Dynamic performance of hierarchical planning systems : modeling and evaluation with dynamic planned lead times

Citation for published version (APA):
Selcuk, B. (2007). *Dynamic performance of hierarchical planning systems : modeling and evaluation with dynamic planned lead times*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR621899

DOI:
10.6100/IR621899

Document status and date:
Published: 01/01/2007

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Dynamic Performance of Hierarchical Planning Systems:

Modeling and Evaluation with Dynamic Planned Lead Times

Barış Selçuk

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Selçuk, Barış

Dynamic performance of hierarchical planning systems: *Modeling and evaluation with dynamic planned lead times* / by Barış Selçuk. – Eindhoven : Technische Universiteit Eindhoven, 2007. – Proefschrift.

ISBN 90-386-0864-0

ISBN 978-90-386-0864-8

NUR 804

Keywords: Hierarchical Planning / Lead Time / Supply Chain / Order Release / Markov Process / Clearing Function

Printed by PrintPartners Ipskamp, Enschede, The Netherlands

Cover designed by Paul Verspaget

Dynamic Performance of Hierarchical Planning Systems:

Modeling and Evaluation with Dynamic Planned Lead Times

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op donderdag 22 februari 2007 om 16.00 uur

door

Barış Selçuk

geboren te Niğde, Turkije

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. A.G. de Kok
en
prof.dr.ir. J.C. Fransoo

Acknowledgements

This thesis is an outcome of research activities that I have conducted jointly with my promoters for the last four years. In addition to hard work and devotion for research, I have had the privilege to live in an international and attractive social environment. I would like to thank all academic and non-academic people who have helped and supported me in some way during this long and challenging journey of my PhD study. In particular, some people deserve special thanks.

Prof.dr.ir. Jan Fransoo has not only been a careful supervisor and a good teacher but also, has become a good friend and a trustful colleague to me. He has indispensable contributions to this thesis. Especially, in defining the problem context, describing the lead time syndrome, and developing new ideas about the concept of clearing function and dynamic lead times, I have learned many inspiring concepts and useful methods from him. Each time I was puzzled with complex models and confusing too many results, Jan helped me find the right direction with his clear descriptions and professional guidance.

I would like to express my thankfulness to prof.dr. Ton De Kok who has supported me in patience and has made this PhD project possible. Our lively and fruitful discussions have opened up new and interesting research subjects, some of which have already been included in this thesis. Ton has had valuable comments in improving the scientific contributions and the practical relevance of this thesis.

I am indebted to dr.ir. Ivo Adan for his directions in developing an exact model for the lead time syndrome and introducing me to matrix geometric methods. I am also thankful to Ivo for being a member of my dissertation committee. His involvement has not only improved the analytical rigorousness in the thesis but also helped me deepen my insight into various topics.

I would like to thank prof.dr.ir. Will Bertrand for being a passionate advocate of system dynamics and showing me that there are various approaches in modeling a problem. His ideas and suggestions have broadened my research horizon and will have effects in shaping my future research directions. I also thank him for accepting to be a volunteer member of my dissertation committee.

In addition, I would like to express my gratitude to prof.dr. Hubert Missbauer and to prof.dr.ir. Gerard Gaalman who kindly accepted to be in my dissertation committee. Their helpful comments and involvement have improved both the content and the presentation of the thesis.

I have always been a proud and happy member of OPAC. It has been a joyful and fruitful working environment with excellent researchers and PhD students, and very helpful department secretaries at OPAC. I thank all the current and former colleagues in the group including Geert-Jan van Houtum, Marco Slikker, Nico Dellaert, Gudrun Kiesmuller, Simme Douwe Flapper, Bram Kranenburg, Pim Ouwehand, Marko Jakšič, Erik Winands, Gergely Mincsovics, Youssef Boulaksil, Koos Huibers, and Ineke Verbakel-Klok.

I have met very pleasant people and made many good friends for the last four years in Eindhoven. I want to thank Mustafa Kemal Doğru for discussing about new ideas and for being a very good friend, and Ulaş Özen for "collaborating" in almost everything including sharing an office room. I also want to mention the names of Barış Pazarbaşı, Okan Oyman, Kanat Çamlıbel, Sandra Balkestein Tan, Tarkan Tan, Ege Ünal Şerbetli, Semih Şerbetli, Kurtuluş Barış Öner, Nesim Erkip, and all the members of Los Rusticos football team who have contributed their support and friendship in an appreciated way.

I want to express my gratitude to those who have long-distance but vital support. I thank my family for being as they are and for their unconditional love. I would not be able to achieve success in many things in my life, including this thesis, without their wholeheartedly support. I also thank the Altaş family. I always felt their faith, sincerity, and understanding as driving forces.

Finally, I want to express my special thanks to a very special person, Aslı, my wife, my mainstay in life, my love. Thank you for your patience, understanding, self-sacrificing support, positive energy, charming smile that makes me forget about all difficulties, and all the things you have done for us that I fail to mention here. Last but not least, thank you for loving me.

Barış Selçuk,

February 2007

Contents

Chapter 1

Introduction: Concepts and Models

In this chapter, we set the stage for the remainder of the thesis. We introduce the concepts and the models that form the basis for our contributions in this thesis. Three different concepts and associated streams of literature can be related to the content of this thesis. These are hierarchical planning systems, lead time management, and clearing function. Within the context of hierarchical planning systems we review and discuss three main approaches: (1) Hierarchical production planning of Hax and Meal (1975), (2) Goods flow and production unit control structure of Bertrand *et al.* (1990), and (3) Organizational planning hierarchies of Schneeweiss (1999). The research on lead time management is discussed from various perspectives including due-date assignment, planned lead times for order release planning, and workload control. Additionally, we provide a detailed overview of the initial studies and the recent improvements on the concept of clearing function. We discuss why clearing functions are important in modeling the dynamic flow times at an aggregate planning level. We then introduce our research questions and a brief description of our research methodology.

1.1 Problem Context

During the past few decades the concept of *Supply Chain Management* (SCM) has become increasingly popular. The drastic changes in market structures, geographical diversity in manufacturing and distribution of goods, and evolutionary developments in information technology (IT) have helped the notion of supply chain become the center of attraction in industry as well as in academia. Simchi-Levi *et al.* (2003) defines SCM as "a set of approaches utilized to efficiently integrate suppliers, manufacturers, warehouses, and stores, so that merchandise is produced and distributed at the right quantities, to the right locations, and at the right time, in order to minimize systemwide costs while satisfying service level requirements". As this definition implies, SCM is a very broad topic. The managerial issues described in this context can be classified into three categories in terms of the scope and the type of problems considered. These are:

Supply Chain Design: It refers to long-term strategic decisions such as the location of factories and warehouses, transportation infrastructure, supplier selection, etc. Supply chain design activities involve choosing what capabilities along the value chain to invest in and develop internally and which to allocate for development by external parties.

Supply Chain Coordination: It refers to the collaboration of mutually independent parties along the supply chain to improve the coordination of their activities as an attempt to increase joint benefits. Supply chain coordination involves design of contracts between suppliers and buyers, as well as the information that is exchanged between them. Collaboration focuses on joint planning, coordination, and process integration between suppliers, customers, and other partners in a supply chain.

Supply Chain Operations: It addresses the problem of matching supply and demand in terms of both volume and product-mix in the mid-term. The focus is on allocating material and production resources through time and within supply chain in order to meet customer needs. This means planning for the right quantities of material to arrive at the right time and place to support production and distribution. It also means maintaining appropriate levels of raw material, work in process (WIP), and finished goods inventories in the correct locations to meet market needs. In the short-term, detailed scheduling of resources is required to meet production requirements. As the day-to-day activities continue, the planning and control system must track the use of resources and execution results to report on material consumption, capacity utilization, completion of customer orders, and other important measures of performance.

In this thesis, we focus our attention on planning problems related to supply chain operations because, most of the reviewing and updating activities are done at this level of planning in the supply chain. Usually, design and coordination choices are rarely revised, and given these choices, the short to mid-term demand or production variations are handled by appropriate revisions in planning supply chain operations. The capability to recognize changes in customer requirements and move them through the value chain is an important dimension of a planning and control system. This requires the ability to monitor system statuses, determine, transmit, revise, and coordinate requirements throughout the supply chain in a dynamic framework.

One response to increasing need for coordination and communication has been the rapid deployment of IT applications, particularly *Enterprise Resource Planning* (ERP) systems. The promise of ERP systems is to provide real-time data availability for coordinated decision making in globally dispersed organizations. These systems are transactional, IT backbone systems that also support various decision-making processes, such as inventory management, production planning, forecasting, etc.

During the seventies and eighties OR applications led to the implementation of tailor-made *Decision Support Systems* (DSS) for supply chains. The required inputs were downloaded from ERP systems and the outputs were uploaded again, either manually or using an IT interface. However, these DSS never raised the same interest with top management as ERP systems. DSS are widely spread across all business functions in particular as homemade spreadsheet programs. The lack of attention of top management with respect to DSS changed fundamentally in the early nineties when the notion of DSS was replaced by the notion of *Advanced Planning Systems* (APS). One of the reasons behind this development was, APS were introduced as standard software applications that provide an integration of different decision making processes among various business functions using real-time data from ERP systems. The structural framework for APS is in general established along the principles of hierarchical planning approach. Thus, it is a relevant topic to investigate the performance of these systems in a dynamic framework. *Hierarchical Planning* (HP) concepts have been developed and widely accepted both in industry and in academia as a management philosophy to decompose a complex planning problem into small and manageable subproblems while considering their interdependencies and coordinating their decision outcomes.

It is very complicated to grasp all the interactions and dependencies between different decision functions in the planning hierarchy and between different stages in the supply chain considering all aspects of production processes, market structure, product-mix, capacity allocation and etc. Thus, we resort to a certain level of detail in modeling supply chain operations, which is made explicit in the following section.

1.2 System under Study

We consider a supply chain that is composed of several production phases and stock points in a serial structure. An illustration for a two-stage system is given in Figure 1.1. Each item produced within the supply chain is kept in a certain stock point either to be used in further production steps or to satisfy end-customer demand, which is stochastic in nature. The term *production unit* refers to a production department, "which on the short term is self-contained with respect to the use of its resources, and which is responsible for the production of a specific set of products (the *production unit end-items*) from a specific set of materials and components (the *production unit start-items*" (Bertrand *et al.* (1990)). We assume the flows of materials between the production units and their stock points occur in batches, and there is an ample supply of raw materials at the most upstream stage of the supply chain.

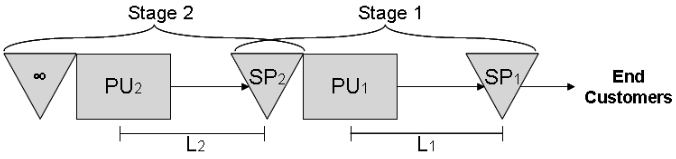


Figure 1.1: A two-stage serial supply chain.

Each stage of the supply chain consists of a production unit and two stock points for the inventory of production unit start-items and of production unit end-items respectively. The coupling between the downstream and upstream stages is established by the inventory of intermediate items which are start-items for a downstream stage (Stage 1 in Figure 1.1) and end-items for an upstream stage (Stage 2 in Figure 1.1). For example, in Figure 1.1, PU_1 produces to final product stock point SP_1 from the intermediate item in stock point SP_2, which is produced by production unit PU_2. Production units are not necessarily situated in the same facility, and in fact, they may refer to different, geographically dispersed business units. The complete process that an item experiences in its production unit is considered at an aggregate level such that it is represented by a single stochastic process. The processing time of a single production unit end-item is assumed to follow a stationary probability distribution. In addition, production unit capacities are not flexible, and the maximum long-term average throughput level of each production unit is fixed to a nominal capacity.

There is a positive duration of time between the moment that an order is released to its production unit and that order is available in its stock point. In practical situations, this duration is not fixed and depends on various important spatial or temporal system characteristics such as the level of process uncertainty, the workload in the production unit and the size of the released order. In this thesis, we define the term that refers to this duration of time as *order flow time* or in short *flow time*.

In planning the release of orders, the flow time is represented by a parameter, which has to be determined in advance of release decisions. This parameter is referred to as *planned lead time* or in short *lead time* throughout the thesis. In managing the supply chain of Figure 1.1, L_2 refers to the planned duration of time to produce a batch of intermediate items in PU_2 and put into stock point SP_2. Similarly, L_1 is the planned duration of time to produce a batch of final products in PU_1 and put into stock point SP_1 to satisfy end-customer demand.

Forecasted information is available for the future end-customer demand which might change during time. The actual demand may be realized differently from its forecasted value. The goal is to satisfy demand by keeping the total costs as low as possible. Costs are generated by keeping finished items in stock including the safety stock against demand and production uncertainties, and also by keeping unfinished items as workload in the production units.

The major planning tasks in a supply chain as we have described here include setting appropriate lead times, deciding on the optimal stock levels together with the optimal usage of materials and resources, and detailed decisions on scheduling of orders released at each stage. Considering all these aspects in a single decision model is not feasible due to huge data processing requirements, even for the simplest case. An HP model is provided by separately considering lead time setting, order release planning, and order scheduling decisions. In the following, we describe the role of each decision function into further detail, and also provide a dynamic framework of planning hierarchies.

Planning Hierarchy

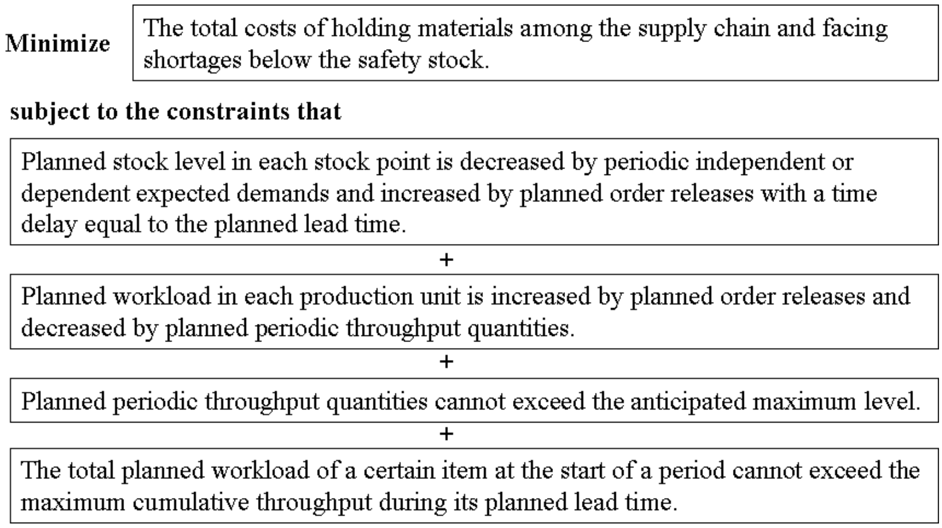
The planning hierarchy modeled in this thesis consists of three separate decision levels: *tactical planning level*, *operational planning level*, and *operational scheduling level*. The lead times are determined at the tactical planning level. This decision function is necessary to instruct the lower level decision models about the time it takes for an order to traverse a production unit. The lead times, exogenous to the operational planning and scheduling levels, are used in releasing orders and balancing inventory for the expected demand, and in determining the due-dates of the released orders. The lead times are set as integer multiples of a period. A period is a prespecified duration of time for which basic input/output rates (i.e., demand and production rates) are defined for all stages of the supply chain. This implies both the operational planning and the operational scheduling levels use identical periods in their models. The method of rolling horizons is applied in order to account for uncertainties in the production and demand processes, and to update planned lead times at each replanning opportunity.

The planned lead times may be determined in a decentralized manner considering each stage independently. This refers to the case that the planned lead times of the items produced in a certain production unit depend only on the information status of that production unit only. Setting planned lead times based on the workload of the production unit or based on the recent occurrences of flow times in that production unit are some examples of this approach. The planned lead times may also be determined in a centralized manner considering the interaction between successive stages.

The operational planning level is modeled within the context of *Supply Chain Operations Planning* (SCOP) as defined in De Kok and Fransoo (2003). Thus, the term *SCOP level* is used as an alternative to the term operational planning level in many places throughout the thesis. "The objective of SCOP is to coordinate the release of materials and resources in the supply network under consideration such that customer service constraints are met at minimal cost" (De Kok and Fransoo (2003)). In accordance to its formal definition, various SCOP models are considered in this thesis. Basically, our SCOP models are *Mathematical Programming* (MP) formulations, where the planning is done periodically for a specified number of periods in the planning horizon. Figure 1.2 provides a meta-model of our SCOP formulation.

In accordance with the rolling horizons method, only the first period's planning decisions are instructed to the operational scheduling level. At this level, detailed, execution related decisions are made in a decentralized manner for each stage of the supply chain. The concern is no longer the planning of material flow between stages, but the scheduling of released orders at each stage and planning the usage of materials at each production unit.

