170\}$ and arrival rates $\lambda \in \{0.8\!:\!0.005\!:\!1\}$. The thick blue line corresponds to the lower bound of sufficient future information $\underline{w}(\lambda)$. $\{x : y : z\}$ denotes the set of numbers in the interval $[x, y]$ with equal spacing y, i.e., $\{1\!:\!2\!:\!7\} = \{1, 3, 5, 7\}$.

above blue line), which follows instantly from our definition of sufficient future information in the previous section. For $w < \underline{w}(\lambda)$, J increases as w decreases. However, for the example displayed in the figure, $J \leq 2$ for all $\lambda \in (0.8, 1)$ and $w \in (0, 170]$ (red area), i.e., we only need a small future distance to ensure

function, is defined as $I_k(y) = \sum_{i=0}^{\infty} \frac{(y/2)^{2i+k}}{m!\Gamma(i+k+1)}$.

that the rate of myopic w-critical arrivals is r.

In the previous section we found that $Q_D^{L,w}$ and $Q_C^{K,w}$ must be recurrent if $w \geq \underline{w}(\lambda)$. However, this does not hold if $w < \underline{w}(\lambda)$. Even if we choose $\Lambda = r$, it could be the case that $\Psi^\infty \nsubseteq \widetilde{\Psi}^w \implies (\Psi^\infty \cap \widetilde{\Psi}^w) \subset \Psi^\infty$. This means that $\widetilde{Q}_D^{L,w}$ and $\widetilde{Q}_C^{K,w}$ must not be recurrent processes, i.e., we have no insights regarding the relevance of the arrivals characterized as modified w-critical for the overall development of the queue length process. Thus, we limit the influence of future information and set $\Lambda = \lambda - (1 - r)$, which is the rate of w-critical arrivals with $w = \infty$ as detailed in Section 4.3.1. Figure 4.8 illustrates the numerically computed future distance versus the arrival rate for different lookahead windows and $\Lambda = \lambda - (1 - r)$. Note that the future

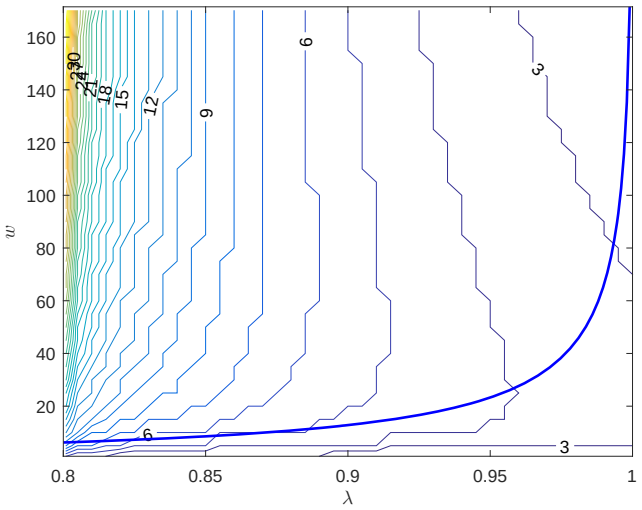


Figure 4.8: Contour plot of $J(r, \lambda, w, \Lambda)$ for $r = 0.2$ and $\Lambda = \lambda - (1 - r)$, different lookahead windows $w \in \{0 : 10 : 170\}$ and arrival rates $\lambda \in \{0.8 : 0.005 : 1\}$. The thick blue line corresponds to $\underline{w}(\lambda)$.

distance between two contour lines is given as the lower value of both lines, i.e., in the area between the contour lines 3 and 4, $J = 3$. On the first sight the contour plot may seem counterintuitive as one could expect that J needs be larger for high arrival rates. However, when looking at the figure, we need to keep the rate of arrivals we want to characterize as w-critical in mind. The rate of arrivals that will be characterized as myopic w-critical

is linearly increasing in λ, $\lim_{t\to\infty} \widetilde{A}^w_\Psi(t)/t = \lambda - (1-r)$. Additionally, the probability that the initial queue length process will drop below $Q_0(T_n)$ at any time $t > T_n + w$ is small if the drift of the initial queue length, i.e., the arrival rate, is comparatively large. For high arrival rates $\lambda \to 1$, the future distance decreases with increasing lookahead window length and as $w \to \infty$, $J \downarrow 0$.

Given $\Lambda = \lambda - (1-r)$ and $J(r, \lambda, w, \Lambda)$, Figure 4.9 shows the resulting rate of myopic w-critical job arrivals for $r = 0.2$. We observe that

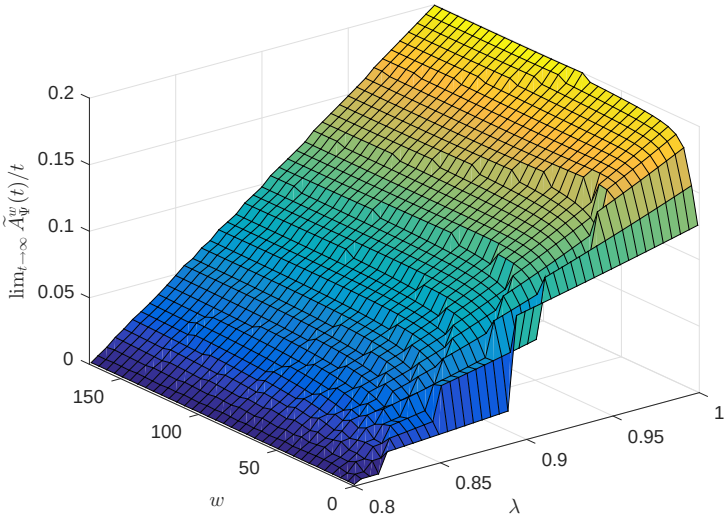


Figure 4.9: Rate of myopic w-critical job arrivals for $r = 0.2$, different lookahead windows $w \in \{0\!:\!10\!:\!170\}$ and arrival rates $\lambda \in \{0.8\!:\!0.005\!:\!1\}$.

$\lim_{t\to\infty} \widetilde{A}^w_\Psi(t)/t \leq \lambda - (1-r)$ for all λ and w. The kinks of the surface originate from J being integer.

Proposition 4.3. *The probability that a myopic w-critical arrival is also ∞-critical is given as*

$$\mathbb{P}(n \in \Psi^\infty | n \in \widetilde{\Psi}^w) = 1 - \left(\frac{1-r}{\lambda}\right)^{\vartheta(r,\lambda,w,\Lambda)},$$

with

$$\vartheta(r,\lambda,w,\Lambda) = J + \sum_{k=J}^{\infty}(k-J)e^{-w(\lambda+1-r)}\left(\frac{\lambda}{1-r}\right)^{k/2} I_k\left(2w\sqrt{\lambda(1-r)}\right).$$

Proof. See Appendix C.1. □

The resulting probability is illustrated in Figure 4.10 for $r = 0.2$. We observe

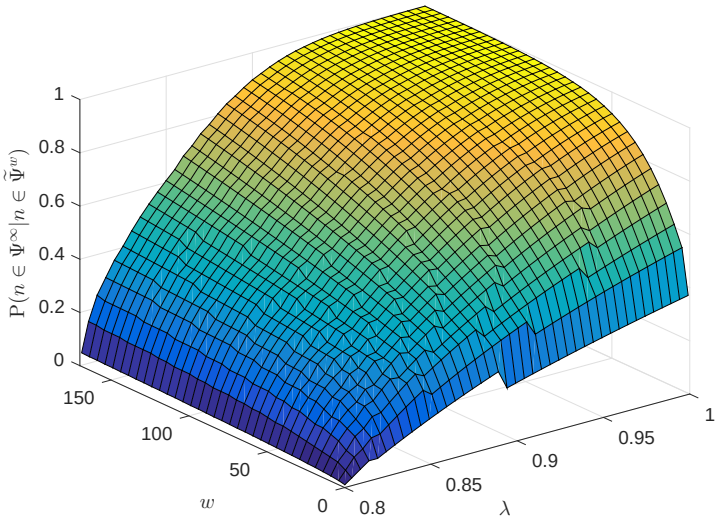


Figure 4.10: Probability that myopic w-critical arrivals is ∞-critical for $\Lambda = \lambda - (1-r)$, lookahead windows $w \in \{0:10:170\}$ and arrival rates $\lambda \in \{0.8:0.005:1\}$.

that the probability that a myopic w-critical arrival is ∞-critical increases in the arrival rate and the lookahead window length. With $J \in \mathbb{Z}_+$, we find that

$$\begin{aligned}
&\mathbb{E}[Q_0(T_n+w) - Q_0(T_n) | Q_0(T_n+w) - Q_0(T_n) \geq 0] \\
&\leq \mathbb{E}[Q_0(T_n+w) - Q_0(T_n) | Q_0(T_n+w) - Q_0(T_n) \geq J] \\
&\implies \mathbb{P}(n \in \Psi^{\infty} | n \in \Psi^{w}) \leq \mathbb{P}(n \in \Psi^{\infty} | n \in \widetilde{\Psi}^{w}),
\end{aligned}$$

i.e., the probability that a myopic w-critical arrival is an ∞-critical arrival

cannot be less than for a w-critical arrival given the same lookahead window length $w < \underline{w}(\lambda)$.

Finally, we can state the modified proactive diversion and capacity control policies for limited future information.

Definition 4.9 (modified proactive diversion policy)**.** *For $w < \underline{w}(\lambda)$ and given $J(r, \lambda, w, \Lambda)$, a job is diverted, $\widetilde{\pi}_D^{L,w}(t) = 1$, if and only if*

i) $\widetilde{Q}_D^{L,w}(t) = L(r, \lambda, w)$, *or*

ii) the arrival is myopic w-critical, $t \in \{T_n\}_{n \in \widetilde{\Psi}^w}$,

and $\widetilde{\pi}_D^{L,w}(t) = 0$ otherwise.

The modified proactive diversion policy essentially remains the same as for sufficient future information, only that we use myopic w-critical arrivals. The reactive threshold level $L(r, \lambda, w)$ needs to be determined numerically after diverting the myopic w-critical arrivals.

Definition 4.10 (modified proactive capacity control policy)**.** *For $w < \underline{w}(\lambda)$ and given $J(r, \lambda, w, \Lambda)$, the modified proactive capacity control policy is such that $\widetilde{\pi}_C^{K,w}(t) = 1$ if and only if*

i) $\widetilde{Q}_C^{K,w}(t) > K(r, p, \lambda, w)$, *or*

ii) $\widetilde{Q}_C^{K,w}(t) > \widetilde{Q}_D^{\infty,w}(t)$,

and $\widetilde{\pi}_C^{K,w}(t) = 0$ otherwise.

The modified proactive capacity control policy is again composed of two threshold levels. The dynamic threshold $\widetilde{Q}_D^{\infty,w}$, the queue length process if all myopic w-critical arrivals are diverted and no reactive threshold is applied, and the static threshold level $K(r, p, \lambda, w)$, which again needs to be determined numerically as illustrated in the next section.

4.4 Numerical Analysis

In this section, we perform a simulation study to investigate (i) if limited future information can be used to reduce the mean queue length when employed

effectively and (ii) how high the benefit is. Three conjectures are formulated to investigate the structural properties of the different policies developed and summarized in this paper. More specifically, we want to verify if policies using (limited) future information lead to a lower mean queue length than the reactive policy. Additionally, we want to compare proactive policies with different lookahead window lengths. Thus, we state two conjectures to compare the proactive policy with limited but sufficient future information to the reactive policy and the proactive policy with infinite future information. The third conjecture compares the modified proactive policy for insufficient future information with the reactive policy.[28] After explaining the experimental setup, we analyze the results of the simulation to verify the conjectures and investigate the benefits of the policies with respect to the mean queue length in numerical terms.

4.4.1 Expected Results

If sufficient future information is available, we expect that the mean queue length is less or equal than if the reactive policy is applied.

Conjecture 4.1. *Given any $r \in (0, 1/2)$, $p > r$ and sufficient future information, $w \geq \underline{w}(\lambda)$,*

$$\mathbb{E}(Q_D^{L,w}) \leq \mathbb{E}(Q_D^{L,0}) \quad \textit{and} \quad \mathbb{E}(Q_C^{K,w}) \leq \mathbb{E}(Q_C^{K,0})$$

for all $\lambda \in (1-r, 1)$, i.e., the proactive diversion and capacity control policies with limited future information outperform their reactive counterparts.

Thus, we want to show that future information is valuable, also if it is only available within a finite lookahead window. Secondly, we want to investigate the performance of the proactive policies for different lookahead window lengths. Therefore, we compare the mean queue length obtained with infinite lookahead with the one obtained using just sufficient future information, $w = \underline{w}(\lambda)$. Intuitively, we expect that the policy using infinite future information outperforms the policy using limited future information.

[28] Each conjecture considers diversion and flexible capacity.

Conjecture 4.2. *Given any $r \in (0, 1/2)$, $p > r$ and minimum sufficient future information, $w = \underline{w}(\lambda)$,*

$$\mathbb{E}(Q_D^{L,\infty}) \leq \mathbb{E}(Q_D^{L,\underline{w}(\lambda)}) \quad \textit{and} \quad \mathbb{E}(Q_C^{K,\infty}) \leq \mathbb{E}(Q_C^{K,\underline{w}(\lambda)})$$

for all $\lambda \in (1-r, 1)$, i.e., the proactive diversion and capacity control policies with infinite lookahead window outperform the policies with limited future information.

