Multi-echelon inventory systems with shipment consolidation

methods for exact analysis and optimization

Malmberg, Filip

2025

Multi-Echelon Inventory Systems with Shipment Consolidation

Methods for Exact Analysis and Optimization

FILIP MALMBERG
PRODUCTION MANAGEMENT | FACULTY OF ENGINEERING | LUND UNIVERSITY

Multi-Echelon Inventory Systems with Shipment Consolidation

Multi-Echelon Inventory Systems with Shipment Consolidation

Methods for Exact Analysis and Optimization

by Filip Malmberg

Thesis for the degree of Doctor of Philosophy

Thesis advisors:
Prof. Johan Marklund, Prof. Peter Berling, Assoc. Prof. Fredrik Olsson

Faculty opponent:
Assoc. Prof. Willem van Jaarsveld,
Eindhoven University of Technology

To be presented, with the permission of the Faculty of Engineering at Lund University, for public criticism at LTH (Room M:E, floor 2, M-Huset) on Friday, the 26th of September 2025 at 10:15.

Organization
LUND UNIVERSITY

Dept. of Industrial and Mechanical Sciences
Division of Production Management
Box 118
SE–221 00 Lund, Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2025-09-26

Sponsoring organization
Formas - a Swedish Research Council for Sustainable Development, project no. 2020-02121

Author(s)
Filip Malmberg

Title and subtitle
Multi-Echelon Inventory Systems with Shipment Consolidation - Methods for Exact Analysis and Optimization

Abstract
Efforts to reduce greenhouse gas emissions, globalized markets with intense competition, and generally rising transportation costs emphasize the importance of achieving economically and environmentally sustainable distribution systems. Many international organizations predict that if no new measures are implemented, global carbon dioxide emissions from freight transport will increase to even higher levels. We cannot afford to wait until new green technology solves the problem for us; we need to start doing what we can with what we have. Increasing the load factor of transport vehicles through freight consolidation, better transport planning, and coordination within the supply chain are therefore important measures in the near term. It is estimated that such improved planning and management could reduce emissions from freight transport by up to one-third.

Thus, there is a significant potential to utilize transport vehicles more effectively and increase load factors through consolidation and joint freight transport, thereby reducing transportation costs and the environmental footprint. At the same time, increased freight consolidation leads to longer lead times and, consequently, higher inventory holding costs to maintain the same customer service levels. To achieve economically and environmentally sustainable solutions, it is therefore essential to have quantitative methods that can balance inventory and shipment consolidation decisions against each other.

This thesis highlights the importance of such methods for achieving more sustainable and cost-effective distribution systems. The developed methods set the stage for practitioners to evaluate alternative replenishment and shipment policies, revealing insights for balancing cost, service, and environmental objectives in multi-echelon inventory and distribution systems under uncertain demand.

In a series of four scientific papers, we analyze centralized One-Warehouse-Multiple-Retailer (OWMR) systems in which non-identical retailers face stochastic demand. For each study, we develop exact analytical methods to obtain the steady-state probability distributions of inventory levels under continuous review replenishment policies, combined with different shipment consolidation strategies. These probabilities are used to evaluate and optimize expected costs, fill rates, and transport-related emissions.

Key words
inventory, inventory management, inventory control, multi-echelon, one-warehouse-multiple-retailer systems, shipment consolidation, delivery policy, stochastic, probability

ISRN
LUTMDN/TMIO-1015-SE

Language
English

ISBN
978-91-8104-614-4 (print)
978-91-8104-615-1 (pdf)

Number of pages
242

Signature ______________________ Date 2025-08-18

Multi-Echelon Inventory Systems with Shipment Consolidation

Methods for Exact Analysis and Optimization

by Filip Malmberg

LUND
UNIVERSITY

Funding information: This research was carried out with financial support from Formas - a Swedish Research Council for Sustainable Development, project no. 2020-02121

Faculty of Engineering, Dept. of Industrial and Mechanical Sciences
Division of Production Management

ISBN: 978-91-8104-614-4 (print)
ISBN: 978-91-8104-615-1 (pdf)
ISRN: LUTMDN/TMIO-1015-SE

Printed in Sweden by Media-Tryck, Lund University, Lund 2025

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

Dedicated to
Setelliten

Contents

List of publications

This thesis is based on the following research papers, referred to by their Roman numerals:

I **Evaluation and Control of Inventory Distribution Systems with Quantity Based Shipment Consolidation**
F. Malmberg, J. Marklund
Naval Research Logistics (NRL), 2023, 70(2), 205-227.
`https://doi.org/10.1002/nav.22090`

II **Exact Analysis of One-Warehouse-Multiple-Retailer Inventory Systems with Quantity Restricted Deliveries**
J. Andersson, **F. Malmberg**, J. Marklund
European Journal of Operational Research (EJOR), 2023, 309(3), 1161-1172
`https://doi.org/10.1016/j.ejor.2023.02.026`

III **Exact Analysis and Control of OWMR Inventory Systems with Joint Order Quantities and Quantity Based Shipment Consolidation**
F. Malmberg, J. Marklund
(Submitted 2025)

IV **Managing Inventories in Sustainable Multi-Echelon Distribution Systems with Hybrid Shipment Consolidation**
F. Malmberg, J. Ralfs, J. Marklund, G. Kiesmüller
(Submitted 2025)

All papers are reproduced with the permission of their respective publishers.

Related publications

1 **Quantity-Based Shipment Consolidation and Delivery Policies in Multi-Echelon Inventory Control**
F. Malmberg
Department of Industrial Management and Logistics
Faculty of Engineering LTH, Lund University
(2022) Thesis for the degree of Licentiate in Engineering

Abstract

Efforts to reduce greenhouse gas emissions, globalized markets with intense competition, and generally rising transportation costs emphasize the importance of achieving economically and environmentally sustainable distribution systems. Many international organizations predict that if no new measures are implemented, global carbon dioxide emissions from freight transport will increase to even higher levels. We cannot afford to wait until new green technology solves the problem for us; we need to start doing what we can with what we have. Increasing the load factor of transport vehicles through freight consolidation, better transport planning, and coordination within the supply chain are therefore important measures in the near term. It is estimated that such improved planning and management could reduce emissions from freight transport by up to one-third.

Thus, there is a significant potential to utilize transport vehicles more effectively and increase load factors through consolidation and joint freight transport, thereby reducing transportation costs and the environmental footprint. At the same time, increased freight consolidation leads to longer lead times and, consequently, higher inventory holding costs to maintain the same customer service levels. To achieve economically and environmentally sustainable solutions, it is therefore essential to have quantitative methods that can balance inventory and shipment consolidation decisions against each other.

This thesis highlights the importance of such methods for achieving more sustainable and cost-effective distribution systems. The developed methods set the stage for practitioners to evaluate alternative replenishment and shipment policies, revealing insights for balancing cost, service, and environmental objectives in multi-echelon inventory and distribution systems under uncertain demand.

In a series of four scientific papers, we analyze centralized One-Warehouse-Multiple-Retailer (OWMR) systems in which non-identical retailers face stochastic demand. For each study, we develop exact analytical methods to obtain the steady-state probability distributions of inventory levels under continuous review replenishment policies, combined with different shipment consolidation strategies. These probabilities are used to evaluate and optimize expected costs, fill rates, and transport-related emissions.

Papers I and III consider systems with quantity-based shipment consolidation to groups of retailers facing Poisson demand processes. We derive exact recursive approaches to determine the inventory level distributions at the retailers and show how to optimize reorder levels and shipment quantities. It is demonstrated that accounting for fixed handling and transportation costs through shipment consolidation is key to reducing overall costs and emissions. The extension in Paper III allows groups of retailers to synchronize their replenishment requests when ordering from the central warehouse. By permitting joint orders from groups of retailers, the warehouse can consolidate shipments with less stock-on-hand. Our analysis shows that this coordinated strategy can significantly outperform the approach from Paper I with base-stock ordering at the retailers, even in the absence of fixed ordering costs. The policy in Paper III includes the Policy in Paper I as a special case and is therefore guaranteed to perform at least as well when optimized.

Paper II analyzes quantity-restricted deliveries motivated by a desire to deliver goods in full packages, pallets, or other load carriers. The retailers face compound Poisson demand, and an exact analysis of the inventory level distributions is derived. The study shows that ignoring these quantitative delivery restrictions can lead to high backorder costs and poor service levels. Optimizing the reorder points while considering delivery restrictions ensures significantly better performance.

Finally, Paper IV studies a hybrid (time-and-quantity-based) shipment consolidation policy that combines periodic transportation schedules with opportunities for intermediate shipments of full load carriers. Notably, the pure time-based and pure quantity-based strategies are both contained within the hybrid policy as special cases. Consequently, when optimized, the hybrid policy offers a performance guarantee over these alternatives. Numerical experiments indicate that substantial emission reductions can be achieved at low cost.

Acknowledgements

First, I would like to express my sincere gratitude to my main supervisor, Professor Johan Marklund, for your continuous support and guidance throughout my Ph.D. Your open door, willingness to discuss both broad research questions and the deepest technical details, and your generous advice have taught me a great deal, for which I am truly grateful.

I want to thank Professor Gudrun Kiesmüller and Jana Ralfs for hosting me at the Technical University of Munich on campus Heilbronn. Our collaboration led to the final paper of this thesis. Thank you both for your warm welcome—and in particular, thank you, Jana, for the extra motivation you brought to our work.

My colleagues at Systecon made it possible to combine consultancy work with doctoral studies. I especially acknowledge Pär Sandin, a valued colleague and friend, whose support has helped me grow as both a researcher and consultant.

I thank all current and former colleagues and fellow Ph.D. students at Production Management and Engineering Logistics—lunches and coffee breaks with you made this journey more enjoyable. I am especially grateful to Ludwig for our side-project collaboration, to Lina for helping me settle into the field of Inventory Management, to Danja for our joint teaching efforts, and to my assistant supervisors, Professor Peter Berling and Associate Professor Fredrik Olsson. I would also like to acknowledge the valuable support given by the past and present administrative team at the division and department. Thank you all for your friendship and support.

I am deeply grateful to my family — my parents, my brother, and his family. To Emma, thank you for your love, patience, and countless ways of making difficult times easier. Finally, to our son, Set — your impending arrival spurred me to finish, though you came two weeks ahead of schedule and transformed my "final two weeks of writing" into nearly a year. It was, without question, the best delay imaginable. In hindsight, I should have known better than to ignore stochastic lead times.

Filip Malmberg, August 2025

Populärvetenskaplig sammanfattning

Klimatavtrycket från logistik och godstransporter är en stor utmaning för vår och kommande generationer. Idag står transportsektorn för en stor andel av de globala utsläppen, och rapporter visar att många transportmedel är ineffektivt utnyttjade. Forskningen i denna avhandling visar att gemensam optimering av lagerstyrning och transporter kan leda till både lägre kostnader och minskade utsläpp bl a genom att öka fyllnadsgraden i de transporter som görs inom försörjningskedjor och distributionssystem. I den här forskningen har vi utvecklat nya matematiska metoder för att analysera och ta fram beslutsundrag för distributionssystem där ett centralt lager försörjer ett antal andra lokala lager, t.ex., återförsäljare, regionala lager, verkstäder eller depåer.

Ökat utnyttjande av transportkapaciteten är i flera europeiska och internationella rapporter omnämnt som en snabbt och effektivt sätt att uppnå bättre miljömässig hållbarhet. Att öka kapacitetsutnyttjandet av transportfordon, exempelvis genom att konsolidera leveranser till flera återförsäljare och/eller från flera ordrar, tenderar dock att resultera i längre och mer varierande ledtider. Detta leder i sin tur till högre lagerrelaterade kostnader. Att balansera kostnader och utsläpp från transporter, med kostnader som är förknippade med lagerhållning, bristnoteringar och restordrar, är därför en central fråga för att uppnå högre kapacitetsutnyttjande.

Den traditionella lagerstyrningslitteraturen har fokuserat på lagerbeslut utan att ta hänsyn till effekterna av olika skeppnings- och transportstrategier. Detta har varit rationellt i en kontext av relativt billiga transporter och en norm av snabba och flexibla leveranser. I de flesta modeller för lagerstyrning antas därför att varje beställd enhet skeppas så snart den finns tillgänglig, vilket om det realiseras i praktiken, kan innebära att det finns en stor potential för att öka fyllnadsgraden för varje transport genom konsolidering och samlastning.

Till skillnad från traditionella metoder gör denna avhandling mer detaljerade modelleringsavväganden för transporterna mellan lagerhållningsplatser för att på så vis möjliggöra billigare och mindre klimatpåverkande totallösningar där både lagerhållning och transporter beaktas.

Lagerstyrning och skeppningskonsolidering i flernivålagersystem, där exempelvis skeppningar från ett centrallager konsolideras till flera återförsäljare/ depåer/siter, är ett område som analyserats i ett antal tidigare publicerade artiklar. Hittills har denna forskning fokuserat på så kallad tidsbaserad skeppningskonsolidering där transporterna avgår enligt ett fast schema, t.ex., dagliga eller veckovisa transporter. I vår forskning introducerar vi, för första gången, möjligheten att analysera kvantitetsbaserad konsolidering i den här typen av lagerstyrningsmodeller. Kvantitetsbaserad skeppningskonsolidering innebär att transporter sker i fasta kvantiteter, som exempelvis kan motsvara en full lastbil, container, eller annan typ av lastbärare. Konsolideringen kan ske över flera ordrar och/eller destinationer. Genom att matcha kvantiteternas storlek mot transportkapaciteten ges möjlighet att nå högre utnyttjandegraden av transporter. I dag utnyttjas bara i genomsnitt 57 % av tillgänglig kapacitet för transporter som bär gods. Denna siffra inkluderar således inte transporter som går tomma för att frakta tillbaka fordon eller containrar. Det finns alltså stor potential i att reducera utsläpp genom att öka fyllnadsgraden i transporterna. Avhandlingen innehåller tre olika modeller för kvantitetsbaserad skeppningskonsolidering som tar hänsyn till olika typer av system och baseras på olika modellantaganden.

Utöver kvantitetsbaserad skeppningskonsolidering introduceras även matematiska modeller för så kallad hybrid skeppningskonsolidering där tidsbaserad och kvantitetsbaserad konsolidering kombineras. Ett sådant system har tidsbaserade transporter i grunden, t.ex. dagliga eller veckovisa leveranser, men om en full skeppningskvantitet efterfrågas (och finns tillgänglig) tidigare kan den skickas direkt. De tidsbaserade kvantiteterna kommer inte vara “fulla” kvantiteter, men regulariteten i tid, gör att dessa enklare kan koordineras med andra transporter och/eller lastas om för vidare transport via exempelvis ett konsolidationscenter.

Trots den relativt låga utnyttjandegraden av transporter görs i praktiken redan en hel del avväganden för transporter och skeppningar i moderna logistiksystem. Ett problem uppstår dock när lagerhållningsbesluten inte tar hänsyn till de beslut som fattas för transporter. I avhandlingen visar vi exempelvis att optimering av lagerhållningsbeslut och lagernivåer utan att ta hänsyn till hur transporter genomförs kan få stora konsekvenser för systemet som helhet. Transportbeslut kan påverka den verkliga ledtiden mellan beställning och leverans. Om ändringen i ledtid underskattas vid beslutsfattande blir konsekvensen alltför låga säkerhetslager och där med höga bristkostnader.

I den här avhandlingen utvecklas matematiska metoder som kan väga lager- och skeppningskonsolideringsbeslut mot varandra för att hitta en optimal lösning för hela distributionssystem, både avseende kostnader och miljöpåverkan. Metoder och modeller som är härledda i den här avhandlingen kan hjälpa företag och organisationer med utformning av strategier och försörjningskedjor samt lägga grunden till verktyg för operationell styrning av ekonomiskt och miljömässigt hållbara lager- och distributionssystem med osäker efterfrågan. Syftet är att stödja företag och försörjningskedjor i att kostnadseffektivt minska sina transportrelaterade koldioxidutsläpp, vilket i sin tur bidrar till att uppnå de globala målen för Hållbar konsumtion och produktion samt Bekämpa klimatförändringarna.

Multi-Echelon Inventory Systems with Shipment Consolidation

Chapter 1

Introduction

For over a century, the use of applied mathematics in inventory management has played a significant role in shaping tools and methods for improving supply chains and industrial processes. At the beginning of the 20th century, batch manufacturing with relatively high setup costs motivated many companies to investigate the most economical size for a production lot. The idea was to balance the economies of scale from batch orders with the increased costs for holding stock until the full replenished quantity has been demanded and arrives. Although a few scholars seem to have arrived at similar results at about the same time, it is today generally accepted that the first to publish a mathematical formula for this problem was Harris (1913). The Economic Order Quantity (EOQ) formula he derived has been widely used throughout the years, and it is still an important concept in courses and textbooks on logistics, production, and inventory management.

Although there are a few similar examples of early work, it was not until the Second World War that the use of mathematical and statistical methods to improve or optimize strategic and operational decision-making really took off. During these difficult times, teams of scientists were formed in the UK and the US to perform *research on operations* that would result in mathematical models that could be used to gain a strategic advantage in the war (Hillier and Lieberman, 2000). The term Operations Research (OR) was coined to describe an interdisciplinary research field that focuses on using mathematical models, statistical methods, and other analytical techniques to solve complex decision-making problems.

Following the war, there was a growing interest in OR within industrial and non-military applications, and an academic discipline of Operations Research emerged, originating from research groups established by military figures in the US and England (Morse et al., 1951, 2003). At this time, the academic field of OR evolved rapidly, and many methods and a body of literature were developed that are significant to this day.

The interest in multi-echelon distribution systems that began in the late 1950s and early 1960s is of particular interest for this thesis. Achieving centralized control in these systems may take advantage of a holistic view that coordinates multiple stock-keeping locations together. Notably, analyzing stochastic multi-echelon systems is complex even for stylized modeling assumptions. In most cases, providing optimal policies for inventory control is seen to be intractable. Consequently, researchers have focused on evaluating systems and optimizing parameters for specific replenishment policies of practical relevance. Throughout the years, advancements within this research have been translated into practical applications, leading to the development of commercial multi-echelon inventory optimization tools adopted by multinational companies (de Kok et al., 2018).