As time proceeds, the result of the execution together with the effects of random events that occurred during the previous period is feedback to the planning system, and the previous plans together with the planned lead times are revised based on the changed information status. Although the method of rolling horizons may result in suboptimal decision making, it is very rele-



Minimize The total costs of holding materials among the supply chain and facing shortages below the safety stock.

subject to the constraints that

Planned stock level in each stock point is decreased by periodic independent or dependent expected demands and increased by planned order releases with a time delay equal to the planned lead time.

+

Planned workload in each production unit is increased by planned order releases and decreased by planned periodic throughput quantities.

+

Planned periodic throughput quantities cannot exceed the anticipated maximum level.

+

The total planned workload of a certain item at the start of a period cannot exceed the maximum cumulative throughput during its planned lead time.

Figure 1.2: The SCOP meta-model.

vant from a real life perspective. It is a common planning strategy employed in various ways within commercial planning and scheduling software mainly utilizing *Material Requirements Planning* (MRP) or *Manufacturing Resources Planning* (MRPII) concepts (cf. Vollmann *et al.* (1997)). The model for decision making process is different in each chapter of the thesis, and is therefore described separately where appropriate. However, the *plan-execute-feedback-(re)plan* cycle is common in all chapters, and is used at different decision levels for updating purposes. This approach leads us to a dynamic framework for hierarchical planning systems, as illustrated in Figure 1.3. It is based on a simple hierarchical structure of two decision levels (top level instructing the bottom level), and is also applicable for general structures. *Dynamic* refers to the fact that the planning process is described as a series of decisions taken in consecutive planning cycles in time, and planning parameters (e.g., planned lead times) are subject to change during the course of time. Accordingly, the performance evaluation is conducted in a dynamic setting. The performance of the planning decisions given for a single problem instance is not only evaluated based on the available information at the time of decision, but based on their effects on the actual system status changing through time, which is unknown at the time of decision.

In Figure 1.3, the parameters of each decision model at each level of the planning hierarchy are subject to changes between the replanning instances t

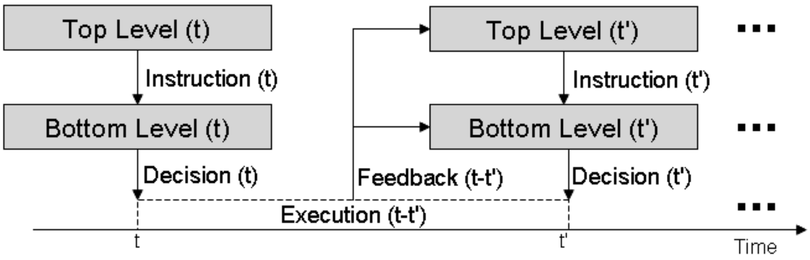


Figure 1.3: Dynamic framework of hierarchical planning.

and t' with $t' > t$. The level of change in a certain parameter depends on the initial system status at time t and on the response to the events that have occurred during period $t - t'$. For example, we may consider updating the parameters for capacity or processing times at an aggregate planning level in response to machine failures in the shop floor.

In this thesis, we aim to shed light on the evaluation of this dynamic framework through scientific results gathered from experimental and analytical studies. Specifically, *we want to provide insights into the performance of hierarchical planning systems in coordinating supply chain operations with dynamic lead times.* The performance is measured along two perspectives:

- *External Performance*: Expressed in terms of average (periodic) costs such that a predefined customer service level is met.
- *Internal Performance*: Expressed in terms of the level of consistency between the higher and lower level decision functions. In particular, it refers to the deviation of the actual delivery dates of the released orders from their planned delivery dates.

The external performance measures are typical ways to evaluate the system in relation to the cost minimization or profit maximization objectives. The internal performance measures are relatively less visible, but at the same time, they are interesting and relevant in evaluating the level of integration between the coordination decisions made at a higher planning level and the execution decisions made at a lower planning level. In many respects, the internal performance measures are relevant tools in understanding the reasons behind the changes in the external performance.

1.3 Related Literature

Three different lines of research can be distinguished within the context of this thesis. These are hierarchical planning in relation to the planning framework presented in the previous section, lead time management in relation to updating the planned lead times, and clearing function in relation to modeling the stochastic production processes at an aggregate planning level. They are reviewed separately in the following sections.

1.3.1 Hierarchical Planning

Hierarchical planning has been a predominant mode for production planning both in academic research and in industrial practice. It is a management philosophy that is based on the decomposition of a large complex planning problem into small and manageable sub-problems. The decision making process in industrial organizations, in general, is modeled as a network of sub-processes with a hierarchically coordinated flow of information in between the processes. From a real-life perspective, it is very relevant to understand and analyze the industrial problems together with their hierarchical interactions.

A vast amount of research has been produced in the context of hierarchical planning. Among these, Anthony (1965) provided the first comprehensive view on the hierarchical framework for organizational decision making with generic descriptions of strategic, tactical and operational level planning problems. This three-level hierarchical framework has been cited as *Anthony's Taxonomy* by many researchers. The computational efficiency introduced by the hierarchical decomposition approach has promoted the use of management science tools for planning large complex organizations. Inspired by the decomposition algorithm of Dantzig and Wolfe (1963) for solving linear programming problems, Ruefli (1971) proposed a generalized goal decomposition model to represent decision making in a three-level hierarchical organization. Similar to Anthony's Taxonomy, these decision levels are identified as central unit, management unit, and operating unit. The central unit coordinates lower level management units by setting goals. Each management unit solves a resource allocation problem based on goals, and provides feedbacks of shadow prices to the central unit so that the latter can evaluate and improve its goal setting policies. Operating units are responsible for generating alternative activity levels for a given resource allocation, and feeds them back to the management units to re-evaluate their decisions. Different from the previous decomposition models, Ruefli's formulation includes the effects of organization structure on the solution, thus, extends the application area of decomposition techniques to include industrial problems.

Although the studies of Anthony and Ruefli emphasized the organizational and algorithmic insights behind the hierarchical approach, Hax and Meal (1975) are the first to give a formal hierarchical planning model and to implement it in a complex production planning and scheduling situation. Thus, the term *Hierarchical Production Planning* (HPP) is generally attributed to their modeling approach. A three level product structure is introduced, where *item* refers to an individual stock keeping unit, and items sharing common tooling and setup characteristics are aggregated to the same *family*, and families sharing common production rates and inventory costs are grouped into the same *type*. Aggregate plans are generated at the product type level, and based on the seasonal stock accumulation of each product type, family run quantities are determined for a shorter time frame, and lastly, the family production run quantities are allocated among the items by equalizing run-out times of individual items. The basic benefits of such a product structure are that demand forecasts are more accurate at the aggregate level and the joint scheduling of items sharing the common setup generates smoother production with fewer disruptions. In addition, there is the reduction in computational and data gathering time introduced by the hierarchical planning methodology. Following the analytical framework of Hax and Meal (1975), the majority of the studies in this area are concentrated on the design of decision models at different levels and developing *perfect* aggregation-disaggregation algorithms. Perfect aggregation refers to the cases where any aggregate plan on product groups and machine groups can be disaggregated to a detailed plan on individual items and machines (Axsäter (1981)). In Bitran and Hax (1977), the HPP methodology of Hax and Meal (1975) is further developed by introducing the concept of *effective demands* to satisfy the feasibility of aggregate plans. The approach in this paper is referred to as *Regular Knapsack Method* (RKM), due to the fact that the family and item disaggregation subsystems are both represented by means of knapsack problems. Various exact and heuristic techniques on the disaggregation problems have been developed in the HPP literature. Examples include Graves (1982) using a lagrangean heuristic, Ari and Axsäter (1988) using dynamic programming, and Leong *et al.* (1989) using a weighted goal programming approach.

Bertrand *et al.* (1990) developed a HP framework by separating out semi-autonomous production units in a multi-stage production-inventory system, and introducing the concept of *Goods Flow Control* (GFC) function. Employing a centralized approach over the entire supply chain, this decision function is responsible for the mutual coordination of the outputs of production units as well as on-time demand satisfaction. Given the operating targets set by the GFC function, each production unit is then modeled separately. The behavior of a production unit as seen from a goods flow perspective is described by the so called *operational characteristics*, which we specify by the concept of clearing function in this thesis. The concepts used in this approach are partly based on certain concepts from MRPII (cf. Bertrand and Wijngaard (1985)), and have led to a considerable amount of academic research (e.g., Fransoo *et al.* (1995), Zäpfel (1996), Raaymakers *et al.* (2000), and Negenman (2000)). The framework is further enhanced by De Kok and Fransoo (2003) in a more detailed way by including the concept of *effectuation lead times.* The effectuation lead time is defined by the duration of time it takes to implement a decision. In our case, the effectuation lead time is represented by the order flow time; any released order can be effectuated at the end of its flow time.

More recently, Schneeweiss (1999) introduced the concept of *organizational planning hierarchies* by emphasizing two important issues: information asymmetry and goal asymmetry. Information asymmetry refers to the assumption that different hierarchical levels possess different information states, and goal asymmetry refers to the case that different hierarchical levels have different, even conflicting, objectives. De Kok and Fransoo (2003) raises the discussion that the existence of effectuation lead times is the primary cause of information asymmetry. The concept of *anticipation* has been developed explicitly in order to take into account the possible outcomes of the influences between different hierarchical levels (cf. Schneeweiss and Schröder (1992), and Schneeweiss (1995)). Two main types of anticipation can be distinguished: reactive and non-reactive anticipation. The reactive anticipation considers a possible reaction of the bottom level to the top level's instructions, whereas the non-reactive anticipation assumes no specific reaction. The reactive anticipation can be further classified as explicit and exact, explicit and approximate, and implicit in terms of the degree with which the bottom level model is represented at the top level model.

Schneeweiss (1999) developed a general framework that HPP models can be considered as a subclass of this framework. HPP models, being mainly capacity oriented, aims at achieving feasible production plans at the item level. In that respect, GFC models may be considered as an extension to the models at the lowest level of HPP framework. Bertrand *et al.* (1990) developed a material oriented approach where capacity is modeled in further detail than it is in HPP models, considering stochastic events at each stage of production.

The hierarchical approach presented in this thesis fits in the framework of Bertrand *et al.* (1990) and Schneeweiss (1999). This is due to our perspective of planning supply chain operations and modeling the dynamic framework described in the previous section. At an aggregate level, we concentrate on the coordination of material flow decisions in a multi-stage production-distribution network with emphasis on modeling with dynamic planned lead times, and at a detailed level, we model order scheduling and shop loading decisions.

In practice, examples of HP application are found in various APS software. A discussion about these systems deserves attention as it is related to possible application areas of our findings in this thesis.

Advanced Planning Systems

In a recent industry report by Gartner (2006) it is stated that the market for *Supply Chain Planning* (SCP) grew by approximately US$40 million to a total of approximately US$741 million in 2005. In 2006, it is forecasted to grow to approximately US$784 million. A supply chain planning suite is an integration of various software modules covering a wide range of applications including supply chain network design and collaboration, capacity and material planning, demand planning, transportation planning, and manufacturing planning and scheduling.

Software vendors such as SAP, J.D. Edwards, I2, Quintiq, and others are now releasing new products under the name APS for coordinating flows, exploiting bottlenecks, keeping due-dates and achieving realistic *Available to Promise* (ATP) records within complex supply chains. The structure and capabilities of each module may vary from one vendor to another, however, the basic structure of APS can be best described by a hierarchical integration of software modules based on the supply chain process that each module plans and its planning horizon. An illustration is provided in Figure 1.4.

APS employ decision support through sophisticated mathematical algorithms (e.g., genetic algorithms, linear programming, etc.) to provide (near) optimal solutions to the supply chain planning problems jointly considering various operating constraints of each supply chain process. Today, in most firms, ERP systems being used for bookkeeping and order processing have been supplemented by APS. This integration provides new challenges and opportunities in the way real-time data availability is used in improving the decision support function of these systems. Any change on the information status may influence the decision outcomes of one or more planning modules, which are then distributed to other planning modules through various coupling mechanisms.

The contents of this thesis may be related to cover **Master Planning** and **Production Planning and Scheduling** modules of APS. **Master Planning** is responsible for the coordination of material flow due to procurement,

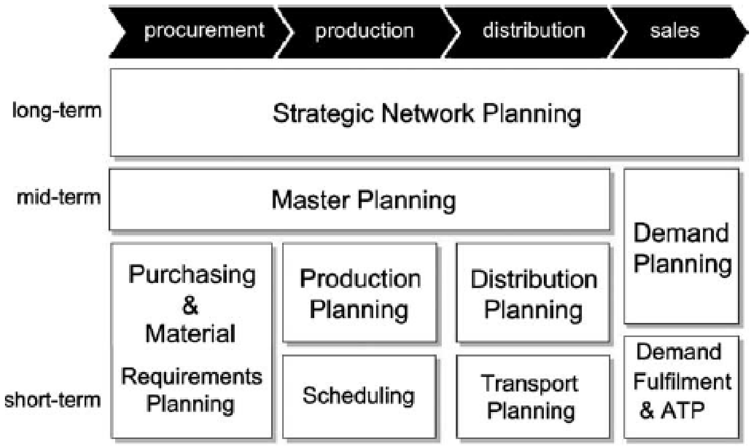


Figure 1.4: Software modules covering APS (Meyr *et al.* (2000b)).

production and distribution activities in the mid-term horizon. Furthermore, master production scheduling is also supported by considering the final product demands. **Production Planning and Scheduling** is usually covered by a single software module (Meyr *et al.* (2000b)). The responsibilities at this level are related to lot sizing, scheduling and shop floor control activities in order to satisfy the production outputs planned by the **Master Planning** level.

1.3.2 Lead Time Management

Lead time management is a very broad field. Three related lines of research may be classified in this field: (1) due-date assignment associated with the existence of internal due-dates, (2) setting planned lead times in planning periodic order releases for a make-to-stock situation, and (3) workload control to satisfy fixed planned lead times.

The research on due-date assignment has mainly evolved by considering individual orders of a job-shop in a make-to-order environment. The earliest models are presented by Conway *et al.* (1967), which is then followed by many others in various directions. In assigning due-dates to a specific order i, the fundamental relationship is

$$d_i = r_i + a_i,$$

where r_i is the release time of order i, a_i is the total allowance for the flow time of order i in the shop, and d_i is the due-date of order i. Given any pair of these three variables, the third one can be determined. The due-date assignment problems are generally described by finding a realistic d_i, through an estimation of the flow time by a_i, given the release time r_i. The key performance indicators are related to the mean, variance or the maximum of earliness, tardiness, lateness, and planned lead times of the orders. Earliness refers to the amount of time that an order is delivered earlier than planned, tardiness refers to the amount of time that an order is delivered later than planned, and lateness refers to both.

The critical question here is how to predict the individual order flow times accurately. This is not an easy task. The actual flow times are shown to be dependent on various system characteristics including order mix, lot sizes, workload levels, machine utilization, setup times, and etc (see Karmarkar (1987), and Karmarkar (1993) for a detailed discussion on these issues). Thus, one has to rely on approximate algorithms. The first rules for due-date assignment are introduced by Conway *et al.* (1967), and are generally based on order characteristics. These are: *Constant Allowance* (CON); lead times of all orders are set to the same constant parameter, *Total Work Content* (TWK); the lead time of an order is set in proportion to the total processing time of the order, and *Number of Operations* (NOP); the lead time of an order is in direct proportion to the number of operations on its routing. Separately, *Processing Plus Waiting Time* (PPW) has also been used as a linear combination of TWK and NOP rules (e.g., Kanet (1986), and Enns (1995)).

This research line has been further developed by Bertrand (1983) who considered the dependency of flow times on the workload and the machine capacity in the shop. Based on the idea of workload dependency new rules have been added to the previous list, such as *Jobs in Queue* (JIQ); the flow time of an order is estimated based on a proportion of the total number of orders in queue on its routing (Chang (1994)). Various techniques have been made available to the research community targeting different aspects of the problem. Examples include Vig and Dooley (1991) in estimating flow times based on a sampling of the flow times of the recently completed orders, Enns (1995) using queueing theory in predicting flow times and controlling tardiness, Hopp and Sturgis (2000) in quoting due-dates to achieve a target percentage of orders completed on-time, and Van Ooijen and Bertrand (2001) in considering the trade-off between the length of the quoted lead times and the delivery reliability from an economic perspective. The common methodology is to consider the assigned due-dates independent of the scheduling mechanism (due to complexity problems), and perform some sensitivity analysis based on simulation. Zijm and Buitenhek (1996) are the first to integrate due-date assignment with scheduling (Shifting Bottleneck Method). Building on the queueing constructs established by Karmarkar *et al.* (1985) and Karmarkar (1987), they proposed a hierarchically structured algorithm iterating between lead time setting and machine scheduling phases until a desired convergence measure is achieved.

