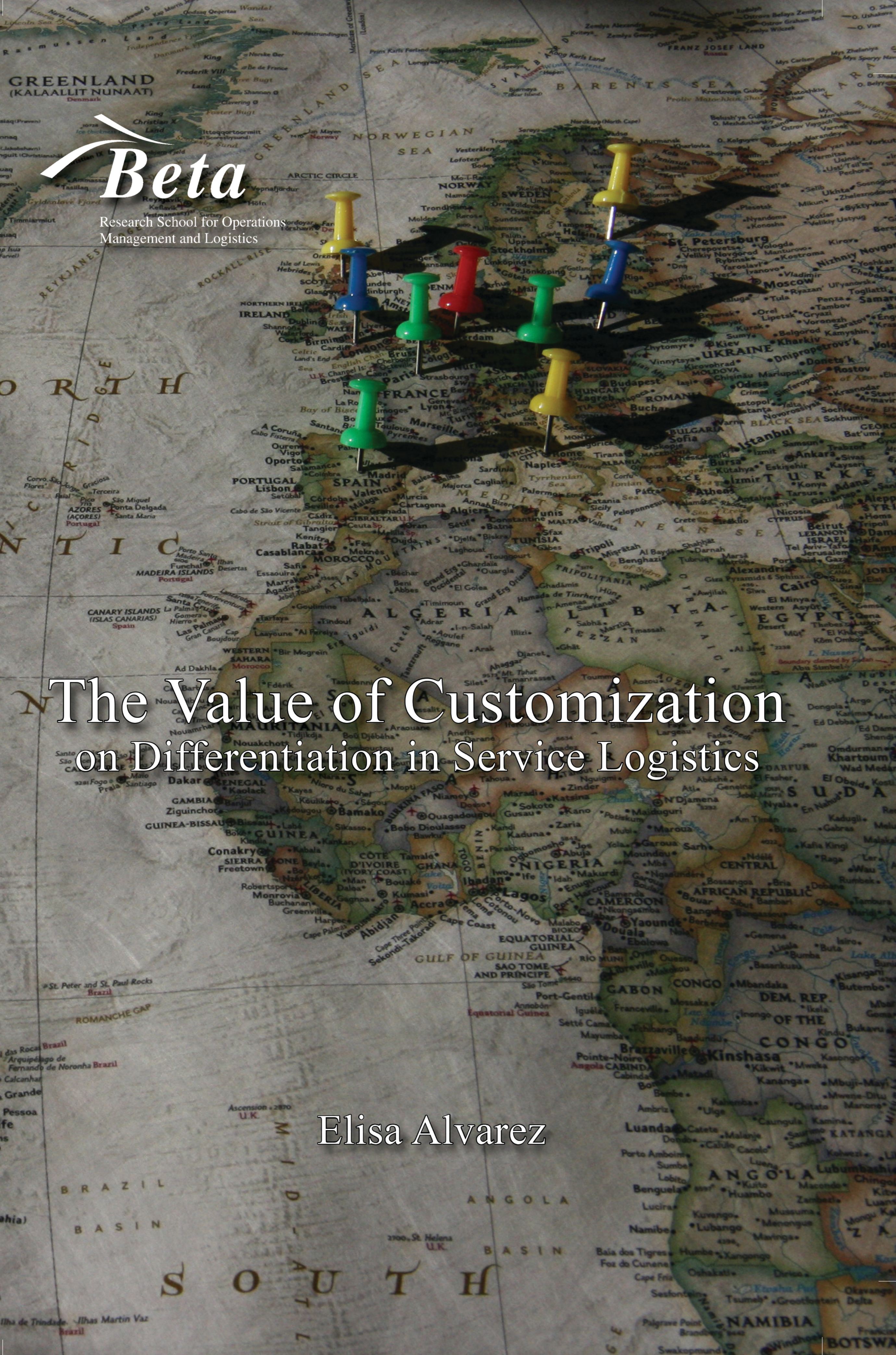


Beta

Research School for Operations Management and Logistics

The Value of Customization on Differentiation in Service Logistics

Elisa Alvarez

THE VALUE OF CUSTOMIZATION
ON DIFFERENTIATION IN SERVICE LOGISTICS

Elisa Monica Alvarez

Dissertation committee

Chairman & secretary	Prof. dr. R.A. Wessel
Promotor	Prof. dr. W.H.M. Zijm
Assistant promotor	Dr. M.C. van der Heijden
Members	Dr. A. Al Hanbali
	Prof. dr. R.J. Boucherie
	Prof. dr. ir. T. Tinga
	Prof. dr. ir. G.J.J.A.N. van Houtum
	Prof. dr. ir. R. Dekker

This thesis is number D-166 of the thesis series of the Beta Research School for Operations Management and Logistics. The Beta Research School is a joint effort of the departments of Technology Management and Mathematics and Computer Science at the Eindhoven University of Technology and the Centre for Telematics and Information Technology at the University of Twente. Beta is the largest research centre in the Netherlands in the field of operations management in technology-intensive environments. The mission of Beta is to carry out fundamental and applied research on the analysis, design and control of operational processes.

This research has been funded by the Service Logistics Forum Research.

Ph.D. thesis, University of Twente, Enschede, the Netherlands

Printed by Proefschriftmaken.nl || Uitgeverij BOXPress

ISBN 978-90-365-3548-9

THE VALUE OF CUSTOMIZATION
ON DIFFERENTIATION IN SERVICE LOGISTICS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
Prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 14 juni 2013 om 14.45 uur

door

Elisa Monica Alvarez

geboren op 31 augustus 1984
te Curaçao

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. W.H.M. Zijm

en de assistent-promotor:

dr. M.C. van der Heijden

Contents

Chapter 1

Introduction

1.1 The golden age of after-sales services

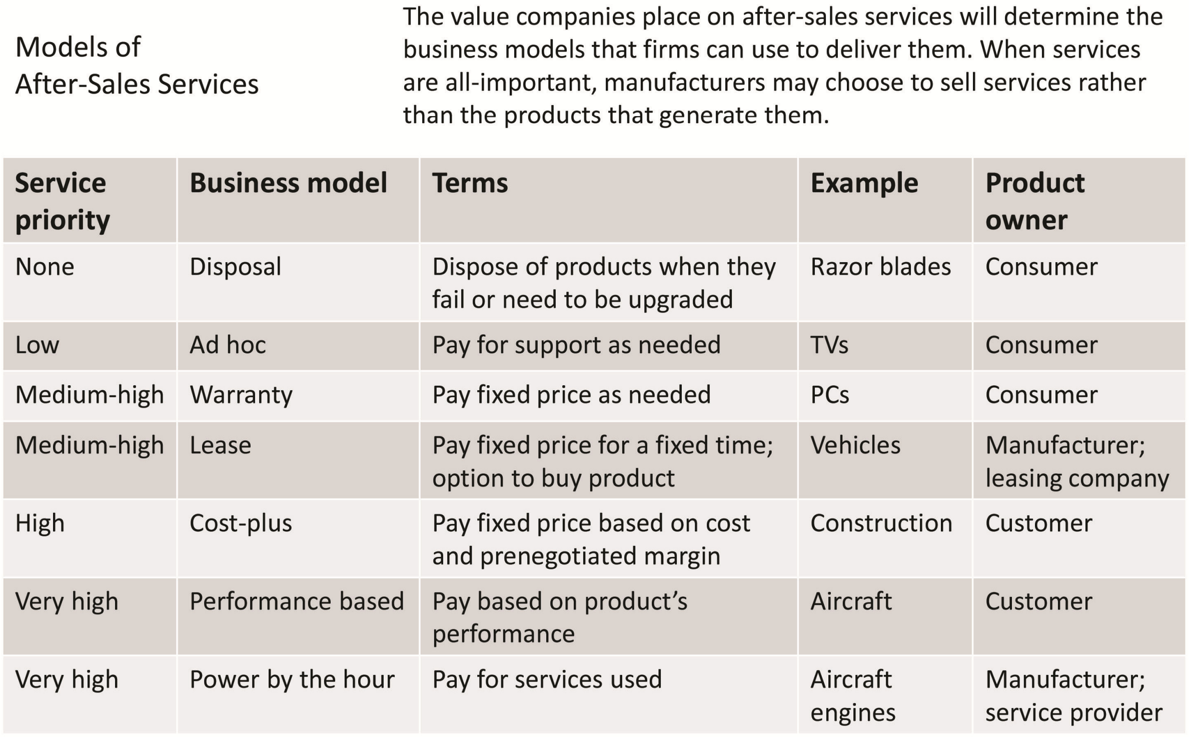
In the current business environment, equipment availability is generally crucial for a company's operations. For such so-called *mission critical equipment* (Kim et al., 2010), an operational failure resulting in downtime is highly undesirable and can indeed be very costly: in the semiconductor industry, for instance, an hour of downtime can amount to a loss of productivity of tens of thousands of euro's (Kranenburg, 2006). Equipment failure can also have severe consequences in terms of safety and security. For instance, malfunctioning parts aboard an aircraft can lead to a crash, as was the case with Turkish Airlines flight 1951 in 2009, where a malfunctioning radio altimeter triggered the crash of the plane, killing nine people.

To protect themselves from such consequences, users of critical equipment require *services for system upkeep* for the time that the system is being used. Examples of such services are inspections and preventive maintenance activities to limit the number of failures that occur, repair activities once failures occur (i.e. corrective maintenance), upgrades and equipment overhauls, and technical support. Often, users are unable or unwilling to provide all these services themselves. Therefore, they demand such services from the equipment suppliers or manufacturers. In turn, suppliers increasingly find these services to be a source of large revenues and profit margins (Cohen et al., 2006). The fact that customers value equipment availability highly means that they are willing to pay an additional fee for these services, often resulting in higher profit margins for suppliers (OEM's) than those received on the sales of the equipment itself. Furthermore, the revenues tend to be constant over the (generally long) life cycle of the system (Oliva and Kallenberg, 2003). Such a constant revenue stream is particularly beneficial for companies whose overall revenue and profits are highly sensitive to economic or market fluctuations. An example of such a company is ASML, a manufacturer of lithography equipment for the semiconductor industry.

As a result of these trends, we are currently in the golden age of after-sales services (Cohen et al., 2006), where revenues can range between 25% and 60% of a company's total revenues (Deloitte, 2007). Service providers can exploit various business models for after-sales services, as shown in Figure 1.1. On the one hand, simple ad hoc models may be considered, where users pay for support as needed. Under such models, also referred to as time and material contracts

(Cohen, 2012), the service provider is compensated based on the amount of resources consumed (e.g. labor, spare parts), with no guarantees being provided on service performance (such as a minimum system uptime fraction). Conversely, more sophisticated business models can be considered, such as *performance based contracting* or *'power by the hour'* type models, where payment to the service provider reflects the value the user attributes to equipment availability. Guajardo et al. (2012) give an example of performance based contracting in the aircraft engine industry, where a service provider is paid in proportion to the number of aircraft flying hours, which is affected by the uptime of the engine. In power by the hour models, the manufacturer often retains equipment ownership, with the user paying for the services used. An example of a company that uses such models is Océ Technologies, a manufacturer of printing equipment in the Netherlands. At Océ, many customers are only interested in having printing capability rather than owning the printing equipment itself. Océ therefore has business models where customers pay an amount of money for every 1000 prints, with the equipment manufacturer retaining ownership and therefore being fully responsible for maintenance. In such sophisticated models, service providers often formalize agreements with their customers in so-called *service contracts.*

The remainder of this chapter is as follows. In Section 1.2, we first discuss service contracts in more detail. We then elaborate on the activities and resources needed for service contract fulfillment in Sections 1.3 and 1.4 respectively. In Section 1.5, we address service level differentiation, which is a key issue in this dissertation. Next, we state our main research goal in Section 1.6, and elaborate on related literature in Section 1.7. In Section 1.8, we state our main contribution and detailed research objectives. To meet these objectives, we require certain techniques which are discussed in Section 1.9. Finally, we present an outline of the remainder of this dissertation in Section 1.10.



Models of After-Sales Services

The value companies place on after-sales services will determine the business models that firms can use to deliver them. When services are all-important, manufacturers may choose to sell services rather than the products that generate them.

Service priority	Business model	Terms	Example	Product owner
None	Disposal	Dispose of products when they fail or need to be upgraded	Razor blades	Consumer
Low	Ad hoc	Pay for support as needed	TVs	Consumer
Medium-high	Warranty	Pay fixed price as needed	PCs	Consumer
Medium-high	Lease	Pay fixed price for a fixed time; option to buy product	Vehicles	Manufacturer; leasing company
High	Cost-plus	Pay fixed price based on cost and prenegotiated margin	Construction	Customer
Very high	Performance based	Pay based on product's performance	Aircraft	Customer
Very high	Power by the hour	Pay for services used	Aircraft engines	Manufacturer; service provider

Figure 1.1 Models for after-sales services (source: Cohen et al., 2006).

1.2 Service contracts and service level agreements

As mentioned, our focus is on complex systems that are crucial to customers' operations. As a result, clear agreements are necessary between the service provider (either the original equipment manufacturer, OEM, or an external service provider) and its customers with respect to the services required. Such agreements are formalized in *service contracts*, in which the service provider specifies the service activities it provides. Service contracts often contain *service level agreements* (SLAs), that may be based on a wide range of performance indicators.

Case studies at Océ Technologies and Philips Healthcare (a manufacturer of medical image processing equipment in the Netherlands) revealed that SLAs were made on the maximum waiting time for an engineer once a failure occurs (i.e. the response time), on the maximum amount of downtime per failure, and on the minimum fraction of system uptime, amongst others. These indicators reflect the importance that customers place on system availability: users of critical equipment are primarily interested in the capability that the equipment provides, which translates into requirements on availability. The specifics of maintaining the equipment, such as resources and activities needed for repair, are not relevant, especially if the user does not own the equipment itself. In practice, we see that increasingly high service levels are being demanded, such as minimum uptime fractions of at least 96%, and average response times of 3 to 4 hours. In the semiconductor industry, response time targets may even be 15 minutes (Cohen et al., 2006). To ensure that the agreements are met, the service contract may also stipulate penalties if the service provider is unable to meet the contract requirements. An example of a company where such penalties apply is Thales Netherlands, a producer of electronic equipment such as radar systems for defense applications (Al Hanbali and Van der Heijden, 2011).

Given the performance indicators agreed upon and the associated service levels, the service provider must design its service fulfillment process such that all agreements are met at minimum costs. This is far from trivial. First, service providers often have little insight in the logistic costs that are necessary to guarantee a specific service level. Indeed, in practice we see that service agreements are made which subsequently cannot be met using the service provider's current service organization. Second, even if a service provider has insight in the relation between service levels and logistic costs, it remains a difficult matter to ensure that all service agreements are met. In particular, the system uptime depends on various factors, some of which are not entirely within the service provider's control. For instance, the system failure rate also depends on the utilization and operating conditions of the system, amongst others. In practice, the customer performance indicators are often translated into more concrete internal performance indicators on different aspects of the fulfillment process, such as the waiting time for engineers or spare parts. Given such more manageable indicators, we present control options to make the trade-off between the required level of service and the associated costs.

We now first elaborate on the fulfillment process itself, specifically the process of repairing systems once a failure occurs (i.e. corrective maintenance). We do realize that fulfillment can also take place through preventive maintenance. However, corrective maintenance is generally much more unpredictable than preventive maintenance and tends to occur at very inconvenient times. Indeed, failures generally occur at moments when the system is being heavily utilized (with failures often being the result of intensive system use). In contrast, preventive maintenance can often be scheduled at convenient moments. For instance, preventive maintenance of baggage handling systems at airports generally occurs at night. Furthermore, preventive maintenance is only beneficial for components that exhibit wear-out (i.e. have an increasing hazard function (Kumar et al., 2000)). For other types of components, preventive maintenance does not improve the system's reliability. For instance, a case study at Océ technologies revealed that certain printers have various (electronic) components with a constant failure rate. Preventive maintenance thus has very limited influence on the failure behavior of these printers.

1.3 The corrective maintenance process

To understand how service providers can effectively offer service to their customers, we consider the corrective maintenance process provided by a manufacturer of healthcare equipment (Figure 1.2).

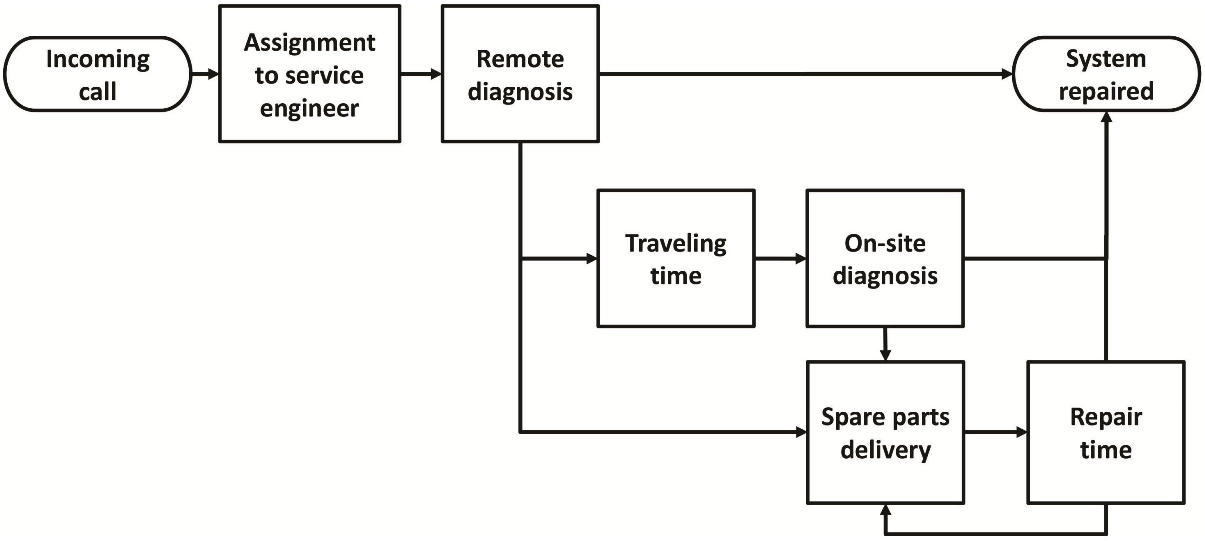


Figure 1.2 Typical steps in a corrective maintenance process (Prakken, 2009).

After a failure occurs, the customer contacts the service provider (in this case the manufacturer), who assigns an engineer to that customer when available. The engineer will first attempt to remotely diagnose the problem, for instance by phone or by logging on to the system. In some cases, a remote repair will be possible. Alternatively, a spare part may be necessary for the repair. This part will then be sent to the customer and the engineer will travel to the customer's site to install it. A third option is that remote diagnosis is not possible. Then, the engineer must diagnose the problem at the customer's site, which could reveal the need for a spare part. This part is then ordered and the engineer will need to return to install the part at a later time. In some cases, the system might still not function after a repair. Then, the diagnosis, supply, and repair steps continue until the system is functioning again.

Clearly, various resources are needed for corrective maintenance, such as service engineers with the right expertise, tools for diagnosis and repair, spare parts to replace failed items, and so forth. The speed of the corrective maintenance process, and hence the availability of the system, thus depends on the availability of these resources during the process. Conversely, a minimum availability requirement can be translated into a maximum delay time in the corrective maintenance process. Specifically, we express availability as follows (Dinesh Kumar, 2000):

$$Availability = \frac{MTBF}{MTBF + MTTR + MLDT} \qquad (1.1)$$

Here, $MTBF$ denotes the mean time between failures, $MTTR$ denotes the mean time to repair (i.e. the actual repair time when all resources and parts are available), and $MLDT$ denotes the mean logistic delay time, which is the time waiting for all resources and parts to become available. Hence, given that the mean time between failures and the mean repair time are known, an availability value can be translated into an allowed delay for obtaining resources. Below, we further elaborate on the resources required in the service fulfillment process.

1.4 Resources needed for service contract fulfillment

In general, various resources are needed for the repair process, such as spare parts, service engineers and tools. A key issue with respect to these resources is that they generally have to be deployed in advance of the actual failures to ensure that service requirements are met (Cohen et al., 2006). However, these resources are scarce and can be very expensive. In practice, modules or parts (e.g. aircraft engines) can have values of well over 100.000 euro. Furthermore, failures of complex and critical systems generally occur infrequently and at random moments in time (Kim et al., 2010). Therefore, service providers find it difficult to determine how many resources they require and where these resources should be located.

1.4.1 Service engineers

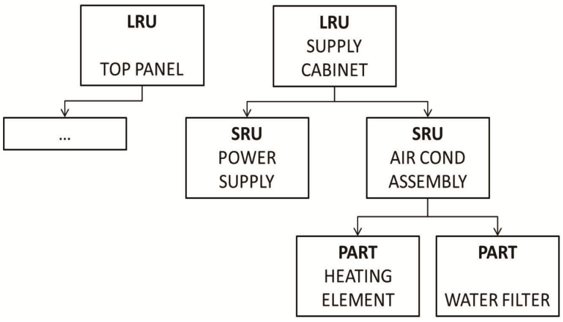
The diagnosis and repair activities are performed by service engineers. In contrast to consumer goods, which are usually serviced at a repair facility, engineers in the capital goods industry generally travel to customers' facilities to perform maintenance (Armistead and Clark, 1991). Therefore, a key performance measure is *response time*, which is the time between the reporting of the failure and the arrival of the engineer at the customer's site (see e.g. Tang et al., 2008).

Jalil (2011) distinguishes three levels at which decisions with respect to service engineers must be made, namely a strategic, tactical, and operational level. At the *strategic* level, the service area must be disaggregated into sub-areas, with each sub-area having a separate pool of service engineers. Usually, the installed base of the service provider (i.e. the set of systems it should service) is dispersed over a large geographical area. As a result, travel distances to customers can be extensive, which makes it impractical – or even infeasible – to service all customers from a single service point. Therefore, various service points are located across the entire service area, with customers mainly being serviced by the closest service point. At the *tactical* level, there is the so-called *manpower planning problem*: The number of engineers and their skills must be determined per sub-area such that response time targets are met at minimal costs. This problem is closely related to engineer *utilization*, i.e. the fraction of time the engineer spends on travel and repair: high utilization rates might result in engineers not meeting response time requirements, whereas a low utilization rate indicates that engineers are not used efficiently (Tang et al., 2008). Finally, at the *operational* level dispatching rules should be chosen for dispatching available engineers to customers. Various dispatching rules have been discussed in the literature, such as first-in-first-out (FIFO) rules and rules that depend on customers' response time targets (e.g. the earliest expiration time, which is comparable to the earliest due date (EDD) in a manufacturing environment, see e.g. Haugen and Hill (1999)). We refer to Jalil (2011) for a literature overview.

1.4.2 Spare parts supply

In the corrective maintenance process, failures are often solved through *repair by replacement*: a defective part is removed from the system and replaced by a functioning one. Any delay in shipping the required part will result in an increased mean logistic delay time, and thus has an impact on availability. Hence, a key performance indicator is the *downtime waiting for parts (DTWP)*, the time that elapses between the moment a part is requested and its availability at the customer's site.

The systems we consider usually have a *multi-indenture product structure*. In spare parts optimization literature, an indenture level indicates a level in the Bill-Of-Materials (BOM) structure at which repairs are performed. At the highest indenture level, we find the first indenture items or so-called *Line Replaceable Units (LRUs)*, usually the modules or main components of which the system is made up. System repair can occur by replacing a faulty LRU by a new one. The faulty LRU, in turn, is sent to a repair shop where it can be repaired by replacing a defective subcomponent, a so-called *Shop Replaceable Unit (SRU)*. Possibly, some SRUs can also be repaired by replacing cheaper parts, and so forth. As a result, we require a diverse and extensive set of spare parts. The supply of these parts occurs through a spare parts supply chain, consisting of various stock points. To service a widely dispersed installed base, stock is kept both at locations close to customers – or possibly even at customers' sites – for fast supply times in case of failures and at central stock locations where stock is pooled both for resupplying the local stock points and possibly for satisfying customer demand through an emergency shipment if the local stock points are out of stock. Such a structure is referred to as a *multi-echelon* structure, where lower echelons (e.g. local stock points) are resupplied by higher echelons (i.e. central locations). Figure 1.3 and Figure 1.4 give examples of the multi-indenture structure and the multi-echelon structure respectively, with the multi-echelon structure based on the setting at Thales Netherlands.



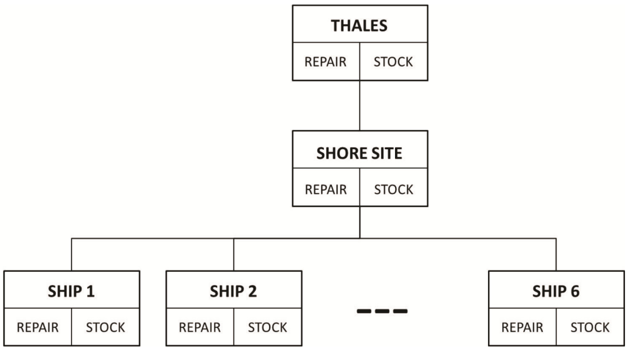


Figure 1.3 Example of a multi-indenture structure.

Figure 1.4 Example of a multi-echelon structure.

The waiting time for parts depends on (i) the *amount* of stock kept in the system at various locations and indenture levels, and (ii) the *throughput times* between stock points and from stock points to customers. With respect to the amount of stock kept, we first note that most components of technologically complex systems are expensive, and therefore are preferably repairable, and generally have low demand rates (so-called *slow movers*). As a result, their inventory is often controlled using one-for-one replenishment policies (i.e. *base stock* $S-1,S$ *policies*), since their holding costs generally outweigh their order costs and they are infrequently requested. For each item, we thus need to determine a base stock level at each of the various locations in the system. A key issue in doing so is that spare parts are only kept in stock to ensure the availability of the system *as a whole*, which is accomplished by determining suitable availability levels for each individual part. As a result, a so-called *system approach* (Sherbrooke, 2004) is required, where stock levels are jointly determined for all items such that a certain system uptime is guaranteed. The idea behind this approach is that inexpensive high demand parts should be available immediately (and thus have high stock levels), while the waiting time for expensive low demand parts can be somewhat longer.

The amount of literature on (multi-item) spare parts optimization models is extensive and dates back to Sherbrooke (1968), who developed the METRIC (Multi-Echelon Technique for Recoverable Item Control) model. Sherbrooke (2004) and Muckstadt (2005) give an overview of further developments in this area. Generic models often consider multi-echelon settings, with demand arriving at locations at the lowest echelon level according to Poisson processes and one-for-one replenishment being used at all locations. All items are repairable (i.e. no condemnation occurs), and repair may occur at various locations, depending on the level of complexity. A key assumption with respect to system evaluation is that the number of outstanding orders at the base follows a Poisson distribution. Optimization occurs by a stepwise increase of stock levels in a greedy 'biggest-bang-for the-buck' manner (also referred to as marginal analysis), where iteratively the option is selected that gives the largest contribution according to some criterion (such as the largest decrease in the mean backorder level per dollar additional investment). These iterations are repeated until some stopping criterion (e.g. a desired mean backorder sum) is satisfied. Various extensions have been made to the METRIC model, notably MOD-METRIC (Muckstadt (1973, 1979)) that incorporates multi-indenture product structures, and VARI-METRIC (Slay (1984), Graves (1985), Sherbrooke (1986)) that more accurately estimates the distribution of outstanding orders at each base.

In addition to stock levels, waiting time is also influenced by the throughput times in the supply chain, such as item repair times or shipment times from upstream locations in the chain to downstream locations. Usually, throughput times can be influenced to some extent at a certain price, for instance by using priority rules in the repair shop, or by satisfying a fraction of demand using a faster (but also more expensive) shipment option instead of regular supply. An example of a company that uses such options is Thales Netherlands. Two common shipment options considered both in literature and in practice are (i) *lateral transshipments,* which are used when a preferred location is out of stock, but a location on the same echelon level has the part on-hand, and (ii) *emergency shipments,* where demand is met from stock at a higher echelon level when the preferred stock facility is out of stock. As a result of these possibilities to reduce throughput times, a trade-off exists between the amount of stock that must be kept and the throughput time: if throughput times are shortened, the same quality of service can be provided with a smaller stock pool.

1.4.3 Tools

Various service tools might also be required during the corrective maintenance process, for instance diagnostic and calibration tools. Like spare parts, tools have low demand rates and can be very expensive (up to hundreds of thousands of euro's). In contrast to spare parts, however, tools are never consumed during repair. Furthermore, various tools are often needed simultaneously (so-called coupling in demand, see Vliegen (2009)). Tools are often combined in so-called tool kits that can be used for multiple types of repairs. We refer to Vliegen (2009) for literature in this area.

Often tools are small and therefore as mobile as spare parts and service engineers. However, certain tools can be very large and therefore difficult to move. For instance, to replace the bogies on trains, specialized equipment is needed to lift the trains. Similarly, ships have to be placed in docks for certain kinds of repairs. The location of such tools influences where components can be repaired, as certain repairs can only occur if these tools are available. This is one of the issues addressed in a so-called level of repair analysis (LORA). The focus of LORA is to determine whether a component should be repaired upon failure and, if so, at what repair location. Decisions are also made with respect to the location of service tools. The objective is to minimize life cycle costs. We refer to Basten (2010) for a review of literature on LORA.

1.5 Service level differentiation

A complication in designing the fulfillment process is that customers value services differently. A system might be crucial for one customer's operations, whereas it is less crucial for another customer. For instance, the uptime of mainframe computers at a stock exchange will be much more crucial than the availability of a computer in a regular office environment. The number of systems a customer uses also determines the severity of downtime. If a company has various printers, for instance, the unavailability of one printer will not have large consequences, especially if employees can easily forward their printing jobs to another printer. In contrast, if a company only has a single printer, downtime will be very undesirable. As a result of these differences, certain customers will require very high service levels (e.g. an uptime fraction of 98% or more), and therefore take an expensive service contract, whereas cost oriented customers can opt for a less expensive contract and accept greater delays (leading to a smaller uptime fraction). The overall customer pool can thus be divided into various customer segments that must all be served from a single organization.

In practice, service providers primarily handle the heterogeneity in service levels by using priority rules when assigning service engineers to customers. At Océ Technologies, for instance, an engineer is always assigned to the highest priority customer. Differentiation in spare parts supply, on the other hand, is handled in one of two extreme ways. At one extreme, all customers are provided with uniform service (a so-called "one-size-fits-all" approach), with the supply process designed to meet high service levels. As a result, this option is very costly, since customers with standard contracts receive better service than their contract requires. In contrast, service levels of premium customers may not always be attained, because the resources may have been used for standard customers instead. Also, standard customers have no incentive to switch to a premium contract if their system becomes more critical in their operations, since they already receive high service. At the other extreme, service providers use differentiated supply chains for each customer segment. For instance, stock for premium customers might be kept close to their sites to minimize downtime, while stock for standard customers is kept at some central location resulting in longer lead times. Such supply chains become difficult to handle when they interact (because they share resources). Also, separate stock piles reduce the benefits of stock pooling compared to the case where all stock is kept centrally (Eppen and Schrage, 1981), which may lead to higher overall stock levels.

The literature on service differentiation in spare parts supply has focused on a solution between these two extremes, with stock rationing or the so-called *critical level policy* being the main differentiation tool. The critical level policy keeps stock for all customers at a central location, but reserves part of this stock for requests from high priority customers. Specifically, demand for lower priority customers is only met from available stock if this stock exceeds a threshold value (i.e. a critical level). We refer to Teunter and Klein Haneveld (2008) for a literature review in this area.

In theory, rationing can lead to large cost savings compared to one-size-fits-all approaches. However, there are practical drawbacks to using this approach. First of all, the service engineers who repair the systems usually are accountable for the speed of repair. As these engineers often have also developed a relationship with their customers – for instance, because they always service the same customers – they opt to use an available part for a non-premium customer anyway. Second, customers have access to stock information at certain service providers. These service providers then become reluctant to refuse a part to non-premium customers when the latter can see that the part is actually available. For example, a case study at Philips Healthcare revealed that critical level policies were not used for this reason at the time of research.

1.6 Research goal

Our aim is to determine the added value of various control options for service contract fulfillment, particularly in settings with differentiated service levels. We mainly focus on control options in spare parts supply. However, as shown in Section 1.2, system downtime depends on various additional resources, such as service engineers. Indeed, a case study at Océ technologies showed that the waiting time for service engineers exceeded the waiting time for spare parts. Therefore, we also investigate the added value of applying differentiation with respect to service engineer utilization.

Our first research area is the application of differentiation on an item level in spare parts supply by selectively **reducing item throughput times** in a multi-echelon multi-indenture setting. To this end, we allow repair times and lead times between locations to be decision variables in addition to the stock levels in the system. Our aim is to investigate whether the reduction of throughput times for certain items can result in a significant reduction in the stock levels required in the system.

Our second research area focuses on applying differentiation in spare parts supply on both an item *and* a customer level. Specifically, we investigate control options to differentiate service to customers based on their service requirements. Our aim is to consider control options that are as effective as critical level policies in terms of cost savings over one-size-fits-all policies, while being easier to implement in practice. Therefore, we investigate the following options:

- **Selective emergency shipments:** in this case, stock is used to satisfy demand from all customers if possible. If a location is out of stock, the service provider often has the option to procure the part at a higher echelon level using an emergency shipment. Generally, emergency shipments are both faster and more expensive than waiting for an item to arrive through regular replenishment. Differentiation occurs by using emergency shipments for specific customer segments and/or types of items.

- **Selective lateral transshipments:** as with selective emergency shipments, service providers sometimes have the option to request transshipments from other locations at the same echelon level in the system to satisfy demand in out-of-stock settings. Differentiation again occurs by limiting the use of this option to premium customers and specific item types.
- **Dedicated customer stocks:** in practice, dedicated stocks of items are sometimes kept at certain customers' locations in addition to stock kept at central stock points to minimize waiting time for spares. So far, the added value of this form of customer differentiation has not been analyzed from a scientific point of view.

A benefit of these options is that stock does not need to be withheld from a non-premium customer when his system fails. Still, differentiated service can be provided by reserving lateral and emergency shipments for high priority customers, or by placing reserved stock directly at those customers' sites. We evaluate the added value of these options by using one-size-fits-all policies and critical level policies as benchmarks. Furthermore, we investigate the added value of combining individual control options for differentiation in one aggregate policy.

Finally, we elaborate on the possibility of applying service differentiation in other activities in the maintenance process. Specifically, we investigate the added value of using **priority mechanisms when assigning service engineers** to customers at a tactical level.

1.6.1 Main research objective

Our main research objective can be formulated as follows:

To develop mathematical models that give insight in the added value of the various control options for the fulfillment of, possibly differentiated, service levels agreed upon with customers.

Key elements of this research objective are:

- *Models:* We focus on developing mathematical models to analyze systems under various control options. These models allow us to evaluate a system for a given control option, resulting in accurate estimates of the system's performance, and determine decision variables such that the (differentiated) service levels are met at minimum costs.
- *Insight in added value:* for each control option, we investigate for what kinds of systems (e.g. for what shipment time values) and item types the option leads to significant cost savings.

- *Service levels:* we focus on system downtime that is caused by lack of resources such as service engineers and spare parts. Specifically, we consider the response times for service engineers and the waiting time for spare parts. In both cases, we focus on *average* times, instead of maximum times that must apply for each failure separately.

1.7 Literature

In this section, we present a detailed overview of the literature related to our area of research. We first focus on the literature related to spare parts supply for expensive slow moving items, starting with a general overview of multi-echelon multi-indenture systems (Section 1.7.1). Subsequently, we focus on literature considering throughput time differentiation (Section 1.7.2), service differentiation (Section 1.7.3), and lateral transshipments and emergency shipments (Section 1.7.4). Finally, we discuss literature on service engineers (Section 1.7.5). At the end of each section, we shortly highlight the key observations for that section.

1.7.1 Multi-echelon multi-indenture systems

We first focus on the setting where unmet demand is backordered, starting with approaches for system evaluation and then proceeding to system optimization. Subsequently, we briefly discuss variants where unmet demand is lost to the system (with such demand being satisfied using alternative options such as emergency shipments).

As mentioned before, the *evaluation* of multi-echelon multi-indenture systems starts with the METRIC model of Sherbrooke. Various extensions have been made to this model, e.g. by Muckstadt (1973, 1979) who incorporates two-indenture product structures. Furthermore, approaches have been developed to more accurately estimate the distribution of the number of outstanding orders (i.e. the pipeline) at each base. A two-moment approximation technique has been considered (Slay (1984), Graves (1986), Sherbrooke (1986)), where the pipeline distribution is approximated by a negative binomial distribution from the pipeline mean and variance. This improved approximation technique culminated in VARI-METRIC, which can be used in multi-echelon multi-indenture settings. Exact optimization approaches have also been developed by Graves (1985) for a multi-echelon single-indenture setting and Rustenburg et al. (2003) for a multi-echelon multi-indenture setting where commonality may apply (i.e. a subassembly may be used in different systems).

In addition to system evaluation, approaches have been developed for *setting stock levels that minimize the total investment*. This has been done both for cost models (with penalty costs for backordered demand incorporated into the cost function) and service models (where a target service level needs to be satisfied), cf. Van Houtum and Zijm (2000). In cost models, the lack of service restrictions allows a multi-item problem to be decomposed into single-item problems that can each be solved individually. In this area, single-item models are therefore considered.

Contributions are, amongst others, by Axsäter (1990), who considers an exact optimization approach, and Gallego et al. (2007), Rong et al. (2010) and Basten and Van Houtum (2013), who consider heuristic methods. In contrast, service models often are multi-item models. Contributions in this area are, amongst others, by Hopp et al. (1999), Caglar et al. (2004), Wong et al. (2007a), and Caggiano et al. (2007). The first three models have restrictions on the aggregate mean waiting time, while Caggiano et al. consider time-based fill rate restrictions (i.e. the minimum fraction of demands that must be met within specific time intervals). Nowicki et al. (2012) developed an approach to improve the efficiency of optimization through marginal analysis in METRIC-type models. Note that cost models are equivalent to service models under certain conditions (Van Houtum and Zijm, 2000).

The literature discussed so far considers models where demand is backordered if it cannot be met from on-hand stock. In contrast, the amount of papers considering multi-echelon *lost sales systems* is much more limited. A key reason for this is that lost sales models are much more difficult to analyze than their backorder counterparts (Bijvank and Vis, 2011). To our knowledge, research in this area is limited to single-indenture two-echelon models, such as Andersson and Melchiors (2001) and Hill et al. (2007). In the literature on emergency shipments we also find such models (see Section 1.7.4). We refer to Bijvank and Vis (2011) for details and further references.

In summary, multi-echelon multi-indenture systems have been studied extensively under *full backordering*, with various (exact and approximate) approaches being given for both system analysis and optimization. Under *lost sales*, in contrast, the types of systems studied are much more limited.

1.7.2 Joint optimization of inventories and throughput times

In the literature of the previous section, throughput times such as repair times and replenishment lead times from higher echelon locations to lower echelon locations are given as parameters. In the last decades, several models have been developed that jointly consider, amongst others, stocking levels and throughput times. We discuss literature at strategic, tactical and operational levels.

At a *strategic* level, joint decisions are made on stock levels and repair locations, taking into account the costs of resources required, as discussed by Alfredsson (1997) and Basten et al. (2012a) amongst others. Papers such as Rappold and Van Roo (2009) and Wu et al. (2011) even combine the spare part stocking problem with facility location analysis, including the location and number of stock points in the system. Rappold and Van Roo (2009) focus on a single-item, single-indenture setting with finite repair capacity, where both the location and capacity of repair facilities are decision variables. Wu et al. (2011) focus on a multi-indenture case where multiple shipment modes are possible from higher to lower echelon locations.

At a *tactical* level, the focus lies on jointly optimizing spare parts levels and repair and supply processes. Both models with multiple shipment options and models with finite repair capacity have been considered, with the latter models focusing both on the number of repair servers needed and the priority setting when repairing items. In the area of *multiple shipment modes*, Verrijdt et al. (1998) consider a single item model to show the impact of emergency repairs if the stock level drops below a certain threshold value. Perlman et al. (2001) consider a single-item, two-echelon model with finite capacity repair shops and assume that emergency repair is applied with a certain probability. Levner et al. (2011) also consider a single-item two echelon model with multiple supply alternatives. They consider both the possibility of repairing items at a repair facility with infinite capacity or purchasing new items at an external supplier. Furthermore, the repair facility has two repair modes (fast and slow), with the fast option being both faster and more expensive. Kutanoglu and Lohiya (2008) consider a single-echelon single-item system with multiple stock facilities where replenishments to a customer can occur through multiple shipment modes (that vary in terms of speed and costs). As a last resort, an emergency shipment from a central facility with infinite stock can be used. Van Utterbeeck et al. (2009) focus on the design of a supply chain, with key decisions being whether the system should be single- or two-echelon, and what kind of supply flexibility should be used (no flexibility, lateral transshipments only, or both lateral and emergency shipments). The models with *finite repair capacities* usually model the repair shops as single or multi-server queues with exponentially distributed repair times (Gross et al., 1983; Diaz and Fu, 1997; Sleptchenko et al., 2003). Finite capacity models are suitable if a service provider has its own repair facilities as opposed to outsourcing repair to an external company. In a multi-echelon multi-indenture setting, Sleptchenko et al. (2005) introduce priority queuing models for the repair shop where the items are assigned to two priority groups (high or low priority). They show that appropriate priority assignment may lead to a significant reduction in the spare part inventory investment. The idea is to prioritize repair of items with high value and small repair times, so that the work-in-process of these items is reduced with limited impact on other items. Adan et al. (2009) use a similar idea when they consider multiple priority classes (>2) in a single-location, single-indenture problem. They develop a method for exact cost evaluation.

At the *operational* level, various priority rules have been examined. These models assume that all resources are given (spare parts stock locations and levels, repair locations and capacities) and search for efficiency gain using (i) *repair priority* rules that determine in what order defective items are repaired, and (ii) *dispatch priority* rules that specify how ready-for-use items are handled. With respect to dispatch priority rules, a further distinction can be made in *item allocation rules* that assign incoming items to outstanding orders for that item and *demand allocation rules* that determine from which location an incoming request for a part will be satisfied. Regarding repair priorities, Hausman and Scudder (1982) discuss several rules in a single-location, three-indenture model. The best rules lead to a backorder reduction equivalent to a 20% reduction in inventories. Scudder (1984) extends this model to the multiple failure case and finds similar results. Pyke (1990) combines repair priorities with item allocation rules in a simulation study and concludes that priority repair improves system performance, whereas allocation rules have limited impact. Caggiano et al. (2006) develop two methods to set repair priorities and item allocation rules in two-echelon networks within a finite planning horizon. They show that significant gains are feasible in a rolling horizon setting. Tiemessen and Van Houtum (2013) show that operational repair priorities may yield about 10% cost reduction on top of static repair priorities in a multi-item, single-location model. Jalil (2011) and Tiemessen et al. (2013) both consider demand allocation in single-item single-echelon systems with multiple warehouses and multiple customer classes (with each customer class having distinct penalty costs). Customer requests can be met by one of the warehouses or can be lost. Key issues are that lower priority requests should not necessarily be satisfied from the nearest warehouse (or from any warehouse) in order to reserve stock for premium requests.

In conclusion, research on throughput time optimization has occurred at various planning levels. At a *tactical level*, both models with multiple shipment modes and models with finite repair capacity have been considered. Most research focuses on *single-item single-indenture* systems.

1.7.3 Service differentiation and multiple demand classes

The literature on multiple demand classes considers systems where the customer base can be segmented into different groups (i.e. demand or customer classes). We focus on models where segmentation results from varying service level requirements or shortage costs. At an operational level, service differentiation has been considered by Jalil (2011) and Tiemessen et al. (2013), as discussed in the previous section. In this section, we discuss literature at a tactical level. In that area, service differentiation is accomplished through *critical level policies*, a concept that has been introduced by Veinott (1965). We first discuss literature in which the optimality of the critical level policy is examined. Subsequently, we focus on the use of critical level policies in *single-item* models under three settings: under full backordering, under lost sales, and in settings where both backordering and lost sales are possible. Finally, we discuss literature where critical level policies are used in a *multi-item* setting.

The critical level policy has been *shown to be optimal* under various settings, such as for Poisson demand under exponential service times (see e.g. Ha (1997b) and De Véricourt et al. (2002) for the backordering case and Ha (1997a) for the lost sales case), under Erlang service times (e.g. Ha (2000) under lost sales and Gayon et al. (2009) under backordering), and under hypo-exponential service times with lost sales (Wieczorek et al., 2011). Other settings where optimality has been proven are, amongst others, assembly systems (Benjaafar et al., 2011), systems where both interarrival times and service times follow an Erlang distribution (ElHafisi et al., 2010), and production systems with capacity restrictions (Zhou et al., 2011). We refer to

Teunter and Klein Haneveld (2008) for a detailed overview of papers in this area. Common – and intuitively clear – findings are that it is never optimal to withhold stock from highest priority customers. Also, critical levels are non-decreasing in priority classes (i.e. if stock should be withheld from a class i customer, it should also be withheld from a class j customer that has lower penalty costs), see e.g. De Véricourt et al. (2002) and Gayon et al. (2009). In some settings, the optimal critical level is state-dependent, e.g. the critical level will be low – or even zero – if a replenishment order will arrive soon.

Critical level policies have been applied in *single-item models with lost sales* (with emergency shipments being used for demand that cannot be met from on-hand stock). Recent contributions are by Dekker et al. (2002) and Kranenburg and Van Houtum (2007b). Dekker et al. (2002) consider a single-item model with N classes, one-for-one replenishment and static critical levels. The authors present optimal solution procedures for both a cost model (where stock and critical level values are determined to minimize the sum of holding and penalty costs) and a service model (where decision variable values are determined to minimize holding costs subject to service level restrictions). Furthermore, the authors present a heuristic for solving the cost model. Kranenburg and Van Houtum (2007b) consider the same cost model as Dekker et al. (2002), but provide three approaches for finding optimal critical levels for given base stock values (with Dekker et al. (2002) providing bounds on the base stock levels themselves). Their approaches are much faster than the complete enumeration approach used by Dekker et al. (2002). Furthermore, Van Jaarsveld and Dekker (2009) prove that two of these approaches (algorithms 1 and 2) are in fact optimal.

Single-item systems with critical level policies under full backordering have been considered both under periodic review (e.g. Möllering and Thonemann (2010) for a setting with 2 customer classes) and under continuous review (e.g. Nahmias and Demmy (1981), Deshpande et al. (2003) and Fadiloğlu and Bulut (2010) for the case with 2 customer classes, and Arslan et al. (2007) and Abouee-Mehrizi et al. (2012) for multiple customer classes). A complicating factor for these models is that positive stock levels and backorders for lower priority classes can occur simultaneously. As a result, authors must keep track of both outstanding orders and backorders for each customer class in order to analyze the system. Furthermore, a backorder clearing mechanism is necessary that specifies how incoming replenishment orders are handled. Naturally, any backorder of high priority customers should be cleared as soon as possible. The clearing of lower priority backorders, in contrast, might need to be delayed in favor of increasing inventory in anticipation of future high priority arrivals. A clearing mechanism that is optimal in certain cases (Ha, 1997b) is priority clearing, where non-premium backorders are only cleared once all premium backorders have been cleared and the inventory level is at least the critical level (see e.g. Fadiloğlu and Bulut (2010)). As this mechanism may result in intractable models (and can lead to very high waiting times for non-premium customers), alternative mechanisms have been considered as well (see e.g. Deshpande et al. (2003), Arslan et al. (2007) and Abouee-Mehrizi et al. (2012)).

A few contributions consider *single-item models with both backordering and lost sales*, both in a continuous-review setting (e.g. Enders et al. (2012), Benjaafar et al. (2010), Van Wijk (2012)) and a periodic review setting (e.g. Tang et al. (2007), Zhou and Zhao (2010a, 2010b)), see Van Wijk (2012) for details. In certain papers, the shipment option used depends on the customer class (e.g. Enders et al. (2012), where premium demand is lost, while non-premium demand is backordered), whereas other papers allow the choice of backordering versus lost sales to only depend on the system state (as in Benjaafar et al. (2010)).

We have only found two papers that consider *critical level policies in a multi-item setting*. One paper is by Kranenburg and Van Houtum (2008), who minimize holding and shipment costs in a multi-item multi-class model with class-dependent waiting time restrictions. Unmet demand is satisfied through emergency shipments. The authors use a solution approach based on decomposition and column generation, combined with greedy heuristics. Pourakbar and Dekker (2012) consider differentiation in an end-of-life problem, when production of certain parts is discontinued and the service provider places a final order quantity of parts for the remaining service life cycle. The authors mainly focus on a single-item setting and show that the optimal policy consists of time-dependent rationing and contract extension thresholds, where the former thresholds indicate whether demands from a customer class should be met or lost, while the latter thresholds indicate whether a contract type should be discontinued from some point in time onwards. They subsequently extend these findings to a multi-item setting.

To our knowledge, the *critical level policy* is the only tool that has been considered so far for service differentiation in spare parts supply at a tactical level. Most research in this area focuses on *single item* models.

1.7.4 Lateral transshipments and emergency shipments

The literature stream on lateral transshipments and emergency shipments consists both of papers in which only one kind of flexibility option is used and papers in which the options are jointly used. We first discuss literature where only emergency shipments are considered. Subsequently, we discuss literature where lateral transshipments are the only flexibility option. Finally, we focus on papers where both options are jointly used.

With respect to literature that *only considers emergency shipments*, we first note the similarity of emergency shipments to lost sales systems, as emergency shipments often occur through different channels than regular replenishments and therefore can be considered lost sales for the regular channel. Lost sales literature dates back to Karush (1957), who considers a single location with Poisson arrivals, one-for-one replenishments and mutually independent replenishment times. Karush shows that the probability of being out of stock is strictly convex as a function of the base stock level. Further contributions are, amongst others, by Feeney and Sherbrooke (1966) and Smith (1977). Literature on emergency shipments dates back to Muckstadt and Thomas (1980), who extend the METRIC model to include emergency shipments. Hausman and Erkip (1994), in turn, extend the work of Muckstadt and Thomas by presenting an improved single-echelon model to approximate their multi-echelon system. Moinzadeh and Schmidt (1991) consider a single location system where the shipment mode (normal or emergency) depends on the amount of on-hand stock and the lead time of outstanding orders. Their work is extended by Aggarwal and Moinzadeh (1994), who consider a two-echelon system. We refer to Bijvank and Vis (2011) for details and further references.

In literature where *lateral transshipments are the sole means of supply flexibility*, demand is backordered when no location at a particular echelon level has stock on-hand. A first contribution in this area is by Lee (1987), who considers a two-echelon model consisting of a depot and various bases. The bases are divided into pools, with lateral transshipments being possible among the (identical) bases in a pool. Various rules are considered for determining from which location to source a transshipment. The restriction of identical bases is relaxed by Axsäter (1990), who also introduces an improved approach for approximating service levels. Specifically, each base is analyzed independently with arrival rates being state-dependent (i.e. when the base has stock, it sees both direct requests and requests for transshipments, and otherwise it only sees direct requests that are backordered) and transshipment rates being approximated by Poisson processes. Through an iterative process, the system performance measures are updated until convergence occurs. Various papers use the same logic for analyzing their systems, also in lost sales settings (e.g. Alfredsson and Verrijdt (1999), Kukreja et al. (2001), Kranenburg and Van Houtum (2009), Van Wijk et al. (2012)). Further contributions under backordering are, amongst others, by Kukreja et al. (2001), Grahovac and Chakravarty (2001), and Tiacci and Saeta (2011).

Lateral transshipments and emergency shipments have also been considered jointly, with initial contributions by Dada (1992) and Alfredsson and Verrijdt (1999), who consider similar two-echelon models. In both models, emergency shipments are only used if lateral transshipments are not possible, which is a common assumption. Recent contributions in this area are, amongst others, by Wong et al. (2007b), who consider a multi-item variant of the Alfredsson and Verrijdt model, Kutanoglu (2008), Kutanoglu and Mahajan (2009) and Reijnen et al. (2009), who focus on time-based service levels (e.g. 60% of demand needs to be met within 2 hours, with 100% of demand met within one day), Kranenburg and Van Houtum (2009), who consider a model in which only a subset of all warehouses can act as a source of transshipments (so-called main warehouses), and Van Wijk et al. (2012), who consider a model in which transshipment requests at any warehouse are only satisfied if the stock level at that warehouse is above a certain threshold (a so-called hold back level, which is similar to a critical level). Literature reviews on lateral transshipments, both under backordering and under emergency shipments, are given by Wong et al. (2006) and Paterson et al. (2011).

Clearly, both lateral and emergency shipments have been considered extensively. Still, most papers focus on system evaluation in a *single-item setting*. Furthermore, neither lateral transshipments nor emergency shipments have been considered as *service differentiation tools* before at a tactical level.

1.7.5 Service engineers

The literature on service engineers primarily consists of *field service models,* where service engineers travel to customers' sites for diagnosis and repair. We discuss literature at the tactical level, as our focus lies on setting staffing levels (i.e., on determining the number of engineers needed to meet all service level targets). First, we discuss papers at a purely tactical level. We do so in two parts: we first discuss papers that consider a single customer class and then the papers that consider multiple customer classes. Subsequently, we discuss papers that consider decisions at both a tactical and an operational level.

At a *purely tactical level,* most papers focus on analyzing the field service network for a given staffing level (i.e. a given number of service engineers). Many papers considering a *single customer class* use queuing models for analysis. Waller (1994) considers a system where each engineer services a separate set of customers. He incorporates spare parts availability in the model through a positive probability of not having the needed part available (which necessitates a second visit from the engineer). Hill et al. (1992) and Tang et al. (2008) consider a system where a set of engineers service various customers. Both papers use state-dependent $M|G|c$ models for the analysis. Furthermore, the papers focus on determining the required number of engineers subject to service restrictions, such as response time targets (Hill et al.) and the fraction of customers serviced within a predetermined time window (Tang et al.). System analysis has also been done using simulation (e.g. Dear and Sherif (2000) and Watson et al. (1998)), with Watson et al. (1998) combining simulation with regression analysis to investigate the relationship between staffing levels, dispatching rules and staff utilization. All papers discussed so far assume that each engineer can repair any system. In practice, however, certain skills are necessary to repair a system, which may depend on the machine type. The skill levels of service engineers then influence what types of repairs they are able to perform, and hence how many engineers are required. Contributions in this area are, amongst others, by Agnihothri et al. (2003) and Colen and Lambrecht (2012), who both consider models with two job types. Both papers use simulation to determine whether service engineers should be dedicated (i.e. specialized in one type of job) or flexible (specialized in both job types).

Of the papers considering *multiple customer classes,* we again find that analysis generally occurs using queuing models, with most papers considering *exponential service times*. Papadopoulos (1996) extends the model by Waller (1994) to a multi-class setting where customers may be serviced by more than one engineer. The author analyses the resulting system as a network of queues using a priority mean value analysis approach (with high priority customers being served before lower priority customers). For an $M/M/c$ queue with various customer classes and a preemptive resume priority discipline, Buzen and Bondi (1983) give exact expressions for the mean waiting time per class when service rates are identical for all classes, with approximate expressions being given for the case that the service rates differ among classes. For the setting with non-preemptive priorities and identical service rates over all classes, Kella and Yechiali (1985) give exact expressions for the Laplace Stieltjes transform (LST) of the waiting time per class. Further contributions considering exponential service times are, amongst others, Peköz (2002) who considers the policy of delaying service to lower priority customers even if an engineer is available to reserve capacity for meeting high priority demand, Sleptchenko et al. (2005) who consider two classes that each consist of multiple customer types, each with distinct arrival and service rates, and Zeltyn et al. (2009) who consider a setting where a subset of the highest priority customer classes has preemptive priority over the remaining classes. We have also found papers that consider *non-exponential service times*. Altinkemer et al. (1998) derive approximations for the mean waiting times per class in an $M/D/c$ non-preemptive priority queue. Harchol-Balter et al. (2005) provide approximate results for the distribution of the number of customers per class in the system (and correspondingly of the time spent in the system) when service times have a phase-type distribution and a preemptive resume priority discipline is used. Wagner (1997) considers a non-preemptive priority model with a generalized Markovian arrival process and a phase-type service time distribution that is identical per class. The author mainly focuses on estimating the mean waiting times per class. Williams (1980) gives approximations for the LST and first two moments of the waiting time per class for a two-class system with non-preemptive priorities and a generalized service time distribution that is identical for both classes. Jagerman and Melamed (2003) provide similar results for the setting with n customer classes and a service time distribution that may differ per class. The authors subsequently use these results to determine the minimal number of servers needed to meet all class-specific service level targets, where the type of service level may vary per customer class.

At a *tactical and operational level*, we find the paper by Gurvich et al. (2008), who consider employee staffing and scheduling in a call center environment (where travel times thus are not considered). Customers belonging to various customer classes, each with a distinct service level requirement, are served by one of various employees. The authors argue that staffing and scheduling can be decomposed into separate problems: the necessary number of employees (i.e. the staffing level) only depends on the total customer arrival rate and the service level requirement of the lowest priority class. As a result, the system can thus be analyzed as a single-class system with one service level requirement when determining the staffing level. Furthermore, customers are assigned to employees through a threshold priority schedule, where a customer of a particular class is only served once all customers with higher priority have been served and the number of unoccupied employees exceeds a class-specific threshold value (similar to a critical level policy in spare parts supply). Other papers, such as Harrison and Zeevi (2005) and Bassamboo (2006) also consider staffing and scheduling in a call-center environment with multiple classes, but they distinguish customer classes based on the type of activity that must be performed (with each pool of servers being able to handle one or more types of activities). In both papers, the decision maker must decide how to handle an incoming customer and what action to take if a server becomes available when customers are waiting. The objective is to minimize the costs of staffing and of customers abandoning the system before service.

Overall, most literature focuses on system *analysis given a certain staffing level*, with analysis generally occurring using queuing models. We have found both papers that consider a single customer class and those that consider multiple classes under various priority mechanisms. Most papers consider Poisson arrivals and exponential service time distributions.

1.8 Contribution and detailed research objectives

1.8.1 Contribution

Overall, our main contribution points are:

- **We consider new (combinations of) control options for applying service differentiation in the fulfillment process.** The literature on service differentiation primarily focuses on the use of critical level policies in spare parts supply. Such policies are difficult to implement in practice, as explained in Section 1.5. Literature on differentiation in areas besides spare parts supply is limited, as shown when discussing literature on service engineers in Section 1.7.5. In addition to considering control options separately, we also investigate the added value of combining various control options.
- **We focus on realistic models for service contract fulfillment.** We mainly consider spare parts supply models with various items. In literature on customer differentiation, in contrast, single-item settings are predominantly considered, with the focus on multi-item settings being limited.