More recent research has expanded the field by relaxing some of the policy restrictions in early work by allowing, e.g., expedited emergency shipments, transshipment between stock locations, and inventory pooling, particularly in the context of spare parts inventory management where demand is uncertain (de Kok et al., 2018). These advancements have re-highlighted the practical significance of multi-echelon systems in achieving cost-efficient supply chains.

Today, we face a new crisis and global threat from global warming and the environmental impact caused by humanity. This seems to be one of the great challenges for ourselves and future generations, and the crisis impacts ecosystems, economies, and societies on a massive scale. It is therefore important that researchers from all fields come together to assist in increasing the efficiency of resource usage, reducing emissions, and reducing the environmental footprint from human activities.

Achieving sustainable *green* inventory management requires balancing the traditional economic objectives, such as minimizing costs and maintaining service levels, with an environmental perspective. According to estimates from the World Economic Forum (Doherty and Hoyle, 2009), approximately 90% of emissions in logistics and transport activities are attributed to transportation, with the remaining 10% arising from facility and building operations (Marklund and Berling, 2024). However, focusing solely on reducing transport emissions can negatively impact the system as a whole. Most actions for greener transport affect transportation lead times and shipment quantities. Consequently, maintaining the same service levels while reducing transport emissions requires increased inventory levels, leading to higher holding costs. In particular, consolidating shipments to reduce shipment frequency and increase transport utilization also leads to changes in the lead time. Thus, if inventories are not adjusted accordingly, the result will be lower service levels and/or unnecessarily high costs (Marklund and Berling, 2024). These are trade-offs that highlight the complexity of green inventory management.

This thesis focuses specifically on inventory distribution systems and how they can be more efficiently managed in these respects. Therefore, we do not explicitly consider the environmental impacts of raw materials or production processes, the sustainability implications of shortages and backorders, the use-phase and consumption of the goods handled, or the phase-out or recycling of goods.

When studying single-echelon inventory systems in isolation, the order quantity from upstream supply is generally directly translated into the amount shipped. This is because potential shortages, which could lead to the need for a partial delivery, do not occur unless upstream activities are explicitly modeled. Therefore, in such models, it is sufficient to adjust the replenishment policies to achieve higher load factors in the replenishment transports. Hence, in the single-echelon literature, where the replenishment orders are directly reflected in the shipments, sustainability aspects are often included as modifications of the long-known EOQ-formula (Bouchery et al., 2012) to achieve higher load factors or transport utilization.

However, as transport and inventory decisions are interlinked, it makes sense to model stock-keeping locations both before and after the transports (i.e., multi-echelon), enabling balancing less flexible transport options with more upstream or downstream inventory. When explicitly modeling a warehouse supplying downstream locations, the risk of shortages makes order policies insufficient for analyzing physical material flows in the system. As noted above, what is available for shipment does not necessarily equal what is ordered, and thus, the risk of shortage introduces a need to consider dispatch policies and replenishment policies separately. As a consequence, in multi-echelon distribution systems, we may leverage collaboration and technology for information sharing to optimize inventory placement and transportation decisions across the distribution system. Using a total cost **and** emissions perspective allows us to achieve cost-efficient emission reductions.

Notably, in practice, transports are often shipped with ample capacity. The World Economic Forum has estimated that within the European Union, the average load factor of vehicles carrying a load (i.e., excluding empty vehicle runs) was merely 57% (Marklund and Berling, 2024). In fact, the trend during the past decades has been decreasing load factors (EEA, 2021a). One reason for decreased capacity utilization is an increasing demand for quick and just-in-time deliveries in lean operations (Wehner, 2018). Notably, the European Environment Agency observes a considerable variation in the load factor among different segments and countries, suggesting that there indeed is a possibility to act for improved utilization in many organizations (EEA, 2021a).

Although a transition into greener transport modes, e.g., electrified transport, is an important measure, such a shift will require both the development and production of new technology, as well as large investments, before it can take effect. In the Nordic countries, the majority of freight transport is carried out on roads. For example, two-thirds of Swedish freight transport was carried out on roads in 2017 (Pinchasik et al., 2019), while the corresponding share in the rest of Europe is approximately fifty percent (EEA, 2021a). Emissions from these transports account for most of the CO_2 emissions from freight transport segments due to the high carbon content in fuels for trucks today. Although the COVID-19 pandemic decreased freight transport temporarily in 2020 (EEA, 2022b,a), the decrease in the European Union was relatively small, in particular for road-based transport (EEA, 2022a). The demand for freight transportation recovered quickly, and with the present vehicle fleets, the emissions of greenhouse gases will continue to increase as the demand for transport and freight services continues to grow, both in general and for road-based transport in particular (EEA, 2021b). The International Energy Agency, Teter et al. (2017), foresees that without any actions, the global truck-based freight logistics will increase emissions of greenhouse gases (GHG) by around 55% until 2050.

This is in direct contrast to the need for reducing emissions and the environmental footprint from operations, and the sustainability community agrees that we cannot afford to wait until technology solves the problem for us; we need to start doing what we can with what we have. Improved planning and control within the supply chain and taking a green perspective on operations management have an immediate potential to reduce the environmental footprint. According to an estimate from the OECD (International Transport Forum, 2015, 2017, 2023), such measures could achieve an overall GHG reduction from freight transport by up to 33%.

Similarly, the International Energy Agency presents an alternative, more optimistic scenario that requires near-term efforts across three key areas (Teter et al., 2017):

1. Improvements of logistics, enabled by data gathering and sharing, to realize some of the potential that underlies system-wide improvements
2. Fuel economy policies to increase the efficiency of trucks through standards and differentiated taxes
3. Support to the use of alternative fuels, such as through RD&D, and support to the build-up of infrastructure

In a scenario where all of the above measures are implemented, the demand for fuel from trucks could instead decrease by around 50% by 2050 with a corresponding reduction in emissions by up to 75%. In this scenario, improvements of logistics enabled by collaboration and data sharing, account for 42% of the fuel demand savings (Teter et al., 2017).

The OECD (International Transport Forum, 2015, 2017, 2023) highlights the importance of improving capacity utilization and minimizing empty transports, for example, by using consolidation centers, sharing resources like trucks and warehouses, and optimizing route planning. Consolidated shipments, where transport of multiple orders is made to multiple locations on a joint carrier, are a potential measure to increase transport utilization as smaller shipments can then share the load carrier with other shipments to the same or other locations (Jackson et al., 1994). The Nordic Council of Ministers lists more efficient use of transport as a key action for a lower environmental footprint from logistics and freight (Pinchasik et al., 2019). Marklund and Berling (2024) draws the same conclusion and emphasize the potential for shipment consolidation to achieve higher load factors. They point out that although a more efficient use of vehicles and carriers may result in fewer transports and thus both lower emissions and transportation costs, it will, in general, also generate longer lead times, less frequent dispatches, and/or less flexible shipments. Thus, shipment consolidation strategies and inventory management are intertwined.

The theoretical foundation for this thesis's research is retrieved from the field of Inventory Theory/Inventory Control within the Operations Research and Management Science area (OR/MS). The uncertainty that these systems face in terms of stochastic demand and supply means that the mathematical analysis needs to incorporate stochastic models and results from probability theory, queueing theory, and mathematical programming.

This thesis will focus on developing methods and models to manage inventory and shipment consolidation decisions across multiple orders and/or destination locations, specifically within a distribution system where a central warehouse serves multiple local warehouses, depots, or retailers, each facing independent stochastic demand processes. The relevance of studying this type of divergent distribution system (One-Warehouse-Multiple-Retailers system, OWMR) is motivated by its frequent occurrences in practice and broad applicability. Furthermore, existing research results and practical insights from integrating transport and inventory decisions remain limited, and there is a gap in the academic literature that can be addressed.

Related literature

Higginson and Bookbinder (1994) define three classes of shipment consolidation policies: time-based, quantity-based, and hybrid (combining time and quantity). These approaches have been studied extensively in single-echelon settings. In the context of multi-echelon distribution systems, the literature has primarily focused on time-based consolidation. Studies analyzing quantity-based shipment consolidation for groups of retailers or hybrid approaches combining time- and quantity-based consolidation remain scarce.

In this overview of related literature, we first review research on shipment consolidation in single-echelon contexts. We then address OWMR inventory systems with shipment consolidation and delivery restrictions, followed by literature on multi-echelon systems that do not explicitly consider shipments. Finally, we summarize the literature's positioning relative to this thesis in a dedicated table and outline its relationship to the author's licentiate thesis.

Single-echelon systems with shipment consolidation

As previously noted, a large body of literature addresses shipment consolidation in single-echelon systems. All three consolidation programs discussed by Higginson and Bookbinder (1994) are encompassed in this research stream.

For example, Çetinkaya and Lee (2000) approximate near-optimal decision parameters for time-based shipment consolidation policies, while Axsäter (2001) provide exact evaluation and analysis for a similar setting. Ralfs and Kiesmüller (2022) explicitly considers transport efficiency by including advance demand information in a single-echelon shipment consolidation model, aiming to increase truck utilization on scheduled shipment days.

Moreover, Çetinkaya and Bookbinder (2003), Chen et al. (2005), and Çetinkaya et al. (2006) investigate both time-based and quantity-based dispatch policies. In addition, Çetinkaya et al. (2006) introduce a hybrid policy in which policy parameters are approximated by combining optimal parameters from quantity-based and time-based policies, respectively.

Mutlu et al. (2010) and Wei et al. (2023) evaluate and compare the performance of time-based, quantity-based, and hybrid consolidation policies with respect to both inventory and transportation decisions. Joint consideration of inventory and shipment decisions in single-echelon systems is also analyzed in, e.g., Cheung and Lee (2002), Toptal et al. (2003), Çetinkaya et al. (2008), and Satır et al. (2018). A finding in the literature on shipment consolidation in single-echelon systems is that quantity-based policies usually outperform time-based policies, and hybrid policies regarding the expected total inventory and shipment costs when assuming the same shipment costs for both shipment options. In theory, quantity-based consolidation maximizes the utilization of the transportation capacity, thereby minimizing transportation costs and emissions. However, it may not necessarily be the most advantageous approach in a larger context when the total cost of a distribution system is considered. For instance, a third-party logistics (3PL) provider can potentially offer a time-based shipment schedule with lower fixed costs through proactive planning, coordination, and consolidation across multiple warehouses and organizations.

OWMR systems with shipment consolidation

Several papers consider time-based shipment consolidation policies and periodic shipments in multi-echelon systems. For example, Marklund (2011) provides exact cost evaluation and optimization of decision parameters for an OWMR system with continuous review and time-based shipment consolidation between a central warehouse and retailers. Howard and Marklund (2011) and Howard (2013) extend this research by investigating OWMR systems that apply state-dependent allocation policies instead of the computationally attractive First-Come-First-Served (FCFS) policy assumed in Marklund (2011). Their findings indicate that while FCFS performs well in general, costs can be reduced in certain situations by using state-dependent policies, particularly when allocation decisions are postponed until shipments approach the retailers. Stenius et al. (2016) generalize the model in Marklund (2011) by allowing for compound Poisson demand. Their approach derives the probability mass functions for the number of back-orders at the central warehouse destined for individual retailers, which are then used to calculate inventory levels at all stock points. In contrast, Johansson et al. (2019) apply real data to develop and evaluate efficient heuristics for optimization and control in systems where exact methods become computationally intractable.

Scheduled shipments at fixed intervals can also be accomplished using periodic review models. By reviewing the inventory system periodically, dispatches can be synchronized to the corresponding period, thus enabling time-based shipment consolidation. In the literature on stochastic divergent multi-echelon systems with periodic review, some studies monitor inventory levels between replenishment periods. These models often assume virtual allocation of retailer orders, see e.g., Axsäter (1993b), Shang and Zhou (2015), and Graves (1996).

Research on OWMR systems with quantity-based shipment consolidation and stochastic demand is relatively scarce. The only work we are aware of, except for the research papers in this thesis, is an unpublished approximation method from a study by Kiesmüller and de Kok (2005). They examine a multi-item, divergent multi-echelon inventory distribution system with compound renewal demand and quantity-based shipment consolidation. The model assumes that upstream warehouses replenish downstream warehouses using consolidated shipments of a fixed quantity. They derive an approximation method to estimate key system characteristics, such as lead time and inventory levels, to derive heuristics for cost estimation and decision-making.

Exact analysis of OWMR systems with quantity-based shipment consolidation and stochastic demand has, to the best of our knowledge, so far only been studied in the papers in this thesis. *Paper I-III* all study systems where quantity-based delivery restrictions are used to ensure that all shipments are dispatched with full load carriers or truckloads.

Moreover, *Paper IV* is the first to analyze an OWMR system combining time-based and quantity-based shipment consolidation policies into a hybrid shipment consolidation policy.

OWMR systems without shipment consolidation

Multiple scientific papers consider exact analysis of similar stochastic continuous review OWMR systems without shipment consolidation and delivery restrictions. Most relevant for this thesis is Axsäter (2000), as it introduces the class of unit-tracking approaches in which the analysis of inventory levels in the system is based on the consideration of whether an arbitrary unit arrives at a retailer before or after the corresponding demand occurs. The approach is used to derive the probability mass function of the inventory level at each stock point, which can be used to derive cost and performance metrics. The paper considers an OWMR system facing compound Poisson demand, with all stock points replenished using (R, Q) policies. Other models that consider exact analysis of OWMR systems with (R, Q) policies and/or compound Poisson demand are found in Axsäter (1993a, 1997), Forsberg (1995, 1997), and Chen and Zheng (1997). Some more recent papers featuring more general replenishment policies, allocation rules, or delivery restrictions, include Marklund (2002, 2006), Moinzadeh (2002), and Axsäter and Marklund (2008). The unit-tracking approach from Axsäter (2000) is different from the traditional unit-tracking approach introduced in Axsäter (1990), where costs and service are considered directly by following an arbitrary unit through the system and deriving the expected time it spends as stock-on-hand and as a backorder, respectively. Axsäter (1990) consider an OWMR system with a base-stock $(S-1, S)$ policy, FCFS allocation, and Poisson demand. Similar systems and policies are modeled in Simon (1971) and Graves (1985) using a different approach. For a more comprehensive overview of the existing literature on stochastic multi-echelon inventory systems, we refer to Federgruen (1993), Axsäter (2003), and de Kok et al. (2018).

Positioning of the thesis in the literature

This thesis presents models for quantity-based and hybrid shipment consolidation within multi-echelon inventory distribution systems. The hybrid policy generalizes both time-based and quantity-based shipment consolidation by allowing these policies as special cases. Thus, it allows the thesis to address all three shipment consolidation approaches discussed in the literature (including time-based consolidation, although not explicitly modeled). The table below presents a small, non-exhaustive summary of how the thesis relates to the existing literature on shipment consolidation in inventory management.

Table 1.1: Schematic positioning of this thesis within the literature on Inventory Control with Shipment Consolidation. The references in the table are not exhaustive.

	Single-echelon	**Multi-echelon**
Time-based	Çetinkaya and Lee (2000) Axsäter (2001) Ralfs and Kiesmüller (2022)	Marklund (2011) Stenius et al. (2016) Stenius et al. (2018) Johansson et al. (2019)
Quantity-based	Chen et al. (2005) Çetinkaya et al. (2008)	Kiesmüller and de Kok (2005) **Paper I-III in this thesis**
Hybrid (time & quantity)	Çetinkaya and Bookbinder (2003) Çetinkaya et al. (2006) Mutlu et al. (2010) Wei et al. (2023)	**Paper IV in this thesis**

The author's Licentiate thesis

This dissertation is a continuation and expansion of the work initiated in the author's Licentiate thesis Malmberg (2022). Although some aspects of the central research question from Malmberg (2022) remain within the scope of this research, the scope has also been expanded to allow for more general policies. As a consequence, the research in this thesis can be used to model more practically relevant policies, enhance the cost-efficiency of inventory systems, and aid in reducing the environmental footprint of supply chains. In some instances, certain portions of the text from Malmberg (2022) have been reproduced or adapted herein to maintain coherence and continuity of the research.

Since the publication of the licentiate thesis, the work in Paper I and Paper II has been published in well-known academic journals in the field of Operations Research. In this process, Paper I has undergone some minor editorial changes, while Paper II has been expanded and contains more stand-alone content. In the licentiate thesis, Paper II was written as a technical note. The work in Papers III and IV constitutes new contributions exclusive to this thesis and has been submitted to well-known academic journals within the field.

Chapter 2

Research objective

The research objective for this thesis is formulated as follows:

> *To develop mathematical models for exact evaluation and sustainable inventory management of stochastic multi-echelon distribution systems with shipment consolidation and delivery policies.*

This chapter is devoted to clarifying and defining the notions in the above formulation.

We consider divergent inventory systems where each stock location may have a maximum of one preceding supplying stock location. This is commonly referred to as a *"distribution system"*. In this thesis, we model *"multi-echelon"* systems with two physical stages, where one centralized location (central warehouse) supplies several other stock locations (often referred to as retailers), which in turn face customer demand. Such distribution systems are often called One-Warehouse-Multiple-Retailer (OWMR) systems and are commonly found in both practice and literature. Note that an OWMR system may be part of a larger supply chain, and although the model does not explicitly include previous echelons, the expected delay and time for material handling should be incorporated in the lead time to the central warehouse. Similarly, the "customers" may not be the end consumers but can also be other facilities, such as maintenance facilities, depots, workshops, or production sites. Chapter 4 provides a more thorough discussion of the modeling aspects of an inventory system.

The systems that we consider face a *"stochastic"* demand process. As a consequence, all state variables in the system are random.

By *"sustainable"* inventory systems, we mean economic and environmental sustainability and, more specifically, systems that are cost-efficient and have limited emissions of greenhouse gases from transportation between echelons.

"develop mathematical models ..." refers to the scope of designing a mathematical model of the inventory distribution system that mimics the real system close enough that results from analyzing the model are useful in the real world. As with any model, this is not necessarily equivalent to making the model as detailed as possible, but instead capturing the characteristics and dynamics that are important for gaining a specific insight or making a decision.

By *"exact evaluation,"* we refer to the derivation of exact (non-approximate) analytical expressions for important state variables in the system. We focus on the probability mass function of the inventory levels at different locations. Given system characteristics and cost parameters, these results can be used to derive expressions for the expected costs, transport emissions, and service measures (such as fill rates) induced by the system. Notably, as distributions are derived for state variables, we are not restricted to specific cost structures, emissions structures, and service measures. However, in this thesis, we consider those most commonly found in the literature and in practice.