The literature on setting planned lead times associated with order release planning is less developed. The traditional approach is to consider planned lead times as fixed inputs, exogenous to the planning system. It is assumed that the planned lead times are set based on some management intuition or experience and the flow times follow this intuition. The MRP logic is based on this assumption of exogenous lead times. The concept of exogenous lead times is further analyzed by De Kok and Fransoo (2003), and by Spitter *et al.* (2005a) and Spitter *et al.* (2005b) within the context of planning supply chain operations with capacity constraints. The majority of the studies on lead time setting approach the problem from a static view, and strive to find *fixed* planned lead times that best fit a stationary situation (e.g., Yano (1987), Molinder (1997), and Enns (2001)), mostly in an MRP context. Hoyt (1978) is the first to criticize the fixed lead times and argue that the lead times should be dynamic, in a sense that they reflect the dynamic operational characteristics of a production process, in particular, by looking at the average queue length and the average output realized recently. This discussion is further enhanced by Kanet in a series of papers; he first investigated the various effects lead times have on a multi-stage production-inventory system (Kanet (1982)), then he emphasized the favorable results in terms of order tardiness achieved by TWK rule (Kanet (1986)). Since then, the research on dynamic planned lead times for supply chain situations has not attracted much attention. Recently, Enns and Suwanruji (2004) modeled exponentially smoothed lead times in a two-stage *Distribution Requirements Planning* (DRP) system, and by simulation, showed the sensitivity of the system to safety lead time factors and lot-sizing choices.

Although they consider different problems at different aggregation levels, the studies we have mentioned thus far discuss the management of lead times from a forecasting perspective with the emphasis on minimizing the impact of the forecasting errors. Another approach is to control the actual flow times in a way to match pre-determined norms (in specific planned lead times). The main motivation comes from the fact that the actual flow times largely depend on the total workload in the manufacturing centers, and one who can control the workload can also control the flow times. Based on this observation, a technique called Input/Output control (Wight (1970)) has become popular. The idea is, roughly, to keep the amount of workload at a constant level by controlling the work order releases to a manufacturing center, which has led to a wide stream of research (e.g., Bertrand and Wortmann (1981), Bechte (1988), Kingsman *et al.* (1989), and Land and Gaalman (1998)). Success stories using this techniques have been reported both in industry and in academic studies (Hopp and Spearman (2000)). On the other hand, a major weakness has been reported in causing long delays for the orders waiting to be released (Tatsiopoulos and Kingsman (1983), and Zijm and Buitenhek (1996)). Although the average and the variability of the time that orders spend in a shop floor can be reduced significantly, the total manufacturing flow time as seen from a higher-level order release perspective may still possess a high level of variability.

Thus, the variability in order flow times are in most cases unavoidable, and it is still an important challenge to incorporate this variability in planning decisions. In this thesis, we aim to model dynamic planned lead times and evaluate their performance. Planned lead times are updated at every replanning opportunity depending on the levels of changes in the system status. In doing so, both forecasting methodology (e.g., exponential smoothing, and JIQ) and a (limited) control approach is employed.

1.3.3 Clearing Function

The idea of clearing was first deployed by Graves (1986), and has been more specifically defined in Karmarkar (1989) (see Karmarkar (1993) for an extensive discussion). The clearing function is based on the fact that production output is a function of workload, and it relates the total workload of a shop floor to the anticipated flow time of the next job to be released. Examples of clearing functions used in the literature are illustrated in Figure 1.5.

The clearing function of Graves (1986) indicates that the amount produced is a constant proportion of WIP, Throughput $= \tan(\sigma) \cdot$ WIP. It is represented by the *Fixed Lead Time* function in Figure 1.5. It assumes infinite nominal capacity and a fixed lead time independent of the WIP level. The fixed planned lead time considered in Graves' model is used in a way to smooth the production output per period. This is different from the way that planned lead times are used in classical MRP systems, where the variability in planned order releases is directly carried onto the manufacturer's planned output process. The *Fixed Capacity* function reveals the assumption that the throughput is independent of the WIP level, and is bounded by a rigid nominal capacity, μ, (e.g., Billington *et al.* (1983), Chung and Krajewski (1984), and Voß and Woodruff (2003)). A typical approach is employed by combining the *Fixed Lead Time* and the *Fixed Capacity* functions, where the throughput is limited according to the available WIP up to a certain level, and beyond that level, it is fixed

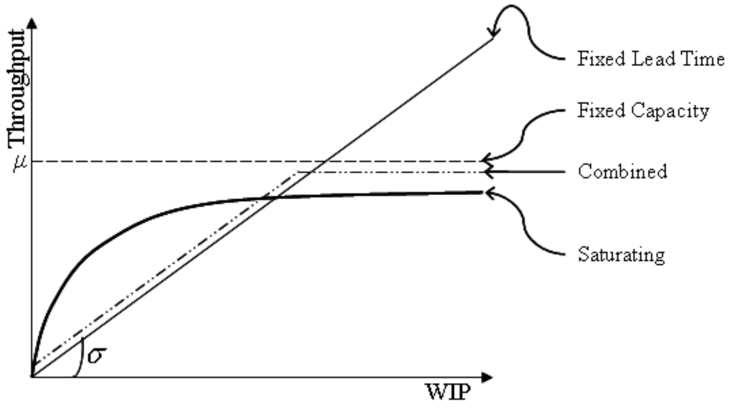


Figure 1.5: Examples of clearing functions (Karmarkar (1989)).

to the nominal production rate implying a finite capacity (e.g., Hackman and Leachman (1989), and Spitter *et al.* (2005b)). The *Saturating* function implies the dynamic behavior of the throughput due to the congestion effect of increasing WIP in the system. It has been derived initially from steady-state queuing constructs (Karmarkar (1989)).

Experimentations with saturating clearing functions based on the theory of steady-state queueing systems have been reported by Zäpfel and Missbauer (1993) in designing efficient workload control systems, by Asmundsson *et al.* (2003) and Asmundsson *et al.* (2004) in improving mathematical programming techniques for aggregate production planning, and by Hwang and Uzsoy (2005) in developing lot sizing models with setup times. Such a non-linear and concave shape for the clearing of WIP is also approximated by Asmundsson *et al.* (2006) and Armbruster *et al.* (2004) by fitting the curve to the experimental results from simulations of practical manufacturing settings. Missbauer (2002) elaborates the saturating function of Karmarkar (1989), and shows some limitations of stationary models especially under time-varying demand. It has been argued that transient analysis of queueing networks should be used to develop more precise models of the dynamic behavior of the production units (see Missbauer (2006) for a recent discussion). Additionally, Riaño (2002) models the cumulative output of a production process in terms of the sum of weighted transformations of the previous inputs to the production process. The transformation function is linear in nature with the weights computed based on an assumed knowledge over a flow time probability distribution.

The concept of clearing function is extensively investigated in this thesis. It is used for anticipation purposes so that models of the behaviors of the production processes controlled by the operational scheduling level present in higher decision levels in terms of predicting the relevant performance indicators such as work-in-process, flow times, etc. We utilize clearing function as a non-reactive anticipation.

1.4 Motivation of the Research

1.4.1 Practical Motivation

From the perspective of practical importance, the research conducted in this thesis is highly relevant for three groups in society: problem owners, software developers, and software implementers.

To the problem owners, the design of supply chain planning hierarchies is highly relevant, and is still largely built on experience and aspect knowledge. The existence of a dynamic framework with update of the state information and accordingly with revisions of planning parameters has not been considered. In relation to rolling horizons, very limited updating activities are put into design, and the configuration of different decision functions is mainly performed in a static setting. However, the performance of hierarchical planning systems is largely dependent on parameters such as the frequency of updating that can be evaluated in a dynamic setting. It relates to the sensitivity of the planning system to the changes in the supply chain status. This thesis provides relevant insights that can be translated into direct assistance when redesigning or reconfiguring hierarchical planning systems in a dynamic setting.

To the software developers, the design of APS is largely directed by theoretical insights on hierarchical planning systems (Zoryk-Schalla (2001)). The current state of the art in APS shows the integration of various software modules based on deterministic optimization algorithms both for high level master planning and for low level production planning and scheduling problems. Stochastic models that explicitly incorporate demand or process uncertainties within the performance measurement (e.g., models based on queueing networks) are absent in APS to date. This thesis provides insights related to modeling the process uncertainty at the master planning level. Although each software module within APS is in itself well-developed, design related issues with regard to the interaction between different modules has to be investigated further with possible improvement alternatives. A number of current paradigms in the theory on hierarchical planning systems (e.g., issues related to anticipation or coupling) need to be researched further especially to get insight into the dynamic performance of these systems. This thesis provides results that can be used as a tool in designing APS.

Data integration between various modules is supported by advanced information technologies, and it is claimed that any new information generated by a module can be automatically transferred to other modules as an input (Meyr *et al.* (2000a)). An important question is to what extent all this information should be used to update plans and planning parameters. Implementation success of APS is highly influenced by the correct parameter setting, such as replanning or updating frequencies. Frequent updating may help being responsive to changing information statuses but at the same time, may cause stability problems in decision-making. In this respect, this thesis provides further insights to assist consultants with successful implementations of APS.

1.4.2 Theoretical Motivation

Although there has been a tremendous interest on hierarchical planning systems for decades, the research on the dynamic performance of such systems is quite scarce. The dynamic framework of planning hierarchies as it is illustrated in Figure 1.3 has not attracted much attention in the research community. Although the actual implementation may occur in a variety of ways, plan-execute-feedback-(re)plan cycle remains as a common strategy in implementing various production planning tools, which has not been evaluated in a hierarchical context.

There have been a few studies that are mainly focused on determining the planning horizons at different hierarchical levels in a rolling horizon setting (e.g., Chung and Krajewski (1986), Chung *et al.* (1988), and Rohde and Wagner (2000)). These studies emphasize the problems in fine-tuning the higher and lower level planning horizons associated with the implementation of rolling horizons. Mainly based on HPP methodology, they consider aggregate capacity planning at a higher level and final product master scheduling at a lower level. Fixed planned lead times are applied in a purely deterministic environment, and their models are not extendable to multi-stage supply chain situations.

In traditional hierarchical planning systems based on mathematical programming models applied in a rolling horizon setting, planning parameters are not subject to change in consecutive replanning opportunities. On the other hand, today's manufacturing organizations are confronted with a high degree of uncertainty in their operations as a result of highly dynamic market conditions, and there is a need to facilitate a dynamic and adaptive hierarchical planning framework. In McKay *et al.* (1995), it is clearly stated that there has been an absence of a critical discussion that deals with the underlying assumptions, implications, and limitations of the HPP paradigm. Additionally, Zijm (2000) criticizes HPP systems by emphasizing drawbacks such as the complexities arising in multi-stage production/inventory structures and the fact that uncertainty at various levels is not incorporated systematically. Such a discussion has been initiated and carried over by a few studies (Bertrand *et al.* (1990), and Schneeweiss (1999)) in a rather conceptual manner, and there is still a lack of formal analysis through numerical and analytical findings. The research conducted in this thesis aims to fill in this gap in the hierarchical planning literature.

The dynamic due-date setting literature includes many examples of elegant techniques in improving the forecast accuracy of order flow times. However, the emphasis has been on the detailed shop floor control activities; the planned lead times determined for every order are used in scheduling orders in a job shop. The multi-level decision hierarchy is ignored through the assumption that the orders are created by an order generation process external to the system under study. Besides, from a supply chain perspective, a single production unit is considered independent of other upstream and downstream entities in the network. However, the concept of internal due-dates is generally attributed to the existence of different interacting units within the production network (Conway *et al.* (1967)). In this way, the arrival and departure of orders at different units can be synchronized by a higher level planning mechanism. The studies on such order release mechanisms on the other hand employ fixed planned lead times (De Kok and Fransoo (2003)). One of the objectives of this thesis is to provide a hierarchical framework by which dynamic planned lead times that are both cost effective and close to the actual flow times can be determined.

There has been a growing awareness in the past few years on various modeling techniques in terms of representing the operational characteristics of production processes at an aggregate planning level. All of the research presented in this line evaluate various approaches in modeling clearing functions for a single-stage manufacturing company. None of the studies explicitly considers the planned lead time in coordinating the order releases in successive stages of the supply chain. There has been a lack of formal analysis in evaluating various impacts that different clearing functions may have on the order delivery performances, and the resulting cost terms. Filling in this gap is one of the motivations of this thesis.

1.5 Research Questions

The research presented in this thesis aims to contribute to the area of developing supply chain planning tools with dynamic planned lead times, and to provide assistance in understanding various impacts that dynamic lead times may have on the performance. In achieving this objective the following questions are posed in this thesis:

1. *What are the performance consequences of updating planned lead times?*
2. *How should planned lead times be updated in an effective way?*
3. *What is the impact of the frequency of updating lead times on the performance?*
4. *What is the impact of the hierarchical coupling mechanism and the level of anticipation on the performance?*
5. *What is the impact of the uncertainty and the utilization level on the performance?*

Updating lead times is a means of disturbing (for the purpose of being flexible) the planning system that is thought to be not realistic anymore. It may have favorable effects as well as drawbacks, which require in-depth analysis. Our first research question is related to this problem. Such an analysis is expected to provide insights into developing effective algorithms in updating planned lead times, and leads us to pose the second research question. The performance of the system is expected to depend on various implementation and design related issues targeted by our third and forth research questions, and also environmental issues related to demand uncertainty and the level of demand that are targeted by the fifth research question.

1.6 Research Methodology

The research questions that we raise in the previous section are addressed in different ways in different chapters of this thesis. Mainly, the research is conducted using computer simulation because, the models we consider require a complex analysis that is not computationally efficient, if not feasible, when it is tackled analytically.

The models for multi-echelon production-inventory systems are generally considered as complex in terms of their analytical tractability. One usually cannot expect to find closed form analytical results unless some simplifying assumptions are made (e.g., Clark and Scarf (1960), and Diks and De Kok (1998)). Incorporating dynamic planned lead times with capacitated production units further increases the problem complexity. Experimental results gathered from computer simulation provide valuable insights towards a formal understanding of such complex systems.

The following steps are employed in structuring our experimental studies. Firstly, a sufficient level of detail in modeling the physical production and distribution system is identified according to our specific research questions targeted in each chapter. Then, a hierarchical structure with proper linkages between its levels are determined by putting emphasis on the flow of information within the planning hierarchy. Update policies depending on the level of reaction to the status feedback from the execution systems are determined. The specific decision functions, the inputs and outputs, and the roles of each level are modeled. High levels of variability and uncertainty are considered both in demand and production processes, and the decision making instances are assigned to specific points in time. In order to guarantee a certain service level for each simulation, an initial run is taken with safety stock equal to zero, then the safety stock adjustment procedure, described in the Appendix of Chapter 2, is employed. The experimental design is set up based on the specific research questions targeted at each chapter, and along the common standards of experimental design for simulation (e.g., Kleijnen and van Groenendaal (1992), and Law and Kelton (2000)). A full factorial design is employed, in order to take into account the effects of each factor separately.

In addition to the simulation studies, we also employ an analytic approach to arrive at closed form results for a described phenomenon. It is based on modeling a production system as a Markov process where our assumption of periodic order releases is replaced by a continuous stream of order generating process. Although a few restrictive assumptions are made to provide brevity in the solutions, strong analytical relationships are established. A certain aspect of updating planned lead times is evaluated. The higher level order release mechanism is formulated by a function of the planned lead time in a rather abstract manner. But, at the same time, it follows the order release dynamics as in the periodic linear programming formulations. The resulting queueing system is modeled by a two-dimensional *Quasi-Birth-and-Death* (QBD) process, and it is solved using the matrix geometric techniques of Neuts (1981). In generating an explicit rate matrix for the QBD, the results of Ramaswami and Latouche (1986) are applied.

1.7 Outline of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 is devoted to understanding the effects that dynamic planned lead times have in planning the operations of a multi-stage supply chain. Numerical insights have been gathered through simulation experiments with emphasis on the update frequency, and the anticipation level at the SCOP function. An earlier version of this paper has appeared as Selçuk *et al.* (2006a).

In Chapter 3, the insights presented in Chapter 2, and the phenomenon of *lead time syndrome* are modeled and analyzed explicitly. The problem space is reduced to a single production unit with continuous release of orders based on a specific relation to the planned lead time. The drawbacks of updating the lead time are analyzed in depth considering the stability and the relevant performance metrics of the system. This chapter is an extension to Selçuk *et al.* (2006b).