Detailed contribution points per research area are as follows:

- **We extend the VARI-METRIC model to a setting where throughput times may be reduced for certain items.** We opt not to model the repair shops by finite capacity (multi-server) queues, because repair capacities often are not fixed or may be fuzzy (repair shops may have other tasks than repair), with flexibility options possibly being available such as working overtime or temporarily hiring personnel. In our model, we may select different options for repair and transportation lead times at different prices, without explicitly modeling capacity. We encountered this situation at Thales Netherlands, which offers both a normal repair and a fast repair option to its customers at different prices. Emergency shipment options also exist that can be applied for certain combinations of items and locations against extra costs.

- **We consider models where lateral transshipments and emergency shipments are only used for a subset of all items and customers.** Lateral transshipments and emergency shipments are generally faster than waiting for items to arrive through regular supply. However, such shipment modes are also more expensive than regular shipments. Therefore, their use will often be limited to high priority customers (who have paid a higher contract price). The types of items will also influence the use of these shipment modes, with lateral and emergency shipments being most beneficial when used for expensive slow moving items. For inexpensive fast movers, in contrast, such shipment modes will be prohibitively expensive. So far, no models have been considered where shipment mode differentiation is possible on both an item and a customer level.
- **We consider a model where dedicated customer stocks can be kept, both under full backordering and under lost sales.** The added value of keeping dedicated stocks has not been investigated in literature before.
- **We develop a new method for analyzing two-echelon models with lost sales.** The option of keeping dedicated stocks in addition to stock at a central location results in an additional echelon level in the supply chain. Under lost sales, no appropriate multi-echelon models yet exist for analyzing such a system.
- **We develop a model for estimating response time distributions when priority mechanisms are used to assign service engineers to customers.** Using these response time distributions, we can both determine the required number of service engineers in a service region for a given set of service level agreements and, conversely, determine what service levels are achievable given the current number of engineers. The situation at Océ Technologies serves as a basis for model development.

1.8.2 Detailed research objectives

Our main research objective translates into 7 detailed objectives. Research objective 1 pertains to differentiation in spare parts supply on an item level through selective throughput time reduction.

1. *To determine whether and when throughput time reduction can lead to large cost savings in general multi-echelon multi-indenture spare parts networks (chapter 2).*

To meet this research objective, we require (i) a procedure to accurately determine the system's performance measures under this control option and (ii) a procedure to set decision variable values (i.e. the stock level and throughput time values in the system) such that the total system costs are minimized subject to service level restrictions. To estimate the added value of throughput time reduction, we compare it to a benchmark model where the system stock levels are the only decision variables (i.e. throughput times are fixed).

Research objectives 2 to 6 pertain to control options for applying differentiation in spare parts supply on both an item level and a customer level.

2. *To determine whether and when the selective use of emergency shipments is an effective control option for applying service level differentiation in spare parts supply (chapter 3).*

3. *To determine whether and when the selective use of lateral transshipments is effective for applying service level differentiation in spare parts supply (chapter 4).*

4. *To find an accurate and fast approach for analyzing a two-echelon model with lost sales (chapter 5).*

5. *To determine whether and when the use of dedicated customer stocks is an effective control option for applying service level differentiation in spare parts supply (chapter 6).*

6. *To investigate the added value of using multiple control options simultaneously for differentiation in spare parts supply (chapters 3, 4 and 6).*

The sub-questions (i) and (ii) for research question 1 also apply to research questions 2, 3, 5 and 6. As indicated in Section 1.8.1, we have developed a new method for analyzing a two-echelon model with lost sales. This model, which we require for analyzing a system under dedicated stocks, is discussed separately in chapter 5. Notice that we also look at the added value of using multiple control options simultaneously for differentiation (research question 6). To investigate the added value of the (combinations of) control options, we compare them to two benchmark models, namely one-size-fits-all policies where no differentiation is used, and critical level policies.

Finally, we also investigate the use of priority mechanisms for applying differentiation in assigning service engineers to customers.

7. *To determine the impact on service level performance of using priority mechanisms for assigning service engineers to customers (chapter 7).*

To meet this objective, we must first be able to determine system performance per customer class when priority mechanisms are used for assigning engineers to customers. Subsequently, we use these performance measures to verify whether all service level agreements have been met. As we will show, it is often not sufficient to only determine the mean response time per customer class, as service levels can also pertain to the overall response time distribution.

In the remainder of the dissertation, we will show that there is indeed significant added value to optimizing throughput times in addition to system stock levels, both when throughput times are only used for differentiation on an item level (research objective 1) and when they are used for differentiation on both an item level and a customer class level (research objectives 2 and 3). Furthermore, we show that (a combination of) control options for service differentiation (research objectives 2 to 6) are nearly as effective as critical level policies, even outperforming critical level policies under specific circumstances. Finally, we show that we can accurately estimate the distribution of service engineer response times when priority mechanisms are used for assigning service engineers to customers (research objective 7).

1.9 Techniques

To determine the added value of the control options, we must be able to evaluate performance measures – such as waiting times for spare parts and service engineers – when an option is used. Furthermore, we require techniques for system optimization under each control option (e.g. finding stock level values that minimize system costs under specific service level restrictions). We now elaborate on the mathematical techniques that we use for these purposes.

1.9.1 Techniques for evaluating system performance

In the research discussed in this dissertation, we commonly make two assumptions, namely (i) that failures occur according to Poisson processes and (ii) that replenishment/response times are exponentially distributed. As previously discussed, both assumptions are very common in after-sales service models. We also find the first assumption often to be valid in practice. The second assumption tends to be less valid in practice (e.g. part replenishment times tend to be close to deterministic). Nevertheless, this assumption is often not restrictive in spare parts models: for instance, earlier spare parts research has shown that system performance tends to be insensitive to the lead time distribution, particularly in lost sales models (see e.g. Alfredsson and Verrijdt (1999)).

Under these assumptions, system evaluation can occur through the use of *continuous-time Markov chains*. *Queuing models*, in particular, are frequently used for performance evaluation. For our purposes, system failures can be modeled as arrivals in a queuing system, with the number of available service engineers or the number of outstanding orders at a warehouse representing the servers (as done in Williams (1980) and Kranenburg and Van Houtum (2008) amongst others). As we assume that the arrivals to the queuing system occur according to a Poisson process, the PASTA property (i.e. Poisson Arrivals See Time Averages) allows us to obtain performance measures such as mean waiting times from the steady-state distribution of, for instance, the inventory levels at various warehouses. In spare parts optimization models, frequently used queuing models are the $M|G|c|c$ queue for lost sales models (i.e. the Erlang loss system) and the $M|G|\infty$ queue for backordering models, see e.g. Karush (1957) and Graves (1985). Both queuing models are insensitive to the service time distribution.

We note that an exact evaluation approach will not always be beneficial for the models we consider: for certain models, exact analysis models will require a lot of computation time, in particular in multi-item settings. Therefore, we also consider approximations that lead to fast and accurate results. Such approximations are quite common, such as the METRIC approximation (Sherbrooke, 2004) or the two-moment approximation (Graves, 1985) to approximate the distribution of the number of outstanding orders at warehouses. Approximations also occur in settings where an exact evaluation approach requires large multi-dimensional Markov chains, such as in lateral transshipment models with various warehouses (see Alfredsson and Verrijdt (1999) for an example). For such models, an iterative approach is used where each warehouse is analyzed separately given the rate at which transshipment requests arrive at that warehouse. This transshipment rate is updated over a number of iterations until convergence occurs (Axsäter, 1990).

1.9.2 Optimization techniques

For the control options with respect to spare parts supply (i.e. research objectives 1 through 6), we require techniques for finding optimal decision variable values. The aim is to minimize system operating costs under class-specific restrictions on the aggregate mean waiting time over multiple items. As mentioned in Section 1.4.2, a system approach is required to ensure a certain availability level. As a result, aggregate waiting time restrictions apply (i.e. we are interested in the waiting time over all items instead of the waiting time per item).

The optimization problems we thus consider are multi-item problems with aggregate waiting time restrictions that are non-linear in the decision variables. Furthermore, the decision variables themselves (such as stock levels) are integer valued. Kranenburg (2006) considers similar optimization problems and notes that those problems are in fact complex knapsack problems for which an optimal polynomial time algorithm most likely does not exist. As our models can be viewed similarly, we expect that no polynomial time solution algorithms exist for our models either. As a result, optimal solution methods will likely be too time-consuming for problems of realistic size (i.e. with various items and/or stock locations).

Therefore, our focus is on heuristic methods, with the objective of finding near-optimal solutions within acceptable time. We consider three types of methods, i.e. greedy heuristics, local search, and Dantzig-Wolfe decomposition. As stated in Section 1.4.2, *greedy heuristics* iteratively select the option that results in the largest contribution according to some criterion (e.g. largest waiting time decrease per dollar additional investment) until some stopping criterion is satisfied. The approach is often considered, as it is simple and easy to implement. Furthermore, under strict convexity properties it may even give optimal solutions. However, its implementation is not always straightforward, especially in settings with multiple customer classes. The key issue then is how to value waiting time reductions for different customer classes. Specifically, it might be more beneficial to improve the service levels of high priority customers instead of those of lower priority customers, making it more difficult to determine which solution gives the largest contribution. Nevertheless, it does provide a feasible solution to the problem, which in turn can serve as input for alternative optimization techniques.

The second technique we consider is *local search*. Local search techniques start from an existing solution (such as the one given by a greedy heuristic) and iteratively try to find better alternatives. In each iteration, a set of solutions is constructed that closely resemble the current solution being considered (the so-called *neighborhood* of the current solution). In spare parts optimization models, a neighborhood may consist of solutions that keep one unit of additional stock for a specific item compared to the current solution. The best alternative is selected according to some criterion. This process is repeated until no further improvement occurs. We refer to Aarts and Lenstra (2003) for an overview.

Finally, we also consider a technique similar to *Dantzig-Wolfe decomposition* (Dantzig and Wolfe (1960), see also Gilmore and Gomory (1961)) which has often been applied to multi-item spare parts optimization problems with similar characteristics (e.g. Kranenburg and Van Houtum (2008), and Wong et al. (2007a)), leading to good results. With the technique, the original nonlinear problem is reformulated to a linear problem (a so-called Master problem), which can subsequently be solved using (integer) linear optimization techniques. Reformulation is accomplished by constructing a set of possible solutions for each item. In this dissertation, we refer to these possible solutions as *item policies*. In a standard spare parts problem, an item policy would indicate the amount of stock kept at the various locations in the system. The reformulated problem then becomes to select for each item exactly one policy from the set, such that costs are minimized, while the original restrictions are still met.

As an example, we present a simple single-location two-item problem (adapted from Kranenburg (2006)). For item i $(i = 1,2)$, m_i denotes the item demand rate. Furthermore, S_i denotes the stock level of item i, with $TC_i(S_i)$ and $EW_i(S_i)$ denoting the related item costs and mean waiting time respectively. The objective is to determine stock levels such that the total costs are minimized, while the aggregate waiting time may not exceed a threshold value W^{max}, i.e.

$$(P1) \quad \min_{S_1,S_2} TC_1(S_1) + TC_2(S_2)$$

$$\text{s.t.} \quad \frac{m_1}{m_1+m_2} EW_1(S_1) + \frac{m_2}{m_1+m_2} EW_2(S_2) \leq W^{max}$$

$$S_1, S_2 \in \mathbb{N}_0$$

This problem can be reformulated as follows: let b_{ir} denote an item policy for item i ($i = 1,2$), with r denoting the policy index. For each policy b_{ir}, $S_{b_{ir}}$ indicates the corresponding base stock level and the binary variable $x_{b_{ir}}$ denotes whether b_{ir} is selected for item i or not ($x_{b_{ir}}$ then equals 1 or 0 respectively). Let B_i denote the set of item policies for item i (so $b_{ir} \in B_i$, with $r = 1,2, \ldots, |B_i|$). We then obtain the following reformulated problem:

$$(P2) \quad \min_{x_{b_{ir}}, i=1,2, r=1,\ldots,|B_i|} \sum_{r=1}^{|B_1|} TC_1(S_{b_{1r}}) x_{b_{1r}} + \sum_{r=1}^{|B_2|} TC_2(S_{b_{2r}}) x_{b_{2r}}$$

$$\text{s.t.} \quad \sum_{r=1}^{|B_1|} \frac{m_1}{m_1+m_2} EW_1(S_{b_{1r}}) x_{b_{1r}} + \sum_{r=1}^{|B_2|} \frac{m_2}{m_1+m_2} EW_2(S_{b_{2r}}) x_{b_{2r}} \leq W^{max} \qquad (P2.1)$$

$$\sum_{r=1}^{|B_i|} x_{b_{ir}} = 1 \qquad i = 1,2 \qquad (P2.2)$$

$$x_{b_{ir}} \in \{0,1\} \qquad i = 1,2, r = 1, \ldots, |B_i|$$

In principle, B_i can consist of an infinite number of item policies. Still, a finite set of policies exists for each item i ($i = 1,2$) such that $(P2)$ and $(P1)$ are equivalent in the sense that both problems have the same optimal solution. Furthermore, if the integrality restriction on $x_{b_{ir}}$ is relaxed in $(P2)$, the solution to the resulting LP-relaxation constitutes a lower bound on the system costs. As a result, the quality of any solution to the original integer problem can be expressed in terms of a gap to this lower bound. Note that the solution to the LP-relaxation can also serve as a starting point for finding a near-optimal integer solution. A further benefit of this decomposition is that the system can be analyzed for each item policy separately, resulting in a distinction between system analysis and optimization. The analysis approach for an item policy does not matter, provided that the related costs and waiting times can somehow be obtained for optimization purposes. Note that an item policy can consist more decision variables than the system stock levels alone: any type of decision variable can be included in the item policy, such as the critical level per location that denotes the amount of stock reserved for high priority customers, or the shipment strategy used (e.g., backordering or emergency shipments) when a location is out of stock.

The main challenge in using decomposition is the selection of item policies for each item. As we just mentioned, the number of decision variables included in an item policy can be vast, resulting in an extensive set of policies to choose from, particularly for problems of realistic size. Then, it will not be viable to consider all policies, since computation times for solving the integer problem quickly increase as the number of policies increases. Furthermore, the system must be evaluated under each considered item policy to obtain the related performance measures. Such an evaluation also takes time, especially for complex systems, such as those where lateral transshipments are allowed among warehouses. We first focus on finding the set of policies such that the optimal solution to the LP-relaxation of $(P2)$ is the same as that of $(P1)$. Subsequently, we describe how to find a near-optimal *integer* solution to $(P1)$.

To find the policy set resulting in an optimal solution to the LP-relaxation of $(P2)$ and $(P1)$, we use *column generation,* which is a technique that is often used to solve problems with a large number of item policies (see e.g., Gilmore and Gomory (1961), Hans (2001), Lübbecke and Desrosiers (2005)). The idea behind column generation is as follows: (1) first, an initial set of policies is constructed for each item which leads to a feasible solution to the LP-relaxation of the reformulated problem when solved with the simplex method. Subsequently, (2) we iteratively add policies to the policy set that have not yet been considered, but could improve the solution value if added. We give further details on policy selection later on. We proceed to add such policies to the policy set until we can show that no further policies exist that would improve the solution value further.

The key to finding interesting policies to add, and to ensuring that we have included sufficient item policies in B_i, lies in duality theory. Below, let (P_{prim}) denote a linear optimization problem with $|J|$ variables and $|K|$ constraints. Furthermore, let (P_{dual}) denote the dual variant of (P_{prim}).

$$\begin{array}{llll}
(P_{prim}) & \min \sum_{j\in J} c_j\lambda_j & & \\
s.t. & \sum_{j\in J} a_{jk}\lambda_j \geq b_k & k \in K \\
 & \lambda_j \geq 0 & j \in J
\end{array}
\qquad\qquad
\begin{array}{llll}
(P_{dual}) & \max \sum_{k\in K} u_k b_k & & \\
s.t. & \sum_{k\in K} u_k a_{kj} \leq c_j & j \in J \\
 & u_k \geq 0 & k \in K
\end{array}$$

Let $\boldsymbol{\lambda}^* = [\lambda_1^*, \lambda_2^*, \ldots]$ denote a solution to (P_{prim}). Duality theory then tells us that $\boldsymbol{\lambda}^*$ is only optimal for (P_{prim}) if the corresponding solution to (P_{dual}) is feasible, see e.g., Winston (2003). Specifically, this means that $c_j - \sum_{k\in K} u_k a_{kj} \geq 0$ for all variables $j \in J$. Conversely, we can thus prove that we have obtained the optimal solution to (P_{prim}) by showing that no variable j exists for which $c_j - \sum_{k\in K} u_k a_{kj} < 0$.

Note that the expression $c_j - \sum_{k\in K} u_k a_{kj}$ in fact denotes the *reduced costs* related to a variable λ_j. Hence, step (2) of column generation focuses on finding at least 1 variable that has negative reduced costs (in which case the solution quality can be improved by including that variable in the problem), or proving that no such policy exists. To compute the reduced costs for any variable, we require values for the dual variables u_k $(k \in K)$. Note that the value of u_k

corresponds to the *shadow price* for constraint k when we solve (P_{prim}). Given that the shadow price denotes the amount by which the solution value increases when the RHS of the constraint increases by 1 unit, the value for u_k will at least be zero in the previous example, as each increase of b_k results in a smaller solution space and thus possibly higher costs. Note that the shadow price will be exactly zero if the restriction is nonbinding.

Applying these general results to problem $(P2)$, we find the following expression for the reduced costs:

$$RED(b_{ir}) = TC_i(S_{b_{ir}}) - \frac{u_1 m_i}{m_1 + m_2} EW_i(S_{b_{ir}}) - u_{i+1} \quad i = 1,2,$$

where $u_1 \leq 0$ denotes the shadow price related to constraint $(P2.1)$ and $u_{2,3} \geq 0$ denotes the shadow prices related to constraints $(P2.2)$. Now, u_1 is non-positive, since an increase of W^{max} can only improve the solution quality (corresponding to the lower solution value). Conversely, u_2 and u_3 are nonnegative: for those constraints, an increase of the RHS values implies that more than 1 policy must be selected per item. Such an adjustment cannot lead to lower costs. Note that $RED(b_{ir})$ does not depend on the costs and waiting times for items $j \neq i$. Hence, we can apply column generation for each item separately. As a result, the overall problem can be decomposed into single-item problems.

Overall, we use the following column generation procedure for the LP relaxation considered in this dissertation:

1. Start with an initial set of item policies that results in a feasible solution to our optimization problem.
2. Using the current set of item policies, solve the restricted LP relaxation of the optimization problem. Derive the shadow price values for each constraint.
3. Use the shadow price values to find per item an item policy that has not been included in the restricted LP problem yet and that has negative (and possibly minimum) reduced costs. If such a policy is found, it is added to the policy set.
4. If we can prove that no policy with negative reduced costs exists for any item, we have found the optimal solution to the LP relaxation. Otherwise, we proceed to step 2.

With the above column generation procedure, we are able to find an optimal solution to the LP-relaxation of problem $(P1)$. Note that this solution constitutes a *lower bound* to the original integer problem. In general, this LP-relaxation solution will not be integer: a linear combination of policies might be selected for a subset of items. We therefore also require approaches to obtain a near-optimal integer solution. We consider two such approaches in this dissertation, both of which use the LP-relaxation solution – and the constructed item policies – as a starting point. The first approach uses the LP-relaxation solution to obtain an initial solution for the local search procedure described earlier in this section. For instance, the LP-relaxation solution can first be rounded to an infeasible solution with low costs, with local search used to obtain a feasible integer solution at minimal additional costs. The second approach starts with the item policies constructed during column generation and uses a solver such as CPLEX to solve the integer variant of problem $(P2)$. As computation times can be extensive under integer optimization, we apply various strategies to keep computation times reasonable, such as removing poor item policies from the problem before optimization, or limiting the time for the solver to optimize the problem.

1.10 Outline

The outline of the dissertation closely follows the research objectives stated in Section 1.8.2. In Chapter 2 (research objective 1), we consider differentiation at an item level by investigating the cost savings that are possible by selectively applying throughput time reduction in multi-echelon multi-indenture spare parts networks. In Chapters 3 through 6, we consider differentiation at both an item and a customer level by considering various (combinations of) control options for applying differentiation in spare parts supply. In Chapter 3 (research objective 2), we focus on the selective use of emergency shipments in a setting with a single warehouse. In Chapter 4 (research objective 3), we subsequently extend this selective emergency shipment model to a multi-warehouse setting where lateral transshipments may be used selectively to satisfy premium customer requests. In Chapter 5 (research objective 4), we present an approach for analyzing a two-echelon model with emergency shipments. The analysis of this model serves as a building block in the multi-item optimization model considered in Chapter 6, where we allow dedicated customer stocks to be kept for meeting differentiated service requirements (research objective 5). Note that research objective 6 – on investigating the added value of combining various control options – is discussed in Chapters 3, 4 and 6. After treating various control options pertaining to spare parts supply, we consider differentiation when assigning service engineers to customers in Chapter 7 (research objective 7). The final chapter in this dissertation is Chapter 8, where we draw our main conclusions and discuss possibilities for further research.

Table 1.1 summarizes the control option(s) considered in each chapter and the levels at which differentiation occurs. In chapters that pertain to multiple control options, the primary option discussed in that chapter is marked by an asterisk. We do not include Chapter 5 in the overview as it serves as a building block for the model in Chapter 6.

1.10. Outline

Chapter	Control options	Differentiation at an item level	Differentiation at a customer level
2	Throughput time reduction options	x	
3	Selective emergency shipments *, critical level policies	x	X
4	Selective emergency shipments, selective lateral transshipments *, critical level policies	x	X
6	Dedicated stocks *, critical level policies	x	X
7	Priority assignment mechanisms		X

Table 1.1 Overview per chapter of the control options considered and the levels at which differentiation occurs.

Chapter 2

Throughput time reduction[1]

2.1 Introduction

In this chapter, we focus on differentiation at an item level in spare parts supply. Specifically, we consider a model where item throughput times for repair and transportation can be reduced at additional costs. As a result, these throughput times also become decision variables in an optimization procedure. Our research is based on a setting we encountered at Thales Netherlands, a supplier of naval radar and combat management systems. For system upkeep during the life cycle, Thales offers service contracts to its customers that contain quantified service levels, such as a maximum response time in case of a failure. To meet these service levels, Thales places initial inventory in the network, with inventory levels optimized using a tool based on VARI-METRIC (Sherbrooke, 2004). If there is evidence during contract execution that the actual service levels are below target (e.g. in terms of downtime waiting for spare parts), Thales has options for intervening at a tactical level, amongst others by (i) buying additional spare parts, (ii) reducing repair shop throughput times, and (iii) reducing transportation times of spare parts. We focus on throughput time (TPT) reduction (of repair and transportation) as an alternative to an investment in spare parts. Earlier literature (Sleptchenko et al., 2005; Adan et al., 2009) has shown that influencing repair TPT for specific items may have a large impact on the total costs.

Inspired by the Thales case, we aim for a realistic model, i.e. a multi-item, multi-indenture, multi-echelon setting. This is in contrast to single-item models found in the literature that can only be used as a building block. As mentioned in the introduction (Section 1.4.2), the product indenture levels indicate the levels at which repair can take place, with modules at the highest indenture level denoted by Line Replaceable Units (LRUs), which in turn can be repaired by replacing subcomponents (so-called Shop Replaceable Units (SRUs)) or parts. Figure 2.1 depicts a multi-indenture structure for the setting at Thales. In the remainder of this chapter, we will use the phrases *parent* and *child* to refer to the relations in the multi-indenture structure: In Figure 2.1, the supply cabinet is the parent of the power supply, and the power supply and air conditioning assembly are children of the supply cabinet. Furthermore, we use the term *item* for

[1] This chapter is based on the paper “Inventory reduction in spare part networks by selective throughput time reduction” by M.C. van der Heijden, E.M. Alvarez, and J.M.J. Schutten, International Journal of Production Economics, 143(2), pp. 509–517.

components at any level in the multi-indenture structure (LRUs, SRUs, parts). Figure 2.2 denotes the multi-echelon structure for Thales. Spare parts may be stocked on board of a naval ship, at the shore organization (close to a harbor), or at Thales Netherlands. We will use the common term *base* for a site where one or more systems are operational and the phrases *supplier* and *customer* for the relations in the multi-echelon structure. In Figure 2.2, Thales is the supplier of the shore site, and the shore site is a customer of Thales. Ready-for-use items are moved from the *upstream* part of the supply chain (Thales) to the *downstream* part (ships).

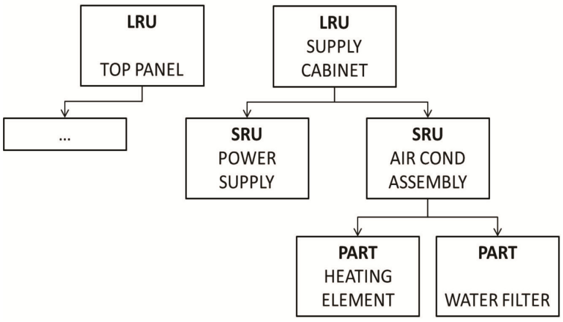


Figure 2.1 A multi-indenture structure.

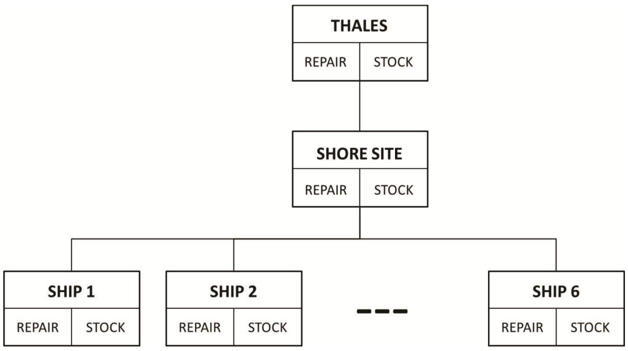


Figure 2.2 A multi-echelon structure.

Our model allows us to both set spare parts inventory levels *and* select different options for repair and transportation lead times at different prices, without explicitly modeling capacity. We encountered this situation at Thales, which offers both a normal repair and a fast repair option at different prices to its customers that do not have service contracts. The same flexibility could be used to optimize the performance for customers having service contracts. This also holds for emergency transportation that Thales may apply for certain combinations of items and locations against additional costs. To gain insight in the impact of TPT reductions, we first develop expressions for the marginal backorder reduction of LRUs at operating sites as a function of the marginal reduction in TPT of each repair and transport in the network. We use these expected number of backorders as a criterion, because their minimization is approximately equivalent to maximizing operational availability (Sherbrooke, 2004). Under the approximation that the pipelines are Poisson distributed, we only need the fill rates of all items in the multi-indenture structure at all locations in the multi-echelon network to compute the marginal backorder reduction as a function of the marginal reduction in repair and transport TPTs. Combining these marginal values with a certain discrete step size for the TPT reductions, we develop a heuristic optimization method to balance the investment in TPT reductions to investment in extra spares. In summary, our contributions to the literature are the following:

1. We consider a simple but practical model for the trade-off between spare part stocks and TPT reduction in repair and transportation, based on pricing of TPT reduction. This model is

suitable for multi-item, multi-echelon, multi-indenture networks as we encountered at Thales Netherlands.

2. We use estimates for the marginal impact of TPT reductions to develop an efficient heuristic method for the simultaneous optimization of spare part inventories and repair and transportation TPTs. We show that significant cost reductions are feasible.
3. We show how the savings depend on the type of problem instance and we characterize the type of policies that we typically find. In particular, we observe that TPT reductions are most profitable *downstream* in the network.
4. We apply our method in a case study at Thales Netherlands and find interesting savings (5.6% on the inventory investment). The restricted options for reduction of TPTs downstream in the network cause lower savings than in the theoretical experiments.

We first define our model in Section 2.2. Section 2.3 shows how we estimate the impact of TPT reduction for given spare part stock levels. This is input for our optimization heuristic (Section 2.4). In Section 2.5, we discuss numerical results for both the case study at Thales Netherlands and a large set of theoretical problem instances. We end up with conclusions in Section 2.6.

2.2 Model, assumptions, and notation

We consider a multi-indenture, multi-echelon spare part network. Our decision variables are spare part inventory levels, and repair and transportation TPTs of all items at all locations in the network. For each combination of item and location, we have a discrete set of TPTs, and costs are attached to each option.

2.2.1 Assumptions

We proceed from the standard assumptions of the VARI-METRIC model (Sherbrooke, 2004):

1. Systems fail according to a stationary Poisson process.
2. All failures are critical, i.e. they cause system downtime.
3. Each item failure is caused by the failure of at most one subcomponent.
4. Repair shops are modeled as $M|G|\infty$ queues, where successive repair TPTs of the same item at the same location are independent and identically distributed.
5. The flow of repair jobs of each item arriving at each location is given. This is modeled as a given fraction of jobs that can be repaired (the rest is forwarded for repair upstream).
6. All items are as good as new after repair.
7. Requests for spare parts are handled First Come, First Serve (FCFS).
8. We use an $(s-1, s)$ replenishment policy for all items at all locations.
9. Any customer stock location has one unique supplier (except the most upstream stock point).

10. Inventories are always replenished from the direct supplier in the multi-echelon structure, i.e. there is no lateral supply between locations at the same echelon.
11. All transportation TPT (or: *order-and-ship* times) are deterministic.
12. There is no commonality among items: the various LRUs do not have any SRU in common, the SRUs do not have any part in common, and so forth.

With respect to TPTs (repair and transportation), we further assume:

13. For each combination of item and location, we have a *discrete set* of TPTs, and costs are attached to each option. This corresponds to the practice at Thales Netherlands, where a limited set of options were available for both repair and transportation TPT (see the case description in Section 2.5.3).

With respect to the latter assumption, we proceed from a standard repair and transportation lead time for each combination of item and location, and we consider options for TPT reductions that we may select at additional costs. Without loss of generality, the additional costs per repair are increasing in the repair TPT reduction, and the same applies to the transportation costs. If not, we ignore inferior (i.e. non-dominant) options.

2.2.2 Notation

We use similar notations as in Sherbrooke (2004) and distinguish input parameters, decision variables, auxiliary variables, and performance measures (output):

Input:

J = set of all locations in the network.

B = set of all bases, i.e. all locations in the network where systems are installed. Note that $B \subset J$.

I = set of all items.

L = set of all LRUs, i.e. all first indenture items, with $L \subset I$.

m_{ij} = demand rate for item i at location j ($i \in I, j \in J$).

r_{ij} = fraction of demand for item i at location j that can be repaired at the same location (the rest is forwarded to the supplier of j for repair).

q_{ki} = fraction of item k failures that is due to a failure of item i.

h_i = costs per year for holding one item i, these costs may include costs of capital, storage and risk, including the obsolescence risk.

$T_{ij}(n)$ = n^{th} option for the repair shop TPT of item i at location j, which is strictly decreasing in n; index $n = 0$ gives the standard repair throughput time.

$O_{ij}(n)$ = n^{th} option for the transportation TPT of item i to location j, which is strictly decreasing in n; index $n = 0$ gives the standard transportation time.

$C_{ij}^{R}(t)$ = costs per repair if the repair shop TPT of item i at location j equals t.

$C_{ij}^{O}(t)$ = costs to move a single item i to location j from its supplier if the transportation TPT equals t.

Note that the demand rates m_{ij} are input for all LRUs $i \in L$ and all bases $j \in B$. We can recursively find the demand rates for all other combinations of item i and location j from $m_{ij} = m_{kj}q_{ki} + \sum_{l \in D_j} m_{il}(1 - r_{il})$, where k is the parent of i and D_j denotes the set of all customers of location j.

Decision variables:

s_{ij} = inventory level for item i at location j.

a_{ij} = index of repair TPT of item i at location j.

b_{ij} = index of transportation TPT of item i to location j.

We denote the matrices of decision variables for all items and all locations in bold face by $\boldsymbol{s}$, $\boldsymbol{a}$ and $\boldsymbol{b}$.

Auxiliary variables:

$f_{ij}(n)$ = probability that the number of items i in the pipeline to location j, i.e. all items in repair or in resupply, equals n; we denote the corresponding mean by μ_{ij}.

Performance measures:

$EBO_{ij}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{b})$ = Expected backorders of item i at location j under policy $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{b})$.

$\beta_{ij}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{b})$ = Fill rate of item i at location j under policy $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{b})$, i.e. the fraction of demand that can be filled from stock on shelf without delay.

2.2.3 Model

As in VARI-METRIC, we aim to balance the operational availability and the costs required for holding spare part inventories, and, in our case, the costs of repair and transportation. As mentioned before, Sherbrooke (2004) uses the sum of the backorders of LRUs at bases (sites where systems are installed) as a proxy for the operational availability. Following this approach, we find the following nonlinear optimization model:

$$(P1) \qquad \min_{\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}} \sum_{i \in I} \sum_{j \in J} \left(h_i s_{ij} + m_{ij} r_{ij} C_{ij}^{R}\left(T_{ij}(a_{ij})\right) + m_{ij}(1 - r_{ij}) C_{ij}^{O}\left(O_{ij}(b_{ij})\right) \right)$$

s.t. $$\sum_{i\in L, j\in B} EBO_{ij}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) \leq EBO^{Target}$$

where EBO^{Target} denotes a target number of LRU backorders at bases corresponding to a certain operational availability. We can interpret this target backorder sum as a maximum on the average number of systems that are down waiting for a spare part. We can set this target as $EBO^{Target} = (1 - availability) * IB$, where IB denotes the total number of systems in the installed base. The expected number of backorders of item i at location j depends on the probability distribution of the number of items in the pipeline $f_{ij}(n)$ and the stock level s_{ij}:

$$EBO_{ij}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) = \sum_{n=s_{ij}+1}^{\infty} \left(n - s_{ij}\right) f_{ij}(n|\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) \qquad (2.1)$$

where the probability of the pipeline $f_{ij}(n|\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$ depends on the repair and transportation TPT of item i at location j, and the probability distribution of the number of backorders (a) of item i at the supplier of location j, and (b) of all children of item i at location j.

Indirectly, the number of backorders of item i at location j depends on all stock levels of item i and all children downwards in the multi-indenture structure, at location j and all locations upstream in the supply chain. The same applies to the repair shop TPTs and transportation TPTs (and hence for the impact of the decision variables a_{ij} and b_{ij}). In METRIC, all pipeline distributions were originally approximated by Poisson distributions. Because this approximation can be quite bad, two-moment approximations for the pipelines have been used in VARI-METRIC (Sherbrooke, 2004). This can be done using negative binominal distributions, because the variance-to-mean ratio of the pipelines are usually ≥1. As a more general solution, we use the method of Adan et al. (1995) to fit a discrete probability distribution function to the first two moments. Hence, we compute an approximation of all backorders using two-moment approximations for the pipeline distributions. VARI-METRIC only considers the stock levels and not the TPT reduction. For optimization, a simple greedy heuristic is typically applied. That is, starting at all stock levels $s_{ij} = 0 \; \forall i,j$, we add in each iteration one item of type i to the stock at a certain location j such that the largest ratio of reduction of the expected number of backorders of LRUs at bases to the additional inventory investment is incurred. In popular terms, this heuristic is referred to as the *biggest bang for the buck* approach.

Problem $(P1)$ is a large nonlinear integer programming problem having three times as many decision variables as VARI-METRIC: Next to the stock levels s_{ij}, we also have to determine the repair and transportation TPTs for all combinations of item i and location j. As VARI-METRIC is an optimization heuristic, it is reasonable to expect that $(P1)$ cannot be solved exactly in a reasonable amount of time for problem instances with a realistic size. So, we focus on optimization heuristics.

2.3 Analysis of TPT reduction

We first specify the impact of TPT reduction on the expected number of backorders of LRUs at the bases. To this end, we initially assume that the pipelines to each location are Poisson distributed (as opposed to the two-moment approximation used in VARI-METRIC). We realize that this approximation can lead to rather poor estimations of the marginal impact of reducing a particular TPT. Still, we believe that this is acceptable, since we only use the *ranking* of the TPTs to decide which TPT reductions are most attractive. Once we have selected the most attractive TPT reduction, we use VARI-METRIC to evaluate the exact impact of this reduction on the system availability, see Section 2.4. Our implicit assumption is that the *ranking* of TPT reductions will not depend on the distribution used to characterize the pipeline. Furthermore, under Poisson distributed pipelines we find simple expressions for the marginal impact of TPT reduction on the reduction of the mean backorder levels as a function of the fill rates in the system.

We now find the partial derivatives of the total expected backorders of LRUs at bases to any mean repair TPT and order-and ship time in the following way. In finding these partial derivatives, we assume that the repair and transport TPTs are continuous variables. During optimization, however, we in fact reduce all TPTs using stepwise functions. Given that the pipelines have a Poisson distribution, we first note that the probability function of the pipeline $f_{ij}(n|\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$ is given by:

$$f_{ij}(n|\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) = \frac{\mu_{ij}^{n} e^{-\mu_{ij}}}{n!} \tag{2.2}$$

where the mean pipeline μ_{ij} depends on the decision variables $(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$. From here on, we will use the shorthand notation $(.)$ if a variable is a function of (some of) the decision variables $(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$. Using elementary calculus, we can derive from (2.1) and (2.2) that

$$\frac{\partial EBO_{ij}(.)}{\partial \mu_{ij}} = \sum_{n=s_{ij}}^{\infty} \frac{\mu_{ij}^{n} e^{-\mu_{ij}}}{n!} \tag{2.3}$$

which equals $1-\beta_{ij}(.)$, so one minus the fill rate. For a single site model, we have that $\mu_{ij} = m_{ij}T_{ij}$, and so we find using the chain rule for differentiation:

$$\frac{\partial EBO_{ij}(.)}{\partial T_{ij}} = \frac{\partial EBO_{ij}(.)}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial T_{ij}} = \left(1-\beta_{ij}(.)\right) m_{ij} \tag{2.4}$$

In a *two-echelon, single-indenture* model with location 0 as the supplier of location j, we have (Sherbrooke, 2004):

$$\mu_{ij} = m_{ij}\left\{r_{ij}T_{ij} + \left(1-r_{ij}\right)\left(O_{ij} + EBO_{i0}(.)/m_{i0}\right)\right\} \tag{2.5}$$

2.3. Analysis of TPT reduction

The chain rule gives us the derivatives to the mean repair TPT and the transportation TPT at location j

$$\frac{\partial EBO_{ij}(.)}{\partial T_{ij}} = \frac{\partial EBO_{ij}(.)}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial T_{ij}} = \left(1 - \beta_{ij}(.)\right) m_{ij} r_{ij} \tag{2.6}$$

$$\frac{\partial EBO_{ij}(.)}{\partial O_{ij}} = \left(1 - \beta_{ij}(.)\right) m_{ij}\left(1 - r_{ij}\right) \tag{2.7}$$

and for the derivative to the mean repair TPT at location 0 we find:

$$\frac{\partial EBO_{ij}(.)}{\partial T_{i0}} = \frac{\partial EBO_{ij}(.)}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial EBO_{i0}(.)} \frac{\partial EBO_{i0}(.)}{\partial \mu_{i0}} \frac{\partial \mu_{i0}}{\partial T_{i0}(.)} = \left(1 - \beta_{ij}(.)\right)\left(1 - \beta_{i0}(.)\right) m_{ij}\left(1 - r_{ij}\right) \tag{2.8}$$

Similarly, we find the partial derivatives of the expected LRU backorders at the bases to all mean repair TPTs and transportation TPT in multi-echelon, multi-indenture networks. To show how, we use $P_{ij,kl}$ for the partial derivative of EBO_{ij} to the mean repair TPT T_{kl}, where

- item k belongs to the multi-indenture structure of item i (i.e. a child of i or a lower indenture item), and
- location l is a location upstream of location j (i.e. the supplier of j, or even more upstream in the multi-echelon structure).

Equivalently, $Q_{ij,kl}$ denotes the partial derivative of EBO_{ij} to the transportation TPT O_{kl}. Then we can recursively compute all partial derivatives under the assumption of Poisson distributed pipelines. Figure 2.3 shows how we compute the partial derivatives of the expected backorders of LRU 0 at base j to the repair TPT of SKU i (child of LRU 0) at location 0 (supplier of j). We obtain this scheme by using the formulas as given in Sherbrooke (2004) (Section 5.5 – 5.7). The equation for $P_{i0,i0}$ (the lower right corner of the figure) has already been explained in formula (2.6). The equation for $P_{ij,i0}$ (the upper right corner of the figure) follows from formula (2.8):

$$P_{ij,i0} = \frac{\partial EBO_{ij}(.)}{\partial T_{i0}} = \frac{\partial EBO_{ij}(.)}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial EBO_{i0}(.)} \frac{\partial EBO_{i0}(.)}{\partial T_{i0}(.)} = \left(1 - \beta_{ij}(.)\right) \frac{m_{ij}\left(1 - r_{ij}\right)}{m_{i0}} P_{i0,i0} \tag{2.9}$$

We can derive the equation for $P_{00,i0}$ (the lower left corner of the figure) analogously:

$$P_{00,i0} = \frac{\partial EBO_{00}(.)}{\partial T_{i0}} = \frac{\partial EBO_{00}(.)}{\partial \mu_{00}} \frac{\partial \mu_{00}}{\partial EBO_{i0}(.)} \frac{\partial EBO_{i0}(.)}{\partial T_{i0}(.)}$$

Using formulas (5.15) and (5.16) in Sherbrooke[2], we find:

$$P_{00,i0} = \left(1 - \beta_{00}(.)\right) \frac{m_{00} q_{0i}}{m_{i0}} P_{i0,i0} \tag{2.10}$$

Finally, we find $P_{0j,i0}$ (the upper left corner of the figure) using formula (5.22) in Sherbrooke:

$$P_{0j,i0} = \frac{\partial EBO_{0j}(.)}{\partial T_{i0}} = \frac{\partial EBO_{0j}(.)}{\partial \mu_{0j}} \left\{ \frac{\partial \mu_{0j}}{\partial EBO_{00}(.)} \frac{\partial EBO_{00}(.)}{\partial T_{i0}(.)} + \frac{\partial \mu_{0j}}{\partial EBO_{ij}(.)} \frac{\partial EBO_{ij}(.)}{\partial T_{i0}(.)} \right\}$$

Using (5.21) in Sherbrooke, this equation can be written as

$$P_{0j,i0} = \left(1 - \beta_{0j}(.)\right) \left\{ \frac{m_{0j}\left(1 - r_{0j}\right)}{m_{00}} P_{00,i0} + 1 \cdot P_{ij,i0} \right\} \tag{2.11}$$

It is straightforward to modify this scheme for the transportation TPT.

$$\begin{tikzcd}
 & \text{LRU } 0 & \text{SRU } i \\
\text{Base } j & \begin{array}{l} P_{0j,i0} = (1-\beta_{0j})P_{ij,i0} \\ +(1-\beta_{0j})\dfrac{m_{0j}(1-r_{0j})}{m_{00}} P_{00,i0} \end{array} & P_{ij,i0} = \dfrac{(1-r_{ij})m_{ij}}{m_{i0}}(1-\beta_{ij})P_{i0,i0} \arrow[l, Rightarrow] \\
\text{Depot } 0 & P_{00,i0} = \dfrac{q_{0i}m_{00}}{m_{i0}}(1-\beta_{00})P_{i0,i0} \arrow[u, Rightarrow] & P_{i0,i0} = r_{i0}m_{i0}(1-\beta_{i0}) \arrow[l, Rightarrow] \arrow[u, Rightarrow]
\end{tikzcd}$$

Figure 2.3 Computation scheme for the partial derivatives of LRU backorders at bases.

We observe that we only need the fill rates to estimate the impact of TPT reduction of all items at all location under the assumption of Poisson distributed pipelines, which is straightforward and fast to compute. Our approach is exact for multi-indenture, multi-echelon networks under Poisson distributed pipelines. However, it is known that the true pipeline distributions may clearly differ from Poisson distributions. We have also observed this, particularly if we need probabilities from the tail of the pipeline distributions. Unfortunately, we could not find reasonable expressions for the partial derivatives under two-moment approximations for the pipelines. In the next section, however, we will see that we do not use the exact values of the partial derivatives, but only use their ranking to select the most promising option (repair or shipment) for TPT reduction.

[2] We note that the indices q_{0i} have a different definition in Sherbrooke than in our paper.

2.4 Optimization heuristic

At first sight, we can easily extend the greedy heuristic for spare part optimization by adding extra options for TPT reduction. We estimate the impact of repair TPT reduction of item i at location j on the total LRU backorders using the partial derivatives as found in the previous section: $\{T_{ij}(a_{ij}) - T_{ij}(a_{ij}+1)\}\sum_{k\in L}\sum_{l\in B} P_{kl,ij}$. This is obviously an approximation, but it gives us a good idea on the impact of TPT reductions. We compare this impact to the additional costs, being the additional repair costs times the number of repairs per year: $\left\{C_{ij}^{R}\left(T_{ij}(a_{ij}+1)\right) - C_{ij}^{R}\left(T_{ij}(a_{ij})\right)\right\} m_{ij} r_{ij}$. So, we have the following simple approximation for backorder reduction per euro $\Delta_R(a_{ij})$ due to repair TPT reduction of item i at location j:

$$\Delta_R(a_{ij}) = \frac{T_{ij}(a_{ij}) - T_{ij}(a_{ij}+1)}{\left\{C_{ij}^{R}\left(T_{ij}(a_{ij}+1)\right) - C_{ij}^{R}\left(T_{ij}(a_{ij})\right)\right\} m_{ij} r_{ij}} \sum_{k\in L}\sum_{l\in B} P_{kl,ij} \tag{2.12}$$

The backorder reduction per euro due to transportation TPT reduction $\Delta_O(b_{ij})$ equals

$$\Delta_O(b_{ij}) = \frac{O_{ij}(b_{ij}) - O_{ij}(b_{ij}+1)}{\left\{C_{ij}^{O}\left(O_{ij}(b_{ij}+1)\right) - C_{ij}^{O}\left(O_{ij}(b_{ij})\right)\right\} m_{ij} (1-r_{ij})} \sum_{k\in L}\sum_{l\in B} Q_{kl,ij} \tag{2.13}$$

We denote the standard backorder reduction per euro from VARI-METRIC, due to adding a spare part i at location j to stock, by $\Delta_S(s_{ij})$. Now a logical extension of the greedy VARI-METRIC heuristic is to add all options for TPT reduction, and to select at each iteration the decision that yields the highest backorder reduction per euro spent. This can be either adding a spare part to stock, or a discrete step reduction in either repair or transportation TPT. Unfortunately, this heuristic does not work well, since TPTs and stock levels are not independent: If we add stocks, the impact of TPT reduction decreases. We typically see that we initially decide to reduce many TPTs, because there are hardly any spare part stocks and so the impact of TPT reduction is high. If spare part stock levels are zero, any hour reduction of TPT is an hour reduction in system down time. When we have added spare parts to stock, we find out that the impact of these TPT reductions decreases, and finally we may even end up with a solution that is worse than the one VARI-METRIC provides while ignoring the options for TPT reduction. So, we have to find another heuristic.

As the problems above are caused by the generally decreasing impact of TPT reduction on the spare part inventories, it seems better to construct a heuristic that considers TPT reduction while stock levels are *de*creasing rather than increasing. The basic idea is the following. First, we apply VARI-METRIC using the standard TPTs $T_{ij}(0)$ and $O_{ij}(0)$. Then, we improve this solution by replacing the spare stock levels having the least added value with TPT reductions having the most added value. The spare part having least added value is the last one we added to stock in the VARI-METRIC algorithm. We search the best (set of) TPT reduction(s) compensating the loss of availability by removing the latter spare parts. If these TPT reductions cost less per year than the holding cost of the removed spare part, we accept the stock level reduction. We continue until no improvement is found. So, our basic algorithm is as follows:

Basic optimization heuristic

1) Initialize the decision variables: $s_{ij} = 0$, $a_{ij} = 0$, $b_{ij} = 0$ $(i \in I, j \in J)$.

2) Use VARI-METRIC to optimize the spare part stock levels for the TPTs $T_{ij}(a_{ij})$ and $O_{ij}(b_{ij})$ $(i \in I, j \in J)$. Keep track of the order in which spare parts are added to stock (item i, location j). Let us denote that list by (i_n, j_n), being the type of item i_n and the location j_n that has been added to stock in iteration n $(n = 1 \dots N)$. Compute partial derivatives $P_{ij,kl}$ and $Q_{ij,kl}$.

3) Consider compensating stock reduction of spare part (i_N, j_N) by TPT reduction. The cost savings per year are h_{i_N}. Set the additional costs for TPT reduction equal to $C^{TR} = 0$. Set $i^* = i_N$ and $j^* = j_N$.

   a. Recalculate the expected backorders and the partial derivatives that have changed (that is, for all combinations of (i) items in the same branch of the multi-indenture structure as i^* (parents and children), and (ii) locations in the same branch of the multi-echelon structure as j^* (customers and suppliers). If the sum of expected LRU backorders at bases is greater than or equal to the target EBO^{Target}, then go to Step 3b, else go to 3c[3].

   b. Select the best TPT reduction from the options a_{ij}, b_{ij} by selecting (i^*, j^*) from

$$(i^*, j^*) = \arg\min_{(i,j)} \left\{ \min\left(\Delta_R(a_{ij}), \Delta_O(b_{ij}) \right) \right\} \tag{2.14}$$

   If the minimum is attained for a repair TPT reduction, then set

$$C^{TR} \coloneqq C^{TR} + \left\{ C^R_{i^*j^*}\left(T_{i^*j^*}(a_{i^*j^*} + 1)\right) - C^R_{i^*j^*}\left(T_{i^*j^*}(a_{i^*j^*})\right) \right\} m_{i^*j^*} r_{i^*j^*},$$

   and $a_{i^*j^*} \coloneqq a_{i^*j^*} + 1$

   else set $C^{TR} \coloneqq C^{TR} + \left\{ C^O_{i^*j^*}\left(O_{i^*j^*}(b_{i^*j^*} + 1)\right) - C^O_{i^*j^*}\left(O_{i^*j^*}(b_{i^*j^*})\right) \right\} m_{i^*j^*}(1 - r_{i^*j^*})$, and $b_{i^*j^*} \coloneqq b_{i^*j^*} + 1$.

   Return to step 3a.

[3] This will never occur in the first iteration, but may happen in next iterations.

c. If $C^{TR} < h_{i_N}$, the costs of TPT reduction are less than the cost savings of removing a spare part, whereas we attain the target backorder level. Accept this stock reduction and go to Step 4. Otherwise, keep item i_N on stock at location j_N and STOP.

4) $N := N - 1$; If $N \geq 1$ and there are still options for TPT reduction left, then consider the next spare part for stock reduction: Go to Step 3.

Because we only have to update a limited number of partial derivatives each time we modify spare part stock levels or TPTs (Step 3a), the algorithm is pretty fast (from a fraction of a second to various minutes, depending on the size of the problem). The basic heuristic stops if it is not cost effective to reduce TPT to compensate for stock reduction of a *single* spare part. A straightforward *extension* is to consider stock reduction of two or more spare parts simultaneously, compensated by pieces of TPT reduction. In principle, we can continue until we run out of either options for spare part reduction or options for TPT reductions, whatever comes first (usually the TPT reductions come first). This may seriously increase the computation times, however. As a compromise, we consider stock reduction of multiple spare parts compensated by one or more pieces of TPT reduction, until the next best marginal effect of TPT reduction according to criterion (2.14) is less that the impact of removing the next spare part, being the total increase in LRU backorders at the bases divided by the decrease in costs h_{i_N}.

An obvious drawback of our heuristic is that the optimality gap is unknown. However, an optimal algorithm is not easy to find. An option is an approach similar to the method by Basten et al. (2012a) for the integration of decisions for repair locations and resource locations (Level of Repair Analysis) and spare part inventories. Such an approach is out of scope for this paper (see also Section 2.6). Advantages of our heuristic are its simplicity and speed, such that we are able to analyze models of realistic size. Moreover, the construction of the heuristic guarantees that we only find solutions that are as least as good as the standard VARI-METRIC procedure without considering TPT reductions.

2.5 Experiment and results

In this section, we design a numerical experiment to analyze the savings that can be obtained using joint optimization of spare part inventories and TPTs and to characterize its type of policies. We present our experimental design in Section 2.5.1, and discuss results in Section 2.5.2. We illustrate our method in a case study at Thales Netherlands (Section 2.5.3).

2.5.1 Experimental design

We focus on two-echelon, two-indenture networks. The holding cost rate is 25% of the item value per year, and the transportation time O_{ij} equals 0.02 years for all items and bases. We vary the size and type of the problem as specified in Table 2.1.

Experimental factor	low value	high value
Number of LRUs	25	100
Average number of SRUs per LRU	0.5	2
Average demand per LRU per base m_{ij} (per year)	0.05	0.25
Number of bases	3	10
Average repair time T_{ij} over all items (year)	0.05	0.25
Repair costs as a percentage of the item value	15%	30%
Transportation costs (€)	100	500
Target availability	0.95	0.99

Table 2.1 Experimental factors.

For each setting, we generate randomly 25 problem instances as follows.

1) We draw the demand per year per base for each LRU m_{ij} $(i \in L, j \in B)$ from a continuous uniform distribution around the mean with minimum demand rate 0.002.

2) We randomly assign the SRUs to LRUs using equal probabilities.

3) If an LRU has one or more SRUs, the probability that no SRU needs to be replaced upon LRU failure is always 0.1, whereas the remaining 0.9 probability mass is allocated to the SRUs based on a continuous uniform distribution (giving the cause probabilities q_{ki}).

4) We draw the *net* value per item from a shifted exponential distribution with lower bound €400 and mean €6000; the *gross* LRU value includes the net values of its SRUs.

5) All items can be repaired at the central depot (r_{ij} = 1 if j represents the central depot). At the bases, the repair probabilities r_{ij} only depend on the item i and are drawn from a continuous uniform distribution on the interval [0.1, 0.9].

In all cases, we consider the following options for TPT reduction (Table 2.2):

Repair		Transportation	
TPT reduction	Cost increase	TPT reduction	Cost increase
25%	40%	50%	100%
50%	100%		
75%	700%		

Table 2.2 Scenarios for repair TPT reduction and transportation TPT reduction.

We use fewer options for transportation TPT reduction, because these times are usually much smaller than repair TPTs. Altogether, our experiment consists of 2^8 (8 experimental factors) * 25 (random problem instances per setting) = 6,400 problem instances.

2.5.2 Numerical results

2.5.2.1 Savings percentage

We compute the cost savings from including throughput time reductions as decision variables in the optimization. That is, we compute the total costs as specified in the goal function of optimization problem $(P1)$ in Section 2.2.3 after optimization to the total costs after Step 1 of our algorithm (i.e. application of VARI-METRIC using standard TPTs only). Over all 6,400 problem instances, we find average cost savings of 19.8%.

Figure 2.4 shows the impact of the experimental factors on the savings, sorted by magnitude of impact. We observe that the average demand per LRU has the highest impact: TPT reduction is particularly profitable if demand is low. This makes sense, because repair and transportation costs increase proportionally in the demand, whereas spare part holding costs increase less than proportionally because of the portfolio effect. Further, the savings percentage decreases with the target availability, the number of LRUs in the system, the mean repair costs, and the mean repair time. The impact of the average availability and the number of LRUs is remarkable. In both cases, the average downtime allowed per LRU decreases. A possible explanation is that low downtime requirements per LRU lead to high spare part stock levels, and then the impact of TPT reduction is relatively low. The other factors (average number of SRUs per LRU, number of operational sites, transportation costs) have a marginal impact on the cost savings. We expect that higher transportation costs would lead to less reduction in order-and ship times and so to less cost savings. We do not see this in the savings percentage, but we see it in the type of policy that we choose. We will discuss these policies in more detail below (Section 2.5.2.2).

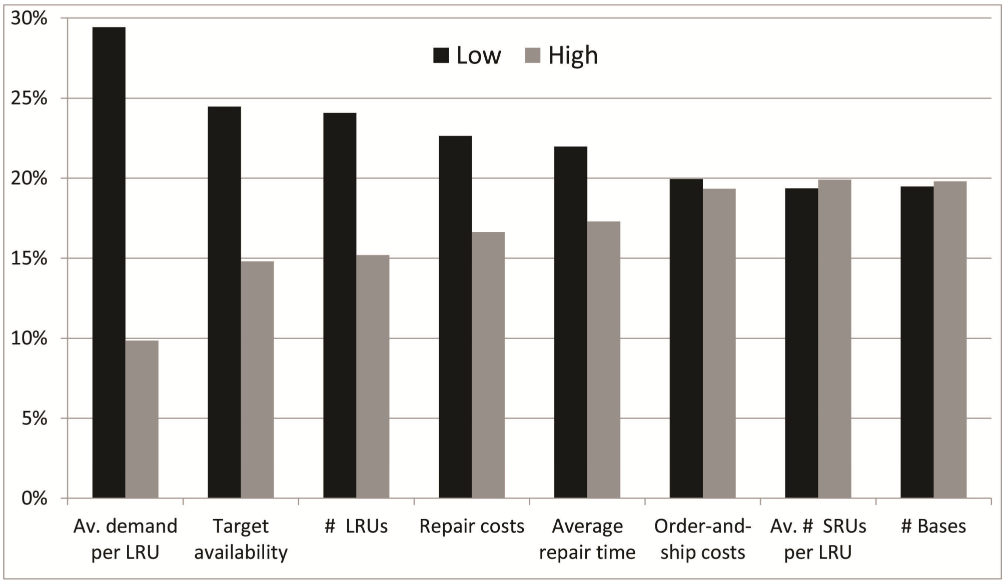


Figure 2.4 Impact of the experimental factors on the average cost savings (see Table 2.1 for the high and low settings per factor).

2.5.2.2 Type of policy

To examine the type of policy we find for the TPTs, we measure the degree of TPT reduction in a single problem instance by the weighted average percentage TPT reduction with the number of (repair or transportation) jobs as weights. We distinguish between the levels in the multi-echelon system and the levels in the multi-indenture structure. Obviously, we find most TPT reduction in the problem instances with the highest savings. Apart from that observation, the following observations are interesting:

- The average reduction in repair TPT is 8.5% for all upstream repairs and 24.8% for all downstream repairs. Clearly, we have most TPT reductions *downstream* in the network.
- We observe most TPT reduction for repairs downstream (at the bases) when repair costs are low and repair times are high (38% reduction).
- We hardly use repair TPT reduction of SRUs at the central depot (6.4% on average). We find the highest reduction in case of few bases and low demand rates (still only 12.7%).
- The average reduction in transportation TPT between central depot and bases is 25%.
- Although the transportation costs have little impact on the savings (see Figure 2.4), they influence the type of policy. The average TPT reduction is 34% if the costs per shipment are €100, and 16% for transportation costs of €500. So, we indeed reduce the transportation TPT less if the costs are higher.

2.5.2.3 Impact of scenarios for TPT reduction

If we only consider repair TPT reductions and no transportation TPT reductions, we still get significant average cost savings, namely 14.9% instead of 19.8%. If we limit the options to transportation TPT reductions however, the average cost savings are 3.3% only. It is remarkable that the joint effect of repair and transportation TPT reductions is larger than the sum of the separate effects.