"inventory management" refers to controlling the inventory system and optimizing decision variables within a reasonable decision policy. Primarily, the objective of the optimization is to minimize the expected inventory and transportation costs subject to constraints on emissions and service levels. The decision variables studied in the papers of this thesis include, e.g., reorder levels, time intervals between periodic shipments, fixed shipment quantities, and, to some extent, order quantities. The complexity of OWMR inventory distribution systems makes it inherently difficult (and perhaps impossible) to find general optimal policies, and no general optimality results are available in the literature. We therefore emphasize that *"optimizing"* means optimizing decision variables for policies that can be used in practice. The decision-making scope of this thesis concerns decisions within replenishment and shipment/delivery policies.

"shipment consolidation and delivery policies" refers to how shipments are dispatched between stock points in the system. The conventional inventory control literature mainly focuses on replenishment and allocation policies. These models typically assume that orders may be split and dispatched in partial deliveries without any restrictions or extra costs related to the number of dispatches or the number of units per shipment. To account for this from both a cost and environmental perspective, we extend the traditional scope by including delivery policies and shipment consolidation in the decision-making.

Chapter 3

Methodology

The choice of research method employed in this thesis stems directly from the research objective as the thesis focuses on developing *"...mathematical models"* for *"inventory control"*.

The standard practice for mathematical/analytical problem-solving and decision-making within operations management in general is found in the field of Operations Research (OR) and Management Science (MS). In particular, using applied mathematics for inventory management and inventory control, often referred to as *inventory theory*, is an established subfield within the OR/MS area. In this dissertation, we consider *stochastic multi-echelon systems* with uncertainty originating from random demand patterns. Consequently, the methodological focus is on *"stochastic models"*, an area within OR where randomness is described mathematically using, for example, probability theory and queueing theory.

A mathematical model is a representation/description of a system using mathematical concepts and expressions. For such a representation to be tractable, the model has to be a simplification of the real-world problem. As with any model building, mathematical modeling is a trade-off between analytical tractability and realism. If a mathematical model is an overly detailed representation of reality, it may be impossible to analyze and draw conclusions from the model. Conversely, if it is too simplistic or stylized, any conclusions drawn from the model may be invalid or a poor fit for the real-world system and therefore of limited relevance.

Thus, when optimizing and analyzing a system using a model, the model needs to be sufficiently detailed so that optimal decisions in the model are also optimal or near-optimal in the real system. Simultaneously, the model must be simplified enough to be tractable and to enable evaluation and optimization of the decision variables. Consequently, one of the main challenges in mathematical modeling is to balance the level of detail in the model to capture the most important aspects that characterize the system and the interactions between different decisions of interest. [1]

Figure 3.1 illustrates the general approach to using mathematical modeling as a tool for decision-making. The figure depicts the real-world system on the left-hand side and the model on the right.

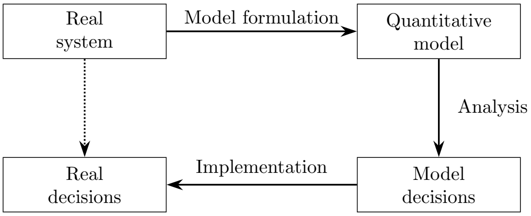

Figure 3.1: The figure illustrates how we can use quantitative models of systems as decision support for making real decisions for the system. In order for modeling decisions to be useful when implemented, the modeling assumptions must be realistic enough that conclusions drawn from the model will be relevant when applied to the real system, Axsäter and Marklund (2009).

Following this process, mathematical modeling for decision making in Operations Research can be summarized in three iterative steps (see, e.g., Axsäter and Marklund (2009) and Hillier and Lieberman (2000)).

The first step is to formulate a model of the real-world phenomenon we wish to analyze. In this first step, we face the trade-off described above between a simple and tractable model on the one hand and the level of detail and complexity on the other. If the objective is to optimize the system based on some performance measures and a set of decisions, these objectives must be formulated together with possible constraints. In this thesis, we do not have a specific case or company in mind; however, by using assumptions that are commonly found in both industry and academia when modeling inventory systems, we ensure that the proposed models have practical relevance. To apply the models proposed in this thesis to problems in a real system, model parameters and system topology must be adjusted to the specific situation.

The next step is to derive expressions and algorithms that evaluate and optimize the performance measures in the system. Approximations and heuristic

[1] There is a classic quote credited to the statistician George Box: “All models are wrong, but some are useful.”

methods may be necessary if an exact solution (and/or finding the optimal decisions) is intractable. Part of the objective of this thesis is to formulate relevant mathematical models that we can evaluate exactly. In general, we start with relatively simple models and increase the complexity and level of detail in the model step-by-step. The complexity is increased until the model fulfills its purpose and valid results are obtained, or when it becomes intractable (or too computationally cumbersome) to derive analytical solutions. For this reason, we may also need to simplify the model or change the assumptions to derive exact expressions for the performance measures we are evaluating. If the analytical or computational complexity is too high, a common approach is to find an approximate evaluation method. Similarly, when optimizing decision variables in the model, it is desirable to find methods that are guaranteed to find the optimal variables; however, sometimes, these may be too computationally cumbersome. For these cases, we may have to rely on faster heuristic methods.

A vital procedure when deriving and implementing mathematical expressions and algorithms is to verify that all the formulas are correct and that optimization schemes are implemented correctly. Although not explicitly stated in the papers, all mathematical models programmed for this thesis have been verified by mimicking them in discrete event simulation software. Comparing the analytical results with simulated results has verified that the mathematical analysis is correct, given the assumed system characteristics.

When applying mathematical modeling for decision-making in real business problems, the final step is to validate the model and implement the decisions in practice. Model validation is the process of evaluating whether the model is an accurate enough description of reality. Thus, decisions, findings, and conclusions obtained from a valid model may also be assumed to work well for the real system. If the model is a poor description of the system, it may be necessary to go back to the first step and reformulate the model to capture the system's dynamics better. It is often not necessary to model all system dynamics in full detail, but rather to focus on those aspects that are influenced by the decision. Simplifying the model to include only the most relevant dynamics can be crucial to making the model tractable while still supporting sound decision-making.

In Section 4, the critical modeling assumptions used in this thesis are discussed, together with some other general considerations within inventory modeling. These assumptions are commonly found in both industrial applications and implementations, as well as the academic literature. Their implications have been thoroughly validated, and the knowledge of how to apply them in the field of inventory control is widespread.

Turning to the scope of this thesis, optimal policies in multi-echelon OWMR systems are inherently difficult to find. With this in mind, and to facilitate the implementation of the models, the focus is on analyzing relatively simple decision policies and decision rules that are common in the literature and industry. This makes it possible to evaluate and optimize the policy parameters analytically and enables the implementation of the policy in a company's existing ERP system. More details regarding the assumptions of the replenishment policies are provided in Section 4.5.

Chapter 4

Inventory modeling

This chapter addresses eight general features commonly used to characterize an inventory system in inventory theory literature. Moreover, the delivery policies that distinguish the models in this thesis from the existing literature are discussed. The aim is to provide a brief overview of these concepts from an inventory modeling perspective to provide a concise introduction for readers less familiar with the field. For comprehensive treatment, the reader is referred to a textbook on the subject, e.g., Silver et al. (2016) or Axsäter (2015).

The features discussed in this chapter are:

1. The **topology** of the inventory system,
2. the **review period** used to monitor the system,
3. the type of customer **demand** the system faces,
4. the **allocation policy** that determines in which sequence demands are served,
5. the **replenishment policies**, i.e., how and when orders are generated (both from outside suppliers and within the system),
6. the **delivery policy** used when dispatching shipments,
7. the **lead times** and the transportation times between stages,
8. and how the **evaluation of system performance** is conducted.

4.1 Topology

An inventory system is a formation of stock-keeping locations/installations that supply both other stock-keeping locations within the system and external customers or locations. The topology of an inventory system describes how these stock points are organized in the system.

The simplest topological structure when modeling an inventory system is the single-echelon system, where we focus on a single stock point. This stock point faces demand from external customers, and stock replenishments are made from an outside supplier/manufacturer, which is not explicitly modeled.

Notably, very few stock points in practice come without some kind of preceding activity or stock point. However, when translating a system to a mathematical model, the level of detail to incorporate depends on many factors. Therefore, studying only part of a supply chain may still be highly relevant. In this case, all activities regarding transportation, material handling, and even risk of shortage, are modeled as part of the lead time from the supplier to the stock point.

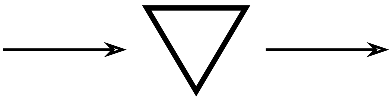

Figure 4.1: A single-echelon system

In multi-echelon systems, at least one additional echelon (stage) in the supply chain is modeled, i.e., we have multiple stock-keeping installations connected to each other. The simplest multi-echelon system structure is a serial system where each stock point has one stock-keeping successor and one stock-keeping predecessor. In this case, we may model the risk of shortage and the uncertainty of the capacity of the immediate predecessor explicitly.

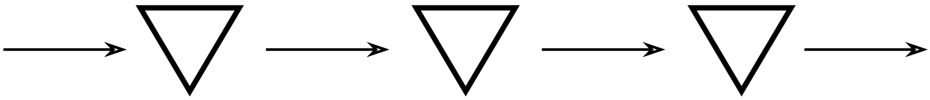

Figure 4.2: Example of a multi-echelon serial system

In a general multi-echelon system, each stock point may have several predecessors and successors. However, there are some special cases commonly referred to in the literature. For instance, an *assembly system* is a system where each stock point can have multiple predecessors but only one successor.

In this thesis, we study divergent multi-echelon inventory **distribution** systems. In these kinds of systems, each installation may only have one predecessor, while multiple successors are allowed. A simple structure for a multi-echelon distribution system is a One-Warehouse-Multiple-Retailer (OWMR) system. For example, modeling a central warehouse supplying several local warehouses or retailers.

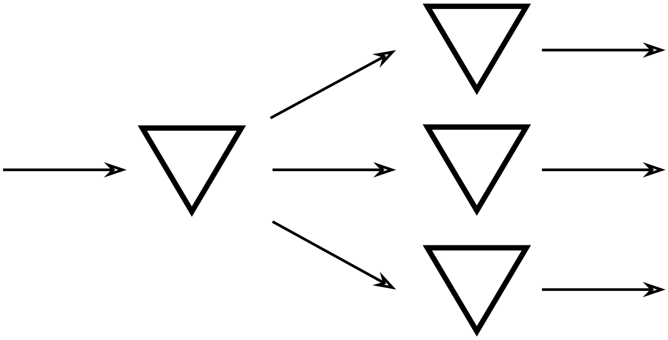

Figure 4.3: Example of a One-Warehouse-Multiple-Retailer multi-echelon distribution system with three retailers

Furthermore, when optimizing decision variables in this thesis, we assume that the inventories are managed centrally, i.e., that there is one decision-maker who controls the whole system. In contrast to centralized inventory control, a decentralized system assumes multiple decision-makers, where the outcome of a decision depends on other agents' decisions.

In line with most of the literature on inventory control for divergent distribution systems, we refer to the stock points at the lowest echelon as retailers. However, this does not necessarily mean that the system deals with consumer goods for end customers. It could just as well be, for instance, an internal spare-part supply system where the retailers correspond to maintenance facilities, depots, or manufacturing sites.

The retailers benefit from the inventory pooling at the central warehouse, which means that the retailers can be supplied at a lower cost. In most cases, when considering multi-echelon systems, the objective is to serve its external customers cost-efficiently. Thus, the service performance at the lowest echelon is of particular interest. By incorporating the central warehouse into the model, the uncertainties of the lead times to the retailers can be modeled in more detail. Thus, we can take a holistic view of where to put the stock to serve customers best when designing our system. Typically, when optimizing inventory levels in OWMR systems, the optimal service level at the central location is much lower than the service at the retailers, Axsäter (2015). However, many companies tend to set service levels at upstream and downstream locations independently using single-echelon techniques, often resulting in excessively high service at the warehouse; even so, we can see a trend of stock-keeping closer and closer to the end consumer also in practice (Silver et al., 2016).

It is worth mentioning that when analyzing a multi-echelon system such as the OWMR system, the central warehouse creates dependencies across the retailers. Thus, the mathematical complexity of analyzing these systems increases a lot compared to single-echelon systems. It may be tempting to model additional echelons (e.g., a stock-location supplying the central warehouse) or the production side of the supply chain to represent reality better. Although this is true in principle, this reflects the trade-off described in Chapter 3 between a model that mimics real-world behavior in detail and analytical tractability. For this reason, this thesis does not address upstream supply chain echelons beyond the central warehouse or variability in lead times from upstream suppliers. Incorporating such elements would increase complexity and reduce analytical tractability without significantly contributing to the research objectives. From a modeling perspective, any delay caused by limited capacity at the outside supplier is incorporated into the lead time to the central warehouse.

Furthermore, it is reasonable to limit the model's scope to consider only the locations where centralized decisions can be applied. For distribution systems with centralized control, the OWMR topology is commonly seen in practice. In these cases, the considered OWMR system may be one of many customers of the external supplier/manufacturer.

The literature on multi-echelon inventory theory also analyzes other structures. For instance, assembly and distribution systems can be generalized into a class of inventory systems allowing stock points to have both multiple stock-keeping successors and predecessors. Although such systems can be seen in practice, for instance, in supply chains that include both the production and the distribution of products (Axsäter, 2015), models for centralized inventory control in these systems are highly complicated. It is often a good approach to break down these systems into more easily analyzed subsystems. Thus, models considering OWMR systems may also be useful for studying a system with a more complex topology.

The design of a distribution system and its topology will, of course, impact the performance in terms of both costs and emissions. In this thesis, the topology and, e.g., facility locations are not part of the decision scope; however, the methods can be used to evaluate different network designs. We refer to Velázquez-Martínez and Fransoo (2024) for insights on how facility locations can be analyzed from both an emission and cost perspective.

4.2 Review period

The review period of an inventory system refers to the time between successive reviews of the system state. As no new knowledge is added during the review period, it also defines the time points at which decisions to take action can be made, e.g., placing an order or dispatching a shipment. The review and monitoring of inventory systems are conducted in the inventory control literature through two primary approaches: continuous review and periodic review.

In a periodic review system, the status review occurs at fixed time intervals, such as daily or weekly reviews. Decisions are then based on the state of the system at these specific time instances. In contrast, in a continuous review system, the system is monitored in real-time, enabling decisions as immediate responses to new information regarding, e.g., demands and replenishment. Modern information systems facilitate automatic access to real-time data, such as, for instance, point-of-sale (POS) and failure information. The use of this kind of system has significantly reduced, or even eliminated, the cost of monitoring the state of an inventory system. In the absence of additional costs for monitoring the system, there is no theoretical advantage to using a periodic review system compared to a more general continuous review when it comes to finding optimal policies. Nevertheless, a periodic review policy can facilitate the design of policies to coordinate material handling and shipment consolidation, particularly for systems with relatively high demand, as it facilitates coordination decisions across multiple facilities, products, or items (Axsäter, 2015).

To take advantage of both continuously monitoring the inventory system and the coordination of periodic dispatches of items, there is a stream of literature on time-based shipment consolidation in continuous review systems. These systems both continuously monitor the inventory system while using periodic, time-based, material handling and coordinated dispatches across multiple retailers and items, see, e.g., Marklund (2011), Stenius et al. (2016), Stenius (2016), or references therein.

In the scope of this thesis, we explicitly consider the coordination of shipments and material handling in systems with centralized control where there are no (or negligible) additional costs associated with monitoring the system continuously. Moreover, the models proposed in the thesis are well-suited for items with relatively low and intermittent demand, e.g., spare parts. For these kinds of items, continuous review is often superior to periodic review, even without the inclusion of delivery policies (Axsäter, 2015).

4.3 Demand

Most real-world inventory systems face some kind of uncertainty in demand. This is particularly true for distribution systems. The objective of this thesis includes modeling uncertainty in the demand structure, i.e., stochastic demand models.

Moreover, the systems modeled in this thesis handle items quantified in discrete units, in contrast to, for instance, fluids that can be handled in any quantity. There are approximation methods to apply models for continuous demand in situations where discrete demand structures are present. Most of these methods approximate the discrete demand during a time period as a continuous random variable. However, such models are generally only suitable when demand is high and the relative impact of a single unit is small (see, e.g., Axsäter (2015)). The papers in this thesis consider only discrete stochastic demand models. These models assume intermittent demand pattern more accurately, which is necessary for items with less frequent (lower) demand.

Discrete stochastic demand in a distribution system can be modeled in several ways. In continuous review systems, each customer's actual arrival time is of interest; thus, it is suitable to model the demand as a stochastic process where the arrivals of customers at different retailers are assumed to follow an arrival process. In probability theory, an arrival process is defined by the stochastic inter-arrival time between customers. When the stochastic inter-arrival times between different arriving customers are independent, the arrival process is called a renewal process.

In this thesis, customer arrivals are assumed to occur according to different Poisson processes. This is a special case of the renewal process where customers at the same retailer arrive with an inter-arrival time modeled by identical, independent, exponentially distributed random variables. The parameters for these distributions may differ among the retailers, but the demand structure remains the same.

Each customer demands a discrete number of units, independent of previous customers' demand quantities. In *Paper I*, *Paper III*, and *Paper IV*, we assume each customer orders only one unit at a time; thus, the demand process is a (*pure*) Poisson process. In *Paper II*, the number of units ordered by a customer follows some arbitrary discrete probability distribution (the distribution may differ among the retailers). Such a demand process is generally referred to as a compound Poisson process, as the individual customers arrive according to a Poisson process and each customer demands a stochastic number of units according to an independent *compounding* distribution.

The Poisson arrival process is often an adequate model for aggregate demand. According to the Palm-Khintchine theorem, the superposition of a finite number of independent equilibrium renewal processes, each with finite but individually small intensity, behaves asymptotically like a Poisson process as the number of streams grows (Heyman and Sobel, 2004). Hence, when demand arises from many independent sources that generate events infrequently, treating the total arrival process as Poisson is theoretically justified and widely adopted in the literature and practice.