Thus, we want to investigate if "the more information, the better" holds for proactive queue control. Eventually, we want to investigate if the modified proactive policies using insufficient future information outperform their reactive, static threshold-based counterparts.

Conjecture 4.3. *Given any $r \in (0, 1/2)$, $p > r$ and insufficient future information, $w < \underline{w}(\lambda)$,*

$$\mathbb{E}(\widetilde{Q}_D^{L,w}) \leq \mathbb{E}(Q_D^{L,0}) \quad \textit{and} \quad \mathbb{E}(\widetilde{Q}_C^{K,w}) \leq \mathbb{E}(Q_C^{K,0})$$

for all $\lambda \in (1-r, 1)$, i.e., the modified proactive diversion and capacity control policies outperform their reactive counterparts.

As one would never use a modified proactive policy if $w \geq \underline{w}(\lambda)$, we only need to compare the resulting mean queue length with the one obtained using no future information. This conjecture investigates the benefit of limited future information given short lookahead windows.

4.4.2 Experimental Setup

The simulation was performed with MATLAB. Each simulation run contains $N = 10.000$ job arrivals (approximately $2N$ events) and we use 50 simulation runs per parameter combination (arrival rate and lookahead window length). Thus, means and standard deviations per parameter combination were obtained by first taking the average of the queue length process of one simulation run and then averaging the 50 simulation runs.

The numerical results provided for the diversion model in the next section were obtained using $r = 0.2$, i.e., a base capacity of $1 - r = 0.8$. Thus, as demanded by Definition 4.1 for feasibility, the diversion rate must be $d = \mathbb{E}\pi_D[\lambda + (1 - r)] \leq r = 0.2$. We use the same base capacity for the flexible capacity model and a contingent capacity $p = 0.4$. Therefore, according to Definition 4.2, the time share the contingent capacity is active must be $\mathbb{E}\pi_C \leq r/p = 1/2$ for feasibility. Simulations with other parameter combinations than $r = 0.2$ and $p = 0.4$ yield the same structural insights as presented in the next section. In conclusion, we assume that the results can be generalized for all combinations of $r \in (0, 1/2)$ and $p > r$.

4.4.3 Results and Interpretation

We will first investigate the benefits of future information for diversion and then move on to capacity control. To numerically investigate if Conjecture 4.1 is true for diversion, we compare the simulated mean queue lengths obtained for the reactive policy and for the proactive policy with sufficient future information. More specifically, we used $w(\lambda) = 2\underline{w}(\lambda)$ to obtain the results displayed in Figure 4.11.[29] The figure suggests that Conjecture 4.1 is true for diversion: Using future information within a finite but sufficiently long lookahead window dominates using no future information with respect to the mean queue length, $\mathbb{E}(Q_D^{L,2\underline{w}(\lambda)}) < \mathbb{E}(Q_D^{L,0})$, for all $\lambda \in \hat{\lambda}$. More specifically, the benefit of future information is smaller for low arrival rates ($\lambda = 0.81$: mean queue length reduced by 39 %), but becomes substantial as the arrival rate increases ($\lambda = 0.99$: improvement of 67 %). The right hand side of the figure shows that both policies are feasible for all arrival rates considered, i.e., the diversion rate is always less or equal than 0.2. It is also interesting to note that we used $L(r, \lambda, 2\underline{w}(\lambda)) = L(r, \lambda, 0) = L(r, \lambda, \infty)$ for all $\lambda \in \hat{\lambda}$.

Figure 4.12 illustrates the results for Conjecture 4.2 and diversion. For each $\lambda \in \hat{\lambda}$, we used the minimum lookahead window such that the policy remains feasible if we set $L(r, \lambda, \underline{w}(\lambda)) = \infty$. The figure suggests that Conjec-

[29] For each $\lambda \in \hat{\lambda}$, 50 simulation runs were used to compute the mean. The range of results corresponds to the minimum and maximum value of these 50 simulation runs.

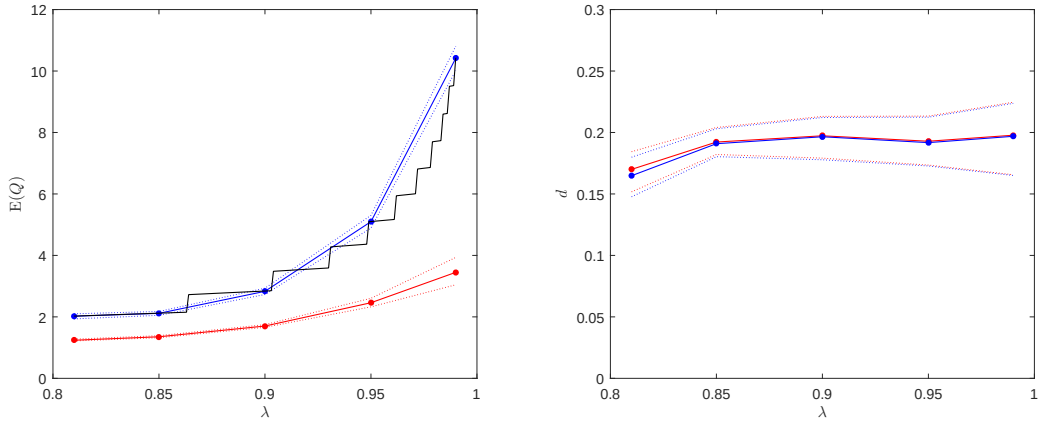


Figure 4.11: Test of Conjecture 4.1 for diversion. The black line corresponds to the analytical results for the reactive policy, the blue line to the mean of the simulated reactive and the red line to the mean of the simulated proactive policy with lookahead window $w(\lambda) = 2\underline{w}(\lambda)$ and $\lambda \in \hat{\lambda} = \{0.81, 0.85, 0.9, 0.95, 0.99\}$. The thin dotted lines show the range of results obtained for 50 simulation runs.

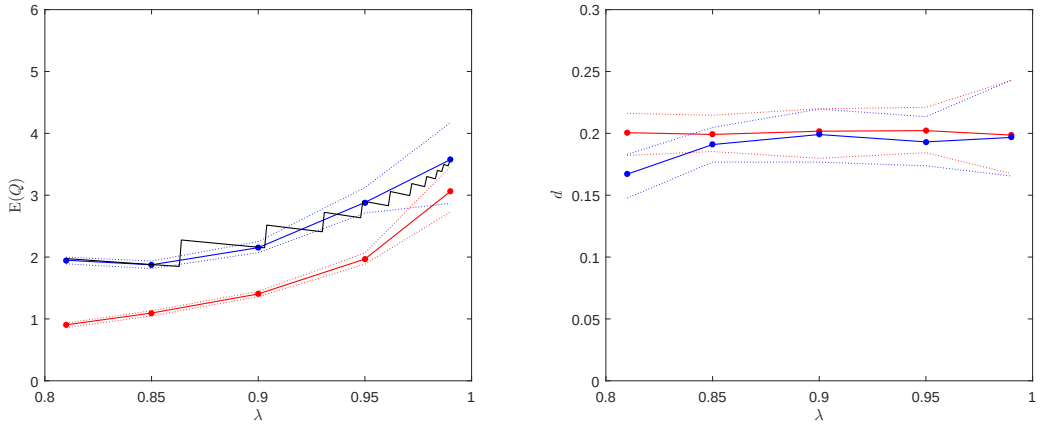


Figure 4.12: Test of Conjecture 4.2 for diversion. The black line corresponds to the analytical result for the proactive policy with infinite lookahead window, the blue line to the mean of the corresponding simulated policy and the red line to the mean of the simulated proactive policy with lookahead window $w(\lambda) = \underline{w}(\lambda)$.

ture 4.2 is false for diversion: The mean queue length is lower if we use less future information, $\mathbb{E}(Q_D^{\infty,\underline{w}(\lambda)}) < \mathbb{E}(Q_D^{L,\infty})$, for all $\lambda \in \hat{\lambda}$. Table 4.1 displays

the numerical results of the mean queue lengths for the different lookahead windows. Indeed, we find that the mean queue length is increasing in the

λ	$\underline{w}(\lambda)$	$\mathbb{E}(Q_D^{L,0})$	$\mathbb{E}(Q_D^{L,\underline{w}})$	$\mathbb{E}(Q_D^{L,2\underline{w}})$	$\mathbb{E}(Q_D^{L,\infty})$
0.81	6.6	2.02±0.01 (2.02)	0.90±0.01	1.23±0.03	1.95±0.03 (1.98)
0.85	8.6	2.11±0.01 (2.12)	1.09±0.02	1.35±0.03	1.88±0.03 (1.88)
0.90	12.8	2.83±0.02 (2.84)	1.40±0.02	1.70±0.04	2.15±0.05 (2.16)
0.95	23.3	5.09±0.06 (5.10)	1.97±0.04	2.46±0.09	2.88±0.09 (2.90)
0.99	66.3	10.41±0.19 (10.43)	3.06±0.14	3.45±0.21	3.57±0.24 (3.57)

Table 4.1: Comparison of mean queue lengths for diversion policies with different lookahead windows. The numbers after the ± correspond to standard deviation of the 50 simulation runs per $\lambda \in \hat{\lambda}$. The numbers in brackets are the analytical results if available.

lookahead window as long as $w \geq \underline{w}(\lambda)$. We find two potential explanations for this counter-intuitive result. First, as it can be seen when looking at the diversion rates displayed in Figure 4.12, if we use a lookahead window of length $\underline{w}(\lambda)$ and a threshold of ∞, the diversion rate is by definition always given as $d = r$. On the other hand, due to the integer constraint for the threshold, the proactive policy with infinite lookahead window but finite threshold cannot always exploit all diversions that could theoretically be made. Secondly, especially for low arrival rates, when using the proactive policy with infinite lookahead window, only few diversions are made due to future information. More specifically, the rate of ∞-critical arrivals is given as $\lambda-(1-r)$, while the rate of diversions made because the queue length process reaches the threshold level is given as $1 - \lambda$. However, if we use a lookahead window of length $\underline{w}(\lambda)$, all ∞-critical arrival are diverted plus others that are also critical when considering a shorter period of time. Thus, all diversions made are based on future information, also if the queue length process is low, and not only if the queue length process reaches a (rather large) threshold level. However, note that $\mathbb{E}(Q_D^{\infty,\underline{w}(\lambda)})$ converges to $\mathbb{E}(Q_D^{L,\infty})$ as $\lambda \to 1$, since $\lim_{\lambda\to 1} \underline{w}(\lambda) = \infty$. If we compare $\mathbb{E}(Q_D^{L,0})$ and $\mathbb{E}(Q_D^{L,\underline{w}})$ for $\lambda = 0.99$, the mean queue length can be reduced by up to 71 % if sufficient future information is available.

Finally, Figure 4.13 illustrates the benefit of limited insufficient future information. The figure suggests that Conjecture 4.3 is true for diversion:

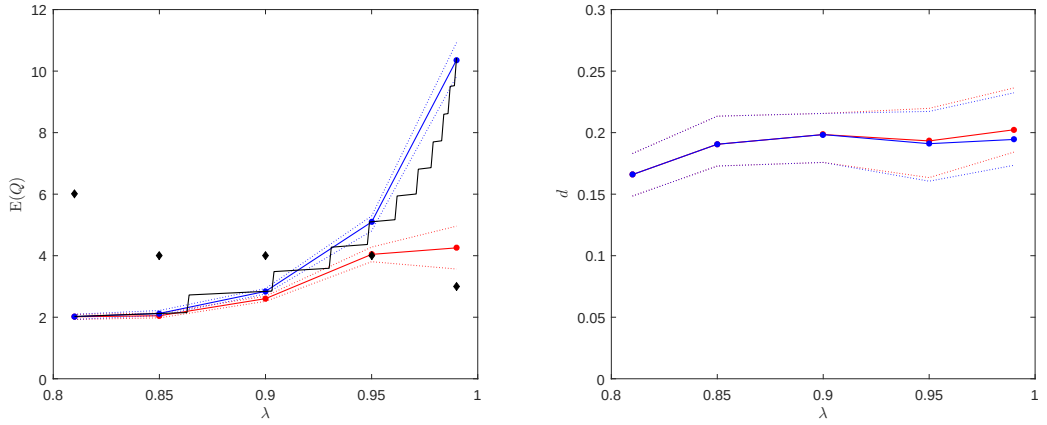


Figure 4.13: Test of Conjecture 4.3 for diversion. The black line corresponds to the analytical results for the reactive policy, the blue line to the mean of the simulated reactive and the red line to the mean of the simulated modified proactive policy with lookahead window $w(\lambda) = \underline{w}(\lambda)/2$. The black diamonds correspond to the future distance $J(r, \lambda, w, \Lambda)$ with $\Lambda = \lambda - (1 - r)$.

The mean queue length obtained when applying the modified proactive policy is less or equal than the one obtained when applying the reactive policy, $\mathbb{E}(\widetilde{Q}_D^{L,\underline{w}(\lambda)/2}) \leq \mathbb{E}(Q_D^{L,0})$, for all $\lambda \in \hat{\lambda}$. It is interesting to observe that $J(r, \lambda, w, \Lambda)$ decreases in λ. This is due to our definition of Λ, which only allows for a limited number of diversions based on future information for low arrival rates and increases in λ. Thus, as the number of diversions made based on future information is low for low arrival rates, the two mean queue lengths coincide. However, with increasing arrival rate, the difference between the two lines increases as more future information is used. For $\lambda = 0.99$ the mean queue length obtained applying the modified proactive queue length is 59 % lower than if the reactive policy is used. Note that we again used $L(r, \lambda, \underline{w}(\lambda)/2) = L(r, \lambda, \infty) = L(r, \lambda, 0)$.