In Chapter 4, we discuss using the clearing function concept in modeling the production process at an order release planning level. The orders are released and scheduled according to their planned lead times, and the evaluation of various types of clearing functions are done in a hierarchical context. In addition to the existing ones, a short-term clearing function based on the probabilistic behavior of the production process is considered. The content of this chapter has been presented in Selçuk *et al.* (2006d).

In Chapter 5, a load dependent lead time update procedure is developed and tested in various settings. The concept of clearing function is embedded into the lead time update procedure. Dynamic and fixed planned lead times are evaluated with emphasis on both the detailed modeling of the clearing behavior and the coupling mechanism within the planning hierarchy. This chapter is based on Selçuk *et al.* (2006c).

Finally, in Chapter 6, we summarize the main contributions of this thesis. In addition, we discuss some directions for further research on the topics covered in this thesis.

Chapter 2

Updating Lead Times in Hierarchical Planning Systems

This chapter is devoted to exploring the opportunities and drawbacks in updating the planned lead times of a multi-stage serial supply chain within a hierarchical planning context. The decision hierarchy, from top to bottom, is composed of lead time setting, operational planning (SCOP), and operational scheduling levels. The planned lead times are determined by exponentially smoothing the previously realized order flow times. Optimal order release decisions are given through a mixed-integer programming formulation at the SCOP level. The objective is to minimize the total inventory costs with a given safety stock level. Orders are then scheduled independently for each production unit in the supply chain. The method of rolling horizons is used to gather data for anticipation and updating purposes.

2.1 Introduction

The majority of the established and widely accepted systems (e.g., MRPII, DRP, and JIT) to manage and control planning activities in a production and distribution environment require an almost perfect environment such as highly reliable manufacturers, deterministic demand, and order flow times that are independent of both workload and order-mix. The planning parameters are usually set according to a priori simulation results, management intuition, or experience. However, it is practically not possible to accurately determine the parameters related to the physical flow of goods within the supply chain, such as the planned lead times. A planned lead time that is set by the management long time ago may not be a valid representation anymore, since order-mix, demand, workload levels, and manufacturing technology are all subject to changes and uncertainties in today's dynamic environment.

In real life, perfect information on the characteristics of the system uncertainty (explicit models for probability distributions) is generally not available, forcing the planner towards being more reactive than pro-active. One obvious, and most popular, way of dealing with uncertainty in a reactive way is to apply a rolling horizon method such that the earlier plans may be revisited and changed in response to unexpected alterations in operating and market conditions. Instead of just applying the same routine procedures with fixed parameters to changed demand forecasts and inventory records, one can also benefit from rolling horizons through elaborating the historical data, and updating planning parameters such as the planned lead times. Planned lead times play a crucial role in managing the flows of materials, especially in multi-stage production/distribution systems, by influencing the order release decisions given at the SCOP level. Therefore, it is important to establish a certain level of consistency between the planned lead times and the flow times realized as a consequence of order release and scheduling decisions.

However, the effect of limited production capacities and uncertainty inherent in the operating environment on the current and the future schedules may not be anticipated at a higher planning level, yielding increased nervousness due to updating the lead times. We expect as the planned lead times are increased, orders are released earlier in larger quantities, and vice-versa if the planned lead times are decreased. In other words, the order(s) planned to be released in a following period are rescheduled and added onto the current period's order in order to prevent the stock-outs that may be caused by longer lead times. This may generate further increase in the planned lead times in the future. In case of a decrease in the planned lead time, the order(s) scheduled to be released in the current period are postponed to later periods to deplete the excess inventory in pipeline caused by shortened lead times. This may generate further decrease in the planned lead times in the future. Such a cyclic relationship between order sizes and dynamic lead times has been conceptually described as *lead time syndrome* by Mather and Plossl (1978) but, the formal analysis of the phenomenon has not been conducted.

This type of cyclic interactions may in general be considered within the context of information-feedback systems, which has been extensively studied in Forrester (1980). An information-feedback system refers to situations where information on environmental status leads to a decision that results in action which affects the environment, and thereby influences future decisions. Information-feedback systems are not necessarily well behaved. "In fact, a complex information-feedback system designed by happenstance or in accordance with what may be intuitively obvious will usually be unstable and ineffective" (Forrester (1980)). There has been a lack of formal analysis on the application of these systems in planning and controlling production-inventory situations.

In this chapter, our primary concern is to provide qualitative insights into how updating the lead times at every replanning epoch in a rolling horizon may affect the performance of a hierarchical planning system for a multi-stage, multi-product serial supply chain situation (see Figure 1.1). The production processes and the final product demands are subject to uncertainty yielding variable workload levels at the production units, and variable inventory levels at the stock points. The planned lead times are updated using exponential smoothing technique over the past realizations of order flow times. In addition to the hierarchical structure with dynamic planned lead times, the supply chain interaction, specifically the dependence of downstream order releases on upstream inventory levels, adds further complexity and opportunities for further insights.

Considerable research has been conducted on the analysis of manufacturing flow times (e.g., Karmarkar (1987), Zijm and Buitenhek (1996), and Lambrecht and Vandaele (1996)). Insights gathered from these studies emphasize the correlation between the actual workloads and the flow times that can be realized under a limited capacity flexibility. A first study on dynamic lead times in an MRP context is Hoyt (1978) who suggested the planned lead times be based on the historical averages of work-center queue sizes and throughput levels. The shop order start date is then determined by offsetting the dynamic lead times, ignoring their effects on order release mechanism. In a more recent study, Vig and Dooley (1991) demonstrated that the flow times from recently completed jobs provide very useful information for internally setting due-dates in a job-shop environment. Our approach is close to Enns and Suwanruji (2004) who adjusted the planned lead times using exponentially smoothed order flow times in a serial supply chain managed by a DRP system, and concentrated mainly on the safety lead time factors and lot sizing approaches in their experimental design. They ignored the nervousness in order release decisions created by updating the planned lead times, and fixed lead times were not considered in their experiments. Insights into the relative effects of dynamic lead times with respect to the fixed lead times based on the hierarchical context and the parameters related to the dynamic framework such as the update frequency are still missing. The purpose of this chapter is to fill in this gap through numerical results obtained by simulation.

2.2 Planning Hierarchy

The planning hierarchy with flows of information in between different decision levels is illustrated in Figure 2.1. At the tactical planning level, the lead times are updated applying the exponential smoothing technique based on the history of order flow times. As it is mentioned in the previous section, using historical averages has intuitively been accepted as a relevant technique in estimating the flow times, and given that the flow times follow a stationary distribution, exponential smoothing is expected to perform well. In our case, we consider stationary demand and production processes. Thus, we expect that the time-series of flow times deviate around a single mean. Only the final product lead times are updated for clarity in the analysis. Final product lead times are influential on the safety stocks, which affect the total costs.

The planned lead times are given to the SCOP and the operational scheduling levels for further, more detailed planning. At the SCOP level, a *Mixed-Integer Programming* (MIP) formulation is solved to determine the optimal quantity of order releases for each item given the current state of inventory levels and the demand forecasts. MIP formulations are proven to be efficient tools in modeling material flow together with capacity utilization decisions (e.g., Billington *et al.* (1983), and Spitter *et al.* (2005a)). With the existence of positive planned lead times, the order release decisions are given to satisfy the periodic demand forecasts in a time-phased approach.

At the operational scheduling level, detailed production schedules are determined for each production unit in a decentralized manner according to *First-Come-First-Serve* (FCFS) strategy. The sizes of the released orders together with their delivery schedules constitute the finalized set of decisions and are given to the execution system such as ERP and *Manufacturing Execution Systems* (MES).

The circled numbers, in accordance with the sequential process of hierarchical decision making, indicate the sequence of information flow within the planning system. Data set (1) refers to the external input to the planning system including status information and performance outputs from the production units and the stock points and the future demand forecasts. The schedule of currently open orders has to be updated in accordance to the capacity restrictions so that the operational limitations of the production units can be represented realistically at the SCOP level. The updated schedule is feedfor-

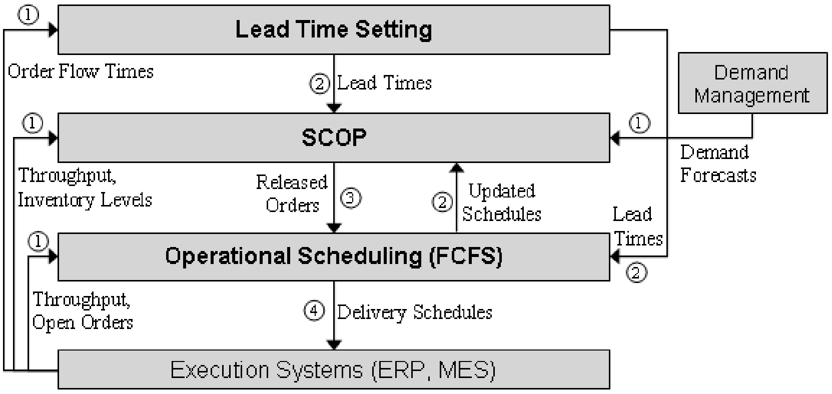


Figure 2.1: Hierarchical planning system.

ward to the SCOP level within data set (2) together with the planned lead times coming from the tactical planning level. This schedule represents the expected delivery date of previously released orders, and is used in releasing new orders and planning inventory at the SCOP level. Data set (3) is composed of the size of the orders released at the start of the current period which are then scheduled in the lower level and finalized within data set (4). With respect to the dynamic framework, data set (1) indicates the feedback from the environment, which influences the decision outcomes within the planning hierarchy.

The static model parameters and the index variables used in the formulations throughout this chapter are as follows:

N = Set of items produced and distributed in the supply chain.
i, j = Item indexes, i, $j = 1, \dots, n$.
a_{ij} = Quantity of item $i \in N$ needed to produce a unit of item $j \in N$.
x_i = Unit order lot-size of item $i \in N$.
τ_i = Processing time for the unit order lot-size of item $i \in N$.
N_e = Set of final products, $N_e = \{i; a_{ij} = 0, \forall j \in N\}$.
u = Production unit index, $u = 1, \dots, m$.
S_u = Set of items produced in production unit u, $u = 1, \dots, m$.
C_u = Total capacity of production unit u measured in terms of time-units available for production during a period, $u = 1, \dots, m$.
T = Forecast horizon for the final product demands.
t = Period index, $t = -\infty, \dots, T - 1$.

We assume the orders are released in integer multiples of a fixed lot size. From a practical perspective these lots sizes may stem from the transportation capacities per truck in a full truck load system, such that the order size refers to the number of truck loads to be transported from a production unit to a stock point. Existence of such flow restrictions motivates differentiation between production of items and delivery of orders to stock points. The planning horizon T is determined so that the cumulative lead time throughout the supply chain does not exceed T. We also assume T is long enough to capture the variability in the planned lead times, which motivates us in setting a practical upper bound for each planned lead time. This is explained in the next section. The period index is used for both decision making about future and incorporate previous decisions into the model. $t < 0$ refers to the past periods, $t = 0$ is the current period, and $t > 0$ refers to the future periods.

The dynamic inputs to the planning system are revised at every replanning opportunity. They include

$D_i(t)$ = Demand forecasts for final products $i \in N_e$ in period $t > 0$.
$I_i(0)$ = Current net inventory level (on-hand minus backorders) of item $i \in N$.
$Q_i(t)$ = The size of an order for item $i \in N$ that were released in a past period $t < 0$, and not yet finished.
$\widetilde{Q}_i(t)$ = Total number of unit lot-sizes of item $i \in N$ that have been released previously and scheduled to be delivered at the start of period $t > 0$.
$Z_i(t)$ = Actual throughput quantity (number of unit lot-size) of item $i \in N$ in period $t < 0$.

The schedule for the open orders, represented by the term $\widetilde{Q}_i(t)$ for $t > 0$, is subject to change by the operational scheduling level depending on the capacity restrictions and the workload status of the production units. The new schedule and the resulting delivery quantities denoted by $\widehat{Q}_i(t)$, is feedforward to the operational planning level as inputs.

2.2.1 Dynamic Planned Lead Time Setting

A logical requirement is that the planned lead times of items produced in a certain production unit should be long enough to capture the current work-in-process in that production unit. Let us define $L_i^{\min}$ as the lower-bound for the planned lead time of item $i \in N_e$. Given that $i \in S_u$,

$$L_i^{\min} = \left\lceil \sum_{j \in S_u} B_j(0)/C_u \right\rceil ,$$

where $B_j(0)$ denotes the current work backlog due to the released orders for item $j \in S_u$ in production unit u measured in terms of time-units, and is computed by

$$B_j(0) = \tau_j \cdot \left(\sum_{s=-\infty}^{-1} (Q_j(s) - Z_j(s)) \right).$$

In addition, an upper-bound for the lead time of an item $i \in N_e$ is given as the planning horizon minus the sum of the lead times of its upstream items. In a two-stage serial system with $j \in N \setminus N_e$ upstream to $i \in N_e$, the upper-bound for the lead time of item i is

$$L_i^{\max} = T - 1 - L_j.$$

Let us denote the flow time of the k^{th} completed order of item i in the previous period by $F_{i,-1}^{(k)}$. Although the actual completion time of any order can be any duration, the flow times are expressed in terms of integer multiple of periods because, the orders are released at the start of a period and delivered at the end. According to the exponential smoothing of the flow times of previously completed orders, the current estimate of the order flow time of item i is

$$\widehat{F}_{i,0} = (1-\zeta)^{\chi_i} \cdot \widehat{F}_{i,-1} + \zeta \cdot \sum_{k=1}^{\chi_i} (1-\zeta)^{\chi_i - k} \cdot F_{i,-1}^{(k)}, \tag{2.1}$$

where $\widehat{F}_{i,-1}$ denotes the estimated order flow time for item i at the start of the previous period, χ_i is the total number of orders of item i completed and delivered in the previous period, and ζ is the smoothing constant that indicates the weight given to the latest occurrences of the order flow times. It should be noted that $\widehat{F}_{i,-1}$ is not an integer, and χ_i is not too large. The former is necessary for smoothing, and the latter is necessary so that the first term on the right-hand-side of Equation (2.1) does not become zero.

Then, the planned lead time of item $i \in N_e$ is expressed as integer multiple of periods, and it is given by

$$L_i = \min\left\{ L_i^{\max}, \max\left\{ L_i^{\min}, \left\lceil \widehat{F}_{i,0} - 0.5 \right\rceil \right\} \right\}.$$

The value of ζ implicitly provides the frequency at which the lead times are updated. This is basically due to the fact that planned lead times are expressed as integer multiple of periods. Let us consider an example to clarify this statement. Consider that $L_i = \widehat{F}_{i,0} = 1$, and we face order flow times of two periods for item i in each of the following five periods. Then, for $\zeta = 0.5$ one unit increase in the planned lead time of item i occurs at the end of period 1, and for $\zeta = 0.2$ one unit increase in the planned lead time of item i occurs at the end of period 3, implying that the lead time of item i is updated less frequently than the case with $\zeta = 0.5$.

Exponential smoothing method provides a relevant flexibility through parameter ζ in modeling the response to the realized flow times. Besides, it is a relevant approach in forecasting flow times due to a high level of correlation between the flow times of consecutive orders especially when the scheduling discipline is FCFS.

2.2.2 SCOP

The objective of the SCOP formulation is to minimize the total inventory costs and the penalty costs for the final product shortages over the planning horizon, where the delivery of order releases are scheduled according to the planned lead times (already determined at the tactical level). An order that is released now is assumed to be available in its stock point after a duration of its planned lead time.

The production units are assumed to have fixed capacities with linear capacity consumption rate per unit lot-size processed. The available capacity levels in the latter periods are anticipated based on the past throughput performance of the production units. It is intuitively clear that a reliable planning system does not load the production units more than they can produce within the given planned lead times. Thus a lead time dependent workload control rule is applied; the total workload planned for a certain production unit is limited in a way to satisfy the planned lead times of the items produced in that production unit.

The cost parameters are

h_i = Unit holding cost of item $i \in N$ in the inventory for one period.
M_i = Unit penalty cost of having shortage for the final product $i \in N_e$ in one period.

M_i is set so high that the system always targets a nonnegative net inventory level for item $i \in N_e$.

The non-negative decision variables are defined as follows:

$I_i^+(t)$ = On-hand inventory of item $i \in N$ at the start of period t, $t = 1, \ldots, T$.
$I_i^-(t)$ = Shortage level of final product $i \in N_e$ at the start of period t, $t = 1, \ldots, T$.
$Q_i(t)$ = Number of unit order lot-size of item $i \in N$ released at the start of period t, $t = 0, \ldots, T-1$.
$U_i(t)$ = Resource utilized for item $i \in N$ in period t, $t = 0, \ldots, T-2$.