This thesis does not address forecasting of demand; the statistical properties of the demand processes are assumed to be known and stationary over time.

4.4 Allocation policies

The allocation policy addresses situations where a choice can be made regarding where a unit is to be delivered. For example, when a stock point receives a replenishment but can fulfill only some, and not all, backorders, a decision must be made about which backorders to satisfy first, i.e., where to allocate the units for best use. Optimally allocating resources is very complicated in an OWMR system. An optimal decision must be based on real-time information about the inventory situation at all stock points, including the pipelines of all outstanding orders. Even if it is possible to take this information into account for a specific replenishment, it is extremely difficult to formulate a set of decision rules (i.e., a policy) that takes this information into account in a meaningful way. For these reasons, general optimality results regarding the allocation policy are very rare in the multi-echelon inventory control literature.

The complexities of describing and implementing state-dependent allocation decision rules have led to the use of simpler policies in practice. Arguably, the most commonly used policy, both in literature and in practice, is the so-called first-come, first-served (FCFS) allocation rule. For instance, using an FCFS policy at the warehouse allocates units to retailers in the sequence that their demands occur, i.e., the demand that occurred first receives the first unit, even though the unit may do better if allocated to another retailer (due to, e.g., lower stock-on-hand or higher backorder costs). All papers in this thesis are analyzed with an FCFS allocation rule. The motivation is that FCFS is generally considered fair, is commonly used in practice, and is tractable.

Special cases where the FCFS policy is both counterintuitive and poorly performing may be found both in conventional OWMR systems and in the systems studied in this thesis. In particular, when analyzing systems with shipment consolidation, there can be a delay between when a single unit is qualified to be shipped from the warehouse (i.e., has been demanded by a retailer and is available at the warehouse) and when the dispatch occurs. This provides additional time during which a reallocation could improve utility. Nevertheless, it has been shown that FCFS performs well in most practical scenarios, including when alternative shipment policies are used. See, for example, Graves (1996), Howard and Marklund (2011), and Howard (2013).

In some systems, there could be specific reasons to deviate from the FCFS policy if there are certain key customers or locations that require a higher service level than others. For instance, Berling et al. (2023) consider an OWMR system in an omnichannel setting, where the warehouse faces demand from both retailers and end customers. They allow the warehouse to reserve units for direct customer demand, thus prioritizing the end consumer over the retail channels where safety stock can be applied.

Paper III deals with issues that can arise from the FCFS allocation rule in systems with quantity-based shipment consolidation to multiple retailers in consolidation groups. In the paper, the priority sequence is altered by allowing a replenishment policy to delay orders, thereby avoiding situations where FCFS allocation performs poorly while retaining mathematical tractability.

4.5 Replenishment policies

Finding optimal decision rules for replenishments in OWMR systems is inherently complex, as the optimal decisions are state-dependent and depend on the optimal allocation policy, which is also state-dependent. Since optimal decision rules to replenish OWMR systems remain unknown, most literature assumes that a simple and practically relevant policy is used and focuses on evaluating costs and determining optimal or near-optimal parameter values for decision variables within these policies. In this section, we discuss some of these replenishment policies. First, we define some commonly used concepts and notions to describe these policies.

A replenishment decision cannot be based solely on the physical stock-on-hand situation. It must also consider the units already ordered from the upstream echelon that have not yet arrived, denoted as outstanding orders, and any unmet demand at the customer or downstream stock point, i.e., backorders (back-ordered units). Throughout this thesis, we assume complete backordering: all demand that cannot be filled immediately is queued and delivered as soon as inventory becomes available. Complete backordering is especially realistic in spare-parts management, where customers are typically willing to wait for the replacement part, as it may not be available elsewhere. A different assumption frequently used in the literature is lost sales, where unmet demand vanishes permanently. This is often a reasonable assumption for standardized consumer products in a retail setting and is not analyzed in this thesis.

To account for the outstanding orders and backorders, the notion of *Inventory position = stock-on-hand + outstanding orders - backorders* is commonly used. The inventory position traditionally reflects the number of demanded units that can be satisfied without placing additional orders from upstream locations. Note that this is not necessarily true in a system with delivery restrictions, as such mechanisms can prohibit units from being dispatched until additional orders are placed and ready for dispatch. This is discussed in more detail in Section 4.6.

In addition, the decision of when and how much to order should consider both the costs of keeping stock-on-hand (holding costs) and the consequences of customers having to wait for their demand to be satisfied (reflected by a shortage cost or service constraint). As both the stock-on-hand and backorder level are to be considered, it is convenient to define the *inventory level*, IL, such that *stock-on-hand* is the positive part of the inventory level ($IL^+ = max(IL, 0)$) and *backorder level* is the negative part of the *inventory level* ($IL^- = max(-IL, 0)$). By denoting the *outstanding orders*: O and the *inventory position*: IP, the relationship between these entities may be summarized as follows:

$$IP = IL + O = IL^+ - IL^- + O. \tag{4.1}$$

Both in practice and in the literature, commonly used replenishment policies include the (R, Q) and the (s, S) policies.

In an (R, Q) policy, an order of exactly Q units (the order quantity) is placed once $IP \leq R$, i.e., the inventory position reaches (or falls below) the reorder point. Since the inventory position also includes outstanding orders, the order will immediately add the quantity, Q, to the inventory position. In cases where the customers (or a succeeding stock point) may place orders of more than one unit at a time, it may be that triggering one order quantity is not enough to bring the inventory position back up above R. In these cases, the (R, Q) policy may be referred to as a (R, nQ) policy. That is, once $IP \leq R$, an order of some discrete multiple of Q units is placed such that the resulting inventory position is $R < IP \leq R + Q$.

In a (s, S)-policy, an order of $S - IP$ units is placed once $IP \leq s$, i.e., once the inventory position reaches (or passes below) s an order is generated such that the inventory position reaches S again. This means that the order quantity may differ if we do not reach the reorder point exactly, for instance, if the demand occurs for multiple units simultaneously or if the review is not continuous.

The special case of the (s, S) policy, where $s = S - 1$, is usually called a base-stock policy or a one-for-one replenishment policy. When using such a policy, demand information is immediately conveyed to the preceding installation. The base stock policy is equivalent to a (R, nQ) policy where $Q = 1$ and $R = S - 1$.

Furthermore, the (s, S) and the (R, Q) policies are equivalent if, for each order cycle, the inventory position reaches the reorder point exactly. This is the case for continuous review models with discrete demand processes if each arriving customer always demands exactly one unit, as in the (pure) Poisson demand process studied in *Paper I*, *Paper III*, and *Paper IV*. However, in *Paper II*, we consider compound Poisson demand. In this case, a customer demanding more than one unit may bring the inventory position below the reorder point, and thus the (s, S) and the (R, Q) policies are not equivalent. All papers in this thesis assume that the warehouse replenishes its stock with a (R, nQ) policy.

The motivation for using an order quantity is often associated with set-up costs for production or for placing an order. However, motivated mainly by transportation costs, the order quantity has also been used to reflect capacity restrictions when shipping the goods (Silver et al., 2016). For instance, there could be a fixed cost per transported container, regardless of the quantity. However, in multi-echelon inventory control, the use of order quantities to reflect restrictions (or costs) for physical handling and moving material may fail to capture the intended features. Most literature on stochastic multi-echelon inventory control in OWMR systems assumes that an order quantity is partially delivered if not all units in an order are available simultaneously. To properly deal with fixed costs and emissions for ordering and transportation, we need to distinguish between ordering and delivery. This is a key aspect of the research presented in this thesis.

4.6 Delivery policies

In most continuous-review inventory models, units are dispatched from an inventory location as soon as they have been demanded and are available at the location. As a result, units ordered together or at the same time will not always be dispatched simultaneously. If demanded units become available at different times (we refer to units that are available at a stock location and have been demanded as *qualified units*), the delivery may be split across different shipments and times. The policies examined in this thesis modify this norm by applying a set of decision rules that constrain shipments and deliveries, potentially improving the distribution process by consolidating shipments across multiple demands and/or orders.

There are multiple ways to consolidate shipments. As mentioned above, Higginson and Bookbinder (1994) identify three classes of policies for shipment consolidation programs: time-based, quantity-based, and hybrid (time- and quantity-based) consolidation policies, all of which can be found both in practice and in the literature.

In an OWMR system with a time-based shipment consolidation policy at the warehouse, shipments are restricted to dispatches at fixed time intervals. This means that all units qualified for shipment within an interval are dispatched at the end of this interval. Consequently, the time instances when shipments will be dispatched are known, allowing for more efficient planning and synchronization of activities across the inventory distribution system.

When a quantity-based shipment consolidation policy is applied at the warehouse, dispatches from the warehouse to the retailers (or retailer groups) are restricted to a specific shipment quantity. Under this policy, units are not shipped until a full quantity is *qualified* for shipment. Compared to the time-based consolidation policies, the advantage of this kind of policy is that, by design, it maximizes the capacity utilization of load carriers and transport vehicles. This reduces transportation costs and emissions by decreasing the need for load carriers and vehicles.

A finding in the single-echelon literature for shipment consolidation is that quantity-based policies usually outperform time-based and hybrid policies in terms of the expected total inventory and shipment costs when assuming the same fixed set of shipment costs for both shipment options. However, this assumption may be questioned - for example, if a third-party logistics (3PL) provider can offer a time-based shipment schedule with lower fixed costs because of improved coordination, advance planning, and consolidation across different companies.

The hybrid policy combines time-based and quantity-based shipment consolidation. Under this policy, the warehouse dispatches full quantities or at fixed time intervals, depending on what occurs first.

All of these decision rules are well known to practitioners and have been used in industry to reduce transport-related costs for a long time (Jackson, 1985). As shipment consolidation and delivery policies are principal concepts in this thesis, Chapter 5 is devoted to a more elaborate description of these issues.

4.7 Lead times

The replenishment lead time is defined as the period from placing an order (of one or several units) until the units arrive and are available at the ordering installation. Consequently, the lead time includes all intermediate activities such as transportation and material handling (e.g., picking, loading, receiving).

The lead time also includes waiting time if, for some reason, the order fulfillment at the preceding installation is delayed. Traditional inventory control theory typically attributes such delays to situations where an upstream installation runs out of stock, i.e., stock-out delays. However, waiting time can also be caused by, e.g., a delivery policy. For example, this may happen if we only ship at specific time slots (periodic review or time-based shipment consolidation) or wait until enough demands have accumulated for a predefined shipment quantity to be complete before we dispatch it (quantity-based shipment consolidation). Both of these situations are studied in this thesis.

In the models analyzed in this thesis, we assume that transportation times (including associated material handling) are deterministic. However, we explicitly model the variations in the lead time caused by stock-outs at the central warehouse and the delivery policies for shipments between the warehouse and the retailers. An outside supplier supplies the warehouse with a constant lead time, and any delays in these replenishments are modeled as part of this fixed lead time, for example, by adding the expected delay.

4.8 Evaluation of system performance

A performance measure is essential for evaluating and comparing various systems, decision policies, or solution methods. The overall objective of decision-makers is often to maximize the system's utility, which is commonly expressed as profit maximization or as minimizing the system's total costs. However, other considerations - such as the environmental impact of different types of emissions - can also exist and are sometimes difficult to quantify in monetary terms.

From a cost perspective, when analyzing an inventory system, we focus only on the costs that vary with the decision variables under consideration to assess the impact of different decisions. It is common to include *holding costs*, *replenishment costs*, and *shortage costs*. The cost of capital is often seen as the primary component of the cost of holding stock (Axsäter, 2015). Inventory ties up capital that could have been used for alternative investments, leading to opportunity costs and interest expenses. However, other costs associated with holding stock-on-hand should also be considered, e.g., material handling, warehousing, obsolescence, insurance, and taxes.

The main reason for orders and shipments to be placed or performed in batches is that replenishment costs include fixed or semi-fixed charges. These costs should reflect any fixed cost of placing and handling an order throughout the system. Replenishment costs usually include production set-up, administrative costs of handing over an order, fixed transportation costs, and costs for material handling. In the systems studied in this thesis, the dispatch and shipment-related costs may be separated from the costs related to actual orders. This is necessary as the quantities dispatched in shipments do not necessarily correspond to the way units have been ordered.

Particularly in systems facing stochastic demand, situations may arise where a stock location cannot immediately satisfy a demand from the available inventory. We refer to such unmet demands as shortages. As previously mentioned, there are two dominant approaches for handling shortages in the literature. If the customer is unwilling to wait, they may cancel their order, resulting in a lost sale. Alternatively, if the customer is willing to wait until the stock is replenished, the demand is backordered. This thesis considers only the case where all unsatisfied demands are backordered, i.e., *complete backordering* is assumed.

This behavior may seem unreasonable for many consumer goods in retail (at least for the end customer); however, there are plenty of examples (including, for instance, spare parts or internal production or replenishment sources) where such an assumption is appropriate. Although customers wait for the replenishment to arrive when shortages occur in a backorder model, this often comes at a cost (otherwise, there would be no need to keep anything in stock). The costs may include, for example, lost uptime, contractual penalties, administrative expenses, discounts, and lost goodwill and associated reputation. As *complete backordering* is assumed, we may refer to these costs as either shortage or backorder costs.

Shortage costs can be complicated to estimate, and in practice, it is common to replace the cost with a target service level. There are several ways to measure service. One of the most common, at least in the inventory literature, is the demand fill rate, which is defined as the proportion of demand that can be served directly from stock-on-hand. The models in this thesis can be applied to both service-level constraints and shortage costs.

Costs and/or service measures are usually the main inventory control objective for decision-makers; however, other performance measures often need to be considered. Today, we see an increased concern about the environment both in society and industry; thus, sustainability is becoming an increasingly important part of the agenda for top management. Decreasing the environmental footprint while remaining cost-effective and maintaining high service levels will benefit both the environment and competitiveness. Although multiple types of emissions are harmful to the environment, greenhouse gases (GHG) from freight transport within a distribution system are the main concern in this thesis. Other parts of a distribution system may, of course, also impact the environment; however, as previously described, when compared to transportation emissions, their contribution to the environmental footprint is often less significant (Marklund and Berling, 2024).

GHG emissions are commonly quantified by their CO_2 equivalent. The general approach in this thesis is to either include the emissions of greenhouse gases as a monetary cost, to have a constraint on the emissions, or to derive Pareto-optimal cost-emission efficient trade-offs, where no solution exists that can reduce one objective without increasing the other. The introduction to *Paper I* discusses emissions modeling in an inventory system in more detail.

Chapter 5

Modeling inventory systems with delivery restrictions and shipment consolidation

As explained above, this thesis focuses on analyzing OWMR systems in which different delivery or shipment consolidation policies restrict deliveries between the warehouse and the retailers. In this chapter, we will further explain the challenges this introduces for modeling the inventory system.

In a system without any delivery restrictions, a unit is shipped as soon as it is available at the warehouse and demanded by a retailer (referred to as units qualified for shipment). The demanded units may still be delayed at the warehouse due to stock-outs (i.e., the demand is backordered); however, as soon as new replenishments arrive at the warehouse, these backorders are cleared according to the applied allocation rule. For example, using a First-Come-First-Served (FCFS) allocation rule, backorders are cleared in the sequence they were received.

However, handling units one by one can be costly, and frequent deliveries in small quantities can increase both costs and carbon footprint in a supply chain.

There is a stream of literature that approaches such quantity-related aspects by considering ordering quantities. For instance, Bouchery et al. (2012) consider order quantities optimized from both a cost and sustainability perspective. However, using order quantities in a multi-echelon setting with stochastic demand does not necessarily guarantee that the physical movement of the good is performed in the same quantity as the order is placed. When ordering a quantity of Q units in a multi-echelon distribution system, all ordered units may not be available and ready to be dispatched from the preceding installation at the same time. Most inventory control models assume that partial deliveries are allowed, meaning that the warehouse will ship any demanded unit as soon as it is available and ready.

In this thesis, we address these coordination issues by applying delivery and shipment consolidation policies that restrict the physical flow of goods to ensure that the intended joint handling of multiple units is achieved. We focus on shipments within a centralized OWMR system. Thus, only shipments from a central warehouse to retailers are subject to the delivery restrictions.

Shipment consolidation can be achieved in multiple ways. When formulating and classifying a delivery policy, it is helpful to focus on how and when units are dispatched from the warehouse. As mentioned above, Higginson and Bookbinder (1994) discusses three classes of policies for shipment consolidation programs: time-based, quantity-based, and hybrid consolidation policies.

In an OWMR system with time-based shipment consolidation, shipments are restricted to be dispatched periodically with a fixed interval. All units that have been demanded and are available at these time epochs are dispatched together. The periodicity of these shipments also facilitates the planning and coordination of the material handling activities. Compared to unrestricted shipping, time-based shipment consolidation can potentially increase the load carrier utilization; however, as the number of units per shipment will be uncertain (depending on how many units have qualified for shipment between two dispatches), there may still be vacant capacity upon dispatch. In practice, periodic dispatches can facilitate distribution pooling from other sites or warehouses, thereby increasing load utilization and economies of scale. Moreover, there are examples in the literature of policies with time-based shipment consolidation that utilize advance demand information to improve capacity utilization by including units that have not yet been demanded by a retailer in the dispatch (see, e.g., Ralfs and Kiesmüller (2022)).

In a quantity-based shipment policy, units are dispatched from the warehouse in specific predetermined shipment quantities. This facilitates higher capacity utilization for the transports, as the quantity can be chosen to match the size of a full load carrier, truckload, or pallet.

A hybrid policy combines time-based and quantity-based shipment consolidation. Under this policy, the warehouse dispatches at fixed time intervals or full quantities, depending on what occurs first. In this thesis, the time-based shipments are fixed to a periodic schedule regardless of whether there are quantity-based shipments in between. This enables the system to get the above benefits of predictability and coordination. In the shipment consolidation literature for single-echelon systems, a hybrid policy is sometimes used slightly differently. Instead of having a fixed schedule for the time-based shipments, the time-based shipments are used as a cap on the shipment delay for the quantity-based shipments. This means that quantity-based shipments are dispatched if the number of qualified units after the most recent dispatch (either time or quantity-based) reaches the shipment quantity before the time interval for the time-based shipment. If the time interval occurs first, a time-based interval is triggered and dispatches all units that have been qualified so far. Although this can make sense in a single-echelon system to reduce long waiting times when the shipments are consolidated to consumers or customers, it does not achieve a fixed periodic schedule that enables efficient coordination and planning.