Next, we analyze the impact of the number of scenarios for TPT reduction (i) by excluding scenarios for repair TPT reduction (we only allow cutting repair TPTs in half at twice the costs), and (ii) by adding scenarios for TPT reduction. In the latter case, we considered the following options for both repair and transportation TPTs as given in Table 2.3. We find that the average savings decrease from 19.8% to 16.0% if we reduce the number of options for TPT reduction. Under additional options, the average savings increase from 19.8% to 22.3%. So, the number of discrete steps in TPT reduction has impact, but it is not very large. We already achieve significant gains with a single alternative option for TPTs only.

Repair		Transportation	
TPT reduction	**Cost increase**	**TPT reduction**	**Cost increase**
10%	10%	10%	10%
25%	40%	25%	40%
50%	100%	50%	100%
60%	300%	60%	300%
75%	700%	75%	700%

Table 2.3 Scenarios for repair and transportation TPT reduction.

2.5.2.4 Three-echelon, three-indenture systems

To examine whether our findings remain valid for other network types, we designed a similar experiment for three-echelon, three-indenture networks. The cost savings are somewhat higher on average (24.8%), but the other findings are similar to two-echelon, two-indenture systems. The only new finding is that we observe a larger impact of the multi-indenture structure on the cost savings. Higher savings are feasible for the combination of more SRUs per LRU and more subcomponents per SRU, so for a "heavier" multi-indenture structure (29.2% savings). We particularly observe a higher reduction in repair TPTs as well as and transportation TPTs downstream (and particularly for LRUs).

2.5.3 Case study

To evaluate our method in a practical setting, we collected data for a part of a radar system at Thales Netherlands. The detailed data are confidential, but we give an outline of the key characteristics below. The data are related to a service contract covering six radar systems onboard of six frigates. Spare parts are supplied in a three echelon system from Thales Netherlands via a shore organization to the frigates. Spare parts may be stocked and repaired at each of the three levels. The subsystem consists of 114 different items, spread among two indenture levels (LRUs and SRUs). The item values vary from a few hundreds of euro's to more than €100,000 (LRU including SRUs). The options for TPT reduction are:

- Repairs at Thales Netherlands can be processed via a "fast channel" at extra labor costs, yielding a repair TPT reduction of 50% on standard values of several months. The extra labor costs are related to the product value and may easily exceed €1,000.

- Transportation TPT from Thales Netherlands to the Shore can be reduced from 14 days to 7 days at limited extra costs (extra transport by an express courier service).

- Transportation TPT from the shore organization to the ships can be reduced from 5 days to 2 days, but this yields huge extra costs (magnitude €10,000), since an additional helicopter flight from the Shore to a frigate on a mission is needed.

Application of our heuristic yields 6.3% savings on the spare part holding costs at extra repair and transportation costs equal to 0.7% of the original inventory investment, so we have a net saving of 5.6%. Note that this is *not* a percentage over the total spare part holding, repair and transportation costs, since we were not able to specify repair and transportation costs for the standard TPTs. In fact, we only need the additional costs of TPT reduction to apply our method. Although the savings are relevant for Thales Netherlands given the amount of money involved, it is clear that the savings are considerably less than the average that we observed in our theoretical experiments. We have the following explanation for this:

- The theoretical experiments show that TPT reduction downstream in the network is usually most profitable. However, Thales Netherlands can only influence repair times at the own site, since both the shore and the ships are part of the customer organization. Therefore, we only considered repair TPT reduction *upstream* in the supply chain.

- The same applies to the transportation TPT: reduction downstream is extremely expensive (helicopter flights) and therefore no realistic option. Only TPT reductions upstream are feasible at reasonable costs.

- We have only two options for repair TPTs, namely either a normal or a fast repair. As shown in Section 2.5.2.3, this reduces the potential for cost savings.

2.6 Conclusions

We developed a heuristic for the joint optimization of spare part inventories and TPTs of repair and transportation based on pricing of TPT reductions for multi-item, multi-echelon, multi-indenture spare part networks. Our heuristic is easy to apply and yields significant cost reductions compared to the standard VARI-METRIC method for spare part optimization where TPTs are fixed. We find that it is particularly profitable to reduce TPTs *downstream* in the supply chain. Repair TPT reduction of lower indenture items upstream in the supply chain is less useful. In a case study at Thales Netherlands, we find a cost reduction of 5.6%, which is somewhat low compared to our theoretical experiments. This is due to the fact that TPT reductions downstream in the Thales network are very expensive because of the special business characteristics (an installed base of radar onboard of frigates).

Our approach is flexible and heavily relies upon the VARI-METRIC method for inventory optimization in multi-echelon, multi-indenture networks. As a consequence, we believe that **known model extensions to VARI-METRIC** can be included in our approach rather easily, thereby relaxing some model assumptions as mentioned in Section 2.2.1. For example, we can include the VARI-METRIC variants to deal with negative binominal demand (assumption 1), differences in item criticality (assumption 2), replenishment order quantities larger than 1 (assumption 8), stochastic order-and-ship times (assumption 11), and commonality among items (assumption 12) (Sherbrooke, 2004). Other model assumptions lead to considerably more complex models, in particular relaxing assumption 10 to include the use of lateral supply between stock points at the same level in the multi-echelon structure. Even disregarding throughput time reductions, a complete approach for lateral supply in general multi-echelon, multi-indenture networks is still missing. Most models consider single or two-echelon networks with a single indenture level only (Paterson et al., 2011). This is a topic for further research.

Other further research would be the development of a method for **exact optimization** of this model to provide a benchmark for the performance of our heuristic. The approach as applied by Basten et al. (2012b) for the joint optimization of the spare part provisioning and Level Of Repair Analysis (LORA) problem seems to be most promising. However, we expect that an exact method require more computation time, so that it will not be suitable to solve problem instances of practical size.

In the subsequent chapter, we consider differentiation at a *customer level* in addition to differentiation at an item level. Specifically, we assume that we have various customer segments that each have distinct service requirements, resulting in the need to differ service over these customer segments.

Chapter 3

Selective emergency shipments[4]

3.1 Introduction

In the previous chapter, we focused on differentiation on an item level in spare parts supply by selectively reducing item throughput times. In this chapter, and Chapter 4 and 6, we apply differentiation on both an item level *and* a customer level. Specifically, we consider new tools at a tactical level to differentiate service to customers based on their service requirements. In this chapter, we focus on selective emergency shipments. Differentiation through selective lateral transshipments and dedicated stocks are discussed in Chapter 4 and 6 respectively.

In the selective emergency shipment model, unmet demand in the supply chain can either be backordered or satisfied using an emergency shipment from a secondary source with infinite supply, with the latter option being both faster and more expensive. We should thus investigate for which combinations of customer segments and item types it is a viable approach. Naturally, emergency shipments will be most beneficial for the customers with the highest service level requirements. However, the characteristics of an item also determine the added value of using emergency shipments: for inexpensive fast moving items, emergency shipments might be too expensive for any customer segment, whereas for expensive slow movers it might be better to minimize stocks and use emergency shipments for all demand.

We make the following contributions to the literature in this chapter: first, we develop two efficient and effective heuristics to find near-optimal stock levels and shipment strategies in a multi-item system with one warehouse and multiple customer segments. Second, we show how to analyze this system for a single item given a warehouse stock level and shipment strategy. We require such an analysis approach as a building block in a multi-item optimization. Third, in an extensive computational experiment, we compare the selective emergency shipment model to two benchmarks, namely (i) the one-size-fits-all strategy where a uniform service fulfillment process is used for all customers and (ii) the critical level policy. We then show that our model leads to clear savings over one-size-fits all strategies and can lead to savings that are close to those found with critical level policies. Finally, we will show that it is very effective to combine selective emergency shipments and critical level policies for service differentiation.

[4] This chapter is based on the paper “The selective use of emergency shipments for service-contract differentiation” by E. M. Alvarez, M. C. van der Heijden, and W. H. M. Zijm, International Journal of Production Economics, 143(2), pp. 518 – 526.

The remainder of the chapter is structured as follows. We state our optimization problem and solution approach in Sections 3.2 and 3.3 respectively. In Section 3.4, we describe how we analyze the system for a single item for the special case with two customer classes. We give the results of the numerical experiment in Section 3.5. In Section 3.6, we formulate conclusions.

3.2 Model

We first present an outline of our model in Section 3.2.1. Next, we discuss the validity of our selection of shipment policies (Section 3.2.2). In Section 3.2.3, we present our model assumptions and notation. We discuss the formal optimization problem in Section 3.2.4.

3.2.1 Model outline

Consider a local warehouse that supplies various types of parts to multiple customer classes, and a central depot with infinite supply that replenishes the local warehouse. All customers have the same system, with each item in the system being critical (i.e. an item failure causes a system failure). Each customer class has a distinct amount of time it is willing to wait for parts on average. The warehouse fills demand from all classes on a first-come-first-served basis. If it is out of stock, the warehouse may backorder the demand or request an emergency shipment from the central depot. We achieve service differentiation by only using emergency shipments for customer classes with tight waiting time restrictions. We expect this to be particularly beneficial for expensive slow movers that often have low fill rates (making the difference between regular and emergency shipment times crucial). Still, it will sometimes be better to avoid stocks altogether and use emergency shipments for all classes. Conversely, for cheap fast movers it is probably better to keep sufficient stock (avoiding expensive emergency shipments) and use full backordering. The shipment mode should thus depend on both the item characteristics and waiting time constraints per customer class.

In addition to the above model, we also consider a model where critical levels and selective emergency shipments are jointly used for differentiation. This combined model only satisfies demand from on-hand stock if it exceeds the critical level for the customer's class. Demand that cannot be met from on-hand stock is satisfied using either backordering or emergency shipments.

The objective in both models is to minimize system holding and shipment costs, under restrictions on the mean aggregate waiting time per class. Firms like Philips Healthcare and Océ Technologies usually have service level requirements with their clients in terms of e.g. average failure resolution times, with delays often being caused by waiting time for spares. Penalties may apply if the supplier violates the agreements, but we have not seen explicit backorder costs in service contracts. Therefore, we do not include penalty costs per unit waiting time in our objective function. Our decision variables are the item stock levels, and the shipment mode (regular, emergency) and critical level for each item and customer class.

3.2.2 Selection of shipment policies

In our model, we do not consider the state of the pipeline when selecting a shipment mode for a particular customer class. However, by incorporating pipeline information in our decision making, we likely find better shipment strategies. For instance, if the pipeline contains many items, the emergency shipment time might exceed the backorder waiting time, making backordering the faster *and* cheaper option. Conversely, if there are few or no items in the pipeline, emergency shipments might be needed to minimize waiting time. Still, we do not consider the complete system state when selecting a shipment mode to keep the notation transparent and reduce computational effort. After all, we are primarily interested in the suitability of selective emergency shipments for differentiation compared to critical level policies and the "one-size-fits-all" approach.

3.2.3 Assumptions and notation

3.2.3.1 Main assumptions

1. *Demand for each item occurs according to a Poisson process.*
2. *An $(S-1,S)$ base stock policy is applied for all items.* In practice, spares often tend to be expensive slow movers. Therefore, holding costs usually dominate ordering costs and hence the optimal order quantity is usually 1.
3. *Regular shipment times from depot to warehouse are exponentially distributed.* This assumption facilitates Markov chain analysis. Also, we show in Section 3.5.3.1 that the system performance measures are insensitive to the lead time distribution.
4. *The shipment time from the local warehouse to the customer is negligible.*
5. *An emergency shipment is shipped directly from central depot to customer (i.e. the shipment does not pass through the local warehouse).*
6. *We consider an infinite horizon.* As a result, the mean waiting time for any customer in class j will equal the average waiting time of class j as a whole.

3.2.3.2 Notation

For each item $i = 1,2,\dots,I$, we denote the mean replenishment lead time by T_i^{reg}, the emergency shipment time by T_i^{em}, the holding costs per time unit by h_i and the additional costs for an emergency shipment over a normal replenishment by SC_i^{em}. The latter cost factor is sufficient, since each request triggers either a normal replenishment or an emergency shipment.

Customers are assigned to classes $j = 1, \dots, J$, each having an upper limit W_j^{max} on the average waiting time for parts. W.l.o.g. we assume that class j has a higher priority than class k ($j < k$), and therefore $W_1^{max} \leq W_j^{max}$ ($j \geq 2$). Class j demand for item i occurs at rate m_{ij} (>0). $M_{.j} = \sum_{i=1}^{I} m_{ij}$ and $M_{i.} = \sum_{j=1}^{J} m_{ij}$ denote the total mean demand for class j and for item i, respectively. The decision variables for item i are:

- The *base stock* level S_i.
- The vector $\boldsymbol{C}_i = [C_{i1}, \dots, C_{iJ}]$ denoting the *critical levels* per class, with C_{ij} denoting the among of stock reserved for classes 1 up to $j-1$. As it is never optimal to withhold stock from the highest priority class, $C_{i1} = 0$.
- The *shipment strategy* D_i, denoting the highest customer class index for which emergency shipments are used in a stock-out situation. D_i is an integer between 0 and J, as emergency shipments are only sensible for higher priority customers.

We combine all variables into an *item policy* $(S_i, D_i, \boldsymbol{C}_i)$ with mean waiting time $EW_{ij}(S_i, D_i, \boldsymbol{C}_i)$ and fill rate $\beta_{ij}(S_i, D_i, \boldsymbol{C}_i)$ for item i and class j as performance indicators.

3.2.4 Formal optimization problem

We express the formal optimization problem $(P1)$ as follows:

$$(P1) \qquad \min_{S_i, D_i, \boldsymbol{C}_i} \sum_{i=1}^{I} \left\{ h_i S_i + SC_i^{em} \sum_{j=1}^{J} m_{ij} \left(1 - \beta_{ij}(S_i, D_i, \boldsymbol{C}_i)\right) I_{\{1 \dots D_i\}}(j) \right\}$$

$$\text{s.t.} \qquad \sum_{i=1}^{I} \frac{m_{ij}}{M_{.j}} EW_{ij}(S_i, D_i, \boldsymbol{C}_i) \leq W_j^{max} \qquad j = 1, \dots, J \qquad (3.1)$$

$$S_i \in \mathbf{N}_0, D_i \in \{0,1,\dots,J\}, C_{ij} \in \{0,1,\dots,S_i\} \qquad i = 1, \dots, I, j = 1, \dots, J$$

We minimize holding and emergency shipment costs with the demand-weighted mean waiting time for class j not allowed to exceed target W_j^{max}. The indicator function $I_{\{1 \dots D_i\}}(j)$ equals 1 if emergency shipments are used for class j (i.e. $j \in \{1 \dots D_i\}$) and 0 otherwise. We compute holding costs over the total stock S_i, including items in the pipeline. It is easy to compute holding costs over the on-hand stock instead (Kranenburg and Van Houtum (2008)).

3.3 Solution approach

Problem $(P1)$ is a nonlinear integer problem that we cannot decompose into separate single-item problems because of the aggregate waiting time restrictions (3.1). This differs from a variant with backorder costs where such a decomposition is possible and the I single-item problems can be solved easily (see Section 3.3.2). We use an approach similar to Dantzig-Wolfe decomposition: We reformulate $(P1)$ to a *linear* integer programming problem and solve its LP-relaxation to find a lower bound. In this section, we specify how the approach can be used for our optimization problem. We first show how to reformulate $(P1)$ to a linear problem and find a lower bound (Sections 3.3.1 and 3.3.2 respectively). As the lower bound is generally fractional, Section 3.3.3 gives two heuristics to find near-optimal integer solutions.

3.3.1 Reformulation to a linear problem

Let B_i be the set of item policies we consider for item i, with $b_{ir} = \big(S_i(r), D_i(r), \boldsymbol{C}_i(r)\big)$ denoting a single item policy in this policy set (so $b_{ir} \in B_i$, with $r = 1,2, \dots, |B_i|$). Furthermore, let the binary variable $x_{b_{ir}}$ indicate whether policy b_{ir} is selected for item i or not ($x_{b_{ir}} = 1$ or 0). We then obtain the linear integer program $(P2)$.

$$(P2) \qquad \min \sum_{i=1}^{I} \sum_{r=1}^{|B_i|} TC_i(b_{ir}) x_{b_{ir}}$$

$$\text{s.t.} \qquad \sum_{i=1}^{I} \sum_{r=1}^{|B_i|} \frac{m_{ij}}{M_{.j}} EW_{ij}(b_{ir}) x_{b_{ir}} \leq W_j^{max} \qquad j = 1, \dots, J \tag{3.2}$$

$$\sum_{r=1}^{|B_i|} x_{b_{ir}} = 1 \qquad i = 1, \dots, I \tag{3.3}$$

$$x_{b_{ir}} \in \{0,1\} \qquad i = 1, \dots, I, r = 1, \dots, |B_i|$$

Here, $TC_i(b_{ir})$ is shorthand for the total costs related to item i under policy b_{ir}, so $TC_i(b_{ir}) = TC_i\big(S_i(r), D_i(r), \boldsymbol{C}_i(r)\big) = h_i S_i(r) + SC_i^{em} \sum_{j=1}^{J} m_{ij} \Big(1 - \beta_{ij}\big(S_i(r), D_i(r), \boldsymbol{C}_i(r)\big)\Big) I_{\{1 \dots D_i(r)\}}(j)$

3.3.2 Finding a lower bound for the total costs

We first solve the LP-relaxation of $(P2)$ with an initial item policy set that results in a feasible solution. Then, we use column generation to iteratively find new item policies that improve the solution if added. We stop once such policies no longer exist. For further details on column generation, we refer the reader to Section 1.9.2. In this model, we find a single initial policy for each item i by setting D_i to 0 and only increasing S_i. Then, we iteratively add the policy with the lowest reduced costs to B_i, if these are negative. The reduced costs for policy b_i, denoted by

$RED(b_i)$, are given by the expression below, with $u_j(\leq 0)$ and $v_i(\geq 0)$ denoting the current shadow prices for constraints (3.2) and (3.3) respectively. For simplicity, we omit suffix r.

$$RED(b_i) = TC(b_i) - \sum_{j=1}^{J} \frac{u_j m_{ij}}{M_{.j}} EW_{ij}(b_i) - v_i$$

For each shipment strategy D_i, we first find the values for S_i and $\boldsymbol{C}_i$ that give minimum reduced costs. Next, we select the item policy with minimum reduced costs over all shipment strategies. A complication in finding an optimal item policy for a shipment strategy is that the reduced costs are rarely convex as a function of S_i and/or $\boldsymbol{C}_i$: the reduced costs are only convex in S_i if $C_{ij} = 0$ for all classes and $D_i = J$. Then, there is no stock rationing and all unmet demand is satisfied through emergency shipments (see Kranenburg and Van Houtum (2007a)). However, from some value onwards, $RED(b_i)$ will only monotonically increase in S_i, irrespective of the values for D_i and $\boldsymbol{C}_i$: as S_i increases, the holding costs increase linearly, whereas the costs related to emergency shipments and waiting time decrease and eventually become negligible. We thus find an upper bound on S_i (and all C_{ij}) when the holding costs outweigh the other cost elements in $RED(b_i)$.

3.3.3 Methods for finding a near-optimal integer solution

We find near-optimal integer solutions using 2 methods: (1) we use the (non-integer) LP relaxation solution as a starting point for local search; (2) we solve *integer* problem $(P2)$ with all policies generated for finding the lower bound. In the literature (e.g. Kranenburg and Van Houtum (2008)), method 1 is often used, but we show in Section 3.5.3.2 that method 2 works better.

3.3.3.1 *Method 1: use a local search algorithm*

Kranenburg and Van Houtum (2008) find an integer solution by first selecting for each item in the LP-relaxation solution the policy b_{ir} with the lowest stock level, subject to $x_{b_{ir}} > 0$. As the resulting solution is usually infeasible, they then iteratively increase the stock level of one item until they find a feasible solution. In each iteration, the item is selected that leads to the largest decrease in the total gap between target and actual mean waiting times per invested euro. Obviously, we have to adapt this method in order to deal with multiple shipment modes. Our criterion for selecting a new item policy also differs from that of Kranenburg and Van Houtum. The latter authors place equal value on reducing either the waiting time for premium customers or that for non-premium customers. We, however, expect that a larger marginal investment is needed to reduce a small waiting time by an amount x compared to reducing a large waiting time by the same amount. Hence, it might be beneficial to select a policy that reduces premium waiting times (that are generally small) by a certain amount over an alternative that reduces non-premium waiting times by a greater amount. To incorporate the fact that the value we place on waiting time reduction may differ per class, we assign weights to the waiting time reduction amounts of each class and then iteratively select the item policy leading to the largest weighted waiting time reduction per extra investment. We now give further details on both the neighborhood construction and policy selection criterion.

We incorporate multiple shipment modes when constructing a *neighborhood* (i.e. the solutions close to the current solution from which we choose a new solution). Our neighborhood contains solutions with either a larger stock level S_i than the current solution or a faster shipment method (i.e. a larger value for D_i). If we increase S_i, we combine this with all values of D_i smaller than or equal to the current shipment strategy provided that the resulting policy has lower waiting times. Similarly, when we increase D_i, we combine this with all values of S_i that are smaller than or equal to the current value insofar that the policy waiting times are lower. This gives us several neighbors for each SKU i. We consider multiple decision variables to increase the probability of finding a feasible solution more quickly.

Furthermore, in our *selection criterion* we place greater value on reducing waiting times of premium customers compared to those of non-premium customers. We do so by assigning a higher weight to a waiting time reduction for the premium class. We realize that it is not trivial to translate the value placed on waiting time reduction in adequate class-specific weights. Still, for simplicity we use the inverse of W_j^{max} as a *weight* for class j. Then, the waiting time reduction for classes with very low waiting time targets gets a correspondingly high weight. We obtain the following expressions for the reduction in waiting time $\Delta \mathrm{W}_i(b'_{ir})$ and the additional investment $\Delta TC_i(b'_{ir})$ compared to the current policy b_{ir}, with b'_{ir} denoting a new item policy:

$$\Delta W_i(b'_{ir}) = \sum_{j=1}^{J} \frac{1}{W_j^{max}} \left\{ \left[\sum_{i=1}^{I} \frac{m_{ij}}{M_{.j}} EW_{ij}(b_{ir}) - W_j^{max} \right]^+ - \left[\sum_{i=1}^{I} \frac{m_{ij}}{M_{.j}} EW_{ij}(b'_{ir}) - W_j^{max} \right]^+ \right\} \tag{3.4}$$

$$\Delta TC_i(b'_{ir}) = \sum_{i=1}^{I} \left(TC_i(b'_{ir}) - TC_i(b_{ir}) \right) \tag{3.5}$$

In (3.4), $[a]^+ = \max\{0, a\}$. When evaluating item policies, we may find alternatives b'_{ir} with both lower costs *and* lower waiting times than b_{ir}. Then, we select the policy with the largest value for $\Delta W_i(b'_{ir})/\Delta TC_i(b'_{ir})$ over those with lower costs instead of over the entire neighborhood. We realize that we do not necessarily select the best item policy in this way. However, the criterion used does not seem to matter, as we observed in a computational experiment with 64 instances: in that experiment, the solutions found if we selected the policy with the largest value of $\Delta W_i(b'_{ir})$ from those with lower costs were exactly the same as those when we used $\Delta W_i(b'_{ir})/\Delta TC_i(b'_{ir})$ as a selection criterion.

3.3.3.2 Method 2: Use of Integer Programming (IP)

When solving the LP-relaxation, we usually only generate 4 to 7 item policies per item. Therefore, we should be able to solve the corresponding integer problem with a commercial solver (we used CPLEX) for most problems of realistic size. However, the generated item policies are not always related, especially for fast moving items. For instance, for a problem instance with two customer classes we have found policies (S_i, D_i, C_{i2}) of (0,1,0) (i.e. no stock, emergency shipments for premium customers only), (9,0,0) and (10,0,0) (i.e. high stock levels, full backordering) for the same item. To test the quality of our method when using the relaxation policy set, we compared the resulting solutions to those found when we include additional item policies in the IP that bridge the gap between the distinct item policies[5]. We found that these additional policies greatly increase computation time, while the solution quality improves only marginally: the average gap to the lower bound drops from 0.041 to 0.038, and the maximum gap drops from 0.259 to 0.258. Therefore, we find that it is sufficient to only use the item policies generated when solving the LP relaxation.

3.4 Evaluating item policies with two customer classes

We use continuous-time Markov chain analysis to find performance measures for an item policy. For simplicity, we limit ourselves to two customer classes in the remainder of this chapter. In Chapter 8, we discuss extensions to more than two classes. We thus consider three shipment strategies: use emergency shipments for both classes ($D_i = 2$), for class 1 only ($D_i = 1$), or not at all ($D_i = 0$). Per item, we have a critical level for the non-premium class (denoted by C_i from now on). Under backordering, we find the expected waiting time by Little's Law: $EW_{ij}(S_i, D_i, C_i) = EBO_{ij}(S_i, D_i, C_i)/m_{ij}$, with $EBO_{ij}(S_i, D_i, C_i)$ being the average number of backorders for item i and class j. When emergency shipments are used, $EW_{ij}(S_i, D_i, C_i)$ equals $\left(1 - \beta_{ij}(S_i, D_i, C_i)\right) T_i^{em}$. Section 3.4.1 first describes the pure selective emergency shipment models, where critical levels are not used. Section 3.4.2 then describes the combined shipment models, which incorporate critical level policies. For simplicity, we omit the item index i and denote the normal replenishment rate by $\mu = 1/T^{reg}$ and the total item demand rate by M instead of $M_{.}$.

3.4.1 Pure selective emergency shipment models

Pure model, emergency shipments for both classes ($D_i = 2$)

We model the pipeline as an Erlang loss system with S servers (Kranenburg and van Houtum, 2007). Let k denote the number in the pipeline. We find:

[5] This was a mid-sized experiment of 100 problem instances with 25, 100 and 400 items.

$$p_k = \left(\frac{M}{\mu}\right)^k \frac{1}{k!} \Big/ \sum_{n=0}^{S} \left(\frac{M}{\mu}\right)^n \frac{1}{n!} \qquad k = 0, \dots, S$$

$$\beta_1(S,D) = \beta_2(S,D) = 1 - p_S$$

Pure model, backorder class 2 demand only ($\boldsymbol{D_i = 1}$)

We define the state as the number of items in the pipeline k, with state k having $[k-S]^+$ class 2 backorders. Figure 3.1 displays the Markov chain. Once the pipeline contains S or more items, class 1 demand is lost to the system. Closed-form expressions for the state probabilities p_k, class 1 fill rate and class 2 mean backorder level are given below the figure.

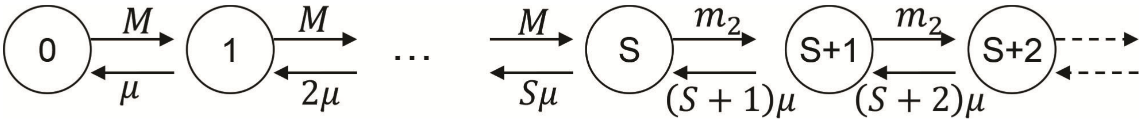


Figure 3.1 Transition diagram for pure model with backordering of class 2 requests.

$$p_0 = \left\{ \sum_{k=0}^{S} \frac{1}{k!}\left(\frac{M}{\mu}\right)^k + \left(\frac{M}{m_2}\right)^S \left(e^{m_2/\mu} - \sum_{k=0}^{S} \frac{1}{k!}\left(\frac{m_2}{\mu}\right)^k \right) \right\}^{-1}$$

$$p_k = \begin{cases} \left(\frac{M}{\mu}\right)^k \frac{1}{k!} p_0 & 1 \le k \le S \\ \left(\frac{M}{\mu}\right)^S \left(\frac{m_2}{\mu}\right)^{k-S} \frac{1}{k!} p_0 & k \ge S+1 \end{cases}$$

$$\beta_1(S,D) = \sum_{k=0}^{S-1} p_k$$

$$EBO_2(S,D) = \left(\frac{M}{m_2}\right)^S p_0 \left\{ \frac{m_2}{\mu}\left(e^{m_2/\mu} - \sum_{k=0}^{S-1} \frac{1}{k!}\left(\frac{m_2}{\mu}\right)^k \right) - S\left(e^{m_2/\mu} - \sum_{k=0}^{S} \frac{1}{k!}\left(\frac{m_2}{\mu}\right)^k \right) \right\}$$

Pure model, backorder demand from all classes ($\boldsymbol{D_i = 0}$)

We use priority backorder clearing: class 1 backorders are cleared before class 2 backorders, even if a class 2 backorder occurred first. Therefore, we need a two-dimensional state space, since the number of backorders per class can even differ among states with the same number of items in the pipeline. We use states (k,l), with k the number in the pipeline and l the number of class 2 backorders. We then have $[[k-S]^+ - l]^+$ class 1 backorders. Figure 3.2 shows the corresponding Markov chain. Demand flows from $(k,0)$ to $(k+1,0)$ until the pipeline contains S items. Then, class 1 demands result in shifts from (k,l) to $(k+1,l)$, while class 2 demands result in shifts from (k,l) to $(k+1,l+1)$. Replenishment flows go from (k,l) to $(k-1,l)$ whenever (k,l) has class 1 backorders. Then, the pipeline decreases, while the number of class 2 backorders remains the same. Flows from (k,l) to $(k-1,l-1)$ only occur when (k,l) only has class 2 backorders.

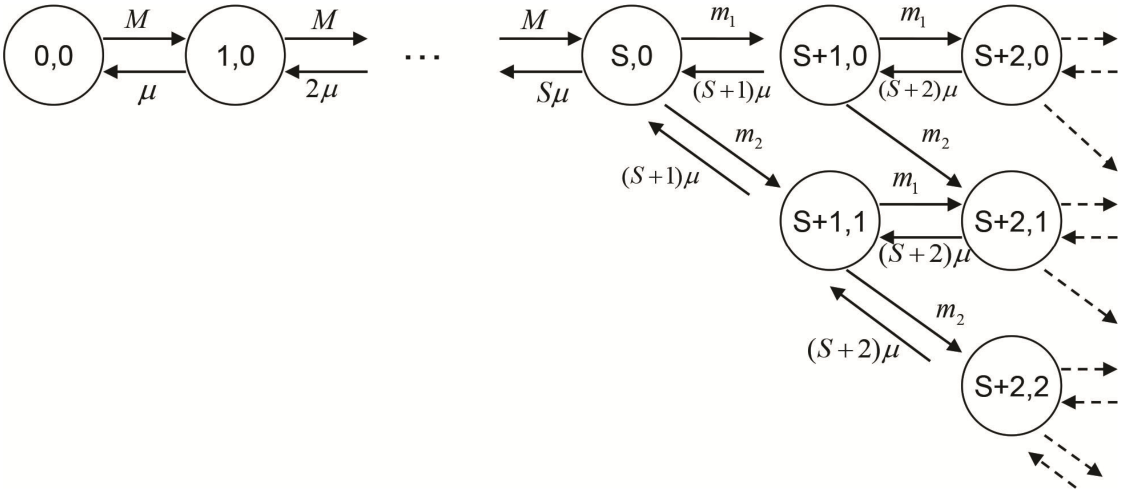


Figure 3.2 Transition diagram for pure model with full backordering.

We could not find analytical expressions for the state probabilities, so we compute an upper bound k^{UB} on the number of items in the pipeline and solve the balance equations numerically. We find k^{UB} by aggregating all demand into a single class and analyzing the resulting $M|M|\infty$ model exactly. Note that the pipeline distribution will be the same as in the two-class model, since the demand and replenishment rates are the same. We find k^{UB} such that $1-\sum_0^{k^{UB}} p_k \leq \varepsilon$ with $\varepsilon = 10^{-8}$. We then find the following expressions for $EBO_j(S,D)$:

$$EBO_1(S,D) = \sum_{k=S+1}^{k^{UB}} \sum_{l=0}^{k-S} (k-S-l)\cdot p_{kl}$$

$$EBO_2(S,D) = \sum_{k=S+1}^{k^{UB}} \sum_{l=0}^{k-S} l\cdot p_{kl}$$

3.4.2 Combined models that incorporate critical levels

We refer to Kranenburg and Van Houtum (2008) for the model with emergency shipments for both classes ($D_i = 2$) and only discuss the (partial) backordering models.

Combined model, backorder demands from all classes ($\boldsymbol{D_i = 0}$)

We follow Ha (1997b), who shows the optimality of only clearing class 2 backorders once all class 1 backorders have been cleared and the on-hand stock is at least C. Our Markov chain (Figure 3.3) consists of states (k, l), with k the number in the pipeline and l the number of class 2 backorders. We can simultaneously have stock on-hand (i.e., fewer than S items in the pipeline) and class 2 backorders, so we cannot derive the number of backorders from the pipeline alone. The Markov chain branches out at $S - C$ or more items in the pipeline (and thus at most C items on-hand): then, class 1 demand is met from stock, with class 2 demand being backordered. Once we are out of stock, we also backorder class 1 demands. Most replenishment flows go from (k, l) to $(k - 1, l)$: we then clear class 1 backorders or increase on-hand stock.

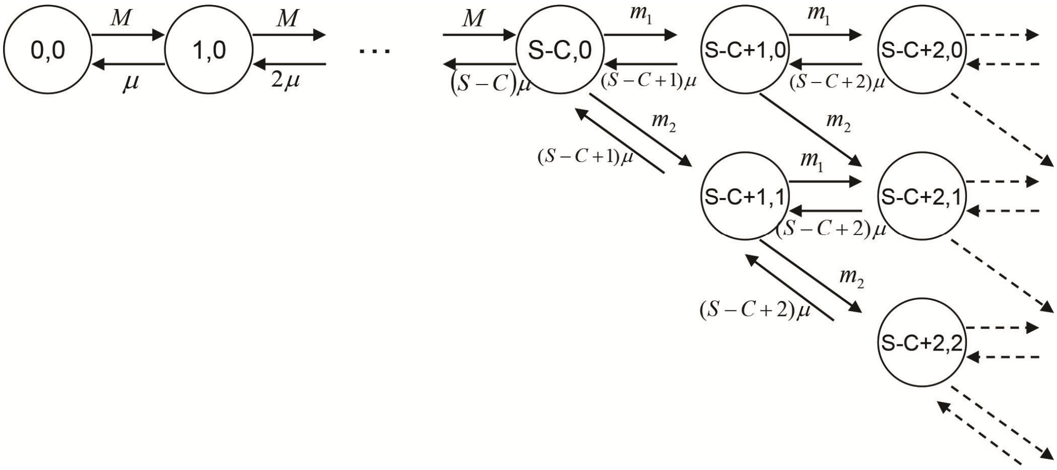


Figure 3.3 Transition diagram for combined model with full backordering.

Since this model is very similar to the pure model with full backordering, we use the same approach to compute k^{UB} and find the state probabilities. We find for $EBO_j(S, D, C)$:

$$EBO_1(S, D, C) = \sum_{k=S+1}^{k^{UB}} \sum_{l=0}^{k-S+C} \max\{0, k - S - l\} \cdot p_{kl}$$

$$EBO_2(S, D, C) = \sum_{k=S-C+1}^{k^{UB}} \sum_{l=0}^{k-S+C} l \cdot p_{kl}$$

This Markov chain (and hence the state probabilities) only depends on the value of $S - C$. Hence, we directly find the performance measures for all other policies with the same value for $S - C$, which greatly reduces the computational burden of analyzing item policies.

Combined model, backorder class 2 demand only $(\boldsymbol{D_i = 1})$

We use state space (k, l) as above, see Figure 3.4. However, once we are out of stock now, denoted by states $(k, k-S)$ with $k \geq S$, the pipeline only increases further from class 2 demand. Increasing on-hand stock to C has priority over clearing class 2 backorders.

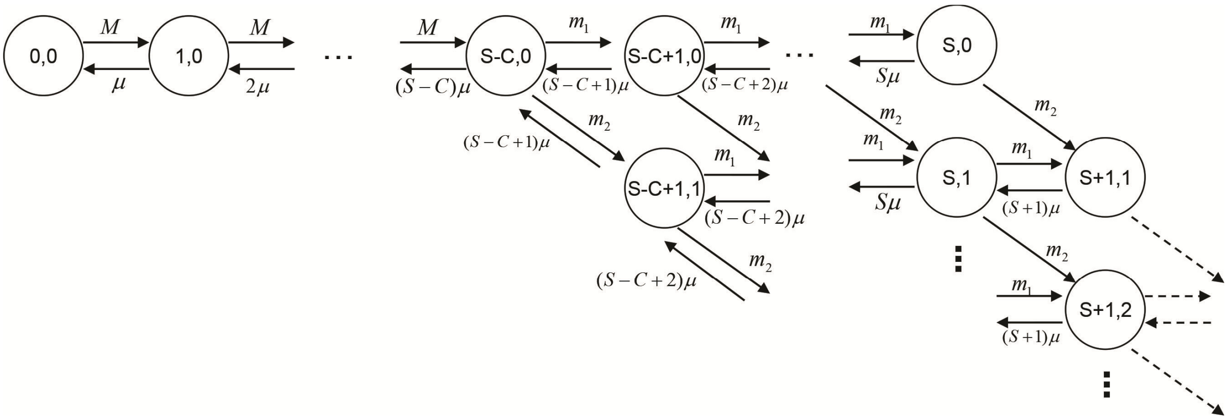


Figure 3.4 Transition diagram for combined model with backordering of class 2 requests.

We determine k^{UB} by aggregating all demand and assuming full backordering. Using this value, we then solve the resulting balance equations. Note that our pipeline bound might be larger than necessary, since we assume full backordering when computing k^{UB}. For the performance measures we find:

$$EBO_2(S, D, C) = \sum_{k=S-C+1}^{k^{UB}} \sum_{l=\max\{0,k-S\}}^{k-S+C} l \cdot p_{kl}$$

$$\beta_1(S, D, C) = 1 - \sum_{k=S}^{k^{UB}} p_{k,k-S}$$

3.5 Computational experiment

We conducted a numerical experiment, for which we state the objectives in Section 3.5.1. Section 3.5.2 specifies the experiment design and Section 3.5.3 the results.

3.5.1 Objectives

Our objectives are: (i) to determine the sensitivity of the system performance measures to the replenishment lead time distribution in the selective emergency shipment model, (ii) to evaluate the two heuristics for obtaining a near-optimal solution (i.e. local search and IP) in terms of solution quality and computation time, (iii) to determine whether and when selective emergency shipments are effective for service differentiation, (iv) to compare selective emergency shipments to critical level policies as differentiation tools, (v) to assess the added value of *jointly* using selective emergency shipments and critical level policies for differentiation.

3.5.2 Experiment design

Table 3.1 shows the tested parameter values, based on the values by Kranenburg and Van Houtum (2008), which are derived from observations in practice. We use $EC_i^{em} = 1000$ as cost normalization. For each combination of parameters 1, 3, 4 and 6, we generate 4 random instances as follows: for each item, a demand rate and holding cost is drawn from uniform distributions on the given intervals, with the correlation between demand rates and holding costs being -0.8. This is realistic, since fast movers tend to have low (holding) costs in practice and vice versa. Except for the demand rates and holding costs, all items in an instance have the same parameter values. We have 3456 instances in total: we have 864 parameter combinations and 4 demand rate/holding cost samples per combination.

	Parameter	Values
1	Number of items I	25, 100, 400
2	Daily demand rate per item $M_{i.}$	$U[0, 0.1]$, $U[0, 0.5]$
3	Fractions of class demand per item $(m_{i1}/M_{i.}; m_{i2}/M_{i.})$	(0.2; 0.8), (0.5; 0.5), (0.8; 0.2)
4	$(T_i^{reg}; T_i^{em})$(in days)	(4; 1), (8; 1), (8;2),(16; 2)
5	Item holding cost interval (per unit per day)	$U[0.02, 19.98]$, $U[0.2, 199.8]$, $U[2, 1998]$
6	Target service levels $(W_1^{max}; W_2^{max})$ (in hours)	(0.5; 2), (0.5; 4), (3; 12), (3; 24)

Table 3.1 Parameter values of the tested instances.

3.5.3 Results

First, we estimate the sensitivity of the performance measures to the lead time distribution. Then, we evaluate the performance of the two heuristics described in Section 3.3.3. Next, we investigate whether and when the emergency shipment strategy has added value over one-size-fits-all strategies. Finally, we compare the emergency shipment policy to the critical level policy and determine the added value of combining both policies.

3.5.3.1 Sensitivity of performance measures to lead time distribution

For a single-class system with emergency shipments, Alfredsson and Verrijdt (1999) show that the system performance measures do not depend on the distribution of the replenishment lead time. We now test whether this observation still holds for our selective emergency shipment model before proceeding to optimization. For 48 problem instances with one item, we used simulation to find mean waiting times per class for both deterministic and exponential lead times. The regular shipment time was 5 or 10 days, the emergency shipment time was 1 or 2 days, and the class demand rates $(m_1; m_2)$ were either $(0.01; 0.04)$ or $(0.1; 0.4)$.

Per shipment strategy, we compute the waiting time deviations as a fraction of the exponentially distributed times. Table 3.2 shows the results for cases with waiting times of at least 10^{-3}. We find that the lead time distribution still has little influence on the waiting times: in general, the average deviations are small. The deviations increase as backordering is used for more classes, particularly for class 1 waiting times, but the differences remain reasonable even under full backordering (we find deviations of 14% when EW_1 is 1.5 and 1.7 respectively). Also, in practice we expect waiting times under backordering to show more variability than those under emergency shipments, especially when items are repaired (as is common for expensive slow movers). Under full backordering, we thus expect exponential shipment times to characterize the supply process more accurately than deterministic times.

Shipment strategy	**Average deviation**		**Maximum deviation**	
	EW_1	EW_2	EW_1	EW_2
Em. shipments for both classes	0.3%	0.1%	1.7%	0.6%
Em. shipments for premium customers only	1.7%	0.3%	9.2%	0.9%
Backordering for both classes	8.8%	0.7%	14.1%	1.5%

Table 3.2 Performance comparison under exponential and deterministic lead times.

3.5.3.2 Performance of the heuristics

We express the solution quality in terms of a relative gap to the lower bound, defined as $TC_H - TC_{LB}/TC_{LB}$, where TC_H gives the solution value of the heuristic (IP or Local Search) and TC_{LB} denotes the lower bound from Section 3.3.2. Table 3.3 shows the solution quality and computation times for different numbers of items, the parameter with most impact. We used a Intel quad core, 2.83 GHz processor. We see that integer programming yields a gap to the lower bound less than half that of local search on average. This gap clearly decreases with the number of items. This is beneficial, because practical instances typically contain hundreds of items. The computation times for both methods are small, although the run times of IP increase greatly with the problem size. Of the overall computation time, the largest fraction is spent on the column generation procedure and hence solving the LP-relaxation (column 'Comp. time LB').

Parameter	**Values**	**gap IP (%)**		**gap LS (%)**		**Comp. time LB (min)**		**Comp. time IP (sec)**		**Comp. time LS (sec)**	
		avg.	max.	avg.	max.	avg.	max.	avg.	max.	avg.	max.
Number of SKUs	25	0.25	2.16	0.51	3.81	0.15	0.38	0.10	0.64	0.01	0.06
	100	0.02	0.11	0.05	0.35	0.62	1.53	0.14	0.73	0.03	0.73
	400	0.00	0.02	0.00	0.08	2.54	6.25	1.48	10.34	0.50	2.66
Overall		**0.09**	**2.16**	**0.19**	**3.81**	**1.10**	**6.25**	**0.58**	**10.34**	**0.18**	**2.66**

Table 3.3 Solution quality and computation times under integer programming (IP) and local search (LS).

3.5.3.3 The added value of using emergency shipments

We compare our selective emergency shipment (SES) policy to a one-size-fits-all policy with emergency shipments for all items (OSFA ES), and a variant where we use either backordering or emergency shipments for an item (OSFA BO+ES). The latter policy differentiates among items, but *not* among customer classes. Backorder clearing is done first-come-first-served. Under one-size-fits-all, we use a single customer class with maximum waiting time W_1^{max}. We use OSFA ES as a benchmark, since this is common for testing critical level policies (e.g. Kranenburg and Van Houtum (2008)). Table 3.4 shows the savings of OSFA BO+ES and our policy compared to OSFA ES. We also show the results for different holding cost intervals, because this parameter has most influence on the savings.

Parameter	**Values**	**savings over OSFA ES (%)**			
		OSFA BO+ES		SES	
		Avg.	Max.	Avg.	Max.
holding cost range interval	[0.02, 19.98]	7.9	38.9	11.7	45.8
	[0.2 - 199.8]	0.2	2.7	1.2	12.6
	[2 - 1998]	0.1	1.9	0.3	3.0
Overall		**2.7**	**38.9**	**4.4**	**45.8**

Table 3.4 Savings of selective emergency shipments (SES) and OSFA BO+ES over OSFA ES.

OSFA BO+ES gives average savings of 2.7% over OSFA ES. Selective emergency shipments yield additional average savings of 1.7%, resulting in average savings of 4.4% over OSFA ES with a maximum of 45.8%. The savings are largest when holding costs are low, because emergency shipments are less appealing then: it is cheaper to keep large stocks and reserve emergency shipments for premium customer demand only. We also find large savings when waiting time restrictions for class 1 demand are loose (7.4% on average).

Figure 3.5 displays on the left the fraction of items assigned to each shipment strategy for OSFA BO+ES and the SES policy. The figure to the right shows the division of items over shipment strategies per holding cost interval for our policy. Both SES and OSFA BO+ES have roughly the same fraction of items where full backordering is used ($D_i = 0$). Clearly, we use the selective shipment strategy ($D_i = 1$) to limit using expensive emergency shipments for both classes ($D_i = 2$). On average, $D_i = 1$ for 20% of the items. This fraction increases to more than 40% when the holding costs are low (see figure on the right).

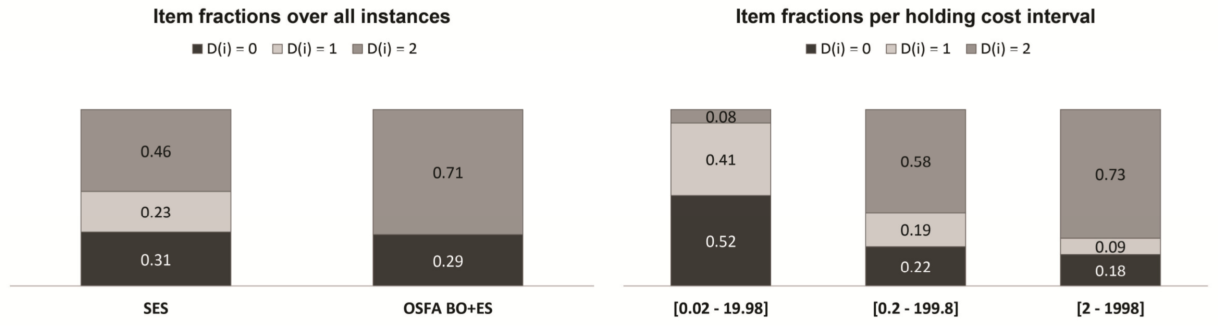


Figure 3.5 Fraction of items per shipment strategy for selective emergency shipments (SES) and OSFA BO + ES (left) and fraction of items per holding cost interval for SES (right).

From Figure 3.5, we thus see that the division of items over shipment strategies varies for different holding cost intervals. However, the figure does not tell us for what *types of items* each strategy is used. Therefore, we show the average item holding costs and demand rates per shipment strategy in Figure 3.6. Clearly, the selective strategy is mainly used for expensive slow movers. Little or no stock is kept of such items, making the shipment mode used crucial for meeting waiting times: use emergency shipments for premium clients and backorder non-premium requests.

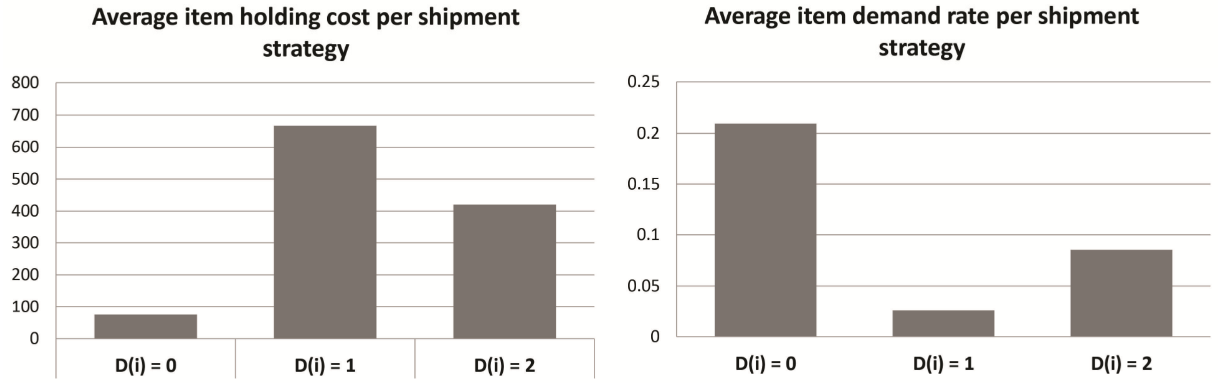


Figure 3.6 Overall average holding costs and demand rates per shipment strategy.

3.5.3.4 Comparison to the critical level policy and a combined policy

We compare our policy to a critical level policy with emergency shipments (CLP ES). Also, we investigate the benefits of combining both policies (CLP+SES). Table 3.5 shows the relative savings of the policies compared to OSFA ES. Critical level policies generally outperform selective emergency shipments, with average savings of 7.9%. This is caused by the mode of differentiation: in critical level policies, premium customers will often obtain a part right away. In contrast, customers need to wait at least T_i^{em} time units for an emergency shipment. The selective emergency shipment policy is also less sensitive to the waiting time restrictions than the critical level policy: the waiting time restriction for class 1 is usually dominant. Then, increasing W_2^{max} has little impact on the solutions found.

Parameter	Values	Savings over OSFA ES (%)		
		SES	CLP ES	CLP + SES
$T_i^{reg} - T_i^{em}$ (days)	4-1	7.6	5.8	16.1
	8-1	4.5	8.5	14.3
	8-2	4.1	7.5	13.1
	16-2	1.5	9.7	12.1
Holding cost range interval	[0.02, 19.98]	11.7	0.5	16.2
	[0.2 - 199.8]	1.2	8.3	10.5
	[2 - 1998]	0.3	14.8	15.0
$W_1^{max} - W_2^{max}$ (hours)	0.5 - 2	1.4	4.4	6.2
	0.5 - 4	1.4	7.6	10.4
	3 - 12	7.4	8.4	16.7
	3 - 24	7.5	11.1	22.4
Overall		**4.4**	**7.9**	**13.9**

Table 3.5 Savings of different policies over OSFA ES.

Selective emergency shipments outperform critical level policies in cases with short regular shipment times, low holding costs, and loose waiting time restrictions for class 1 demand. Then, it is viable to meet (a part of) the demand through the regular channel instead of expensive emergency shipments. Indeed, the fraction of items for which D_i is 0 or 1 is relatively high then (for the holding costs we can see this in Figure 3.5). Note that selective emergency shipments do not outperform CLP ES for the given waiting time restrictions, but this happens if we further increase W_1^{max}. Under the mentioned conditions, the base stock levels with CLP ES tend to be high to avoid expensive emergency shipments.

Obviously, the combined policy works best. Still the additional gain is surprisingly large: it exceeds the combined savings of the individual policies. The reason is that under CLP ES, the mean waiting time *for class 2 customers* tends to be considerably below the target; the class 1 target is usually the bottleneck. By including selective emergency shipments, we are able to push the actual performance of low priority customers closer to the target (0.04% instead of 29% in the experiments with 100 items).

3.6 Conclusions

In this chapter, we considered the selective use of emergency shipments as a tool for applying service differentiation in spare parts supply. If demand could not be satisfied from on-hand stock at the warehouse, it was either backordered or satisfied through an emergency shipment, with the chosen shipment mode depending on the item type and customer class. We showed how to accurately analyze this system for a single item under various shipment strategies.

Furthermore, we developed two heuristics for multi-item optimization that are accurate (average gaps to the lower bound well below 1%) and fast. Clearly, greedy approaches are not necessary to find good solutions: integer programming with limited columns is simple and works well. In an extensive computational experiment, we showed selective emergency shipments have clear added value, with average savings of 4.4% compared to one-size-fits-all policies. The approach also outperforms critical level policies when holding costs are low, premium waiting times are not very tight, and regular shipment times are short. Then, emergency shipments are very expensive and should thus be avoided. Differentiation through selective emergency shipments is most useful for expensive slow movers, since the approach has most impact when little or no stock is kept of an item. Finally, we find large savings (13.9% on average) by jointly using critical levels and selective emergency shipments for differentiation in spare parts supply.

In the next chapter, we expand the selective emergency shipment model to allow for lateral transshipments for premium customers. We expect that the addition of selective lateral transshipments could be beneficial as a differentiation tool as such transshipments are generally both faster and less expensive than emergency shipments.

Chapter 4

Selective lateral transshipments[6]

4.1 Introduction

In the previous chapter, we have shown that selective emergency shipments can be beneficial as a tool for applying service differentiation. In this chapter, we extend the selective emergency shipment model by also allowing *lateral transshipments for premium customers* in addition to the earlier backordering and emergency shipment options. In practice, a warehouse that is out of stock can often obtain the needed item from a neighboring warehouse that still has the item on-hand (see e.g. Kranenburg and Van Houtum, 2009). We do not allow this option for non-premium requests to avoid that such a lateral transshipment depletes stock that could have been used for meeting a premium request arriving just a bit later. Demand that cannot be met from on-hand stock (either directly or through lateral transshipments) is either backordered or satisfied using emergency shipments from a location with infinite supply. Throughout the chapter, we use the terms “lateral transshipments” and “transshipments” interchangeably.

To our knowledge, lateral transshipments have only been considered as a service differentiation tool at an operational level, i.e., with stock levels given as input. Furthermore, contributions in this area – Jalil (2011) and Tiemessen et al. (2013) – are limited to single-item models. In this chapter, in contrast, we consider a multi-item model at a tactical level (i.e. we consider stock levels as decision variables) where lateral transshipments may only be used to satisfy premium customer requests. The use of lateral transshipments for differentiation likely has significant added value: these shipments are generally both faster and less expensive than emergency shipments. Hence, if there is added value to using selective emergency shipments for differentiation, it will likely be beneficial to use selective transshipments for this purpose as well. Furthermore, such a form of stock pooling can also result in lower overall stock levels in the supply chain (Paterson et al., 2011). A complication, however, is that the feasibility of a lateral transshipment depends on the stock levels at other warehouses, whereas emergency shipments are always possible if the central warehouse is assumed to have infinite supply (as is the case in most literature as well as in the previous chapter). Hence, we investigate under what conditions lateral transshipments are beneficial and how often each shipment option (i.e. lateral transshipments, emergency shipments, backordering) is used in a multi-item setting. To do so,

[6] This chapter is based on the working paper “Service differentiation through selective lateral transshipments” by E. M. Alvarez, M. C. van der Heijden, I.M.H. Vliegen, and W. H. M. Zijm.

we require a single-item building block that has not been considered in literature so far, namely an approach to analyze the model when transshipments are used for premium customers only, both when unmet demand is backordered and when (some) unmet demand is satisfied using emergency shipments. Furthermore, as we found large savings when combining selective emergency shipments with critical level policies, we also combine selective emergency shipments *and* selective transshipments with critical level policies. Our detailed contributions are:

1. For a system where multiple warehouses each receive requests from two customer classes, we give an analysis approach for a single item under lateral transshipments for premium customers. We also extend this approach for the combination with critical levels.
2. We develop an optimization approach similar to Dantzig-Wolfe decomposition for the overall multi-item model and show that this approach is fast and gives good quality solutions. Although such an approach has been used for solving similar problems before, amongst others in the previous chapter, its application is not straightforward for our problem, since we have a large number of control options.
3. In an extensive computational experiment, we show (i) that there is significant added value to using selective transshipments in addition to selective emergency shipments, especially in settings with slow moving items and (ii) that the combination of selective transshipments and selective emergency shipments is a good alternative to using critical level policies.

The chapter is structured as follows: we describe the system in Section 4.2 and present in Section 4.3 an analysis approach for this system when transshipments are used for premium requests. This single-item analysis approach serves as a building block for solving the multi-item optimization problem that we address in Section 4.4, where we also present the solution approach. We then discuss extensions to a model where a critical level policy is combined with selective transshipments and emergency shipments in Section 4.5. In Section 4.6, we discuss our extensive computational experiment. Finally, we draw conclusions in Section 4.7.

4.2 Model

4.2.1 Model outline

We consider a multi-item network of multiple local warehouses and a central depot with infinite supply. Each warehouse has its own customer base consisting of 2 customer classes, specifically premium and non-premium customers. Per customer class, there is a maximum amount of time that customers of that class are willing to wait on average for parts. Naturally, the premium class has the strictest waiting time requirement. Direct requests at a warehouse (i.e. from its own customer base) are met from stock at the warehouse if possible, with a replenishment request being sent to the central depot (i.e. we consider a continuous-time, one-for-one replenishment policy).

A *premium* customer request that cannot be met from on-hand stock may be satisfied through a *lateral transshipment* from another warehouse. We consider a model where transshipments are only used for a subset of warehouses and items, with the selection of the most appropriate subset being part of the multi-item optimization problem (Section 4.4). If transshipments of a specific item are *not* allowed at a warehouse, that warehouse can neither request the item at another warehouse nor receive transshipment requests. Not allowing transshipments may be justified if a warehouse is far away from other warehouses or if an inexpensive fast moving item is considered (for which a lateral transshipment is relatively expensive). In contrast, if transshipments are allowed, the warehouse can both send and receive transshipment requests. On-hand stock is always used to satisfy an incoming transshipment request, i.e. no stock is reserved for direct requests. A warehouse issues transshipment requests to other warehouses in a predetermined order of the warehouses. Such an order is common in practice and will depend on shipment times and costs among warehouses.

If a request cannot be met either from stock at the direct warehouse or through a lateral transshipment, it is either *backordered* or met using an *emergency shipment* from the central depot. Based on these shipment options, we consider the following three shipment strategies:

1. *Full backordering:* Premium and non-premium requests are backordered, with backorders cleared *first-come-first-served*. Premium backorders thus do not receive higher priority. Notice that this clearing strategy differs from the priority clearing strategy used in Chapter 3 for the selective emergency shipments model.
2. *Emergency shipments for premium customers only:* we backorder non-premium requests.
3. *Emergency shipments for all customers.*

We do not allow premium requests to be backordered when non-premium requests are met through emergency shipments. Notice that the third shipment strategy allows emergency shipments for non-premium customers, while cheaper lateral transshipments are not allowed for this customer class. Our reason for still considering this strategy lies in the availability of the various alternatives: emergency shipments can always be performed due to the infinite capacity of the central depot: it will never prevent a premium request arriving a bit later from being filled. In contrast, since the feasibility of a lateral transshipment depends on the stock levels at other warehouses, the use of this shipment mode for a non-premium request could ensure that a premium request arriving at a later moment might *not* be satisfied. For the non-premium class, we therefore only consider emergency shipments as a fast shipment option.