In conclusion, the numerical analysis suggests that Conjectures 4.1 and 4.3 are true for diversion. However, Conjecture 4.2 is false, i.e., using less future information than available can be beneficial with respect to the mean

queue length.

We subsequently test the three conjectures for capacity control. Figure 4.14 shows the results for Conjecture 4.1. The results suggest that Con-

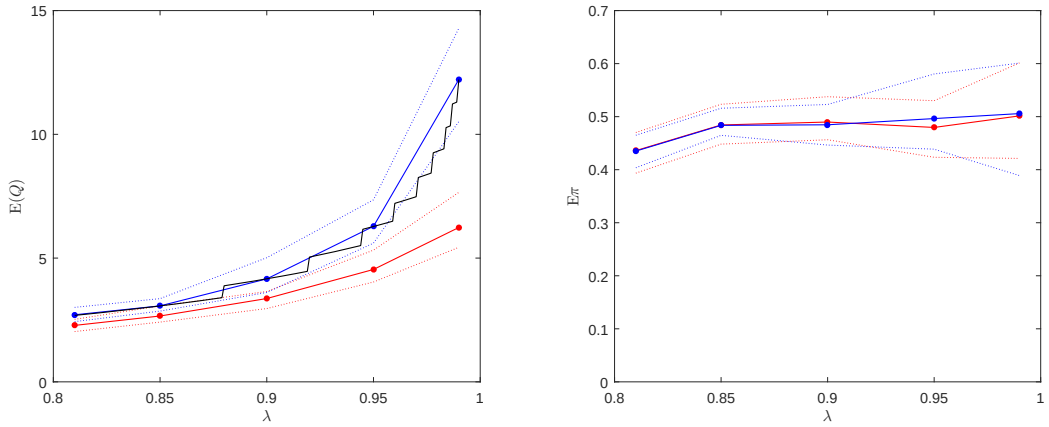


Figure 4.14: Test of Conjecture 4.1 for capacity control. The black line corresponds to the analytical result for the reactive policy, the blue line to the mean of the simulated reactive and the red line to mean of the simulated proactive policy with $w(\lambda) = 2\underline{w}(\lambda)$.

jecture 4.1 is true: Using future information available within a limited but sufficiently large lookahead window outperforms the reactive capacity control policy, $\mathbb{E}(Q_C^{K,2\underline{w}(\lambda)}) < \mathbb{E}(Q_C^{K,0})$, for all $\lambda \in \hat{\lambda}$. The right hand side of the figure shows that both policies are feasible, i.e., that the time share the contingent capacity is active is always less or equal than r/p. Note that we used $K(r, p, \lambda, 2\underline{w}(\lambda)) = K(r, p, \lambda, \infty)$ to obtain these results.

Next, Figure 4.15 shows the results for Conjecture 4.2 and capacity control. Similar as for diversion, we can observe that the conjecture is false: Using less future information can lead to a lower mean queue length than using future information until infinity, $\mathbb{E}(Q_C^{\infty,\underline{w}(\lambda)}) < \mathbb{E}(Q_C^{K,\infty})$, for all $\lambda \in \hat{\lambda}$. The mean queue lengths for different arrival rates and lookahead window lengths are also displayed in Table 4.2. As for diversion, we observe that the queue length increases in the lookahead window as long as $w \geq \underline{w}(\lambda)$. The reasons for this counter-intuitive finding follow from the equivalent result for the diversion policy. With a lookahead window of length $\underline{w}(\lambda)$ and no

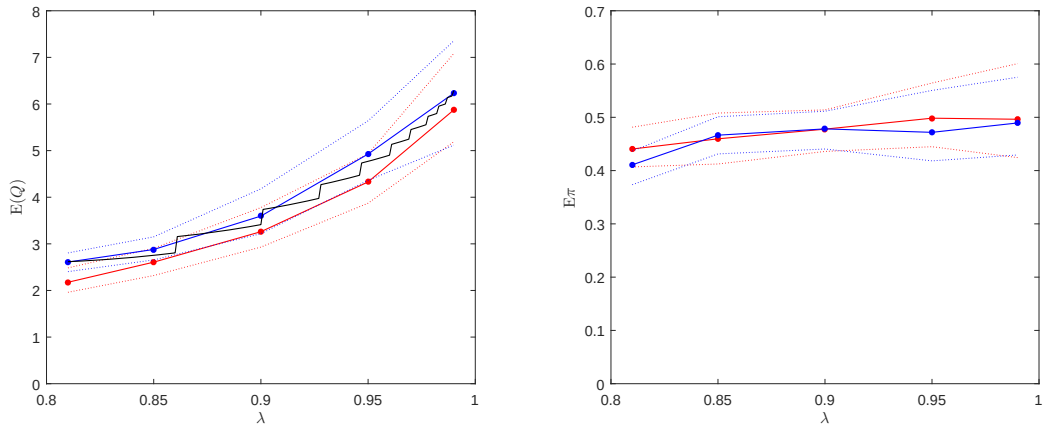


Figure 4.15: Test of Conjecture 4.2 for capacity control. The black line corresponds to the analytical result for the proactive policy with infinite lookahead window, the blue line to the mean of its simulated counterpart and the red line to mean of the simulated proactive policy with $w(\lambda) = \underline{w}(\lambda)$.

λ	$\underline{w}(\lambda)$	$\mathbb{E}(Q_C^{K,0})$	$\mathbb{E}(Q_C^{K,\underline{w}})$	$\mathbb{E}(Q_C^{K,2\underline{w}})$	$\mathbb{E}(Q_C^{K,\infty})$
0.81	6.6	2.71±0.12 (2.68)	2.18±0.11	2.28±0.12	2.61±0.10 (2.62)
0.85	8.6	3.07±0.11 (3.07)	2.61±0.14	2.66±0.14	2.88±0.11 (2.75)
0.90	12.8	4.16±0.29 (4.16)	3.26±0.20	3.37±0.18	3.60±0.19 (3.41)
0.95	23.3	6.28±0.38 (6.28)	4.33±0.29	4.54±0.31	4.93±0.28 (4.77)
0.99	66.3	12.22±0.69 (12.21)	5.87±0.46	6.22±0.48	6.23±0.44 (6.29)

Table 4.2: Comparison of mean queue lengths for capacity control policies with different lookahead windows.

threshold ($K = \infty$), the contingent capacity is active at time t if and only if $Q_C^{\infty,\underline{w}(\lambda)}(t) > Q_D^{\infty,\underline{w}(\lambda)}(t)$. Thus, as $Q_D^{\infty,\underline{w}(\lambda)}(t) \leq Q_D^{\infty,w}(t)$, for all $w > \underline{w}(\lambda)$, the contingent capacity is generally activated earlier if the lookahead window is shorter. Also, the amount of time the contingent capacity is active due to future information is larger if the lookahead window is short and the contingent capacity is not only activated if the queue length process reaches a rather large threshold value. However, when comparing these results with the results obtained for the equivalent diversion problem, we observe that the benefits of using a shorter lookahead window are not as distinct for capacity

control as for diversion. This is due to the tardiness of the flexible capacity model. While a job is simply deleted from the queue if it is diverted, it needs to be served in the flexible capacity model, and the rate of additional service tokens produced by the contingent capacity is low, here $p = 0.4$. The largest improvement can be observed if we consider $\mathbb{E}(Q_C^{K,0})$ and $\mathbb{E}(Q_C^{K,\underline{w}})$, where the mean queue length is reduced by 52 % for $\lambda = 0.99$ (versus 71 % for diversion).

Finally, Figure 4.16 displays the numerical results to validate Conjecture 4.3 for capacity control. Note that we used $K(r, p, \lambda, \underline{w}(\lambda)/2) = K(r, p, \lambda, \infty)$,

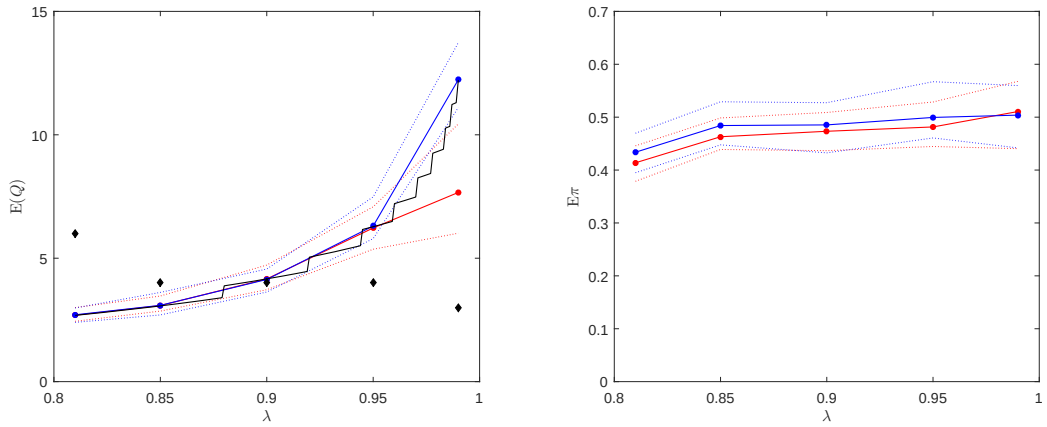


Figure 4.16: Test of Conjecture 4.3 for capacity control. The black line corresponds to the analytical results for the reactive policy, the blue line to the mean of the simulated reactive and the red line to the mean of the simulated modified proactive policy with lookahead window $w(\lambda) = \underline{w}(\lambda)/2$. The black diamonds correspond to the future distance $J(r, \lambda, w, \Lambda)$ with $\Lambda = \lambda - (1 - r)$.

for all $\lambda \in \hat{\lambda} \setminus \{0.95\}$. Only for $\lambda = 0.95$ we used $K(r, p, \lambda, \underline{w}(\lambda)/2) = K(r, p, \lambda, \infty) - 1$. As for diversion, we find that Conjecture 4.3 is true for capacity control: Using limited insufficient future information can be beneficial, $\mathbb{E}(\widetilde{Q}_C^{K,\underline{w}(\lambda)/2}) \leq \mathbb{E}(Q_C^{K,0})$, for all $\lambda \in \hat{\lambda}$. We see that the mean queue lengths (more or less) coincide for $\lambda \in \{0.81, 0.85, 0.9\}$. Starting at $\lambda = 0.95$, the mean queue length is lower for the modified proactive policy than for the reactive policy. Again, the benefits of future information are not as pronounced for capacity control as for diversion (an improvement of 37 % for $\lambda = 0.99$ versus 59 % for diversion), but still significant.

In conclusion, we found that Conjectures 4.1 and 4.3 are true for capacity control and diversion. However, for both operating modes, Conjecture 4.2 is false. This means that proactive policies with limited future information can outperform proactive policies with infinite future information. This interesting result is good news for all practical applications of proactive diversion and capacity control policies, as future information is never available within an infinite lookahead window in real life.

We used $r = 0.2$ and $p = 0.4$ to obtain these numerical examples. However, the analysis sheds light on structural properties of the policies and the same insights can be obtained for all other choices of $r \in (0, 1/2)$ and $p > r$.

4.5 Conclusion and Outlook

We have shown that predictive information extracted from data can be used to prescribe actions that lead to improved operating and therefore economic performance. More precisely, future information can be employed to reduce the mean queue length for diversion and capacity control, also if it is only available within a finite lookahead window (of sufficient or insufficient length). Additionally, we found an interesting result: "less can be more", i.e., using shorter lookahead windows can be better than using (infinitely) long lookahead windows because the proactive policy then actually uses more of the future information and does not rely on static threshold values, especially for low arrival rates. Therefore, it can even be beneficial to artificially limit the lookahead window length to obtain low mean queue lengths. Also, since future information until infinity is never available in real life, this result indicates that proactive diversion and capacity control policies can be valuable for practical applications.

There exist multiple opportunities for future research regarding proactive diversion and capacity control policies. First of all, the analytical treatment of the proactive policies with limited future information and the resulting mean queue lengths remain open for further analysis. Also, there could be different policies yielding even better results as we did not investigate the

performance of the policies with respect to optimality. Another opportunity for future research is to integrate diversion and capacity control: How should a system that allows for diversion and different capacity levels be controlled if future information is available? Finally, it could be interesting to investigate the robustness of the proactive policies with respect to observational noise as future information will in practical applications mostly not be available with infinite precision.

5 Summary and Conclusion

This dissertation has aimed at investigating the effects of future demand information on capacity decisions. To this end, it has been shown that, if used appropriately, predictive information can lead to significant cost and waiting time improvements. As argued in Chapter 1, businesses in the age of cyber-physical systems and digitization focus on acquiring and analyzing data. However, they often miss to employ the data to prescribe actions that help to improve their operating and economic performance. Therefore, in this thesis, this last step of actually using data has been made: Motivated by an aircraft engine MRO (maintenance, repair and overhaul) service facility, we have investigated how information about future job arrivals can be used to plan and control the capacity of the facility. Additionally, we have developed insights regarding the benefits that can be expected if this newly available data is used.

In the first part of the thesis, Chapter 2, the aircraft engine MRO's facility is described as a network of GI/G/1 queues. First, the optimal capacity per workstation is determined such that total costs composed of capacity and tardiness-related penalty costs are minimized. Furthermore, a framework for collaborative maintenance management is developed that leads to advanced information regarding future engine arrivals. We find that this information can be used to reduce mean service times as well as the variability of service and interarrival times, resulting in reduced total maintenance costs. The article thus provides a decision-making framework whether a company should invest in the required technologies that are needed to obtain advanced information or not.