In the OWMR systems in this thesis, the hybrid policy instead has a fixed schedule for the time-based shipments; however, if full truck loads can be dispatched before these time instances, it makes no sense to wait until the time-based dispatch occurs, and therefore, these are dispatched in between. If, for example, a load carrier (e.g., a truck) with capacity Q is used for transports from the warehouse to a retailer or retailer group, all quantity-based dispatches will occur with exactly this capacity. The time-based shipments will be dispatched with a quantity less than Q, meaning that a truck with available capacity Q will always be able to pick up these shipments.

In a similar system, with a pure time-based shipment consolidation policy, the quantity to be dispatched can exceed Q, requiring the use of multiple load carriers. From this perspective, it is sensible to ship the full load carriers as soon as possible (i.e., as soon as Q units have qualified for shipment) rather than waiting for the periodic dispatch.

Notably, all of these policies generalize the traditional literature, modeling OWMR systems without shipment consolidation. By using a shipment quantity of a single unit in the quantity-based or hybrid shipment consolidation policy described above, all units will be shipped as soon as they qualify (note that if units qualify for shipment to the same retailer or retailer group at the same time, they may still be shipped together). The pure time-based shipment consolidation policy with a time between the dispatches that approach zero can also, in theory, be used to resemble this situation; however, in practice, there are more straightforward and more suitable methods to analyze an OWMR system without shipment consolidation, see e.g., Axsäter (1990, 2000).

Moreover, the hybrid policy generalizes both the time-based and quantity-based shipment consolidation policies. Using an infinite shipment quantity ensures that only time-based dispatches occur. Similarly, using an infinite time between time-based dispatches ensures that only quantity-based shipments occur.

All three policies, and the kinds of decision rules included in these, are well known to practitioners and have been used in industry to reduce transport-related costs for a long time (Jackson, 1985). Notably, when delivery restrictions and shipment consolidation policies are used, the lead time is affected by a variable shipment delay. Failing to consider the delivery restrictions in a system leads to underestimating the lead time between the warehouse and the retailers, as the shipment delay is not accounted for. If optimizing reorder levels based on an underestimated lead time, we expect to see high backorder costs (or alternatively undershooting fill rates) and, therefore, a substantial cost increase. Jointly considering inventory and delivery policies is therefore essential to optimize the system with respect to the delivery restrictions. Being able to analyze a stock location both before and after the transport makes it possible to place the stock where it is of best use, making the multi-echelon OWMR structure suitable for the analysis.

As discussed in Chapter 1, there is a stream of scientific literature that analyzes both time-based, quantity-based, and hybrid shipment consolidation in a single-echelon system motivated by a Vendor Managed Inventory (VMI) setting, see, e.g., Çetinkaya et al. (2008) and references within. In addition, for OWMR systems similar to those considered in this thesis, time-based policies have been analyzed in a couple of previous articles (see, e.g., Marklund (2011), Stenius et al. (2016, 2018), and Johansson et al. (2019)). However, the literature on quantity-based consolidation policies and hybrid (time **and** quantity-based) policies in multi-echelon systems is scarce. To the best of our knowledge, the papers in this thesis are the first to provide exact analysis and evaluation methods for models with these policies in stochastic OWMR systems.

The first paper, *Paper I*, analyzes a quantity-based shipment consolidation policy, where the warehouse consolidates shipments to groups of retailers (referred to as consolidation groups). The quantity-restricted policy only allows shipments to a retailer group once a specified number of units corresponding to a full load carrier (e.g., a vehicle, container, pallet, etc.) is available. Thus, a shipment is not dispatched until the retailers in the consolidation group have demanded a specified quantity, and these units are available at the warehouse. Since all shipments to a specific consolidation group have precisely the quantity specified, the model can be used as a tool to increase transport utilization and decrease the environmental impact.

In the second paper, *Paper II*, the retailers are not grouped into consolidation groups. Thus, the delivery restriction applies only to a single retailer. This may be perceived as a more restrictive model; however, compared to *Paper I*, we consider a more general demand process (compound Poisson demand instead of pure Poisson demand), and we also allow the retailers to place orders in specified quantities according to (R, nQ)-policies. It allows us to extend the cost structure with additional elements, e.g., ordering costs or discounts. Thus, the model in *Paper II* generalizes many aspects of *Paper I*; however, with respect to the shipment consolidation, *Paper II* is more restrictive than *Paper I.*

The policy in *Paper III* extends the policy from *Paper I* by introducing the possibility of using a joint order quantity for the retailer group. The purpose of the joint order quantity in this context is primarily to improve the allocation sequence and thereby reduce inventory costs in the system. A numerical study shows that there may be significant savings from using the improved policy in *Paper III* compared to *Paper I.*

Paper IV considers the most general delivery policy, using a *hybrid shipment consolidation policy* for dispatches from the warehouse to groups of retailers. The system structure shares most of the characteristics of the systems analyzed in *Paper I* and *Paper III.*

It is interesting to note that, although all papers have different possibilities in terms of shipment consolidation, all four models coincide in the special case of: pure Poisson demand, only one retailer per consolidation group, base-stock ordering policies at each of these retailers, and a pure quantity-based shipment consolidation policy to the respective retailer (i.e., an infinite time between time-based shipments when using the hybrid policy in *Paper IV*).

Table 5.1 provides an overview of the modeling features used in this thesis, highlighting key differences and similarities between the four papers and offering a simplified overview of the different policies and assumptions used. More detailed summaries of the papers are provided in Chapter 6.

Table 5.1: Comparison of different papers on various modeling assumptions (CW = central warehouse).

	Paper I	Paper II	Paper III	Paper IV
Demand structure				
Poisson demand	x	x	x	x
Compound Poisson demand		x		
Shipment policy from CW				
Quantity-based	x	x	x	x
Time-based				x
Hybrid (time- & quantity-based)				x
Replenishment structure				
(R,Q) at CW & (S-1,S) at retailers	x			x
(R,Q) at CW & (R,Q) at retailers		x		
(R,Q) at CW & (R,Q) at retailer groups			x	
Performance measures				
Shortage costs	x	x	x	x
Emissions	x	x		x
Fill rate	x	x	x	x
Exact evaluation	x	x	x	x

Chapter 6

Summary of papers

The following chapter provides summaries of the four scientific papers included in this dissertation, stating assumptions, problem formulations, overviews of the modeling and analysis approaches, and some key results. To maintain the connection between the summary and the respective article, the notations in each summary follow those in the articles. As the notations in the research papers vary slightly from one to another, the notations in the summaries will also vary slightly.

6.1 Paper I

Evaluation and Control of Inventory Distribution Systems with Quantity Based Shipment Consolidation

In *Paper I*, we consider a one-warehouse-multiple-retailer inventory system where the non-identical retailers are grouped into consolidation groups that are supplied with joint shipments of a fixed quantity following a fixed milk-run route.

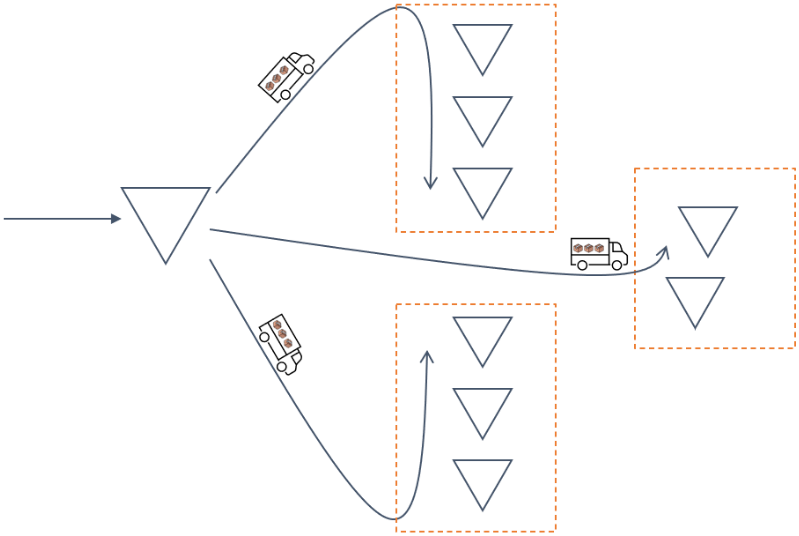

Figure 6.1: An illustration of the topology of the OWMR system when there are three retailer groups with 3, 2, and 3 retailers per group respectively.

The shipments are dispatched from the central warehouse as soon as the retailers in a group have ordered a specific number of units, referred to as the shipment quantity, **and** all of these units are available at the warehouse. As a result, all shipments to a retailer group correspond precisely to the shipment quantity, for example, a full truckload, a container, or another type of load carrier. Thus, the shipment consolidation policy has the potential to maximize the capacity utilization of the load carrier upon leaving the warehouse, thereby reducing the system's transportation costs and environmental footprint.

The system is centralized with an integrated supply chain information system offering free information sharing and access to real-time point-of-sale data at the warehouse. Consequently, there are no incentives for batch ordering at the retailers. Therefore, they immediately convey the demand information to the warehouse (this behavior corresponds to the case of base-stock policies). Fixed costs and emissions associated with handling and transporting goods from the warehouse to the retailers are reflected in the quantity-based shipment consolidation policy.

We assume that the transportation times between the warehouse and the retailers are constant. However, the lead times are stochastic due to potential delays at the warehouse caused by stock-outs or shipment delays from waiting for the full shipment quantity to be reached. Units that are delayed due to the shipment policy are referred to as *reserved stock-on-hand*.

The warehouse places orders from an outside supplier/manufacturer according to a (R, Q)-policy. The outside supplier is not explicitly modeled, and deliveries are made with a constant lead time. Furthermore, complete backordering and FCFS allocation are assumed at all stock points. The analysis focuses on a single-item setting; however, with some restrictions on the assumptions, an extension to a multi-item model is also provided.

The paper considers performance measures such as expected warehouse holding costs for both reserved and unreserved stock-on-hand, expected retailer holding costs, expected backorder costs, and expected shipment costs (cost per shipment). In addition to these long-run average costs, we also consider the fill rate service measure at each retailer and emissions generated by transports from the warehouse to the retailers. We show that all these metrics can be obtained from the probability distribution of the inventory levels at the retailers.

Deriving these probabilities to enable cost, service, and emission evaluation is the primary focus and contribution of the paper. However, we also provide an optimization routine for the expected costs in the system. In this optimization problem, the environmental aspects can be handled as either an emission constraint or additional shipment costs. The optimization also accounts for the customer service level by either including service constraints or backorder costs at the retailers.

The analysis approach in this paper is to consider an arbitrary customer arrival at some point in time, τ, and analyze how many orders that are outstanding at this time. The probability distribution of the number of outstanding orders can then be transformed into the distribution of the inventory level. The analysis is conducted for retailer N in retailer group M. The same procedure can be applied for all other retailers.

A key insight that enables our analysis is that the last unit to be ready for shipment, among the units in a shipment, experiences no shipment delay. Therefore, this last unit behaves exactly as in a system without a shipment consolidation policy (i.e., when unrestricted partial deliveries are applied). In other words, the dispatch time for an arbitrary unit is determined by when the last unit in the same shipment is both available at the warehouse and demanded by a retailer in the same group.

Consequently, by keeping track of both an *arbitrary unit* and the *last unit* in the same shipment quantity, we can determine the probability that the *arbitrary unit* is outstanding or not at time τ.

We introduce a priority list of the units that have been ordered at retailer N before τ. The higher the priority number, the earlier the order occurred. If an arbitrary unit, with priority n, has been delivered before τ, there can be at most $n-1$ outstanding orders at this time. Thus, the analysis focuses on whether such unit arrives before or after τ.

The unit will be part of some shipment containing exactly Q^M units, all going to retailers in retailer group M. Among these Q^M units, u, were ordered after the n^{th} unit in the priority list. Furthermore, among the u units, k are destined to the same retailer as the n^{th} unit. With these notations, and given u and k, the following probabilities are equivalent

1. Pr(the n^{th} unit in the priority list has been delivered at time τ)
2. Pr(the $n-k^{th}$ unit in the priority list has been delivered at time τ)
3. Pr(at most $n-k-1$ units are outstanding at time τ)
4. $Pr(O_N \leq n-k-1)$.

Because retailer N acts according to a base stock policy with base-stock level, S_N, we know that the outstanding orders, O_N, plus the inventory level, IL_N, equals the base-stock level, S_N. Consequently,

$$Pr(O_N \leq n-k-1) = Pr(IL \geq S_N - (n-k-1)). \tag{6.1}$$

Now as $Pr(IL = j) = Pr(IL \geq j) - Pr(IL \geq j+1)$ we focus on deriving Pr(at most $n-k-1$ units are outstanding at time τ given u and k). Thereafter, by taking the expectation over u and k, we obtain the unconditional probability, *Pr(*the n^{th} unit in the priority list has been delivered at time $\tau) = Pr(IL \geq S_N - (n-1))$.

The batch quantity at the warehouse further complicates the problem, as the warehouse delay (due to the risk of stock-outs) will depend on the warehouse inventory position associated with the considered last unit in the batch. To handle this issue, we introduce the notion of an inventory position corresponding to a specific unit. At the time the central warehouse places an order for a batch of units, we say that each unit has a corresponding inventory position, indicating which demand this unit will satisfy. If the inventory position for a specific unit is y, the unit will serve the y^{th} retailer demand, counting from when the order was placed.

Therefore, we must also condition the number of outstanding orders on the inventory position corresponding to the *last unit* in the shipment, IP_0^{lu}. Note that it is only the warehouse delay for this last unit that is of interest to find out whether a shipment has arrived at a retailer or not.

The final expression we use to evaluate the probability mass of the inventory level is

$$\begin{aligned} &Pr(IL_N \geq S_N - (n-1)) = \\ &\quad \text{P(unit } n \text{ has been delivered at time } \tau \text{)} = Pr(O_N \leq n-1) = \\ &\quad\quad \sum_{y=R_0+1}^{R_0+Q_0} \sum_{u=0}^{Q^M-1} \sum_{k=0}^{u} \Big(\xi(n-k-1|y,u,k) \cdot H(y,u,k) \Big) \end{aligned} \tag{6.2}$$

where,

$\xi(n-k-1|y,u,k)$ Probability that there are at most $n-k-1$ outstanding orders at retailer N (at time τ) given that $IP_0^{lu} = y$, the *considered* n^{th} *unit* has shipment position u and that k of these u units are destined to retailer N

$H(y,u,k)$ Probability that $IP_0^{lu} = y$, the *considered* n^{th} *unit* has shipment position u and that k of these u units are destined to retailer N

Because the *last unit* behaves in the same way with and without a consolidation policy, $\xi(n-k-1|y,u,k)$ is derived in the same way as in a system with unrestricted partial deliveries. When we derive $H(y,u,k)$, we find that both the inventory position y and the shipment position u are uniform over the attainable state space and that the probability that k out of the u units belongs to retailer N follows a binomial distribution. We provide an exact recursive method for determining these probabilities and the retailers' inventory level distributions.

As described before, this allows us to evaluate the expected inventory and shipment costs, fill rates, and transport emissions for the entire system.

An optimization method to minimize the total costs both with and without service and emission constraints is provided by deriving optimality bounds on the decision variables (reorder points and shipment quantities). The principle for these bounds is to determine a value of the decision variable from which a change could never decrease the cost further than the so-far best solution found. The lower bounds are dynamic in the sense that, as the algorithm finds better and better solutions, the bounds become tighter.

We also provide a computationally efficient heuristic, *the Economic Shipment Quantity* (ESQ), to find near-optimal values of the shipment quantity given the cost structure described earlier. The heuristic represents an adapted *EOQ* model where we assume deterministic demand and ignore the possibility of having backorders. In several numerical examples, we see that the *ESQ* heuristic performs very well. When applying the *ESQ* model to set shipment quantities, the optimization procedure is reduced to finding optimal reorder levels, which makes the procedure much faster.

Moreover, further numerical experiments and sensitivity analyses demonstrate that failing to account appropriately for shipment-quantity constraints during system-parameter optimization can lead to significant extra cost. Furthermore, we see that it is often possible to increase the shipment quantity without a substantial increase in costs if the longer lead time is adequately compensated for by optimizing the reorder levels in the system. Higher shipment quantities lead to less frequent dispatches, and fixed shipment quantities may increase the utilization of the shipments. Thus, imposing delivery restrictions may be a cost-effective tool to decrease an inventory system's environmental footprint.

6.2 Paper II

Exact Analysis of One-Warehouse-Multiple-Retailer Inventory Systems with Quantity Restricted Deliveries

Paper II studies a divergent multi-echelon inventory distribution system where a single warehouse supplies an arbitrary number of retailers, i.e., an OWMR system as illustrated in Figure 6.2.

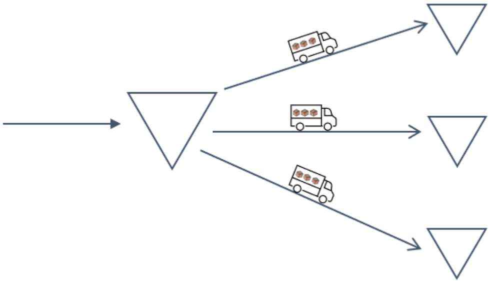

Figure 6.2: An illustration of the topology of the OWMR system for the special case of three retailers.

The retailers face stochastic demand, according to a compound Poisson process, and use continuous review (R, nQ) policies to place replenishment orders from the warehouse. For these replenishments, the warehouse applies a quantity-based delivery policy that restricts the shipments from the warehouse to each retailer to fixed quantities. More precisely, a shipment from the warehouse to a retailer is not dispatched until the retailer has demanded the complete shipment quantity and it is available as stock-on-hand at the warehouse.

All demands that cannot be served from stock-on-hand are backordered and served on a first-come-first-served basis. As a result, replenishment orders may be delayed at the warehouse due to both stock-outs and delivery restrictions. We assume a fixed transportation time from the warehouse to the retailers; however, the lead times are stochastic because both stock-out delays and delivery delays are stochastic. The warehouse replenishes its stock from an outside supplier with a continuous review (R, nQ) policy and a constant lead time.