The shipment strategy may vary per item and warehouse. In Chapter 3, we have shown that the suitability of a shipment strategy depends on the characteristics of the item: full backordering is particularly beneficial for relatively inexpensive items with high demand rates, while emergency shipments are more suitable for premium requests for expensive slow moving items. As demand rates differ per warehouse, the shipment strategy may also vary among warehouses. The lateral and emergency shipment times do not have a specific distribution: we only use the mean shipment times in our model.

Figure 4.1 shows a single-item three-warehouse example where transshipments are only allowed among warehouses 1 and 2. The shipment strategies differ per warehouse: warehouse 2 uses full backordering, warehouse 1 uses emergency shipments for premium requests only, and warehouse 3 uses emergency shipments for all requests.

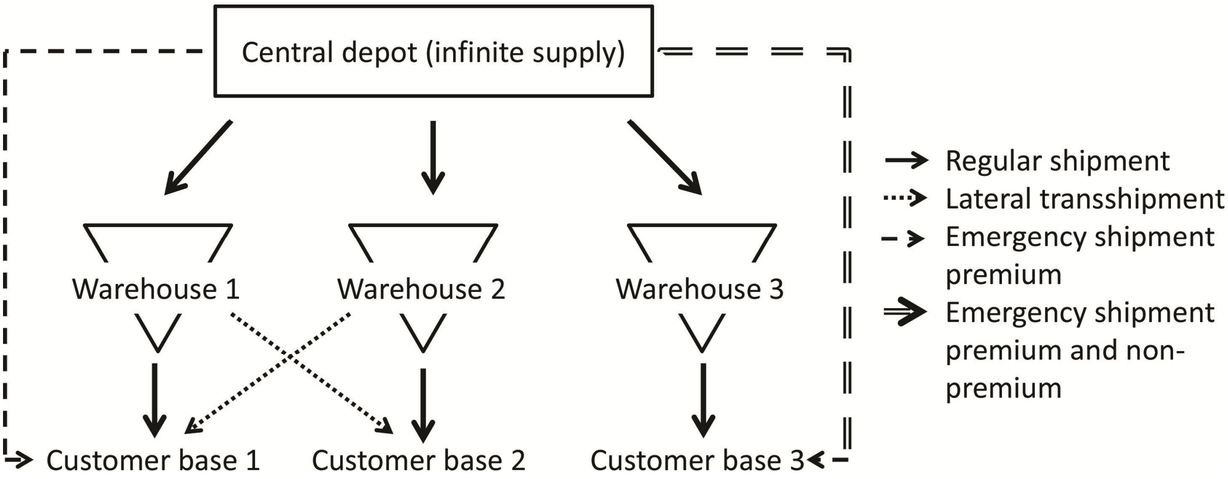


Figure 4.1 Example system with 3 warehouses.

4.2.2 Key assumptions and notation

We make the following assumptions:

- All direct requests arrive according to mutually independent Poisson processes.
- The replenishment lead time to any warehouse is exponentially distributed. This assumption facilitates system analysis using continuous-time Markov chains. The system performance measures also tend to be insensitive to the lead time distribution, especially when emergency shipments are used for both classes (see e.g. Alfredsson and Verrijdt (1999) or Chapter 3).
- The shipment time from any warehouse to a customer is negligible.
- Lateral transshipments are faster than emergency shipments and also have lower shipment costs. As a result, they are preferred over emergency shipments.

- Lateral and emergency shipments are sent directly to the customer and do not pass the warehouse first.
- Emergency shipment requests originate from the warehouse that needs the item: a second warehouse cannot request the item and then forward it to the warehouse that actually needs it.

We have K warehouses that each receive requests for I items from class 1 (premium) customers and class 2 (non-premium) customers. On average, class j customers ($j = 1,2$) are willing to wait at most W_j^{max} time units for spares. Direct requests for item i ($i = 1,..,I$) from class j customers at warehouse k ($k = 1, \dots, K$) occur at rate m_{ijk}, with $M_{jk} = \sum_{i=1}^{I} m_{ijk}$ denoting the total direct demand from class j customers arriving at warehouse k and $M_k = M_{1k} + M_{2k}$ denoting the total direct demand arriving at warehouse k. The mean replenishment lead time of item i to warehouse k is denoted by T_{ik}^{reg}, the emergency time by T_{ik}^{em}, and the lateral transshipment time from warehouse l by T_{ilk}^{lat} (with $T_{ilk}^{lat} \leq T_{ik}^{em} \leq T_{ik}^{reg}$ $\forall i, l, k$). Warehouse k issues transshipment requests to other warehouses in the order specified by $\sigma_k = \{\sigma_k(1), \dots, \sigma_k(K-1)\}$, with $\sigma_k(n)$ being the n-th warehouse in the sequence. Note that σ_k is the same for all items, as the order will only depend on the shipment distances and costs among warehouses. Also, σ_k only indicates the order in which we try to find a transshipment source. Whether a warehouse can actually serve as a transshipment source for warehouse k also depends on the decision whether transshipments are allowed from that warehouse, and on the available stock at the time of a request. The holding cost parameter h_i denotes the item i unit costs per time unit, identical for all warehouses. Emergency and lateral shipments of item i to warehouse k occur at additional costs C_{ik}^{em} and C_{ilk}^{lat} over the costs of a regular replenishments, with l denoting the warehouse sourcing the item. We assume that $C_{ilk}^{lat} \leq C_{ik}^{em}, \forall l$, as this generally holds in practice.

We have three decision variables for each combination of item i and warehouse k, i.e. (i) the base stock level S_{ik}, (ii) the lateral transshipment strategy L_{ik} denoting whether transshipments are allowed for that item and warehouse (L_{ik} then equals 1), and (iii) the shipment strategy D_{ik} which denotes the highest customer class for which emergency shipments are used. In a setting with two customer classes, D_{ik} can take on three values: 0 (full backordering), 1 (emergency shipments for premium customers only), and 2 (emergency shipments for all customers). On a system level, the variables are denoted by vectors $\boldsymbol{S}_i = [S_{i1}, \dots, S_{iK}]$**,** $\boldsymbol{L}_i = [L_{i1}, \dots, L_{iK}]$ and $\boldsymbol{D}_i = [D_{i1}, \dots, D_{iK}]$. We aggregate all variables in an *item policy* $(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i)$. As performance measures, we have $EW_{ijk}(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i)$, the expected class-$j$ waiting time for item i at warehouse k, and $TC_{ik}(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i)$, the total relevant costs for item i at warehouse k. The relevant costs consist of holding costs and extra costs for lateral and emergency shipments over regular replenishments.

4.3 Analysis

4.3.1 Approach

In this section, we focus on the special case where transshipments are allowed among all warehouses (i.e. $\boldsymbol{L}_i = [1, \ldots, 1]$). The analysis under alternative values for $\boldsymbol{L}_i$ is straightforward: if $L_{ik} = 0$, warehouse k can be analyzed individually, as it does not send or receive transshipments of item i. An exact analysis with continuous-time Markov chains is intractable for more than 2 warehouses: we have to keep track of the inventory level at each warehouse separately to determine when transshipments are needed and where stocks are available. Solving such a Markov chain leads to very large computation times for systems with many warehouses. Therefore, we use a decomposition approach in which we analyze each warehouse separately and iteratively update the demand rates arising from lateral transshipments. Such an approach has led to accurate results for related models (Axsäter (1990), Alfredsson and Verrijdt (1999), and Van Wijk et al. (2012)). A key approximation in this decomposition approach is that transshipment requests arrive according to *Poisson processes* with a known rate. Then, each warehouse can be analyzed separately, resulting in fill rate estimates. In turn, the fill rates at all warehouses determine the rate at which transshipment requests occur. Using a similar rationale, we develop an iterative procedure to analyze a system where lateral transshipments are only possible for a subset of all customers. We also assume that all warehouses operate independently of each other, which allows us to compute, amongst others, the fraction of demand met through transshipments as simple products of warehouse fill rates. Obviously, this assumption does not hold in reality. We include these dependencies to some extent by iteratively updating the transshipment rates among warehouses.

Section 4.3.2 lists further notation for computing $EW_{ijk}(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i)$ and $TC_{ik}(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i)$. Section 4.3.3 gives the main analysis steps, and Section 4.3.4 details the analysis of a warehouse. Section 4.3.5 evaluates the approach performance. We omit suffix i, as we consider a single item only.

4.3.2 Additional notation

We introduce the notation below, which applies for each warehouse k and customer class j (when applicable). The term 'demand at warehouse k' refers to the direct demand at that warehouse.

- e_k: the rate at which transshipment requests arrive. We use e_k to analyze the warehouses and estimate the system performance measures.
- $\beta_{jk}(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$: the fill rate (i.e. the fraction of demand met directly from stock).

- $\alpha_{jlk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$: the fraction of demand met through transshipments from a warehouse l, with $\alpha_{2lk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = 0$ (transshipments are not used for non-premium customers). Also, $\alpha_{1lk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = 0$ when L_l or L_k equals 0: then, no transshipments are sourced from l.
- $\gamma_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$: the fraction of demand met through emergency shipments, with $\gamma_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = 0$ if $j > D_k$ (then, emergency shipments are not allowed for that class).
- $EBO_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$: the mean backorder level, with $EBO_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = 0$ if $j \le D_k$.

Using these performance measures, we find $EW_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ and $TC_k(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ as follows:

$$EW_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = EBO_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})/m_{jk} + \gamma_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})T_k^{em} + \sum_{l\in\sigma_k} \alpha_{jlk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})T_{lk}^{lat} \tag{4.1}$$

$$TC_k(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = hS_k + \sum_{l\in\sigma_k} \alpha_{1lk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})m_{1k}C_{lk}^{lat} + \sum_{j=1}^{2} \gamma_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})m_{jk}C_k^{em} \tag{4.2}$$

The first term of $EW_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ arises from backordering (using Little's formula), whereas the second and third term denote the waiting time arising from emergency and lateral transshipments. Note that $TC_k(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ consists of holding costs (which are computed over both on-hand stock and items in the pipeline), and the costs for using lateral and emergency shipments if applicable.

4.3.3 Main analysis steps

Our main analysis steps are:

1. **Initialization:** $e_k = 0, k = 1..K$, so we initially ignore lateral transshipments.
2. **Warehouse analysis:** Given the current value of e_k, compute fill rates $\beta_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ and the expected number of backorders $EBO_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ for all warehouses k and classes j.
3. **Update** the transshipment rates e_k $\forall k$ given the current values of $\beta_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$
4. **Finish:** Stop if the change in e_k is smaller than some small $\forall k$. Otherwise, go to step 2.

We discuss Step 2 in more detail in Section 4.3.4. We update e_k in step 3 as follows: let e_{kl} denote the rate at which transshipment requests arrive at warehouse k from warehouse l. If $k = \sigma_l(n)$ for any positive integer n, k receives transshipment requests from l when l and all warehouses $\sigma_l(1)$ up to $\sigma_l(n-1)$ are out of stock. Assuming independence among warehouses, we find:

$$e_{kl} = m_{1l}\left(1 - \beta_l(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})\right)\prod_{x=0}^{n-1}\left(1 - \beta_{\sigma_l(x)}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})\right) \tag{4.3}$$

$$e_k = \sum_{l|k\in\sigma_l} e_{kl} \tag{4.4}$$

We obtain $\alpha_{1kl}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ and, if applicable, $\gamma_{jl}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ from equations (4.5) and (4.6) respectively. We find α_{1kl} by multiplying the fraction of premium demand at l forwarded to k (i.e. e_{kl}/m_{1l}) by the probability that this demand can be met from on-hand stock at k (i.e. $\beta_k(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$). Note that $\gamma_{jl}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = 0$ when $j > D_l$, as specified in Section 4.3.2.

$$\alpha_{1kl}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \beta_k(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})e_{kl}/m_{1l} \tag{4.5}$$

$$\beta_l(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) + \gamma_{jl}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) + \sum_{k\in\sigma_l}\alpha_{1kl}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = 1 \tag{4.6}$$

4.3.4 Detailed analysis of a single warehouse

We find the fill rate and the expected number of backorders per class from the distribution of the number of items in the pipeline to the warehouse by using a continuous time Markov chain. For simplicity, we drop index k and denote $\mu = 1/T^{reg}$. Let $\lambda = m_1 + m_2 + e$ denote the demand rate including transshipment requests when the warehouse has stock on-hand, and $\theta(D)$ the demand rate under shipment strategy D when the warehouse is out of stock. We find for $\theta(D)$:

- $\theta(2) = 0$: all demand is lost (i.e. met through lateral or emergency shipments).
- $\theta(1) = m_2$: premium demand is met through lateral or emergency shipments.
- $\theta(0) = \pi_1 m_1 + m_2$: premium requests are backordered when the item cannot be obtained elsewhere in the system, which coincides with all warehouses in σ being out of stock. Hence, the probability π_1 of a premium backorder equals $\prod_{l\in\sigma}\big(1-\beta_l(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})\big)$.

Note that the warehouse never backorders transshipment requests when it is out of stock, i.e., $\theta(D)$ does not contain the transshipment rate e. Figure 4.2 shows the Markov chain depicting the number of items in the pipeline. At S or more items in the pipeline, the arrival rate becomes $\theta(D)$. When $D = 0$, the states $S + i$, $i \geq 1$, have i backorders in total for both class 1 and class 2 requests. The exact disaggregation of these backorders over the classes is not required for estimating the pipeline distribution, as backorders are cleared first-come-first-served.

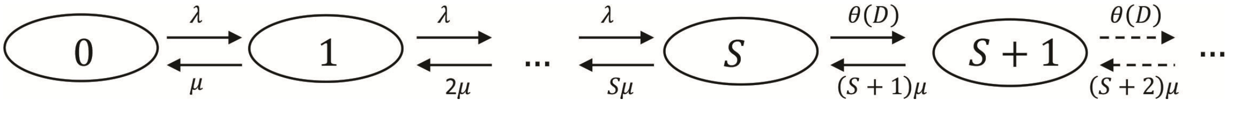


Figure 4.2 Markov chain of the number of outstanding orders at the warehouse under shipment strategy $\boldsymbol{D}$.

Under full emergency shipments ($D = 2$), the Markov chain simplifies to an Erlang loss system with S servers. Using the notation $\rho = \lambda/\mu$, we thus have (see e.g. Gross et al. (2008)):

$$\beta_1(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \beta_2(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = 1 - \frac{\rho^S/S!}{\sum_{w=0}^{S}\rho^w/w!} \tag{4.7}$$

Under (partial) backordering ($D = 0{,}1$), we solve balance equations to find the steady-state probabilities p_n of n outstanding orders. With ρ as before, and ρ_1 equal to $\theta(D)/\mu$, we find:

$$p_0 = \left\{\sum_{w=0}^{S} \frac{\rho^w}{w!} + \left(\frac{\lambda}{\theta(D)}\right)^S \left(e^{\rho_1} - \sum_{w=0}^{S} \frac{\rho_1^w}{w!}\right)\right\}^{-1} \tag{4.8}$$

$$p_n = \rho^{\min\{n,S\}} {\rho_1}^{[n-S]^+} \frac{1}{n!} p_0 \tag{4.9}$$

$$\beta_1(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \beta_2(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \sum_{n=0}^{S-1} p_n \tag{4.10}$$

$$EBO(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \left(\frac{\lambda}{\theta(D)}\right)^S p_0 \left\{\rho_1 \left(e^{\rho_1} - \sum_{n=0}^{S-1} \frac{\rho_1^n}{n!}\right) - S\left(e^{\rho_1} - \sum_{n=0}^{S} \frac{\rho_1^n}{n!}\right)\right\} \tag{4.11}$$

Under partial backordering ($D = 1$), (4.11) denotes the non-premium mean backorder level $EBO_2(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$. Under full backordering ($D = 0$), (4.11) denotes the total mean backorder level. As premium and non-premium backorders occur at rates $\pi_1 m_1$ and m_2 respectively, we have:

$$EBO_1(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \frac{EBO(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})\pi_1 m_1}{\pi_1 m_1 + m_2} \tag{4.12}$$

$$EBO_{2k}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \frac{EBO(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) m_2}{\pi_1 m_1 + m_2} \tag{4.13}$$

4.3.5 Quality of the analysis approach

We evaluate our method by comparison to simulation for three performance measures: $\alpha_{1k}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \sum_{l \in \sigma_k} \alpha_{1lk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$, and $\beta_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ and $EW_{jk}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$, ($j = 1{,}2$). We test 32 problem instances with either 6 or 18 warehouses and transshipments allowed at all warehouses (i.e. $L_k = 1\ \forall k$). For the simulation, we used a replication/deletion approach with at least 0.3 million requests for both premium and non-premium customers (average values are 1 million premium and 5 million non-premium requests). Table 4.1 shows the relative deviation of our method to simulation. The method is very accurate for slow movers and systems with many warehouses. We thus expect the approach to be accurate for practical instances. In systems with 6 warehouses and low stock levels (resulting in fill rates below 50%), the estimate of the transshipment fraction $\alpha_{1k}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ can be poor. Still, in practice it will not occur that stocks of fast movers are low: these items have a high contribution to the overall waiting time. Therefore, waiting times for these items should be low. The maximum computation time for an instance is 12 milliseconds. Clearly, our approach is accurate and requires little computation time. As a result, it will be a suitable building block for solving multi-item problem of the next section. We refer to the appendix at the end of this chapter for more details.

Settings			Average deviation				Maximum deviation			
M_k	K	S_k	β_k	α_k	EW_{1k}	EW_{2k}	β_k	α_k	EW_{1k}	EW_{2k}
0.05	6	1	0%	1%	1%	0%	0%	2%	5%	0%
		2	0%	0%	0%	0%	0%	1%	1%	0%
	18	1	0%	0%	0%	0%	0%	0%	0%	0%
		2	0%	0%	0%	0%	0%	1%	1%	1%
0.5	6	4	2%	8%	4%	1%	6%	20%	9%	5%
		8	0%	1%	1%	0%	0%	1%	1%	1%
	18	4	0%	1%	1%	0%	2%	3%	4%	1%
		8	0%	1%	1%	1%	0%	1%	1%	1%

Table 4.1 Relative deviations of the analysis approach to simulation.

4.4 Problem description and optimization

Problem $(P1)$ minimizes the total system costs $TC(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D})$ under restrictions on the mean aggregate waiting times per customer class *and* warehouse. A high waiting time at one warehouse thus cannot be compensated by a low waiting time at another warehouse, although such a variant (e.g. if a customer can be serviced from multiple warehouses) can be analyzed in a similar way.

$$(P1) \qquad \min TC(\boldsymbol{S},\boldsymbol{L},\boldsymbol{D}) = \sum_{i=1}^{I}\sum_{k=1}^{K} TC_{ik}(\boldsymbol{S}_i,\boldsymbol{L}_i,\boldsymbol{D}_i)$$

$$\text{s.t.} \qquad \sum_{i=1}^{I} \frac{m_{ijk}}{M_{jk}} EW_{ijk}(\boldsymbol{S}_i,\boldsymbol{L}_i,\boldsymbol{D}_i) \leq W_j^{max} \qquad j = 1,2, k = 1,\dots,K \qquad (4.14)$$

$$S_{ik} \in N_0, L_{ik} \in \{0,1\}, D_{ik} \in \{0,1,2\} \qquad (4.15)$$

As in the selective emergency shipment model of Chapter 3, we use an approach similar to Dantzig-Wolfe decomposition to solve $(P1)$. First, we reformulate the non-linear problem $(P1)$ to a linear problem by focusing on *item policies* as decision variables. The reformulated problem becomes to select one item policy for each item such that the system costs are minimized with the waiting time requirements still being met. Let B_i denote the set of item policies for item i, with $b_{ir} = \left(\boldsymbol{S}_i(r), \boldsymbol{L}_i(r), \boldsymbol{D}_i(r)\right)$ denoting a single item policy in B_i (i.e. $b_{ir} \in B_i$, with $r = 1,2,\dots,|B_i|$). Furthermore, let $x_{b_{ir}}$ be a binary variable indicating whether b_{ir} is selected for item i (then, $x_{b_{ir}} = 1$). We then find linear problem $(P2)$:

$$(P2) \qquad \min \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{r=1}^{|B_i|} TC_{ik}(b_{ir})x_{b_{ir}}$$

$$\text{s.t.} \qquad \sum_{i=1}^{I}\sum_{r=1}^{|B_i|}\frac{m_{ijk}}{M_{jk}}EW_{ijk}(b_{ir})x_{b_{ir}} \leq W_j^{max} \qquad j = 1,2, k = 1, \dots, K \qquad (4.16)$$

$$\sum_{r=1}^{|B_i|} x_{b_{ir}} = 1 \qquad i = 1, \dots, I \qquad (4.17)$$

$$x_{b_{ir}} \in \{0,1\} \qquad i = 1, \dots, I, r = 1, \dots, |B_i| \qquad (4.18)$$

If B_i contains all item policies, $(P2)$ and $(P1)$ are equivalent and have the same optimal solution. Also, we find a lower bound on the costs by solving the LP-relaxation of (P2). Our challenge is the selection of policies to include in B_i for each item i, which is far from trivial: contrary to the selective emergency shipment model of the previous chapter, each policy $b_{ir} \in B_i$ now refers to a *multi*-location problem. As we will show, an exact decomposition into single location problems is not possible under lateral transshipments. So we face a large set of relevant item policies. Furthermore, policy evaluation may take significant time when transshipments are allowed. The careful selection of item policies is thus crucial: we should select the minimal number of policies such that we still find a (near-) optimal solution to $(P2)$ and its LP-relaxation.

4.4.1 Solving the LP-relaxation

First, we first construct an initial set of item policies. Subsequently, we iteratively add policies to the policy set using column generation until no further interesting policies can be found.

4.4.1.1 Constructing an initial policy set

An initial policy set should lead to a feasible solution to the *integer* problem $(P2)$. One option to find such a set is to select a policy per item i such that $EW_{ijk}(b_{ir}) \leq W_j^{max}$ for each class j and warehouse k, guaranteeing that $\sum_{i=1}^{I}\frac{m_{ijk}}{M_{jk}}EW_{ijk}(b_{ir}) \leq W_j^{max}$. However, that option may lead to relatively large stock levels. Instead, we look for a policy over all items simultaneously. We use a "biggest-bang-for-the-buck" algorithm, where we satisfy all unmet demand using emergency shipments, i.e. $L_{ik} = 0$ and $D_{ik} = 2$. This is justified since we only need a reasonable feasible solution as starting point for optimization. In each step of our algorithm, we increase the stock level S_{ik} by one unit at the item-warehouse combination (i, k) that leads to the greatest added value. We continue until all waiting time restrictions are met. To choose an option (i, k), we compute the decrease in waiting time relative to the extra investment needed. We find $\Delta W(\boldsymbol{S}_i + U_{ik})$, the decrease in waiting times for a unit stock increase at (i, k) (denoted by $\boldsymbol{S}_i + U_{ik}$), as follows:

$$\Delta W(\boldsymbol{S}_i + U_{ik}) = \sum_{j=1}^{2}\sum_{k=1}^{K}\left\{\left(\sum_{i=1}^{I}\frac{m_{ijk}}{M_{jk}}EW_{ijk}(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i) - W_j^{max}\right)^+ - \left(\sum_{i=1}^{I}\frac{m_{ijk}}{M_{jk}}EW_{ijk}(\boldsymbol{S}_i + U_{ik}, \boldsymbol{L}_i, \boldsymbol{D}_i) - W_j^{max}\right)^+\right\} \tag{4.19}$$

Here $[a]^+ = \max\{0, a\}$, which ensures that we only consider waiting time reductions above their respective thresholds. The extra investment $\Delta TC(\boldsymbol{S}_i + U_{ik}) = TC(\boldsymbol{S}_i + U_{ik}) - TC(\boldsymbol{S}_i)$ follows from (4.2). Note that options (i, k) may exist where both waiting times *and* costs decrease: a stock increase may lead to lower waiting times and fewer transshipments and emergency shipments (resulting in lower shipment costs). Then, we select the option with the largest $\Delta W(\boldsymbol{S}_i + U_{ik})$ among the options with lower costs. Otherwise, we select the option with the largest $\Delta W(\boldsymbol{S}_i + U_{ik})/\Delta TC(\boldsymbol{S}_i + U_{ik})$.

During the procedure, we obtain a new item policy each time we adjust the stock level for one item and warehouse. With the exception of *dominated* policies that have both higher costs and higher waiting times at all warehouses than at least one other policy, each obtained item policy b_{ir} is included in the policy set B_i. This means that the initial policy set might contain more than one policy for each item: we expect that having many policies in the initial policy set reduces the amount of time needed for generating additional policies through column generation.

4.4.1.2 Finding additional policies through column generation

Through column generation, we iteratively add the policy with minimal reduced costs to the policy set if these costs are negative. We stop once we cannot find further policies with negative reduced costs. Let u_{jk} (≤ 0) and v_i (≥ 0) denote the shadow price values for constraints (4.16) and (4.17) respectively, resulting from solving (P2) for a given set of item policies (see Section 1.9.2 for details). The reduced costs $Z_i(b_i)$ for a policy b_i are now found as follows, with suffix r (i.e. the policy index) omitted for simplicity:

$$Z_i(b_i) = Z_i(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i) = \sum_{k=1}^{K}\left\{TC_{ik}(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i) - \sum_{j=1}^{2}u_{jk}\frac{m_{ijk}}{M_{jk}}EW_{ijk}(\boldsymbol{S}_i, \boldsymbol{L}_i, \boldsymbol{D}_i)\right\} - v_i \tag{4.20}$$

Notice that it is far from trivial to find the item policy with the lowest reduced costs for an item i. If transshipments are allowed at a warehouse k, the performance at that warehouse depends on the rates at which it sends and receives transshipment requests. Hence, the optimal values for S_{ik} and D_{ik} depend on the values of the decision variables at the other warehouses where transshipments are allowed. As a result, we can only guarantee optimality if all warehouses are jointly optimized.

For problems of realistic size, however, optimization over all warehouses jointly requires too much time: even for small instances with 10 items and 4 warehouses, the computation times can amount to three days. Instead, we opt for an approximate disaggregation of the overall problem into K single warehouse problems. Specifically, we can optimize the decision variable values at a warehouse k separately, if the decision variable values at the other warehouses are given. Clearly, the choice of variable values at warehouse k will influence the optimal values at other warehouses. Therefore, we iteratively optimize each warehouse until convergence occurs.

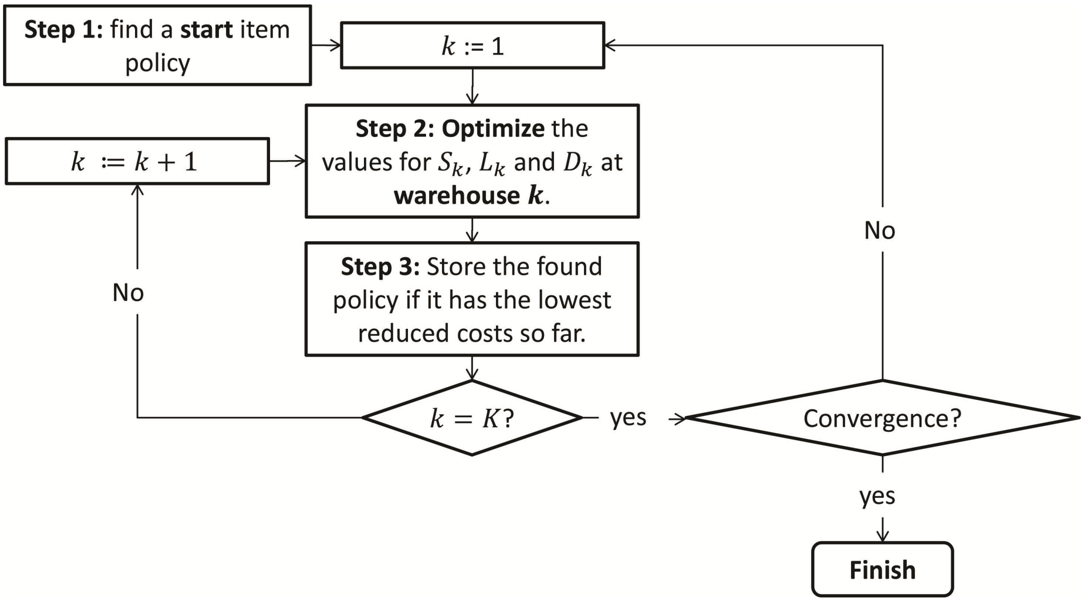

Step 1: find a start item policy
$k := 1$
$k := k + 1$
Step 2: Optimize the values for S_k, L_k and D_k at warehouse k.
Step 3: Store the found policy if it has the lowest reduced costs so far.
No
No
$k = K$?
yes
Convergence?
yes
Finish

Figure 4.3 Column generation approach to find a near-optimal item policy for a particular item.

Figure 4.3 shows the main column generation steps for a single item. We omit suffix i in the figure and the rest of the section. First, we construct a start (i.e. initial) item policy. This policy serves as input for optimizing the decision variables at warehouse 1 a first time (i.e. the decision variable values for warehouses $l > 1$ serve as input for optimizing the values for warehouse 1). Then, we iteratively optimize decision variable values at a warehouse k, with the variable values of warehouses $n \neq k$ fixed to their most recent values. Each time we find a new item policy, we verify whether it has the lowest reduced costs so far and store it if this is the case. In an iteration, all warehouses in the system are considered. Convergence occurs when the decision variable values for all warehouses remain unchanged from one iteration to the next. We now give details on steps 1 and 2, with $(\boldsymbol{S}^*, \boldsymbol{L}^*, \boldsymbol{D}^*)$ being the best item policy found overall.

Step 1: finding a start item policy for the column generation procedure.

We can find a start policy in two extreme ways: either we allow transshipments at all warehouses (i.e. $L_k = 1\ \forall k$) or we do not allow them at any warehouse ($L_k = 0\ \forall k$). The advantage of the second option is that we can easily find good values for the remaining decision variables S_k and D_k, since each warehouse can be optimized separately. On the other hand, the first option will likely result in a more suitable start policy: we expect it to be easiest to move from a policy where transshipments are allowed at all warehouses to one where transshipments are only allowed at a subset of warehouses. In contrast, a move from a policy where transshipments are not used to one where transshipments are allowed can only occur if it is beneficial to transshipment among two or more warehouse (transshipments will not occur if they are only allowed at one warehouse).

These arguments prompt us to combine the options to find a start policy: first, we set $L_k = 0$ and optimize values for S_k and D_k $\forall k$. Then, we set $L_k = 1$ $\forall k$ to obtain the start policy. In this way, we easily find values for S_k and D_k, while still obtaining a start policy where transshipments are allowed among all warehouses. Note that the values found for S_k and D_k result in a valid item policy both when $L_k = 0$ and when $L_k = 1$. Therefore, we analyze the system under both settings and store the policy with the lowest reduced costs $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ as the best policy so far $(\boldsymbol{S}^*, \boldsymbol{L}^*, \boldsymbol{D}^*)$.

Given that $L_k = 0$, we first optimize S_k for each value of $D_k \in \{0,1,2\}$ separately. Subsequently, we select the combination (S_k, D_k) leading to the lowest value for $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$. Given a value for D_k, we start with $S_k = 0$. We then iteratively increase S_k by one unit until a further increase has no benefit. Each time we increase S_k, we store the combination (S_k, D_k) if it leads to the lowest value for $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ so far (denoted by $Z^{min}(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$). A further increase of S_k has no benefit once $h(\sum_{n=1}^{K} S_n + 1) - v \geq Z^{min}(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$. Then, the minimal reduced costs for $S_k + 1$ (consisting of the system holding costs minus the item shadow price) already exceed the best reduced costs found so far. Note that the actual reduced costs for $S_k + 1$ will be larger than that minimum value, as we ignore the shipment and waiting time costs.

Step 2: optimizing decision variable values at warehouse $\boldsymbol{k}$.

Our aim is to find the values for S_k, L_k and D_k that minimize the reduced costs $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ in the entire system. We do so, because the decision variable values at warehouse k influence the service levels at all warehouses. This influence can be significant: in particular, if stock is mainly (or even only) kept at warehouse k, the value of L_k is crucial, since it influences whether other warehouses have access to this stock.

We first optimize the decision variable values at k for each value of L_k separately. We then select the combination (S_k, L_k, D_k) that minimizes $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$. Note that when $L_k = 0$, the optimal values for S_k and D_k are the same as those found when looking for the start item policy (step 1), as the warehouse is not influenced by transshipment requests from other warehouses. When $L_k = 1$, we use the approach described in step 1 to find optimal values for S_k and D_k.

Given values for S_k, L_k and D_k, we have two options for estimating $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$. The first option is to use the full analysis approach of Section 4.3. This option gives the most accurate estimate of $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$, but it is also time-consuming, especially since we need to analyze the system for various combinations of S_k, L_k and D_k. The second option is to use a partial analysis approach for optimization – where we only analyze warehouse k (as in Section 4.3.4) and update the estimates of $\alpha_{jl}(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ and $\gamma_{jl}(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ for the other warehouses $l \neq k$ in the system through equations (4.5) and (4.6) – and subsequently use the full analysis approach to estimate $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$. This option is much faster than the first option (we now only need to analyze one warehouse at a time during optimization), while it still gives sufficiently good solutions, as we will show in Section 4.4.1.3. Furthermore, we still find an accurate value for $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ for the chosen values of S_k, L_k and D_k. Therefore, we use this option for optimizing the decision variables values and estimating $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$.

After optimization, we determine whether the newfound policy is the best so far (i.e., step 3) and store it if this is the case.

4.4.1.3 Quality of the obtained lower bound

We cannot guarantee that our column generation procedure always finds the item policy with the lowest reduced costs. Hence, we cannot ensure that we find the true optimal solution to the LP-relaxation of $(P2)$. Therefore, we compare the lower bound found with our column generation procedure to the lower bound when using an optimal column generation approach based on complete enumeration. As the latter approach is time-consuming, we only test small problem instances. We tested 128 problem instances, each with 5 or 10 items, and 2 or 4 warehouses. The remaining parameter values have been marked by an asterisk in Table 4.3 (Section 4.6.1).

I	K	**Relative deviation to true LB**	
		Average	Maximum
5	2	0.24%	2.23%
	4	0.13%	1.23%
10	2	0.24%	2.29%
	4	0.06%	0.46%

Table 4.2 Relative deviation to the true lower bound.

From Table 4.2, we see that our approach indeed does not always find the true lower bound. Still, the deviation is at most 2.29%. Also, the deviations decrease in the number of warehouses and items, with the deviation being at most 0.46% for instances with 10 items and 4 warehouses. We thus expect the lower bound estimate to be accurate for larger instances that occur in practice.

4.4.2 Finding a near-optimal integer solution

The solution to the LP-relaxation might be fractional, with a combination of item policies being selected for certain items. Therefore, we need an approach to find a near-optimal solution to the integer problem $(P2)$. A simple option would be the intelligent rounding of the fractional values $x_{b_{ir}}$ of the LP-relaxation solution. However, such rounding will not be trivial, as we can have many items for which multiple policies are used: $(P2)$ has $2K + I$ constraints, leading to $2K + I$ item policies b_{ir} being basis variables (i.e. where $x_{b_{ir}} > 0$). For each item, at least one policy will be selected. We thus can have up to $2K$ items for which multiple policies are selected. Also, it is generally known that intelligent rounding does not give good results in 0-1 (i.e., binary) integer programming. Furthermore, even if rounding is used to find a starting point for a local search procedure, the resulting solution is usually inferior to that obtained when solving the integer problem using linear optimization software such as CPLEX, as shown in the previous chapter. Therefore, we also solve $(P2)$ using CPLEX.

The policy set used for solving the LP-relaxation serves as a starting point for the integer problem policy set, as this set has worked well in the previous chapter. From the LP-relaxation set, we remove all dominated policies (i.e. policies with both higher costs and waiting times than at least one other policy) and all policies b_i where $\frac{m_{ijk}}{M_{jk}} EW_{ijk}(b_{ir})$ exceeds W_j^{max} for at least one item i and warehouse k (the aggregate waiting time $\sum_{i=1}^{I} \sum_{r=1}^{|B_i|} \frac{m_{ijk}}{M_{jk}} EW_{ijk}(b_{ir}) x_{b_{ir}}$ then also exceeds W_j^{max}).

Still, computation times remain extensive under this smaller policy set and can amount to several hours. To decrease computation times, we consider two options, namely (i) further reducing the number of item policies per item or (ii) setting a limit on the time for CPLEX to find a solution. We choose for option (ii) because computation times remained large under option (i), irrespective of the criterion used for removing item policies (e.g. when reduced costs of a policy exceed a certain threshold). Also, the solutions found could be very poor, such as a gap to the lower bound of 14%. Option (ii) outperformed option (i) both in solution quality and computation times. The reason is that CPLEX often finds a good solution in the first few minutes, with improvements being minor from then on. Most time is spent on evaluating options that turn out to be infeasible. In an experiment with 80 problem instances – with 20 to 50 items and 10 to 20 warehouses – we considered time limits from 15 to 60 minutes. We found that a limit of 15 minutes already works well, with an average gap to the lower bound of 0.85%. Further improvements in quality were negligible under larger time limits (e.g., under a limit of 60 minutes the average gap reduced to 0.84%).

4.5 Extension to a model with critical levels

We now extend the model of Section 4.2 to include positive *critical levels*, i.e. where an amount of stock can be reserved for premium requests (either direct or transshipment requests). We let C_{ik} denote the critical level for item i at warehouse k, with $\boldsymbol{C}_i = [C_{i1}, \dots, C_{iK}]$ denoting the system critical levels. As before, premium requests may be met through transshipments when the direct warehouse is out of stock. However, warehouses with positive critical levels must always use emergency shipments to satisfy all (premium and non-premium) requests that cannot be satisfied through stock or transshipments. In other words: $D_{ik} = 2$ if $C_{ik} > 0$. We choose this model for its simplicity: as we shall see, the combination of critical levels with emergency shipments leads to a simple analysis of a warehouse. Also, it remains a realistic model: critical levels are mainly beneficial for expensive slow movers, as found both in the previous chapter and by Kranenburg and Van Houtum (2008). For such items, all unmet demand is generally satisfied through emergency shipments.

We can easily extend the analysis and optimization approaches for the combined model. In the analysis approach, the main steps and the computation of the transshipment rates (Section 4.3.3) remain the same. We must only be able to analyze a single warehouse under a critical level policy with emergency shipments. For the optimization procedure, we require a slight modification to the column generation method. Specifically, we must be able to optimize decision variable values – including the critical level – at a single warehouse. We discuss the warehouse analysis in Section 4.5.1 and the optimization in Section 4.5.2. For simplicity, suffixes i and k are omitted.

4.5.1 Warehouse analysis

Figure 4.4 shows the Markov chain of the number of outstanding orders, with λ and μ as in Section 4.3.4. Non-premium demand is lost once $S - C$ orders or more are outstanding (equivalent to having at most C items on-hand). This Markov chain is similar to that of Kranenburg and Van Houtum (2008) (the difference is that they do not consider transshipments, with e thus being zero).

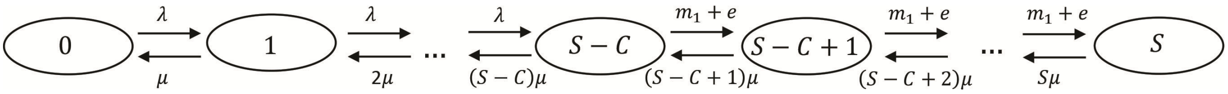


Figure 4.4 Markov chain of the pipeline at the local warehouse under a critical level policy with emergency shipments.

The steady-state probabilities p_n and fill rate values β_j ($j = 1{,}2$) follow directly from the balance equations. We refer to Kranenburg and Van Houtum (2008) for further details.

4.5.2 Warehouse optimization

Optimization occurs in a way similar to the procedure in Section 4.4.1.2, except that we must also optimize C for any combination (S, L, D) when $D = 2$. Given values for S and L, and $D = 2$, we find an optimal value for C as follows: starting with $C = 0$, we iteratively increase C by one unit, with the value leading to the lowest $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D}, \boldsymbol{C})$ being stored. We keep increasing C until either (i) $C = S$ (we can reserve at most the base stock level) or (ii) $\beta_1(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D}, \boldsymbol{C}) \geq 1 - \varepsilon$, with

ε a specified tolerance. As C increases, the service level at premium customers improves (leading to lower reduced costs for those customers) at the expense of non-premium customers (for whom we find higher reduced costs). Overall, $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D}, \boldsymbol{C})$ will thus first decrease and then increase. Still, we are unable to prove convexity of $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D}, \boldsymbol{C})$ in C. However, once $\beta_1(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D}, \boldsymbol{C})$ is close to 1, we can be certain that the reduced costs for premium customers will barely decrease further.

As before (see step 2 of Section 4.4.1.2), we have two options for estimating $Z(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D}, \boldsymbol{C})$ for given values of S, L, D and C, i.e. (i) a more accurate but time-consuming option of analyzing the entire system, and (ii) a faster but less accurate option of analyzing a single warehouse and only updating the values of $\alpha_j(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ and $\gamma_j(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{D})$ for the other warehouses. We use the first option for the model with critical levels only. This model serves as a benchmark for evaluating the model with lateral transshipments and emergency shipments as the only differentiation tools. The first evaluation option results in a stronger benchmark, as it generally gives better solutions.

4.6 Computational experiment

In an extensive computational experiment, we investigate (i) the performance of our optimization approach in terms of solution quality and computation time, (ii) the added value of the selective transshipment approach by comparing it to alternative differentiation approaches, and (iii) the suitability of the various shipment and transshipment strategies.

4.6.1 Experiment design

We construct 1024 problem instances, with T_{ilk}^{lat} always equal to 1 day and C_{ik}^{em} equal to 1000. Table 4.3 gives the other parameter values. The asterisks specify the values considered when evaluating the quality of our lower bound estimate (Section 4.4.1.3). Shipment times and costs are the same for all items and warehouses in a problem instance, with the lateral times and costs equal for any warehouse pair. Using a uniform distribution, the holding costs h_i are randomly drawn on the specified interval. Below, we detail how we obtain values for demand rates m_{ijk}.

	Parameter	Value
1	I	$20, 50$
2	K	$10, 20$
3	T_{ik}^{reg} (days)	$8^*, 16^*$
4	T_{ik}^{em} (days)	$2^*, 4^*$
5	$[W_1^{max}; W_2^{max}]$ (hours)	$[0.5; 2]^*$, $[3; 24]^*$
6	C_{ilk}^{lat}	$100^*, 500$
7	Avg. M_{ik} – interval (p. day)	$[0.002; 0.05]^*$,$[0.002; 0.5]^*$
8	Avg. fraction premium $frac_p$	$0.2^*, 0.5$
9	h_i – interval (p. day)	$[0.1; 10]^*$,$[0.1; 50]^*$

Table 4.3 Tested parameter values.

Our demand rates m_{ijk} should differ among warehouses *and* items, with the *overall* fraction of premium demand in the system equal to $frac_p$. We find m_{ijk} in three steps: first, (1) we draw a value on the M_{ik}–interval (using a uniform distribution) to obtain the average demand rate for item i at one warehouse. By multiplying this value by K we find the total *system* demand rate M_i. Then, (2) we find the total premium demand in the system M_i^p by multiplying M_i by $frac_p$, with M_i^n denoting the remaining non-premium demand. Finally, (3) we disaggregate M_i^p and M_i^n over the warehouses to obtain m_{ijk}. Each warehouse is assigned a fraction of M_i^p and M_i^n (using a normal distribution), with normalization ensuring that $\sum_{k=1}^{K} m_{i1k} = M_i^p$ and $\sum_{k=1}^{K} m_{i2k} = M_i^n$.

The parameter values used by Kranenburg and Van Houtum (2008, 2009) formed the basis for our values, as their values are based on practice. We consider items that have both high and low values, and high and low demand rates. The annual demand rates vary between 0.7 units and 183 units. In practice, an item's annual holding cost is a fraction (roughly 25%) of its value. In this study, we thus consider item values between 146 and 73000 euro's.

For simplicity, a warehouse k sends transshipment requests to other warehouses in the same order in all problem instances: $\sigma_k = \{k + 1, k + 2, \dots, K, 1,2, \dots\}$. So, if warehouse k is out of stock, it first requests an item at warehouse $k + 1$, then at warehouse $k + 2$, etc. We consider this order to balance transshipment streams among warehouses such that no warehouse is always the first to receive transshipment requests from the other warehouses. In practice, the sequence σ_k will depend on the shipment costs and distances among the warehouses.

For each combination of parameters in Table 4.3, we construct 2 sets of item demand rates and holding costs to ensure that our results are not dependent on the specific values of one sample. Combined with $2^9 = 512$ possible parameter combinations, we thus have 1024 instances in total.

4.6.2 Performance of the optimization procedure

Table 4.4 shows the solution quality – expressed as a relative gap to the lower bound estimate – and computation times of the optimization procedure. We used a Dell optiplex 760 with Intel quad core 2.83 GHz processor. Overall, the relative gap is 0.8% on average, with a maximum of 5.5%. The average and maximum gap decrease greatly as the number of items increases. We therefore expect the approach to work very well in realistic settings with many items. The average instance computation time is 12 minutes, with the maximum being 34 minutes. The computation time mainly increases with the number of items and warehouses in an instance.

Parameter	Values	Gap to lower bound estimate (%)		Computation time (min.)	
		Average	Maximum	Average	Maximum
I	20	1.3	5.5	7	21
	50	0.3	1.3	17	34
K	10	0.6	2.9	7	16
	20	1.0	5.5	16	34
Grand Total		**0.8**	**5.5**	**12**	**34**

Table 4.4 Solution quality and computation times of optimization procedure.

4.6.3 Comparison to alternative differentiation approaches

We compare the selective transshipment model (ST_SES) to two alternatives:

1. A **selective emergency shipment model (SES)**, which is the special case of ST_SES with transshipments not allowed for any item or warehouse
2. The **selective transshipment model with critical levels (CLP_ST_SES)** of Section 4.5.

The added value of both ST_SES and CLP_ST_SES is expressed in terms of relative cost savings to SES, shown in Table 4.5. Notice that ST_SES has significant savings compared to SES: the average savings are 14% and can amount up to 34%. The savings are particularly large for instances with many slow moving items; for fast movers, lateral transshipments are not beneficial, as we will see in Section 4.6.4. Savings are also large when emergency shipment times are large and waiting times are not very strict, although the influence of these parameters is mainly large in settings with expensive slow movers.

Parameter	Values	Average savings over SES		Maximum savings over SES	
		ST_SES	CLP_ST_SES	ST_SES	CLP_ST_SES
T_{ik}^{em}	2	12%	12%	28%	28%
	4	17%	17%	34%	35%
$[W_1^{max}; W_2^{max}]$	$[0.5; 2]$	11%	11%	19%	19%
	$[3; 24]$	18%	19%	34%	35%
Max. M_{ik}	0.05	19%	20%	34%	35%
	0.5	9%	10%	20%	20%
Grand Total		**14%**	**15%**	**34%**	**35%**

Table 4.5 Relative savings of ST_SES and CLP_ST_SES over SES.

The savings of CLP_ST_SES are similar to those of ST_SES. Clearly, there is little benefit to also allowing stock reservation for premium customers. The reason for this is that ST_SES is already able to differentiate very effectively: the aggregate waiting times per class j are close to their thresholds W_j^{max}. Adding critical levels therefore does not lead to extra savings.

4.6.4 Suitability of shipment and transshipment strategies

For each combination (L_{ik}, D_{ik}), Figure 4.5 shows the overall fraction of items and warehouses for which that combination is used. Clearly, lateral supply is very suitable for meeting premium requests: overall, transshipments are allowed at 96% of all item-warehouse combinations. For the remaining 4% of combinations where transshipments are not allowed, we always use full backordering. This is logical: if it is not beneficial to use lateral transshipments for premium customers, it will also not be beneficial to use more expensive (and slower) emergency shipments for this class (or thus for the non-premium class). We can thus limit the combinations (L_{ik}, D_{ik}) that we should consider during optimization. The instances where lateral transshipments are not beneficial have many inexpensive fast moving items, high transshipment costs and loose waiting time restrictions, making transshipments expensive and unnecessary.

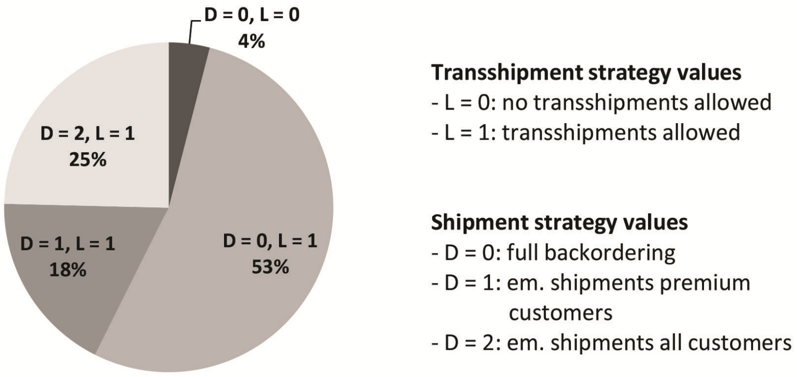


Figure 4.5 The fraction of items and warehouses using a particular (trans-)shipment combination.

Overall, full backordering ($D = 0$) is the most frequently used shipment strategy (see Figure 4.5). Still, the added value of each shipment strategy depends heavily on the shipment times and type of item, as shown in Figure 4.6. Full backordering ($D = 0$) is especially beneficial when emergency shipments are slow relative to regular shipments, and when items are mostly cheap fast movers. Then, that strategy is used for roughly 85% of all items and warehouses. This coincides with the findings of the previous chapter. Clearly, it is beneficial to consider backordering in addition to emergency shipments, even though it is common in both literature and business for emergency shipments to be the only shipment mode.

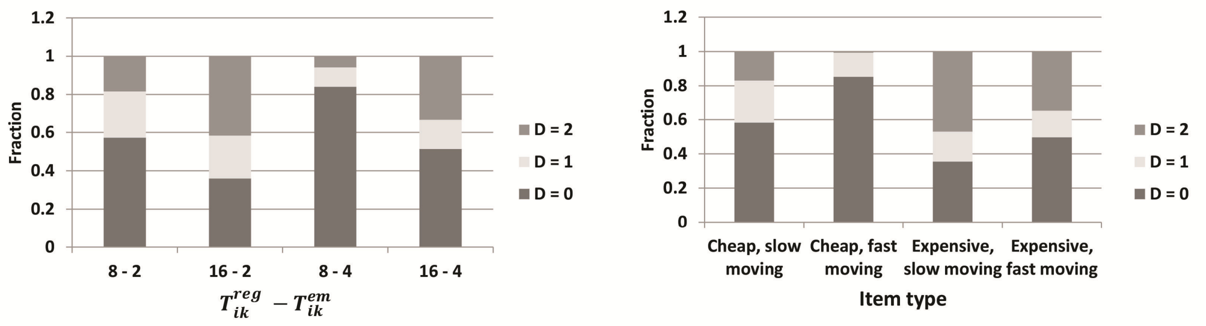


Figure 4.6 The influence of shipment times (left) and item type (right) on the use of various shipment strategies.

Figure 4.7 shows for various problem instances how the strategies (L_{ik}, D_{ik}) are distributed over the items in each instance. We focus on instances with an M_{ik}-interval of $[0.002; 0.5]$ and a holding cost interval of $[0.1; 50]$; the results are similar for other parameter values. As expected, neither lateral transshipments nor emergency shipments are used for inexpensive fast movers, with both transshipments and (partial) emergency shipments used for expensive slow movers.

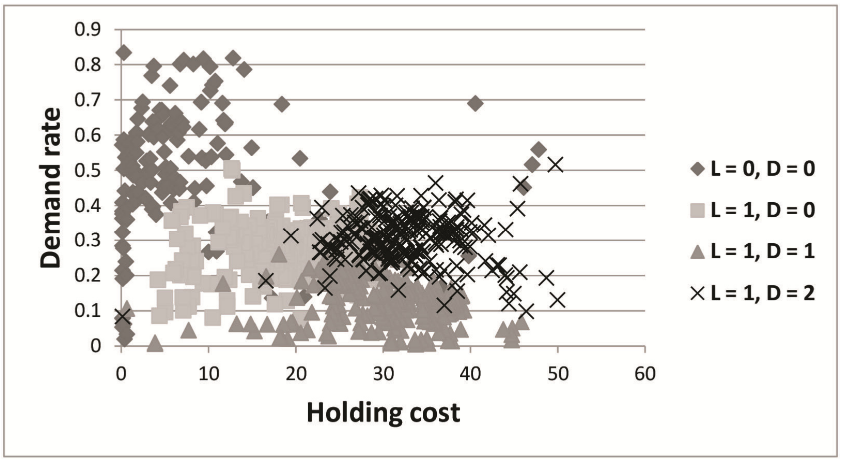


Figure 4.7 Item characteristics per (trans-)shipment strategy.

4.7 Conclusions

We considered a system with two customer classes where lateral transshipments and emergency shipments are both used selectively for service differentiation purposes. For a single-item setting, we developed an analysis approach when selective transshipments may only be used for premium requests. We also developed an approach similar to Dantzig-Wolfe decomposition to optimize the multi-item system under class-specific waiting time restrictions. Key conclusions are:

- Our analysis approach is accurate and fast.
- Our multi-item solution approach gives near-optimal solutions in little computation time.
- Selective lateral transshipments lead to significant cost savings when combined with selective emergency shipments. The savings are 14% on average and can amount to 34%. The savings can be particularly large (19% on average) if we have many expensive slow movers.
- Using critical levels besides selective (trans-)shipments does not lead to significant extra gains. Clearly, the combination of selective transshipments and emergency shipments is a good alternative to critical level policies. Furthermore, the former combination is also easier to implement in practice.
- Backordering should also be considered as a shipment option in spare parts settings. This is in contrast to the practice of always using emergency shipments for unmet demand.

From the findings in this and the previous chapter, our conjecture is that significant cost savings can be obtained by using any two differentiation tools jointly, such as critical levels and selective emergency shipments (Chapter 3) or selective transshipments and selective emergency shipments. Combining three differentiation tools does not lead to additional benefits, but clearly “two out of three (options) is not bad”. This flexibility to choose differentiation tools allows service providers to select those tools that are easiest to implement (with critical level policies possibly not being used in favor of options with fewer practical drawbacks).

So far, we have applied service differentiation by considering multiple shipment options (i.e. lateral transshipments, emergency shipments, and backordering). In the subsequent chapters, we consider an alternative differentiation tool for spare parts supply, namely the use of dedicated customer stocks. In Chapter 5, we first present an analysis approach for a two-echelon system under lost sales. We require such an approach to analyze a system under dedicated stocks with emergency shipments. The analysis approach of Chapter 5 serves as a building block in Chapter 6, where we discuss multi-item optimization under dedicated stocks. There, we also evaluate the added value of using dedicated stocks for service differentiation.

Appendix: detailed performance analysis approach

We now give details on the comparison of our analysis approach to simulation from Section 4.3.5. Table A1 shows the parameter values tested. In all instances, the shipment strategies are spread evenly over the warehouses, i.e. one third of all warehouses uses full backordering, one third uses emergency shipments for premium customers only, etc. The demand rates and shipment times are the same at all warehouses, with a fraction $frac_p$ of demand coming from premium customers. We let large demand rates coincide with large stock levels.

Parameter	Values	
K	6; 18	
$\left[T_k^{reg}, T_{lk}^{lat}, T_k^{em}\right]$	[8,1,2]	
Premium fraction $frac_p$	0.1; 0.2; 0.3; 0.5	
M_k	0.05	0.5
S_k	1; 2	4; 8

Table A1 Parameter values considered for testing the analysis approach.

Table A2 shows the simulated and computed values for various performance measures, computed as averages over all warehouses (e.g. β_{avg} shows the average warehouse fill rate).

Settings					β_{avg}		α_{avg}		EW_{1-avg}		EW_{2-avg}	
Case	$\boldsymbol{K}$	$\boldsymbol{M_k}$	$\boldsymbol{frac_p}$	$\boldsymbol{S}$	**Sim**	**Analytic**	**Sim**	**Analytic**	**Sim**	**Analytic**	**Sim**	**Analytic**
1	6	0.05	0.1	1	0.68	0.68	0.32	0.32	0.32	0.32	1.14	1.14
2	6	0.05	0.1	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
3	6	0.05	0.2	1	0.67	0.67	0.33	0.33	0.33	0.33	1.16	1.16
4	6	0.05	0.2	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
5	6	0.05	0.3	1	0.66	0.66	0.33	0.34	0.34	0.34	1.18	1.18
6	6	0.05	0.3	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
7	6	0.05	0.5	1	0.65	0.65	0.34	0.35	0.37	0.36	1.21	1.22
8	6	0.05	0.5	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
9	6	0.5	0.1	4	0.50	0.50	0.48	0.48	0.52	0.51	1.23	1.23
10	6	0.5	0.1	8	0.96	0.96	0.04	0.04	0.04	0.04	0.06	0.06
11	6	0.5	0.2	4	0.49	0.48	0.48	0.50	0.56	0.54	1.22	1.22
12	6	0.5	0.2	8	0.96	0.96	0.04	0.04	0.04	0.04	0.06	0.06
13	6	0.5	0.3	4	0.47	0.46	0.48	0.51	0.59	0.56	1.20	1.21
14	6	0.5	0.3	8	0.96	0.96	0.04	0.04	0.04	0.04	0.06	0.06
15	6	0.5	0.5	4	0.44	0.42	0.47	0.55	0.67	0.62	1.19	1.22
16	6	0.5	0.5	8	0.96	0.96	0.04	0.04	0.04	0.04	0.05	0.05
17	18	0.05	0.1	1	0.68	0.68	0.32	0.32	0.32	0.32	1.14	1.14
18	18	0.05	0.1	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
19	18	0.05	0.2	1	0.67	0.67	0.33	0.33	0.33	0.33	1.16	1.16
20	18	0.05	0.2	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
21	18	0.05	0.3	1	0.66	0.66	0.34	0.34	0.34	0.34	1.18	1.18
22	18	0.05	0.3	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
23	18	0.05	0.5	1	0.65	0.65	0.35	0.35	0.35	0.35	1.22	1.22
24	18	0.05	0.5	2	0.94	0.94	0.06	0.06	0.06	0.06	0.15	0.15
25	18	0.5	0.1	4	0.50	0.50	0.50	0.50	0.50	0.50	1.23	1.23
26	18	0.5	0.1	8	0.96	0.96	0.04	0.04	0.04	0.04	0.06	0.06
27	18	0.5	0.2	4	0.48	0.48	0.52	0.52	0.52	0.52	1.23	1.22
28	18	0.5	0.2	8	0.96	0.96	0.04	0.04	0.04	0.04	0.06	0.06
29	18	0.5	0.3	4	0.46	0.46	0.54	0.54	0.55	0.54	1.23	1.22
30	18	0.5	0.3	8	0.96	0.96	0.04	0.04	0.04	0.04	0.06	0.06
31	18	0.5	0.5	4	0.40	0.40	0.59	0.60	0.62	0.60	1.25	1.25
32	18	0.5	0.5	8	0.96	0.96	0.04	0.04	0.04	0.04	0.05	0.05

Table A2 Detailed comparison between simulation and our analysis approach.