The second article, Chapter 3, deals with actually using information regarding the arrival and service times of specific future jobs for the control of a

flexible capacity. We assume that the MRO can switch between a base and a high capacity. Thus, proactive capacity control policies taking future information into account are developed and compared to a reactive policy that relies on a static threshold. It turns out that using future information can significantly reduce the mean queue length, i.e., the jobs' waiting time, especially for high arrival rates. Thus, using future information available through cyber-physical systems to prescribe capacity decisions leads to improved operating performance. However, we assume that future information is available within a lookahead window of infinite length, which is never the case with practice applications.

Thus, the third article, Chapter 4, extends the second article for the case of finite lookahead windows. Additionally, we do not only consider the control of a flexible capacity, but also admission control: Which jobs should be diverted, if a given number of jobs can be diverted? This corresponds to the case where the MRO uses a subcontractor to avoid demand spikes. Proactive diversion and capacity control policies that can also be used for limited future information are developed. A numerical analysis suggests that the policies relying on lookahead windows of finite length still outperform their reactive counterparts. Interestingly, the results also indicate that using less future information can be beneficial compared to using infinite future information, especially for lower arrival rates.

In conclusion, all three parts of the thesis support the hypothesis that information generated by cyber-physical systems can be used to optimize operating and economic performance of enterprises if applied effectively. Although the motivation of this thesis is a very concrete application, the models developed and theoretical insights gained are very holistic and can be applied to a variety of settings where capacity decisions meet data.

There exist many opportunities for future research that can be categorized in practical and more theoretical topics. On the practical side, finding suitable applications and adapting the models presented here to meet the applications' requirements could be one opportunity. Furthermore, it could be interesting to solve a similar problem through appropriate optimization models rather than queues. Given that information about future jobs is available, can we

achieve the same or even better results using deterministic optimization tools? Additionally, we have assumed that the information regarding future jobs is perfect. Therefore, it could be interesting to investigate how robust the policies are with respect to observational noise, which is almost always present in practice.

On the theoretical end, integrating capacity control and diversion could be an interesting avenue for further research. How can the service provider benefit from future information if he can switch between two (or more) capacity levels and divert jobs? In addition, given a proactive diversion or capacity control policy, another opportunity is to determine optimal capacities or the maximum diversion rate minimizing a cost function composed of capacity and waiting costs?

Generally speaking, queueing with future information is a very recent and still to a large extent unstudied topic that just started to gain attention. Thus, since future information will become more and more available through the snowballing use of predictive models, this topic will attract and be of interest for an increasing professional and academic audience.

A Appendix of Chapter 2

A.1 Proofs of Section 2.3

Proof of Proposition 2.1. In the first part of the proof, we show that the mean sojourn time approximation function $S(\mu)$, $S : \mathbf{R} \mapsto \mathbf{R}$, as defined in (2.3) is convex and strictly monotonically decreasing in $\mathbf{dom}\ S = \{\mu \mid \mu > \lambda\} \subset \mathbf{R}_{++}$. We start by showing convexity of $g(\mu)$ as defined in (2.2). For $ca > 1$, $g(\mu) = 1$, convexity follows instantly. For $ca \in [0, 1]$, since $g(\mu) = \exp\{h(\mu)\}$ is convex if $h(\mu)$ is convex, we need to show that

$$h(\mu) = \frac{-2(1 - ca)(\mu - \lambda)}{3\lambda(ca + cs)}$$

is convex. With some reformulation, $h(\mu)$ can be rewritten as an affine function $h(\mu) = a\mu + b$ with

$$a = \frac{2(ca - 1)}{3\lambda(ca + cs)} \qquad \text{and} \qquad b = \frac{2(1 - ca)}{3(ca + cs)}.$$

Since affine functions are always convex, $h(\mu)$ and therefore $g(\mu)$ are convex for $ca \in [0, 1]$. Furthermore, since $h(\mu), h'(\mu) \leq 0$ (h' denotes the first-order derivative of h) for $ca \in [0, 1]$ and $g(\mu) = 1$ for $ca > 1$, $g(\mu)$ is non-increasing (or monotonically decreasing) in $\mathbf{dom}\ S$.

Next, since the product of two convex, non-increasing functions on an interval in $\mathbf{R}$ is convex, we show that $f(\mu) = 1/(\mu(\mu - \lambda))$ is convex and non-increasing in $\mathbf{dom}\ S$. Convexity can easily be shown by checking the second-order condition, $\nabla^2 f(\mu) \succeq 0,\ \forall \mu \in \mathbf{dom}\ S$. Since $f(\mu)$ is a composition $f = h \circ g$ of two scalar functions $h(g) = 1/g$ and $g(\mu) = \mu(\mu - \lambda)$ in $\mathbf{dom}\ S \subset \mathbf{R}_{++}$,

the second-order condition can be expressed with the first- and second-order derivatives f', g', h' and f'', g'', h'', respectively, as $f''(\mu) = h''(g(\mu))g'(\mu)^2 + h'(g(\mu))g''(\mu) \geq 0$,

$$f''(\mu) = \underbrace{\frac{2(2\mu-\lambda)^2}{(\mu^2-\mu\lambda)^3}}_{>0} - \underbrace{\frac{2}{(\mu^2-\mu\lambda)^2}}_{>0} \overset{?}{\geq} 0.$$

In order for $f''(\mu)$ to be positive, the first term must be larger than the second term for all $\mu \in \mathbf{dom}\ S$. By multiplying with $(\mu^2-\mu\lambda)^3$ on both sides and some rearranging we end up with the inequality $3\mu^2 - \lambda\mu + \lambda \geq 0$ which strictly holds since $\mu > \lambda$ and $\lambda > 0$. Therefore, $f(\mu)$ is strictly convex. As the first derivative of $f(\mu)$ is negative,

$$f'(\mu) = -\frac{2\mu-\lambda}{(\mu^2-\mu\lambda)^2} < 0 \qquad \forall \mu \in \mathbf{dom}\ S,$$

the function is strictly monotonically decreasing. Therefore, we can conclude that $1/(\mu(\mu-\lambda))g(\lambda,\mu,ca,cs)$ is convex and strictly monotonically decreasing in μ.

Since the multiplication with a non-negative multiplier $(ca+cs)\lambda/2 > 0$ reveals convexity, the sum of two convex functions is convex and $\partial^2/\partial\mu^2(1/\mu) > 0$, $\forall \mu \in \mathbf{dom}\ S$, we can conclude that the approximation function for the mean sojourn time S is convex in $\mathbf{dom}\ S$. Since $\partial/\partial\mu(1/\mu) < 0$, $\forall \mu \in \mathbf{dom}\ S$, it also follows that S is strictly monotonically decreasing. This concludes the first part of the proof.

The sum of separable approximations of the mean sojourn times $S_j(\mu_j)$ is convex in $\mathbf{dom}\ \mathcal{C} = \bigcup_{j\in\mathcal{J}} \mathbf{dom}\ S_j$ since all functions $(S_j(\cdot))_{j\in\mathcal{J}}$ are convex functions. For all $e \in \mathcal{E}$, subtraction of a constant S_e^T, application of the $\max\{0,\cdot\}$ function and multiplication with a non-negative multiplier $\gamma_e > 0$ retains convexity. Therefore, with the linear convex left hand side of the objective function, $c^\top\mu = \sum_{j\in\mathcal{J}} c_j\mu_j$, as well as the convex right hand side, the objective function is convex in $\mathbf{dom}\ \mathcal{C}$.

Finally, since the left hand side of the objective function is strictly monotonically increasing and the right hand side is monotonically decreasing (not

strictly due to $\max\{0, \cdot\}$) in **dom** $\mathcal{C}$, at least one optimal vector $\mu^\star \in$ **dom** $\mathcal{C}$ exists which concludes the proof. □

Proof of Proposition 2.2. From the definition of *Case 1* we know that $\mathcal{E}_= = \emptyset$, i.e,

$$\nexists\, e \in \mathcal{E},\ \sum_{j \in \mathcal{J}_e} S_j(\mu_j^{\star 1}) = S_e^T.$$

Taking this and relation (2.4) into account, the capacity allocation (CAP) can be expressed as

$$\text{minimize} \quad \mathcal{C}^1(\mu) = c^\top \mu + \sum_{e \in \mathcal{E}_>} \gamma_e \Big[\sum_{j \in \mathcal{J}_e} S_j(\mu_j) - S_e^T \Big]$$

and the subdifferential (2.5) of $\mathcal{C}^1(\mu)$ simplifies to the ordinary gradient,

$$\partial\, \mathcal{C}^1(\mu) = \nabla \mathcal{C}^1(\mu) = c^\top + \sum_{e \in \mathcal{E}_>} \gamma_e \nabla \sum_{j \in \mathcal{J}_e} S_j(\mu_j).$$

In order to find potential solutions of the optimization problem, $\nabla \mathcal{C}^1(\mu^{\star 1}) = 0$, we need to find vectors $\mu^{\star 1} \in$ **dom** $\mathcal{C}$ that solve

$$c^\top = - \sum_{e \in \mathcal{E}_>} \gamma_e \nabla \sum_{j \in \mathcal{J}_e} S_j(\mu_j^{\star 1}) \tag{A.1}$$

for all possible combinations of $\mathcal{E}_>$. Since $\mathcal{E}_> \neq \emptyset$ and we demand $\bigcup_{e \in \mathcal{E}_>} \mathcal{J}_e = \mathcal{J}$, this are at most $E^2 - 1$ possible combinations, each requiring the solution of J equations. Therefore, we need to solve at most $J(E^2 - 1)$ equations in total.

Since the solution of Equation (A.1) is independent of S_e^T, we need to check the conditions

$$\forall e \in \mathcal{E}_>, \sum_{j \in \mathcal{J}_e} S_j(\mu_j^{\star 1}) > S_e^T$$

and

$$\forall e \in \mathcal{E}_<, \sum_{j \in \mathcal{J}_e} S_j(\mu_j^{\star 1}) < S_e^T$$

in order to verify that a solution obtained through the approach above is a

feasible minimizer of the problem. This concludes the proof. □

Proof of Proposition 2.3. The proof is based on the Karush-Kuhn-Tucker (KKT) conditions of a reformulation of the optimization problem. If the minimizer of $\mathcal{C}$ is such that $\mathcal{E}_{=} = \mathcal{E}$ (i.e., $\mathcal{E}_{>} \cup \mathcal{E}_{<} = \emptyset$), problem (CAP) can be reformulated as the optimization problem

$$\begin{aligned}
&\text{minimize} && c^\top \mu && && \text{(A.2)}\\
&\text{subject to} && \sum_{j \in \mathcal{J}_e} S_j(\mu_j) - S_e^T = 0 && \forall e \in \mathcal{E} && \text{(A.3)}\\
& && \lambda_j - \mu_j \leq 0 && \forall j \in \mathcal{J} && \text{(A.4)}
\end{aligned}$$

If we refer to the equality constraints (A.3) as $h_e(\mu),\ e \in \mathcal{E}$, the inequality constraints (A.4) as $f_j(\mu),\ j \in \mathcal{J}$, the objective function as $f_0(\mu) = c^\top \mu$ and if we introduce the KKT multipliers η_j and ν_j, the KKT conditions of the optimization problem are defined as

$$\begin{aligned}
f_j(\mu^{\star 2}) &\leq 0, && \forall j \in J && \text{(A.5)}\\
h_e(\mu^{\star 2}) &= 0, && \forall e \in \mathcal{E} && \text{(A.6)}\\
\eta_j^\star &\geq 0, && \forall j \in J && \text{(A.7)}\\
\eta_j^\star f_j(\mu^{\star 2}) &= 0, && \forall j \in J && \text{(A.8)}\\
\nabla f_0(\mu^{\star 2}) + \sum_{j=1}^{J} \eta_j^\star \nabla f_j(\mu^{\star 2}) &+ \sum_{e=1}^{E} \nu_e^\star \nabla h_e(\mu^{\star 2}) = 0. && && \text{(A.9)}
\end{aligned}$$

We instantly see that $\eta_j^\star = 0,\ \forall j \in J$ from the complementary slackness condition (A.8) and constraint (A.4). Therefore, condition (A.9) for the optimal solution of the optimization problem can be expressed as

$$c^\top = -\sum_{e=1}^{E} \nu_i^\star \nabla \sum_{j \in \mathcal{J}_e} S_j(\mu_j^{\star 2}). \tag{A.10}$$

With this equation and condition (A.6) we get a system of $J + E$ equations which can be solved to obtain the primal and dual optimal points $\mu^{\star 2} \in \mathbf{dom}\ \mathcal{C}$ and $(\eta^\star, \nu^\star)$, respectively (where $\eta^\star = 0$).