In this general OWMR system, we provide an efficient recursive method to evaluate the probability mass function (pmf) of the inventory level at each stock point (the retailers and the central warehouse). The inventory level distributions can be used to assess relevant performance measures, e.g., expected holding costs, expected backorder costs, and/or fill rates. Performance evaluation is, of course, also key for optimizing decision variables such as reorder levels or shipment quantities.

Using different quantity restrictions on shipments has been common in practice for a long time (Jackson, 1985). It is often activities within the material handling that restrict the shipment quantities. Handling physical goods is costly, and using delivery batch quantities that correspond to the size of the load carrier or package may save both time and money; however, there are multiple other relevant scenarios. For instance, to achieve economies of scale in material handling and transportation, the delivery quantity may also be chosen as integer multiples of the preferred load carrier. Another choice could be to let delivery quantities correspond to full truckloads, thus maximizing the utilization of the transports. Compared to most existing literature on OWMR systems, which usually assumes unrestricted partial deliveries, the ability to analyze an inventory system for this class of quantity-restricted shipment policies makes it possible to model more general transportation costs and emission structures with fixed charges.

The model in *Paper II* allows for order quantities at the retailers. A delivery batch quantity equal to the order quantity means that orders are only dispatched in full, i.e., no partial deliveries are allowed. To the best of our knowledge, we are the first to provide an exact analysis of complete deliveries in the type of OWMR system with (R, nQ) policies that we consider. Furthermore, using a delivery quantity of a single unit represents unrestricted partial deliveries. Since we allow for an arbitrary choice of the delivery batch size (both lower, equal, and higher than the order quantity), our proposed method can be used to evaluate both these scenarios that are commonly found in the literature and in practice. However, we may also look at other cases where the delivery quantity and the order quantity differ.

The analysis in the paper generalizes the work of Axsäter (2000), which looks at a similar system under the assumption of unrestricted partial deliveries. This corresponds to a special case of our presented work where the delivery quantity is one.

One key finding in *Paper II* is that the analysis can be performed by modifying some of the probabilities calculated as an intermediate step in Axsäter (2000). We show that, with our modification, the more general case with restrictions on the delivery quantities can be computed with the same computational effort as the sub-problem with unrestricted partial deliveries. As we build upon probabilities from Axsäter (2000), we may benefit from the same efficient recursive method to compute these probabilities.

In the analysis, the positive inventory level (i.e., the stock-on-hand) at the warehouse is divided into *reserved-* and *unreserved stock-on-hand*. *Reserved stock-on-*

hand is the stock-on-hand that a retailer has demanded but where the delivery policy delays the shipment. The delay results from the shipment's having to wait for dispatch until a complete delivery batch quantity is both demanded and available as stock-on-hand at the warehouse. Notably, when excluding the *reserved stock-on-hand*, the inventory level at the warehouse is unaffected by the delivery batch size. As a result, the method from Axsäter (2000) can be applied directly to obtain the corresponding pmf. However, new methods are presented to derive the respective probabilities for both the *reserved stock-on-hand* at the central warehouse and the inventory levels at the retailers.

To facilitate the exposition of the paper, we note that, as the warehouse and the N retailers use fixed order quantities, all units that are moving in the system will do so in multiples of the greatest common divisor (GCD) of all these quantities. We call this entity a sub-batch, and denote the sub-batch quantity with $q = \text{GCD}(Q_0, Q_1, Q_2, \ldots, Q_N)$. Without loss of generality, we may treat the sub-batches as the smallest entity in the system and assume that the delivery batch quantity is chosen as a multiple of this quantity, q.

In the analysis, we consider one retailer at a time, and the method is presented for retailer N in the paper; however, it applies to any retailer. We define q_N^d as the number of sub-batches that are used as a delivery batch quantity for shipments to retailer N.

The analysis determines the inventory level at the time, τ, of an arbitrary customer arrival. As a result of the PASTA[1] property of Poisson arrivals (Wolff, 1982), this is equivalent to finding the stationary inventory level distribution. Moreover, if the inventory position is known, the inventory level distribution at this time can be directly determined from the inventory position and the number of outstanding orders. To analyze the system, we introduce a priority list for the sub-batches. The priority of the sub-batches represents the sequence in which the orders have been placed. The sub-batch with the lowest priority, priority 1, is the sub-batch that was most recently ordered before τ.

If an arbitrary sub-batch with priority n, referred to as "sub-batch n," has been delivered before τ, there can be at most $n-1$ outstanding sub-batches at this time. Consequently, by deriving the probability that "sub-batch n" has been delivered to the corresponding retailer before τ, we can obtain the inventory level distribution for the retailer. We define the following probabilities

[1] Poisson Arrivals See Time Averages

$F(m-(n-1)q|m)$ the probability that “sub-batch n” has been delivered before τ, given $IP_N = m$, when a delivery batch size of one sub-batch is used (i.e., corresponding to the case with unrestricted partial deliveries)

$F_{q_N^d}(m-(n-1)q|m)$ the probability that “sub-batch n” has been delivered before τ, given $IP_N = m$, when a delivery batch size of q_N^d sub-batches is used.

$F(\cdot|\cdot)$ is defined in the exact same way as in Axsäter (2000), while $F_{q_N^d}(\cdot|\cdot)$ are the corresponding probabilities for the case with quantity-based delivery restrictions. A key insight that enables the efficient evaluation technique is that the last unit in the delivery batch (i.e., the unit that was ordered most recently) behaves in the same way as in a system without shipment consolidation. Thus, this last unit behaves exactly as in the model by Axsäter (2000). We use these insights to transform $F(\cdot|\cdot)$ into $F_{q_N^d}(\cdot|\cdot)$.

To perform the transformation, we use the notation, $D(n)$, to represent the number of sub-batches in the same delivery batch as ”sub-batch n,” serving demands with lower priority than the demands served by ”sub-batch n”. We observe that the event $\{$sub-batch n has arrived at $\tau\}$ is equivalent to the event $\{$sub-batch $n-D(n)$ has arrived at $\tau\}$. As the sub-batch with priority $n-D(n)$ triggers the dispatch of the whole shipment, it acts in the same way as in a system without delivery restrictions.

As a result, we may condition on the random variable $D(n)$ to transform $F(\cdot|\cdot)$ into $F_{q_N^d}(\cdot|\cdot)$. Furthermore, to compute the inventory levels at the retailers, we use the relationship that the inventory position equals the inventory level plus the outstanding orders and take the expectation over all possible inventory positions.

A nice result is that, although there are some intrinsic combinatorial complexities to account for when determining $D(n)$, these complications cancel out when deriving the inventory level probabilities. As a result, we arrive at a simple expression, combining the probabilities $F(\cdot|\cdot)$ which in turn are computed following the same recursive procedure as presented in Axsäter (2000). Equation (6.3) gives the pmf for the inventory level at retailer N.

$$Pr(IL_N = j) = \frac{\gamma_N}{q_N^d \cdot Q_N} \sum_{r=max\left(0,\left\lceil \frac{1-j}{q\cdot\gamma_N}\right\rceil\right)}^{\left\lfloor \frac{Q_N-j}{q\cdot\gamma_N}\right\rfloor} \big(F(j|j+r\cdot q) - F(j+q_N^d\cdot q|j+r\cdot q)\big) \quad (6.3)$$

In the above equation, γ_N is a constant given by the greatest common divisor of the order quantity and the delivery batch quantity at retailer N (expressed in sub-batches); computationally, this more general formulation given by 6.3 remains comparable to the special case treated in Axsäter (2000).

In addition to the pmf of the inventory level at the retailers, we derive the pmf of the reserved stock-on-hand at the warehouse. Furthermore, in many applications, both in industry and the literature, the performance measures for a system are based on averages. Thus, for practical applications of our model, we also present a simple formula, given by Equation (6.4), for the expected reserved stock-on-hand for an arbitrary retailer, retailer N.

$$E[W_N] = E[IL_N|q_N^d = 1] - E[IL_N|q_N^d] \tag{6.4}$$

In Equation 6.4, $E[IL_N|q_N^d = 1]$ is the expected inventory level at retailer N when unrestricted partial deliveries are applied, while $E[IL_N|q_N^d]$ is the expected inventory level with quantity-based delivery restrictions.

Numerical tests show that in a multi-echelon setting, retailer reorder points need to reflect delivery batch constraints; neglecting them increases backorder costs, and service levels may suffer. Although the effect is less significant when the warehouse maintains a high service level, an optimized centralized OWMR system usually runs the warehouse with a surprisingly low service level. Therefore, accounting for delivery constraints in multi-echelon inventory control is very important.

6.3 Paper III

Exact Analysis and Control of OWMR Inventory Systems with Joint Order Quantities and Quantity Based Shipment Consolidation

In the third paper, the model from *Paper I* is revisited to find improved policies for more efficient control of the inventory distribution system. We consider an OWMR system where the central warehouse uses an (R, nQ) policy to replenish its inventory from an outside supplier with constant lead time. The retailers face stochastic demand according to independent Poisson processes and are grouped into consolidation groups supplied with joint shipments of a fixed quantity following a fixed milk-run route. Transports from the warehouse to the retailers are made with constant transportation times; however, the lead time (i.e., the time from when an order is placed until the ordered unit arrives) is stochastic due to possible stock-outs at the warehouse and shipment delays when awaiting a full shipment quantity to be ready for dispatch.

The quantity-based shipment consolidation policy ensures that all shipments to a retailer group, e.g., group m, are dispatched with the same quantity of Q_d^m units. Q_d^m can be different for different retailer groups and typically represents the size of a load carrier, a full pallet, etc. Just as in *Paper I*, dispatches to group m take place once Q_d^m units have been demanded by the retailers in the consolidation group and **all** of these units are available at the warehouse (in *Paper I* this is referred to as units being qualified for shipment). The warehouse allocates units to different retailers and retailer groups according to a First-Come-First-Served (FCFS) allocation rule.

FCFS in multi-echelon inventory distribution systems is generally considered a well-performing and fair allocation policy. Moreover, the tractability of FCFS allocation makes it very attractive when designing parameterized policies that can be evaluated, optimized, and implemented in practice.

However, FCFS allocation is clearly not optimal, as the inventory situation may change between the time a demand takes place and the time a unit is ready to be shipped or has arrived at the customer. In particular, in Paper *I*, there can be units qualified for shipment that are not dispatched, as they are waiting for a full shipment to qualify. These units are referred to as *reserved stock-on-hand*, as they are reserved for specific retailers under the FCFS allocation rule. The FCFS allocation of the *reserved stock-on-hand* to different retailer groups has a drawback with a non-negligible impact on the system performance. If several reserved units for different retailer groups are on hand simultaneously, there might already be enough in total to fill a full shipment for one group. Yet, because each unit is locked to a specific group according to FCFS, no dispatch occurs until a group's own reserved stock reaches its full-shipment quantity.

To address the inefficiencies regarding the reserved stock-on-hand in the policy from Paper *I*, Paper *III* adds the possibility for a retailer group to place joint replenishment orders. More precisely, the retailers in group m place an order of Q_m units when Q_m customer demands have accumulated within the group. This means that the FCFS allocation at the warehouse can differ from the sequence in which customer orders arrive in the system. A consequence of this new joint ordering policy is that, for $Q_m = Q_d^m \; \forall \; m$, there will never be reserved stock-on-hand at the warehouse for more than one retailer group at the same time.

In *Paper I*, the retailers act under individual base-stock policies $(S-1, S)$, motivated by a centralized system without any incentives (no costs associated) for placing orders in batches. The motivation for introducing a joint order quantity, Q_m, is not an explicit ordering cost, but to alter how units are prioritized for shipment without losing the computational tractability offered by FCFS. The drawback is that the warehouse cannot act directly on information about demand at the retailers. Setting Q_m is a balance between reducing the reserved stock-on-hand and allowing the warehouse to take advantage of immediate point-of-sale (POS) information. Note that costs and economies of scale in the material handling and shipping are reflected in the delivery quantity, Q_d^m, and the cost for such delivery.

Although the FCFS allocation has been shown to perform reasonably well in practice for many types of systems, including systems with time-based shipment consolidation (Howard, 2013), there may be some apparent disadvantages in particular scenarios. For instance, in Malmberg and Marklund (2023), situations may occur when there are in total enough units available at the warehouse to satisfy outstanding orders with full shipments, but as the units are reserved for different retailer groups, no dispatch can take place.

Of course, the issue can be addressed by reallocating the priority of the outstanding orders among the different retailer groups. However, retaining the FCFS allocation rule's simplicity and tractability from a computational point of view has an advantage. Fairness and its widespread use in practice are also common arguments to motivate its use.

The paper presents an exact, computationally efficient, recursive method to derive the inventory level distributions at each stock point for any combination of shipment quantities, joint replenishment quantities, reorder levels, and order quantities under this new policy.

The presented method allows for joint optimization of the inventory and shipment consolidation decisions for a wide range of performance measures including expected inventory holding and backorder costs, expected transportation costs, and emissions. It also enables evaluation and consideration of constraints on service levels and transport emissions.

The exact approach is based on carefully formulating an equivalent three-echelon model of the two-echelon, OWMR system. This allows us to leverage results from Andersson et al. (2023) and Axsäter (2000) to analyze the system and arrive at the probability mass functions for the following random variables

- unreserved stock-on-hand at the warehouse, defined as IL_0^+,
- reserved stock-on-hand at the warehouse, defined as $W_m \ \forall \ m$,
- the number of outstanding orders for retailer group m, defined as $-IL_m \ \forall \ m$.

Having these in place, we can extend the analysis to incorporate the individual retailers, using binomial disaggregation and the law of total probability. Z_n denotes the number of outstanding units (at the warehouse) for retailer N

$$P(Z_N = z) = \sum_{y=z}^{\infty} Bin(z, y, \lambda_N/\lambda^M) \cdot Pr(-IL^M = y). \tag{6.5}$$

We can then add X_N, i.e., the number of units that are ordered by retailer N during the transportation time (which is independent of Z_N), to get the total number of outstanding orders **at** retailer N. We get the inventory level by the relationship $IL_N = S_N - Z_N - X_N$.

The reformulated *three-echelon* system is illustrated in Figure 6.3 for an example with two retailer groups. The additional *virtual-echelon* with M *virtual warehouses* (one for each retailer group) is placed between the central warehouse and the retailers. The purpose of the *virtual echelon* is to coordinate the unit orders from the retailers and place a joint replenishment order at the warehouse when Q^m unit orders have been accumulated; thus, the *virtual warehouses* are not allowed to hold any stock-on-hand. Moreover, it is sufficient for the shipment consolidation policy to act between the warehouse and the *virtual warehouse*, restricting dispatches to Q_d^m units for *virtual warehouse m*. The transport time between the central warehouse and the *virtual warehouse* is zero. Once the units are available at the *virtual warehouse*, the units are shipped to the respective retailers in the group with individual transportation times reflecting a milk-run delivery route.

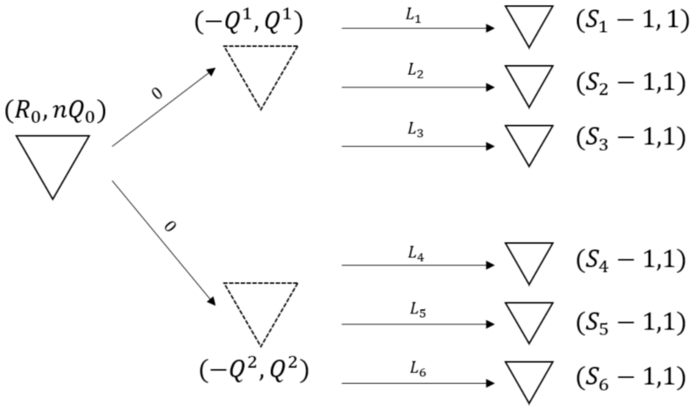

Figure 6.3: A system with six retailers divided into two retailer groups reformulated as an equivalent three-echelon model by inclusion of a *virtual-echelon* consisting of two *virtual-warehouses*. The delivery policy is, in the transformed system, applied between the central warehouse and the *virtual-warehouse*. This means that shipments to *virtual-warehouse* m are made with a delivery policy such that only full shipments of Q_d^m units are made to *virtual-warehouse* m, and the transportation time from the warehouse to the *virtual-warehouse* is always zero.

A numerical study illustrates that allowing joint replenishment orders from the retailer groups can lead to significant cost savings compared to systems where the retailers place independent base stock orders with the warehouse, as in Malmberg and Marklund (2023). The average cost reduction in our study is about 15%. As there are no fixed costs for placing orders in the system, these savings are only due to better coordination of the replenishment, allocation, and shipment decisions in the system, resulting in less reserved stock at the warehouse. Another insight from the study is that setting the joint replenishment quantities equal to the shipment quantities is often the best choice, but not always.

6.4 Paper IV

Managing Inventories in Sustainable Multi-Echelon Distribution Systems with Hybrid Shipment Consolidation

In *Paper IV*, we study a multi-echelon inventory distribution system with a shipment consolidation policy combining quantity-based and time-based dispatches. In line with the existing literature on shipment consolidation in single-echelon systems, we refer to this more general delivery policy as a hybrid shipment consolidation policy, as it can be used to model both pure time-based shipment consolidation, pure quantity-based shipment consolidation, and systems where both time-based and quantity-based shipment consolidation are combined.

We model a One-Warehouse-Multiple-Retailer (OWMR) inventory system where the shipments are consolidated from the warehouse to distinct groups of retailers (similar to the systems in *Paper I* and *Paper III*) using the hybrid policy.

The hybrid policy has fixed time-based transportation schedules for the individual retailer groups. Each time these transports are dispatched, the units that have been demanded by the retailers and are available at the warehouse are shipped on a joint shipment from the warehouse to the retailers. In addition, there is also a possibility to ship additional full shipment quantities (constituting, e.g., a full truckload, a full pallet, or another load carrier) between the periodic shipments if there are enough units available at the warehouse and demanded by the retailers for a full shipment quantity to be dispatched.

Note that this hybrid policy encompasses the standard time-based and quantity-based shipment consolidation policies as special cases. By letting the time between time-based shipments be very long ($T \to \infty$), we get the quantity-based policy, used in *Paper I* and *Paper III.* Similarly, by setting a very high quantity threshold for the quantity-based dispatches ($Q \to \infty$), the policy will follow a time-based policy studied, for example, in Marklund (2011).