The term $U_i(t)$ provides the planned duration in a period that is allocated for the production of item $i \in N$ in period t. We additionally note that the intermediate items do not have external, independent demands, and their dependent demands cannot be recorded as backorders for later periods. Therefore, $I_i^-(t) = 0$ and $D_i(t) = 0$, for all $i \in N \setminus N_e$ and $t = 0, \ldots, T-1$.

The SCOP formulation is as follows:

$$\text{Min.} \sum_{i=1}^{n}\sum_{t=1}^{T}\left(h_i \cdot I_i^+(t) + M_i \cdot I_i^-(t)\right) \tag{2.2}$$

s.t.

$$I_i^+(t+1) - I_i^-(t+1) = I_i(0) + \sum_{s=1}^{t}\widehat{Q}_i(s)\cdot x_i + \sum_{s=L_i}^{t} Q_i(s-L_i)\cdot x_i$$

$$-\sum_{s=0}^{t} D_i(s) - \sum_{s=0}^{t}\sum_{j=1}^{n} a_{ij}\cdot Q_j(s)\cdot x_j, \quad i \in N,\, t = 1, \ldots, T-1 \tag{2.3}$$

$$\sum_{s=0}^{t} U_i(s) \leq B_i(0) + \sum_{s=0}^{t}\tau_i \cdot Q_i(s), \quad i \in N,\, t = 0, \ldots, T-2 \tag{2.4}$$

$$B_i(0) + \sum_{s=0}^{t}\tau_i \cdot Q_i(s) \leq \sum_{s=0}^{t+L_i-1} U_i(s), \quad i \in N,\, t = 0, \ldots, T-L_i-1 \tag{2.5}$$

$$\sum_{i \in S_u} U_i(t) \leq C_u, \quad u = 1, \ldots, m,\, t = 0, \ldots, T-2 \tag{2.6}$$

Constraint set (2.3) balances the material flow between consecutive planning periods. Constraint set (2.4) imposes the input-output relationship for the production units. The cumulative amount of resource allocated for the production of a certain item cannot be larger than the resource requirement of cumulative quantity released for that item. Constraint set (2.5) ensures that any release will be produced within its given planned lead time. Finally, constraint set (2.6) sets capacity restrictions for each production unit.

Ignoring constraint sets (2.4) through (2.6) generates a formulation for a DRP system. Having $M_i >> h_i$ implies that the order release decisions are driven by demand forecasts and material availability. So, the objective is to keep as low inventory as possible while avoiding backorders, which yields a compact expression for the optimal order release decisions for the first period. Given that item $j \in N_e$ is produced from item $i \in N \backslash N_e$, the order release for item j in the current period is formulated by

$$Q_j(0) = \min\left\{\frac{I_i(0)}{a_{i,j}\cdot x_j}, \left(\frac{\sum_{s=0}^{L_j} D_j(s) - IP_j(0)}{x_j}\right)^+\right\},$$

where $IP_j(0)$ is the current inventory position of item j just before the order for item j is released. It is the current net inventory level of item j plus the total quantity of item j that have been in process in its production unit. When the inventory for item i is not large enough, the amount of forecasted shortage for item j, $\sum_{s=0}^{L_j} D_j(s) - IP_j(0) - \frac{I_i(0)}{a_{i,j}}$, is transferred to the next periods' releases to avoid future projected stock-outs. When there is an ample stock of item i, then the release decisions for item j follow a periodic review base-stock policy with the base-stock level equal to the total forecasted demand during the planned lead time plus the review period. The safety stock is taken az zero as in the SCOP formulation of (2.2)-(2.6). For brevity we relax the assumption that $Q_j(0)$ is integer.

In Lambrecht *et al.* (1984) and Buzacott *et al.* (1992), it has already been shown that MRP systems used for control of production release and parts ordering in manufacturing are equivalent to somewhat generalized base-stock systems, with the key difference being that MRP systems make decisions at each level using an echelon target stock that includes a forecast of future final demand. A similar approach can also be used for the description of DRP logic in continuous time. This will provide us some preliminary insights into the impacts of updating the lead times on the order release decisions.

A Continuous Time Analysis of DRP Logic

In this section, we concentrate on a two-stage structure with items i and j as the upstream and downstream items respectively. It is also assumed that the inventory position of an item is always less than or equal to the expected demand during its lead time. It should be clear throughout the text that this assumption eases the presentation and does not restrict the results. The following analysis can also be applied to serial supply chains of various sizes.

Assuming that the review period is indefinitely small we can carry our discussion about DRP onto the continuous space. Let us define $\lambda_i(t)$ and $\lambda_j(t)$ as the expected release rates of unit lot-sizes of items i and j respectively at time $t \geq 0$. They are the continuous time equivalent of $Q_i(t)$ and $Q_j(t)$. The expected total demand for item j during a period of time between now and a future time $t > 0$ is denoted by $d_j(0, t)$, and $d_j(t)$ is the expected demand rate for item j at time $t > 0$. Then,

$$\lambda_j(0) = \min\left\{\frac{I_i(0)}{a_{i,j} \cdot x_j}, \lambda_j^*(0)\right\}, \tag{2.7}$$

where $\lambda_j^*(0)$ is the release rate of item j given that there are infinitely many item i in stock,

$$\lambda_j^*(0) = \frac{d_j(0,\, L_j) - IP_j(0)}{x_j}. \tag{2.8}$$

Considering that there is no shortage for the raw material of item i, the order releases for item i are $\lambda_i(t) = \lambda_i^*(t)$ for all $t \geq 0$. The shortage in the stock of item i creates a transient effect on the release of item j with $\lambda_j(0) < \lambda_j^*(0)$. Due to the ample supply assumption at the upstream stage, it is expected that this effect lasts for at most L_i duration of time,

$$\int_0^{L_i} \lambda_j(t)dt = \int_0^{L_i} \lambda_j^*(t)dt.$$

Thus, we can identify a time-point $0^+ \leq L_i$ such that $\lambda_j(t) = \lambda_j^*(t)$ for all $t \geq 0^+$.

Consider that after an expected release of order for item j at time $t \geq 0^+$ the next order is expected to be released at time $t + \Delta t$. Then,

$$\begin{aligned}
\lambda_j(t+\Delta t)\cdot x_j &= d_j(t+\Delta t,\, t+\Delta t+L_j) - IP_j(t+\Delta t) \\
&= d_j(t+\Delta t,\, t+\Delta t+L_j) - (d_j(t,\, t+L_j) - d_j(t,\, t+\Delta t)) \\
&= d_j(t+L_j,\, t+L_j+\Delta t),
\end{aligned}$$

which yields

$$\lim_{\Delta t \to 0} \lambda_j(t+\Delta t) = \lambda_j(t) = \frac{d_j(t+L_j)}{x_j},\ t \geq 0^+. \tag{2.9}$$

As a result, it is shown that the release rates for item j at time $t \geq 0^+$ as given at time zero chase the demand rates with a time lag equal to the planned lead time of item j. We call these as the expected steady-state release rates.

Equations (2.8) and (2.9) demonstrate the fact that a change in the lead time of item j at time $t = 0$ creates a significant transient effect in the release rate of item j, and the effect on the expected steady-state release rates depends on the stationarity of the assumed demand process. When L_j is changed by ΔL_j then $\lambda_j^*(0)$ is changed by

$$\Delta\lambda_j^*(0) = \frac{d_j(L_j,\, L_j+\Delta L_j)}{x_j}.$$

We note that $d_j(L_j,\, L_j+\Delta L_j) = -d_j(L_j+\Delta L_j,\, L_j)$ for $\Delta L_j < 0$. The change in the expected steady-state release rate of item j at time $t \geq 0^+$ is

$$\Delta\lambda_j^*(t) = \frac{d_j(t+L_j+\Delta L_j) - d_j(t+L_j)}{x_j},\ t \geq 0^+.$$

Given that the demand for item j has a stationary distribution with a fixed mean, $d_j(t) = d$, then the effects of updating the lead time of item j are formulated as

$$\begin{aligned} \Delta\lambda_j^*(0) &= \frac{\Delta L_j \cdot d}{x_j}, & (2.10) \\ \Delta\lambda_j^*(t) &= 0,\, t \geq 0^+. & (2.11) \end{aligned}$$

From Equation (2.7) it is understood that, in a multi-stage structure, dependence of the downstream item release on the upstream item availability in stock generates a smoothing effect when the lead time is increased. Instead of releasing the total effect of lead time increase (see Equation (2.10)), only a portion of it is released based on $I_i(0)$, and the rest is carried onto the next releases in $0 < t \leq 0^+$. Equation (2.11) indicates that, for a stationary demand process, the lead time update effect is only transient. However, this is based on the assumption that the lead time is fixed during the rest of the time.

The release rate for item i at time zero depends on the expected release rates of its downstream item j. That is,

$$\lambda_i^*(0) = \frac{\lambda_j^*(0,\, L_i) \cdot a_{i,j} \cdot x_j - IP_i(0)}{x_i},$$

which is rewritten by

$$\lambda_i^*(0) = \frac{d_j(0,\, L_i + L_j) \cdot a_{i,j} - IP_j(0) \cdot a_{i,j} - IP_i(0)}{x_i}. \tag{2.12}$$

The term $d_j(0,\, L_i + L_j) \cdot a_{i,j}$ is the total estimated demand for item j, expressed in terms of units of upstream item i, during a period of cumulative supply chain lead time. Similarly, $IP_j(0) \cdot a_{i,j}$ is the downstream item inventory position at time zero expressed in terms of units of item i. Equation (2.12) clearly states that the release rate of the upstream item i is given according to a generalized base-stock policy such that the echelon target stock for item i is the total expected final product demand, in terms of units of item i, during the cumulative echelon lead time of $L_i + L_j$. As a result, a change in the planned lead time of the downstream item j has a direct effect on the release pattern of the upstream item i.

2.2.3 Released Order Scheduling

The planned lead times used for scheduling purposes are provided by the tactical level, and the solution of the SCOP formulation provides the optimal order releases for every item over the entire planning horizon. Due to the rolling horizons method, only the first planning period's ordering decisions are passed to the operational scheduling level. Order crossovers are not allowed, which means orders are processed in accordance to the FCFS strategy. However, between the orders released in the same period a random selection rule is applied in determining their production sequences.

Lot-for-lot lot-sizing strategy is applied at the production units. In addition, lot splitting is not allowed, that is, the items produced within an order are not sent to the corresponding stock point until the entire production lot has been processed. This is done in order to be consistent with the material flow assumption (constraint set (2.3)) applied at the operational planning level.

At each replanning opportunity new state information from the execution system may result in a change in the schedule of open orders at each production unit in accordance with the capacity restrictions. The new schedules, corrected according to the new status, are then feedforward to the operational planning level as inputs. The idea is to feed the SCOP model with realistic schedules so that capacity restrictions can be anticipated effectively in making new order release decisions. This role for the operational scheduling level can only be identified in a dynamic framework.

For every item $i \in S_u$, the schedules for open orders are updated according to the following procedure based on the current workload in the production unit.

1. Set the latest expected completion time of all open orders in production unit u as $t^c = \min\{t;\, t \cdot C_u \geq \sum_{j \in S_u} B_j(0)\}$. Go to Step 2.

2. For all $i \in S_u$, set $\widehat{Q}_i(t^c) = \sum_{s=t^c+1}^{T-1} \widetilde{Q}_i(s)$, and $\widehat{Q}_i(t) = \widetilde{Q}_i(t)$ for $t \leq t^c$. Go to Step 3.

3. For all $i \in S_u$, set $\widehat{Q}_i(1) = \sum_{s=-\infty}^{0} \widetilde{Q}_i(s)$. Stop.

The logic behind this procedure is to reschedule all the orders ahead of their planned schedules to their latest expected completion time, and to reschedule all the currently late orders to the end of the current period. The former is done in Step 2 of the procedure, and the latter is done in Step 3 of the procedure. The schedule for the set of orders that does not fit into one of these cases is kept unchanged. Note that $t^c = L_i^{\min}$, which directly implies that order crossovers are avoided through the schedule updates before making new release decisions at the operational planning level.

2.3 Simulation Experiments

2.3.1 Setting

A serial supply chain is considered. There are two final products produced from two separate intermediate items, and each unit of a final product requires one unit of its intermediate item, $N = \{1, 2, 3, 4\}$, $N_e = \{1, 2\}$, $a_{3,1} = 1$, $a_{3,2} = 0$, $a_{4,1} = 0$, $a_{4,2} = 1$. Items sharing common resources interfere causing a high level of variability in order flow times, which motivates using dynamic planned lead times. To keep the complexity at a reasonable level, the production units both for the intermediate and for the final production stages are assumed to possess identical capacities. The available capacity per period for each of the production units is $C_u = 100$ time-units/period, $u = 1, 2$. Figure 2.2 provides an illustration of the product and process structure.

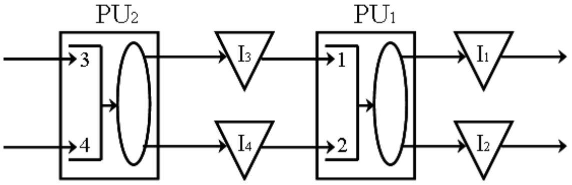


Figure 2.2: The product and process structure of two-stage serial supply chain.

The unit lot-sizes are identical, $x_i = 50$, for all $i \in N$. The processing time per unit lot-size follows an exponential distribution with mean $\tau_i = 50$ time-units, $i \in N$. At this point, it is important to restate that the detailed production process characteristics are not modeled. Considering complex flow of material within the shop floor, machine breakdowns, maintenance, etc., a highly variable production process may become realistic. Given that the average demand rates for the final products are identical, then the utilization of each production unit is shared identically between the items produced in that unit. Demand forecasts throughout the entire planning horizon are kept at a fixed level equal to the mean demand. A 90% utilization is set for each production unit, thus $D_i(t) = 45$ units/period, for all $i \in N_e$. However, the actual demand may as well deviate from the forecast.

The planned lead times for the intermediate items are considered as static parameters. Preliminary experiments showed that the average flow times are realized close to two periods, and with an additional one period of safety lead time, the system performs reasonably well. Therefore, the fixed lead times through PU_1 and PU_2 are set equal to three periods. The cost parameters are taken as follows: $h_3 = h_4 = 1.0$ and $h_1 = h_2 = 1.5$. The total costs presented in the results are computed disregarding the penalty costs of shortages.

The initial planned lead times for the final products are set equal to three periods. The initial backorders for all final products are set to zero and the initial inventories of the final products are equal to the average demand during the initial planned lead time in order to shorten the warm-up period in the simulation. In a similar manner, the initial inventories for an intermediate item is equal to the initial planned lead time for that item multiplied by the unit lot-size of its final product. The production units are initially considered as empty.

2.3.2 Design

Different aspects of the problem have been considered in designing our experiments. We mainly concentrate on the evaluation of dynamic planned lead times with respect to fixed planned lead times. Additionally, we consider aspects regarding the supply characteristics, the decision functions in the hierarchy, the update frequency, and the demand uncertainty. The list of design factors and their corresponding levels of treatment in the experiments are given in Table 2.1.

Table 2.1: Experimental design factors.

Factors	Treatments	Number of Treatments
Lead Time, L_i, $i \in N_e$	3, dynamic	2
Intermediate Stock, SP_i, $i \in N \backslash N_e$	∞, $< \infty$	2
Operational Planning Model	SCOP, DRP	2
SCV of Demand, SCV_D	0.25 (low), 0.50 (high)	2
Smoothing Parameter, ζ	0.10 (low), 0.50 (high)	2

The concept of planned lead time especially plays a crucial role in multi-stage systems, such that the upstream stage deliveries have an effect on the downstream order releases. To evaluate the significance of such a supply chain effect two different situations are evaluated. $SP_i = \infty$, for $i \in N \backslash N_e$, refers to the situation where the production of final products does not starve due to shortages of upstream intermediate items. It simply refers to a single-stage structure with ample supply of raw materials. Differently, $SP_i < \infty$, for $i \in N \backslash N_e$, refers to the situation where the upstream material availability plays a role in limiting the order releases for the downstream items. As also implied by Equation (2.7), we expect that such a supply chain effect would be significant in determining the performance with dynamic lead times.

The role of the planned lead times depends on the extent to which the operational limitations of the production units are anticipated at the operational planning level. Absence of anticipation on the production capacities leads to a DRP approach where the order sizes depend only on the demand forecasts and the current inventory position. Under the DRP approach, planned lead times determine the probability that an order will be available at the due-date established by the planning system (cf. Enns and Suwanruji (2004)). Short planned lead times decrease the delivery performance due to tardy orders. On the other hand, long lead times cause excessive inventory carried in the supply chain. SCOP implies more anticipation on the operational limitations of the production units, and applies a control mechanism to avoid unattainable order releases. Therefore, shorter planned lead times cause order deliveries become more reliable due to more strict workload limitations. However, we expect to see large variations in the inventory levels because of the additional delays in transferring demands as workloads to the production units. Thus, different operational interactions are in charge depending on what kind of an operational planning model we apply, and its effects on the performance of dynamic planned lead times are worth analyzing.