From constraint (A.6) we know that $\mathcal{E}_> = \emptyset$. Another condition to verify that the obtained solution is optimal is

$$c_j \leq -\sum_{e \in \mathcal{E}^j} \gamma_e \frac{\partial S_j(\mu_j^{\star 2})}{\partial \mu_j} \qquad \forall j \in \mathcal{J}, \tag{A.11}$$

where $\mathcal{E}^j \subseteq \mathcal{E}$ is the set of product families where node j is contained in all paths, $j \in \mathcal{J}_e,\ \forall e \in \mathcal{E}^j$. This condition is necessary since $\mu^{\star 2}$ needs to be such that there is a kink at $\mathcal{C}(\mu^{\star 2})$ where $0 \in \partial \mathcal{C}(\mu^{\star 2})$ (also, γ_e is not considered in the optimization problem). This concludes the proof. □

A.2 Proofs of Section 2.4

Proof of Proposition 2.4. We start by showing that the penalty cost term is weakly decreasing in the improvement. On all paths without penalty we instantly know that

$$\sum_{j \in \mathcal{J}_e} S_j(\xi_j \widetilde{\mu}_j) = \sum_{j \in \mathcal{J}_e} S_j(\mu_j^\star) = S_e^T \quad \forall e \in \mathcal{E}_=.$$

Therefore, we need to show that on all paths where a penalty occurs

$$\sum_{j \in \mathcal{J}_e} S_j(\xi_j \widetilde{\mu}_j) < \sum_{j \in \mathcal{J}_e} S_j(\mu_j^\star) \quad \forall e \in \mathcal{E}_>$$

holds. Since $S_j(\cdot)$ is strictly monotonically decreasing we can prove the statement by showing that $\xi_j \widetilde{\mu}_j > \mu_j^\star,\ \forall j \in \mathcal{J}$ where $\xi_j > 1$.

Following the proof of Proposition 2.2 for the updated cost function $\widetilde{\mathcal{C}}(\mu)$ as defined in (UCAP) we get

$$c^\top = -\sum_{e \in \mathcal{E}_>} \gamma_e \nabla \sum_{j \in \mathcal{J}_e} S_j(\xi_j \widetilde{\mu}_j) \tag{A.12}$$

for the solution of the differentiable optimization problem. Intuitively, when comparing this result to Equation (A.1), one could think that the updated *Case 1*-optimal solution is given by $\widetilde{\mu} = A^{-1}\mu^{\star 1}$. Although we can simply

replace μ_j by $\xi_j\mu_j$ to compute $S_j(\xi_j\mu_j)$, we have to explicitly evaluate the first-order derivative of $S_j(\xi_j\mu_j)$ in order to compute the gradient of the updated cost function. The original first-order derivative of $S_j(\mu_j)$ is given by

$$\frac{\partial S_j(\mu_j)}{\partial \mu_j} = -\left\{\frac{1}{\mu_j^2} + \frac{(ca_j + cs_j)\lambda_j}{2\mu_j(\mu_j - \lambda_j)} \exp\left\{-\frac{2(1-ca_j)(\mu_j - \lambda_j)}{3\lambda_j(ca_j + cs_j)}\right\}\right. \\ \left. \cdot \left[\frac{2\mu_j - \lambda_j}{\mu_j(\mu_j - \lambda_j)} + \frac{2(1-ca_j)}{3\lambda_j(ca_j + cs_j)}\right]\right\}. \tag{A.13}$$

The updated first-order derivative of $S_j(\xi_j\mu_j) = \widetilde{S}_j(\mu_j)$ is given by

$$\frac{\partial \widetilde{S}_j(\mu_j)}{\partial \mu_j} = -\left\{\frac{1}{\xi_j\mu_j^2} + \frac{(ca_j + cs_j)\lambda_j}{2\xi_j\mu_j(\xi_j\mu_j - \lambda_j)} \exp\left\{-\frac{2(1-ca_j)(\xi_j\mu_j - \lambda_j)}{3\lambda_j(ca_j + cs_j)}\right\}\right. \\ \left. \cdot \left[\frac{2\xi_j^2\mu_j - \xi_j\lambda_j}{\xi_j\mu_j(\xi_j\mu_j - \lambda_j)} + \frac{2\xi_j(1-ca_j)}{3\lambda_j(ca_j + cs_j)}\right]\right\}. \tag{A.14}$$

Since $\partial_{\mu_j} S_j(\mu_j)$ and $\partial_{\mu_j} \widetilde{S}_j(\mu_j)$ are differentiable and strictly monotonically increasing in $\mathbf{dom}\ \widetilde{\mathcal{C}}$ we can show that $\xi_j\widetilde{\mu}_j > \mu_j^\star,\ \forall j \in \mathcal{J}_e,\ e \in \mathcal{E}_>$ where $\xi_j > 1$ by inserting $\mu_j = \mu_j^\star/\xi_j$ in equation (A.14) and comparing it to equation (A.13) with $\mu_j = \mu_j^\star$. If

$$-\frac{\partial \widetilde{S}_j\left(\frac{\mu_j^\star}{\xi_j}\right)}{\partial \mu_j} > -\frac{\partial S_j(\mu_j^\star)}{\partial \mu_j} \tag{A.15}$$

holds we can conclude that $\widetilde{\mu}_j > \mu_j^\star/\xi_j$ or $\xi_j\widetilde{\mu}_j > \mu_j^\star$. The substitutions yield

$$-\frac{\partial S_j(\mu_j^\star)}{\partial \mu_j} = \underbrace{\frac{1}{(\mu_j^\star)^2}}_{(1a)} + \underbrace{\frac{(ca_j + cs_j)\lambda_j}{2\mu_j^\star(\mu_j^\star - \lambda_j)}}_{(1b)} \underbrace{\exp\left\{-\frac{2(1-ca_j)(\mu_j^\star - \lambda_j)}{3\lambda_j(ca_j + cs_j)}\right\}}_{(1c)} \\ \cdot \left[\underbrace{\frac{2\mu_j^\star - \lambda_j}{\mu_j^\star(\mu_j^\star - \lambda_j)}}_{(1d)} + \underbrace{\frac{2(1-ca_j)}{3\lambda_j(ca_j + cs_j)}}_{(1e)}\right]$$

and

$$-\frac{\partial \widetilde{S}_j\left(\frac{\mu_j^\star}{\xi_j}\right)}{\partial \mu_j} = \underbrace{\frac{\xi_j}{(\mu_j^\star)^2}}_{(2a)} + \underbrace{\frac{(ca_j+cs_j)\lambda_j}{2\mu_j^\star(\mu_j^\star-\lambda_j)}}_{(2b)} \underbrace{\exp\left\{-\frac{2(1-ca_j)(\mu_j^\star-\lambda_j)}{3\lambda_j(ca_j+cs_j)}\right\}}_{(2c)} \cdot \left[\underbrace{\frac{2\xi_j\mu_j^\star-\xi_j\lambda_j}{\mu_j^\star(\mu_j^\star-\lambda_j)}}_{(2d)} + \underbrace{\frac{2\xi_j(1-ca_j)}{3\lambda_j(ca_j+cs_j)}}_{(2e)}\right].$$

When comparing the terms in the two equations we see that $(2a) > (1a)$, $(2b) = (1b)$, $(2c) = (1c)$, $(2d) > (1d)$ and $(2e) > (1e)$ if we assume $\xi_j > 1$. Therefore, inequality (A.15) holds and the first part of the proof is concluded.

The second part of the proof is similar to the first part. In order to show that $c^\top\widetilde{\mu} < c^\top\mu^\star$ we show that $\widetilde{\mu}_j < \mu_j^\star,\ \forall j \in \mathcal{J}$ where $\xi > 1$. This means, following the arguments of the proof of the first statement, we need to show that

$$-\frac{\partial \widetilde{S}_j(\mu_j^\star)}{\partial \mu_j} < -\frac{\partial S_j(\mu_j^\star)}{\partial \mu_j}. \tag{A.16}$$

The substitution of $\mu_j = \mu_j^\star$ in $\widetilde{S}_j(\mu_j)$ yields

$$-\frac{\partial \widetilde{S}_j(\mu_j^\star)}{\partial \mu_j} = \underbrace{\frac{1}{\xi_j(\mu_j^\star)^2}}_{(3a)} + \underbrace{\frac{(ca_j+cs_j)\lambda_j}{2\xi_j\mu_j^\star(\xi_j\mu_j^\star-\lambda_j)}}_{(3b)} \underbrace{\exp\left\{-\frac{2(1-ca_j)(\xi_j\mu_j^\star-\lambda_j)}{3\lambda_j(ca_j+cs_j)}\right\}}_{(3c)} \cdot \left[\underbrace{\frac{2\xi_j^2\mu_j^\star-\xi_j\lambda_j}{\xi_j\mu_j^\star(\xi_j\mu_j^\star-\lambda_j)}}_{(3d)} + \underbrace{\frac{2\xi_j(1-ca_j)}{3\lambda_j(ca_j+cs_j)}}_{(3e)}\right].$$

When comparing the terms in the two equations we see that $(3a) < (1a)$, $(3b) < (1b)$, $(3c) < (1c)$, $(3d) < (1d)$ and $(3e) > (1e)$ if we assume $\xi_j > 1$. While the first three inequalities are easy to see the fourth inequality needs some basic calculus which we do not show here for conciseness.

In order for inequality (A.16) to hold we finally have to show that the influence of the last term $(3e) > (1e)$ is insignificant in comparison to the four other terms. Therefore, we multiply it with $(3b)$ and $(1b)$, respectively, and

compare the products. With some rearranging we end up with

$$\underbrace{\frac{1-ca_j}{3\mu_j^\star(\xi_j\mu_j^\star-\lambda_j)}}_{(3b)\cdot(3e)} < \underbrace{\frac{1-ca_j}{3\mu_j^\star(\mu_j^\star-\lambda_j)}}_{(1b)\cdot(1e)}$$

which concludes the second part of the proof. □

Proof of Proposition 2.5. For *Case 1* problems, where a penalty is incurred for all product families, we see from equation (2.6) that $\widetilde{\mu}_j = \mu_j^\star$ for all work stations j where $\xi_j = 1$. For *Case 2* problems, where no penalty is incurred for any product family, we can conclude from the equality constraint (A.3) that $\widetilde{\mu}_j < \mu_j^\star$ for all work stations j where $\xi_j = 1$ in order for the actual approximate mean total sojourn times to be equal to the contractually defined maximum mean sojourn times.

The lower and upper bounds for the updated optimal capacities at work stations j where an improvement $\xi_j > 1$ is imposed, $\widetilde{\mu}_j/\mu_j^\star \in (1/\xi_j, 1)$, can be found directly from the proof of Proposition 2.4 independent of the contractually defined maximum mean sojourn times determining the differentiability of the total costs function in the optimum. This concludes the proof. □

Proof of Proposition 2.6.

i) We need to show that $S_j(\cdot)$ is strictly monotonically increasing in cs_j, i.e., $\partial_{cs_j} S_j(\mu_j) > 0,\ \forall \mu_j \in \mathbf{dom}\ \mathcal{C}$.

$$\begin{aligned}
&\frac{\partial}{\partial cs_j}\left[\frac{1}{\mu_j} + \frac{(ca_j+cs_j)\lambda_j}{2}\frac{1}{\mu_j(\mu_j-\lambda_j)}\exp\left\{\frac{-2(1-ca_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)}\right\}\right] \\
&= \left[\underbrace{\frac{\lambda_j}{2\mu_j(\mu_j-\lambda_j)}}_{>0} + \underbrace{\frac{2(1-ca_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)^2}}_{\geq 0}\right]\underbrace{\exp\left\{\frac{-2(1-ca_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)}\right\}}_{>0} > 0
\end{aligned}$$

The proof that $S_j(\cdot)$ is strictly monotonically increasing in ca_j is omitted as it is similar to the preceding proof.

ii) We need to show that $\partial ca_j S_j(\cdot) > \partial cs_j S_j(\cdot)$, where

$$\frac{\partial}{\partial ca_j}\left[\frac{1}{\mu_j} + \frac{(ca_j + cs_j)\lambda_j}{2}\frac{1}{\mu_j(\mu_j - \lambda_j)}\exp\left\{\frac{-2(1-ca_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)}\right\}\right]$$
$$= \left[\frac{\lambda_j}{2\mu_j(\mu_j-\lambda_j)} + \frac{2(1+cs_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)^2}\right]\exp\left\{\frac{-2(1-ca_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)}\right\}.$$

Therefore, the inequality simplifies to

$$\frac{2(1+cs_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)^2} \stackrel{?}{>} \frac{2(1-ca_j)(\mu_j-\lambda_j)}{3\lambda_j(ca_j+cs_j)^2}$$

which is true since $1 + cs_j > 1 - ca_j$. This concludes the proof.

□

A.3 Queueing Network Parameter Analysis

The algorithm described here is a variation of the Queueing Network Analyzer (QNA) algorithm described by Whitt 1983 in [44] and enhanced by Bitran et al. 1996 in [5], modified to fit the requirements of this paper. We first summarize the input parameters required for the network analysis and not yet introduced in Section 2.3.

The external arrival rate to the system from customer type e is denoted as λ_e. The squared coefficient of variation (ratio of the variance to the squared mean, SCV) of the external interarrival times of customer type e is given by ca_e. τ_j denotes the expected service time at work station j and q_{ij} the share of the total arrivals at node i that are routed to node j.

Whereas some input parameters such as λ_e, ca_e, $\mu_j = \tau_j^{-1}$, cs_j and q_{ij} are given exogenously, the arrival rates λ_j and the according SCV need to be computed prior to the network analysis. Since we assume deterministic routing, the arrival rates λ_j can simply be calculated as the sum of the external expected arrival rates of all engine or customer classes arriving at work station

j,

$$\lambda_j = \sum_{e \in \mathcal{E}} \sum_{i \in \mathcal{J}_e} \lambda_e 1\{i = j\}. \tag{A.17}$$

In order to determine the variability parameters of the arrivals, we need to consider three different processes: superposition of arrivals, departures from nodes and (deterministic) splitting of departures.