Chapter 5

Analysis of a two-echelon system under lost sales[7]

5.1 Introduction

In Section 1.8.1, we mentioned that no suitable models yet exist to accurately analyze a two-echelon model under lost sales. However, such a model is a necessary building block to determine whether dedicated customer stocks are an effective differentiation tool. Specifically, when stock can be kept at customer sites in addition to stock at a central stock point, we get an additional echelon level in the supply chain. By developing an analysis approach for this system, we satisfy our fourth research objective. Subsequently, we use this approach in Chapter 6 to determine the effectiveness of dedicated stocks as a differentiation tool.

We consider a single-item two-echelon spare parts supply system, consisting of a central depot and multiple local warehouses. Demand arrives at each local warehouse according to a Poisson process. Each location uses a continuous review, one-for-one replenishment policy for inventory control. Demand that cannot be met from stock is served using an emergency shipment from an external source with infinite supply and is thus lost to the system. As discussed in Section 1.7.1, the analysis of this lost-sales inventory system is more complex than its counterpart under full backordering. In particular, the analysis of the central depot is complex, since (i) the order process is not Poisson, and (ii) the order arrival rate depends on the inventory states of the warehouses: warehouses only generate replenishment orders if they have stock on hand.

In the literature, solutions have been found for specific cases. Andersson and Melchiors (2001) consider a model where demand at a local warehouse is lost if the warehouse is out of stock, even if the depot still has stock on hand. They approximate the arrival process at the depot by a Poisson distribution with a rate that depends on the fill rates at the warehouses. Given these fill rates, they compute the mean waiting time for replenishment orders at the depot. This waiting time is input for the computation of the fill rates at the warehouses. This yields an iterative procedure that gives reasonably accurate results in general, but often does not converge when a lot of stock is kept at the central depot, with little stock kept locally. Such a setting is very

[7] This chapter is based on the working paper “On two-echelon inventory systems with Poisson demand and lost sales” by E. M. Alvarez and M. C. van der Heijden.

common for expensive slow movers to benefit from risk pooling. Seifbarghy and Jokar (2006) consider a model similar to that of Andersson and Melchiors under an (R, Q) policy, i.e. orders are placed in batches of size Q whenever the inventory position reaches a level R. The analysis approach is similar as well. The authors only consider cases with identical warehouses and limit their experiments to settings with high service levels (fill rates of 90% or more). Alternatively, Hill et al. (2007) explicitly use arrival rates that depend on the number of outstanding orders at the depot. They assume that (i) each local warehouse may only have one outstanding order at any time, and (ii) the shipment time from depot to warehouse is at least the central depot lead time. The second assumption is particularly restrictive, since upstream lead times tend to exceed downstream lead times in practice.

The papers mentioned above use rather simple approximations for the analysis of the central depot, ignoring the fact that demand is not Poisson distributed there, and that the demand rate at the depot depends on the inventory levels at all warehouses. In this paper, we develop improved approximations for the service levels using a more accurate analysis of the order arrival rates and the pipeline at the central depot. Although we also assume that demand arrives at the depot according to Poisson processes, we find that the accuracy of our approximations is high and does not depend on the stock levels in the system. As a result, our approach works well for both high and low service levels. We typically encounter both in the optimization of multi-item spare part inventory systems. Furthermore, the approach can handle settings with non-identical local warehouses and no assumptions are made on the maximum number of outstanding orders.

To facilitate the analysis, we make one key assumption on the product flows that seems very reasonable from a practical perspective: demand at a local warehouse is only lost if it cannot be met from local stock, central stock or any replenishment order in transit between the depot and the local warehouse. The logic is that a shipment from the depot to any warehouse is generally faster than an emergency shipment from a (remote) external supplier. Therefore, the latter option should not be used if the depot still has stock on hand. This setting also applies in the chapter on dedicated stocks: there, we assume that all customers are in the vicinity of the stock point, making a regular shipment to any customer faster than an emergency shipment.

For each warehouse, we give approximations for (i) the fraction of demand met using the two-echelon system and (ii) the related mean waiting time. The remaining demand is met through emergency supply and faces the related supply time as delay. Then, we can compute both the overall expected downtime due to spare parts unavailability and the system costs for the item being considered. Both indicators are key performance measures in the multi item problem of the next chapter.

In the next section we define our model and give our analysis. We present the results from numerical experiments in Section 5.3, and give our conclusions in Section 5.4.

5.2 Model

5.2.1 Notation and assumptions

Consider a single-item two-echelon network consisting of a central depot and K local warehouses (Figure 5.1). We use index 0 for the central depot and indices $1 \dots K$ for the local warehouses. Demand occurs at local warehouse k according to a Poisson process with rate m_k. The inventory at stock point k, $k = 0, \dots, K$, is controlled using an $(S_k - 1, S_k)$ installation stock policy. The transportation time between the central depot and local warehouse k is deterministic and equal to L_k. The replenishment lead time to the central depot is assumed to be exponentially distributed with mean L_0, which facilitates an analysis using Markov chains. Also, the performance of such lost sales models is not very sensitive to the lead time distribution, see Alfredsson and Verrijdt (1999).

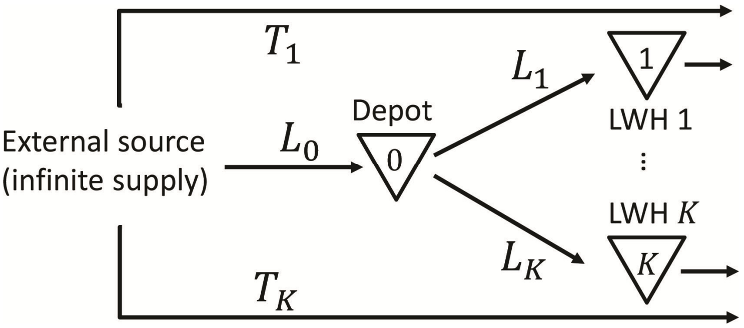


Figure 5.1 A graphical representation of the supply system.

Demand arriving at local warehouse k is served through the two-echelon network if an item is available at the local warehouse, the central depot or in the transport pipeline between the two. Otherwise, the demand is satisfied using an external emergency source at additional costs. The emergency lead time has an arbitrary probability distribution with mean T_k. We consider the setting where the depot is located close to the local warehouses – which may occur in practice when the local warehouses represent car stocks or stock points at customer sites – whereas the distance from the external source to the depot is much longer. Therefore, we assume that $L_k < T_k < L_k + L_0$ $\forall k$: as emergency shipments occur from the external source, the emergency shipment time will exceed the shipment time from the depot. Incidentally, even if $L_k > T_k$, it may still be beneficial to only use emergency shipments once the depot is out of stock to limit expensive emergency shipment costs. We use the common assumption that the external emergency source has infinite capacity.

Our performance indicators are the fraction of warehouse k demand that is satisfied through the regular channel, α_k, and the expected waiting time for demand satisfied through the regular channel $E[W_k]$ $(k = 1 \dots K)$. These performance indicators enable us to evaluate:

- the total relevant costs per year as $h_0 S_0 + \sum_{k=1}^{K}\{h_k S_k + C_k(1-\alpha_k)m_k\}$, where h_k denotes the unit holding costs per year at stock point k and C_k the additional emergency shipment costs to local warehouse k (above the costs of regular supply);
- the downtime waiting for parts $DTWP_k$ at local warehouse k as $\alpha_k E[W_k] + (1 - \alpha_k)T_k$.

5.2.2 Analysis

We find $E[W_k]$ by noting that waiting time when using the regular channel only arises when demand is backordered while waiting for a part that is either on stock at the depot, or in transit between depot and warehouse and still unassigned to other demand. Hence, $E[W_k]$ follows from Little's Law, i.e. $E[W_k] = E[BO_k]/\alpha_k m_k$, with BO_k denoting the number of items backordered at warehouse k.

In turn, we find both $E[BO_k]$ and α_k by analyzing the central depot, specifically the distribution of the *number of backorders at the depot* destined for local warehouse k, which we denote by BO_0^k. This is the critical and novel part of our analysis. Depot backorders occur when a local warehouse sends a replenishment request to the depot when the depot is out of stock. The distribution of BO_0^k allows us to determine the distribution of the number of outstanding orders (i.e. the pipeline) at warehouse k, which in turn allows us to determine the distribution of BO_k. From the distribution of the depot backorders BO_0^k, we also directly obtain α_k: once the depot is out of stock, at most S_k additional demands can still be met through the regular channel (either directly from warehouse stock or from items in the transport pipeline). Hence, once we have S_k depot backorders for warehouse k, further demand at that warehouse is lost. We thus find α_k as follows:

$$\alpha_k = \Pr\{BO_0^k < S_k\} = 1 - \Pr\{BO_0^k = S_k\} \tag{5.1}$$

We first show how to find the distribution of BO_0^k. Then, we show how to find $E[BO_k]$. For the analysis, we also need PI_k, the number of outstanding orders at each location $(k = 0, \dots, K)$.

5.2.2.1 <u>*Distribution of BO_0^k, the number of backorders at the central depot for warehouse k*</u>

We find distribution of BO_0^k by conditioning on the number of outstanding orders at the depot PI_0. Under full backordering, PI_0 has a Poisson distribution, with the depot backorders being disaggregated over warehouses using a binomial distribution (i.e. given that the depot has y_0 outstanding orders, the conditional probability $\Pr\{BO_0^k = x_k | PI_0 = y_0\}$ of having x_k backorders at the depot for warehouse k follows a binomial distribution with $y_0 - S_0$ trials and "success probability" $q_k = m_k / \sum_{h=1}^{K} m_h$, see e.g. Graves (1985)). However, in our model the arrival rate at the depot becomes state-dependent once the depot is out of stock (i.e. $PI_0 > S_0$). Then, the probability of a warehouse k being out of stock increases with PI_0, resulting in an increasingly large fraction of arrivals being lost to the system (and hence to the depot). Also, $\Pr\{BO_0^k = x_k|PI_0 = y_0\}$ no longer follows a binomial distribution, as we show later on.

We now first show how to compute $\Pr\{BO_0^k = x_k|PI_0 = y_0\}$. Subsequently, we use these conditional probabilities to compute the state-dependent arrival rates at the depot and find the distribution of PI_0. Finally, we find the distribution of BO_0^k as follows:

$$\Pr\{BO_0^k = x_k\} = \sum_{y_0=0}^{S^{tot}} \Pr\{BO_0^k = x_k|PI_0 = y_0\} \Pr\{PI_0 = y_0\} \qquad (5.2)$$

In (5.2), S^{tot} is shorthand notation for $\sum_{k=0}^{K} S_k$. Note that PI_0 cannot exceed this value: the central depot can have at most S_0 orders outstanding for replenishing its stock. Once the depot is out of stock, further outstanding orders can only be realized if a warehouse still has stock on-hand, which results in backorders at the depot for that warehouse. As the number of depot backorders for warehouse k cannot exceed S_k, we find in an upper bound on PI_0 of $S_0 + \sum_{k=1}^{K} S_k$ items.

Step 1: compute $Pr\{BO_0^k = x_k|PI_0 = y_0\}$

Note that $\Pr\{BO_0^k = 0|PI_0 = y_0\}$ equals 1 when $y_0 \leq S_0$, as these orders are only meant for replenishing depot stock. For $y_0 > S_0$, we have in total $y_0 - S_0$ backorders at the depot which must be disaggregated over the warehouses. As each warehouse k can have at most S_k backorders at the central depot and the total number of backorders at the depot may add up to $\sum_{k=1}^{K} S_k$, it is clear that the total number of depot backorders cannot be disaggregated over the warehouses using a binomial distribution. For instance, if $y_0 - S_0 = 5$ depot backorders are to be allocated over 3 warehouses, with $S_k = 2$ for each warehouse, the only possibilities (BO_0^1, BO_0^2, BO_0^3) are $\{(1,2,2), (2,1,2), (2,2,1)\}$.

If $S_k \to \infty\, \forall k$, the joint distribution of BO_0^k $(k = 1..K)$ when $PI_0 = y_0$ is a K-category multinomial distribution with $y_0 - S_0$ trials and success probabilities $q_k = m_k / \sum_{h=1}^{K} m_h$. Let us denote this multinomial probability distribution by

$$p(x_1, .., x_K|PI_0 = y_0) = Pr\{BO_0^1 \leq x_1, .., BO_0^K \leq x_K|PI_0 = y_0\} \qquad (5.3)$$

Clearly, the joint probability distribution for *finite* values of S_k has a *truncated* multinomial distribution, as we condition on the upper bounds S_k for the depot backorders for warehouse k:

$$p_{S_1,\dots,S_K}(x_1, .., x_K|PI_0 = y_0) = \frac{Pr\{BO_0^1 \leq x_1, .., BO_0^K \leq x_K|PI_0 = y_0\}}{Pr\{BO_0^1 \leq S_1, .., BO_0^K \leq S_K|PI_0 = y_0\}} = \frac{p(x_1, .., x_K|PI_0 = y_0)}{p(S_1, .., S_K|PI_0 = y_0)} \qquad (5.4)$$

We thus find the marginal probability density functions $Pr\{BO_0^k = x_k|PI_0 = y_0\}$, where $x_k \leq S_k \; \forall k$, as below:

$$Pr\{BO_0^k = x_k|PI_0 = y_0\} = \\ p_{S_1,..,S_K}(S_1,..,x_k,..,S_K|PI_0 = y_0) - p_{S_1,..,S_K}(S_1,..,x_k - 1,..,S_K|PI_0 = y_0) \qquad (5.5)$$

In summary, we are able to compute $Pr\{BO_0^k = x_k|PI_0 = y_0\}$ from the multinomial distribution $p(x_1,..,x_K|PI_0 = y_0)$. To quickly evaluate $p(x_1,..,x_K|PI_0 = y_0)$, we use the powerful method by Levin (1981). Let $n = y_0 - S_0$ denote the total number of backorders at the depot for all warehouses jointly. Levin (1981) shows that:

$$p(x_1,..,x_K|PI_0 = y_0) = \frac{n!}{v^n e^{-v}}\{\prod_{k=1}^{K} Pr\{X_k \leq x_k\}\}Pr\{W = n\} \qquad (5.6)$$

In (5.6), v may be any real number, the X_k are independent Poisson distributed random variables with means $v \cdot q_k$, and W is a sum of independent truncated Poisson random variables. That is, $W = \sum_{k=1}^{K} Z_k$, where Z_k has a truncated Poisson distribution with mean $v \cdot q_k$ and upper bound x_k. As Levin (1981) states, v is a tuning parameter which may be chosen for convenience and numerical stability. He suggests setting $v = n$, because then Stirling's approximation can be used to compute the first term in (5.6): $\frac{n!}{v^n e^{-v}} \approx \sqrt{2\pi n}$. Finally, $\Pr\{W = n\}$ can either be evaluated using explicit convolutions (as we do in our numerical experiments) or using a Normal approximation.

<u>Step 2. Compute $Pr\{PI_0 = y_0\}$</u>

Given y_0 outstanding orders at the depot, we find the arrival rate of new requests $M(y_0)$ as follows: with probability $1 - \Pr\{BO_0^k = S_k|PI_0 = y_0\}$, new demand at warehouse k can be met from the regular channel (either from stock at warehouse k or the depot, or because there is an item in transit from the depot to warehouse k). Then, the depot sees requests from warehouse k at rate m_k. Otherwise, the depot sees no arrivals from that warehouse. We thus find:

$$M(y_0) = \sum_{k=1}^{K} m_k(1 - \Pr\{BO_0^k = S_k|PI_0 = y_0\}) \qquad (5.7)$$

Note that $M(y_0) = \sum_{k=1}^{K} m_k$ when $y_0 < S_0$, as the depot still has stock on-hand then. By approximating the arrival process by a Poisson process, we can model PI_0 as a continuous-time Markov chain (Figure 5.2). In the figure, $\mu = 1/L_0$ denotes the regular replenishment rate.

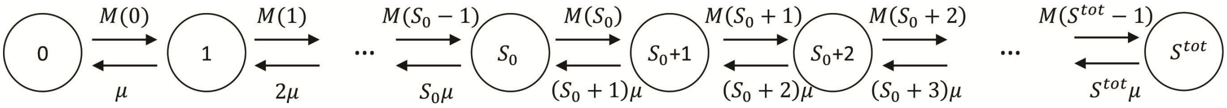


Figure 5.2 The Markov chain characterizing the number of units in resupply at the central depot.

5.2.2.2 Finding $E[BO_k]$

We find $\mathrm{E}[BO_k]$ from the distribution of the number of outstanding orders PI_k to warehouse k. As in Graves (1985), PI_k ($k \geq 1$) consists of (i) the items backordered *at the depot* for warehouse k (i.e. BO_0^k), and (ii) the items in transit from the depot to warehouse k, which we denote by Q_k. Under full backordering, these two elements are independent of each other (Graves, 1985) and Q_k thus has the same distribution as that of new demand at warehouse k during the lead time. Then, PI_k can be computed as a convolution of BO_0^k and Q_k. However, in this model BO_0^k and Q_k are *not* mutually independent: once $BO_0^k = S_k$ (i.e., there are S_k backorders at the depot for warehouse k), the transport pipeline to warehouse k cannot increase further until one of the S_k depot backorders is cleared. Furthermore, new requests at warehouse k are lost to the system in this state.

Despite this complication, we can still find an accurate approximation for PI_k by assuming that (i) Q_k still has the same distribution as new demand arriving at warehouse k during the lead time and that (ii) BO_0^k and Q_k are mutually independent when $BO_0^k < S_k$. As a result, when $BO_0^k < S_k$ we find PI_k as a convolution of BO_0^k and Q_k, with Q_k having a Poisson distribution with parameter $m_k L_k$. If $BO_0^k = S_k$, any demand arriving at warehouse k will be served using an emergency shipment from an external supplier and thus be lost to the system, with Q_k thus being 0. We realize that Q_k might actually be greater than 0, since there might be items in transit from the depot to warehouse k that are destined for clearing previous backorders at warehouse k. Still, as these items will not contribute to the replenishment of warehouse k stock, they may be ignored when computing the pipeline to warehouse k. Overall, we find:

$$\Pr\{PI_k = y_k\} = \begin{cases} \sum_{x_k=0}^{\min\{S_k-1, y_k\}} \Pr\{BO_0^k = x_k\} \Pr\{Q_k = y_k - x_k\}, \ y_k \neq S_k \\ \sum_{x_k=0}^{S_k-1} \Pr\{BO_0^k = x_k\} \Pr\{Q_k = y_k - x_k\} + \Pr\{BO_0^k = S_k\}, \ y_k = S_k \end{cases} \tag{5.8}$$

From this distribution, we subsequently find $E[BO_k]$ as follows:

$$E[BO_k] = \sum_{y_k=S_k+1}^{\infty} (y_k - S_k) \Pr\{PI_k = y_k\} \tag{5.9}$$

To truncate the sum in (5.9), we set an upper bound on PI_k of $S_k + g_k^{UB}$, with S_k the bound on BO_0^k and g_k^{UB} being an accurate upper bound on Q_k. We find g_k^{UB} when $1 - \Pr\{Q_k \leq g_k^{UB}\} < \varepsilon$.

5.3 Numerical experiment

5.3.1 Experiment objectives

First, we investigate whether our technique for analysing the central depot leads to more accurate estimations of the performance measures $DTWP_k$ and α_k. To this end, we compare our model to that of Andersson and Melchiors (2001) which has similar characteristics. For a fair comparison, the transportation time L_k from the depot to any warehouse k is set to 0 and the warehouse base stock levels S_k to values greater than 0. Then, both models are equivalent: a negligible transportation time results in direct item arrivals at a warehouse if the depot has stock available. Hence, in our model demand is only lost if a warehouse is out of stock, which is similar to the Andersson and Melchiors model (AM model in short). We use discrete-event simulation as a benchmark for both models, with model accuracy expressed in terms of a relative deviation to simulated values.

We also investigate the accuracy of our model when $L_k > 0$. In practice, L_k will be relatively short, given a setting with a regional depot that supplies warehouses near that depot. The shipment time to the depot will then greatly exceed the shipment time from the depot to any warehouse. As our model has not been considered in literature before, we only use simulation as a benchmark.

Per problem instance, we performed a single long simulation run consisting of a warm-up period to reach steady-state and a data collection interval. The data collection interval was chosen such that the total demand at each local warehouse would be large. Each warehouse received at least half a million requests, with the average number of requests per warehouse being over 3 million.

5.3.2 Comparison to Andersson and Melchiors when $L_k = 0$ at all warehouses

We test 32 instances with $L_k = 0$, each with 5 or 20 warehouses. In all instances, L_0 equals 10 days, and T_k equals 2 days. The remaining parameter values are given in Table 5.1 and Table 5.2. The demand rates in Table 5.2 are specified for groups of warehouses: warehouses 1 through 4 always have the same demand rates, as do warehouses 5 through 8, etc. The stock levels depend on the demand rate heights: for slow movers we use smaller stock levels than for fast movers.

Parameters	**Values**	
$[m_1, m_2, m_3, m_4, m_5]$	[0.05,0.05,0.05,0.05,0.05]; [0.01,0.02,0.04,0.08,0.1]	[0.5,0.5,0.5,0.5,0.5]; [0.1,0.2,0.4,0.8,1]
S_0	1; 2	5; 10
$S_k (k \geq 1)$	1; 2	2; 4

Table 5.1 Tested settings for problem instances with K=5 warehouses.

Parameters	Values	
$[m_1 - m_4, m_5 - m_8, m_9 - m_{12}, m_{13} - m_{16}, m_{17} - m_{20}]$	[0.05,0.05,0.05,0.05,0.05]; [0.01,0.02,0.04,0.08,0.1]	[0.5,0.5,0.5,0.5,0.5]; [0.1,0.2,0.4,0.8,1]
S_0	4; 8	20; 80
$S_k (k \geq 1)$	1; 2	2; 4

Table 5.2 Tested settings for problem instances with K=20 warehouses.

Table 5.3 shows the accuracy of α_k (i.e. the fraction of demand met through the regular channel) and the computation times, with our model denoted as OM. As α_k fully determines $DTWP_k$ (i.e. the waiting time when using the regular channel is zero), we do not show the accuracy of $DTWP_k$ as well. Clearly, our method is very accurate: all deviations are at most 0.1%. The AM model, in contrast, performs much worse and does not even converge for 3 problem instances (2nd column). For that model, the performance measures are estimated over those instances where convergence did occur. The instances where convergence did not occur all had a high stock level at the depot combined with low local stock levels. In those instances, the value of α would sometimes iterate between 0.6 and 0.99. The computation times of both methods are a fraction of a second, where we note that those of our method clearly increase with the amount of stock kept locally.

	# not converged AM model	Avg. deviation α_k		Max. deviation α_k		Avg. comp. time (millisec)		Max. comp. time (millisec)	
K		OM	AM	OM	AM	OM	AM	OM	AM
5	1	0.02%	0.74%	0.10%	3.29%	4	4	4	4
20	2	0.01%	0.30%	0.06%	1.68%	24	11	47	16

Table 5.3 The accuracy and computation times of our method (OM) and that of Andersson and Melchiors (AM), $L_k = 0$.

5.3.3 Accuracy of our model for instances with positive transportation times

We test 64 instances, with all parameters except L_k as before. For L_k, we consider 0.5 and 1.5 days. Table 5.4 shows the accuracy for $DTWP_k$ (i.e. the overall downtime for parts). Because the average deviation is influenced by a few large deviations for waiting times close to zero, we also show the median and the 75th and 90th percentiles. Relatively large deviations, such as 3.1%, pertain to very small waiting times: e.g. a simulated value of 0.00035 and a computed value of 0.00034. The accuracy of α_k and the computation times are the same as in Table 5.3: neither estimate depends on the value of L_k. Clearly, the $DTWP$-estimates are also very accurate: 90% of all values has a deviation below 0.6%.

K	Avg. deviation $DTWP_k$	Max. deviation $DTWP_k$	Median	75th percentile	90th percentile
5	0.25%	3.06%	0.10%	0.24%	0.56%
20	0.21%	2.84%	0.08%	0.22%	0.58%

Table 5.4 The accuracy of our method for cases where $L_k > 0$.

5.4 Conclusions

We developed a simple and accurate approximation for a two-echelon inventory model with Poisson demand and lost sales. In contrast to existing methods, we do not require restrictive assumptions, and our approach works very well for a broad range of settings. Our model's simplicity arises from the fact that no iterative procedure is needed as in Andersson and Melchiors (2001). In the next chapter, we will therefore use the model as a building block in a multi-item model for service differentiation using dedicated customer stocks.

Chapter 6

Dedicated customer stocks[8]

6.1 Introduction

In this chapter, we meet our fifth research objective by investigating the added value of dedicated customer stocks as a service differentiation tool. A supplier may then keep stock of certain items at premium customers' sites in addition to stock at some central location. Dedicated stocks are often used in practice, because of its simplicity. A company specializing in baggage handling systems at airports, for instance, often uses such stocks to ensure fast reaction times when a failure occurs. Also, if the shipment time from the central stock point to a customer exceeds the maximum time that this customer is willing to wait for parts, it might even be necessary to keep some items stock at the customer's site. Still, no research has yet been done on the savings possible with this approach: we expect the benefits from risk pooling to be smaller for this approach than for the case where all stock is kept centrally. We thus investigate if and when the approach has added value.

In an extensive computational experiment, we investigate the added value of dedicated customer stocks by comparing it to two benchmark differentiation approaches, namely the one-size-fits-all approach and the critical level policy. Such a comparison has not been made in the literature before. Furthermore, we investigate a model where dedicated stocks and critical level policies are jointly used for differentiation. As we have seen in Chapters 3 and 4, it may be very beneficial to combine differentiation tools. By considering the combination of dedicated stocks and the critical level policy, we can determine whether this combination also leads to significant savings and under what conditions each individual strategy (dedicated stocks, critical level policies) works best.

We focus on a multi-item system with one warehouse that serves various customers, with each customer belonging to either a premium or a non-premium class. Of this system, we consider two variants, i.e. (i) a variant in which all unmet demand is *backordered* and, (ii) a variant in which unmet demand is satisfied through an *emergency shipment* from a central stock point with infinite supply. In both cases, we aim to minimize the costs of the system subject to

[8] This chapter is based on the paper "Service differentiation in spare parts supply through dedicated stocks" by E. M. Alvarez, M. C. van der Heijden, and W. H. M. Zijm, to appear in Annals of Operations Research.

restrictions on the waiting times for spares. As in Chapters 3 and 4, we solve this problem by using an approach similar to Dantzig-Wolfe decomposition.

As a building block for multi-item optimization, we must be able to compute performance measures for a single item under each differentiation strategy (e.g. dedicated stocks, critical level policy) and shipment option (backordering or emergency shipments). Under the dedicated stocks strategy with emergency shipments, the supply chain fits in the framework that has been described in 0. Therefore, we can analyze that system using the approach of Chapter 5. We give analysis details for the remaining strategies and shipment options in this chapter.

The chapter is organized as follows: We present our model in Section 6.2. and discuss the optimization approach for this model in Section 6.3. In Section 6.4, we discuss system analysis for a single item under the various differentiation strategies and shipment modes. We then test the model in an extensive experiment (Section 6.5). Finally, we draw conclusions in Section 6.6.

6.2 Model description

6.2.1 Outline

We consider a warehouse supplying various items to customers in the vicinity of that warehouse. All items are critical: any item failure causes a system failure. Customers can be partitioned into two customer classes, a *premium* and a *non-premium* class, with a distinct target service level applying for each customer class in terms of a maximum time a customer of that class will wait for spares.

We consider four strategies for meeting all service requirements at minimal costs:

- **One-size-fits-all (OSFA):** all stock is kept at the warehouse, with demand met from on-hand stock if it is available. Premium and non-premium requests are handled in the same way, fulfilling the tightest service level requirement.
- **Dedicated stocks (DS):** stock of certain items may be kept at a customer's facility next to stock at the warehouse. Demand is satisfied from customer stock if possible, with a replenishment request being sent to the warehouse. At the warehouse, all demand is satisfied from on-hand stock if possible; we do not reserve a part of the warehouse stock for meeting premium requests.
- **Critical level policy (CLP):** all stock is kept at the warehouse. This stock is always used for meeting premium requests if available, whereas non-premium requests are only met if the warehouse stock exceeds a critical level.
- **A combined strategy with dedicated stocks and critical levels (COMBO):** the individual strategy (DS or CLP) may vary per item, with only one strategy selected per item (i.e. DS and CLP may not be used for the same item).

The rationale behind COMBO is that the added value of each individual strategy depends on an item's characteristics: DS is likely most beneficial for inexpensive fast movers, since they are often needed and inexpensive to keep in stock. Then, dedicated stocks avoid downtime arising from the shipment time between warehouse and the customer's facility. Conversely, for expensive slow movers it might be better to centralize stocks and differentiate through CLP. The contribution of items with a low demand rate to the overall waiting time for parts is small, which limits the need for fast reaction times for those items. Furthermore, the risk pooling effect is high due to the high item value.

In all strategies, one-for-one replenishment is used at all stock locations. Of each strategy, we consider both (1) a *backorder* variant, and (2) an *emergency shipment* variant. In the backorder variant, first-come-first-served backorder clearing is used under OSFA and DS. Under CLP, *priority* backorder clearing is used, with non-premium backorders only cleared after all premium backorders have been cleared and the warehouse stock level is at least the critical level. In the emergency shipment variant, there exists a central location with infinite capacity upstream in the supply chain that may supply parts directly to the customer site. As customers are located close to the warehouse, we assume that a shipment from the warehouse to any customer is faster than an emergency shipment. Therefore, emergency shipments are only used if both the customer and the warehouse are out of stock, *and* if all items in transit between the warehouse and customer have already been reserved for earlier requests.

In both literature and practice, emergency shipments are often used to satisfy demand that cannot be met from on-hand stock at the closest warehouse (see e.g. Kranenburg and Van Houtum, 2008). However, in Chapters 3 and 4 we have shown that emergency shipments can be very costly in certain circumstances, e.g. when items have relatively low holding costs. In those circumstances, we concluded that backordering is the preferred shipment mode. Therefore, we investigate the differentiation strategies in this paper both under backordering and under emergency shipments.

Figure 6.1 depicts the system, with the grey triangles specifying the differentiation options (either dedicated stocks or critical levels at the warehouse). We have regular and emergency supply channels, with replenishments for backordered requests occurring through the regular channels. Figure 6.2 shows the scheme for handling incoming requests. This scheme applies both for DS and CLP strategies, under both backordering and emergency shipments.

We aim to minimize the system's holding costs and additional emergency shipment costs under constraints on the mean aggregate waiting time per customer. Our decision variables are the warehouse stock level, the strategy per item (i.e. one-size-fits-all, dedicated stocks, critical level policy), and – if applicable – the stock levels at customer sites and the warehouse critical level.

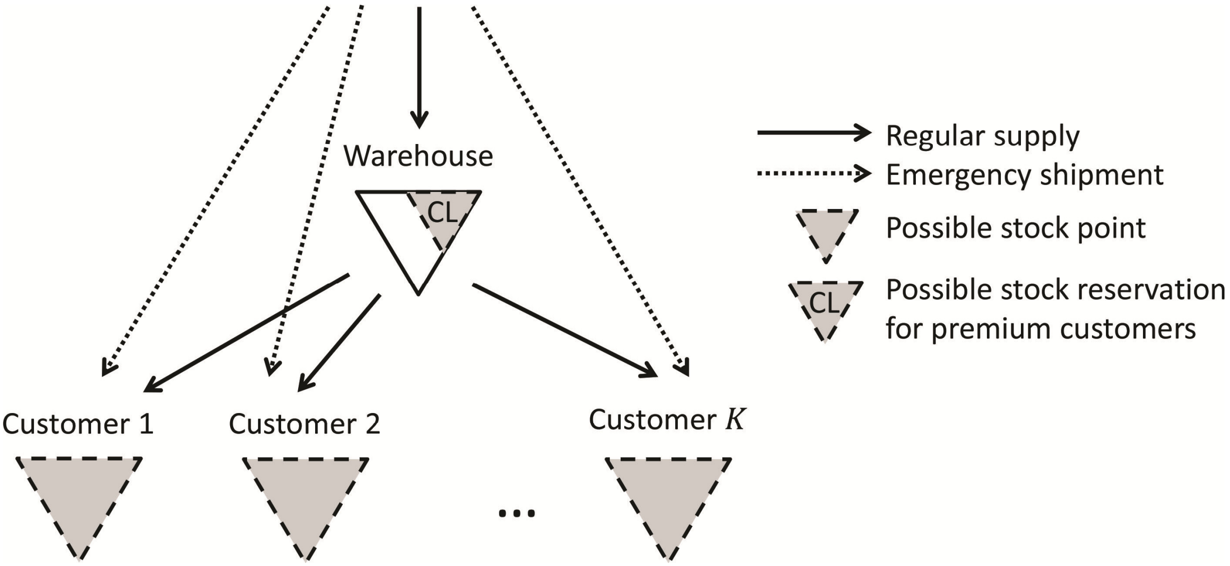


Figure 6.1 Depiction of the system.

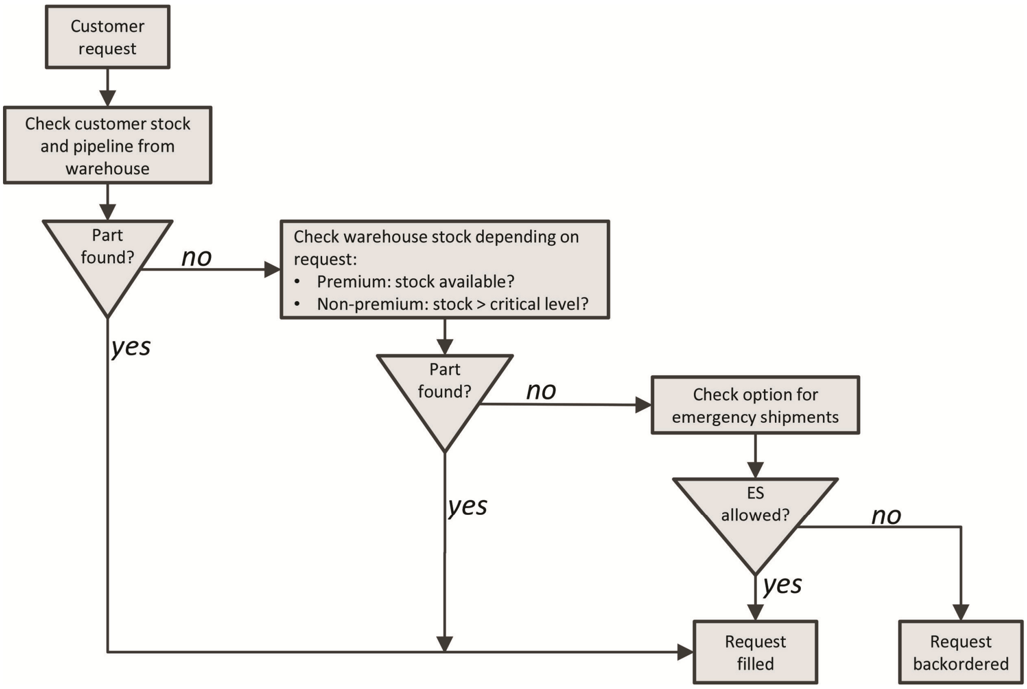

Customer request
Check customer stock and pipeline from warehouse
Part found?
no
yes
Check warehouse stock depending on request:
• Premium: stock available?
• Non-premium: stock > critical level?
Part found?
no
yes
Check option for emergency shipments
ES allowed?
no
yes
Request filled
Request backordered

Figure 6.2 Steps for handling an incoming customer request.

6.2.2 Assumptions and notation

6.2.2.1 Model assumptions

- All demand occurs according to mutually independent Poisson processes.
- The shipment time from the warehouse to any customer is deterministic.

- The emergency shipment time to a customer is deterministic. This is most realistic, although variability could be included in the model (then we use the mean only).
- The regular shipment time to the warehouse is exponentially distributed. Although deterministic shipment times are generally more realistic, this assumption facilitates a performance evaluation based on Markov chain analysis. Also, inventory models for slow moving parts tend to be quite insensitive to lead time variability (Alfredsson and Verrijdt (1999)).

6.2.2.2 Notation

We keep stock of I items for K customers; index 0 refers to the warehouse, and indexes $1, \dots, K$ to the customers. Each customer is either a premium (class 1) customer or a non-premium (class 2) customer, with a class j customer ($j = 1,2$) willing to wait at most W_j^{max} time units on average for any item ($W_1^{max} \leq W_2^{max}$). We let $q(k)$ denote the class to which customer k belongs. Customer k requests item $i = 1, \dots, I$ at rate m_{ik}, with $M_k = \sum_{i=1}^{I} m_{ik}$ denoting the total demand rate from customer k. For each item i, the shipment time to the warehouse is denoted by T_{i0}^{reg}, the mean regular shipment time from the warehouse to customer k is denoted by T_{ik}^{reg} and the emergency shipment time from the central depot to customer k is denoted by T_{ik}^{em} (>T_{ik}^{reg}). Finally, for each item i we denote the unit holding costs per time unit at location k (i.e. including the warehouse) by h_{ik} and the additional costs of an emergency shipment compared to a regular replenishment at customer k by EC_{ik}^{em}. We only require the additional shipment costs, since each demand triggers either a regular or an emergency shipment.

For each item i, we have as decision variables (1) the base stock level S_{ik} at each location k ($k = 0, \dots, K$), with $\boldsymbol{S}_i = [S_{i0}, \dots, S_{iK}]$ denoting the system stock levels, and (2) the critical level C_i denoting the amount of warehouse stock reserved for premium customers. Note that $0 \leq C_i \leq S_{i0}$, since we cannot reserve more items than we have in stock at the warehouse. We combine all variables for item i in an item policy $(\boldsymbol{S}_i, C_i)$. For each item policy, we have as performance measures the expected waiting time $EW_{ik}(\boldsymbol{S}_i, C_i)$ and the fraction of demand met through emergency shipments $\gamma_{ik}(\boldsymbol{S}_i, C_i)$ for item i and customer k, and the total costs $TC_i(\boldsymbol{S}_i, C_i)$ for item i. We now express problem $(P1)$ as follows:

$$(P1) \qquad \min \sum_{i=1}^{I} TC_i(\boldsymbol{S}_i, C_i) = \sum_{i=1}^{I} \sum_{k=0}^{K} h_{ik} S_{ik} + \sum_{i=1}^{I} \sum_{k=1}^{K} \gamma_{ik}(\boldsymbol{S}_i, C_i) m_{ik} EC_{ik}^{em}$$

$$\text{s.t.} \qquad \sum_{i=1}^{I} \frac{m_{ik}}{M_k} EW_{ik}(\boldsymbol{S}_i, C_i) \leq W_{q(k)}^{max} \qquad k = 1, \dots, K$$

$$S_{ik}, C_i \in \mathbf{N}_0, C_i \leq S_{i0} \qquad i = 1, \dots, I, k = 0, \dots, K$$

As mentioned, our system costs consist of holding costs and, if applicable, additional emergency shipment costs. Under backordering, $\gamma_{ik}(\boldsymbol{S}_i, C_i)$ will be 0, and thus the total costs will only consist of holding costs then. Holding costs are computed over the stock in the entire system, including items in transit to the customers. However, the model can be adjusted to compute holding costs over on-hand stock only. Each customer k has a restriction on the mean aggregate waiting time over all items, with m_{ik}/M_k being the fraction of item i waiting time that contributes to the aggregate waiting time. Note that the waiting time threshold $W_{q(k)}^{max}$ depends on the customer's class.

6.3 Solution approach

As in Chapters 3 and 4, we solve $(P1)$ by using an approach based on decomposition and column generation which closely resembles Dantzig-Wolfe decomposition, i.e. we reformulate $(P1)$ to a linear integer programming problem and solve its LP-relaxation to obtain a lower bound. Then, we obtain a near-optimal integer solution by solving the integer problem itself. Section 6.3.1 gives the reformulated variant of $(P1)$. Sections 6.3.2 and 6.3.3 detail how to find a lower bound and near-optimal integer solution respectively.

6.3.1 Reformulation to a linear problem

We obtain the linear variant of $(P1)$ by considering a set of item policies for each item. Our decision problem becomes to select one item policy for each item such that the system costs are minimized while the waiting time restrictions per customer are still met. Let B_i denote the set of item policies for item i and let b_{ir} denote a single item policy $\big(\boldsymbol{S}_i(r), C_i(r)\big)$ in set B_i, i.e. $b_{ir} \in B_i$ with $r = 1,2,\dots,|B_i|$. The binary variable $x_{b_{ir}}$ specifies whether b_{ir} is selected for item i ($x_{b_{ir}}$ then equals 1). The reformulated problem $(P2)$ becomes:

$$\min \sum_{i=1}^{I} \sum_{r=1}^{|B_i|} TC_i(b_{ir}) x_{b_{ir}}$$

$$\text{s.t.} \quad \sum_{i=1}^{I} \sum_{r=1}^{|B_i|} \frac{m_{ik}}{M_k} EW_{ik}(b_{ir}) x_{b_{ir}} \leq W_{q(k)}^{max} \qquad k = 1,\dots,K \tag{6.1}$$

$$\sum_{r=1}^{|B_i|} x_{b_{ir}} = 1 \qquad i = 1,\dots,I \tag{6.2}$$

$$x_{b_{ir}} \in \{0,1\} \qquad i = 1,\dots,I, r = 1,\dots,|B_i|$$

6.3.2 Lower bound

To solve the LP-relaxation of $(P2)$, we must determine what item policies to include in B_i for item i. We first construct an initial policy set to solve the LP-relaxation a first time. Subsequently, we use column generation to iteratively find unconsidered item policies that further improve the solution value. We proceed in this way until we cannot find any more relevant policies. In Section 6.3.2.1, we show how we find an initial policy set. In Section 6.3.2.2, we give the column generation problem and the main steps in solving this problem. Finally, in Section 6.3.2.3 we give the formal column generation procedure.

6.3.2.1 Creating an initial set of policies for each item

We find an initial policy set in a similar way as for the selective transshipment model (see Section 4.4.1.1), i.e. we construct a policy set over all items simultaneously that results in a feasible solution to problem $(P2)$. We limit ourselves to policies without critical levels, i.e. $C_i = 0 \ \forall i$ irrespective of the strategy (DS or CLP, backordering or emergency shipments) considered. Iteratively, we add stock at the item-location combination resulting in largest reduction in waiting times per euro extra costs. Note that stock can be placed at the warehouse or at one of the customer locations, depending on the differentiation strategy considered. If options exist that lead to both lower waiting times *and* costs, we select the option with the greatest reduction in waiting times among those with lower costs. We proceed in this manner until all waiting time restrictions are met.

We denote a stock increase at item-location (i,k) by $\boldsymbol{S}_i + U_{ik}$, with $\Delta W(\boldsymbol{S}_i + U_{ik})$ denoting the related decrease in waiting times and $\Delta TC(\boldsymbol{S}_i + U_{ik}) = TC_i(\boldsymbol{S}_i + U_{ik}) - TC_i(\boldsymbol{S}_i)$ the extra investment. The expression $\Delta W(\boldsymbol{S}_i + U_{ik})$ is given below, with $[a]^+ = \max\{0, a\}$. Note that we only focus on the amounts by which the aggregate waiting times exceed their respective thresholds, as our aim is only to find a feasible solution.

$$\Delta W(\boldsymbol{S}_i + U_{ik}) = \\ \sum_{n=1}^{K} \left\{ \left[\sum_{i=1}^{I} \frac{m_{in}}{M_n} EW_{in}(\boldsymbol{S}_i) - W_{q(n)}^{max} \right]^+ - \left[\sum_{i=1}^{I} \frac{m_{in}}{M_n} EW_{in}(\boldsymbol{S}_i + U_{ik}) - W_{q(n)}^{max} \right]^+ \right\} \tag{6.3}$$

Note that we obtain a new item policy whenever we change one stock level value. As in the selective transshipment model, we include all these policies in policy set B_i to limit the amount of time needed for generating additional policies later on. We realize that some of these policies might be poor options. Therefore, when looking for an integer solution later on, we first remove any poor policies from our policy set before optimizing the integer problem. Section 6.3.3 gives further details.

6.3.2.2 The column generation problem

In the column generation step, we iteratively look for unconsidered item policies that have negative reduced costs. In each iteration, we find the policy with minimum reduced costs for each item i and we add this policy to B_i if these reduced costs are negative. We proceed in this way until we cannot find any policy with negative reduced costs. We give further details on column generation in Section 1.9.2.

6.3. Solution approach

The reduced costs $RED_i(b_{ir})$ related to policy b_{ir} are given by equation (2), with $u_k \leq 0$ and $v_i \geq 0$ denoting the shadow prices associated with (6.1) and (6.2) respectively.

$$RED_i(b_{ir}) = RED_i(\boldsymbol{S}_i(r), C_i(r)) = TC_i(\boldsymbol{S}_i(r), C_i(r)) - \sum_{k=1}^{K} \frac{u_k m_{ik}}{M_k} EW_{ik}(\boldsymbol{S}_i(r), C_i(r)) - v_i \qquad (6.4)$$

For three strategies, the column generation procedure has been discussed before. Specifically, the procedure for DS under backordering has been described in Wong et al. (2007a), that for CLP under backordering has been described in Chapter 3, and that for CLP under emergency shipments has been described in Kranenburg and Van Houtum (2008). We now focus on the column generation procedure for DS under emergency shipments. For the combination of DS and CLP (i.e. COMBO), we use both the procedures for DS and CLP. As we only consider DS, we omit the critical level $C_i(r)$ in the remainder of this section. Also, we omit suffix r for ease of notation.

Under DS with emergency shipments, we find a complication for column generation that does not occur for the variant with backordering. Specifically, we cannot make an exact decomposition over customers in the emergency shipment variant, because the service level (i.e. fill rates and waiting times) at a customer may depend on the stock levels at other customers, see also Chapter 5. If the warehouse stock level is positive (i.e. $S_{i0} > 0$), a stock increase at a customer k improves that customer's service level at the expense of the service levels at other customers. Let us consider the special case where S_{ik} increases from 0 to 1, while $S_{ih} = 0, h \neq k$. When $S_{ik} = 0\ \forall k \geq 1$, a customer's demand is only met through regular supply if the warehouse has stock on-hand. Otherwise, an emergency shipment is used. Figure 6.3 shows the Markov chain of the on-hand stock level at the warehouse (with $M_i^{tot} = \sum_{k=1}^{K} m_{ik}$ and $\mu_{i0} = 1/T_{i0}^{reg}$). If $S_{ik} = 1$, we can reach all states. In contrast, state -1 cannot be reached if $S_{ik} = 0$, since there are no regular replenishment orders from customer k then.

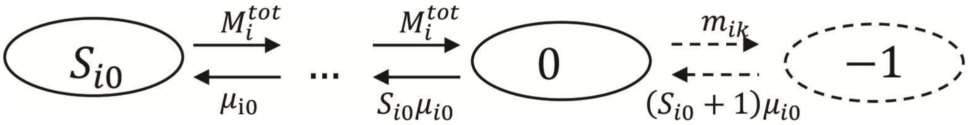


Figure 6.3 Markov chain of the warehouse stock level when there is at most 1 unit of dedicated stock.

When $S_{ik} = 1$, the steady-state probability of being in states S_{i0} up to 0 decreases compared to a similar setting with $S_{ik} = 0$. Hence, an increase of S_{ik} causes a decrease in the warehouse fill rate. As a consequence, the service levels at all customers decrease, except at customer k because of the additional unit of stock.

Despite the complications described above, we can find the policy with minimum reduced costs by determining upper bounds on the stock levels S_{ik}, $k \geq 0$. As a result, we limit the solution space that we need to consider. The four observations below allow us to find these upper bounds, with observations 1 and 2 focusing on the upper bound for S_{i0}, while observations 3 and 4 focus on the bounds for the customer stock levels S_{ik} $(k \geq 1)$. Once we have described how to obtain all bounds, we detail how we find the best item policy in the restricted solution space. From now on, we let Z denote the minimum reduced costs (i.e., $Z = \min_{b_i} RED_i(b_i) = \min_{\boldsymbol{S}_i} RED_i(\boldsymbol{S}_i)$).

Observation 1: *We can find an upper bound S_{i0}^{max} on S_{i0}.* Let $RED_i^*(S_{i0})$ denote the minimal reduced costs for a given S_{i0}, calculated over the stock levels at all customer locations $S_{ik}(k \geq 1)$. We thus have that $Z = \min_{S_{i0}} RED_i^*(S_{i0})$. Equation (6.4) shows that $RED_i^*(S_{i0}) \geq h_{i0}S_{i0} - v_i \; \forall S_{i0}$, since the total reduced costs include the holding costs at the warehouse, and $u_k \leq 0$. We ignore the reduced cost elements related to the various customers (i.e., the elements $h_{ik}S_{ik}$, $\gamma_{ik}(\boldsymbol{S}_i)m_{ik}EC_{ik}^{em}$, and $\frac{u_k m_{ik}}{M_k} EW_{ik}(\boldsymbol{S}_i)$ per customer k), as these depend on the other stock levels as well. Conversely, we can determine an upper bound on Z, which we denote by RED_i^{UB}. We find RED_i^{UB} as the reduced costs of any item policy. Overall, we have the following relations:

$$\min_{S_{i0}} h_{i0}S_{i0} - v_i \leq \min_{S_{i0}} RED_i^*(S_{i0}) = Z \leq RED_i^{UB}$$

From the above relation, we conclude that it is not beneficial to consider values of S_{i0} for which $h_{i0}S_{i0} - v_i > RED_i^{UB}$. Also, it is not beneficial if $h_{i0}S_{i0} - v_i > 0$, since we aim to find an item policy with negative reduced costs. Hence, we find S_{i0}^{max} as the smallest value of S_{i0} for which $h_{i0}S_{i0} - v_i$ exceeds $\min\{RED_i^{UB}, 0\}$.

We choose RED_i^{UB} as minimal reduced costs over S_{i0} given that all customer stocks S_{ik} are zero $(k \geq 1)$. Then, the resulting reduced cost function is easy to optimize, as it is convex in S_{i0}, see Kranenburg and Van Houtum (2007).

Observation 2: *We can further limit the relevant values for S_{i0}.* By tightening the lower bound on $RED_i^*(S_{i0})$ as given in Observation 1, we can further limit the values of S_{i0} that we must consider. To obtain this bound, we first give an explicit expression for $RED_i^*(S_{i0})$, i.e.,

$$RED_i^*(S_{i0}) = h_{i0}S_{i0} - v_i + \min_{S_{ik}, k \geq 1} \sum_{k=1}^{K} \left\{ h_{ik}S_{ik} + \gamma_{ik}(\boldsymbol{S}_i)m_{ik}EC_{ik}^{em} - \frac{u_k m_{ik}}{M_k} EW_{ik}(\boldsymbol{S}_i) \right\} \tag{6.5}$$

Notice that $h_{i0}S_{i0} - v_i$ has a fixed value. Hence, if we are able to find a lower bound on the third element of $RED_i^*(S_{i0})$ (i.e., the element that depends on the customer stock levels S_{ik} $(k \geq 1)$), we also get a lower bound on $RED_i^*(S_{i0})$.

Given that we are able to optimize the reduced costs for a *single customer* k, we obtain a lower bound on the third element of $RED_i^*(S_{i0})$ by optimizing the reduced costs per customer and subsequently summing over the values obtained per customer, i.e.,

$$\min_{S_{ik},k\geq 1} \sum_{k=1}^{K} \left\{h_{ik}S_{ik} + \gamma_{ik}(\boldsymbol{S}_i)m_{ik}EC_{ik}^{em} - \frac{u_k m_{ik}}{M_k}EW_{ik}(\boldsymbol{S}_i)\right\} \geq \sum_{k=1}^{K} \min_{S_{in}\ (n\geq 1)} RED_{ik}(S_{i0},\boldsymbol{S}_i),$$

where

$$RED_{ik}(S_{i0},\boldsymbol{S}_i) = h_{ik}S_{ik} + \gamma_{ik}(\boldsymbol{S}_i)m_{ik}EC_{ik}^{em} - \frac{u_k m_{ik}}{M_k}EW_{ik}(\boldsymbol{S}_i).$$

Note that $RED_{ik}(S_{i0},\boldsymbol{S}_i)$ is smallest when $\gamma_{ik}(\boldsymbol{S}_i)$ and $EW_{ik}(\boldsymbol{S}_i)$ are small, which correspond to a high service level at customer k. As shown at the start of this section, the service level at a customer k deteriorates when stock is placed at other customers. Hence, the service level at k is highest when no stock is kept at other customers. We thus find $\min_{S_{in}\ (n\geq 1)} RED_{ik}(S_{i0},\boldsymbol{S}_i)$ by setting $S_{in} = 0$ $(n \neq k, 0)$ and computing $RED_{ik}(S_{i0},\boldsymbol{S}_i)$ over interval $S_{ik} \in [0, \dots, S_{ik}^{MAX}]$, with S_{ik}^{MAX} following from observation 3.

Observation 3: *We can find a rough upper bound S_{ik}^{MAX} on $S_{ik}(k \geq 1)$.* An increase of S_{ik} can only benefit the service level at customer k. Hence, we find S_{ik}^{MAX} once the additional holding costs of increasing S_{ik} outweigh the maximum reduction in that customer's emergency shipment and waiting time costs, i.e. S_{ik}^{MAX} is the smallest S_{ik} for which $h_{ik} > \gamma_{ik}(\boldsymbol{S}_i)m_{ik}EC_{ik}^{em} - \frac{u_k m_{ik}}{M_k}EW_{ik}(\boldsymbol{S}_i)$. To ensure that S_{ik}^{MAX} is sufficiently large, we need upper bounds on $\gamma_{ik}(\boldsymbol{S}_i)$ and $EW_{ik}(\boldsymbol{S}_i)$. We find such bounds by assuming that demand at customer k can only be met from on-hand stock at that customer (i.e. that customer has no access to warehouse stock). Then, we have the worst-case scenario in terms of service level. The resulting system can be analyzed as an Erlang-loss system with S_{ik} servers.

Observation 4: *For a given value of S_{i0}, we can find a tighter upper bound on S_{ik} $(k \geq 1)$, denoted by $S_{ik}^{max}(S_{i0})$.* As in observation 3, we find $S_{ik}^{max}(S_{i0})$ when the holding costs of increasing S_{ik} exceed the emergency shipment and waiting time costs of customer k. Compared to S_{ik}^{MAX} (in Observation 3) , we now use more accurate values for $\gamma_{ik}(\boldsymbol{S}_i)$ and $EW_{ik}(\boldsymbol{S}_i)$, which we find by also considering the stock kept at other locations in the system. Specifically, we set all other customer stock levels S_{in} $(n \neq k, 0)$ to S_{in}^{MAX} and then we determine $\gamma_{ik}(\boldsymbol{S}_i)$ and $EW_{ik}(\boldsymbol{S}_i)$. As the service level at customer k is lowest when the stock levels at other customers are large, the values for $\gamma_{ik}(\boldsymbol{S}_i)$ and $EW_{ik}(\boldsymbol{S}_i)$ will still be sufficiently large.

In addition to these observations, we empirically find that the optimal value of S_{ik} for a given S_{i0}, denoted by $\hat{S}_{ik}(S_{i0})$, generally lies between two thresholds $S_{ik}'(S_{i0})$ and $S_{ik}''(S_{i0})$. We find

$S'_{ik}(S_{i0})$ as the value of $S_{ik} \in \{0, \dots, S_{ik}^{max}(S_{i0})\}$ that minimizes $RED(\boldsymbol{S}_i)$ when all other customer stock levels S_{in} $(n \neq k, 0)$ are set to their upper bounds $S_{in}^{max}(S_{i0})$. Similarly, we find $S''_{ik}(S_{i0})$ as the optimal S_{ik} when all other customer stock levels are set to their lower bounds $S_{in}^{min}(S_{i0}) = 0$. Since we optimize S_{ik} in two extreme cases (the remaining customer stocks are either at their maximum or at their minimum), we expect the optimal value of S_{ik} to lie between $S'_{ik}(S_{i0})$ and $S''_{ik}(S_{i0})$. Note that $S'_{ik}(S_{i0})$ and $S''_{ik}(S_{i0})$ in fact give us new bounds on $\hat{S}_{ik}(S_{i0})$. We can thus repeat the mentioned steps (i.e. we can find new values for $S'_{ik}(S_{i0})$ and $S''_{ik}(S_{i0})$) by updating $S_{ik}^{min}(S_{i0})$ and $S_{ik}^{max}(S_{i0})$. We proceed in this way until the bounds stabilize (either because the values for $S_{ik}^{min}(S_{i0})$ and $S_{ik}^{max}(S_{i0})$ no longer change or because $S'_{ik}(S_{i0}) = S''_{ik}(S_{i0})$ for all customers k).

Overall, the procedure works as follows: We increase S_{i0} from zero up to S_{i0}^{max} with step size 1. In each step, we first verify whether it is beneficial to consider that value of S_{i0} (see observation 2) and, if so, we compute $S'_{ik}(S_{i0})$ and $S''_{ik}(S_{i0})$ for each customer k. Then, we look for the combination of customer stock levels that has minimum reduced costs, given $S'_{ik}(S_{i0}) \leq S_{ik} \leq S''_{ik}(S_{i0})$.

6.3.2.3 The formal steps of the column generation procedure

Full column generation procedure

1. Find S_{i0}^{max} from observation 1.
2. For each customer k, find a rough upper bound S_{ik}^{MAX} on the optimal stock level (observation 3).
3. For each $S_{i0} \in \{0, \dots, S_{i0}^{max}\}$ do:
    a. Determine whether the tighter lower bound for $RED_i^*(S_{i0})$ (see observation 2) is below the best reduced cost upper bound so far (either $\min\{RED_i^{UB}, 0\}$ from observation 1 or the most recent value for Z). If not, skip steps 3b through 3e.
    b. Find a tighter upper bound $S_{ik}^{max}(S_{i0})$ on the optimal stock level for customer k (see observation 4).
    c. Find thresholds $S'_{ik}(S_{i0})$ and $S''_{ik}(S_{i0})$ for each customer k.
    d. Find the customer stock combination $[S_{i1}, \dots, S_{iK}]$ that minimizes $RED_i(\boldsymbol{S}_i)$, with $S'_{ik}(S_{i0}) \leq S_{ik} \leq S''_{ik}(S_{i0})$.
    e. If the solution is the best so far, store it. Also store the related reduced costs as Z. If $h_{i0}(S_{i0}+1) - v_i$ (i.e. the lower bound on the reduced costs for $S_{i0}+1$) exceeds Z, exit the procedure.