The frequency of updating is also an important design issue in dynamic systems. In this chapter, this is included in the experiments via the smoothing constant ζ. Since the planned lead times are integer valued, a low value of ζ results in a low frequency of update, and a high value of ζ results in a high frequency of update. Usually, a significant change in the update frequency is not possible within a certain range of ζ values. The ζ values of 0.10 and 0.50 in Table 2.1 are respectively chosen as representatives of a low frequency range and a high frequency range.

In addition to such design related issues, the level of uncertainty in demand is considered such that both of the final products possess identical demand processes following a Gamma distribution. The squared coefficient of variation in the actual demand can be either low, $SCV_D = 0.25$, or high, $SCV_D = 0.50$. The low variation is modeled by a Gamma(4, 45/4) distribution, and the high variation is modeled by a Gamma(2, 45/2) distribution.

A full factorial design is employed where a total of 24 different experimental treatments are simulated in 15 repetitions, each during 5250 periods. Each replication is performed with different random number streams and the same set of random number streams is used between different treatments. The data for the first 250 periods is discarded as part of a warm-up duration. Welch's procedure (see Law and Kelton (2000) for a complete description) is applied to compute the warm-up duration.

2.3.3 Results

We are mainly interested in the external and the internal performance measures. The external measures are related to the final product demand satisfaction such as the safety stock levels and the total inventory holding cost in order to guarantee a certain service level, such as the final product demand fill rate. The target fill rate is set equal to 98% for each final product. For each setting an initial simulation run is taken with safety stock levels equal to zero (see SCOP formulation in Section 2.2.2), then the safety stock adjustment procedure in the Appendix of this chapter is applied in order to set the fill rates at the desired level. The internal performance measures are related to the delivery of the released orders, such as the average flow times, the forecast error of the planned lead times, and the percentage of orders that are tardy. In this way, we will be able to derive relationships between updating the lead times, the inner dynamics of the planning system, and their cost implications. The forecast error in the planned lead times is modeled as the mean squared deviation of the planned lead times from the order flow times.

Tables 2.2 and 2.3 provide the performance measures in terms of the relative increase caused by updating the lead times respectively for situations with $SCV_D = 0.25$ and $SCV_D = 0.50$. We use the following abbreviations in this section in interpreting the simulation results.

$SS\%$ = Percentage increase in the sum of the safety stocks of two final products due to updating the lead times.

$TC\%$ = Percentage increase in the total cost due to updating the lead times.

$L\%$ = Percentage increase in the average planned lead times of the orders for the final products due to updating the lead times.

$F\%$ = Percentage increase in the average flow times of the orders for the final products due to updating the lead times.

$\Delta L\%$ = Percentage increase in the forecast error of the planned lead times of the final products due to updating the lead times.

$\Delta\Pi$ = Difference in the percentage of tardy orders for final products between the static and the dynamic cases.

We consider the static case as a base case, and evaluate the performance of the dynamic lead times with changing update frequencies in comparison to the base case. The values with ”†” refer to the cases where we can reject the hypothesis that the dynamic and the static case values are different according to a 95% confidence level.

Table 2.2: The relative performance of the dynamic lead times over the static lead times, $SCV_D = 0.25$.

	$SP_i < \infty$, $i \in N \backslash N_e$				$SP_i = \infty$, $i \in N \backslash N_e$			
	SCOP		DRP		SCOP		DRP	
	$\zeta = 0.1$	$\zeta = 0.5$	$\zeta = 0.1$	$\zeta = 0.5$	$\zeta = 0.1$	$\zeta = 0.5$	$\zeta = 0.1$	$\zeta = 0.5$
$SS\%$	25.07	36.26	25.79	45.98	48.97	77.25	39.16	71.10
$TC\%$	3.64	$1.75^{\dagger}$	5.72	5.10	7.13	9.75	11.22	12.45
$L\%$	-3.86	$0.21^{\dagger}$	5.33	7.20	-3.94	4.09	12.59	19.32
$F\%$	5.45	21.04	11.93	26.01	6.92	29.46	21.76	45.82
$\Delta L\%$	-12.65	-20.28	1.56	-15.05	-14.15	-19.24	17.76	0.64
$\Delta\Pi$	2.47	4.36	0.89	2.89	2.91	5.05	1.68	4.59

Table 2.3: The relative performance of the dynamic lead times over the static lead times, $SCV_D = 0.50$.

	$SP_i < \infty$, $i \in N \backslash N_e$				$SP_i = \infty$, $i \in N \backslash N_e$			
	SCOP		DRP		SCOP		DRP	
	$\zeta = 0.1$	$\zeta = 0.5$	$\zeta = 0.1$	$\zeta = 0.5$	$\zeta = 0.1$	$\zeta = 0.5$	$\zeta = 0.1$	$\zeta = 0.5$
$SS\%$	6.30	13.91	4.61	20.62	15.19	33.26	7.87	30.01
$TC\%$	$1.07^{\dagger}$	$0.45^{\dagger}$	2.58	5.16	4.46	8.71	8.81	12.75
$L\%$	$-0.27^{\dagger}$	3.58	12.47	14.63	$0.25^{\dagger}$	9.14	26.32	35.30
$F\%$	5.15	20.23	11.39	26.11	7.13	30.61	26.65	55.80
$\Delta L\%$	-2.86	-10.07	10.05	-6.85	-2.45	-6.80	44.36	24.43
$\Delta\Pi$	1.63	3.85	-1.37	1.12	2.03	4.57	-1.03	3.13

One obvious result is that, in all settings, the dynamic lead times increase the total cost of the supply chain. In a situation with limited supply from the upstream stage, when the order releases are controlled via the SCOP model, the increase in the total cost mostly fails to be significant. In all other cases, there is a significant increase up to almost 13%. This result supports our intuition stated in Section 2.1. Taking periodic feedbacks of realized order flow times and then updating the planned lead times in order to keep them as close as possible to the flow times generates worse performance. During an inventory shortage period, backorders are increased further due to longer order flow times because, they increase the congestion and delay the planned deliveries. When the shortage period is over, the inventory is increased extensively because, the pipeline has already been filled up with large orders due to long planned lead times. Updating the lead times creates increased variations in inventory levels, which increases the holding costs. An illustration of this phenomenon is provided in Figure 2.3 for a single final product. $I^{(VL)}$ and I refer to the on-hand inventory levels with dynamic and fixed lead times respectively. Their values are smoothed for illustration purposes. VL denotes the dynamic planned lead time values of that final product.

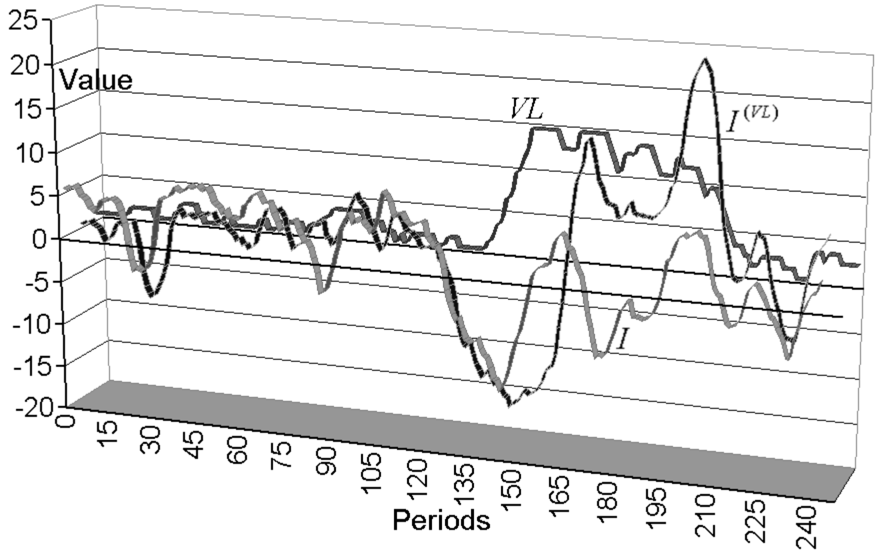


Figure 2.3: High inventory variation caused by dynamic planned lead times.

There is a severe temporary supply shortage between approximately 100^{th} and 150^{th} periods, which yields increased lead times and in turn increased order flow times. After the 150^{th} period, what is put into the pipeline started to be delivered more quickly yielding an increase in the inventory level. When one looks at the behavior of inventory levels between the 100^{th} and 200^{th} periods, $I^{(VL)}$ oscillates in much larger amplitude than I does. This phenomenon is simply due to the fact that the correlation between planned lead times and order flow times cannot be modeled at the order release planning level. The simulation results indicate a strong correlation between these two factors with a correlation coefficient that is always greater than 0.97. Long order flow times trigger further increase in planned lead times yielding an uncontrolled situation in updating the lead times. Starting from a level of two periods at period 130, the planned lead time increases steadily and steeply and reaches its maximum limit at period 150. Further (analytical) analysis of this phenomenon is carried out for a simpler setting in Chapter 3 of this thesis.

Tables 2.2 and 2.3 provide various insights into the effects of updating the planned lead times in a multi-stage production-distribution environment. As the system is controlled by the DRP approach, $TC\%$ and $\Delta L\%$ become significantly larger than those under SCOP. This means updating the lead times is more degrading in terms of the cost performance and the consistency of the planned delivery schedules under DRP than it is under SCOP. This is because, SCOP has more detailed anticipation about the operational characteristics of the production units (specifically about attainable workload levels) when compared to DRP, which disregards capacity. This causes the order releases to have more erratic pattern as the lead times are updated under DRP because, limitations on workload levels in the SCOP formulation smooth the effects of dynamic lead times. Especially under high demand variability and with limited upstream supply, $TC\%$ fails to be significant when the SCOP formulation is applied. Additionally, the decreased forecast errors in these cases even suggests SCOP together with the dynamic planned lead times as a favorable planning tool.

Another interesting result is that both $L\%$ and $\Delta\Pi$ are positive under DRP (except the cases with $SCV_D = 0.50$ and $\zeta = 0.1$), which seems to contradict the intuition that delivery becomes more reliable with longer planned lead times. The fundamental assumption behind this intuition is the independence of the release mechanism from the changed lead times, which is violated in our study due to the hierarchical nature of decision making.

More safety stock is needed in order to satisfy the target fill rate when ζ is increased from 0.1 to 0.5, since the amplification in order sizes is stronger with more frequent updates during temporary shortage periods. In a DRP managed environment, $SS\%$ almost doubles under $SCV_D = 0.25$ and increases almost 4 times under $SCV_D = 0.50$ with more frequent updating. When the SCOP formulation is applied, the system is less sensitive to higher update frequency than it is when DRP is applied. $L\%$ and $F\%$ increase significantly with the update frequency because, frequent updating imposes larger variance in the ordering pattern. Since the flow times of successive orders are highly correlated, an increased smoothing constant improves the forecast accuracy, $\Delta L\%$. This is because, more weight is given to the latest occurrences of order flow times when ζ is larger. On the other hand, $\Delta\Pi$ increases significantly with more frequent updating due to the amplification effect in the order releases during the shortage periods. These two effects play contradictory roles in determining the total average inventory kept in the system, because larger $\Delta\Pi$ implies larger safety stocks, whereas smaller $\Delta L\%$ implies less excess stocks.

Therefore, the effect of update frequency on the total costs is not very obvious, especially for situations with limited supply of intermediate items. When there is unlimited availability of intermediate items, the erratic behavior becomes more significant, and higher update frequency yields more costly planning outputs.

The interdependencies between the downstream and the upstream order releases, together with the effect of updating the lead times have been formulated in Equations (2.10) and (2.12) for a DRP managed environment. The simulation results are in line with the intuition gathered from this formal analysis. All of the performance measures degrade when the final product order releases are planned with the assumption of infinite supply of intermediate items. This reveals the idea that in a situation with increased variability it is favorable to smooth the erratic planning outputs either through intelligent tools such as SCOP or naturally through the multi-stage structure of the supply chain.

2.4 Conclusion

In this chapter we have provided various insights into the application of dynamic planned lead times in managing the material flow within a two-stage serial supply chain. Our purpose was to provide simulation results that would yield a better understanding of the inner dynamics of updating the planned lead times used in a hierarchical planning system. The results can be analyzed from different angles including the supply chain structure, hierarchical anticipation, and the update frequency. In this chapter, we have shown the following:

- Updating the lead times in response to the latest occurrences of order flow times generates erratic order releases and large variations in inventory levels.
- The degrading effects of dynamic lead times can be smoothed by improving the anticipation, at the operational planning level, on the operational limitations of the production units. Ignoring capacity causes worse performance of dynamic lead times.
- A higher frequency of updating the lead times decreases the deviation of flow times form the lead times but, increases the nervousness in the delivery schedules, and the percentage of tardy orders increases.
- The specific supply chain structure is very significant in determining the relative performance of using dynamic planned lead times. Upstream inventory smoothes the downstream variability in order releases.

Through our findings from the simulation experiments performed in this chapter, we have identified interesting research questions that are worth to be studied. First of all, the erratic order release phenomenon due to updating the lead times should be analyzed in-depth with analytical results. This may help in stronger understanding of the results we have described in this chapter. The level of anticipation at the operational planning level, and its impacts on the external and the internal performance measures need to be analyzed further with more detailed anticipation functions. Although our findings in this chapter do not support the use of dynamic planned lead times, it is still intuitively appealing that a responsive planning mechanism needs to be developed in order to better manage the supply chains. These issues are the topics covered in the following chapters of this thesis.

Appendix to Chapter 2

Safety Stock Adjustment Procedure

The safety stock adjustment procedure applied in this thesis is adapted from Kohler-Gudum and De Kok (2001). This procedure enables the determination of safety stocks that ensure target service levels in simulation studies of inventory systems. Various kinds of service measures have been considered in Kohler-Gudum and De Kok (2001). The procedure we applied is based on *demand fill rate* measure; the fraction of demand satisfied directly from stock.

Let SS_0 denote the initial choice of the safety stock, and $\Psi_e(SS_0)$ is the (empirical) variable for the net stock at the end of a period based on the simulation of the system with the safety stock SS_0. Similarly, $\Psi_b(SS_0)$ denotes the net stock at the beginning of an arbitrary period, immediately after (possible) arrival of a replenishment. In order to determine the relevant probability distributions, define ψ_0 and ψ_K respectively as the minimum and the maximum recorded net stock value for $\Psi_e(SS_0)$ during the simulation. The intermediate values can then be determined by

$$\psi_k = \psi_0 + \frac{k}{K}\left(\psi_K - \psi_0\right), \quad k = 1, \ldots, K-1$$

where $K+1$ is a chosen number of probabilities for the distribution of $\Psi_e(SS_0)$. In our case, $K = \psi_K - \psi_0$ so that all the integer values between ψ_0 and ψ_K are considered in the distribution.

Let us define $y_k = \Pr\left\{\Psi_e(SS_0) \leq \psi_k\right\}$, and $z_k = \Pr\left\{\Psi_b(SS_0) \leq \psi_k\right\}$. Accordingly, the average backorder level, $\bar{B}(\psi_k)$, depending on a safety stock adjustment quantity of ψ_k is computed by

$$\bar{B}(\psi_k) = \sum_{j=0}^{k-1}(y_j - z_j), \quad k = 1, \ldots, K-1.$$

Define ψ_{s^*} as the adjustment quantity that guarantees the target fill rate s^*, $0 \leq s^* \leq 1$. Then, $(1-s^*)\bar{D} = \bar{B}(\psi_{s^*})$, where $\bar{D}$ denotes the average demand per period. From y_k and z_k values we must find ψ_s such that $\bar{B}(\psi_{s-1}) \leq (1-s^*)\bar{D} \leq \bar{B}(\psi_s)$. Consequently, ψ_{s^*} can be found by linear interpolation:

$$\psi_{s^*} = \frac{\left((1-s^*)\bar{D} - \bar{B}(\psi_{s-1})\right)\psi_s + \left(\bar{B}(\psi_s) - (1-s^*)\bar{D}\right)\psi_{s-1}}{\bar{B}(\psi_s) - \bar{B}(\psi_{s-1})},$$

and the new safety stock value is $SS^* = SS_0 - \psi_{s^*}$. The simulation repeated with SS^* gives us a demand fill rate of s^*.