Superposition of arrivals

With $\lambda_{ij} = \lambda_i q_{ij}$, $\rho_j = \lambda_j \tau_j$ and with the interarrival time variability at node j from node i, ca_{ij}, the variability parameters of a superposition of arrivals can be approximated as

$$ca_j = w_j \sum_{i=1}^{J} \frac{\lambda_{ij}}{\lambda_j} ca_{ij} + 1 - w_j, \tag{A.18}$$

where

$$w_j = \frac{1}{1 + 4(1-\rho_j)^2(v_j - 1)} \tag{A.19}$$

$$v_j = \frac{1}{\sum_{i=1}^{J} (\frac{\lambda_{ij}}{\lambda_j})^2}. \tag{A.20}$$

Departures

The variability of the departure process from a node depends on the variability of the arrivals and the service times. It can be approximated as

$$cd_j = \rho_j^2 cs_j + (1 - \rho_j^2) ca_j. \tag{A.21}$$

Deterministic splitting of departures

For the deterministic splitting process, we do not use the approximation developed by Whitt 1983 in [44] for Markovian routing,

$$cd_i = p_i cd + 1 - p_i, \tag{A.22}$$

but the convex combination developed by Whitt 1994 in [45] which is an approximation of the complex Erlang numerical procedure proposed by Bitran 1988 in [7] for $ca_i \leq 1$,

$$cd_1 = p_1 cd + p_1(1 - p_1)ca_2 + (1 - p_1)^2 ca_1. \tag{A.23}$$

Note that the subscript 1 denotes the currently observed customer class $e_1 \in \mathcal{E}$ while subscript 2 represents the superposition of all other classes present at the node, $\mathcal{E}_2 = \mathcal{E} \setminus e_1$.

With these three approximations, all necessary parameters of the modified KLB equations (2.3) are known which can then be applied to compute the expected sojourn times in the queueing network.[30]

[30] Simulations with JMT's JSIMgraph were used to verify the results computed with the approximations. For the different scenarios described in Section 5, the average relative devition between the total mean sojourn times computed using the QNA approximations and the results obtained from the simulation was approximately 4%, and did never exceed 11% for any scenario. As found in Wu and McGinnis (2013), the QNA approximations perform best for loads of around 80%, which is in the range of the loads in our scenarios [47]. Therefore, if networks with higher system loads (heavy traffic) are considered, more suitable approximations should be employed (e.g., algorithms using the intrinsic ratio developed by Wu and McGinnis (2012, 2013) [46, 47]).

B Appendix of Chapter 3

B.1 Proofs of Section 3.4

Proof of Theorem 3.1: Since we model the facility of the service provider as an M/M/1 queue, all interarrival and service times are i.i.d. exponentially-distributed random variables. Thus, we can use find the optimal threshold by considering the resulting continuous-time Markov chain. As our objective is to minimize the time-average queue length, we want the threshold value K to be as small as possible. Consequently, with Definition 3.1 in mind, the optimal feasible threshold can be determined as

$$\begin{aligned} K &= \min\left\{n \in \mathbb{Z}_+ : \mathbb{E}\pi_R^K \leq {}^r\!/_p\right\} \\ &= \min\left\{n \in \mathbb{Z}_+ : \mathbb{P}(Q > n) \leq {}^r\!/_p\right\}, \end{aligned}$$

since the expected time share the high capacity is active is equal to the probability that the queue length is greater than the threshold value,

$$\mathbb{E}\pi_R^K = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}\{Q(t) > K(r,p,\lambda)\}\,\mathrm{d}t = \mathbb{P}(Q > K).$$

As illustrated in Figure 3.3, we can model the queueing system as a continu-ous-time Markov chain with different traffic rates at different nodes. The balance equations are given as

$$\begin{aligned} \nu(n|\pi_R^K) &= \rho_1^n \nu(0|\pi_R^K), && n = 0, 1, \ldots, K-1, \\ \nu(n|\pi_R^K) &= \rho_1^K \rho_2^{n-K} \nu(0|\pi_R^K), && n = K, K+1, \ldots, \end{aligned}$$

where $\rho_1 = \lambda/(1-r) > 1$ and $\rho_2 = \lambda/(1-r+p) < 1$. Since there is an offset of one when comparing the number of jobs in the system L to the queue length (the job being currently process is counted in the former, but not in the latter), we can define $\mathbb{P}(Q > K) = \mathbb{P}(L > K) = 1 - \mathbb{P}(L \leq K)$. Since we use service tokens to model the queue, the number of jobs in the queue is the same as the number of jobs in the system, see Section 3.3. From the balance equations we can therefore determine

$$\begin{aligned}\mathbb{P}(Q > K) &= 1 - \mathbb{P}(L \leq K) \\ &= 1 - \sum_{n=0}^{K} \rho_1^n \nu(0|\pi_R^K) \\ &= 1 - \frac{\rho_1^{K+1} - 1}{\rho_1 - 1} \nu(0|\pi_R^K). \end{aligned} \tag{B.1}$$

We compute the steady-state probability that the system is empty with the normalization equation

$$\sum_{n=0}^{K-1} \rho_1^n \nu(0|\pi_R^K) + \sum_{n=0}^{\infty} \rho_1^K \rho_2^n \nu(0|\pi_R^K) = 1.$$

Thus, we find

$$\begin{aligned}\nu(0|\pi_R^K) &= \left[\sum_{n=0}^{K-1} \rho_1^n + \rho_1^K \sum_{n=0}^{\infty} \rho_2^n \right]^{-1} \\ &= \left[\frac{\rho_1^K - 1}{\rho_1 - 1} + \frac{\rho_1^K}{1 - \rho_2} \right]^{-1}. \end{aligned} \tag{B.2}$$

Using the derived relations we can solve

$$\mathbb{P}(Q > \hat{K}) \leq \frac{r}{p}$$

for $\hat{K}$ directly, resulting in

$$\hat{K} \geq \log_{\frac{\lambda}{1-r}} \frac{r(r-1)(\lambda + r - p - 1)}{p\lambda} \frac{1}{1-\lambda}.$$

Therefore, with $\mathcal{K}(r,p,\lambda) = \frac{r(r-1)(\lambda+r-p-1)}{p\lambda}$ and for integer threshold values, $K = \lceil \hat{K} \rceil$, we obtain the optimal feasible threshold as given in Equation (3.1). This concludes the proof of the theorem. □

Proof of Theorem 3.2: The time-average queue length $\mathcal{Q}(r,p,\lambda,\pi_R^K)$ is related to the mean number of states in the system as $\mathcal{Q}(r,p,\lambda,\pi_R^K) = \mathbb{E}(Q|\pi_R^K) = \mathbb{E}(L|\pi_R^K)$, i.e., the mean number of occupied states. With

$$\mathbb{E}(Q|\pi_R^K) = \sum_{n=0}^{K-1} n\rho_1^n \nu(0|\pi_R^K) + \sum_{n=0}^{\infty}(n+K)\rho_1^K\rho_2^n\nu(0|\pi_R^K)$$

and by using some well-known geometric series relations we obtain

$$\mathcal{Q}(r,p,\lambda,\pi_R^K) = \left[\frac{(K-1)\rho_1^{K+1} - K\rho_1^K + \rho_1}{(\rho_1-1)^2} + \frac{\rho_2 + K(1-\rho_2)}{(1-\rho_2)^2}\rho_1^K\right]\nu(0|\pi_R^K).$$

We next show that the time-average queue length is increasing in K. Fix $r \in (0,1),\ p > r$ and $\lambda \in (1-p,1)$. We need to show that

$$\begin{aligned}
&\mathcal{Q}(r,p,\lambda,\pi_R^{K+1}) - \mathcal{Q}(r,p,\lambda,\pi_R^K) = \mathbb{E}(L|\pi_R^{K+1}) - \mathbb{E}(L|\pi_R^K)\\
&= \underbrace{\frac{K\rho_1^{K+2} - (K+1)\rho_1^{K+1} + \rho_1}{(\rho_1-1)^2}\nu(0|\pi_R^{K+1})}_{1a}\\
&+ \underbrace{\frac{\rho_2 + (K+1)(1-\rho_2)}{(1-\rho_2)^2}\rho_1^{K+1}\nu(0|\pi_R^{K+1})}_{1b} - \underbrace{\frac{(K-1)\rho_1^{K+1} - K\rho_1^K + \rho_1}{(\rho_1-1)^2}\nu(0|\pi_R^K)}_{2a}\\
&+ \underbrace{\frac{\rho_2 + K(1-\rho_2)}{(1-\rho_2)^2}\rho_1^K\nu(0|\pi_R^K)}_{2b} > 0.
\end{aligned}$$

We first show that $1a > 1b$,

$$\begin{aligned}
&\frac{K\rho_1^{K+2} - (K+1)\rho_1^{K+1} + \rho_1}{(\rho_1-1)^2}\nu(0|\pi_R^{K+1})\\
&\overset{?}{>} \frac{(K-1)\rho_1^{K+1} - K\rho_1^K + \rho_1}{(\rho_1-1)^2}\nu(0|\pi_R^{K+1}).
\end{aligned} \tag{B.3}$$

With

$$\nu(0|\pi_r^K) = \frac{(\rho_1 - 1)(1 - \rho_2)}{(\rho_1^K - 1)(1 - \rho_2) + \rho_1^K(\rho_1 - 1)} \tag{B.4}$$

we can reformulate inequality (B.3) as

$$\frac{K\rho_1^{K+2} - (K+1)\rho_1^{K+1} + \rho_1}{(\rho_1^{K+1} - 1)(1 - \rho_2) + \rho_1^{K+1}(\rho_1 - 1)}$$

$$\geq \frac{K\rho_1^{K+1} - (K+1)\rho_1^K + 1}{(\rho_1^K - 1)(1 - \rho_2) + \rho_1^K(\rho_1 - 1)} \overset{?}{>} \frac{(K-1)\rho_1^{K+1} - K\rho_1^K + \rho_1}{(\rho_1^K - 1)(1 - \rho_2) + \rho_1^K(\rho_1 - 1)}$$

$$\implies K\rho_1^{K+1} - (K+1)\rho_1^K + 1 \overset{?}{>} (K-1)\rho_1^{K+1} - K\rho_1^K + \rho_1$$
$$\implies \rho_1^K(\rho_1 - 1) > \rho_1 - 1$$

Since $\rho_1 > 1$, the inequality holds. Next, we show that $2a > 2b$. With Equation (B.4) we find that

$$\frac{\rho_1[\rho_2 + (K+1)(1 - \rho_2)]}{(\rho_1^{K+1} - 1)(1 - \rho_2) + \rho_1^{K+1}(\rho_1 - 1)}$$

$$\geq \frac{\rho_2 + (K+1)(1 - \rho_2)}{(\rho_1^K - 1)(1 - \rho_2) + \rho_1^K(\rho_1 - 1)} \overset{?}{>} \frac{\rho_2 + K(1 - \rho_2)}{(\rho_1^K - 1)(1 - \rho_2) + \rho_1^K(\rho_1 - 1)}$$

$$\implies \rho_2 + (K+1)(1 - \rho_2) > \rho_2 + K(1 - \rho_2).$$

Since $\rho_2 < 1$, the inequality holds. With $1a > 1b$ and $2a > 2b$ it follows that $1a + 1b - 2a - 2b > 0$ which proves that the time-average queue length is increasing in K.

Therefore, if we consider the limit $\lambda \to 1$ we find that

$$\nu(0|\pi_r^K) \sim \mathcal{O}\left(\rho_1^{-(K+1)}\right)$$

and

$$\frac{(K-1)\rho_1^{K+1} - K\rho_1^K + \rho_1}{(\rho_1 - 1)^2} + \frac{\rho_2 + K(1 - \rho_2)}{(1 - \rho_2)^2}\rho_1^K \sim \mathcal{O}\left(K\rho_1^{K+1}\right).$$

Thus,

$$\mathcal{Q}(r,p,\lambda,\pi_R^K) \sim \mathcal{O}(K), \quad \text{as } \lambda \to 1.$$

Finally, we can conclude that

$$\begin{aligned} K(r,p,\lambda) &\sim \log_{\frac{1}{1-r}} \mathcal{K}(r,p,1)\frac{1}{1-\lambda}, \\ \Rightarrow \mathcal{Q}(r,p,\lambda,\pi_R^K) &\sim \mathcal{O}\left(\log_{\frac{1}{1-r}} \frac{1}{1-\lambda}\right), \end{aligned}$$

as $\lambda \to 1$. This concludes the proof of Theorem 3.2. □

Proof of Proposition 3.1: From the proof of Theorem 3.1 we find that the time-average queue length increases as $\nu(0|\pi_R^K)$ decreases, i.e., when the probability that the system is empty is low. Therefore, if we fix $r \in (0,1)$ and $\lambda \in (1-r,1]$, we can find p such that $\nu(0|\pi_R^K)$ is minimal. Assume the we can choose $K^\star \in \mathbb{Z}_+$ freely, then we want to determine $\min_{p>r,K^\star\in\mathbb{Z}_+} \nu(0|\pi_R^{K^\star})$. If we compute $\partial_K \nu(0|\pi_R^{K^\star}) = 0$ we find that $K^\star = 0$. Thus we can evaluate

$$\min_{p>r} \nu(0|\pi_R^0) = 1 - \frac{\lambda}{1-r+p}$$

and find that $\nu(0|\pi_R^0) \to \min!$ for $p \downarrow r$, $\lim_{p\downarrow r} \nu(0|\pi_R^0) = 1-\lambda$. The expected queue length of an M/M/1 queue with constant service rate is given as

$$\mathbb{E}(L_{\mathrm{M/M/1}}) = \frac{\rho}{1-\rho}$$

with $\rho = \lambda/\mu < 1$.Therefore, we need to show that

$$\lim_{p\downarrow r} \mathcal{Q}(r,p,\lambda,\pi_R^{K(r,p,\lambda)}) = \mathbb{E}(L_{\mathrm{M/M/1}}) \overset{\mu=1}{=} \frac{\lambda}{1-\lambda}. \tag{B.5}$$

As we chose $K^\star$ freely to determine for which p the reactive policy performs worst, we have to confirm that it coincides with the definition given in Theo-

rem 3.1 for $p \downarrow r$,

$$\lim_{p\downarrow r} \log_{\frac{\lambda}{1-r}} \mathcal{K}(r,p,\lambda)\frac{1}{1-\lambda} = \log_{\frac{\lambda}{1-r}} \frac{1-r}{\lambda} = -1$$

$$\implies \lim_{p\downarrow r} K(r,p,\lambda) = \lim_{\epsilon\downarrow 0} \lceil -1+\epsilon \rceil = 0.$$

With $\lim_{p\downarrow r} \rho_2 = \lambda$, we can compute

$$\lim_{p\downarrow r} \mathcal{Q}(r,p,\lambda,\pi_R^0) = \frac{-\rho_1+\rho_1}{(\rho_1-1)^2}(1-\lambda) + \frac{\rho_2}{(1-\rho_2^2)}(1-\lambda) = \frac{\lambda}{1-\lambda}$$

which is the same as $\mathbb{E}(L_{\mathrm{M/M/1}})$ with $\mu = 1$. This concludes the proof. □

B.2 Proofs of Section 3.5

Proof of Lemma 3.3: Let $\lambda \to 1$. We need to show that, for any feasible policy, the number of service tokens generated by the contingent capacity p until time t cannot exceed the number of NOB arrivals. From Lemma 2.1 we know that no service tokens can be wasted after the all-time minimum of Q_0 has been reached.[31] We can prove the statement by showing that a service token will be wasted if there exists a time $t \in \mathbb{R}_+$ where $S_2(\varpi(t)) > A_\Psi^\infty(t)$.