Next, we give further details on steps 1 through 3c.

Step 1. Finding S_{i0}^{max}.

1. Determine an upper bound RED_i^{UB} on the reduced costs.
    a. Set all customer stocks S_{ik} to zero $(k \geq 1)$.
    b. Find the S_{i0} that minimizes $RED_i(\boldsymbol{S}_i)$. Set RED_i^{UB} to this value.
2. Find S_{i0}^{max} as the smallest S_{i0} for which $h_{i0}S_{i0} - v_i$ exceeds $\min\{RED_i^{UB}, 0\}$.

Step 2. Finding S_{ik}^{MAX} for each customer k.

1. Consider an Erlang Loss system with S_{ik} servers and replenishment rate $\mu_{ik} = 1/(T_{ik}^{reg} + T_{i0}^{reg})$. Our performance measures now only depend on S_{ik}: we find $\gamma_{ik}(S_{ik})$ as the probability of all servers being occupied, with $EW_{ik}(S_{ik})$ being equal to $T_{ik}^{em}\gamma_{ik}(S_{ik})$.
2. Find S_{ik}^{MAX} as the smallest S_{ik}-value for which h_{ik} exceeds $\gamma_{ik}(S_{ik})m_{ik}EC_{ik}^{em} - \frac{u_k m_{ik}}{M_k}EW_{ik}(S_{ik})$. From that moment, the reduced costs cannot improve further.

Step 3a. Compute a tighter lower bound on the minimal reduced costs $RED_i^*(S_{i0})$.

1. For each customer k do:
    a. Set all other customer stocks S_{in} $n \neq k, 0$ to 0.
    b. Find the value for $S_{ik} \in [0, \dots, S_{ik}^{MAX}]$ that minimizes $RED_{ik}(S_{i0}, \boldsymbol{S}_i)$.
2. The bound on $RED_i^*(S_{i0})$ now equals $h_{i0}S_{i0} + \sum_{k=1}^{K} \min_{S_{in}\,(n\geq 1)} RED_{ik}(S_{i0}, \boldsymbol{S}_i) - v_i$.

Step 3b. Finding $S_{ik}^{max}(S_{i0})$ for any S_{i0}.

1. Set all other customer stocks S_{in} $n \neq k, 0$ to S_{in}^{MAX}.
2. $S_{ik}^{max}(S_{i0})$ is the smallest S_{ik} for which $h_{ik} > \gamma_{ik}(\boldsymbol{S}_i)m_{ik}EC_{ik}^{em} - \frac{u_k m_{ik}}{M_k}EW_{ik}(\boldsymbol{S}_i)$.

Step 3c. Finding $S'_{ik}(S_{i0})$ and $S''_{ik}(S_{i0})$.

1. Set all customer lower bounds $S_{ik}^{min}(S_{i0})$ $(k \geq 1)$ to 0.
2. Find $S'_{ik}(S_{i0})$ for each customer k.
    a. Set all other customer stocks S_{in} $n \neq k, 0$ to $S_{in}^{max}(S_{i0})$.
    b. Find $S'_{ik}(S_{i0})$ as the $S_{in} \in \{S_{ik}^{min}(S_{i0}), \dots, S_{ik}^{max}(S_{i0})\}$ that minimizes $RED_i(\boldsymbol{S}_i)$.
3. Find $S''_{ik}(S_{i0})$ for each customer k.
    a. Set all other customer stocks S_{in} $n \neq k, 0$ to $S_{in}^{min}(S_{i0})$.
    b. Find $S''_{ik}(S_{i0})$ as the $S_{in} \in \{S_{ik}^{min}(S_{i0}), \dots, S_{ik}^{max}(S_{i0})\}$ that minimizes $RED_i(\boldsymbol{S}_i)$.
4. Exit if (i) $S'_{ik}(S_{i0}) = S''_{ik}(S_{i0})$ $\forall k \geq 1$, or (ii) neither $S'_{ik}(S_{i0})$ nor $S''_{ik}(S_{i0})$ has changed compared to the previous iteration for any customer. Otherwise, set $S_{ik}^{min}(S_{i0})$ to $\min\{S'_{ik}(S_{i0}), S''_{ik}(S_{i0})\}$ and $S_{ik}^{max}(S_{i0})$ to $\max\{S'_{ik}(S_{i0}), S''_{ik}(S_{i0})\}$ and proceed to step 2.

We cannot guarantee that the procedure just described always finds the policy with minimal reduced costs, since the optimal value for S_{ik} given S_{i0} does not always lie between thresholds $S'_{ik}(S_{i0})$ and $S''_{ik}(S_{i0})$ (as assumed in step 3d). Still, in a computational experiment of 100 instances – with the number of items either 20 or 50 and the number of customers either 8 or 16 – we always found the same lower bound as when we used a complete enumeration procedure for column generation (where all relevant combinations of stock levels were considered).

6.3.3 Near-optimal integer solution

The optimal solution to the LP-relaxation might be fractional, i.e. it might be that a combination of item policies has been selected for certain items. Therefore, we also require an approach to find a near-optimal integer solution. As in Chapter 3 and 4, we obtain such a solution by solving the *integer* problem $(P2)$ using a limited set of item policies.

As before, we start with the set of item policies generated when solving the LP-relaxation of $(P2)$. This policy set might contain many item policies: when constructing our initial policy set (Section 6.3.2.1), we included all found policies in B_i. We also added additional policies during column generation. Such a large policy set is not an issue when solving an LP-relaxation, but computation times might explode when solving the integer problem. Therefore, we eliminate all *dominated* item policies from the LP-relaxation set before solving the integer problem. Dominated policies have both higher costs and higher waiting times than at least one other policy in the policy set. As a result, such policies will never be chosen and they can thus be eliminated from the policy set without sacrificing solution quality.

6.4 Evaluation of an item policy

We now shortly describe how we can obtain performance measures for an item policy using Markov chain analysis. We do so for any differentiation strategy (DS, CLP) and shipment option (backordering, emergency shipments). A detailed description of the evaluation procedure for each policy type has been given before either in other chapters of this dissertation or in the referred literature.

Under CLP, we only need to analyze the warehouse to obtain the needed performance measure values, as this is the only location where stock may be kept. Indeed, under emergency shipments we find performance measures directly from the distribution of the *pipeline* to the warehouse: if we have fewer than $S_{i0} - C_i$ items in the pipeline, demands from both customer classes are satisfied from warehouse stock. Otherwise, only demand from premium customers is met from on-hand stock if possible, with non-premium demand being covered by emergency shipments. Kranenburg and Van Houtum (2008) further detail how an item policy can be analyzed under lost sales. In contrast, under backordering we require both the number of items in the pipeline and the number of class 2 backorders to analyze the system. We described the evaluation procedure for this system in Chapter 3.

Under DS, we can keep stock at both the warehouse and the customers. As a result, we obtain a two-echelon system for analysis purposes. Under full backordering, the analysis of such a system has been considered by Graves (1985) and Wong et al. (2007a) amongst others. Those authors first analyze the warehouse to obtain per customer the distribution of items outstanding at the warehouse (i.e. the items that still need to be shipped to that customer). Using this distribution, the authors determine the distribution of the pipeline to each customer, which consists of the items outstanding for that customer at the warehouse and the number of items in transit from the warehouse to that customer. Finally, the authors use these pipeline distributions to determine the expected number of backorders at each customer (resulting in an expected waiting time through Little's Law). Under emergency shipments, analysis is complicated by the fact that the distribution of outstanding items at the warehouse depends on the availability of stock in the entire system. We described the analysis approach for this system in Chapter 5.

6.5 Computational experiment

In this section, we describe our computational experiment. We give the objectives in Section 6.5.1, the experiment design in Section 6.5.2, and the results in Section 6.5.3.

6.5.1 Experiment objectives

First, we investigate the performance of our optimization approach for the DS strategy with emergency shipments in terms of solution quality and computation time. Second, we determine the added value of using dedicated customer stocks for differentiation by comparing the results under DS to those under a one-size-fits-all strategy (OSFA) and those under CLP. Finally, we consider the added value of the COMBO strategy where the differentiation mode (i.e. DS or CLP) may differ per item.

6.5.2 Experiment design

Table 6.1 shows the parameter values we used for our problem instances.

Parameter	Values
Number of items I	20; 100
Number of customers K	8; 16
Percentage premium customers (% of K)	12.5; 25
$(W_1^{max}; W_2^{max})$ (hours)	(2;8); (2;16)
Intervals for demand rates m_{ik} (per day)	[0.002 – 0.025]; [0.002 –0.075]
Intervals for holding costs h_i (per day)	[0.1 –10]; [0.1 – 100]
T_{i0}^{reg} (days)	5; 15
T_{ik}^{reg} (hours)	0.5; 1.5
T_{ik}^{em} (as a % of T_{i0}^{reg})	10; 20
EC_{ik}^{em} (per shipment)	1000

Table 6.1 Parameter settings in problem instances.

We express some values in days and others in hours, with 1 day equal to 24 hours. Except for the demand rates and holding costs, the parameter values are the same for all items, and if applicable for all customers, in a problem instance. The demand rates m_{ik} and holding costs h_i are randomly drawn from uniform distributions on the specified intervals. For simplicity, we use the same holding cost value at all locations in a problem instance. Also, we expect that the main factors influencing the holding cost value – such as opportunity cost on the investment and risk of obsolescence – do not depend on the item's location.

Our parameter values are partially based on those by Wong et al. (2007a) who consider a similar setting. Note that T_{ik}^{reg} has very small values compared to the other shipment times: under CLP, the mean waiting time for each customer will be at least T_{ik}^{reg}. Hence, we can only find solutions under CLP if T_{ik}^{reg} is smaller than W_1^{max}. Our demand rates have been chosen such that we mainly consider slow moving items: the maximum demand rate per year is 27 units. Furthermore, case studies at multiple companies have shown that spare parts can be very expensive, with values exceeding 100,000 euro's. We assume that an item's annual holding costs are 25% of its value – a common assumption in practice – and thus consider item values up to 146,000 euro's.

For each combination of parameters 1 through 9, we create 3 samples of demand rates and holding costs, thereby ensuring that our results are not sensitive to the specific values of one sample. In total, we have 2304 instances: 3 (samples) * 2^8 (parameters 1..8) = 768 instances for each of the strategies (i) backordering, (ii) emergency shipments with T_{ik}^{em} as 10% of T_{i0}^{reg}, (iii) emergency shipments with T_{ik}^{em} as 20% of T_{i0}^{reg}.

6.5.3 Results

We discuss the results for each experiment objective in a separate section.

6.5.3.1 Performance of the optimization approach

We determine the solution quality of the optimization approach by comparing the integer solutions found to their lower bounds. The solution quality is expressed as a gap to the lower bound, i.e. $(TC_{IP} - TC_{LB})/TC_{LB}$ with TC_{LB} and TC_{IP} respectively denoting the lower bound and integer solution value.

Table 6.2 summarizes the results on solution quality and computation times. The solution quality is very good, with gaps below 1%. Furthermore, for instances with many items the maximum gap is only 0.1%. The approach will thus work well for practical instances. The computation time of an instance is roughly 3 minutes on average, with over 98% of the instances having a computation time below 30 minutes. The maximum computation time is 218 minutes. Computation times are largest when there are many customers, and when item demand rates and the shipment time T_{i0}^{reg} to the warehouse are large.

Parameter	Values	Gap to lower bound		Computation time (mins)	
		Average	Maximum	Average	Maximum
I	20	0.2%	1.0%	0.7	80
	100	0.0%	0.1%	4.7	218
K	8	0.1%	0.8%	0.3	4
	16	0.1%	1.0%	5.0	218
T_{i0}^{reg}	5	0.1%	1.0%	0.8	8
	15	0.1%	0.8%	4.6	218
m_{ik} – interval	[0.002 – 0.025]	0.1%	1.0%	1.0	12
	[0.002 – 0.075]	0.1%	0.6%	4.4	218
Overall		0.1%	1.0%	2.7	218

Table 6.2 Performance of the optimization approach for DS with emergency shipments.

6.5.3.2 The added value of dedicated stocks (DS)

We compare the solutions under dedicated stocks to those under one-size-fits-all (OSFA) and under critical level policies (CLP). OSFA serves as a benchmark as it is a special case of both DS and CLP. We express the added value of the latter two strategies in terms of a relative cost saving over OSFA $(TC_{OSFA} - TC_{DIFF})/TC_{OSFA}$. Here, TC_{OSFA} denotes the costs under OSFA and TC_{DIFF} those under a differentiation strategy. Figure 6.4 shows the savings under both backordering and emergency shipments. Under backordering, the average savings under DS and CLP are 13% and 19% respectively (with maximum savings for both strategies close to 40%). Under emergency shipments, the average savings are 5% for each strategy.

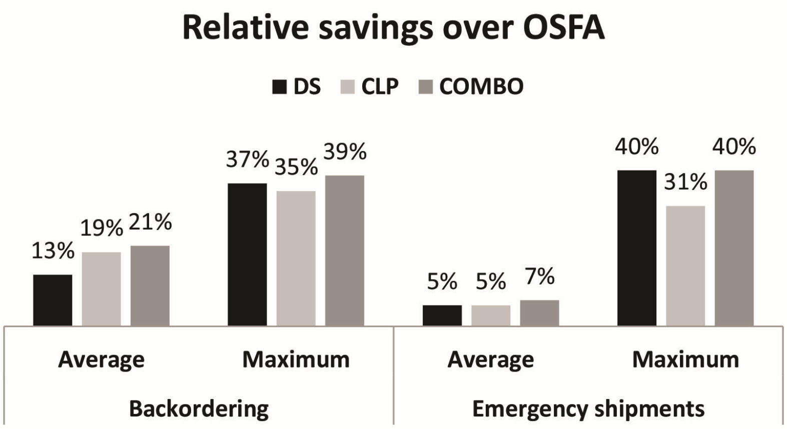


Figure 6.4 Relative savings of DS, CLP and COMBO over OSFA.

Although CLP generally outperforms DS under backordering, Figure 6.5 shows that DS is particularly beneficial when there are relatively few (premium) customers and when the shipment time from the warehouse to the customers is large. In those circumstances, the savings are comparable to those under CLP. When there are few premium customers, we need little dedicated stock to effectively apply differentiation. Storing items at customer sites also becomes more interesting (and possibly even necessary) when shipment times to customers are relatively large.

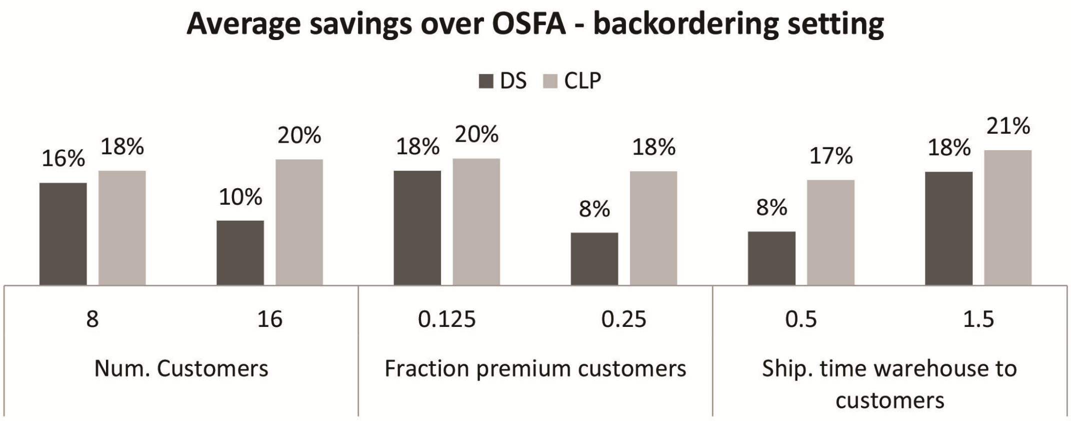


Figure 6.5 Average savings of DS and CLP over OSFA – backordering setting.

Under emergency shipments, the parameter values of an instance heavily influence the height of the savings for both DS and CLP, as shown in Figure 6.6. A key observation is that it is clearly not beneficial to use either DS or CLP in combination with emergency shipments for inexpensive items. For such items, emergency shipments are relatively expensive. This prompts the model to keep high stock levels simply to limit emergency shipment costs as opposed to keeping stock to reduce waiting times. Indeed, we then find that the aggregate waiting times per customer are much lower than the corresponding thresholds. In contrast, when item holding costs are high, stock is mainly kept to satisfy premium waiting time requirements. As a result, the aggregate waiting times for premium customers are very close to W_1^{max}. DS and CLP also lead to large savings when the regular shipment times – both to the warehouse and from the warehouse to the customers – are relatively large.

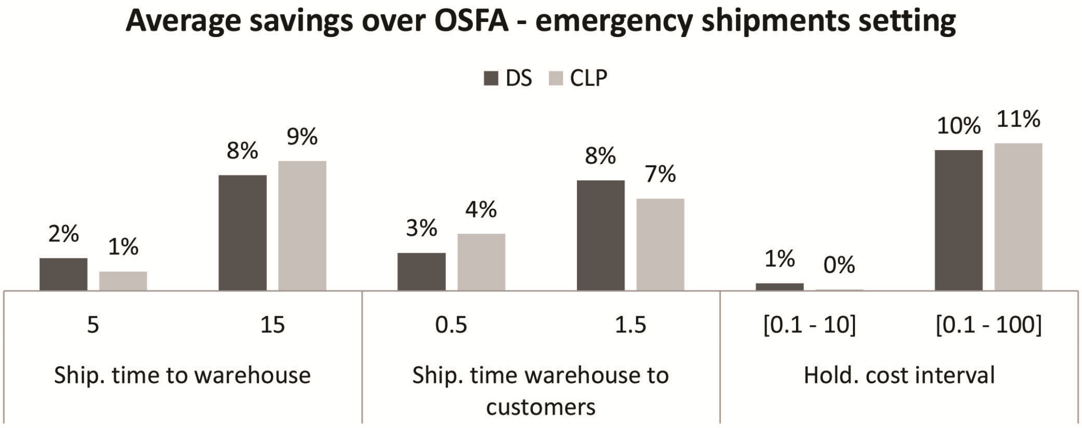


Figure 6.6 Average savings of DS and CLP over OSFA - emergency shipments setting.

Overall, the savings under DS are close to those under CLP. In some cases, the savings under the two strategies are even comparable, especially when emergency shipments are used for demand that cannot be met from the system. Of the two strategies, we expect DS to be the easiest one to implement in practice. Also, DS might be the only viable option if the shipment times to the customers exceed the waiting time thresholds. For instance, if T_{ik}^{reg} is 2 hours while W_1^{max} equals 1 hour, some stock must be kept at premium customers' sites to ensure that the average waiting times are within target. In that setting (with W_2^{max} either 4 or 8 hours and the other parameters as in Table 6.1), the average fraction of items kept at premium customers' sites increases greatly compared to the setting with the original shipment times and waiting time thresholds, see Figure 6.7. In the figure, 'BO' stands for backordering, while 'ES' stands for emergency shipments. Notice that we rarely keep stock at non-premium sites, even when we reduce W_2^{max} by half.

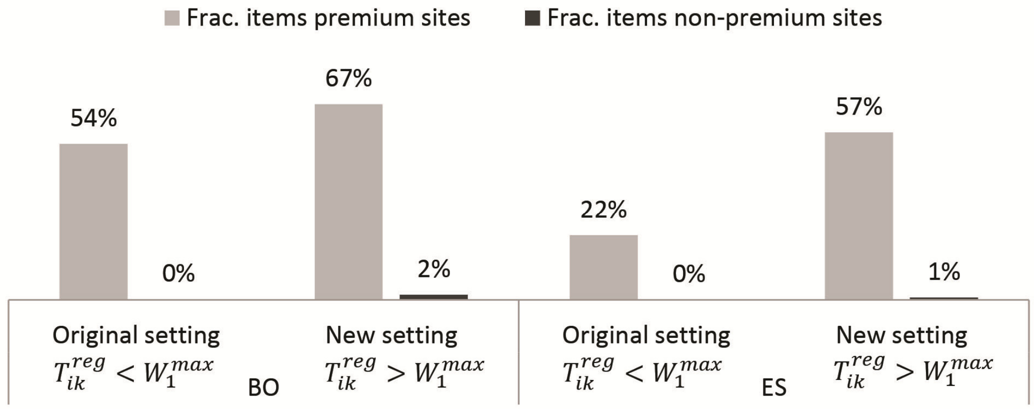


Figure 6.7 Fraction of items kept a customer sites both for the original parameter setting and for a setting with tighter waiting time restrictions. 'BO': backordering, 'ES': emergency shipments.

For the original problem setting (Table 6.1), we analyzed the items kept at customer sites. We found that dedicated stocks are mainly kept of inexpensive, fast moving items. Figure 6.8 shows both the average holding costs and demand rate of items kept at premium customers' sites compared to the overall mean item holding cost and demand rate. We ignore stock kept at non-premium sites, as this rarely occurs. The results are based on instances with holding cost interval [0.1 – 100] and demand rate interval [0.002 – 0.075]; the figures are similar for different intervals.

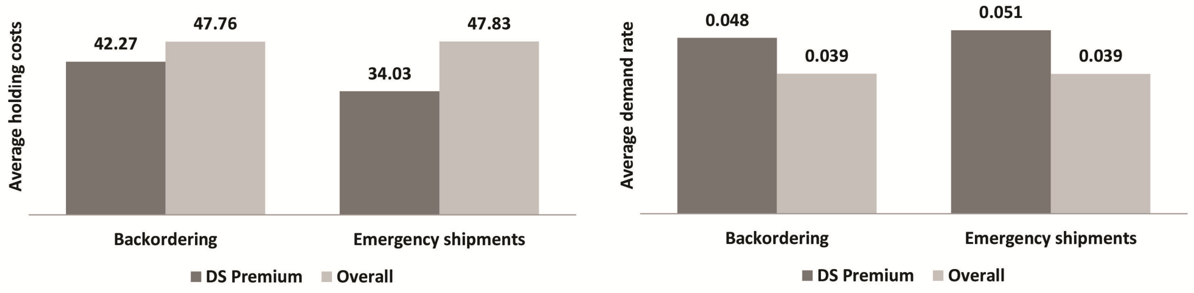


Figure 6.8 Characteristics of items kept at premium customers' sites (compared to the overall values).

6.5.3.3 The COMBO strategy

Figure 6.4 shows the savings of the COMBO strategy. Notice that these savings are not much larger than those under DS or CLP. Still, a mix of dedicated stocks and critical level policies is used in many solutions: under emergency shipments, such a mix is used for 30% of all instances overall, and for 58% of instances with relatively high holding costs (i.e. holding cost interval $[0.1 - 100]$). Under backordering, such a mix is used for over 70% of all instances.

In the instances where both dedicated stocks are critical level policies are used, we use dedicated stocks for roughly 20% of the items, and critical level policies for roughly 49% of the items. However, the actual frequency with which a strategy is used depends on the parameter values, see Figure 6.9.

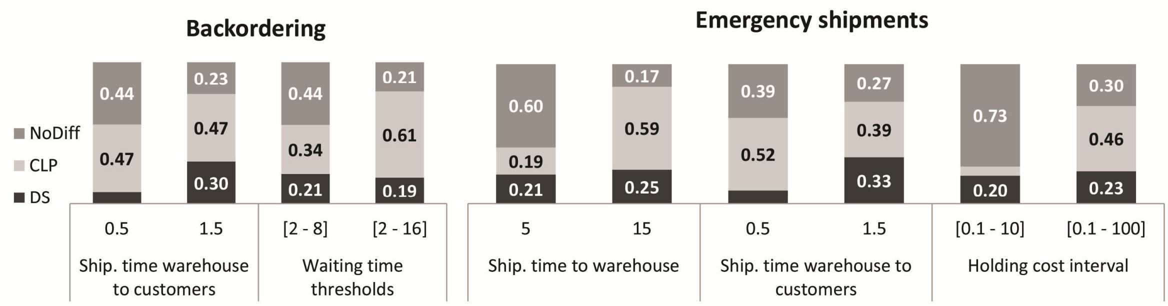


Figure 6.9 Average fraction of items per strategy (incl. the option of not using differentiation) in the combo approach – instances where both DS and CLP are used.

A detailed analysis of the instances where a combination of DS and CLP is used shows that DS is mainly used for very cheap items, while CLP is used for relatively expensive items. Figure 6.10 shows this for the instances with a holding cost interval of $[0.1 - 100]$; we find similar figures for other holding cost intervals. This observation is logical: if items are inexpensive, it is best to keep them at customer sites to minimize waiting time. Conversely, it will be too expensive to keep high value items at all sites. Therefore, stock should be centralized, with a critical level policy used for differentiation purposes. We were unable to draw clear conclusions on the item demand rates per strategy.

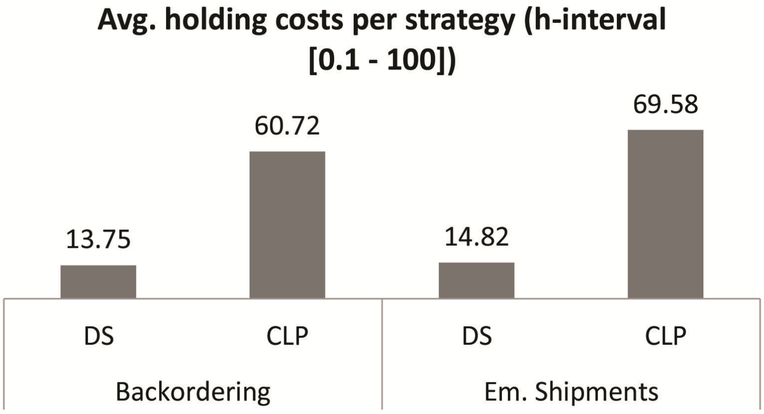


Figure 6.10 The average item holding costs per differentiation strategy.

6.6 Conclusions

In this chapter, we investigated the use of dedicated customer stocks as a differentiation tool. For a multi-item two-class system with dedicated stocks and emergency shipments, we developed an optimization approach that works well: the integrality gaps are always below 1% and for instances with many items the maximum gap is even 0.1%. Furthermore, computation times are 3 minutes on average and remain below 30 minutes for most problem instances we tested. Other key conclusions that we draw from this chapter are:

- **Dedicated stocks have significant added value.** The average savings of DS compared to an approach where no differentiation is used (i.e. OSFA) are 13% under backordering and 5% under emergency shipments, with maximum savings equal to 37% and 40% respectively. Furthermore, the savings obtained under DS are comparable to those under CLP (who has average savings of 19% and 5% under backordering and emergency shipments respectively).
- **Under emergency shipments, both dedicated stocks and critical level policies only have added value if holding costs are relatively high.** If holding costs are low, stock is kept to avoid expensive emergency shipments. As a result, no differentiation takes place, as both premium and non-premium aggregate waiting times are below their thresholds.
- **Dedicated stocks are very beneficial, if not necessary, when the shipment time to customers is large.** As shipment times to customers increase, it might no longer be possible to only keep stock centrally if customers have high service requirements. So far, this fact has been largely ignored in literature on critical level policies, where the shipment time to customers is assumed to be negligible as it is often much smaller than the shipment time to the warehouse.
- **We find relatively small additional gains under the combined strategy (COMBO) compared to DS or CLP.** The practical relevance of this observation is that dedicated stocks indeed have significant added value, as we do not find much greater savings by adding critical levels.
- **Under the combined strategy, we keep dedicated stocks of inexpensive items, while using critical level policies for expensive items.** Keeping dedicated stocks of inexpensive items greatly reduces waiting time at little expense. Conversely, it is too expensive to keep high value items at customer sites. Instead, stock should be centralized to benefit all customers, with some stock reserved for premium customers.

So far, we did now allow **both DS and CLP to be used for the same item**. Still, we do not expect further savings if this option had been available: for the parameter values in Table 6.1, we compared our COMBO strategy under backordering to a variant where CLP and DS could be used for the same item. The savings of the new strategy where at most 0.6% of the original

COMBO strategy. We expect that the lack of savings is caused by the fact that DS is beneficial for a different set of items than CLP: as shown in Section 6, DS is beneficial for inexpensive items, while CLP is used for expensive items.

In this chapter, and Chapters 3 and 4, we have considered tools for applying service differentiation in spare parts supply. However, in Section 1.2 we have shown that the overall system downtime may depend on various other resources, amongst others service engineers. In Chapter 7, we therefore consider priority mechanisms when assigning service engineers to customers as a differentiation tool.

Chapter 7

Priority mechanisms when assigning service engineers to customers[9]

7.1 Introduction

In the previous chapters, we investigated various control options for applying differentiation in spare parts supply. However, a service provider also depends on other resources besides spare parts when providing service to its customers. Indeed, in the printing and copying equipment industry for instance the availability of service engineers is the main bottleneck in ensuring that all service level agreements are met. In this chapter, we therefore consider human resources, where we focus on the assignment of a set of engineers to a group of customers with varying service level requirements, thereby meeting our 7th research objective.

When a customer's system breaks down, an engineer diagnoses the cause of the failure and then repairs the system. A key performance indicator is the *response time,* i.e., the time between the reporting of a failure and the arrival of the engineer at the customer's site. Naturally, the response time is influenced by the manner in which service engineers are assigned to customers. In this chapter, we focus on *priority assignment*, i.e., an available engineer is assigned to the customer with the highest priority as opposed to the customer that has been waiting longest. As a result, customers with high service level requirements exhibit low response times at the expense of other customers. Given this assignment mechanism, we aim to accurately estimate the waiting times for the various classes of customers, with the customer's class indicating the required level of service. As we aim for a high probability that service level targets are met, mean waiting times alone are insufficient: We need the *waiting time distribution* per customer class. Then, combined with the travel time to customers, we have an estimate of the response times per customer class, and hence of the service provider's performance on his response time targets. In the remainder of the chapter, we assume that travel times to customers are known. Therefore, these times can be ignored, with the response times only depending on the waiting times for engineers to become available.

[9] This chapter is based on the working paper "Approximations for the waiting time distribution in an $M/G/c$ priority queue" by A. Al Hanbali, E.M. Alvarez and M.C. van der Heijden.

7.1. Introduction

We estimate the waiting time distribution per customer class by modeling the system as a multi-class, non-preemptive $M/G/c$ priority queue with identical service time distributions in the distinct classes. We consider this system for the following reasons:

- **Poisson arrivals:** In practice, complex systems seem to have a constant hazard function, since failures arise from various causes, thus appearing completely random. Such randomness holds even more in systems with a high electronic content, where material-based fatigue does not occur. Therefore, Poisson arrivals are often a valid assumption. We have observed such behavior for printing and copying equipment amongst others, and Jardine and Tsang (2006) give additional cases where this assumption is reasonable in Section 3.5.5.
- **Non-preemptive priorities:** Once an engineer has been assigned to a customer, he will first service that customer before proceeding to another, even if a higher-priority customer appears in the meantime. Hence, we consider a non-preemptive discipline.
- **Equal service time distributions**: We consider the setting where all customers have similar types of systems. As a result, the failure behavior of the system, and hence the distribution of the time to repair the system, will be the same at all customers.

As shown in Section 1.7.5, the literature on multi-class $M/G/c$ queues with a non-preemptive priority service discipline is limited. For the case where $c = 1$, there is far more literature. To our best knowledge, only three papers consider a model similar to ours: Wagner (1997) uses an approximate approach based on matrix geometric methods to primarily estimate the mean waiting time per class in a multi-class model with a generalized Markovian arrival process and a phase-type service time distribution. In contrast, Williams (1980) and Jagerman and Melamed (2003) both estimate the waiting time distributions per class in an $M/G/c$ queue. Williams focuses on 2 priority classes that have identical service time distributions, while Jagerman and Melamed (2003) consider multiple classes with the service rates possibly differing among classes. Both papers make the following approximations: (i) the delay probability in an $M/G/c$ queue is approximated by the same probability in an $M/M/c$ queue with equal arrival rates and service rate, and (ii) when all servers are occupied, the Laplace-Stieltjes Transform (LST) of the service time in an $M/G/c$ queue is approximated by that of the service time in an $M/G/1$ queue when the server works c times as fast as in the original $M/G/c$ queue. Williams (1980) states that the approximations above are exact both for the single server $M/G/1$ and the multi-server $M/M/c$ queue. Hence, it follows that the mean waiting time for a class-k customer satisfies the following well-known scaling approximation, which can easily be derived by conditioning on the waiting time when all servers are occupied, see, e.g., Buzen and Bondi (1983):

$$\frac{\mathbb{E}[W_k(M/G/c)]}{\mathbb{E}[W_k(M/M/c)]} \approx \frac{\mathbb{E}[W_k(M/G/1)]}{\mathbb{E}[W_k(M/M/1)]},$$

where the server in the $M/G/1$ and $M/M/1$ queues works c times faster than in the related $M/G/c$ and $M/M/c$ queues. This scaling approximation also holds for the second moment of the waiting time.

Neither Williams (1980) nor Jagerman and Melamed (2003) validate the quality of their results. Still, we found that Williams' method can be inaccurate, especially in settings with many servers. Our main contributions in this chapter are: (i) we *refine the approximation assumption* of Williams (1980) and Jagerman and Melamed (2003), and from that we obtain very accurate methods to estimate the waiting time distribution per class in a system with multiple priority classes. As we will show in a computational experiment, our methods generally outperform Williams' method, particularly for the highest priority classes. Also, (ii) we present options to simplify the analysis such that large systems (with many servers and a phase-type service time distribution with many phases) can still be quickly analyzed with a limited decrease in accuracy. Finally, (iii) we apply our methods to determine service level performance in a practical setting.

In the remainder of the chapter, we first describe our model in Section 7.2. There, we also globally present the analysis approach for this model. A key building block of the approach is the analysis of a single-class system, which we give in Section 7.3. We give extensions for speeding up the computations in Section 7.4. In Section 7.5, we evaluate our analysis methods and extension options in an extensive numerical experiment. In Section 7.6, we apply the best variant to a case study. Finally, we draw our main conclusions in Section 7.7.

7.2 Model description and main analysis steps

We present our model with notation in Section 7.2.1, and provide the analysis in Section 7.2.2.

7.2.1 Model description

We consider a non-preemptive $M/G/c$ priority queue with K classes. Customers of class k have priority over those of classes $j \geq k+1$ (i.e., class 1 customers have the highest priority). Class k customers arrive according to a Poisson process with rate λ_k, with $\lambda = \sum_{k=1}^{K} \lambda_k$ denoting the total arrival rate. All customers have the same service time distribution, with $\mathbb{E}[S]$ denoting its mean, cv_S^2 its squared coefficient of variation, $S(t)$ the cumulative distribution, and $\tilde{S}(s)$ the Laplace-Stieltjes transform (LST). The utilization rate per class is denoted by $\rho_k = \frac{\lambda_k \mathbb{E}[S]}{c}$, with $\rho = \sum_{k=1}^{K} \rho_k$. We assume that the queue is stable (i.e., that $\rho < 1$) and that all moments of the service time are finite.

Our aim is to estimate the following performance measures:

- The *delay probability* π_w, i.e., the steady-state probability that all servers are occupied. This probability does not depend on the priority mechanism used.

- The first two moments of the *conditional waiting time* CW_k per class k given that all servers are occupied.

Then, we can fit a reasonable class of distributions on these performance measures to approximate the distribution of the overall waiting time per class. A continuous distribution on which data is commonly – and accurately – fitted is the gamma distribution. In Appendix C, we detail how the parameters for the gamma distribution can be selected to approximate the waiting time distribution per class.

7.2.2 Main analysis steps

For π_w, a fairly accurate approximation is the delay probability in an $M/M/c$ queue, i.e., π_w can be written as Erlang's C formula, see, e.g., Tijms (2003). We now first describe how to obtain the first two moments of CW_1, i.e., the conditional waiting time for class 1. Next, we focus on the conditional waiting time moments of the remaining classes.

To find $\mathbb{E}[CW_1]$ and $\mathbb{E}[CW_1^2]$ we use the following arguments. Given that we consider a non-preemptive service discipline, it does not matter what type of customers are being served when a class 1 customer arrives to find all servers busy. Also, new arrivals from classes 2 up to K have no impact on the waiting time for class 1. Therefore, we obtain $\mathbb{E}[CW_1]$ and $\mathbb{E}[CW_1^2]$ as the first two moments of the conditional waiting time in a single-class $M/G/c$ queue with arrival rate λ_1.

To obtain $\mathbb{E}[CW_k]$ and $\mathbb{E}[CW_k^2]$ for classes $k \geq 2$, we use an argument similar to Williams (1980) and Cohen (1969). We first sketch what happens when a tagged customer of class k arrives to the system when all servers are occupied. Upon arrival, he will see N_1 customers, say, of classes $i \leq k$ that are already waiting to be served. The waiting time of the tagged customer will thus at least consist of the time needed to clear these N_1 customers from the queue, which we denote by T_1. During T_1, new customers of classes $i < k$ may arrive that have priority over the tagged customer. Let N_2 denote the number of higher priority customers that arrive in the time that the first N_1 customers are cleared from the queue. While these N_2 customers are being cleared, new higher priority customers may arrive, and so forth. Overall, the waiting time for the tagged class k customer thus consists of two elements: (i) the time T_1 to clear all N_1 customers of classes $i \leq k$ that were already present in the queue and (ii) the time T_2 to clear those customers of class $i < k$ that arrive while the tagged customer is waiting, starting with the N_2 customers that arrive while the first N_1 are being cleared. Note that T_1 and T_2 are not strictly consecutive, as the higher priority customers that arrive while the tagged customer is waiting may also have priority over some of the N_1 customers that were already present in the system. The values T_1 and T_2 simply denote the workloads associated with clearing the initial N_1 customers and clearing all higher priority customers that arrive after the tagged customer respectively. Obviously, T_2 and T_1 are strongly correlated: If T_1 is large, N_2 will be large and so will T_2.

We compute T_1 as the conditional waiting time in a single-class $M/G/c$ queue with arrival rate $\lambda_k^* = \sum_{i=1}^{k} \lambda_i$. By conditioning on T_1, we can evaluate the distribution of N_2, and then approximate T_2 as the *residual busy period* in a single-class $M/G/c$ queue with arrival rate λ_{k-1}^*. Here we define the residual busy period as the period until all higher priority customers have left the queue, starting with N_2 higher priority customers in the queue, one server just starting with service, and the other $c-1$ servers busy with servicing a customer for some unknown time. We approximate the residual service time of those $c-1$ customers in service by the equilibrium excess of the service time as it is known from renewal theory. Furthermore, we approximate the residual busy period length by the sum of N_2 independent and identically distributed busy periods that each start with an arrival of one customer to the queue. This approximation is exact for the $M/G/1$ and $M/M/c$ queues, see, e.g., Tijms (2003) and Riordan (1962).

Let Z_k be the random variable that denotes the conditional waiting time in an $M/G/c$ queue with arrival rate λ_k^*, with $\tilde{Z}_k(s)$ being the related LST. Similarly, let B_{k-1} and $\tilde{B}_{k-1}(s)$ be the random variable and LST of the busy period of an $M/G/c$ queue with arrival rate λ_{k-1}^*. Note that Z_k corresponds to T_1, while $T_2 = \sum_{i=0}^{N_2} B_{k-1,i}$, where $B_{k-1,i}$ are i.i.d. copies of B_{k-1}. As an approximation, we can now express the conditional waiting time for a class k customer as $CW_k = Z_k + \sum_{i=0}^{N_2} B_{k-1,i}$ with the related LST $\widetilde{CW}_k(s)$ as follows, see, e.g., Williams (1980):

$$\widetilde{CW}_k(s) \approx \tilde{Z}_k\left(s + \lambda_{k-1}^*\left(1 - \tilde{B}_{k-1}(s)\right)\right). \tag{7.1}$$

By taking the first two derivatives at point zero, we find the first two moments of CW_k, $k = 2, \dots, K$:

$$\mathbb{E}[CW_k] = (1 + \lambda_{k-1}^* \mathbb{E}[B_{k-1}])\mathbb{E}[Z_k], \tag{7.2}$$

$$\mathbb{E}\left[CW_k^2\right] = \lambda_{k-1}^* \mathbb{E}\left[B_{k-1}^2\right]\mathbb{E}[Z_k] + (1 + \lambda_{k-1}^* \mathbb{E}[B_{k-1}])^2 \mathbb{E}\left[Z_k^2\right]. \tag{7.3}$$

In the above equations, we indeed see that the length of the residual busy period is influenced by the time needed to clear all $i \leq k$ customers that were initially present in the queue. In expression (7.2), for instance, $\lambda_{k-1}^* \mathbb{E}[Z_k]$ is the expected number of higher priority customers N_2 that arrive while the first N_1 customers are being cleared.

Note that $\mathbb{E}[Z_k]$ and $\mathbb{E}[Z_k^2]$ denote the first two moments of the conditional waiting time in a single-class $M/G/c$ queue with arrival rate λ_k^*, $k = 2, \dots, K$. Similarly, $\mathbb{E}[B_k]$ and $\mathbb{E}[B_k^2]$ denote the first two moments of the busy period in a single-class $M/G/c$ queue with arrival rate λ_k^*,

$k \leq K-1$. Hence, we obtain the first two moments of the conditional waiting time for each customer class – including class 1 – from the analysis of a single-class $M/G/c$ queue.

In Section 7.3, we detail how we can analyze a single-class $M/G/c$ queue, resulting in the first two moments of the conditional waiting time and the first two moments of the busy period.

7.3 Detailed analysis of a single class $M/G/c$ system

We now show how to analyze a single-class $M/G/c$ queue with arrival rate λ as a building block for the multi-class model (note that we do not need class index k in this section). In Section 7.3.1, we show how to compute the first two moments of the conditional waiting time CW. In Section 7.3.2, we describe the computation of the first two moments of the busy period B, i.e., the period in which all servers are occupied.

7.3.1 Computation of $\mathbb{E}[CW]$ and $\mathbb{E}[CW^2]$

We consider two approximate methods to obtain $\mathbb{E}[CW]$ and $\mathbb{E}[CW^2]$, which are both based on Section 9.6.2 in Tijms (2003). The first method, which we denote by AVA1[10], is discussed in Section 7.3.1.1, whereas the second, denoted by AVA2, is discussed in Section 7.3.1.2.

In both AVA1 and AVA2, we obtain performance measures for the $M/G/c$ queue from those for other queues, specifically the $M/M/c$ and $M/D/c$ queues. We denote a performance measure V for the $M/M/c$ queue and the $M/D/c$ queue by $V(exp)$ and $V(det)$ respectively.

7.3.1.1 AVA1

We can find the first two moments of the waiting time (both conditional and unconditional) by using the distributional form of Little's law (see Bertsimas and Nakazato, 1995, Theorem 1), i.e.,

$$\mathbb{E}[CW] = \frac{\mathbb{E}[CL_q]}{\lambda}, \tag{7.4}$$

$$\mathbb{E}[CW^2] = \frac{\mathbb{E}[CL_q(CL_q-1)]}{\lambda^2}. \tag{7.5}$$

In (7.4) and (7.5), CL_q denotes the number of customers waiting in the queue given that all servers are occupied. Note that the distributional form of Little's law does not hold for the sojourn times of the customers in the system, i.e., the sum of the customer's waiting time and service time: in an $M/G/c$ queue, customers may overtake each other during service, ensuring that assumption 2 in Theorem 1 (Bertsimas and Nakazato, 1995) is not necessarily satisfied.

For the $M/G/c$ queue, Tijms (2003) proposes an approximation for the generating function $P_q(z)$ of the *unconditional* number of customers waiting in the queue L_q, see equation (9.6.22) in Tijms (2003). The approximation is based on the following two assumptions: (i) if fewer than c

[10] Where the letters AVA are the initials of the developers of the approximations.

servers are occupied in the $M/G/c$ queue, that queue may be treated as an $M/G/\infty$ queue, and (ii) when all servers are occupied, the $M/G/c$ queue may be treated as an $M/G/1$ queue where the single server works at a rate that is c times as fast as the servers in the original $M/G/c$ system. For both the $M/G/\infty$ and the $M/G/1$ queue, the remaining service time of any busy server is distributed as the equilibrium excess time in a renewal process with the service times as interoccurrence times, see Section 9.6.2 in Tijms (2003).

By taking the first derivative of $P_q(z)$ at $z = 1$, Tijms (2003) finds, without giving the derivation, an expression for $\mathbb{E}[L_q]$ as a linear function of $\mathbb{E}[L_q(exp)]$. Note that it is nontrivial to find this function. Therefore, we describe how this can be done in Appendix A, where we also give the derivation for $\mathbb{E}[CL_q(CL_q - 1)]$ as a function for $\mathbb{E}[CL_q(CL_q - 1)(exp)]$, i.e., equation (7.9). We now use the assumption that π_W is the same in the $M/G/c$ and $M/M/c$ queue and Little's Law to find that $\frac{\mathbb{E}[L_q]}{\mathbb{E}[L_q(exp)]} = \frac{\mathbb{E}[CL_q]}{\mathbb{E}[CL_q(exp)]} = \frac{\mathbb{E}[CW]}{\mathbb{E}[CW(exp)]}$. We thus obtain the following linear relation between $\mathbb{E}[CW]$ and $\mathbb{E}[CW(exp)]$:

$$\frac{\mathbb{E}[CW]}{\mathbb{E}[CW(exp)]} = (1-\rho)\gamma_1 \frac{c}{\mathbb{E}[S]} + \frac{\rho}{2}(1 + cv_S^2), \tag{7.6}$$

where γ_1 is given by:

$$\gamma_1 = \int_0^\infty \left(1 - S_e(t)\right)^c dt, \tag{7.7}$$

with $S_e(t)$ denoting the equilibrium excess distribution function of the service time, i.e.,

$$S_e(t) = \frac{1}{\mathbb{E}[S]} \int_0^t \left(1 - S(u)\right) du. \tag{7.8}$$

Note that γ_1 can be interpreted as the expectation of $\min(S_e^1, \ldots, S_e^c)$, where S_e^i, $i = 1, \ldots, c$, are i.i.d random variables with common probability distribution $S_e(t)$.

Similarly, we find a linear relation between $\mathbb{E}[CL_q(CL_q - 1)]$ and $\mathbb{E}[CL_q(CL_q - 1)(exp)]$, and hence between $\mathbb{E}[CW^2]$ and $\mathbb{E}[CW(exp)^2]$:

$$\frac{\mathbb{E}[CW^2]}{\mathbb{E}[CW(exp)^2\,]} = \frac{\lambda^2(1-\rho)^2}{\rho^2}\gamma_2 + \frac{\lambda(1-\rho)}{2}(cv_S^2 + 1)\gamma_1 + \frac{\rho^2}{4}(cv_S^2 + 1)^2 + \frac{\rho(1-\rho)}{6}\frac{\mathbb{E}[S^3]}{\mathbb{E}[S]^3}, \tag{7.9}$$

where γ_2 is given by:

$$\gamma_2 = \int_0^\infty t\left(1 - S_e(t)\right)^c dt. \tag{7.10}$$

Similar to γ_1, $2\gamma_2$can be interpreted as the second moment of $\min(S_e^1, \dots, S_e^c)$. This can easily be verified via partial integration of the right-hand side of (7.10), see, e.g., Tijms (2003), Section 5.1.2.

For $\mathbb{E}[CW(exp)]$ and $\mathbb{E}[CW(exp)^2]$, expressions can easily be found, see, e.g., Section 5.1.2 in Tijms (2003):

$$\mathbb{E}[CW(exp)] = \frac{\mathbb{E}[S]}{c(1-\rho)}, \tag{7.11}$$

$$\mathbb{E}[CW(exp)^2] = \frac{2\mathbb{E}[S]^2}{c^2(1-\rho)^2}. \tag{7.12}$$

7.3.1.2 <u>AVA2</u>

In this method, we estimate both $\mathbb{E}[CW]$ and $\mathbb{E}[CW^2]$ as a weighted average of the waiting time moments in an $M/D/c$ and an $M/M/c$ queue, with the mean service time in the latter queues being equal to $\mathbb{E}[S]$. We use the squared coefficient of variation of the service time cv_S^2 as weight when computing $\mathbb{E}[CW]$ and α, defined by (7.15) below, as weight when computing $\mathbb{E}[CW^2]$. We find:

$$\mathbb{E}[CW] = (1 - cv_S^2)\mathbb{E}[CW(det)] + cv_S^2\mathbb{E}[CW(exp)], \tag{7.13}$$

$$\mathbb{E}[CW^2] = (1-\alpha)\mathbb{E}[CW(det)^2] + \alpha\mathbb{E}[CW(exp)^2]. \tag{7.14}$$

We derive expression (7.13) from expression (9.6.24) in Tijms (2003). In contrast, we develop expression (7.14) ourselves, where we determine the expression for α, given in (7.15), such that it is exact for $c = 1$. Given that $c = 1$, we obtain analytical expressions for $\mathbb{E}[CW]$ and $\mathbb{E}[CW^2]$ under any service time distribution by using the Pollaczek-Khintchine formula. Note that the expression for α is exact for both the $M/M/c$ and the $M/D/c$ queue, with $\alpha = 1$ for exponential service times and $\alpha = 0$ for deterministic service times.

$$\alpha = \frac{1}{10-\rho}\left(2(1-\rho)\frac{\mathbb{E}[S^3]}{\mathbb{E}[S]^3} + \frac{3\rho\mathbb{E}[S^2]^2}{\mathbb{E}[S]^4} - 2 - \rho\right). \tag{7.15}$$

The expressions for $\mathbb{E}[CW(exp)]$ and $\mathbb{E}[CW(exp)^2]$ are given by the equations (7.11) and (7.12) respectively. We find expressions for $\mathbb{E}[CW(det)]$ and $\mathbb{E}[CW(det)^2]$ from the LST of the unconditional waiting time in an $M/D/c$ queue, see, e.g., Riordan (1962):

$$\mathbb{E}\left[e^{-s\mathbb{E}[S]^{-1}W}\right] = \frac{(1-\pi_w)s}{(c\rho)^c e^{-s} - (c\rho - s)^c}\prod_{i=1}^{c-1}(u_i - s), \tag{7.16}$$

where $u_i = c\rho(1 - z_i)$, and z_i, $i = 0, \dots, c-1$, are the c roots of $z^c = e^{c\rho(z-1)}$, with $|z_i| \le 1$ and $z_0 = 1$. Note that (7.16) does not use this latter root. The roots z_i ($i \ge 1$) can easily be computed recursively: starting with $z_i^{(0)} = 0$, $z_i^{(n+1)}$ can be computed as a function of $z_i^{(n)}$ until convergence occurs (see equation (14) in Janssen and Van Leeuwaarden, 2008). Moreover, the roots z_i are known in closed-form as an infinite sum (Janssen and Van Leeuwaarden, 2008). In Janssen and Van Leeuwaarden, we also find an expression for the waiting probability π_w in the $M/D/c$ queue, which we denote by $\pi_w(det)$:

$$\pi_w(det) = 1 - \frac{c(1-\rho)}{\prod_{i=1}^{c-1}(1-z_i)}.$$

By multiplying both sides of (7.16) by $(c\rho)^c e^{-s} - (c\rho - s)^c$ and taking the second and third order derivatives of the resulting expression, we find that:

$$\mathbb{E}[CW(det)] = \frac{1}{\lambda\pi_w}\left(\frac{c\rho^2 - c + 1}{2(1-\rho)} + \sum_{i=1}^{c-1}\frac{1}{1-z_i}\right), \tag{7.17}$$

$$\begin{aligned}\mathbb{E}[CW(det)^2] = {} & \frac{c^2\rho^3 - (c-1)(c-2) + 3\lambda(c\rho^2 - c + 1)\pi_w\mathbb{E}[CW(det)]}{3\lambda^2\pi_w(1-\rho)} \\ & + \frac{2}{\lambda^2\pi_w}\sum_{i=1}^{c-2}\frac{1}{1-z_i}\sum_{l=i+1}^{c-1}\frac{1}{1-z_l}.\end{aligned} \tag{7.18}$$

7.3.2 Computation of $\mathbb{E}[\boldsymbol{B}]$ and $\mathbb{E}[\boldsymbol{B}^2]$

We now show how to compute the first two moments of the busy period. Both in this section, and in the computational experiments, we restrict ourselves to $M/Ph_m/c$ queues, i.e., queues where the service time has a phase type distribution with m phases. A phase type distribution characterizes the time until absorption in an absorbing Markov chain with a finite state space given that the chain starts in an initial transient (non-absorbing) state. Such a distribution is characterized by the tuple $(\boldsymbol{\beta}, \boldsymbol{V}, V^0)$, where the $\boldsymbol{\beta}$ is a row vector of size m indicating the initial state probability vector, i.e., element j in $\boldsymbol{\beta}$ denotes the probability of starting in state $j = 1, \dots, m$, $\boldsymbol{V}$ is an m-by-m matrix denoting the transition rates among transient states, and V^0 is a column vector of size m denoting the transition from the transient to the absorbing state. The two-phased Coxian-2 distribution, for instance, can be characterized as follows:

$$(\boldsymbol{\beta}, \boldsymbol{R}, R^0) = \left((1 \quad 0), \begin{pmatrix} -\mu_1 & p\mu_1 \\ 0 & -\mu_2 \end{pmatrix}, \begin{pmatrix} (1-p)\mu_1 \\ \mu_2 \end{pmatrix}\right). \tag{7.19}$$

The class of phase type distributions is rich in the sense that it allows us to cover a broad range of coefficients of variation for the service time distribution. In particular, the mixed generalized Erlang distribution, i.e., a distribution that is a generalized Erlang-n distribution with probability q_n, $n = 1, .., m$, allows us to model both variables with any value for cv_S^2. A special case of the mixed generalized Erlang distribution is the Coxian distribution, where the Coxian-2 distribution, for instance, can model a distribution with $cv_S^2 \geq 0.5$, see, e.g., Marie (1980).

The busy period can be seen as the first passage time of the queue-length process from the moment there are c customers in the system to that when there are $c - 1$ customers in the system. Let $\boldsymbol{Q}$ denote the generator matrix of the queue length process. For an $M/Ph_m/c$ queue, $\boldsymbol{Q}$ can be characterized as follows. An element (i,j) in $\boldsymbol{Q}$ denotes the transitions from level i (with a *level* being the set of states with a queue length size i) to level j.

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{A}_1^0 & \boldsymbol{A}_0^0 & 0 & 0 & 0 & \cdots & & & & \\ \boldsymbol{A}_2^1 & \boldsymbol{A}_1^1 & \boldsymbol{A}_0^1 & 0 & 0 & \cdots & & & & \\ & \ddots & \ddots & \ddots & 0 & \cdots & & & & \\ & 0 & \boldsymbol{A}_2^{c-1} & \boldsymbol{A}_1^{c-1} & \boldsymbol{A}_0^{c-1} & 0 & \cdots & & & \\ & \cdots & 0 & \boldsymbol{A}_2^{c} & \boldsymbol{A}_1 & \boldsymbol{A}_0 & 0 & \cdots & & \\ & & \cdots & 0 & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & 0 & \cdots & \\ & & & \cdots & 0 & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & 0 & \\ & & & & \cdots & 0 & \ddots & \ddots & \ddots & \end{pmatrix}. \tag{7.20}$$

In $\boldsymbol{Q}$, $\boldsymbol{A}_0 = \lambda \boldsymbol{I}$, $\boldsymbol{A}_1 = -\lambda\boldsymbol{I} + \oplus_{i=1}^{c} \boldsymbol{V}$, and $\boldsymbol{A}_2 = \oplus_{i=1}^{c} V^0 \boldsymbol{\beta}$, with $\boldsymbol{I}$ being the identity matrix of size m^c and $\oplus_{i=1}^{c} \boldsymbol{V} = \boldsymbol{V} \oplus \ldots \oplus \boldsymbol{V}$, see, e.g., Neuts (1981). Note that $\boldsymbol{Q}$ is a Quasi-Birth Death process that is homogeneous for levels strictly larger than c. This property also holds for the $M/Ph_m/1$ queue. Therefore, the busy period results of $M/Ph_m/1$ also hold for $M/Ph_m/c$ by setting $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$ as defined before. Neuts (1981, Section 3.3) studies the busy period of phase type single server queues using a matrix analytical approach. We shall now apply Neuts' approach to derive the first two moments of the busy period in an $M/Ph_m/c$ queue. Let $\boldsymbol{G}$ denote an m^c-by-m^c matrix where entry (j, j') denotes the conditional probability that the queue length process, starting in level $i+1$ $(i \geq c)$ at state j at time zero, reaches level i for the first time in state j'. Note that the entries in $\boldsymbol{G}$ are independent of i due to the homogeneous property of $\boldsymbol{Q}$ for levels greater than c. The matrix $\boldsymbol{G}$ is the minimal solution of the following quadratic matrix equation:

$$\boldsymbol{G} = \boldsymbol{C}_0 + \boldsymbol{C}_2\boldsymbol{G}^2, \tag{7.21}$$

where $\boldsymbol{C}_0 = -(\boldsymbol{A}_1)^{-1}\boldsymbol{A}_2$ and $\boldsymbol{C_2} = -(\boldsymbol{A}_1)^{-1}\boldsymbol{A_0}$. Note that $\boldsymbol{C}_0$ is the transition probability matrix that the queue-length process jumps from level $i+1$ to i, $i \geq c$, and $\boldsymbol{C}_2$ the transition probability matrix that the queue length process jumps from level i to $i+1$, $i \geq c$. The matrix $\boldsymbol{G}$ is stochastic, i.e., $\boldsymbol{G}e = e$. Moreover, it is the unique solution of (7.21) if the queue is stable (Neuts, 1981, Th. 3.3.2). In the remainder of this section, we assume that the queue is stable, i.e., that $\rho < 1$. Therefore, $\boldsymbol{G}$ can be computed recursively. Let $\boldsymbol{G}_n$ denote the estimate of $\boldsymbol{G}$ after iteration n. We then find:

$$\boldsymbol{G}_{n+1} = \boldsymbol{C}_0 + \boldsymbol{C}_2(\boldsymbol{G}_n)^2, \quad n \geq 1,$$

where $\boldsymbol{G}_1 = \boldsymbol{C}_0$. The above equation is proven to converge, see Th. 3.3.1 in Neuts (1981).

From $\boldsymbol{G}$, we are able to derive the first two moments of the busy period B. Let bp_1 denote a column vector of size m^c with the j-th entry being equal to the mean conditional busy period given that the busy period starts in level c in state j. Similar to the way in which Neuts derives the busy period moments from $\boldsymbol{G}$, we find the following expression for bp_1 from Eq. (3.3.23) and (3.3.36) in Neuts (1981).

$$bp_1 = -(\boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_0\boldsymbol{G})^{-1}e. \tag{7.22}$$

Note that the matrix $\boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_0\boldsymbol{G}$ is nonsingular since it can be written as a product of two nonsingular matrices, see Neuts (1981), Th. 3.3.3.