Chapter 3

Lead Time Syndrome: Formulation and Analysis

Updating planned lead times in response to changing workload levels leads to erratic ordering behavior, resulting in even larger variability in work-in-process and flow times. This phenomenon is called the lead time syndrome. Although it has been conceptually defined and intuitively accepted, formal analysis has not been conducted in any prior study. The objective of this chapter is to provide a clearer and more formal understanding of this phenomenon by enriching our numerical findings in Chapter 2 through analytical results. A single-stage, single-item produce-to-order situation is considered with the order releases sensitive to the planned lead time. The situation is modeled by a two-dimensional Markov process that is solved by using the matrix-geometric methods. Analytical results on the utilization level and the variability in the system are presented in relation to various design parameters.

3.1 Introduction

Since Conway *et al.* (1967) the problem of lead time setting, mainly for make-to-order situations, has attracted much attention. This is due to the fact that the lead time is a fixed planning parameter that refers to a dynamic and uncertain frame on a continuous time axis. Numerous techniques have been developed of which the most popular ones utilize order characteristics together with dynamic shop-load information. The majority of the studies in this line (e.g., Wein (1991), Enns (1995), and Hopp and Sturgis (2000)) are conducted in a make-to-order job-shop environment to set lead times for *externally* generated orders according to service-related performance measures such as the length of the lead time, tardiness, earliness, etc. One of the basic assumptions is that the order characteristics are determined externally and independent from the planned lead time, which generally holds true for engineer-to-order situations. However, for example in batch processing industries where orders are released and processed in large batches of production items, the orders are generally released according to some anticipated knowledge on the total demand levels during the planned lead times. In a multi-stage production-inventory system, one would expect that the planned lead times are used for coordination purposes between stages. In the previous chapter, we have illustrated the erratic order release patterns for such a situation where the planned lead times are based on the exponential smoothing of the order flow times.

Mather and Plossl (1978) are the first to describe the lead time syndrome as a vicious cycle between lead time update and order release decisions. It is argued that closing the gap between the lead time and the order flow times by updating the lead time results in uncontrolled order release pattern. As the lead time gets longer, orders must be released earlier to cover increased expected demand during the longer lead time, leading to longer queues of production backlog and thus, flow times get longer, which causes again a longer lead time. It results from the fact that in releasing the orders, the dynamic effect on future lead times and on future orders is ignored. It is suggested that the lead time syndrome causes instability, and should be avoided. This reasoning has become one of the main arguments for controlling flow times within predetermined norms instead of forecasting them (e.g., Plossl (1988), Kingsman *et al.* (1989), Zäpfel and Missbauer (1993), and Breithaupt *et al.* (2002)). However, an important issue that needs to be clarified at this point is that Mather and Plossl (1978) and their followers described this phenomenon in a very general view only conceptually without a clear evidence of its existence and without any formal results. The simulation results in Chapter 2 indicate the existence of this phenomenon especially when the orders are released by a DRP system. In this chapter, we provide analytical results on the consequences of this phenomenon in a simpler setting.

The lead time syndrome is becoming increasingly relevant due to the opportunities of frequent information exchange enabled by recent advances in data processing and storage technologies. Ubiquitous information such as inventory levels, work-in-process, and shop conditions at various stages of a supply chain may continuously be available to update planning parameters. An important question is to what extent all this information should be used to update plans and planning parameters. Insights, based on the lead time syndrome, would suggest to make limited use of these updating capabilities due to the increased variability in the order releases in response to the operational changes.

Based on the insights gathered from the previous chapter, let us describe the lead time syndrome by a simple example. Consider a production unit and a downstream stock point where the orders for a single item are released based on the DRP logic with ample supply of raw materials. An illustration of the lead time syndrome is provided in Table 3.1 for a duration of 12 periods. In each period, the following sequence of events occurs: the lead time is set, the order for that period is released, produced items are delivered to the stock point, and demand is realized. The downstream customer demand is deterministic with a fixed level of 3 units/period, and the periodic production may vary with an expected quantity of 3 units/period. Without loss of generality, assume the initial lead time is two periods with zero on-hand inventory and with a total workload of six units, which together imply that an order of 3 units puts the DRP system in balance. Therefore, as Equation (2.9) suggests in a continuous time frame, orders of sizes equal to the level-demand are released in each period in the static case. In the dynamic case, the lead time is set as the minimum number of periods within which the total workload is expected to be finished. Orders are released to raise the inventory position up to a level equal to the expected demand during a duration of planned lead time plus one period.

Table 3.1: Order lead time sheet.

Period	Workload	Inventory	Lead Time	Order	Production
1	6	0	2	3	3
2	6	0	2	3	0
3	9	-3	3	6	3
4	12	-3	4	6	3
5	15	-3	5	6	3
6	18	-3	6	6	6
7	18	0	6	3	6
8	15	3	5	0	3
9	12	3	4	0	3
10	9	3	3	0	3
11	6	3	2	0	0
12	6	0	2	3	3

In period 1, the production unit finishes three units, and with a lead time of two periods, receives three units of order, while keeping the total workload at six units. In period 2, production cannot occur due to some temporary breakdowns, and the total workload increases to nine units. Therefore, the lead time is increased to three periods. The order in period 3 now includes the static case order plus an additional period's demand, adding up to six units, thereby increasing the total workload to 12 units. As a result, the lead time is increased to four periods, and in period 4 an order of six units is placed again. The vicious cycle continues, and in period 6 the lead time is increased to six periods, and the workload to 18 units. In this period the production unit works faster and produces 6 units. Thus, the lead time is not changed in period 7 and a static order of 3 units is placed. The production goes faster again in period 7, decreasing the workload to 15 units and the lead time to five periods in period 8. The new order now includes the static order minus the excess of one period demand, which results in the cancelation of the static order. This causes a further decrease in the workload and therefore in the lead time. The cyclic effect continues, and in week 11, the lead time becomes two periods again.

From Table 3.1, one should notice that, although there is a fixed and deterministic demand each period, starting from a balanced DRP, a temporary (one period) shortage of 3 units in the production leads to the workload level to increase from 6 units to 18 units within 4 periods. Again starting from a balanced DRP, a temporary (one period) increase of 3 units in the production quantity leads to the workload level to decrease from 18 units to 6 units within 4 periods. In short, the deviation in the production quantities is amplified and carried onwards the workload levels through erratic order releases caused by updating the lead time. This behavior seems quite arbitrary for a rational planner but, the phenomenon is generally considered to be a relevant problem in real life decision support systems (cf. IBM (1972)). As they have been outlined by the example in Table 3.1, the basic dynamics of this behavior stems from the presence of inter-dependent activities being reactive with improper anticipation to the changes in each other. Economics studies based on some specific cases of producers' and consumers' behaviors in response to the changing expectations on future market conditions have reported the existence of similar phenomenons (e.g., the pig-cycle in Coase and Fowler (1937)).

In this chapter, our objective is to provide formal insights into the various effects of the lead time syndrome on the production unit that faces a random ordering pattern. As it is shown by the example in Table 3.1, lead time syndrome has a significant effect on the workload level, and an explicit model of this phenomenon may yield interesting results about some related performance measures such as utilization, flow times, and etc. At the production unit level, the process can be considered as a single-stage single-item produce-to-order situation where the orders are released sensitive to the planned lead time. In order to avoid the curse of dimensionality, we assume that the status information downstream to the production unit is invisible to the order generating process. That is, order releases simply chase the demand in addition to responding to the changes in the planned lead time. Deviations in the on-hand inventory kept in various downstream or upstream stages of the supply chain are not considered. For example, consider a production unit with ample supply of raw materials where the production orders are released based on the point-of-sale transactions. For each unit lot-size of realized demand an order of equal size is released, and the order size is adjusted in response to the change in the planned lead time such that one period increase/decrease in the planned lead time generates a fixed increase/decrease in the order size. In Table 3.1, ignoring the inventory levels and releasing the orders in this manner does not make any differences on the workload levels and on the order release pattern. Therefore, the simplified problem setting is a realistic representation to evaluate the consequences of the lead time syndrome from the production unit control perspective. Analytical results are then provided for two issues raised by the lead time syndrome:

- How is the stability of the production unit affected when the lead time is updated?
- What is the effect of the update frequency on the performance of the production unit?

3.2 Problem Setting

We model a single-stage single-item produce-to-order situation. Under static conditions, orders of unit lot-size are placed according to a Poisson process with rate λ. Due to updating the lead time, orders can be released in multiple number of unit lot-sizes. During the rest of this chapter, each unit lot-size is referred to as a single job in the queueing system (production unit), which is composed of a single queue to backlog the released jobs and a single shop that actually processes the jobs. w denotes the number of jobs being processed (WIP) in the shop, and the total workload is $\hat{w} = w + b$, where b is the number of jobs waiting in the production backlog to be loaded to the shop floor. There is a WIP limit, $\bar{w}$, that indicates the size and the speed of the shop floor with respect to a single job. The shop is assumed to handle at most $\bar{w}$ jobs at the same time, and when the shop is totally loaded, arriving jobs are put in the backlog queue, and loaded to the shop each time a job is completed and leaves the shop. The backlog queue is modeled as a single-server system with FCFS processing discipline. The shop is considered as a single entity where the WIP can be cleared within exponentially distributed time intervals with rate μ. An illustration of our queueing system is given in Figure 3.1. The following examples may further clarify the situation.

Example 3.1 *The shop has a single machine that operates with exponentially distributed processing times with rate μ jobs/time-unit. There is a buffer in front of the machine that can store at most $\bar{w} - 1$ jobs at the same time.* □

Example 3.2 *The shop has multiple manufacturing centers. Jobs can follow different routes in the shop. The possibility that different jobs interfere (setups, processor sharing, etc.) with each other depends on the WIP level in the shop. Each job is processed fast when the WIP level is low and vice-versa. Jobs are processed in parallel with independent, identically distributed exponential processing times with mean w/μ time-units. Since the minimum of the exponentials is also exponential with rate equal to the sum of the rates, the overall processing rate is μ jobs/time-unit irrespective of the WIP level in the shop.* □

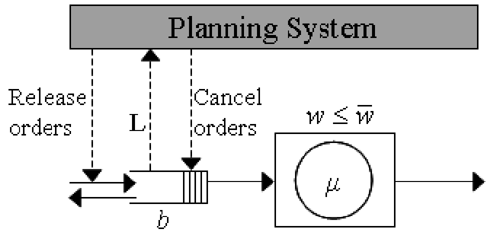


Figure 3.1: Single-server queueing system with dynamic lead time.

Figure 3.1 provides the coupling mechanism between the planning system and the physical situation. The flow of information is represented by dashed lines, and the solid lines refer to the physical flow of goods. The planning system is responsible for determining the number of jobs released and the production authorization of the jobs in the backlog queue with the existence of continuous feedback about the status information to update the planned lead time.

The planned lead time is determined based on the expected flow time of the last job currently residing in the backlog. When the backlog queue is empty, $\hat{w} \leq \bar{w}$, the lead time is set to a fixed level, $L_{\min}$, referring to the minimum estimated flow time. When there are jobs waiting in the backlog, then an estimate of the waiting time in the backlog is added to $L_{\min}$. From the memoryless property of exponential processing times it is straightforward that, at any point in time, the expected duration of time to clear b jobs through the production process is b/μ. Here, the lead time is determined based on a logic similar to those procedures that have been widely applied in the literature such as TWK, and JIQ (e.g., Conway *et al.* (1967), Kanet (1986), Vig and Dooley (1991), and Chang (1994)). The waiting time is set by a management constant multiplied by the term, b/μ. Since μ is constant, without loss of generality, we can write the lead time as a function of the total workload level as follows:

$$L = L_{\min} + \lfloor \alpha \cdot b \rfloor,$$

where α is a management constant, and the lead time is an integer. In words, the lead time is based on the management perceptions on the range of the number of jobs that can be cleared in a certain time frame. Implicitly, α refers to the update frequency. For greater α, the lead time is updated more frequently, and for smaller α, the lead time is updated less frequently. We define the reciprocal of the update frequency, $r = 1/\alpha$, as the amount of increase or decrease in the number of jobs in the backlog, b, in order to have one unit of increase or decrease in the lead time respectively. The update parameter r is used for modeling purposes, and is integer valued. Each time a change is triggered in the planned lead time, it is for one unit. Accordingly, the update parameter r has to be greater than or equal to two. This is also intuitively clear because, usually, the backlog level needs to be changed by more than one job in order to consider a change in the lead time.

In a realistic setting, although the planning process is a very complex task including human intervention, the underlying relationship as has been described in Section 3.1 and in Chapter 2 still remains valid. Depending on the situation, the degree of the reaction to the change in the lead time may vary. As the insight provided by Equation (2.10), the response to the change in the lead time depends on the traffic intensity, λ. In a general view, we can model the response through two different decision functions: $h^{+}(\lambda)$ and $h^{-}(\lambda)$. When the lead time is increased, additional $h^{+}(\lambda)$ units are ordered, and when there is a decrease in the lead time, excess $h^{-}(\lambda)$ units waiting in the production backlog are canceled. From the perspective of the production unit the jobs are canceled, but from the systems perspective the jobs do not disappear but are suspended from the production backlog until the time new orders are released. Given that the change in the lead time is limited to at most one period, Equation (2.10) implies $h^{+}(\lambda) = h^{-}(\lambda) = \lambda$. However, due to the various factors in the planning process, the response can be different. For example, the planner may anticipate a trendy increase/decrease in the lead time and over-reacts by setting $h^{+}(\lambda) > \lambda$, and $h^{-}(\lambda) > \lambda$. On the other hand, the planner may react to dampen the variability, and smoothes the production orders by setting $h^{+}(\lambda) < \lambda$, and $h^{-}(\lambda) < \lambda$. In this study, for modeling purposes, the response to the change in the lead time is assumed as follows:

$$h^+(\lambda) = \begin{cases} 1, & \text{with probability}\,\beta \\ 0, & \text{with probability}\,1-\beta \end{cases}$$

and

$$h^-(\lambda) = \begin{cases} 1, & \text{with probability}\,\gamma \\ 0, & \text{with probability}\,1-\gamma \end{cases}$$

Considering the lead time is only increased or decreased by one period, the change in the order size is limited to one job. We fix the response function independent of the traffic intensity. This is based on the practical intuition that planners apply simple and similar procedures to the same events instead of solving complex decision functions. Besides, the degree to which one is reactive to the change in the lead time is incorporated into our model. For greater β and γ values, it is assumed that the planning system is highly sensitive to the change in the planned lead time, and for smaller β and γ values the response is smoothed. $\beta = \gamma = 0$ implies the non-reactive case that the system never responds to the change in the planned lead time, and $\beta = \gamma = 1.0$ implies the full-reactive case that the system always responds to the change in the planned lead time. In a practical setting, such as the DRP formulation in the previous chapter, the response function is generally expected to be symmetric with some positive response probability, $\beta = \gamma > 0$. In the following analysis, we consider a generalized formulation including response functions that may both be symmetric and asymmetric.

3.3 Markov Process

3.3.1 Description

Define $\hat{w}(t)$ and $b(t)$ as the total number of jobs in the system and the number of jobs in the backlog at time t respectively. We model this queueing system as a two-dimensional Markov process defined by $\{X_r(t), t \geq 0\}$, $X_r(t) = (\hat{w}(t), L_r(t))$, where $L_r(t) = L_{\min} + \lfloor \frac{b(t)}{r} \rfloor$ is the lead time set according to the update parameter r. We use r as a subscript because it determines the characteristics of the process. In this Markov process, an arrival refers to a release of a job by the planning system, and a departure refers to a completion of a job in process in the shop. When $b(t) = r \cdot (L_r(t) - L_{\min}) + r - 1$, an arrival triggers an increase in the lead time and an additional job is ordered immediately with probability β. When $b(t) = r \cdot (L_r(t) - L_{\min})$, a departure triggers a decrease in the lead time and a job in the production backlog is canceled immediately with probability γ. As long as $b(t)$ remains in between, the system behaves as an ordinary $M^{\lambda}|M^{\mu}|1$ system. The length of the $M^{\lambda}|M^{\mu}|1$ region (consisting of r states) is short for frequently updated lead times and long for less frequently updated lead times. An illustration of the process $\{X_r(t), t \geq 0\}$ with $r = 4$ is provided in Figure 3.2.

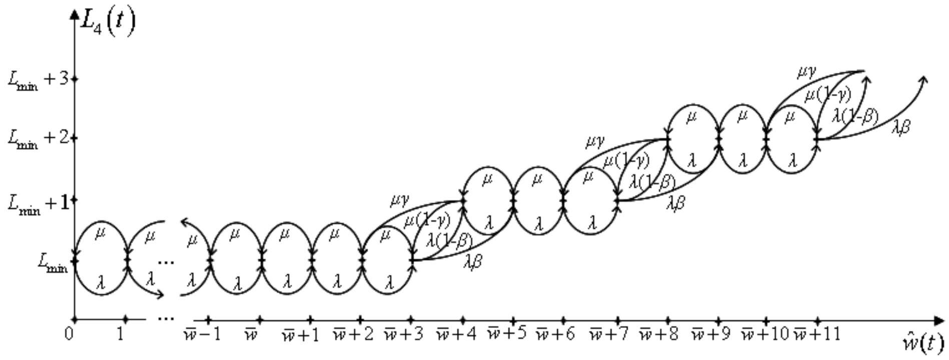


Figure 3.2: Transition rate diagram for the process $\{X_4(t), t \geq 0\}$.