In order to exceed the number of NOB arrivals at time t_1, there must be a time $t_2 < t_1$ where $S_2(\varpi(t_2)) = A_\Psi^\infty(t_2)$. Also, we assume that $S_2(\varpi(t_3)) \leq A_\Psi^\infty(t_3)$ for all $t_3 < t_2$, i.e., we consider the first excursion of $S_2(\varpi(t))$ above $A_\Psi^\infty(t)$. If the queue is empty at time t_2, $Q_1(t_2) = Q_2(t_2) = \delta Q(t_2) = 0$, any service token that is generated before the next arrival, $t_1 \in (t_2, \inf\{x > t_2 : A(x) = A(x-) + 1\})$, will be wasted. Now assume that the queue is not empty at time t_2, $Q_1(t_2) = Q_2(t_2) > 0$ and $\delta Q(t_2) = 0$. Since Q_1 is a

[31] When controlled manually, a policy can be such that the number of service tokens generated by the contingent capacity exceeds the number of NOB arrivals, but still yields a finite time-average queue length. For each service token that is wasted, the time-average queue length after the waste is increased by 1. On the other hand, as we assume that a policy is a prescription controlling the contingent capacity based on the current and future state of the system, the number of events where a service token is wasted goes to ∞ as $t \to \infty$. Therefore, when considering a prescriptive policy, for all $t \in \mathbb{R}_+, S_2(\varpi(t)) \leq A_\Psi^\infty(t)$.

recurrent random walk and NOB arrivals can only occur when $Q_1(t) = 0$, no NOB arrival will occur before the next drop of Q_1 to 0. Therefore, if we denote $t_4 = \inf\{x > t_2 : Q_1(x) = 0\}$ as the time where Q_1 next drops to 0, we know that, in the time interval $t \in [t_2, t_4]$, the number of service tokens produced by the base capacity equals the number of jobs currently in queue plus all additionally arriving jobs, $Q_1(t_2)+A(t_4)-A(t_2)-[S_1(t_4)-S_1(t_2)] = 0$. Thus, if an additional service token is produced at time $t_1 \in [t_2, t_4]$, a service token has to be wasted, although not necessarily the service token produced by the contingent capacity. This concludes the proof. □

Proof of Lemma 3.4: We need to show that there exists no feasible policy $\pi \in \Pi^\infty$ such that the number of service tokens produced by the contingent capacity p until time t exceeds the number of service tokens produced when applying the solely forward-looking policy π_F^∞. If a policy π is supposed to generate more service tokens than the policy π_F^∞, the contingent capacity must be switched on earlier. We can prove the statement by showing that the probability that a service token produced by the contingent capacity will be wasted is larger than 0 by doing so.

By Lemma 3.3, as the number of service tokens generated by the contingent capacity until time t is bounded by $A_\Psi^\infty(t)$, the contingent capacity must always be switched off as soon as $S_2(\varpi(t)) = A_\Psi^\infty(t)$. The policy π_F^∞ is such that the contingent capacity is activated as soon as a NOB arrival occurs (and the contingent capacity was not active before the NOB arrival). Therefore, consider a time t_1 where the queue is empty and a NOB arrival occurs. Then,

$$
\begin{aligned}
&Q_1(t_1-) = 0, \quad Q_1(t_1) = 0,\\
&Q_2(t_1-) = 0, \quad Q_2(t_1) = 1,\\
&\delta Q(t_1-) = 0, \quad \delta Q(t_1) = 1,\\
&\pi_F^\infty(t_1-) = 0, \quad \pi_F^\infty(t_1) = 1,\\
&S_2(\varpi_F^\infty(t_1-)) = A_\Psi^\infty(t_1-), \quad S_2(\varpi_F^\infty(t_1)) = A_\Psi^\infty(t_1) - 1.
\end{aligned}
$$

Denote $t_2 = \sup\{x < t_1 : Q_2(x) = Q_2(x-) - 1\}$ as the time of the last

service token generation, where the queue dropped to 0. Now, assume that the contingent capacity is switched on at a time $t_3 \in (t_2, t_1)$ between the time of the last generation of a service token and the time of the next NOB arrival. Then, with the Markov property, the probability that at least one service token will be produced by the contingent capacity with rate p is given as

$$\mathbb{P}[S_2(\varpi(t_1)) > S_2(\varpi(t_2))] = \sum_{k=1}^{\infty} \frac{p^k (t_1 - t_3)^k}{k!} e^{-p(t_1 - t_3)} > 0.$$

If the event realizes, $S_2(\varpi(t)) > A_{\Psi}^{\infty}(t)$ and the service tokens will be wasted as there is no job to be served in the queue. Thus, the earliest time we can activate the contingent capacity is t_1, the time of the next NOB arrivals. As this corresponds to the time where the solely forward-looking policy π_F^{∞} switches on the high capacity, the proof is concluded. □

B.3 Proofs of Section 3.6

Proof of Proposition 3.2: In order the show that the proactive policy outperforms the asymptotically optimal policy it must be true that, for all $\lambda \in (1 - r, 1]$,

$$\left[\frac{(\widetilde{K} - 1)\widetilde{\rho}_1^{\widetilde{K}+1} - \widetilde{K}\widetilde{\rho}_1^{\widetilde{K}} + \widetilde{\rho}_1}{(\widetilde{\rho}_1 - 1)^2} + \frac{\widetilde{\rho}_2 + \widetilde{K}(1 - \widetilde{\rho}_2)}{(1 - \widetilde{\rho}_2)^2} \widetilde{\rho}_1^{\widetilde{K}} \right] \widetilde{\nu}(0|\pi_R^{\widetilde{K}}) \leq \frac{1 - r}{\lambda - (1 - r)}.$$

This inequality holds as the right side is the expected number of jobs in an M/M/1 queue with service rate λ and arrival rate $(1 - r)$ and the left part is the expected number of jobs in an M/M/1 queue with threshold policy with arrival rate $(1 - r)$ and service rates $\{\lambda, \lambda + p\}$. When $\lambda \to 1$ and $\widetilde{K} \to \infty$, the two expressions coincide. This concludes the proof. □

B.4 Proofs of Section 3.7

Proof of Proposition 3.3: With finite lookahead, all *real NOB* arrivals will be identified, plus additional ones (*virtual*). For real NOB arrivals we know that

$U(Q_0, n, \infty) = U(Q_0, n, W(n)) = \infty$. Thus, virtual NOB arrivals are all for which we have $U(Q_0, n, W(n)) = \infty$ but there exists $\{j \in \mathbb{Z}_+ : Q_0[n+j] = Q_0[n]-1\}$, i.e., $U(Q_0, n, \infty) < \infty$. The number of arriving jobs left (all arrivals but real NOB arrivals) that could additionally be identified as virtual NOB arrivals is given as $(1-r)$. The proactive (and solely forward-looking) policy will identify all jobs as NOB arrivals for which the busy-period is longer than the lookahead window. It is well known (see, e.g., Asmussen (2008), Section III.8c, Corollary 8.7 [2]) that the busy-period distribution of an M/M/1 queue with arrival rate α, service rate β and load $\gamma = \alpha/\beta$ is given by the density

$$g_{\alpha,\beta}(t) = \frac{\gamma^{-1/2}}{t} e^{-(\alpha+\beta)t} I_1\left(2t\sqrt{\alpha\beta}\right),$$

where $I_1(y)$ denotes the modified Bessel function of the first kind of order one,

$$I_1(y) = \sum_{i=0}^{\infty} \frac{(y/2)^{2i+1}}{k!(k+1)!}.$$

Let $F_{\alpha,\beta}(t) = \int_0^t g_{\alpha,\beta}(x)\,\mathrm{d}x$ be the cumulative distribution function of the busy-period. Then, the time-average number of NOB arrivals given a finite lookahead window of length $w < \infty$ can be computed as

$$\begin{aligned}\lim_{t\to\infty} \frac{A_\Psi^w(t)}{t} &= \lambda - (1-r) + \chi(w) \\ &= \lambda - (1-r) + (1-r)[1 - F_{1-r,\lambda}(w)] \\ &= \lambda - (1-r)F_{1-r,\lambda}(w).\end{aligned}$$

This concludes the proof. □

C Appendix of Chapter 4

C.1 Proofs of Section 4.3

Proof of Proposition 4.1: The proposition follows instantly from equation (4.1). The rate of w-critical arrivals can be at most r, i.e.,

$$\lambda - (1-r)F_{1-r,\lambda}(w) \leq r.$$

Reformulation yields the expression stated in the proposition. □

Proof of Proposition 4.2: The rate of w-critical job arrivals is equal to the probability that a random job arrival $n \in \Phi(Q_0)$ will be characterized as w-critical. With $\xi(t) = \min_{s\in[0,w]} Q_0(t+s)$, this probability is given as

$$\begin{aligned}
\mathbb{P}(n \in \widetilde{\Psi}^w) &= \mathbb{P}[\xi(T_n) \geq Q_0(T_n) \wedge Q_0(T_n + w) \geq Q_0(T_n) + J] \\
&\overset{(a)}{=} \mathbb{P}[\xi(T_n) \geq Q_0(T_n)]\mathbb{P}[Q_0(T_n + w) \geq Q_0(T_n) + J | \xi(T_n) \geq Q_0(T_n)] \\
&\overset{(b)}{=} \mathbb{P}[\xi(T_n) \geq Q_0(T_n)] \\
&\quad \times \mathbb{P}[Q_0(T_n + w) \geq Q_0(T_n) + J | Q_0(T_n + w) \geq Q_0(T_n)] \\
&\overset{(c)}{=} \mathbb{P}[\xi(T_n) \geq Q_0(T_n)] \frac{\mathbb{P}[Q_0(T_n + w) - Q_0(T_n) \geq J]}{\mathbb{P}[Q_0(T_n + w) - Q_0(T_n) \geq 0]}. \qquad \text{(C.1)}
\end{aligned}$$

Equality (a) follows from the dependency of the two events. When considering this dependency, we only need to take into account $Q_0(T_n + w) \geq Q_0(T_n)$ which follows instantly from $\xi(T_n) \geq Q_0(T_n)$, leading to equality (b). Finally,

equality (c) follows from Kolmogorov's definition of conditional probability. However, if we denote $Q_0(T_n + w) - Q_0(T_n)$ as the random variable N we find that $\mathbb{P}(N \geq y | N \geq x)$, $y \geq x$, can be rewritten as $\mathbb{P}(N \geq y \wedge N \geq x)/\mathbb{P}(N \geq x) = \mathbb{P}(N \geq y)/\mathbb{P}(N \geq x)$ since $(N \geq x) \subseteq (N \geq y)$.