Similar to bp_1, let bp_2 also be a column vector of size m^c with the j-th entry equal to the second moment of the conditional busy period that starts in level c in state j. We derive bp_2 from eq. (3.3.26) in Neuts (1981) by using the fact that $\boldsymbol{G}e = e$:

$$bp_2 = -2(\boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_0\boldsymbol{G})^{-1}(\boldsymbol{A}_0\boldsymbol{M}_1 + \boldsymbol{I})bp_1, \tag{7.23}$$

where the matrix $\boldsymbol{M}_1$ is the minimal, unique and nonnegative solution of the following equation:

$$\boldsymbol{M}_1 = -(\boldsymbol{A}_1)^{-1}\boldsymbol{G} + \boldsymbol{C}_2(\boldsymbol{G}\boldsymbol{M}_1 + \boldsymbol{M}_1\boldsymbol{G}).$$

This matrix equation can be solved recursively by starting with an initial solution that is equal to the zero matrix and using an iteration procedure similar to that for computing matrix $\boldsymbol{G}$.

We now obtain the first two moments of the busy period by multiplying bp_1 and bp_2 by the joint distribution of the remaining service times on the servers when a busy period starts. At the start of a busy period, there is exactly one server that just started service. For the other $c-1$ servers, we use the common approximation, see, e.g., Tijms (2003), that the remaining service time on each server has a distribution equal to that of the remaining service time in equilibrium, where the service times are assumed to be independent among all servers. Given that the service times are phase-type distributed, we find the equilibrium distribution of the remaining service time on any server by considering the probability of being in each phase, since the time spent in any phase is exponentially distributed. Overall, the initial distribution of the joint phases of customers in service at the start of a busy period equals $\boldsymbol{\beta}\otimes(\otimes_{i=1}^{c-1} z^*)$, with z^* equal to the following expression, see, e.g., Lemma 1 in Al Hanbali et al. (2012):

$$z^* = -\frac{1}{\mathbb{E}[S]}\boldsymbol{\beta} \cdot \boldsymbol{V}^{-1}.$$

7.4 Extensions to speed up the analysis methods

As we show in Section 7.5.2.1, it can be time-consuming to estimate the two moments of the busy period, particularly in problem instances with many servers and service time distributions with low values for cv_S^2 (corresponding to distributions with many phases). Therefore, we present three options for reducing the overall computation time. We describe the options in Sections 7.4.1 through 7.4.3.

7.4.1 Option 1: Scaling the service time distribution

We scale the service time distribution based on the number of servers when estimating the first two moments of the busy period. Specifically, we replace the $M/Ph_m/c$ queue by a $M/Ph_m/3$ queue where the service rate in each phase is $\frac{c}{3}$ times as fast as in the original system. As an example, for a 6-server queue where the service time has a Coxian-2 distribution, we now find:

$$(\boldsymbol{\beta}, \boldsymbol{S}, S^0) = \left((1 \quad 0), \begin{pmatrix} -2\mu_1 & 2p\mu_1 \\ 0 & -2\mu_2 \end{pmatrix}, \begin{pmatrix} 2(1-p)\mu_1 \\ 2\mu_2 \end{pmatrix}\right) \tag{7.24}$$

By limiting the number of servers to 3, we obtain small matrices when computing $\mathbb{E}[B_k]$ and $\mathbb{E}[B_k^2]$. As a result, the computation times for 3-server instances remain below 1 second for service time distributions with up to 4 phases. We refer the reader to Section 7.5.2.1 for details. In contrast, when there are at least 6 servers, the computation times quickly explode. Therefore, we scale the service time distribution for instances with more than 3 servers.

7.4.2 Option 2: Estimating $\mathbb{E}[\boldsymbol{CW_k}]$ and $\mathbb{E}\left[\boldsymbol{CW_k^2}\right]$ for class $\boldsymbol{k}$ $(1 < k < K)$ through interpolation from those of class 1 and class $\boldsymbol{K}$ customers

Our second option is to estimate the waiting time moments for class k customers, $1 < k < K$, from those of class 1 and class K customers. Then, we do not require values for $\mathbb{E}[B_{k-1}]$ and $\mathbb{E}[B_{k-1}^2]$ to compute $\mathbb{E}[CW_k]$ and $\mathbb{E}[CW_k^2]$. In fact, we only need to compute $\mathbb{E}[B_{K-1}]$ and $\mathbb{E}[B_{K-1}^2]$ to estimate the waiting time moments for the lowest priority class K. Note that this approximation can only be used for problem instances with at least 3 classes, as we require the waiting time moments of at least 2 classes, i.e., those for class 1 and class K, to estimate the moments for the remaining classes.

We obtain $\mathbb{E}[CW_k]$ and $\mathbb{E}[CW_k^2]$, $1 < k < K,$ from the moments of classes 1 and K as follows:

$$\mathbb{E}[CW_k] = r_{1k}\mathbb{E}[CW_1] + (1 - r_{1k})\mathbb{E}[CW_K], \tag{7.25}$$

$$\mathbb{E}\left[CW_k^2\right] = r_{\ 2k}\mathbb{E}[CW_1^2] + (1 - r_{2k})\mathbb{E}[CW_K^2]. \tag{7.26}$$

We obtain the interpolation factors r_{jk} ($j = 1,2$, $k \in \{2, \dots, K-1\}$) by solving equations (7.25) and (7.26) for the $M/M/c$ queue. Let $\sigma_k = \sum_{i=1}^{k} \rho_i$ be shorthand notation for the utilization rate for classes 1 up to k. Using the formulas for the waiting time moments per class in Kella and Yechiali (1985), we find the following expressions for r_{1k} and r_{2k}:

$$r_{1k} = \frac{(1-\sigma_1)}{(1-\sigma_k)(1-\sigma_{k-1})} \frac{(1-\sigma_K)(1-\sigma_{K-1}) - (1-\sigma_k)(1-\sigma_{k-1})}{(1-\sigma_K)(1-\sigma_{K-1}) - 1 + \sigma_1}. \tag{7.27}$$

$$r_{2k} = \frac{(1-\sigma_1)^2}{(1-\sigma_k)^2(1-\sigma_{k-1})^3} * \frac{(1-\sigma_k\sigma_{k-1})(1-\sigma_K)^2(1-\sigma_{K-1})^3 - (1-\sigma_k)^2(1-\sigma_{k-1})^3(1-\sigma_K\sigma_{K-1})}{(1-\sigma_K)^2(1-\sigma_{K-1})^3 - (1-\sigma_1)^2(1-\sigma_K\sigma_{K-1})}. \tag{7.28}$$

Note that r_{12} and r_{22} only depend on the values of $\rho_k (k = 1,..,K)$.

7.4.3 Option 3: Extrapolation for service time distributions with low variability

When the service time variability is low (i.e., $cv_S^2 \leq 0.2$), the approach of Section 7.2 may result in large computation times. Then, we must fit a phase type distribution with many phases to characterize the service time, e.g., an Erlang-10 distribution when cv_S^2 is 0.1. To gain efficiency, we may use *extrapolation*, i.e., we estimate the conditional waiting time moments for a distribution with a low cv_S^2 from those of distributions with larger values for cv_S^2.

We use a least squares approach to fit a function on a set of support points, with a support point denoting the known waiting time moment value for a given cv_S^2 (and thus serving as input for extrapolation). Given that the conditional waiting time moments increase monotonically in cv_S^2, it is reasonable to fit a monotonically increasing function, such as a linear or exponential function, on the support points.

7.5 Computational experiment and results

We performed an experiment aimed at the validation of our methods. Section 7.5.1 contains our experiment objectives and design. We validate our analysis methods and extension options in Sections 7.5.2 and 7.5.3 respectively.

7.5.1 Experimental design

We use discrete-event simulation as a benchmark for validation. We use a replication-deletion approach with a warm-up period of 1 million arrivals and multiple runs of 1 million arrivals each. The threshold of 1 million ensures that we have observed an extensive number of events. After each run, we compute the relevant performance measures over all arrivals after the warm-up period (and not only the arrivals during the most recent run). Let $\mathbb{E}[X(j)]$ denote the value of a performance measure after the j-th run. The simulation stops once convergence occurs, i.e., $\frac{\mathbb{E}[X(j)] - \mathbb{E}[X(j-1)]}{\mathbb{E}[X(j-1)]} < 0.05\%$ for all performance measures. For the two-class instances, the minimal number of runs per instance was 23, with the average being 51. Both the simulations and the analysis using our methods have been performed on a Dell optiplex 760 computer with Intel quad core, 2.83 GHz processor, with our methods implemented in Maple 14.

Our test bed consists of 648 problem instances, 324 with two customer classes and 324 with three classes. Table 7.1 shows the parameter values considered. The asterisks in the table pertain to the subset of instances on which extension option 3 (i.e., extrapolation) was tested (see Section 7.5.3.2). To obtain the class arrival rates λ_k, we compute the total arrival rate λ as $\rho c/\mathbb{E}[S]$ and disaggregate λ over the classes using the ratios λ_k/λ. For the squared coefficient of variation $cv_S^2 \leq 0.5$, we fit an Erlang- n distribution to $\mathbb{E}[S]$ and cv_S^2. For $cv_S^2 = 0.75$, we use a Coxian-2 distribution with $\mu_1 = \frac{2}{\mathbb{E}[S]}$, $p = \frac{0.5}{cv_S^2}$, and $\mu_2 = \mu_1 p$, see Marie (1980).

	Parameter	**Values for theoretical problem instances**
1	c	3*, 6, 9*
2	ρ	0.8*, 0.9, 0.95*
3	$\mathbb{E}[S]$ (hours)	1.25* 2.5, 5, 10*
4	cv_S^2	0.25, 0.5, 0.75
5	Division two-class instances $\left(\frac{\lambda_1}{\lambda};\frac{\lambda_2}{\lambda}\right)$	(0.1; 0.9)*, (0.3; 0.7), (0.5; 0.5)*
6	Division three-class instances $\left(\frac{\lambda_1}{\lambda};\frac{\lambda_2}{\lambda};\frac{\lambda_3}{\lambda}\right)$	(0.1; 0.2; 0.7), (0.2; 0.3; 0.5) , $\left(\frac{1}{3};\frac{1}{3};\frac{1}{3}\right)$

Table 7.1 Parameter values considered for theoretical problem instances.

7.5.2 Method validation

We first show in Section 7.5.2.1 that we obtain good results when a scaled service time distribution is used for finding the first two moments of the busy period (i.e., extension option 3). Then, we validate AVA1 and AVA2 with scaling on problem instances with 2 and 3 customer classes in Section 7.5.2.2.

7.5.2.1 The impact of scaling the service distribution

We show the performance of AVA1 (see Section 7.3.1) both with and without scaling (the findings are similar for AVA2), where we only consider the cases with 2 classes and 6 servers. We omit the 9-server instances, because we are unable to estimate the busy period moments without scaling when $cv_S^2 = 0.25$, as the required matrices become too large to evaluate then.

Table 7.2 shows the average and maximum relative error to simulation (rows 'Avg. RE' and 'Max. RE' respectively) for the first two moments of B_1 (the busy period when there are only class 1 arrivals) and CW_2. We conclude that the mean busy period $\mathbb{E}[B_1]$ remains accurate under scaling. Also, although $\mathbb{E}[B_1^2]$ is less accurate under scaling, the greater inaccuracy has no impact on the second moment of the class-2 waiting time. Indeed, the relative error for $\mathbb{E}[CW_2]$ is comparable under scaling and non-scaling, whereas the errors for $\mathbb{E}[CW_2^2]$ are smallest under scaling. A closer observation of the results shows that $\mathbb{E}[B_1]$ and $\mathbb{E}[B_1^2]$ are generally underestimated (for each waiting time moment, 80% of all values are underestimated), whereas the first two moments of CW_2 are still generally overestimated (60% on average). Clearly, the underestimation of $\mathbb{E}[B_1]$ and $\mathbb{E}[B_1^2]$ is compensated to some extent by the approximation we use to compute $\mathbb{E}[CW_2]$ and $\mathbb{E}[CW_2^2]$. The estimates for $\mathbb{E}[CW_2]$ remain accurate for a larger number of servers, as shown in Table 7.7 for three-class instances with 9 servers.

	$\mathbb{E}[B_1]$		$\mathbb{E}[B_1^2]$		$\mathbb{E}[CW_2]$		$\mathbb{E}[CW_2^2]$	
	Scaled	**Unscaled**	**Scaled**	**Unscaled**	**Scaled**	**Unscaled**	**Scaled**	**Unscaled**
Avg. RE	0.2%	0.3%	5.0%	0.5%	0.8%	0.9%	1.5%	1.7%
Max. RE	0.6%	1.3%	10.5%	2.1%	3.1%	3.0%	5.8%	6.5%

Table 7.2 Solution quality with and without scaling for method AVA1.

Scaling is also very fast: the time to compute the busy period moments is at most 0.9 seconds. In contrast, the non-scaled variant has an average computation time of 17 minutes for cases with 6 servers and a cv_S^2 of 0.25. For the 9-server instances with $cv_S^2 = 0.25$, the resulting matrices are so large that we obtain memory errors. As a result, we even cannot compute the busy period moments without scaling. We therefore use scaling from now on.

7.5.2.2 Validation of AVA1 and AVA2

We evaluate the accuracy of AVA1 and AVA2 by comparison to Williams' method (1980) and to simulation. Table 7.3 and Table 7.4 show the overall relative error to simulation for the mean conditional waiting time per class and the second moment of the conditional waiting time respectively. In both tables, 'Will' denotes the results using Williams' method.

		$\mathbb{E}[CW_1]$			$\mathbb{E}[CW_2]$			$\mathbb{E}[CW_3]$		
		AVA1	**AVA2**	**Will**	**AVA1**	**AVA2**	**Will**	**AVA1**	**AVA2**	**Will**
2-class setting	**Avg. RE**	0.8%	1.4%	13.1%	0.8%	0.6%	1.4%	-	-	-
	Max. RE	3.5%	5.1%	29.2%	3.3%	3.8%	6.9%	-	-	-
3-class setting	**Avg. RE**	0.6%	1.6%	14.2%	1.1%	1.2%	9.3%	1.0%	1.0%	1.0%
	Max. RE	2.9%	5.0%	29.4%	4.2%	4.8%	25.1%	5.1%	5.6%	5.6%

Table 7.3 Relative error per method for the mean conditional waiting time per class.

		$\mathbb{E}[CW_1^2]$			$\mathbb{E}[CW_2^2]$			$\mathbb{E}[CW_3^2]$		
		AVA1	**AVA2**	**Will**	**AVA1**	**AVA2**	**Will**	**AVA1**	**AVA2**	**Will**
2-class setting	**Avg. RE**	2.0%	2.8%	24.8%	1.5%	1.5%	2.0%	-	-	-
	Max. RE	8.5%	9.4%	55.0%	7.9%	8.3%	9.3%	-	-	-
3-class setting	**Avg. RE**	1.8%	2.6%	27.4%	2.5%	2.3%	15.5%	2.2%	2.5%	1.4%
	Max. RE	7.6%	10.1%	55.6%	10.0%	10.0%	45.4%	12.3%	13.0%	7.6%

Table 7.4 Relative error per method for the second moment of the conditional waiting time per class.

In general, AVA1 and AVA2 both clearly outperform Williams' method. The latter method gives particularly poor results for class 1 customers. For this class, Williams' method always severely underestimates the first two moments of the waiting time, leading to an overestimation of the service level (and hence risking that in practice insufficient servers will be deployed, such that SLAs are not met). In contrast, AVA1 and AVA2 overestimate 40% and 50% of class 1 waiting time moments on average in the two-class instances (the fractions are even higher in the three-class instances). Still, William's method works very well for the lowest priority class. In fact, that method is very accurate for the class 3 waiting time moments, even giving the most accurate values for $\mathbb{E}[CW_3^2]$. For all methods, accuracy is largest when estimating the mean waiting times compared to the second moments. Based on a further investigation of the results, we conclude:

- **AVA1 gives the most accurate results, especially on the class 1 waiting time moments.** For the remaining classes, AVA1 gives comparable or better results than AVA2 and performs significantly better than Williams' method, except for the lowest priority class (see below). The accuracy of AVA1 is influenced most by the squared coefficient of variation, an issue that we discuss in more detail later on. AVA1 is also most accurate when the low priority customers are a large fraction of the total demand rate, particularly for the second moment of the class-3 waiting time (with the average relative error decreasing from 3.6% to 1% as λ_3/λ increases from $\frac{1}{3}$ to 0.7). The value of c also influences the accuracy of AVA1, but does so in different ways for each class. In particular, the relative error increases with c for the class 1 waiting times, while those for class 2 decrease in the three-class instances. In the worst case, the average relative error per performance measure is around 4%.
- **For the lowest priority class, Williams' method works very well under high loads, large fractions of class 1 customers and few servers.** Then, the accuracy of Williams' method is comparable to – and often better than – that of AVA1 and AVA2 for the conditional waiting time moments of class K. In the two-class instances, for example, the average relative error on $\mathbb{E}[CW_2^2]$ then equals 1.8% as opposed to the 3% error found with AVA1 and AVA2.
- **In general, the accuracy of AVA2 increases as c decreases.** For the lower priority classes, the relative errors are then equal to, or smaller than, those with AVA1.
- **AVA2 outperforms the other methods on class K when ρ is low**. On the mean waiting time $\mathbb{E}[CW_K]$ $(K = 2{,}3)$, for instance, the relative error with AVA2 is 0.5%. The second best method is AVA1 with a relative error of 1%.

We also find that all methods become much more accurate as cv_S^2 increases to 1. For AVA1, for instance, the average relative error decreases from 4.6% to 1% in the most striking case. Surprisingly, AVA2 does not outperform AVA1 for class 1 even when cv_S^2 is low. This finding does not change if we incorporate the fact that π_w has a slightly different value for the $M/D/c$ queue than for the $M/M/c$ queue (which leads to slightly different weights when determining the conditional waiting times for the $M/G/c$ queue from those of the former two queues).

Table 7.5 shows the computation times for the two-class instances, which include the times needed for computing the busy period moments. The computation time is a fraction of a second on average, and at most a few seconds. Williams' method even has negligible computation time, since the waiting time moments are found using analytical expressions. Therefore, this method may be beneficial for estimating the conditional waiting time moments of class K.

	AVA1	AVA2	Williams
Average time (sec)	0.21	0.14	0.00
Maximum time (sec)	3.66	2.40	0.00

Table 7.5 Computation times per method for the two-class instances.

A final interesting finding from our analysis is that the squared coefficient of variation cv_{CW}^2 of the conditional waiting time over all classes increases to 1 as the utilization rate ρ increases. Similarly, the squared coefficient of variation $cv_{CW_K}^2$ of the conditional waiting time for the lowest priority class also tends to move to 1 with the increase of ρ. For the remaining classes, the squared coefficient of variation of the conditional waiting time remains constant in ρ. These findings apply to both AVA1 and AVA2. Further details can be found in Appendix B.

7.5.3 Performance of extension options 2 and 3

7.5.3.1 Performance of extension option 2: interpolation over customer classes

Table 7.6 shows the relative error of AVA1 and AVA2 in estimating $\mathbb{E}[CW_2]$ and $\mathbb{E}[CW_2^2]$, both under the original variant (i.e., using equations (7.2) and (7.3) of Section 7.2.2, denoted by 'Orig' in the table) and under interpolation (i.e., Section 7.4.2, denoted by 'IntPol' in the table). For the mean conditional waiting time $\mathbb{E}[CW_2]$, the solution quality of both variants is similar. For the conditional second moment $\mathbb{E}[CW_2^2]$, the results are clearly worse under interpolation.

	AVA1				**AVA2**			
	$\mathbb{E}[CW_2]$		$\mathbb{E}[CW_2^2]$		$\mathbb{E}[CW_2]$		$\mathbb{E}[CW_2^2]$	
	Orig	**IntPol**	**Orig**	**IntPol**	**Orig**	**IntPol**	**Orig**	**IntPol**
Avg. RE	1.1%	1.3%	2.5%	4.7%	1.2%	1.1%	2.3%	4.5%
Max. RE	4.2%	5.7%	10.0%	15.4%	4.8%	4.6%	10.0%	14.6%

Table 7.6 Comparison of original analysis method to the interpolation variant for class 2 waiting time moments.

Still, Table 7.7 shows that interpolation, combined with AVA1, may be effective if the number of servers is small. Interpolation also works well for a low utilization rate, both in combination with AVA1 and with AVA2.

		Avg relative error AVA1				**Avg relative error AVA2**			
		$\mathbb{E}[CW_2]$		$\mathbb{E}[CW_2^2]$		$\mathbb{E}[CW_2]$		$\mathbb{E}[CW_2^2]$	
		Orig	**IntPol**	**Orig**	**IntPol**	**Orig**	**IntPol**	**Orig**	**IntPol**
c	**3**	1.7%	0.7%	4.0%	1.6%	0.7%	1.0%	1.3%	2.4%
	6	1.0%	1.1%	2.3%	4.6%	1.2%	1.0%	2.5%	4.4%
	9	0.6%	2.2%	1.3%	7.8%	1.6%	1.3%	3.1%	6.6%
ρ	**0.8**	1.4%	0.7%	3.1%	3.1%	1.8%	0.8%	3.2%	2.9%
	0.9	1.0%	1.3%	2.3%	5.0%	1.0%	1.1%	2.1%	4.8%
	0.95	0.9%	1.9%	2.2%	6.0%	0.7%	1.4%	1.6%	5.6%

Table 7.7 Comparison of original analysis method to the interpolation variant for specific parameter values.

7.5.3.2 *Performance of extension option 3: using extrapolation when service variability is low*

We use extrapolation to analyze distributions with $cv_S^2 \in \{0, 0.1, 0.2\}$, as computation times explode when the phase-type service time distributions have more than, say, 5 phases. To this end, we use at most four distributions to construct support points, i.e., those with $cv_S^2 \in \{0.25, 1/3, 0.5, 1\}$. We consider all combinations of at least 2 support points. Overall, we thus have $\sum_{i=2}^{4} \binom{4}{i} = 11$ strategies, where a strategy denotes the set of support points considered.

We test each strategy on 16 two-class problem instances. The tested parameter values have been marked by an asterisk in Table 7.1. We obtain our support points by using AVA1. For each combination of strategy and problem instance, we fit both a linear and an exponential function on the data points. A first look at the waiting time values shows that both the first and second moment of the conditional waiting time seem to be linear in cv_S^2, see Figure 7.1 for the first two moments of CW_2 in one problem instance (the results are similar for other instances). For completeness, we also fit an exponential function on the data, as this function is also monotonically increasing.

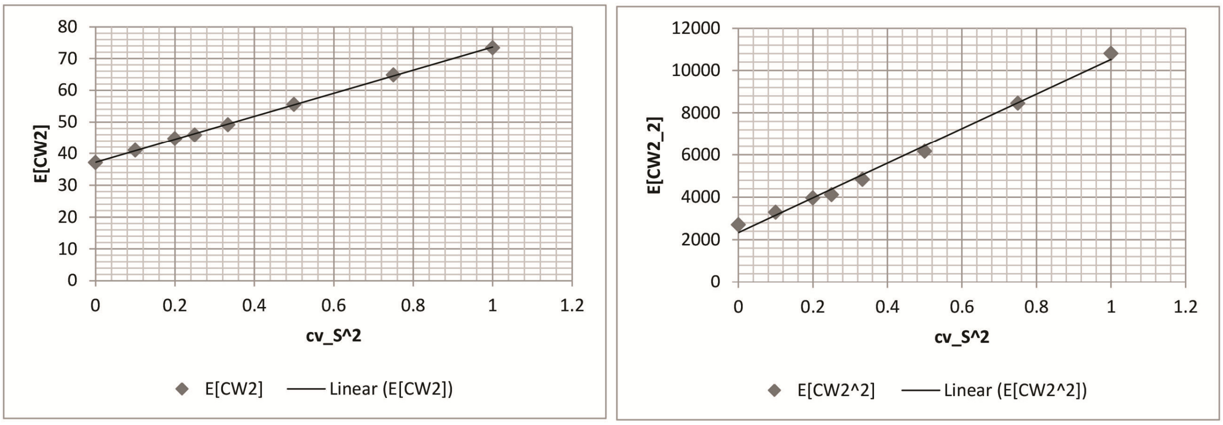


Figure 7.1 The first two moments of the conditional waiting time for class 2 (i.e., CW_2) as functions of cv_S^2 for a single problem instance.

Overall, accuracy is largest when we use support points with low squared coefficients of variation, particularly when estimating the second moment of the conditional waiting time per class. Table 7.8 shows the average relative error on $\mathbb{E}[CW_2]$ and $\mathbb{E}[CW_2^2]$ over all problem instances for the best strategies with 2, 3, and 4 support points (strategies 1, 2, and 3 respectively), with 'LIN' denoting a fit with a linear function and 'EXP' that with an exponential function. The observations are similar for $\mathbb{E}[CW_1]$ and $\mathbb{E}[CW_1^2]$. First, note that the accuracy of all strategies increase with cv_S^2: irrespective of the support points used or type of function fitted, the results are most accurate when $cv_S^2 = 0.2$. We also draw the following conclusions:

- **Accuracy does not necessarily increase if we use more support points.** Indeed, accuracy then decreases for the second moment of the conditional waiting time, irrespective of the function type. We expect that the additional support points are increasingly far away from the points we wish to estimate. Hence, they do not provide further accuracy.
- **With two support points, we obtain similar accuracy when fitting a linear or exponential function.** As the number of support points increases, however, the linear function is most accurate for the mean conditional waiting time, while the exponential function is most accuracy for the second moment of the waiting time.

		Avg. RE CW_2						Avg. RE CW_2^2					
		$cv_S^2 = 0$		$cv_S^2 = 0.1$		$cv_S^2 = 0.2$		$cv_S^2 = 0$		$cv_S^2 = 0.1$		$cv_S^2 = 0.2$	
	Support points	**LIN**	**EXP**	**LIN**	**EXP**	**LIN**	**EXP**	**LIN**	**EXP**	**LIN**	**EXP**	**LIN**	**EXP**
1	0.25 - 0.33	3%	4%	2%	3%	2%	2%	8%	7%	4%	4%	2%	2%
2	0.25 - 0.33 - 0.5	3%	5%	2%	3%	1%	2%	13%	9%	6%	5%	2%	3%
3	0.25 - 0.33 - 0.5 - 1	3%	8%	2%	5%	1%	2%	32%	16%	17%	9%	7%	4%

Table 7.8 Aggregate relative errors under various strategies when estimating $\mathbb{E}[CW_2]$ and $\mathbb{E}[CW_2^2]$.

Overall, we find the best results when using two support points that have a low squared coefficient of variation, where a linear and an exponential function provide similar accuracy. Still, the method is not sufficiently accurate for estimating performance when $cv_S^2 = 0$: then, the maximum relative error to simulation can amount to 15% and 20% under the exponential and linear function respectively. For larger values of cv_S^2, the accuracy is reasonable, with a maximum relative error of 10% for both types of functions.

From Sections 7.5.2 and 7.5.3, we conclude that analysis method AVA1 generally is most accurate, with Williams' method being a good alternative only for estimating the conditional waiting time moments of class K. We also found that the scaling of the service time distribution is accurate and fast, and is indeed a necessary tool for analyzing large problem instances in reasonable time. The remaining extension options work well under specific conditions. We now apply a subset of these methods – specifically AVA1 with scaling (extension option 1) and interpolation (extension option 3) – to a case at a manufacturer of printing and copying equipment that has two types of customers. For simplicity, we do not consider further analysis methods, such as that of Williams. We also omit extension option 2 (i.e., the estimation of class 2 waiting time moments from those of classes 1 and 3), since we consider a two-class setting.

7.6 Case study

We now consider one service region with two customer classes that each have distinct service level requirements on the *overall* (i.e., unconditional) waiting time: the waiting time for the premium class should *always* be below 3 hours, while the *average* waiting time for the non-premium class should remain below 3.5 hours. Table 7.9 gives the remaining parameter values.

Parameter	Values for case study
ρ	0.93
$\mathbb{E}[S]$ (hours)	2.3662
cv_S^2	0.2161
Division in classes $(\lambda_1/\lambda; \lambda_2/\lambda)$	(0.15; 0.85)

Table 7.9 Parameter values for the case study.

A service region is generally serviced by 4 engineers. In Section 7.6.1, we therefore first evaluate performance under that setting. We shall see that the service target for class 2 cannot be met then. In Section 7.6.2, we therefore consider two strategies for meeting all service level targets.

7.6.1 Performance under the current capacity

First, we compute the first two moments of the conditional waiting time per class. To this end, we use linear interpolation with the waiting time moments in an Erlang-5 distribution (with $cv_S^2 = 0.2$) and an Erlang-4 distribution (with $cv_S^2 = 0.25$) as support points[11]. We expect to find accurate results in this way, as our value for cv_S^2 lies between 0.2 and 0.25. Given $\mathbb{E}[CW_k]$ and $\mathbb{E}[CW_k^2]$ for both classes k, we estimate both the distribution of W_1 (the overall class-1 waiting time) and the mean overall class-2 waiting time $\mathbb{E}[W_2]$ (see Appendix C for details). Table 7.10 shows the conditional waiting time moments per class and the performance on the overall waiting time targets. Although the class-1 requirement is almost always met, the mean waiting time for class 2 is far larger than 3.5 hours.

	$\mathbb{E}[CW_1]$	$\mathbb{E}[CW_1^2]$	$\mathbb{E}[CW_2]$	$\mathbb{E}[CW_2^2]$	$\Pr\{W_1 \leq 3\ hours\}$	$\mathbb{E}[W_2]$
AVA1	0.54	0.50	6.10	72.65	0.999	5.18

Table 7.10 The first two moments of the conditional waiting time per class and the performance on the service level targets ($c = 4$).

[11] Incidentally, we are also able to fit an Coxian–5 distribution to the service parameters.

7.6.2 Options for meeting the service level targets

We have two options to reduce the class-2 waiting time, while ensuring that the class-1 waiting time never exceeds 3 hours. First, we can increase the number of servers. Alternatively, we may consider a more dynamic priority mechanism for service engineer assignment. As class 1 customers always have priority over class 2 customers at present, it may be that the class-1 waiting times are lower than required at the expense of the class-2 waiting times. Therefore, we prefer a mechanism where a new class 1 customer does not have priority over a class 2 customer that has already been waiting for a certain amount of time. Still, system analysis quickly becomes complicated under such a priority mechanism. Therefore, we emulate a softer priority mechanism as follows: An arriving class 2 customer is treated as a class 1 customer with a probability p, with p being any value between 0 and 1. The value of p influences the waiting times of both classes: as p increases, a fraction of class 2 customers experiences a lower waiting time, which might reduce the overall waiting time for that class. Conversely, class 1 customers now occasionally need to wait for an 'upgraded' class 2 customer, which can increase the class-1 waiting times.

We now use the following approach to determine values for c and p:

1. Set c to its original value. In our case study c will thus equal 4.
2. For the current value of c, compute the service level targets both when (A) no class 2 customer is treated as a class 1 customer (corresponding to $p = 0$), and when (B) all customers are treated equally, i.e., $p = 1$.
3. Depending on the outcome of the previous step, do the following:
    a. If the targets for both classes are met under either (A) or (B), STOP.
    b. If the target for class 1 is not met under (A), it will certainly not be met for $p > 0$. Conversely, if the class-2 target is not met under (B), it will not be met for $p < 1$. In both cases, increase c by 1 unit and proceed to step 2.
    c. If the class-1 target is met under (A), while the class-2 target is met under (B), it might be possible to meet both targets by setting $p > 0$. Proceed to step 4. Otherwise, increase c by 1 unit and proceed to step 2.
4. Use bisection to check whether a value for p exists such that the service targets are satisfied for both classes. Proceed until either all targets are satisfied (we then STOP), or the class-1 target is no longer satisfied (we then increase c by 1 and go to step 2).

For our case study setting, we require 5 servers to meet both service level targets (Table 7.11). Increasing p when $c = 4$ has no benefit here, as we still are not able to meet the class-2 target even when $p = 1$. This is because the low priority customers comprise the bulk of the workload: reducing their waiting time has a strong impact on the waiting time of high priority customers.

c	p	$\mathbb{E}[CW_1]$	$\mathbb{E}[CW_1^2]$	$\mathbb{E}[CW_2]$	$\mathbb{E}[CW_2^2]$	π_w	$\Pr\{W_1 \leq 3\ hours\}$	$\mathbb{E}[W_2]$
4	0	0.54	0.50	6.10	72.43	0.85	0.999	5.18
4	1	5.26	53.25	5.26	53.25	0.85	0.508	4.47
5	0	0.44	0.33	1.39	3.56	0.45	1.000	0.63
5	1	1.24	2.72	1.24	2.72	0.45	0.967	0.56

Table 7.11 Performance on service level targets for various control options.

Overall, the impact of p depends on the type of service level considered, as shown in Figure 7.2. We base the figure mainly on the case study values (Table 7.9), with only cv_S^2 adjusted to 0.2 for simplicity. In the left figure, $\mathbb{E}[W_2]$ decreases slightly with p, while $\mathbb{E}[W_1]$ explodes for large values of p. The picture is different for the waiting time percentiles (where the figure on the right denotes per class the 90th percentile, i.e., the value X such that $\Pr\{W_k \leq X\} = 0.9$ for $k = 1{,}2$). Specifically, the class-2 percentile function initially *increases* with p. This occurs because the variability of W_2 may increase with p, since a fraction of class 2 customers is now treated as a class 1 customer (with a corresponding low waiting time), while the remaining class 2 customers have an increasingly high waiting time.

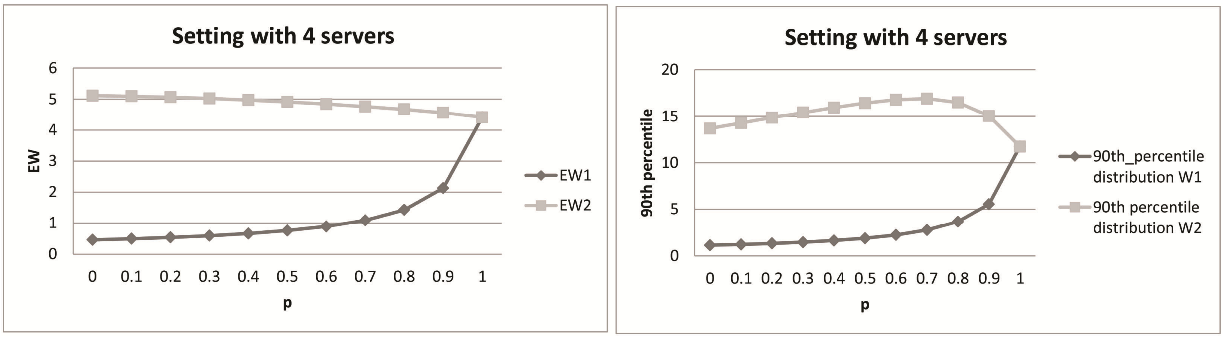


Figure 7.2 The impact of p on the mean waiting time and waiting time percentiles per class.

The impact of p also depends on the value of c and on the distribution of the total demand rate λ over the classes, as shown in Figure 7.3 and Figure 7.4 respectively for the mean waiting times per class (the conclusions are similar for the waiting time percentiles per class). For clarity, we normalize the waiting times in Figure 7.3 on the mean waiting time when $p = 0$ to show the relative impact on the waiting times. Such normalization is not needed in Figure 7.4, as all functions have the same waiting time value when $p = 1$ (which corresponds to a single-class system). In Figure 7.3, we see that the impact of p on $\mathbb{E}[W_1]$ is particularly large when c is small. In contrast, the impact of p on $\mathbb{E}[W_2]$ is relatively constant for varying values of c. Conversely, the value of λ_1/λ ('frac1' in Figure 7.4) has little impact on $\mathbb{E}[W_1]$, whereas the impact on $\mathbb{E}[W_2]$ is significant: clearly, the use of p as a priority mechanism is most beneficial for reducing $\mathbb{E}[W_2]$ in settings where λ_1/λ is large. We note that the same does not hold for the class-2 waiting time percentile, since this function simply increases more strongly with p then.

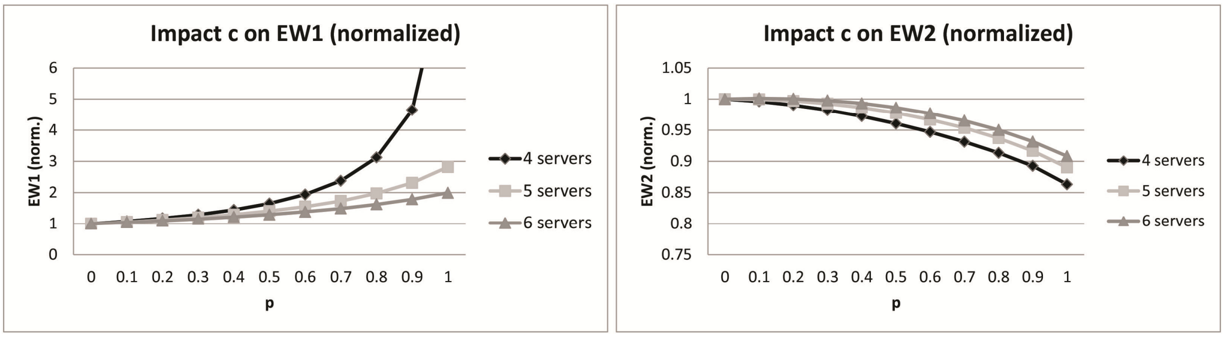


Figure 7.3 The impact of c on the mean waiting times per class (with the waiting times normalized to the value when $p = 0$).

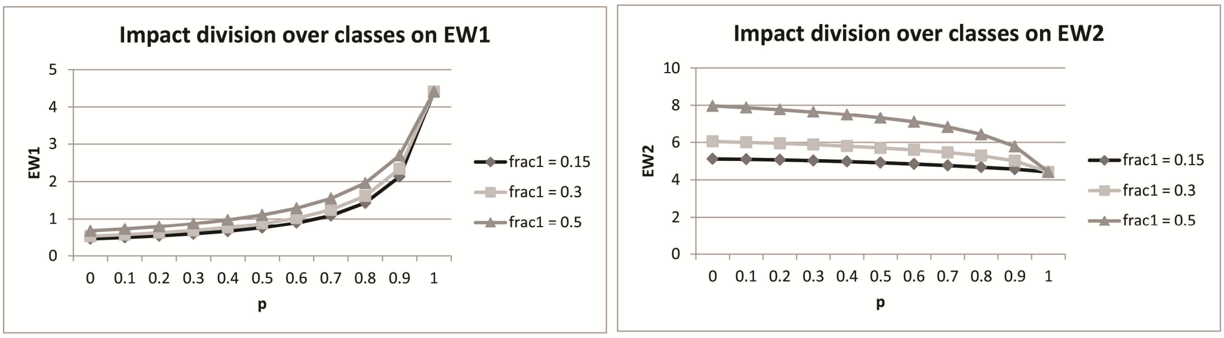


Figure 7.4 The impact of λ_1/λ on the mean waiting times per class (where 'frac1' denotes the value of λ_1/λ).

Overall, our analysis methods enable a service provider to accurately estimate performance on various types of service levels. In particular, he is now able to characterize the *distribution* of the waiting time per class from the first and second moment of the conditional waiting time per class. The service provider can use these methods both to estimate service level performance for a given number of engineers and, conversely, *to determine what service levels he can guarantee to his customers*. In this case study, for instance, the service provider must consider whether it is beneficial to guarantee a mean waiting time of at most 3.5 hours to his lowest priority customers, since he then requires a fifth service engineer to satisfy all targets.

7.7 Conclusions

In this chapter, we considered an $M/G/c$ queue with K classes and a non-preemptive service discipline. For this system, we developed two main methods to obtain the first two moments of the waiting time per class given that all servers are busy. We also presented three options for reducing computation times. We applied the various approaches to an extensive set of theoretical instances and to a case study at a manufacturer of printing and copying equipment. Our main conclusions are:

- **Overall, AVA1 is the most effective analysis method.** AVA1 generally gives the most accurate results, especially when estimating the conditional waiting time moments of the highest priority class. Furthermore, the computation time of the method is on average a fraction of a second and at most 4 seconds for settings with two customer classes.
- **In some settings, Williams' method may be a good alternative for finding the conditional waiting time moments of the lowest priority class only.** Williams' method can be more accurate than AVA1 for the conditional waiting time moments of class K, for instance in systems with high loads or few servers. As Williams' method is also very fast, it is a good alternative for class K waiting times, especially when there are 3 or more customer classes.
- **The scaling of the service time distribution is an effective option for reducing the analysis time.** The scaling of the service time distribution generally leads to accurate results: under AVA1, the average relative error to simulation for any performance measure remains below 2.5%, while the maximum relative error is 12.3%. Scaling also greatly reduces analysis time in settings with 6 or more servers and a complex service time distribution with 4 or more phases. Indeed, scaling is even necessary for analyzing queues with 9 or more servers.
- **The analysis methods allow a service provider to accurately estimate his performance on various types of service levels.** Given that the methods compute both the mean and second moment of the conditional waiting time per class, a service provider is able to estimate the distribution of the overall waiting time besides the corresponding mean value. As a result, he is able to evaluate his performance on various types of service levels and, more importantly, determine what service levels he can feasibly promise to his customers.

In the model considered in this chapter, all customer classes have the same service time distribution. Still, it might be that the service time distribution varies per customer segment, for instance if an engineer can service multiple types of systems that each require different service times, while the system type is not evenly distributed over the customer classes. It would thus be an interesting area of further research to allow the service time distribution to vary per customer segment. Such an extension will likely result in a significant increase in complexity. For instance, the distribution of the remaining service time of any busy server will now depend on the type of customer being served by that server.

In this dissertation, we have considered various control options for applying differentiation in the service fulfillment process, both in spare parts supply and in the assignment of service engineers to customers. In the next chapter, we draw our key conclusions and discuss options for further research on a broader scale.

Appendix A: The first two queue length moments in an $M/G/c$ queue

In this appendix, we briefly indicate how $\mathbb{E}[L_q]$ and $\mathbb{E}[L_q(L_q-1)]$ can be obtained from the generating function $P_q(z)$, given by equation (9.6.22) in Tijms (2003). As mentioned before, we obtain $\mathbb{E}[L_q]$ and $\mathbb{E}[L_q(L_q-1)]$ by taking the first and second derivative in $P_q(z)$ in point $z = 1$. Still, the resulting expressions initially seem very complex. Fortunately, these expressions consist of elements that can be greatly simplified, hence resulting in simple analytical expressions for $\mathbb{E}[L_q]$ and $\mathbb{E}[L_q(L_q-1)]$.

After differentiating $P_q(z)$ in $z = 1$, we find for $\mathbb{E}[L_q]$ and $\mathbb{E}[L_q(L_q-1)]$:

$$\mathbb{E}[L_q] = \frac{c}{\mathbb{E}[S]}(1-\rho)\pi_w\left(\frac{I_5}{1-\lambda\, I_1} + \frac{I_4\,\lambda\, I_2}{(1-\lambda\, I_1)^2}\right), \tag{7.29}$$

$$\mathbb{E}[L_q(L_q-1)] = \frac{c}{\mathbb{E}[S]}(1-\rho)\pi_w\left(\frac{I_6}{1-\lambda\, I_1} + \frac{2\, I_5\,\lambda\, I_2}{(1-\lambda\, I_1)^2} + \frac{2\, I_4\,\lambda^2\, I_2^2}{(1-\lambda\, I_1)^3} + \frac{I_4\,\lambda\, I_3}{(1-\lambda\, I_1)^2}\right), \tag{7.30}$$

where I_1 through I_6 pertain to the integrals listed in equations (7.31) to (7.36). Note that each integral can be greatly simplified, as shown below. Details on the derivations are given afterwards.

$$I_1 = \int_0^\infty \big(1 - S(c\cdot t)\big)dt = \int_0^\infty \big(1 - S(u)\big)\frac{du}{c} = \frac{\mathbb{E}[S]}{c}. \tag{7.31}$$

$$I_2 = \int_0^\infty \big(1 - S(c\cdot t)\big)\lambda t\, dt = \lambda\int_0^\infty (1 - S(u))\frac{u\, du}{c\ c} = \frac{\lambda}{c^2}\int_0^\infty (1 - S(u))\, u\, du = \frac{\lambda\mathbb{E}[S^2]}{2c^2}. \tag{7.32}$$

$$I_3 = \int_0^\infty \big(1 - S(c\cdot t)\big)\lambda^2 t^2\, dt = \frac{\lambda^2}{c^3}\int_0^\infty (1 - S(u)\, u^2\, du = \frac{\lambda^2\mathbb{E}[S^3]}{3c^3}. \tag{7.33}$$

$$I_4 = \int_0^\infty \left(1 - \frac{\int_0^t (1 - S(u))du}{\mathbb{E}[S]}\right)^{c-1} \big(1 - S(t)\big)dt = \frac{\mathbb{E}[S]}{c}. \tag{7.34}$$

$$I_5 = \int_0^\infty \left(1 - \frac{\int_0^t (1 - S(u))du}{\mathbb{E}[S]}\right)^{c-1} \big(1 - S(t)\big)\,\lambda t\, dt = \rho\cdot\gamma_1, \tag{7.35}$$

$$I_6 = \int_0^\infty \left(1 - \frac{\int_0^t (1 - S(u))du}{\mathbb{E}[S]}\right)^{c-1} \big(1 - S(t)\big)\,\lambda^2 t^2\, dt = 2\lambda\cdot\rho\cdot\gamma_2, \tag{7.36}$$

where γ_1 and γ_2 are defined by (7.7) and (7.10) respectively.

The rewriting of I_1 is trivial. For I_2, we find that $\int_0^\infty (1 - S(u))\, u\, du = \frac{1}{2}\mathbb{E}\left[S^2\right]$ through integration by parts. In a similar way, we obtain I_3. For I_4, we first rewrite $1 - \frac{\int_0^t (1-S(u))du}{\mathbb{E}[S]}$ as $Y(t)$ (i.e., $Y(t) = 1 - \frac{\int_0^t (1-S(u))du}{\mathbb{E}[S]}$). We then find:

$$I_4 = -\int_0^\infty \left(Y(t)\right)^{c-1} \cdot Y'(t)\ \mathbb{E}[S]\, dt = -\mathbb{E}[S]\left[\frac{\left(Y(t)\right)^c}{c}\right]_0^\infty = \frac{\mathbb{E}[S]}{c}.$$

Finally, to obtain simple expressions for I_5 and I_6, we again substitute $1 - \frac{\int_0^t (1-S(u))du}{\mathbb{E}[S]}$ by $Y(t)$. We find for I_5:

$$I_5 = -\int_0^\infty \left(Y(t)\right)^{c-1} \cdot Y'(t) \cdot \mathbb{E}[S]\, \lambda t\, dt = -\lambda \mathbb{E}[S] \int_0^\infty t \cdot \left(Y(t)\right)^{c-1} \cdot Y'(t)\, dt.$$

By integrating the latter integral by parts, we find the simplified expression for I_5. In a similar way, we find the expression for I_6.

By dividing the simple expressions for $\mathbb{E}\left[L_q\right]$ and $\mathbb{E}\left[L_q\left(L_q - 1\right)\right]$ by those for $\mathbb{E}\left[L_q(exp)\right]$ and $\mathbb{E}\left[L_q\left(L_q - 1\right)(exp)\right]$ respectively, we obtain expressions (7.6) and (7.9) in Section 7.3.1.1.

Appendix B: The influence of the utilization rate on the squared coefficient of variation of the conditional waiting time

As mentioned in Section 7.5.2.2, we observe a relationship between the utilization rate ρ and the squared coefficient of variation of the conditional waiting time, both over all classes and for the lowest priority class. Figure 7.5 and Figure 7.6 summarize the results under AVA1 for the two-class and three-class problem instances respectively. The results are similar for AVA2.

Note that the squared coefficient of variation c_{CW}^2 of the overall waiting time increases to 1 as ρ increases. The squared coefficient of variation $c_{CW_K}^2$ of the lowest priority class K also moves to 1 as ρ increases, with $c_{CW_K}^2$ usually decreasing with ρ. For the remaining classes k, the squared coefficient of variation $c_{CW_k}^2$ is constant in ρ. Still, the value of $c_{CW_k}^2$ for those classes increases with the coefficient of variation of the service time.

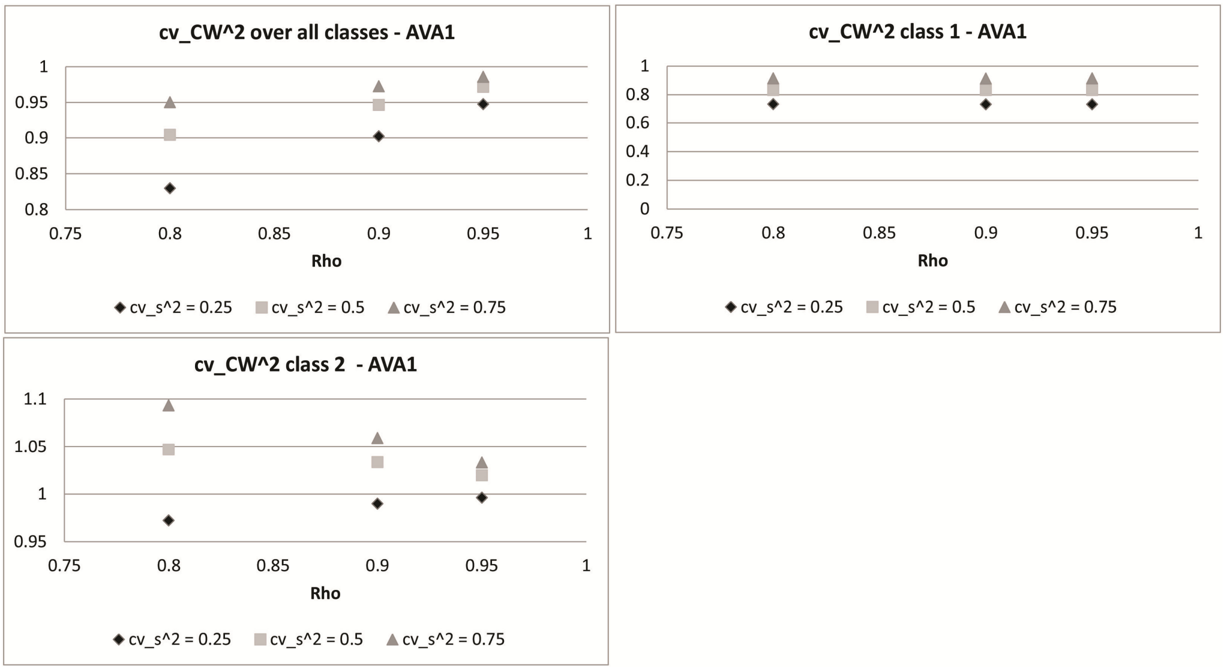


Figure 7.5 The relationship between the utilization rate ρ and the squared coefficient of variation of the conditional waiting times for two-class instances.

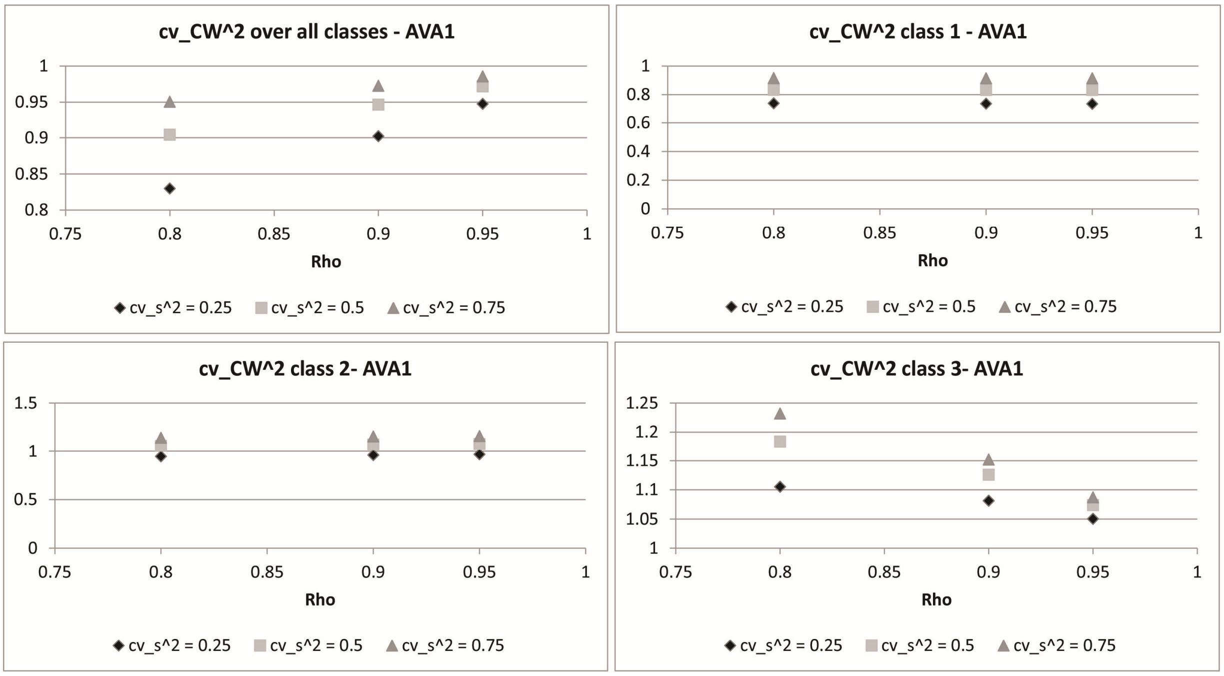


Figure 7.6 The relationship between the utilization rate ρ and the squared coefficient of variation of the conditional waiting times for three-class instances.

Appendix C: Finding the distribution of the overall waiting time, including the first two moments

Given that we have π_w (i.e., the delay probability) and the first two moments of CW_k, we can obtain the first two moments of the *unconditional* waiting time W_k as follows:

$$\mathbb{E}[W_k] = \pi_w \mathbb{E}[CW_k] \tag{7.37}$$

$$\mathbb{E}\left[W_k^2\right] = \pi_w \mathbb{E}\left[CW_k^2\right] \tag{7.38}$$

By fitting the first two moments of CW_k to a gamma distribution, we can also approximate the distribution of CW_k, and hence that of W_k. For the distribution of CW_k we find:

$$Pr\{CW_k \leq x\} = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \quad , \tag{7.39}$$

with $$\theta = \left(\frac{\mathbb{E}[CW_k]}{\mathbb{E}[CW_k^2] - E[CW_k]^2}\right)^{-1} \quad \text{, and} \tag{7.40}$$

$$k = \frac{\mathbb{E}[CW_k]^2}{\mathbb{E}[CW_k^2] - E[CW_k]^2} \tag{7.41}$$

Subsequently, we find the distribution of W_k from that of CW_k:

$$P\{W_k = 0\} = 1 - \pi_w \tag{7.42}$$

$$P\{W_k \leq x\} = P\{W_k = 0\} + P\{CW_k \leq x\}\pi_w \tag{7.43}$$

Chapter 8

Conclusions and further research

In this chapter, we draw our key conclusions, give guidelines for applying differentiation policies in practice, and discuss areas for further interesting research.

8.1 Conclusions

In this section, we reflect on the research objectives of Section 1.8.2. We start with the first main research objective:

1. *To determine whether and when throughput time reduction can lead to large cost savings in general multi-echelon multi-indenture spare parts networks.*

We consider throughput time reduction in Chapter 2. We first developed expressions for the marginal backorder reduction at operating sites as a function of the marginal reduction of throughput times in the network. As a result, we are able to estimate the marginal impact of throughput time reduction on system availability. Subsequently, we used these impact estimates to develop an efficient heuristic for simultaneously optimizing spare part inventories and repair and transportation throughput times. Using this heuristic, we find significant cost reductions compared to the standard VARI-METRIC method with fixed throughput times: 20% on average in an extensive numerical experiment with theoretical problem instances. We also tested the model on a case study at Thales Netherlands, where we find a net saving of 5.6%. The gap in savings between the theoretical cases and the Thales case is caused by the fact that Thales has limited options for throughput time reduction, especially at the shore and at the operating sites (i.e., the frigates), whereas in general it is most profitable to reduce throughput times for LRUs downstream in the supply chain.

Now, we discuss the research objectives that pertain to differentiation in spare parts supply on both an item level and a customer level. We considered three differentiation strategies besides the critical level policy: selective emergency shipments (research objective 2), selective lateral transshipments (research objective 3), and dedicated customer stocks (research objective 5). Furthermore, we considered combinations of differentiation strategies (research objective 6). Under each strategy (or combination of strategies), we analyzed the resulting system using continuous-time Markov chains. The system analysis under dedicated stocks with emergency shipments was considered as research objective 4. For multi-item optimization, we used an approach similar to Dantzig-Wolfe decomposition. Note that the application of this approach was far from trivial both for the selective transshipment model and the dedicated stocks model with emergency shipments, as system performance then depends on the stock levels at all stock points in the system. Also, in the selective transshipment model we have distinct service level requirements for each customer class *and* warehouse, making it more difficult to find a near-optimal integer solution.

We draw separate conclusions on each research objective.

2. *To determine whether and when the selective use of emergency shipments is an effective control option for applying service level differentiation in spare parts supply*

In Chapter 3, we show that the selective emergency shipment model has added value as a differentiation tool, with savings of 4.4% on average over a one-size-fits-all policy where all customers receive uniform service. The average savings even increase to 11.7% when we have many low value items, as emergency shipments are very expensive in that case. The selective emergency shipment model then uses full backordering for many items. In contrast, a critical level policy with emergency shipments keeps very high stock levels then to avoid emergency shipments. As a result, the selective emergency shipment model outperforms the critical level policy in that setting and in other settings where emergency shipments are not beneficial.

3. *To determine whether and when the selective use of lateral transshipments is effective for applying service differentiation in spare parts supply.*

In Chapter 4, we extend the selective emergency shipment model (Chapter 3) to consider lateral transshipments for premium customers only. We then show that selective transshipments are an effective tool for differentiation. On average, such transshipments are used for 96% of all item-warehouse combinations. We give further details when discussing research objective 6, specifically the combination of selective transshipments with selective emergency shipments. As in Chapter 3, we further find that it is beneficial to consider backordering as a shipment option besides emergency shipments: on average, emergency shipments for both classes are only used for 25% of the warehouses and items. For the remaining item-warehouse combinations, backordering is at least used for the non-premium class (and often for both classes). Also, we find that expensive shipment options are reserved for high value items if a warehouse is out of stock: requests for inexpensive items are backordered, and those for more expensive items are met through lateral or emergency shipments.