To analyze the system, we first derive the probability distributions for the number of outstanding orders at the central warehouse for each retailer group at an arbitrary point in time, τ'. From this probability, we can then calculate the probability for a specific inventory level, IL_i, at the individual retailers for a given base-stock level. Once we have $Pr(IL_i = j)$ for all stock locations i and inventory levels j, we may use these probabilities to obtain performance measures such as expected costs, fill rates, and emissions.

In the following, we focus on providing some insights into the method to derive the distribution of the number of outstanding orders at the warehouse for an arbitrary retailer group (group m), at time τ'.

We start by enumerating all the units that have been ordered prior to τ' in a list such that the unit ordered most recently before τ' is the first unit, the unit ordered most recently prior to the first unit is the second unit, and so on. As we assume a FCFS allocation rule at all stock locations, this list constitutes a *priority list* where the order of the units in the list represents the priority of the units. The higher the priority, the more likely it is that the unit has already been dispatched (i.e., is not outstanding) for retailer group m at time τ'.

Similarly to Papers I, II, and III, we consider an arbitrary unit in this priority list, the n^{th} unit. We know that the probability that this unit has been dispatched before τ' is equal to the probability that there are at most $n-1$ outstanding units at the warehouse for retailer group m, at time τ'. Thus, to obtain the probabilities of interest, it suffices to determine whether the n^{th} unit is dispatched from the warehouse before or after τ'.

We now turn to the necessary conditions for the considered n^{th} unit to be dispatched before τ' given the hybrid shipment consolidation policy. There are two options for shipments. Units can either be shipped on a periodic time-based shipment (TBS), or they can be shipped with a full shipment quantity in a quantity-based shipment (QBS). For a unit to be part of a shipment, the unit must be both available at the warehouse and have been demanded by a retailer in the retailer group when the dispatch occurs. As in Paper I, we refer to available units that have already been demanded as *qualified* for shipment.

Qualified units may be delayed at the warehouse if they are waiting for the next dispatch to occur (either a TBS or a QBS); however, only units that are qualified for shipment can be subject to shipment with one of the two shipment options.

To continue, we define three additional time instances in Table 6.1.

Table 6.1: Important time instances to consider in the analysis

$\tau' - L_0$	the time instance when a unit must have been ordered from the outside supplier in order to be available at the warehouse at time τ'
$\tau' - T_I$	the time instance when the most recent time-based shipment, prior to τ' was dispatched
$\tau' - T_I - L_0$	the time instance when a unit must have been ordered from the outside supplier in order to be available at the warehouse at time $\tau' - T_I$

The time instances defined above are all essential to determine whether the considered n^{th} unit is dispatched before τ' or not. Figure 6.4 and Figure 6.5 illustrate the time instances from Table 6.1 depending on whether $T_I > L_0$ or $T_I \leq L_0$

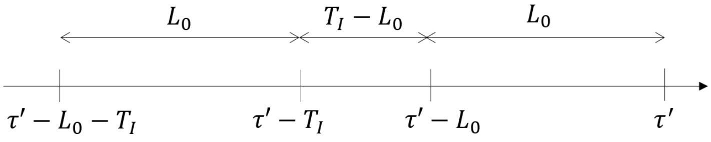

Figure 6.4: Case A: $T_I > L_0$

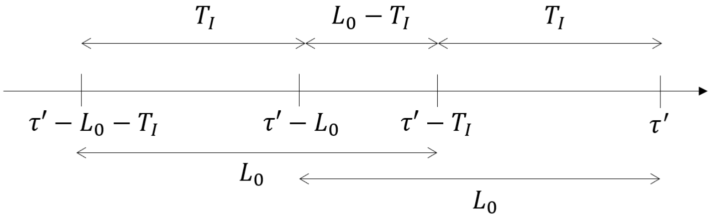

Figure 6.5: Case B: $T_I \leq L_0$

We first note that all units qualified for shipment before $\tau' - T_I$ will be dispatched at or before this time instance, as a time-based dispatch occurs at that time. For this to be the case, for the considered n^{th} unit , the unit must have been ordered by a retailer from the central warehouse before $\tau' - T_I$, and ordered by the warehouse from the outside supplier before $\tau' - T_I - L_0$. We define the event that the considered n^{th} unit has been dispatched from the warehouse before or at $\tau' - T_I$ as *Shipping Option 1*, {ShipOp1}.

Units that are not dispatched with {ShipOp1} can still be dispatched before τ' with a QBS if enough units qualify for shipments between the most recent TBS and τ' to fill a full shipment quantity.

To derive how many additional units need to be qualified for shipment for a QBS, including the considered n^{th} unit , to be dispatched. We first note that it is the *last unit*, i.e., the unit with the lowest priority, among the units included in this QBS, that triggers the dispatch. We derive the probability that this unit is qualified for shipment before τ'. If it is, we also know that the considered n^{th} unit will be part of a QBS dispatched after $\tau' - T_I$ but before τ'. We refer to this event as {ShipOp2}. When looking at the different time instances, we can observe that the sequence differs depending on whether T_I is smaller than or larger than L_0. Two distinct cases, Case A and Case B, are formulated to handle this. In this brief summary, we will avoid going into details regarding the sequence of these time instances; therefore, for the sake of simplicity, we will limit this summary to the case referred to as Case A in the paper, where $\tau' - L_0 \leq \tau' - T_I$. This case is less complicated to analyze compared to Case B.

By definition of the two shipping options we now formulate

$$\begin{aligned} &Pr(\text{the considered considered } n^{th} \text{ unit is dispatched before } \tau') \\ &\qquad\qquad = Pr\left(\{\text{ShipOp1}\} \cup \{\text{ShipOp2}\}\right) \end{aligned} \tag{6.6}$$

The general principle for deriving (6.6) is to compute the joint probability mass function *(pmf)* of the number of demands that take place in each time interval defined by the time instances from above. We distinguish between the demand originating from retailer group m and the other retailer groups.

Moreover, in the joint *pmf*, we also include the probability for a specific number of backorders at the warehouse at time τ', originating from retailer group m, as well as the warehouse inventory position at time $\tau' - L_0 - T_I$. The latter is needed to determine when a unit is available at the warehouse since it influences when the warehouse places the order from the outside supplier.

The joint *pmf* is referred to as $p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, b_0^M, ip\right)$, where $\boldsymbol{\alpha}$ is a vector with the number of demands from retailer group M in each time interval, and $\boldsymbol{\beta}$ is a vector with the number of demands in each time interval from the other retailer groups. Moreover, b_0^M denotes the backorders for group M, and ip is the inventory position at the central warehouse.

The derivation of the joint pmf is relatively straightforward. In principle, we have to compute the *pmf* of the number of Poisson arrivals in disjoint time intervals. However, the challenge is that, since τ' is an arbitrary time instance, the time before τ' when the most recent TBS occurred, i.e., T_I, is a random variable. Since many of the time intervals depend on T_I, there is a joint dependency on T_I for how many arrivals occur in each time interval.

Having the *pmf* in place (it is derived in the Appendix of the paper) we focus on the conditional probability of (6.6), i.e.,

$$Pr\left(\{\text{ShipOp1}\} \cup \{\text{ShipOp2}\} | \boldsymbol{\alpha}, \boldsymbol{\beta}, b_0^M, ip\right). \tag{6.7}$$

As the events are disjoint, we may treat the probabilities for {ShipOp1} and {ShipOp2} separately. Pr\{ShipOp1\} is relatively easy to obtain as it can be directly determined whether this event occurs or not, given the variables that we condition on in (6.7). Thus, the probability will be determined as either one (if it occurs) or zero (if it does not occur) based on the values $\boldsymbol{\alpha}, \boldsymbol{\beta}, b_0^M, ip$.

To exemplify, in order for the considered n^{th} unit to have been demanded before $\tau' - T_I$ (a condition for {ShipOp1}) we need n to be greater than the number of demands for retailer group m in the interval $(\tau' - T_I, \tau')$. Similarly, in order for the considered n^{th} unit to be available before $\tau' - T_I$, the number of backorders for retailer group M at time τ', i.e., b_0^M, cannot exceed the priority of the considered n^{th} unit at time $\tau' - T_I$, i.e., it cannot exceed n minus the demand from retailer group M in $(\tau' - T_I, \tau')$. If both conditions are fulfilled, the probability that {ShipOp1} occurs is one.

The derivation of {ShipOp2} becomes more complex as there are more probabilistic relationships and special cases to consider. The idea is to identify which unit would be the *last unit* in a quantity-based shipment that includes the considered n^{th} unit. We then determine whether this unit is qualified for shipment before or after τ'. Identifying if the last unit has been demanded by a retailer in group M is relatively straightforward; however, determining if it is also available at the warehouse is more challenging since the warehouse ordering process is dependent on all demands in the system and not only the demands at retailer group M.

To deal with this, we derive the probability distribution for how many of the demands from the other retailer groups occur before the demand from retailer group M that will be satisfied with the *last unit* in the QBS. If we know this quantity, and the inventory position at the warehouse at a fixed time instance (in this case ip), we can determine which unit triggered the warehouse order from the outside supplier that will satisfy the demand for the *last unit.* If this order was placed before $\tau' - L_0$, we know that the unit will also be available at τ'.

To give a detailed and precise description, a lot of additional notations and concepts are needed. We, therefore, refer to *Paper IV* for the full derivation.

Once the conditional probabilities from (6.7) and the joint pmf have been obtained for both the case where $T_I \geq L_0$ (referred to as Case A) and $T_I < L_0$ (Case B) we get a final expression for the probability of interest in (6.6) by taking the expectation over all possible values.

$$\begin{aligned}
&Pr(\text{the considered } n^{th} \text{ unit is dispatched before } \tau') = \\
&\quad Pr(\text{Case A}) \sum_{(\boldsymbol{\alpha^A},\boldsymbol{\beta^A},ip,b_0^M)} \big(p_A(\text{ShipOp1}|\boldsymbol{\alpha^A}, b_0^M) \\
&\qquad\qquad + p_A(\text{ShipOp2}|\boldsymbol{\alpha^A}, \boldsymbol{\beta^A}, ip, b_0^M)\big) \cdot p_A(\boldsymbol{\alpha^A}, \boldsymbol{\beta^A}, ip, b_0^M) \\
&\quad + Pr(\text{Case B}) \sum_{(\boldsymbol{\alpha^B},\boldsymbol{\beta^B},ip,b_0^M)} \big(p_B(\text{ShipOp1}|\boldsymbol{\alpha^B}, b_0^M) \\
&\qquad\qquad + p_B(\text{ShipOp2}|\boldsymbol{\alpha^B}, \boldsymbol{\beta^B}, ip, b_0^M)\big) \cdot p_B(\boldsymbol{\alpha^B}, \boldsymbol{\beta^B}, ip, b_0^M)
\end{aligned} \tag{6.8}$$

The exact evaluation of probability mass functions for inventory levels can be used to derive expected costs and transport emissions in the system. In addition, we provide an optimization procedure to optimize reorder levels (R_0 and $\mathbf{S}$), time-based shipment intervals ($\mathbf{T}$), and shipment quantities ($\mathbf{Q}$) in the system. The optimization problem to minimize the total expected cost, TC, is formulated both with and without a side constraint on the total expected transport emissions, TTE.

$$\begin{aligned}
\min \quad & TC(R_0, \mathbf{S}, \mathbf{T}, \mathbf{Q}) \\
s.t. \quad & TTE(R_0, \mathbf{T}, \mathbf{Q}) \leq \text{Emissions target}
\end{aligned} \tag{6.9}$$

A numerical study shows that the hybrid policy can yield significant savings compared to a pure time-based or pure quantity-based shipment consolidation policy. Moreover, analysis of Pareto-optimal combinations of total expected costs and total expected transport emissions shows that our proposed policy can considerably reduce emissions, at the expense of only a small cost increase. In Figure 6.6 on page 60, we illustrate this for one of the examples in the study. The graph provides a scatter plot plotting the expected total cost (TC) and the expected total transport emissions (TTE) for the Pareto-optimal combinations of $R_0, \mathbf{S}^*, \mathbf{T}, \mathbf{Q}$. That is, the graph omits any combination that has a lower cost at the same or higher emissions, or vice versa, lower emissions at the same or higher cost. In this particular example, reducing the transport emissions by 32% from the cost-optimal solution on the far right of the graph, only increases the total expected costs by 1.0%.

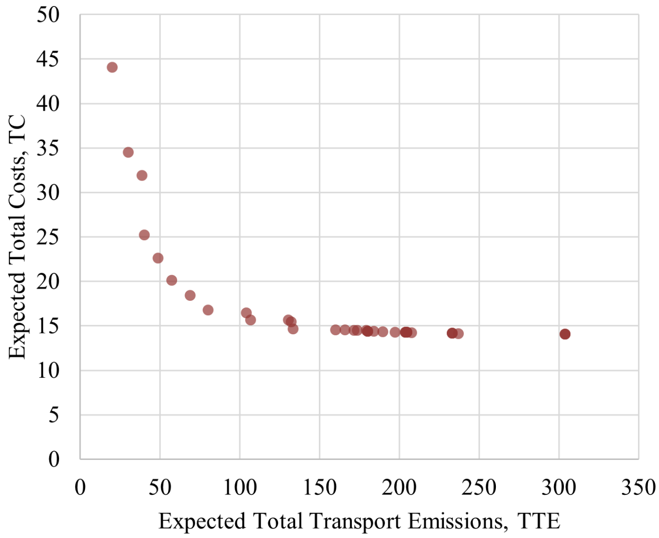

Figure 6.6: The scatter plot contains the Pareto-optimal cost/emissions pairs of $(TC(R_0, \mathbf{S}^*, \mathbf{T}, \mathbf{Q}), TTE(R_0, \mathbf{T}, \mathbf{Q}))$ for Problem 1 in the numerical study. Cost-minimizing base-stock levels, $\mathbf{S}^*$, are used at the retailers as these do not affect the transport emissions. By looking at Pareto-optimal combinations of R_0, $\mathbf{T}$ and $\mathbf{Q}$ with respect to expected costs and emissions, we can find solutions to (6.9) that minimize the total expected costs for a given constraint on the expected total transport emissions. Note that the point on the far right in the plot represents the cost-optimal point, while the point at the top is emissions-optimal. It can be seen that in this particular example, reducing the transport emissions by 32% from the cost-optimal solution only increases the total expected costs by 1.0%.

Chapter 7

Author contribution to each research paper

The work conducted for each paper is divided into seven steps:

1. **Problem identification and modeling assumptions.** Define the managerial problem; set the system scope and modeling assumptions to balance realism with tractability.
2. **Devise a solution strategy.** Select analytical techniques and algorithmic procedures so the model yields accurate results with a practical computational effort.
3. **Analytical and methodological development.** Derive exact mathematical expressions, formulate and prove propositions and lemmas. Find optimization procedures, and (where necessary) heuristics to facilitate optimization.
4. **Software implementation.** Translate the model and mathematical expression into an evaluation and optimization software module.
5. **Verification.** Verify formulas and code by mimicking the model in a discrete-event simulation environment and checking that analytical and simulated results coincide.
6. **Numerical experimentation.** Design computational experiments, run them, and analyze the results to draw managerial insights.
7. **Document mathematical expressions and write the paper.** Draft the manuscript, coordinate and incorporate co-author feedback, and revise the paper in response to peer review.

In Paper I, I contributed to the methodological development and problem-solving by deriving and documenting some proofs. I was responsible for implementing the solution and programming an evaluation and optimization software. I verified the software via discrete-event simulation, conducted numerical experiments, and analyzed the results. Moreover, I wrote the paper together with my supervisor and prepared the LaTeX submission for the journal. I participated in the review process and updated numerical studies and the manuscript, addressing comments and insights from the review team.

In Paper II, I participated in deriving the main solution, some of the proofs, and the mathematical exposition of the solution. I was responsible for implementing the solution in an evaluation and optimization software. I verified the software via discrete-event simulation. I was responsible for the numerical study and worked on designing it, conducting the experiments, and analyzing the results. I participated in writing the paper and incorporated my co-authors' comments to prepare a LaTeX submission to the journal. I participated in the review process and updated the manuscript following the review team's comments.

In Paper III, I conceptualized and identified the problem and determined suitable modeling assumptions. I was responsible for the methodological development, devising the solution strategy, and deriving the mathematical results. I derived the proofs and documented the solution. I implemented the mathematical model in software and performed simulations to verify it in a discrete-event simulation model. As lead author, I drafted most parts of the manuscript.

In Paper IV, I participated in identifying the problem and was responsible for determining detailed model assumptions. I was responsible for the methodological development, devising the solution strategy, deriving the mathematical solutions and proofs, and documenting these in the first draft of the paper. I participated in programming a mathematical software and performed discrete-event simulations to verify the mathematical model and the solution software. I collaborated with my co-authors to design a numerical study and analyze the results. I collaborated closely with my co-authors on drafting the remaining parts of the paper and prepared the final draft for submission by incorporating co-author feedback.

Chapter 8

Main contributions and outline for future work

This thesis explores methods for exact analysis of inventory level distribution in One-Warehouse-Multiple-Retailer (OWMR) inventory distribution systems with quantity-based and hybrid (time-and-quantity-based) shipment consolidation policies, considering different replenishment policies and stochastic demand processes.

In the research conducted for this thesis, we show how to evaluate expected costs and emissions in different systems for given sets of policy parameters and provided optimization procedures allowing for both backorder costs and fill rate constraints, as well as constraints on transport emissions.

The thesis successfully addresses a gap identified in the literature by focusing on quantity-based and hybrid shipment consolidation policies within OWMR inventory distribution systems. Thus, this work complements the existing body of research on time-based consolidation, providing new insights and methodologies to enhance decision-making in these areas. Across the four papers in this thesis, a recurring conclusion can be identified:

> *Jointly considering shipment and inventory decisions is important for achieving both economic and environmental objectives*

Moreover, the research provides a general insight that slight deviations from a "cost-optimal" solution, such as fewer, consolidated shipments, can translate into substantial environmental benefits without sacrificing service levels and with a relatively low impact on total costs. However, to fully capitalize on these benefits, it is essential to take delivery constraints into account when optimizing reorder levels; otherwise, there will be increased shortages, leading to higher costs or failure to meet service targets.