From the proof of Proposition 4.1 we know that

$$\mathbb{P}[\xi(T_n) \geq Q_0(T_n)] = \lim_{t\to\infty} \frac{A^w_\Psi(t)}{t} = \lambda - (1-r)F_{1-r,\lambda}(w).$$

For the right hand side of expression (C.1) we can use the fact that

$$\mathbb{P}[Q_0(t+w) - Q_0(t) \geq x] = \mathbb{P}[\underbrace{(A(t+w) - A(t))}_{N_1} - \underbrace{(S_1(t+w) - S_1(t))}_{N_2} \geq x],$$

with the Markov property, can be computed as the difference between two Poisson random variables N_1 and N_2 with rates $\nu_1 = w\lambda$ and $\nu_2 = w(1-r)$, respectively. The difference between two Poisson random variables is described by the Skellam distribution which is defined as

$$\mathbb{P}(N_1 - N_2 = k) = e^{-(\nu_1+\nu_2)} \left(\frac{\nu_1}{\nu_2}\right)^{k/2} I_k\left(2\sqrt{\nu_1\nu_2}\right),$$

with $I_k(\cdot)$ being the modified Bessel function of the first kind of order k. Therefore, we can express

$$\begin{aligned}
\mathbb{P}[Q_0(t+w) - Q_0(t) \geq x] &= \mathbb{P}(N_1 - N_2 \geq x) \\
&= \sum_{k=x}^{\infty} \mathbb{P}(N_1 - N_2 = k) \\
&= \sum_{k=x}^{\infty} e^{-w(\lambda+1-r)} [\lambda/(1-r)]^{k/2} I_k\left(2w\sqrt{\lambda(1-r)}\right).
\end{aligned}$$

Bringing the different parts together, we receive

$$\lim_{t\to\infty} \frac{\widetilde{A}^w_\Psi(t)}{t} = [\lambda - (1-r)F_{1-r,\lambda}(w)] \frac{\sum_{k=J}^{\infty} [\lambda/(1-r)]^{k/2} I_k\left(2w\sqrt{\lambda(1-r)}\right)}{\sum_{k=0}^{\infty} [\lambda/(1-r)]^{k/2} I_k\left(2w\sqrt{\lambda(1-r)}\right)}.$$

Rearranging terms yields the expression stated in the proposition and the

proof is concluded. □

Proof of Proposition 4.3: The proof is based on the properties of the underlying transient random walk Q_0. If we are currently at time T_n and $n \in \widetilde{\Psi}^w$, we know that $Q_0(t) \geq Q_0(T_n), \forall t \in [T_n, T_n + w]$, and $Q_0(T_n + w) \geq Q_0(T_n) + J$ (we use J instead of $J(r, \lambda, w, \Lambda)$ for conciseness). Denote $\mathcal{M} = \max_{t \in [T_n+w, \infty)} [Q_0(T_n+w) - Q_0(t)] \in \mathbb{Z}_+$ as the maximum excursion of Q_0 below $Q_0(T_n + w)$ after time $T_n + w$. The maximum excursion is geometrically distributed,

$$\mathbb{P}(\mathcal{M} \geq a) = \left(\frac{1-r}{\lambda}\right)^a,$$

with $a \in \mathbb{Z}_+$ [2]. Therefore, we can compute

$$\begin{aligned}
&\mathbb{P}(n \in \Psi^\infty | n \in \widetilde{\Psi}^w) \\
&= 1 - \mathbb{P}[\mathcal{M} \geq \mathbb{E}(Q_0(T_n + w) - Q_0(T_n) | Q_0(T_n + w) - Q_0(T_n) \geq J)] \\
&= 1 - \left(\frac{1-r}{\lambda}\right)^{\vartheta(r,\lambda,w,\Lambda)},
\end{aligned}$$

with $\vartheta(r, \lambda, w, \Lambda) = \mathbb{E}(Q_0(T_n + w) - Q_0(T_n) | Q_0(T_n + w) - Q_0(T_n) \geq J)$. Following the proof of Proposition 4.2, we find

$$\vartheta(r, \lambda, w, \Lambda) = \mathbb{E}[N_1 - N_2 | N_1 - N_2 \geq J(r, \lambda, w, \Lambda)].$$

With $J = J(r, \lambda, w, \Lambda)$ we obtain

$$\begin{aligned}
\vartheta(r, \lambda, w, \Lambda) &= J + \mathbb{E}(N_1 - N_2 - J)^+ \\
&= J + \sum_{k=J}^{\infty} (k - J) \mathbb{P}(N_1 - N_2 = k) \\
&= J + \sum_{k=J}^{\infty} (k - J) e^{-w(\lambda+1-r)} \left(\frac{\lambda}{1-r}\right)^{k/2} I_k\left(2w\sqrt{\lambda(1-r)}\right).
\end{aligned}$$

This concludes the proof. □

List of Figures

List of Tables

Bibliography

[1] Gad Allon and Jan A. Van Mieghem. Global dual sourcing: Tailored base-surge allocation to near-and offshore production. *Management Science*, 56(1):110–124, 2010.

[2] Søren Asmussen. *Applied probability and queues*. Springer Science & Business Media, 2008.

[3] René Bekker, O.J. Boxma, and J.A.C. Resing. Queues with service speed adaptations. *Statistica Neerlandica*, 62(4):441–457, 2008.

[4] René Bekker, G. M. Koole, Bo Friis Nielsen, and Thomas Bang Nielsen. Queues with waiting time dependent service. *Queueing Systems*, 68(1):61–78, 2011.

[5] Gabriel R. Bitran and Reinaldo Morabito. State-of-the-art survey: Open queueing networks: Optimization and performance evaluation models for discrete manufacturing systems. *Production and Operations Management*, 5(2):163–193, 1996.

[6] Gabriel R. Bitran and Reinaldo Morabito. An overview of tradeoff curves in manufacturing systems design. *Production and Operations Management*, 8(1):56–75, 1999.

[7] Gabriel R. Bitran and Devanath Tirupati. Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Science*, 34(1):75–100, 1988.

[8] Gabriel R. Bitran and Devanath Tirupati. Tradeoff curves, targeting

and balancing in manufacturing queueing networks. *Operations Research*, 37(4):547–564, 1989.

[9] Onno Johan Boxma, A.H.G. Rinnooy Kan, and Mario van Vliet. Machine allocation problems in manufacturing networks. *European Journal of Operational Research*, 45(1):47–54, 1990.

[10] Stephen Boyd, J. Duchi, and L. Vandenberghe. Subgradients. *Lecture notes of EE364b, Stanford University, Spring Quarter*, 2015.

[11] Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient methods. *Lecture notes of EE392, Stanford University, Autumn Quarter*, 2003.

[12] Justin Boyle, Melanie Jessup, Julia Crilly, David Green, James Lind, Marianne Wallis, Peter Miller, and Gerard Fitzgerald. Predicting emergency department admissions. *Emergency Medicine Journal*, 29(5):358–365, 2012.

[13] James R. Bradley. A Brownian approximation of a production-inventory system with a manufacturer that subcontracts. *Operations Research*, 52(5):765–784, 2004.

[14] Kurt M. Bretthauer and Murray J. Côté. Nonlinear programming for multiperiod capacity planning in a manufacturing system. *European Journal of Operational Research*, 96:167–197, 1996.

[15] Nasuh C. Buyukkaramikli, J. Will M. Bertrand, and Henny P.G. van Ooijen. Periodic capacity management under a lead-time performance constraint. *OR Spectrum*, 35(1):221–249, 2013.

[16] Zhi-Long Chen and Nicholas G. Hall. Supply chain scheduling: Conflict and cooperation in assembly systems. *Operations Research*, 55(6):1072–1089, 2007.

[17] Thomas B. Crabill. Optimal control of a service facility with variable exponential service times and constant arrival rate. *Management Science*, 18(9):560–566, 1972.

[18] Euclides da Conceicao Pereira Batalha. Aircraft engines maintenance costs and reliability. Master's thesis, Universidade Nova de Lisboa, 2012.

[19] Noah Gans, Haipeng Shen, Yong-Pin Zhou, Nikolay Korolev, Alan McCord, and Herbert Ristock. Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing & Service Operations Management*, 17(4):571–588, 2015.

[20] Andy Geer, Stuart Ellis, Jerry Goodwin, Dave Benbow, and Mike Page. Rolls-Royce: Trent engines. `http://innovationnow.raeng.org.uk/innovations/default.aspx?item=15`. Accessed: 2014-10-15.

[21] Wallace J. Hopp, Mark L. Spearman, Sergio Chayet, Karen L. Donohue, and Esma S. Gel. Using an optimized queueing network model to support wafer fab design. *IIE Transactions*, 34(2):119–130, 2002.

[22] Oliver Ibe. *Markov processes for stochastic modeling*. Newnes, 2013.

[23] McKinsey Global Institute. The age of analytics: Competing in a data-driven world, December 2016.

[24] Wolfgang Krämer and M. Langenbach-Belz. Approximate formulae for the delay in the queueing system GI/G/1. In *Proceedings ITC*, volume 8, pages 235–1, 1976.

[25] Julian Kurz. Capacity planning for a maintenance service provider with advanced information. *European Journal of Operational Research*, 251(2):466–477, 2016.

[26] Julian Kurz. Queueing with limited future information. *Working Paper*, December 2016.

[27] Julian Kurz and Richard Pibernik. Flexible capacity managmemt with future information. *Working Paper*, November 2016.

[28] Jiyeon Lee and Jongwoo Kim. A workload-dependent M/G/1 queue under a two-stage service policy. *Operations Research Letters*, 34(5):531–538, 2006.

[29] Helen Mayhew, Tamim Saleh, and Simon Williams. Making data analytics work for you–instead of the other way around. *McKinsey Quarterly*, October 2016.

[30] Reinaldo Morabito, Mauricio C. de Souza, and Mariana Vazquez. Approximate decomposition methods for the analysis of multicommodity flow routing in generalized queuing networks. *European Journal of Operational Research*, 232(3):618–629, 2014.

[31] Claudio Rogerio Negri da Silva and Reinaldo Morabito. Performance evaluation and capacity planning in a metallurgical job-shop system using open queueing network models. *International Journal of Production Research*, 47(23):6589–6609, 2009.

[32] P.J. García Nieto, E. Garcia-Gonzalo, F. Sánchez Lasheras, and F.J. de Cos Juez. Hybrid PSO–SVM-based method for forecasting of the remaining useful life for aircraft engines and evaluation of its reliability. *Reliability Engineering & System Safety*, 138:219–231, 2015.

[33] Raul Poler, Jorge E. Hernandez, Josefa Mula, and Francisco C. Lario. Collaborative forecasting in networked manufacturing enterprises. *Journal of Manufacturing Technology Management*, 19(4):514–528, 2008.

[34] Christoph Reményi and Stephan Staudacher. Systematic simulation based approach for the identification and implementation of a scheduling rule in the aircraft engine maintenance. *International Journal of Production Economics*, 147:94–107, 2014.

[35] J. G. Shanthikumar and John Buzacott. Open queueing network models of dynamic job shops. *International Journal Of Production Research*, 19(3):255–266, 1981.

[36] Joel Spencer, Madhu Sudan, and Kuang Xu. Queuing with future information. *The Annals of Applied Probability*, 24(5):2091–2142, 2014.

[37] Shaler Stidham. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, 30(8):705–713, 1985.

[38] Shaler Stidham Jr. Analysis, design, and control of queueing systems. *Operations Research*, 50(1):197–216, 2002.

[39] Armin Stranjak et al. A multi-agent simulation system for prediction and scheduling of aero engine overhaul. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*, pages 81–88. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[40] Jianzhong Sun, Hongfu Zuo, Wenbin Wang, and Michael G. Pecht. Application of a state space modeling technique to system prognostics based on a health index for condition-based maintenance. *Mechanical Systems and Signal Processing*, 28:585–596, 2012.

[41] Yan Sun, Bee Hoon Heng, Yian Tay Seow, and Eillyne Seow. Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine*, 9(1):1, 2009.

[42] Lotfi Tadj and Gautam Choudhury. Optimal design and control of queues. *Top*, 13(2):359–412, 2005.

[43] Fabian Taigel. Secure collaborative forecasting using supervised machine learning. Master's thesis, Julius-Maximilians-Universität Würzburg, May 2015.

[44] Ward Whitt. The queueing network analyzer. *Bell System Technical Journal*, 62(9):2779–2815, 1983.

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

[46] Kan Wu and Leon McGinnis. Performance evaluation for general queueing networks in manufacturing systems: Characterizing the trade-off between queue time and utilization. *European Journal of Operational Research*, 221(2):328–339, 2012.

[47] Kan Wu and Leon McGinnis. Interpolation approximations for queues in series. *IIE Transactions*, 45(3):273–290, 2013.

[48] Kuang Xu. Necessity of future information for admission control. *Operations Research*, 63(5):1213–1226, 2015.

[49] Kuang Xu and Carri W. Chan. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management (MSOM)*, 2016.

[50] Qiushi Zhu, Hao Peng, and Geert-Jan van Houtum. A condition-based maintenance policy for multi-component systems with a high maintenance setup cost. *OR Spectrum*, June 2015.

[51] Antonio Zilli, Julian Kurz, et al. D24.2 – Business Modelling. Technical report, PRACTICE: Privacy-Preserving Computation in the Cloud, May 2015.

Eidesstattliche Erklärung (Statement of Academic Integrity)

Hiermit erkläre ich gemäß § 6 Abs. 2 Nr. 2 der Promotionsordnung der wirtschaftswissenschaftlichen Fakultät der Universität Würzburg, dass ich diese Dissertation eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters angefertigt habe. Ausgenommen davon sind jene Abschnitte, bei deren Erstellung ein Koautor mitgewirkt hat. Diese Abschnitte sind entsprechend gekennzeichnet und die Namen der Koautoren sind vollständig und wahrheitsgemäß aufgeführt. Bei der Erstellung der Abschnitte, bei denen ein Koautor mitgewirkt hat, habe ich einen signifikanten Beitrag geleistet, der meine eigene Koautorschaft rechtfertigt.

Außerdem erkläre ich, dass ich außer den im Schrifttumsverzeichnis angegebenen Hilfsmitteln keine weiteren benutzt habe und alle Stellen, die aus dem Schrifttum ganz oder annähernd entnommen sind, als solche kenntlich gemacht und einzeln nach ihrer Herkunft nachgewiesen habe.

Kornwestheim, den 28. Januar 2017

Julian Frederick Kurz