4. *To find an accurate and fast approach for analyzing a two-echelon model with lost sales.*

As discussed in the introduction, we obtain a two-echelon model when dedicated stocks are allowed as a differentiation tool: at the higher echelon level we find warehouse stock and at the lower level we have customer stock points. Under lost sales, no suitable approaches yet existed to analyze such a two-echelon model: existing approaches use simple approximations to analyze the warehouse and ignore the fact that the demand rate at the warehouse depends on the stock levels at the customers. By using a more accurate analysis of the warehouse arrival rates and pipeline, we developed a highly accurate model in Chapter 5 that works well in a large number of settings: deviations to simulation remain below 0.6% for 90% of all performance measure observations. Our approach is also fast (at most 47 milliseconds), making it a suitable building block for optimizing a multi-item system with dedicated stocks.

5. *To determine whether and when the use of dedicated customer stocks is an effective control option for applying service level differentiation in spare parts supply.*

We consider the use of dedicated stocks for service level differentiation in Chapter 6. In a computational experiment, we compared this differentiation strategy to both the one-size-fits-all approach and the critical level policy, both for a system under full backordering and one with emergency shipments. We show that dedicated stocks have significant added value, with average savings of 13% and 5% under backordering and emergency shipments respectively. Also, the savings are of the same order of magnitude as those found with the critical level policy. Dedicated stocks are particularly beneficial when the shipment time to customers is large. Indeed, it might even be necessary to keep dedicated stocks in that case to meet high service requirements. Under emergency shipments, both dedicated stocks and critical levels should only be used in settings with many expensive items.

6. *To investigate the added value of using multiple control options simultaneously for differentiation in spare parts supply.*

We discuss combinations of differentiation strategies in Chapters 3, 4, and 6. In Chapter 3, we show that large savings can be obtained by jointly using selective emergency shipments and critical level policies: the savings of that combination over a one-size-fits-all approach are 13.9% on average. Compared to the selective emergency shipment model, we also find large savings (14% on average) by also using selective lateral transshipments for differentiation (see Chapter 4). However, the combination of the latter strategy with critical levels does not result in additional gains, which we also show in Chapter 4. A more detailed observation shows that the first two combinations result in effective differentiation strategies (with the overall waiting times close to their respective targets). Therefore, the addition of critical levels to the selective transshipment model does not lead to additional benefit. The added value of combining dedicated customer stocks with the critical level policy is also small, as shown in Chapter 6.

Our final research objective pertains to the use of priority mechanisms for applying differentiation in assigning service engineers to customers.

7. *To determine the impact on service level performance of using priority mechanisms for assigning service engineers to customers.*

We discuss this objective in Chapter 7, where we consider a multi-class system with multiple servers where high priority customers have non-preemptive priority over lower priority customers. For this model, we developed two main methods to obtain the first two moments of the waiting time per customer class given that all servers are busy, which in turn allow us to estimate the overall waiting time distribution per class. Furthermore, we presented three options for reducing the computation time. Overall, we find that analysis method AVA1 combined with scaling of the service time distribution results in high accuracy and fast computation times. When applied to a case study, this combination accurately estimated the performance on various types of service levels. Also, the method serves as a useful tool for service providers to set realistic service level targets for their customers.

8.2 Guidelines for applying differentiation policies

In this dissertation, we have presented various control options for implementing service differentiation in the fulfillment process. We now offer guidelines on applying these differentiation policies in practice, where we focus on the policies related to spare parts supply. Our aim is to globally indicate how the suitability of a policy depends on the characteristics of the items being considered. Naturally, a service provider must also consider the influence of other parameters such as replenishment times when deciding what policies to use.

Figure 8.1 shows which policies should be considered for varying types of items. In general, we see clear links between item type and policy characteristics,. First, notice that the **item value influences the shipment mode** used. When the item value is low, holding costs are low as well. As a result, emergency shipments are very expensive and should not be used. As the item value increases, it becomes increasingly beneficial to use lateral transshipments and emergency shipments. Second, we find that the **demand rate influences the mode of differentiation**. For slow movers, the most suitable differentiation option – if any – is selective transshipments. We then only need to keep stock at a few locations and use transshipments to satisfy premium requests at other locations. If stock only needs to be kept at one location, the system in fact simplifies to a two-echelon model where the depot also serves customers directly. Still, even lateral transshipments may be too expensive when the item value is also low. Then, the only beneficial option is full backordering. Notice that the critical level policy also has little added value when an item's demand rate is low: as we keep little stock of that item, we have few options for applying differentiation in that way. Conversely, for fast movers it is beneficial to consider the critical level policy. The benefit of the remaining differentiation modes depends on the item value as well: dedicated stocks should be limited to low value items, whereas selective emergency shipments are useful for high value items.

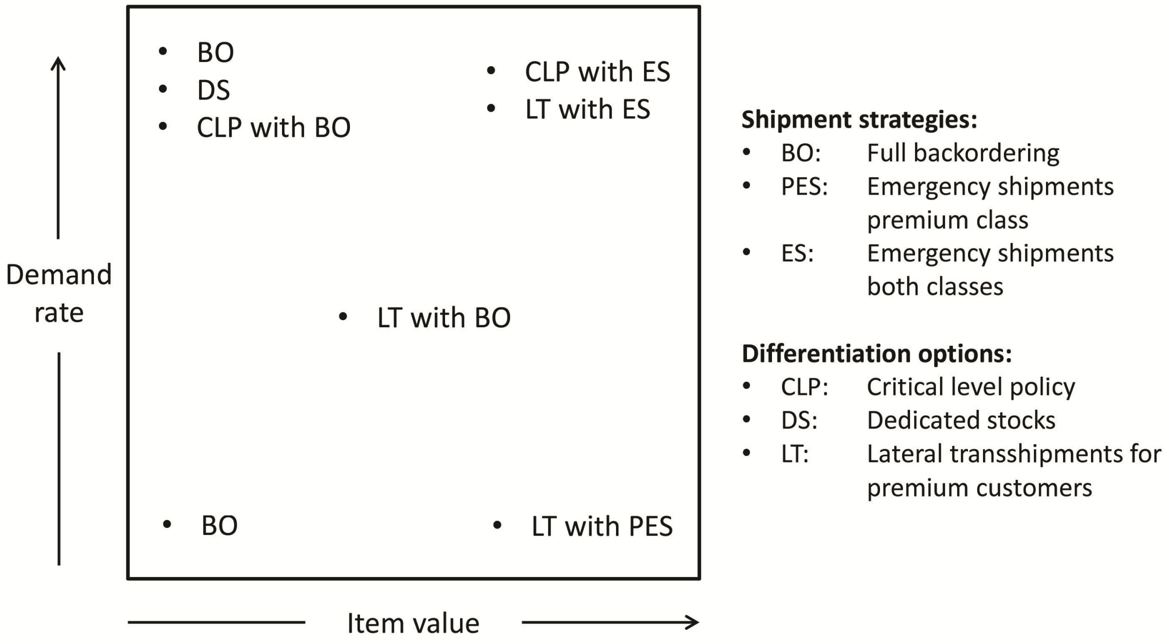


Figure 8.1 Suitability of differentiation policies for item types.

Figure 8.1 gives a general relationship between item characteristics and policy suitability. In order to specify what policy must be used for a specific item, we also require the remaining parameters of a problem instance. For instance, if regular replenishment times are very large compared to emergency shipment times, the use of emergency shipments might greatly limit the stock levels that must be kept. As a result, this shipment mode might also be used for items of moderate value. Conversely, emergency shipments will only be used for the highest value items, if any, when regular replenishment times are sufficiently short to meet the service level targets.

8.3 Discussion and further research areas

We now discuss areas for further research. As we performed research at a tactical planning level, we first discuss extension options at this level. Subsequently, we discuss extension options at the operational and strategic level.

8.3.1 Extensions at the tactical level

In practice, a service provider may have three or four customer segments. In that case, the service differentiation models need to be **extended to more than two customer classes.** Such an extension can occur in two ways: either (i) at a system level, or (ii) at an item level. At a *system level*, we may opt to have at most 2 customer classes for each item policy – which simplifies system analysis for an item policy – while the assignment of customers to classes may differ across item policies. The optimization approach then remains similar as before, we simply need to account for the assignment of customers to classes when selecting item policies and we must consider more service level restrictions (i.e., one for each class). Note that a customer might only be assigned to the high priority class for a subset of items. Then, decisions are also required at an operational level to determine whether a customer order for multiple items which have distinct priorities should be treated as a high priority or a low priority order. It is also possible to extend the models at an *item level*, i.e., by allowing more than 2 customer classes per item policy. Such an extension is not straightforward under (partial) backordering with priority backorder clearing, especially when combined with a critical level policy. Then, each class adds an extra dimension to the state space of the Markov chain to keep track of both the pipeline and the number of backorders per class. Incidentally, we doubt whether an extension at item level will result in serious additional savings: Kranenburg and Van Houtum (2008) have shown that a smart assignment of customers to classes (with at most 2 classes per item policy) can result in savings that are close to those of allowing 5 classes per item policy.

A second interesting area for further research is on the **incorporation of state information when selecting a shipment mode.** For instance, if we know that the pipeline to a warehouse contains many items, it might be both faster and cheaper to backorder incoming demand instead of using expensive lateral transshipments and emergency shipments. As a result, we might reduce system costs (either by keeping lower stock levels or by limiting expensive shipment modes). However, to do so, information on the system state must be available. In practice, however, such information often is not available, either because it is not monitored or because it is not easily accessible. Furthermore, the inclusion of state information will greatly complicate system analysis, and hence system optimization.

The incorporation of state information could also be interesting when assigning service engineers to customers. Specifically, it would be interesting to consider a **dynamic priority assignment mechanism** that considers both the priority and current waiting time of customers when deciding where to send an engineer. In Chapter 7, for instance, we have seen that a static priority mechanism where premium customers always have priority over non-premium customers can be undesirable, since the waiting times for non-premium customers might become excessively large. In such a setting, it might be better to 'upgrade' a non-premium customer to a premium status once he has waited for more than certain amount of time. Naturally, system analysis will be greatly complicated under such priority mechanism. Also, research must also be done on the circumstances under which such a priority mechanism is effective: in Section 7.6, we considered a simple mechanism where an arriving non-premium customer could become a premium customer with a probability p, with p being a decision variable. Then, we found that it was not always interesting to set $p > 0$, particularly in settings with relatively few premium customers and few servers.

The possibilities for further research mentioned so far pertain to the research described in this dissertation. However, we see further interesting research areas that do not pertain to spare parts or service engineers. In particular, we still see options for considering **preventive maintenance** as a control option. In practice, preventive maintenance often occurs at the same time as corrective maintenance, i.e., a service engineer replaces components that are likely to fail soon when he comes to repair a component that has already failed. However, there are benefits both for the customer and for the service provider to using preventive maintenance as a control option: a key advantage of preventive maintenance is that it can be planned, as opposed to corrective maintenance. Therefore, preventive maintenance can be scheduled at times that are convenient for all parties. Also, the service provider can then ensure that all required resources are available at the time that maintenance must be performed, which minimizes system downtime. We even see possibilities for using preventive maintenance as a differentiation tool in settings with multiple customer classes. For instance, we might choose to monitor the systems of premium customers and perform preventive maintenance before a failure occurs. In contrast, for non-premium customers we either do not perform preventive maintenance or do so less frequently than for premium customers to limit monitoring and preventive maintenance costs. To use such a differentiation approach, the service provider must be able to monitor the system's state to detect impending failures. Also, the benefit of such an approach will depend on the costs for preventive maintenance relative to the benefit of fewer corrective maintenance visits.

Finally, to our knowledge little research has been performed that considers **multiple resources or processes simultaneously.** For instance, in spare parts models the availability of other resources such as service engineers is generally ignored. However, maintenance can only be performed if all resources are available. Similarly, the use of preventive maintenance as a control option will influence the time at which engineers and spare parts are needed. By considering multiple resources and processes, we can better estimate how long a system is down and hence whether we attain our guarantees on uptime.

8.3.2 Extensions at the operational level

In this dissertation, we have described various control options for service fulfillment at a tactical level. Still, service providers must implement these control options at the operational level as well. One research area is the **operational planning of resources:** given the quantity and deployment of resources, one must determine how to use these resources for system upkeep. A key issue is that a service provider may use his resources in a way that deviates from the decisions made at a tactical level. For instance, if the service provider knows that there is only one unit of an item in stock at the warehouse – with this unit reserved for a premium customer – he may still opt to use the unit for an incoming non-premium request if he knows that a new unit will arrive shortly thereafter. Similarly, if a non-premium customer has been waiting for an engineer for a very long time, it might be beneficial to assign the next available engineer to that customer instead of to a newly arriving premium customer. Further research should thus show under what circumstances such deviations are beneficial. Note that the use of resources will depend on the **operational management of service contracts**, taking into account the current performance in the ongoing service period For instance, consider the setting where a service provider has only one unit of a part requested by two customers. Overall, the availability requirements of these customers may be 98% and 95% respectively. However, when the end of the contract period is near, it may be that the attained availability for the first customer is 100% (i.e., that customer has had no failures), whereas the availability for the second customer is 93%. Then, it is likely more logical to provide the part to the second customer.

8.3.3 Extensions at the strategic level

We first see opportunities for extending research on the **development of service contracts**. Specifically, if a service provider needs to draw up a service contract for a new customer, he must be able to define the correct *type* of service level agreements as well as the correct *values* of service levels that he can agree upon given his current set of service contracts and his service organization. This issue still remains far from trivial: one must then know the frequency at which systems fail and the resources/costs required to repair these failures. In terms of resources, we have considered the costs related to spare parts and service engineers. As shown in Chapter 1, however, a service provider requires many more resources, such as those for diagnosis and preventive maintenance. Furthermore, many service providers have little insight in the failure rates of their systems, even for their current contracts.

A second option for further research is the consideration of **performance indicators that better match with customers' needs**. In the spare parts models considered in this dissertation, and those considered in the literature in general, the performance targets pertained to mean values (e.g. a mean overall waiting time). In practice, however, customers generally wish to know that the waiting time will be reasonable each time the system fails. A target will then be required on the distribution of the waiting time, as we considered in the service engineer model of Chapter 7. To our best knowledge, only a few papers, such as Caggiano et al. 2007, have considered such targets before in spare parts models. A target on the distribution of a performance indicator will also be necessary if we aim to determine performance over a specific interval as opposed to a long-term average. Such targets on performance over an interval are often seen in practice, where customers are interested, for instance, in system availability over a year. However, to analyze performance over an interval, we require models that are highly complex, see e.g. Al Hanbali and Van der Heijden (2011), who estimate the distribution of interval availability (i.e. system availability over a specific interval). In conclusion, various areas for interesting research remain.

Bibliography

- Aarts, E., and Lenstra, J.K., (2003). *Local search in combinatorial optimization*. Princeton University Press.

- Abouee-Mehrizi, H., Baron, O., and Berman, O., (2012). Customer differentiation in capacitated multi-echelon inventory systems. Working paper. Available online: http://www-2.rotman.utoronto.ca/opher.baron/files/Multi-Echelon%20Inventory%20ManagementGMR.pdf

- Adan, I.J.B.F., van Eenige, M.J.A., and Resing, J.A.C., (1995). Fitting discrete distributions on the first two moments, *Probability in the Engineering and Informational Sciences,* 9(4), pp. 623-632.

- Adan, I.J.B.F., Sleptchenko, A., and van Houtum, G.J., (2009). Reducing costs of spare parts supply systems via static priorities, *Asia-Pacific Journal of Operational Research,* 26(4), pp. 559-585.

- Aggarwal, P.K., and Moinzadeh, K., (1994). Order expedition in multiechelon production/distribution systems. *IIE Transactions,* 26(2), pp. 86–96.

- Agnihothri, S.R., Mishra, A.K., and Simmons, D.E., (2003). Workforce cross-training decisions in field service systems with two job types. *Journal of the Operational Research Society*, 54(4), pp. 410-418.

- Alfredsson, P., (1997). Optimization of multi-echelon repairable item inventory systems with simultaneous location of repair facilities, *European Journal of Operational Research,* 99, pp. 584–595.

- Alfredsson, P., and Verrijdt, J., (1999). Modeling emergency supply flexibility in a two-echelon inventory system. *Management science,* 45, pp. 1416 – 1431.

- Al Hanbali, A., (2001). Busy period analysis of the level dependent $PH/PH/1/K$ queue. *Queueing Systems: Theory and Applications,* 67(3), pp. 221 – 249.

- Al Hanbali, A., de Haan, R., Boucherie, R., and van Ommeren, J.-K., (2012). Time-limited polling systems with batch arrivals and phase-type service times. *Annals of Operations Research,* 198(1), pp. 57 – 82.

- Al Hanbali, A., and van der Heijden, M.C., (2011). Interval availability analysis of a two echelon, multi-item system. *Beta working paper series,* 359, http://beta.ieis.tue.nl/node/1967

- Al Hanbali, A., Alvarez, E.M., and van der Heijden, M.C., (2013). Approximations for the waiting time distribution in an $M/G/c$ priority queue. *Beta working paper series,* 411, http://beta.ieis.tue.nl/node/2084

- Altinkemer, K., Bose, I., and Pal, R., (1998). Average waiting time of customers in an $M/D/k$ queue with nonpreemptive priorities. *Computers & Operations Research*, 25(4), pp. 317 – 328.

- Alvarez, E.M., and van der Heijden, M.C., (2011). On two-echelon inventory systems with Poisson demand and lost sales. *Beta working paper series,* 366, http://beta.ieis.tue.nl/node/1978

- Alvarez, E.M., van der Heijden, M.C., Vliegen, I.M.H., and Zijm, W.H.M., (2012). Service differentiation through selective lateral transshipments. *Beta working paper series,* 395, http://beta.ieis.tue.nl/node/2050

- Alvarez, E.M., van der Heijden, M.C., and Zijm, W.H.M., (2013b). The selective use of emergency shipments for service-contract differentiation. *International Journal of Production Economics,* 143(2), pp. 518 – 526.

- Alvarez, E.M., van der Heijden, M.C., and Zijm, W.H.M., (2013b). Service differentiation in spare parts supply through dedicated stocks. *Annals of Operations Research*, in press, DOI: 10.1007/s10479-013-1362-z

- Andersson, J., and Melchiors, P., (2001). Two-echelon inventory model with lost sales, *International Journal of Production Economics,* 69(3), pp. 307-315.

- Armistead, C., and Clark, G., (1991). A framework for formulating after-sales support strategy. *International Journal of Physical Distribution & Logistics Management,* 21(9), pp. 22-29.

- Arslan, H., Graves, S.C., and Roemer, T.A., (2007). A single-product inventory model for multiple demand classes. *Management Science,* 53, pp. 1486–1500.

- Axsäter, S., (1990). Modelling emergency lateral transshipments in inventory systems. *Management Science,* 36(11), pp. 1329–1338.

- Axsäter, S., Kleijn, M., and de Kok, T.G., (2004). Stock rationing in a continuous review two-echelon inventory model, *Annals of Operations Research,* 126, pp. 177-194.

- Bassamboo, A., Harrison, J.M., and Zeevi, A., (2006). Design and control of a large call center: Asymptotic analysis of an LP-based method. *Operations Research*, *54*(3), pp. 419-435.

- Basten, R.J.I., (2010). *Designing logistics support systems: Level of repair analysis and spare parts inventories.* PhD thesis, Beta Research School, University of Twente, Enschede, The Netherlands.

- Basten, R.J.I., and van Houtum, G.J., (2013). Near-optimal heuristics to set base stock levels in a two-echelon distribution network, *International Journal of Production Economics,* 143(2), pp. 546–552.

- Basten, R.J.I., van der Heijden, M.C., Schutten, J.M.J., and Kutanoglu, E., (2012a), An approximate approach for the joint problem of level of repair analysis and spare parts stocking. *Annals of Operations Research*: in press, pp. 1-25.

- Basten, R.J.I., van der Heijden, M.C., and Schutten, J.M.J., (2012b). Joint optimization of level of repair analysis and spare parts stocks. *European Journal of Operational Research*, 222(3) , pp. 474-483.

- Benjaafar, S., ElHafsi, M., and Huang, T., (2010). Optimal control of a production-inventory system with both backorders and lost sales. *Naval Research Logistics*, *57*(3), pp. 252-265.

- Benjaafar, S., ElHafsi, M., Lee, C-Y., Zhou, W., (2011). TECHNICAL NOTE—Optimal Control of an Assembly System with Multiple Stages and Multiple Demand Classes. *Operations research*, 59(2), pp. 522-529.

- Bertsimas, D., and Daisuke, N., (1992). Transient and busy period analysis of the $GI/G/1$ queue: the method of stages. *Queuing Systems*, 10(3), pp. 153 – 184.

- Bertsimas, D., and Nakazato, D., (1995). The distributional little's law and its applications. *Operations Research*, 43(2), pp. 298 – 310.

- Bijvank, M., and Vis, I.F.A., (2011). Lost-sales inventory theory: A review. *European Journal of Operational Research* 215(1), pp. 1-13.

- Buzen, J., and Bondi, A., (1983). Response times of priority classes under preemptive resume in $M/M/m$ queues. *Operations Research*, 31(3), pp. 456 – 465.

- Caglar, D., Li, C-L., and Simchi-Levi, D., (2004). Two-echelon spare parts inventory system subject to a service constraint, *IIE Transactions*, 36(7), pp. 655 — 666.

- Caggiano, K.E., Muckstadt, J.A., and Rappold, J.A., (2006). Integrated real-time capacity and inventory allocation for reparable service parts in a two-echelon supply system, *Manufacturing and Service Operations Management,* 8, 292 - 319.

- Caggiano, K.E., Jackson, P.L., Muckstadt, J.A., and Rappold, J.A., (2007). Optimizing Service Parts Inventory in a Multiechelon, Multi-item Supply Chain with Time-based Customer Service-Level Agreements. *Operations Research,* 55, pp. 303 – 318.

- Cohen, J.W., (1969) *The Single Server Queue*, Section III. 3.8(i). North-Holland Amsterdam.

- Cohen, M., Agrawal, N., and Agrawal, V., (2006). Winning in the aftermarket. *Harvard Business Review,* 84(5), pp. 129–138.

- Cohen, M.A., (2012). *Product Performance Based Business Models: A Service Based Perspective.* 45th Hawaii International Conference on System Science (HICSS).

- Colen, P.J., and Lambrecht, M.R., (2012). Cross-training policies in field services. *International Journal of Production Economics*, 138, pp 76 – 88.

- Cosmetatos, G., (1976). Some approximate equilibrium results for the multi-server queue ($M/G/r$). *Operational Research Quarterly*, pp. 615 – 620.

- Dada, M., (1992). A two-echelon inventory system with priority shipments. *Management Science,* 38(8), pp. 1140–1153.

- Dantzig, G.B., and Wolfe, P., (1960). Decomposition principle for linear programs. Operations research, 8, pp. 101-111.

- Dear, R.G., and Sherif, J.S., (2000). Using simulation to evaluate resource utilization strategies. *Simulation*, 74(2), pp. 75-83.

- Dekker, R., Hill, R.M., Kleijn, M.J., and Teunter, R.H., (2002). On the $(S-1,S)$ lost sales inventory model with priority demand classes, *Naval Research Logistics,* 49, pp. 593-610.

- Deloitte (2007). *The Service Revolution in Global Manufacturing Industries*. http://www.deloitte.com/view/en_TR/tr/industries/manufacturing/bcff5c968b0fb110VgnVCM100000ba42f00aRCRD.htm#

- Deshpande, V., Cohen, M. A., and Donohue, K., (2003). A Threshold Inventory Rationing Policy for Service-Differentiated Demand Classes. *Management Science*, 49(6), pp. 683-703.

- De Véricourt, F., Karaesmen, F., and Dallery, Y., (2002). Optimal stock allocation for a capacitated supply system. *Management Science,* 48, pp. 1486 – 1501.

- Diaz, A. and Fu, M.C., (1997). Models for multi-echelon repairable item inventory systems with limited repair capacity, *European Journal of Operational Research,* 97, pp. 480–492.

- Drekic, S., and Woolford, D.G., (2005). A preemptive priority queue with balking. *European journal of operational research*, 164(2), pp. 387-401.

- ElHafsi, M., Camus, H., and Craye, E., (2010). Managing an integrated production inventory system with information on the production and demand status and multiple non-unitary demand classes. *European Journal of Operational Research* 207(2), pp. 986-1001.

- Enders, P., Adan, I., Scheller-Wolf, A., and van Houtum, G.J., (2012). Inventory Rationing for a System with Heterogeneous Customer Classes. *Flexible Services and Manufacturing Journal*, in press, DOI: 10.1007/s10696-012-9148-1

- Eppen, G., and Schrage, L., (1981). Centralized ordering policies in a multi-warehouse system with lead times and random demand. In: *Multi-level Production/Inventory Control Systems: Theory and Practice*, North-Holland, L.B. Schwarz, Editor, pp. 51–67.

- Fadıloğlu, M.M., and Bulut, Ö., (2010). An embedded Markov chain approach to stock rationing. *Operations Research Letters*, 38(6), pp. 510-515.

- Feeney, G.J., and Sherbrooke, C.C., (1966). The (S − 1, S) Inventory Policy Under Compound Poisson Demand. *Management Science,* 12(5), pp. 391-411.

- Gallego, G., Özer, Ö., and Zipkin, P.H., (2007). Bounds, heuristics, and approximations for distribution systems. *Operations Research,* 55(3), pp. 503-517.

- Gayon, J.-P., De Véricourt, F., and Karaesmen, F., (2009). Stock rationing in an $M/E_r/1$ multi-class make-to-stock queue with backorders. *IIE Transactions,* 41, pp. 1096 – 1109.

- Gilmore, P.C., and Gomory, R.E., (1961). A linear programming approach to the cutting-stock problem. *Operations research*, 9, pp. 849 – 859.

- Grahovac, J., and Chakravarty, A., (2001). Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items. *Management Science,* 47, pp. 579-594.

- Graves, S.C., (1985). A multi-echelon inventory model for a repairable item with one-for-one replenishment, *Management Science,* 31(10), pp. 1247 – 1256.

- Gross, D., Miller, D.R., and Soland, R.M., (1983), A closed queuing network model for multi-echelon repairable item provisioning, *IIE Transactions,* 15(4), pp. 344-352.

- Gross, D., Shortle, J.F., Thompson, J.M., and Harris, C.M., (2008). *Fundamentals of queuing theory,* 4th edition, John Wiley & Sons, Hoboken, New Jersey.

- Guajardo, J.A., Cohen, M.A., Kim, S-H., and Netessine, S., (2012). Impact of performance-based contracting on product reliability: an empirical analysis. *Management Science*, 58(5), pp. 961-979.

- Gurvich, I., Armony, M., and Mandelbaum, A., (2008). Service-level differentiation in call centers with fully flexible servers. *Management Science*, 54(2), pp. 279-294.

- Ha, A.Y., (1997a). Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science,* 43, 1093 – 1103.

- Ha, A.Y., (1997b). Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics,* 44, 457-472.

- Ha, A.Y., (2000). Stock rationing in an $M/E_k/1$ make-to-stock queue. *Management Science,* 46, pp. 77 – 87.

- Hans, E.W., (2001). *Resource loading by branch-and-price techniques.* PhD thesis, Beta Research School, University of Twente, Enschede, The Netherlands.

- Harchol-Balter, M., Osogami, T., Scheller-Wolf, A., and Wierman, A., (2005). Multi-server queueing systems with multiple priority classes. *Queueing Systems*, 51(3), pp. 331-360.

- Harrison, J.M., and Zeevi, A., (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, 7(1), pp. 20-36.

- Haugen, D.L., and Hill, A.V., (1999). Scheduling to Improve Field Service Quality. *Decision Sciences*, 30(3), pp. 783-804.

- Hausman, W.H., and Scudder, G.D., (1982). Priority scheduling rules for repairable inventory systems, *Management Science,* 28, 1215-1232.

- Hausman, W.H., and Erkip, N.K., (1994). Multi-echelon vs. single-echelon inventory control policies for low demand items. *Management science,* 40, pp. 597 – 602.

- Hill, A.V., (1992). An Experimental Comparison of Dispatching Rules for Field Service Support. *Decision Sciences*, *23*(1), pp. 235-249.

- Hill, A.V., March, S.T., Nachtsheim, C.J., and Shanker, M.S., (1992). An approximate model for field service territory planning. *IIE Transactions,* 24(1), pp. 2-10.

- Hill, R.M., (2007). Continuous-review, lost-sales inventory models with Poisson demand, a fixed lead time and no fixed order cost, *European Journal of Operational Research*, 176(2), pp. 956-963.

- Hill, R.M., Seifbarghy, M., and Smith, D.K., (2007). A two-echelon inventory model with lost sales, *European Journal of Operational Research,* 181, pp. 753-766.

- Hopp, W.J., Zhang, R.Q., and Spearman, M.L., (1999). An easily implementable hierarchical heuristic for a two-echelon spare parts distribution system. *IIE Transactions*, 31(10), pp. 977-988.

- Jagerman, D.L., and Melamed, B., (2003). Models and approximations for call center design. *Methodology and Computing in Applied Probability,* 5(2), pp. 159-181.

- Jalil, M., Zuidwijk, R., Fleischmann, M., and Nunen, J.V., (2010). Revenue Management and Spare Parts Logistics Execution. Research Paper ERS-2010-XXX-LIS, Erasmus Research Institute of Management (ERIM).

- Jalil, M.N., (2011). Customer information driven after sales service management: lessons from spare parts logistics, PhD thesis, Erasmus Research Institute of Management (ERIM), http://repub.eur.nl/res/pub/22156/

- Janssen, A., and van Leeuwaarden, J., (2008). Back to the roots of the $M/D/s$ queue and the works of erlang, crommelin and pollaczek. *Statistica Neerlandica*, 62(3), pp. 299 – 313.

- Jardine, A.K.S., and Tsang, A.H.C., (2006). *Maintenance, replacement, and reliability: theory and applications.* CRC press.

- Karush, W., (1957). A queuing model for an inventory problem. *Operations Research,* 5, pp. 693–703.

- Keilson, J., and Servi, L., (1990). The distributional form of little's law and the fuhrmann cooper decomposition. Operations Research Letters, 9(4), pp. 239 – 247.

- Kella, O., and Yechiali, U., (1985). Waiting times in the non-preemptive priority $M/M/c$ queue. *Stochastic Models*, 1(2), pp. 257 – 262.

- Kim, S-., Cohen, M.A., Netessine, S., and Veeraraghavan, S., (2010). Contracting for Infrequent Restoration and Recovery of Mission-Critical Systems. *Management Science,* 56(9), pp. 1551–1567.

- Kranenburg, A.A., (2006). *Spare parts inventory control under system availability constraints.* PhD thesis, Beta Research School, Eindhoven University of Technology, Eindhoven, The Netherlands.

- Kranenburg, A.A., and van Houtum, G.J., (2007a), Effect of commonality on spare parts provisioning costs for capital goods, *International Journal of Production Economics*, 108(1-2), pp. 221-227.

- Kranenburg, A.A., and van Houtum, G.J., (2007b), Cost optimization in the (S - 1, S) lost sales inventory model with multiple demand classes, *Operations Research Letters*, 35(4), pp. 493-502.

- Kranenburg, A.A., and van Houtum, G.J., (2008), Service differentiation in spare parts inventory management, *Journal of the Operational Research Society*, 59(7), pp. 946-955.

- Kranenburg, A.A., and van Houtum, G.J., (2009), A new partial pooling structure for spare parts networks, *European Journal of Operational Research*, 199(3), pp. 908-921.

- Kukreja, A., Schmidt, C., and Miller, D., (2001). Stocking decisions for low-usage items in a multilocation inventory system. *Management Science,* 47(10), pp. 1371–1383.

- Kumar, D., (2000). *Reliability maintenance and logistic support: a life cycle approach.* Springer.

- Kutanoglu, E., (2008). Insights into inventory sharing in service parts logistics systems with time-based service levels. *Computers & Industrial Engineering*, 54(3), pp. 341–358.

- Kutanoglu, E., and Lohiya, D., (2008). Integrated inventory and transportation mode selection: A service parts logistics system. *Transportation Research Part E: Logistics and Transportation Review,* 44(5), pp. 665-683.

- Kutanoglu, E., and Mahajan, M., (2009). An inventory sharing and allocation method for a multi-location service parts logistics network with time-based service levels. *European Journal of Operational Research,* 194(3), pp. 728–742.

- Lee, H.L., (1987). A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science,* 33(10), pp. 1302–1316.

- Levin, B., (1981). A representation for multinomial cumulative distribution functions, *The Annals of Statistics,* 9(5), pp. 1123-1126.

- Levner, E., Perlman, Y., Cheng, T.C.E., and Levner, I., (2011), A network approach to modeling the multi-echelon spare-part inventory system with backorders and interval-valued demand, *International Journal of Production Economics,* 132(1), pp. 43-51.

- Lübbecke, M.E., and Desrosiers, J., (2005). Selected topics in column generation. *Operations Research*, *53*(6), pp. 1007-1023.

- Marie, R.A., (1980). Calculating equilibrium probabilities for $\lambda(n)/c_k/1/n$ queues. In *Proceedings of Performance '80,* Toronto*,* Canada, pp. 117-125.

- Moinzadeh, K., and Schmidt, C.P., (1991). An (S 1, S) inventory system with emergency orders. *Operations Research,* 39(2), pp. 308–321.

- Möllering, K.T., and Thonemann, U.W., (2010). An optimal constant level rationing policy under service level constraints. *OR Spectrum,* 32, pp. 319-341.

- Muckstadt, J.A., (1973). A Model for a Multi-Item, Multi-Echelon, Multi-Indenture Inventory System. *Management Science,* 20, pp. 472-481.

- Muckstadt, J.A. (1979). A three-echelon, multi-item model for recoverable items. *Naval Research Logistics Quarterly,* 26(2), pp. 199–221.

- Muckstadt, J.A., and Thomas, L.J., (1980). Are multi-echelon inventory methods worth implementing in systems with low-demand-rate items? *Management Science,* 26, pp. 483 – 494.

- Muckstadt, J.A., (2005). *Analysis and algorithms for service part supply chains*, Springer.

- Munnik, M., (2011). *Research on the management of service level agreements at Océ: A queueing model which can predict the waiting times of corrective maintenance jobs at the planning department.* Master thesis, University of Twente, Enschede, The Netherlands.

- Nahmias, S., and Demmy, S., (1981). Operating characteristics of an inventory system with rationing. *Management Science*, 27, pp. 1236–1245.

- Neuts, M.F., (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Johns Hopkins University Press.

- Nowicki, D.R., Randall, W.S., and Ramirez-Marquez, J.E., (2012), Improving the computational efficiency of metric-based spares algorithms, *European Journal of Operational Research*, 219(2), pp. 324-334.

- Oliva, R., and Kallenberg, R., (2003), Managing the transition from products to services, *International Journal of Service Industry Management*, 14(2), pp. 160-172.

- Öner, K.B., Kiesmüller, G.P., and van Houtum, G.J., (2010), Optimization of component reliability in the design phase of capital goods, *European Journal of Operational Research,* 205, pp. 615-625.

- Page, E., (1972). *Queueing theory in OR*. Butterworths.

- Papadopoulos, H.T., (1996). A field service support system using a queueing network model and the priority MVA algorithm, *Omega,* 24(2), pp. 195-203.

- Paterson, C., Kiesmüller, G., Teunter, R., and Glazebrook, K., (2011). Inventory models with lateral transshipments: A review. *European Journal of Operational Research,* 210(2), pp. 125–136.

- Peköz, E.A., (2002). Optimal policies for multi-server non-preemptive priority queues. *Queueing systems*, 42(1), 91-101.

- Perlman, Y., Mehrez, A., and Kaspi, M., (2001). Setting expediting repair policy in a multi-echelon repairable-item inventory system with limited repair capacity, *Journal of the Operational Research Society,* 52, pp. 198–209.

- Pourakbar, M., and Dekker, R., (2012). Customer differentiated end-of-life inventory problem. *European Journal of Operational Research*, 222, pp. 44–53.

- Prakken, M., (2009). *Customer based fill rates: the relationship between fill rates and service contracts at Philips Healthcare*. Master thesis, University of Twente, Enschede, The Netherlands.

- Pyke, D.F., (1990). Priority repair and dispatch policies for repairable-item logistics systems, *Naval Research Logistics,* 37, pp. 1–30.

- Rappold, J.A. and van Roo, B.D., (2009), Designing multi-echelon service parts networks with finite repair capacity, *European Journal of Operational Research*, 199(3), pp. 781-792.

- Reijnen, I.C., Tan, T., and Van Houtum, G.J., (2009). Inventory planning for spare parts networks with delivery time requirements. *Beta working paper series,* 280, http://beta.ieis.tue.nl/node/1461

- Riordan, J., (1962). *STOCHASTIC SERVICE SYSTEMS*. Wiley.

- Rong, Y., Bulut, Z., and Snyder, L.V., (2010). Heuristics for Base-Stock Levels in Multi-Echelon Distribution Networks. Available at SSRN: http://ssrn.com/abstract=1475469

- Rustenburg, J.W., Houtum, G.J., and Zijm, W.H.M., (2003). Exact and approximate analysis of multi-echelon, multi-indenture spare parts systems with commonality. In: J.G. Shanthikumar, D. D. Yao, W. H. M. Zijm (Eds.), *Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains*, pp. 143-176.

- Scudder, G.D., (1984). Priority scheduling and spares stocking for a repair shop: the multiple failure case, *Management Science,* 30(6), 739-749.

- Seifbarghy, M., and Jokar, M.R.A., (2006). Cost evaluation of a two-echelon inventory system with lost sales and approximately Poisson demand, *International Journal of Production Economics,* 102, pp. 244-254.

- Sherbrooke, C.C., (1968). METRIC: A Multi-Echelon Technique for Recoverable Item Control. *Operations Research,* 16, pp. 122-141.

- Sherbrooke, C.C., (1986). VARI-METRIC: Improved Approximations for Multi-Indenture, Multi-Echelon Availability Models. *Operations Research,* 34, pp. 311-319.

- Sherbrooke, C.C., (2004). *Optimal inventory modeling of systems* 2nd edition, Kluwer Academic Publishers.

- Slay, F.M., (1984). *VARI-METRIC: An Approach to Modelling Multi-Echelon Resupply when the Demand Process is Poisson with a Gamma Prior*. Logistics Management Institute, Washington, D.C. Report AF301-3.

- Sleptchenko, A., van der Heijden, M.C., and van Harten, A., (2003). Trade-off between inventory and repair capacity in spare part networks, *Journal of the Operational Research Society,* 54(3), pp. 263- 272.

- Sleptchenko, A., van der Heijden, M.C., and van Harten, A., (2005). Using repair priorities to reduce stock investment in spare part networks, *European Journal of Operational Research,* 163, pp. 733-750.

- Sleptchenko, A., van Harten, A., and van der Heijden, M.C., (2005). An exact solution for the state probabilities of the multi-class, multi-server queue with preemptive priorities. *Queueing Systems*, 50(1), pp. 81-107.

- Smith, S.A., (1977). Optimal Inventories for an (S − 1, S) System with No Backorders. *Management Science,* 23(5), pp. 522-528.

- Tang, Y., Xu, D-S., and Zhou, W-H., (2007). Inventory rationing in a capacitated system with backorders and lost sales. In *2007 IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 1579-1583.

- Tang, Q., Wilson, G.R., and Perevalov, E., (2008). An approximation manpower planning model for after-sales field service support. *Computers & Operations Research*, 35(11), pp. 3479-3488.

- Teunter, R.H., and Klein Haneveld, W.K., (2008). Dynamic inventory rationing strategies for inventory systems with two demand classes, Poisson demand and backordering. *European Journal of Operational Research,* 190(1), pp. 156–178.

- Tiacci, L., and Saetta, S., (2011). Reducing the mean supply delay of spare parts using lateral transsshipments policies. *International Journal of Production Economics,* 133(1), pp. 182-191.

- Tiemessen, H.G.H. and van Houtum, G.J., (2013). Reducing costs of repairable inventory supply systems via dynamic scheduling. *International Journal of Production Economics*, 143(2), pp. 478–488.

- Tiemessen, H.G.H., Fleischmann, M., van Houtum, G.J., van Nunen, J.A.E.E., and Pratsini, E., (2013). Dynamic demand fulfillment in spare parts networks with multiple customer classes. *European Journal of Operational Research*, 228(2), pp. 367–380.

- Tijms, H.C., (2003). *A first course in stochastic models*. Wiley.

- Topkis, D.M., (1968), Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes, *Management Science,* 15*,* pp. 160-176.

- Van der Heijden, M., van Harten, A., and Sleptchenko, A., (2004). Approximations for markovian multi-class queues with preemptive priorities. *Operations Research Letters*, 32(3), pp. 273 – 282.

- Van der Heijden, M., Alvarez, E.M., and Schutten, J.M.J., (2013). Inventory reduction in spare part networks by selective throughput time reduction. *International journal of production economics,* 143(2), pp. 509–517.

- Van Jaarsveld, W., and Dekker, R., (2009). *Finding optimal policies in the (S-1, S) lost sales inventory model with multiple demand classes.* Report Econometric institute EI 2009-14, Erasmus University Rotterdam.

- Van Houtum, G.J., and Zijm, W.H.M., (2000). On the relation between cost and service models for general inventory systems. *Statistica Neerlandica*, 54(2), pp. 127-147.

- Van Utterbeeck, F., Wong, H., Van Oudheusden, D., and Cattrysse, D., (2009). The effects of resupply flexibility on the design of service parts supply systems, *Transportation Research Part E,* 45(1), pp. 72 – 85.

- Van Wijk, A.C.C., (2012). *Pooling and polling: Creation of Pooling in Inventory and Queueing Models.* PhD thesis, Beta Research School, Eindhoven University of Technology, Eindhoven, The Netherlands.

- Van Wijk, A.C.C., Adan, I.J.B.F., and Van Houtum, G.J., (2012). Approximate evaluation of multi-location inventory models with lateral transshipments and hold back levels. *European Journal of Operational Research,* 218(3), pp. 624–635

- Veinott, A.F., (1965). Optimal Policy in a Dynamic, Single Product, Nonstationary Inventory Model with Several Demand Classes. *Operations Research,* 13, pp. 761 – 778.

- Verrijdt, J., Adan, I., and de Kok, A.G., (1998). A trade-off between emergency repair and inventory investment, *IIE Transactions,* 30, pp. 119-132.

- Vliegen, I.M.H., (2009). *Integrated planning for service tools and spare parts for capital goods.* PhD thesis, Beta Research School, Eindhoven University of Technology, Eindhoven, The Netherlands.

- Wagner, D., (1997). Analysis of mean values of a multi-server model with non-preemptive priorities and non-renewal input. *Stochastic Models*, 13(1), pp. 67-84.

- Waller, A.A.W., (1994). A queueing network model for field service support systems. *Omega*, 22(1), pp. 35-40.

- Watson, E.F., Chawda, P.P., McCarthy, B., Drevna, M.J., and Sadowski, R.P., (1998). A Simulation Metamodel for Response-Time Planning. *Decision Sciences*, 29(1), pp. 217-241.

- Wieczorek, A., Bušić, A., and Hyon, E., (2011). Critical level policies in lost sales inventory systems with different demand classes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6977 LNCS, pp. 204-218.

- Williams, T., (1980). Nonpreemptive multi-server priority queues. *Journal of the Operational Research Society*, pp. 1105 – 1107.

- Winston, W.L., (2004). *Operations research: applications and algorithms*, 4th edition. Brooks/Cole – Thomson Learning, Belmont, USA.

- Wong, H., Van Houtum, G., Cattrysse, D., Oudheusden, D., (2006). Multi-item spare parts systems with lateral transshipments and waiting time constraints. *European Journal of Operational Research,* 171(3), pp. 1071–1093.

- Wong, H., Kranenburg, B., van Houtum, G.J., and Cattrysse, D., (2007a). Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse. *OR spectrum*, 29, pp. 699-722.

- Wong, H., Van Oudheusden, D., and Cattrysse, D., (2007b). Two-echelon multi- item spare parts systems with emergency supply flexibility and waiting time constraints. *IIE Transactions,* 39, pp. 1045 – 1057.

- Wu, M.-., Hsu, Y.-. and Huang, L.-., (2011), An integrated approach to the design and operation for spare parts logistic systems, *Expert Systems with Applications,* 38(4), pp. 2990-2997.

- Zeltyn, S., Feldman, Z., and Wasserkrug. S., (2009). Waiting and sojourn times in a multi-server queue with mixed priorities. *Queueing Systems*, 61(4), pp. 305-328.

- Zhou, W., Lee, C.Y., and Wu, D., (2011). Optimal control of a capacitated inventory system with multiple demand classes. *Naval Research Logistics*, 58(1), pp. 43-58.

- Zhou, Y., and Zhao, X., (2010a). A two-demand-class inventory system with lost-sales and backorders. *Operations Research Letters*, 38(4), pp. 261-266.

- Zhou, Y., and Zhao, X., (2010b). Optimal policies of an inventory system with multiple demand classes. *Tsinghua Science and Technology*, 15(5), pp. 498-508.

Acknowledgements

Nearly five years ago, I was asked whether I would like to start a PhD research project. At the time, I chose to do so, foremost because I wanted to challenge myself and also because I would otherwise keep wondering whether I would have had the capability to complete such a project successfully. Although it has been a great challenge indeed, I am still satisfied with my choice, as it has given me the opportunity to work on a variety of interesting problems with people from a broad range of disciplines. Also, these four years have taught me a great deal about myself and the things I value, both in my working life and private life.

Still, I could not have completed this dissertation successfully without the help and support of a great number of people. First of all, I am very grateful to Matthieu as my daily supervisor. Throughout these years, I could always count on him to free up time to discuss any number of research points with me when I needed it. I also appreciate the extensive and detailed comments on my papers and dissertation chapters, even though the 'Track changes' function ensured that I could never quite see my own text back afterwards. I am also very grateful to Henk as my promotor for giving thorough feedback on my papers and dissertation chapters, in particular his ability to pinpoint those errors and unclear texts that would go unnoticed by everyone else. I am also very grateful to all members in the PhD committee, both for reading my dissertation and subsequently discussing various aspects of this dissertation with me during my defense, and to the members of the Service Logistics Forum Research (SLFR), both for funding part of my research and for providing me with a platform to present my work.

One of the things I value greatly in my working life is a pleasant working environment. Such an environment has been readily provided these past years by my colleagues at IEBIS. I am grateful to all of you for the pleasant coffee breaks, lunches and occasional dinners, during which a variety of (sometimes peculiar) topics have been discussed. I am especially grateful to Marco, Ahmad and Ingrid, who have contributed to chapters in this dissertation. Marco, you were the one who suggested that I should consider a PhD project. You were also briefly my second supervisor at the start of the project. I like the fact that we have supervised a Master project together, ultimately resulting in Chapter 2 of this dissertation. I also appreciate the many hours you spent explaining issues pertaining to Delphi and CPLEX to me. Ahmad, I highly appreciate our collaboration, which resulted in Chapter 7 of this dissertation. I especially value the time you took to (repeatedly) explain various concepts of queuing theory to me and your drive to continuously improve our work. Ingrid, our collaboration resulted in Chapter 4 of this dissertation. I really appreciate the enthusiasm that you showed in tackling the challenges of the related research. I am also very grateful for your willingness to discuss various topics with me, both research-related and otherwise. I still hope that the day will come when you attend one of my presentations.

I am further grateful for a wonderful group of friends and family who have continuously showed an interest in my work, and have provided me with welcome distractions in the form of tea breaks, shopping trips, cheese fondues, rehearsals and recordings with Ola Caribense, and holidays. Elisabeth, I am particularly grateful that you could always sympathize, as you were in a similar position a few months ago. Niki and Dito, you were always willing to listen to my frustrations, and you made the greatest effort to understand my models in detail. Finally, Roan, I truly appreciate your patience and support all these years, and I look forward to see what new adventures the future holds for both of us.

Samenvatting

Voor kapitaalintensieve bedrijven en organisaties zijn de continue beschikbaarheid en het ongestoord functioneren van productiefaciliteiten en systemen van cruciaal belang. Als voorbeelden valt te denken aan medische apparatuur, defensiematerieel, vliegtuigen en hoogwaardige productiesystemen. Een storing aan deze systemen kan leiden tot een fors productie- en inkomstenverlies of tot het ontstaan van onveilige situaties. Goed onderhoud en het snel verhelpen van eventuele storingen van deze systemen zijn dus van groot belang. De gebruikers van deze systemen besteden het onderhoud in toenemende mate uit aan de systeemfabrikant, waarbij afspraken over de te leveren diensten worden vastgelegd in een servicecontract. Een dergelijk contract bevat vaak zogenaamde *service level agreements* die het gewenste serviceniveau aangeven. Voorbeelden hiervan zijn een minimale beschikbaarheid van het systeem over een bepaald tijdsinterval, een maximale reactietijd in het geval dat een storing optreedt, en een maximale tijd tussen melding en oplossing van de storing.

Als er een storing optreedt, wijst de systeemfabrikant (of een extern onderhoudsbedrijf) een service engineer toe aan de klant om de storing te verhelpen. Vaak maakt de engineer hiervoor gebruik van reserve-onderdelen, waarbij onderhoud plaatsvindt door een defect onderdeel in het systeem te vervangen door een nieuw onderdeel (*repair by replacement*). Het defecte onderdeel kan zelf vaak worden gerepareerd door een defect subonderdeel te vervangen, enzovoorts. Systemen hebben doorgaans een dergelijke meerlaagse productstructuur, in de literatuur ook wel bekend als een *multi-indenture structuur*. Om tijdig een defect onderdeel te kunnen vervangen, beschikt de fabrikant vaak over een netwerk van voorraadpunten waar reserveonderdelen bewaard kunnen worden. Om schaalvoordelen binnen het voorraadbeheer ("risk pooling") tot stand te brengen is een centrale voorraadlocatie nuttig, terwijl anderzijds verschillende lokale magazijnen vlakbij de gebruikers nuttig zijn om korte levertijden te kunnen realiseren. Dit leidt doorgaans tot een combinatie van centrale en lokale voorraadpunten, waarbij centrale magazijnen vooral dure langzaamlopers op voorraad houden en lokale magazijnen vooral goedkope snellopers. De voorraden bij de lokale magazijnen wordt dan aangevuld vanuit een centraal magazijn. We spreken in dat geval van een *multi-echelon structuur*. Een aantal voorraadpunten wordt ook gebruikt om (vooral dure) defecte onderdelen te repareren. De beschikbaarheid van engineers en onderdelen bepaalt in hoge mate de tijd die nodig is om een storing te verhelpen, en dus de kans om het afgesproken serviceniveau te halen. Natuurlijk zijn er kosten verbonden aan alle mensen en middelen die nodig zijn voor de instandhouding (service engineers, voorraadpunten, voorraden, reparatie, transport, etc.).

In dit proefschrift onderzoeken wij verschillende logistieke opties om te kunnen voldoen aan prestatieafspraken die in servicecontracten zijn gemaakt. Hierbij richten wij ons met name op situaties waarin het gewenste serviceniveau per klant verschilt: de éne groep klanten kan behoefte hebben aan een duur servicecontract dat een hoog serviceniveau biedt, terwijl andere klanten tevreden zijn met een goedkoper contract en een lager serviceniveau. Op dit moment bestaan er nog weinig goede methoden om met een dergelijke differentiatie in serviceniveaus om te gaan: het leveren van een uniforme hoge service aan alle klanten is duur en biedt standaardklanten een te hoge service, mogelijk ten koste van klanten die juist hogere eisen stellen. Differentiatie kan in theorie wel tot stand worden gebracht door voorraad te reserveren voor hoge prioriteitsklanten (de zogenaamde *critical level policy*). Helaas blijkt deze strategie in de praktijk slecht werkbaar te zijn, onder andere omdat service engineers niet bereid zijn om onderhoud uit te stellen voor een standaardklant als een onderdeel feitelijk beschikbaar is. Daarnaast vinden leveranciers het doorgaans lastig om aan klanten te moeten melden dat een benodigd onderdeel weliswaar op voorraad is, maar dat ze dat onderdeel op basis van een relatief goedkoop contract niet willen uitleveren.

Het onderzoek in dit proefschrift is verdeeld in drie delen. Het eerste deel, beschreven in hoofdstuk 2, betreft de beschikbaarheid van onderdelen waarbij differentiatie alleen op itemniveau wordt toegepast. Het tweede deel, beschreven in hoofdstukken 3 tot en met 6, betreft de beschikbaarheid van onderdelen met differentiatie op zowel item- als klantniveau. Het laatste deel, beschreven in hoofdstuk 7, betreft de beschikbaarheid van service engineers waar een prioriteitsmechanisme gebruikt wordt bij het toewijzen van engineers aan klanten.

In hoofdstuk 2 onderzoeken wij differentiatie op itemniveau bij het leveren van onderdelen. Wij beschouwen een multi-indenture multi-echelon model waarin zowel de voorraadniveaus op verschillende locaties als de doorlooptijden voor het repareren en vervoeren van onderdelen beslisvariabelen zijn. In een theoretisch experiment laten wij zien dat de besparingen van dit model gemiddeld 20% kunnen zijn ten opzichte van een standaard VARI-METRIC model waarin de doorlooptijden vast staan. Verder vonden wij een besparing van 5.6% in een case study bij Thales Nederland, waar de mogelijkheden voor doorlooptijdverkorting beperkt waren.

In de hoofdstukken 3 tot en met 6 onderzoeken wij differentiatie op zowel item- als klantniveau bij het leveren van onderdelen. Wij richten ons op drie differentiatietechnieken:

- **Het selectief gebruik van spoedleveringen:** Doorgaans bestaat het netwerk om voorraden van onderdelen aan te houden uit een centraal magazijn en meerdere lokale magazijnen die elk hun eigen verzorgingsgebied (regio) hebben. Een spoedlevering houdt in dat een klant snel (en duurder) wordt beleverd vanuit een centraal voorraadpunt (hetzij het centrale magazijn of de producent van het onderdeel) als het “eigen” lokale magazijn buiten voorraad is.

- **Het selectief gebruik van laterale leveringen:** Net als spoedleveringen, worden laterale leveringen alleen gebruikt als een klant in een bepaalde regio niet vanuit zijn eigen lokale magazijn bediend kan worden. Bij een laterale levering wordt de klant echter beleverd vanuit een *lokaal magazijn* van een naburige regio. Vaak is een laterale levering zowel sneller als goedkoper dan een spoedlevering. Deze levering onttrekt echter wel voorraad aan het naburige magazijn, waardoor dit laatste magazijn vervolgens wellicht niet aan alle klantvraag in de eigen regio kan voldoen.
- **Het aanhouden van (deel)voorraden bij klanten:** Een fabrikant kan er ook voor kiezen om van bepaalde snellopende items wat voorraad bij de klant neer te leggen, zodat storingen waarvoor deze onderdelen nodig zijn extra snel kunnen worden verholpen.

Wij analyseren zowel de toegevoegde waarde van de afzonderlijke methoden, als die van combinaties van meerdere methoden. Daarbij beogen wij een zo goed mogelijke afweging te maken tussen de te bereiken serviceniveaus en de relevante kosten.

Het selectief gebruik van spoedleveringen wordt behandeld in hoofdstuk 3. In dat hoofdstuk analyseren wij een model waarin een fabrikant de mogelijkheid heeft om tot een spoedlevering over te gaan als het lokale magazijn geen voorraad heeft. We veronderstellen daarbij dat het onderdeel altijd vanuit een centraal magazijn of door een externe leverancier te leveren is, zij het tegen (zeer) hoge kosten. Het nut van deze optie hangt zowel af van het type onderdeel dat gevraagd wordt als van het type klant. Met dit model vinden wij een besparing van gemiddeld 4.4% ten opzichte van de situatie waarin geen differentiatie gebruikt wordt. Ook zien wij dat de combinatie van selectieve spoedleveringen en voorraadreservering (critical level policies) grote toegevoegde waarde heeft, met een gemiddelde kostenbesparing van 14%. De combinatie van de twee differentiatieopties blijkt een groter effect te hebben dan de som der delen. In hoofdstuk 4 breiden wij het spoedleveringsmodel uit met de mogelijkheid voor laterale leveringen van nabijgelegen magazijnen, waarbij deze laterale leveringen alleen voor hoge-prioriteitsklanten gebruikt mogen worden. Het toevoegen van laterale leveringen levert een extra kostenbesparing van gemiddeld 14% op ten opzichte van het model waar alleen selectieve spoedleveringen voor differentiatie gebruikt worden. Verdere uitbreiding van het model met de critical level policy leverde geen significante extra besparing op.

In hoofdstuk 6 onderzoeken wij de mogelijkheid om voorraden op locatie bij de klanten als differentiatietechniek te gebruiken. Wij analyseren daar een multi-item model met één magazijn en meerdere klanten, zowel voor het geval dat alle vraag wordt nageleverd indien er geen voorraad is, als voor het geval dat een spoedlevering gebruikt wordt wanneer zowel de klant als het magazijn buiten voorraad zijn en er geen onderdelen reeds onderweg zijn van het magazijn naar de klant toe. In het laatste geval is voor de analyse van het systeem een twee-echelon model nodig dat nog niet in de literatuur was geanalyseerd. Wij beschrijven dit model en de bijbehorende analyse in hoofdstuk 5. Vergeleken met de setting waarin alle klanten uniforme service krijgen levert het gebruik van klantvoorraden een gemiddelde besparing op van 13% voor het geval dat alle vraag wordt nageleverd en 5% voor de situatie waarin spoedleveringen zijn toegestaan. Het combineren van klantvoorraden met de critical level policy levert verder geen significante extra besparing op.

In hoofdstuk 7 onderzoeken wij tenslotte het gebruik van een prioriteitsmechanisme bij het toewijzen van service engineers aan klanten. We beschouwen een model met meerdere klanttypen waarin een beschikbare engineer steeds aan de klant met de hoogste prioriteit toegewezen wordt. Voor dit model ontwikkelen wij methoden waarmee we nauwkeurig en snel de kansverdeling van de wachttijd kunnen bepalen per klanttype. Deze methoden passen wij vervolgens toe op een case study bij een fabrikant van kopieer- en printsystemen om de prestaties op bepaalde serviceniveaus te kunnen toetsen.

About the author

Elisa Alvarez was born in Willemstad, Curaçao, on August 31st , 1984. In 2002, she completed her secondary school education at the Radulphus college in Curaçao, after which she moved to the Netherlands to study Industrial Engineering and Management at the University of Twente. She obtained her Master’s degree in 2008, after performing a Master’s assignment at ORTEC on the development of a generic scheduling approach for a production scheduling system.

In 2007, Elisa performed a Capita Selecta assignment for the Service Logistics Forum Research (SLFR) to investigate what challenges companies experience when developing service contracts and subsequently striving to meet the service level agreements stated in these contracts. The results of this assignment prompted Elisa to start a PhD research project in 2008 with the aim to investigate various options for applying differentiation in after-sales services. Elisa executed this project in the department of Industrial Engineering and Business Information Systems at the University of Twente, under the supervision of prof. dr. W.H.M. Zijm and dr. M.C. van der Heijden. The results of her studies are presented in this dissertation.