Paper IV constitutes the thesis's most comprehensive modeling contribution. Introducing a hybrid time- and quantity-based shipment consolidation policy generalizes the existing time-based and quantity-based shipment consolidation policies, enabling cost savings and emission reductions compared to the pure policies. Moreover, Paper IV provides a performance guarantee over the time-based and quantity-based shipment consolidation polices, as both are special cases of the hybrid policy. A numerical study investigates the cost performance of the pure time- and quantity-based policies relative to the dominant hybrid policy. The study shows an average cost increase of about 6% for both policies, with a maximum of 18% for the time-based and 28% for the quantity-based policies.

The methods and mathematical models in this thesis can ultimately be developed into decision-support tools to aid companies in improving the design and management of multi-echelon distribution systems and achieve more cost-effective and environmentally sustainable supply chain operations. The methods are most suitable for relatively intermittent demand, as the exact nature of the analysis is computationally intensive, in particular when the demand-to-lead-time ratio is high. This underscores that finding and comparing good approximations and heuristics to deal with the evaluation and optimization of the system is a possible direction for future research, building on the exact methods developed in this thesis.

A natural extension of the papers in this thesis would be to allow for compound Poisson processes at the retailers in Papers I, III, and IV. In general, the challenge with compound Poisson demand in these models is that the disaggregation of backorders cannot be accomplished using a binomial distribution. As a standardized distribution for this problem remains unknown, we need to address these issues differently. The method introduced for Paper IV allows us to track backorders in more detail over time. Interestingly, it appears to be relatively straightforward from a theoretical point of view to generalize this model to compound Poisson demand process at the expense of further increasing the complexity of the derivation and the computational effort. Moreover, as Paper IV generalizes Paper I, such an approach would, in theory, be valid also for Paper I. Note, however, that when using Paper IV to evaluate the delivery policy from Paper I, the computational effort is much higher than if using the method from Paper I directly. Moreover, the model in Paper I would also, when facing compound Poisson demand, have the same challenges related to the reserved stock-on-hand, as addressed in Paper III.

A different direction of future research would be to consider multi-item shipment consolidation. This is something that would have great value for many practical applications where many different types of items are consolidated together on joint shipments via, e.g., a 3PL. It looks promising to use a methodology similar to the procedures presented in Paper IV to address these types of systems. The main challenge is to consider multiple independent central warehouses within the same shipment consolidation program. Another direction for analysis of multi-item systems could be to replace the central warehouse with a production site that manufactures different item types in batches with a joint buffer. Although some challenges from *independent item-types* would disappear, new challenges would naturally arise.

Moreover, finding new policies that can leverage the analytical techniques derived for Paper IV can be of interest in various other contexts. For instance, it looks very promising to use the method in a system with a time-based shipment consolidation policy with other types of decisions and policies; one such example would be to include the possibility of having emergency dispatches to avoid shortages between the *regular* time-based dispatches.

Papers I, III, and IV assume predetermined retailer groups and fixed vehicle routing. While the proposed methods can be applied to evaluate and compare different grouping and routing configurations, the choice of retailer groups and the associated vehicle routes is currently outside the optimization scope. Extending the models to jointly determine inventory levels, shipment consolidation policies, retailer group assignments, and vehicle routing decisions would constitute a valuable direction for future research. Such an extension would bridge the gap to the literature on stochastic inventory–routing problems.

Another extension could be to investigate other types of shipment restrictions from an inventory management perspective. The literature on shipment consolidation has primarily focused on time-based, quantity-based, and hybrid policies; however, emerging technologies such as electric vehicles, advances in autonomous transport, and multi-modal shipment options may motivate alternative delivery restrictions. These developments could change the cost structure, capacity constraints, and scheduling flexibility of transportation resources, which may provide a rationale for other delivery restrictions. Beyond this, future research could explore models where shipment and order parameters adapt dynamically over time, for example, in response to demand fluctuations, transport capacity changes, or the stock situation. It could also be valuable to consider systems that employ different types of trucks or load carriers depending on the situation, enabling additional cost–emission trade-offs and operational efficiencies.

References

Jonas Andersson, Filip Malmberg, and Johan Marklund. Exact analysis of one-warehouse-multiple-retailer inventory systems with quantity restricted deliveries. *European Journal of Operational Research*, 309(3):1161–1172, 2023. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2023.02.026. URL `https://www.sciencedirect.com/science/article/pii/S0377221723001583`.

Sven Axsäter. Simple solution procedures for a class of two-echelon inventory problems. *Operations Research*, 38(1):64–69, 1990.

Sven Axsäter. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Operations research*, 41(4):777–785, 1993a.

Sven Axsäter. Optimization of order-up-to-s policies in two-echelon inventory systems with periodic review. *Naval Research Logistics*, 40(2):245–253, 1993b.

Sven Axsäter. Simple evaluation of echelon stock (r, q) policies for two-level inventory systems. *IIE transactions*, 29(8):661–669, 1997.

Sven Axsäter. Exact analysis of continuous review (r, q) policies in two-echelon inventory systems with compound poisson demand. *Operations research*, 48 (5):686–696, 2000.

Sven Axsäter. A note on stock replenishment and shipment scheduling for vendor-managed inventory systems. *Management Science*, 47(9):1306–1310, 2001.

Sven Axsäter. Supply chain operations: Serial and distribution inventory systems. *Handbooks in operations research and management science*, 11:525–559, 2003.

Sven Axsäter. *Inventory control*, volume 225. Springer, 2015.

Sven Axsäter and Johan Marklund. Optimal position-based warehouse ordering in divergent two-echelon inventory systems. *Operations Research*, 56(4):976–991, 2008.

Sven Axsäter and Johan Marklund. Decision sciences. In *Encyclopedia of Library and Information Sciences*, pages 1450–1457. CRC Press, 2009.

Peter Berling, Lina Johansson, and Johan Marklund. Controlling inventories in omni/multi-channel distribution systems with variable customer order-sizes. *Omega*, 114:102745, 2023.

Yann Bouchery, Asma Ghaffari, Zied Jemai, and Yves Dallery. Including sustainability criteria into inventory models. *European Journal of Operational Research*, 222(2):229–240, 2012.

Sıla Çetinkaya and James H Bookbinder. Stochastic models for the dispatch of consolidated shipments. *Transportation Research Part B: Methodological*, 37 (8):747–768, 2003.

Sıla Çetinkaya and Chung-Yee Lee. Stock replenishment and shipment scheduling for vendor-managed inventory systems. *Management Science*, 46(2):217–232, 2000.

Sıla Çetinkaya, Fatih Mutlu, and Chung-Yee Lee. A comparison of outbound dispatch policies for integrated inventory and transportation decisions. *European Journal of Operational Research*, 171(3):1094–1112, 2006.

Sıla Çetinkaya, Eylem Tekin, and Chung-Yee Lee. A stochastic model for joint inventory and outbound shipment decisions. *IIE Transactions*, 40(3):324–340, 2008.

Fangruo Chen and Yu-Sheng Zheng. One-warehouse multiretailer systems with centralized stock information. *Operations Research*, 45(2):275–287, 1997.

Frank Y Chen, Tong Wang, and Tommy Z Xu. Integrated inventory replenishment and temporal shipment consolidation: A comparison of quantity-based and time-based models. *Annals of Operations Research*, 135(1):197–210, 2005.

Ki Ling Cheung and Hau L Lee. The inventory benefit of shipment coordination and stock rebalancing in a supply chain. *Management Science*, 48(2):300–306, 2002.

Ton de Kok, Christopher Grob, Marco Laumanns, Stefan Minner, Jörg Rambau, and Konrad Schade. A typology and literature review on stochastic multi-echelon inventory models. *European Journal of Operational Research*, 269(3): 955–983, 2018.

Sean Doherty and Seb Hoyle. Supply chain decarbonization: The role of logistics and transport in reducing supply chain carbon emissions. In *World Economic Forum, Geneva*, 2009.

EEA. *Indicator assessment - Load factors for freight transport.* European Environmental Agency, 2021a. URL `https://www.eea.europa.eu/data-and-maps/indicators/load-factors-for-freight-transport/load-factors-for-freight-transport-1`.

EEA. *Indicator assessment - Indicator assessment - Passenger and freight transport demand in Europe.* European Environmental Agency, 2021b. URL `https://www.eea.europa.eu/data-and-maps/indicators/passenger-and-freight-transport-demand/assessment-1`.

EEA. Transport and environment report 2022. digitalisation in the mobility system: Challenges and opportunities. *European Environment Agency*, 2022a. URL `https://www.eea.europa.eu/publications/transport-and-environment-report-2022`.

EEA. Reducing greenhouse gas emissions from heavy-duty vehicles in europe. *European Environment Agency*, 2022b. URL `https://www.eea.europa.eu/publications/co2-emissions-of-new-heavy`.

Awi Federgruen. Centralized planning models for multi-echelon inventory systems under uncertainty. *Handbooks in operations research and management science*, 4:133–173, 1993.

Rolf Forsberg. Optimization of order-up-to-s policies for two-level inventory systems with compound poisson demand. *European Journal of Operational Research*, 81(1):143–153, 1995.

Rolf Forsberg. Exact evaluation of (r, q)-policies for two-level inventory systems with poisson demand. *European Journal of Operational Research*, 96(1):130–138, 1997.

Stephen C Graves. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science*, 31(10):1247–1256, 1985.

Stephen C Graves. A multiechelon inventory model with fixed replenishment intervals. *Management Science*, 42(1):1–18, 1996.

Ford W. Harris. How many parts to make at once. *Operations Research. Vol. 38,No 6 (Nov. Dec. 1990), pp. 947-950. Reprinted from Factory, The Magazine of Management, Volume 10, Number 2*, 1913.

Daniel P. Heyman and Matthew J. Sobel. *Stochastic Models in Operations Research, Volume 1: Stochastic Processes and Operating Characteristics*. Dover Publications, 2004. ISBN 9780486432595.

James K Higginson and James H Bookbinder. Policy recommendations for a shipment-consolidation program. *Journal of Business Logistics*, 15(1):87, 1994.

Frederick S. Hillier and Gerald J. Lieberman. *Introduction to Operations Research*. McGraw Hill Higher Education, 7 edition, 2000. ISBN 9780071181631.

Christian Howard. New allocation policies for divergent inventory systems with real-time information and shipment consolidation. *Howard, Christian., Real-time Allocation Decisions in Multi-echelon Inventory Control, Doctoral Thesis, Div. of Production Management, Dept. of Industrial Management and Logistics, Lund University, Faculty of Engineering*, 2013.

Christian Howard and Johan Marklund. Evaluation of stock allocation policies in a divergent inventory system with shipment consolidation. *European Journal of Operational Research*, 211(2):298–309, 2011.

International Transport Forum. *ITF The Carbon Footprint of Global Trade - Tackling Emissions from International Freight Transport*. OECD, November 2015. URL `https://www.itf-oecd.org/sites/default/files/docs/cop-pdf-06.pdf`.

International Transport Forum. *ITF Transport Outlook 2017*. OECD, 2017. doi: https://doi.org/https://doi.org/10.1787/9789282108000-en. URL `https://www.oecd-ilibrary.org/content/publication/9789282108000-en`.

International Transport Forum. *ITF Transport Outlook 2023*. OECD, 2023. doi: https://doi.org/https://doi.org/10.1787/b6cc9ad5-en. URL `https://www.oecd-ilibrary.org/content/publication/b6cc9ad5-en`.

George C Jackson. A survey of freight consolidation practices. *Journal of Business Logistics*, 6(1), 1985.

George C Jackson, Jeffrey J Stoltman, and Audrey Taylor. Moving beyond trade-offs. *International Journal of Physical Distribution & Logistics Management*, 1994.

Lina Johansson, Danja R. Sonntag, Johan Marklund, and Gudrun P. Kiesmüller. Controlling distribution inventory systems with shipment consolidation and compound poisson demand. *European Journal of Operational Research*, 2019. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2019.06.045.

Gudrun P. Kiesmüller and Ton de Kok. *A multi-item multi-echelon inventory system with quantity-based order consolidation.* Beta, Research School for Operations Management and Logistics, 2005.

Filip Malmberg. *Quantity-Based Shipment Consolidation and Delivery Policies in Multi-Echelon Inventory Control.* Lund University, 2022. [Thesis for the degree of Licentiate in Engineering].

Filip Malmberg and Johan Marklund. Evaluation and control of inventory distribution systems with quantity based shipment consolidation. *Naval Research Logistics (NRL)*, 70(2):205–227, 2023. doi: https://doi.org/10.1002/nav.22090. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.22090`.

Johan Marklund. Centralized inventory control in a two-level distribution system with poisson demand. *Naval Research Logistics*, 49(8):798–822, 2002.

Johan Marklund. Controlling inventories in divergent supply chains with advance-order information. *Operations Research*, 54(5):988–1010, 2006.

Johan Marklund. Inventory control in divergent supply chains with time-based dispatching and shipment consolidation. *Naval Research Logistics (NRL)*, 58 (1):59–71, 2011.

Johan Marklund and Peter Berling. Green inventory management. In Yann Bouchery, Charles J Corbett, Jan C Franso, and Tarkan Tan, editors, *Sustainable Supply Chains*, pages 143–177. Springer, 2 edition, 2024. ISBN 978-3-031-45564-3. doi: 10.1007/978-3-031-45565-0_6.

Kamran Moinzadeh. A multi-echelon inventory system with information exchange. *Management Science*, 48(3):414–426, 2002.

Philip M. Morse, George E. Kimball, and Patrick M.S. Blackett. Methods of operations research. *Physics Today*, 4(11):18–20, 1951.

Philip M. Morse, George E. Kimball, and Saul I. Gass. *Methods of operations research.* Courier Corporation, 2003.

Fatih Mutlu, Sıla Çetinkaya, and James H Bookbinder. An analytical model for computing the optimal time-and-quantity-based policy for consolidated shipments. *IIE Transactions*, 42(5):367–377, 2010.

Daniel R. Pinchasik, Inger B. Hovi, Inge Vierth, Anna Mellin, Heikki Liimatainen, and Niels B. Kristensen. Reducing CO2 emissions from freight. *TemaNord*, January 2019. doi: 10.6027/tn2018-554. URL `https://norden.diva-portal.org/smash/get/diva2:1277299/FULLTEXT01.pdf`.

Jana Ralfs and Gudrun P Kiesmüller. Inventory management with advance demand information and flexible shipment consolidation. *OR Spectrum*, 44 (4):1009–1044, 2022.

Benhür Satır, Fatih Safa Erenay, and James H Bookbinder. Shipment consolidation with two demand classes: Rationing the dispatch capacity. *European Journal of Operational Research*, 270(1):171–184, 2018.

Zhijie Shang, Kevin H .and Tao and Sean X. Zhou. Optimizing reorder intervals for two-echelon distribution systems with stochastic demand. *Operations Research*, 63(2):458–475, 2015.

Edward R. Silver, David F. Pyke, and Douglas J. Thomas. *Inventory and Production Management in Supply Chains, Fourth Edition.* Taylor & Francis, 2016. ISBN 9781466558618.

Richard Macey Simon. Stationary properties of a two-echelon inventory model for low demand items. *Operations Research*, 19(3):761–773, 1971.

Olof Stenius. *Exact Methods for Multi-echelon Inventory Control: Incorporating Shipment Decisions and Detailed Demand Information.* Doctoral thesis (compilation), Lund univiersty, Division of Production Management, 2016.

Olof Stenius, Ayşe Gönül Karaarslan, Johan Marklund, and Ton de Kok. Exact analysis of divergent inventory systems with time-based shipment consolidation and compound poisson demand. *Operations research*, 64(4):906–921, 2016.

Olof Stenius, Johan Marklund, and Sven Axsäter. Sustainable multi-echelon inventory control with shipment consolidation and volume dependent freight costs. *European Journal of Operational Research*, 267(3):904–916, 2018.

Jacob Teter, Pierpaolo Cazzola, Timur Gül, and E Mulland. The future of trucks. *International Energy Agency*, 1(1):1–166, 2017.

Ayşegül Toptal, Sıla Çetinkaya, and Chung-Yee Lee. The buyer-vendor coordination problem: modeling inbound and outbound cargo capacity and costs. *IIE transactions*, 35(11):987–1002, 2003.

Josué C. Velázquez-Martínez and Jan C. Fransoo. *Green Network Design and Facility Location*, pages 179–194. Springer International Publishing, Cham, 2024. ISBN 978-3-031-45565-0. doi: 10.1007/978-3-031-45565-0_7. URL `https://doi.org/10.1007/978-3-031-45565-0_7`.

Jessica Wehner. Energy efficiency in logistics: An interactive approach to capacity utilisation. *Sustainability*, 10(6):1727, 2018.

Bo Wei, Sıla Çetinkaya, and Daren BH Cline. Inbound replenishment and outbound dispatch decisions under hybrid shipment consolidation policies: An analytical model and comparison. *Transportation Research Part E: Logistics and Transportation Review*, 175:103135, 2023.

Ronald W Wolff. Poisson arrivals see time averages. *Operations Research*, 30 (2):223–231, 1982.

Scientific papers

Paper I: Evaluation and Control of Inventory Distribution Systems with Quantity Based Shipment Consolidation

Naval Research Logistics (NRL), 2023, 70(2), 205-227.

The paper has received certified awards for being among the journal's top 10 most-cited papers compared to articles published between 1st January 2022 – 31st December 2023 and ranking in the top 10% of the most downloaded papers published between 1st January 2022-31st December 2022.

Paper II: Exact Analysis of One-Warehouse-Multiple-Retailer Inventory Systems with Quantity Restricted Deliveries

European Journal of Operational Research (EJOR), 2023, 309(3), 1161-1172

Paper III: Exact Analysis and Control of OWMR Inventory Systems with Joint Order Quantities and Quantity Based Shipment Consolidation

(Submitted 2025)

Paper IV: Managing Inventories in Sustainable Multi-Echelon Distribution Systems with Hybrid Shipment Consolidation

(Submitted 2025)

All papers are reproduced with the permission of their respective publishers.

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

Lund University

Division of Production Management
Industrial and Mechanical Sciences
Faculty of Engineering

ISBN 978-91-8104-614-4
ISRN LUTMDN/TMIO-1015-SE

9 789181 046144