The process has a QBD structure in the diagonal direction of the ($\hat{w}(t)$, $L_r(t)$) coordinates. For the brevity of the presentation, we decompose the overall process into two adjoint parts. The first part is composed of the state space with $\hat{w}(t) < \bar{w}$ and $L_r(t) = L_{\min}$. We denote this process by $\{N(t), t \geq 0\}$, where $N(t) = \hat{w}(t) < \bar{w}$ is the number orders being processed in the shop. The second part of the process is a QBD and defined by $\{Y_r(t), t \geq 0\}$, $Y_r(t) = (\hat{L}_r(t), \hat{b}_r(t))$, $\hat{L}_r(t) = L_r(t) - L_{\min}$, and $\hat{b}_r(t) = b(t) - r\hat{L}_r(t)$. Accordingly, level l of this QBD process is composed of r states with state space as

$$\{(l,\, 0), (l,\, 1), \ldots, (l,\, r-2), (l,\, r-1)\}, \, l = 0, 1, \ldots$$

The transition rate diagram of Figure 3.2 can be transformed into a simpler representation as provided in Figure 3.3.

Throughout the rest of this chapter, we will mainly concentrate on the explicit solution of the QBD process $\{Y_r(t), t \geq 0\}$ and use $\{N(t), t \geq 0\}$ in the normalization equations. To conveniently describe the infinitesimal generator of the process, $\{Y_r(t), t \geq 0\}$, we employ the following notation for r-dimensional square matrices. We denote the identity matrix as $I^{(r)}$, the right and left shift matrices as $T_R^{(r)}$ and $T_L^{(r)}$ respectively. So, $\left(T_R^{(r)}\right)_{i,j} = \delta_{i+1,j}$ and

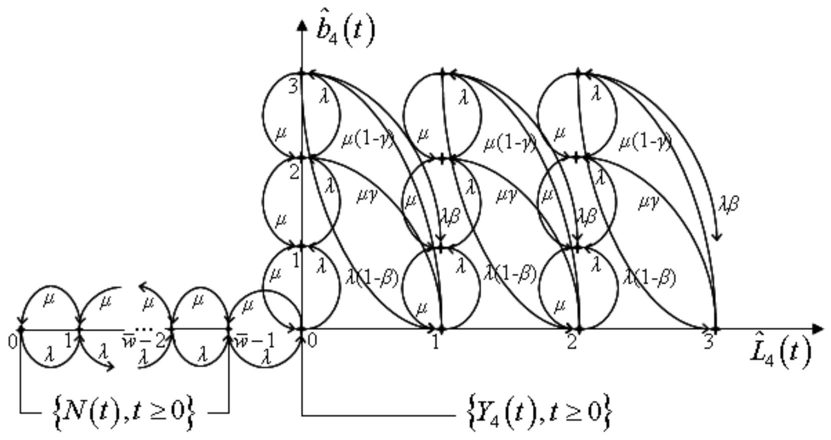


Figure 3.3: Transition rate diagram for the processes $\{N(t), t \geq 0\}$ and $\{Y_4(t), t \geq 0\}$.

$\left(T_L^{(r)}\right)_{i+1,j} = \delta_{i,j}$ for $i, j = 0, 1, \ldots, r-1$, where $\delta_{i,j}$ denotes the Kronecker delta. Furthermore, we define the r-dimensional unit column-vector on the k^{th} coordinate by $e_k^{(r)}$, $k = 0, 1, \ldots, r-1$. Given that the states are in lexicographic order, the generator $Q^{(r)}$ of $\{Y_r(t), t \geq 0\}$ is

$$Q^{(r)} = \begin{bmatrix} B_0^{(r)} & A_0^{(r)} & 0 & 0 & \cdots \\ A_2^{(r)} & A_1^{(r)} & A_0^{(r)} & 0 & \cdots \\ 0 & A_2^{(r)} & A_1^{(r)} & A_0^{(r)} & \cdots \\ 0 & 0 & A_2^{(r)} & A_1^{(r)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}.$$

$Q^{(r)}$ is partitioned into r-dimensional square matrices that provide the transition rates between and within the levels of the QBD process. $B_0^{(r)}$ provides the transition rates between the states of level 0. $A_0^{(r)}$ is the transition rate matrix from level l to level $l+1$, $l = 0, 1, \ldots$. Similarly, $A_2^{(r)}$ is the transition rate matrix from level l to level $l-1$, and $A_1^{(r)}$ provides the transition rates within level l, $l = 1, 2, \ldots$. We represent these matrices as follows:

$$\begin{aligned}
A_0^{(r)} &= \lambda e_{r-1}^{(r)} \left((1-\beta) e_0^{(r)} + \beta e_1^{(r)} \right)^T, \\
A_1^{(r)} &= \lambda T_R^{(r)} + \mu T_L^{(r)} - (\lambda + \mu) I^{(r)}, \\
A_2^{(r)} &= \mu e_0^{(r)} \left(\gamma e_{r-2}^{(r)} + (1-\gamma) e_{r-1}^{(r)} \right)^T, \\
B_0^{(r)} &= A_1^{(r)} + \mu e_0^{(r)} \left(e_0^{(r)} \right)^T.
\end{aligned}$$

3.3.2 The Golden Ratio of Stability

Theoretically, for $\bar{w} \to \infty$ or $r \to \infty$ the lead time is never updated, and the stability condition is $\lambda/\mu < 1$. We define

$$\rho = \lambda/\mu$$

throughout the rest of this chapter. It is the utilization of the shop floor in the static case.

For finite update parameter, r, and finite WIP limit, $\bar{w}$, the stability condition of the QBD process, $\{Y_r(t), t \geq 0\}$, can be derived from *Neuts' mean drift condition* (cf. Neuts (1981)). The Markov process defined by the generator $Q^{(r)}$ is ergodic (stable) if and only if

$$\pi^{(r)} A_0^{(r)} e^{(r)} < \pi^{(r)} A_2^{(r)} e^{(r)}, \tag{3.1}$$

where $e^{(r)}$ is the r-dimensional column vector of ones, and the r-dimensional row vector $\pi^{(r)} = \left(\pi_0^{(r)},\ \pi_1^{(r)},\ \ldots,\ \pi_{r-1}^{(r)} \right)$ is the steady-state probability vector of the Markov process with generator $A^{(r)} = A_0^{(r)} + A_1^{(r)} + A_2^{(r)}$. So,

$$\pi^{(r)} A^{(r)} = \mathbf{0}^{(r)}, \quad \pi^{(r)} e^{(r)} = 1,$$

where $\mathbf{0}^{(r)}$ is the r-dimensional row vector of zeros.

Condition (3.1) has an intuitive interpretation. The generator $A^{(r)}$ describes the behavior of the Markov process $\{Y_r(t), t \geq 0\}$ in the vertical direction, $\hat{b}_r(t)$. Weighted by the steady state probabilities in the vertical direction, if the mean drift to the left, $\pi^{(r)} A_2^{(r)} e^{(r)}$, is greater than the mean drift to the right, $\pi^{(r)} A_0^{(r)} e^{(r)}$, then the process is stable. Condition (3.1) reduces to

$$\frac{\pi_{r-1}^{(r)}}{\pi_0^{(r)}} \cdot \rho < 1. \tag{3.2}$$

In order to explicitly determine the stability condition (3.2) for every update parameter r, we need to have a detailed look at the Markov process defined by the generator $A^{(r)}$. It has r states, and its transition rate diagram is given in Figure 3.4. Utilizing the global balance principle, we can formulate the stability condition of $\{Y_r(t), t \geq 0\}$ as in the following theorem:

Theorem 3.1 *For every update parameter $r = 2, 3, \ldots$, the Markov process defined by $\{Y_r(t), t \geq 0\}$ is stable if the following condition holds:*

$$\rho^r + \beta\rho^{r-1} - \gamma\rho < 1. \tag{3.3}$$

Proof See the Appendix at the end of this chapter.

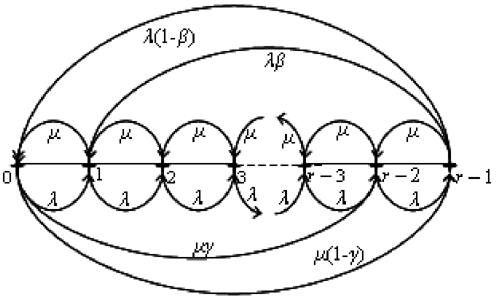


Figure 3.4: Transition rate diagram for the Markov process with generator $A^{(r)}$.

Condition (3.3) provides us with a polynomial function of degree r to determine the stability condition, and solving that polynomial is a very complex task for $r > 3$. However, we can characterize the stability condition by looking at monotonicity properties of the stability function,

$$s(\rho,\, r,\, \beta,\, \gamma) = \rho^r + \beta\rho^{r-1} - \gamma\rho - 1.$$

Condition (3.3) can be rewritten as $s(\rho,\, r,\, \beta,\, \gamma) < 0$, through which the following corollary can be derived from Theorem 3.1:

Corollary 3.1 *The stability condition of the process $\{Y_r(t), t \geq 0\}$ can be rewritten as*

$$\rho < \rho^*,$$

where ρ^* is the unique positive root of $s(\rho,\, r,\, \beta,\, \gamma)$ for given r, β, and γ values, and is bounded by the golden range;

$$\Phi - 1 \leq \rho^* \leq \Phi,\, \Phi = \frac{1+\sqrt{5}}{2}.$$

Proof See the Appendix at the end of this chapter.

The range of values that ρ^* can take is named as *golden range* after the golden ratio, $\Phi \approx 1.618033989$, a well-known irrational number in mathematics. The golden ratio (also known as divine proportion or mean and extreme ratio) has its roots defined by Euclid ca. 300 B.C. (cf. WolframMathWorld (2007)). A line segment is said to be divided into its mean and extreme ratios if the whole segment is to the bigger segment as the bigger segment is to the smaller segment (see Figure 3.5).

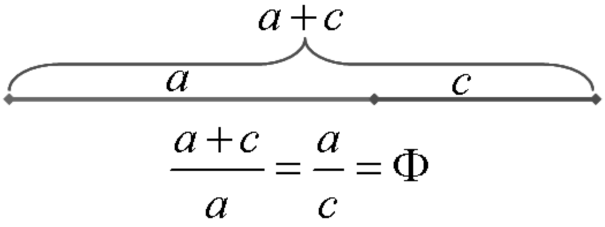


Figure 3.5: Mean and extreme ratios of a line segment of length $a + c$.

The golden ratio has been referred quite frequently in explaining the structure of certain historical and artistic figures such as the Great Pyramid of Giza, and the Da Vinci paintings (cf. Livio (2002)). Surprisingly, it is also used in this thesis in formulating the stability condition of the QBD process $\{Y_r(t), t \geq 0\}$. Furthermore, the golden range of stability can be divided into its mean and extreme ratios at point $\rho^* = 1$. This partition generates two separate regions that emphasize different aspects of stability caused by dynamic planned lead times. $\Phi - 1 \leq \rho^* < 1$ indicates the situation that updating the lead time causes increased congestion in the production unit, and $1 < \rho^* \leq \Phi$ indicates the situation that updating the lead time causes decreased congestion in the production unit.

Corollary 3.1 implies that the stability condition becomes tighter as ρ^* decreases and looser as ρ^* increases. The range with which the utilization level in the static case is increased without causing instability at the production unit in the dynamic case is smaller as the stability condition is tighter, and bigger as the stability condition is looser. ρ^* depends on the design variables β, γ, and r. For $\beta = \gamma$ it is obvious that $\rho^* = 1$ independent of the update parameter r, which insights into the fact that the stability of the system is always retained as long as the response function is indifferent between an increase and a decrease in the planned lead time. An intuitive explanation is that the increase in the total workload due to placing additional orders in response to the increased lead time is balanced by the decrease in the total workload due to the order cancelations in response to the decreased lead time. The frequency of updating the lead time varies depending on the choice of the update parameter r ($\alpha = 1/r$) but, the rates of order addition and order cancelation stay balanced in the long-run. Therefore, the stability is retained. The first order derivative of $s(\rho,\, r,\, \beta,\, \gamma)$ in β implies that as β is increased the stability condition becomes tighter. This is because, as β is increased, on average more jobs are released to the production unit. On the other hand, as γ is increased the stability condition becomes looser because of the greater number of job cancelations.

3.3.3 Steady-State Distribution

For the derivations done throughout the rest of this chapter we assume the queueing system is stable. That is, $\rho^r + \beta\rho^{r-1} - \gamma\rho < 1$.

Let z_n, be the steady state probability that the process $\{N(t), t \geq 0\}$ is in state $\hat{w}(t) = n$.

$$z_n = z_0\rho^n,\ n = 0, 1, \ldots, \bar{w} - 1.$$

Let $p_l^{(r)}$ denote the vector of equilibrium probabilities for the level l of the QBD process $\{Y_r(t), t \geq 0\}$. So,

$$p_l^{(r)} = \Big(p^{(r)}(l,\, 0),\ p^{(r)}(l,\, 1),\ \ldots,\ p^{(r)}(l,\, r-2),\ p^{(r)}(l,\, r-1)\Big)\,,\ l = 0, 1, \ldots,$$

where $p^{(r)}(l,\, j)$ is the steady state probability that the QBD process is in state $(l,\, j)$. The equilibrium equations for this process are

$$\begin{aligned} p_0^{(r)}B_0^{(r)} + p_1^{(r)}A_2^{(r)} &= \mathbf{0}^{(r)}, && (3.4)\\ p_{l-1}^{(r)}A_0^{(r)} + p_l^{(r)}A_1^{(r)} + p_{l+1}^{(r)}A_2^{(r)} &= \mathbf{0}^{(r)},\ l = 1, 2, \ldots. && (3.5) \end{aligned}$$

The coupling relation between the processes $\{N(t), t \geq 0\}$ and $\{Y_r(t), t \geq 0\}$ is given by

$$p^{(r)}(0, 0) = z_0\rho^{\bar{w}}, \tag{3.6}$$

which yields the normalization equation,

$$\frac{p^{(r)}(0,0)(1-\rho^{\bar{w}})}{\rho^{\bar{w}}(1-\rho)} + \sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = 1. \tag{3.7}$$

Given that the Markov process with generator $Q^{(r)}$ is ergodic, the equilibrium probability vectors are determined by deploying the matrix-geometric form,

$$p_l^{(r)} = p_0^{(r)} \left(R^{(r)}\right)^l, \; l = 0, 1, \ldots,$$

where the r-dimensional rate matrix $R^{(r)}$ is the minimal-nonnegative solution of the matrix-quadratic equation,

$$A_0^{(r)} + R^{(r)} A_1^{(r)} + \left(R^{(r)}\right)^2 A_2^{(r)} = \mathbf{0}^{(r)\times(r)}. \tag{3.8}$$

Matrix geometric methods, initiated by Neuts (1981), serve as a powerful framework to analyze and (approximately) solve large classes of stochastic processes of $M|G|1$ type in a unified manner. In order to solve for the steady state properties of the process, one should determine the rate matrix $R^{(r)}$ that solves Equation (3.8). The problem of finding an explicit rate matrix is still a developing research area. Structural results have been provided in Ramaswami and Latouche (1986) for the QBD processes with transition matrices of rank 1. Van Leeuwaarden and Winands (2005) describe a class of QBD processes for which an explicit rate matrix can be found. Based on the results of Ramaswami and Latouche (1986), we provide an explicit solution for the rate matrix $R^{(r)}$ of the QBD process $\{Y_r(t), t \geq 0\}$ for every update parameter, r. It is given in the following theorem:

Theorem 3.2 *Given the Markov process with generator $Q^{(r)}$ is ergodic, the rate matrix $R^{(r)}$ that exactly solves the matrix quadratic equation (3.8) for every $r = 2, 3, \ldots$ is given by*

$$R^{(r)} = \begin{bmatrix} \boldsymbol{0}^{(r)} \\ \boldsymbol{0}^{(r)} \\ \vdots \\ \boldsymbol{0}^{(r)} \\ R_{r-1}^{(r)} \end{bmatrix}, \tag{3.9}$$

where

$$R_{r-1}^{(r)} = \left(\rho,\ \rho(\rho+\beta),\ \rho^2(\rho+\beta),\ \ldots,\ \rho^{r-2}(\rho+\beta),\ \frac{\rho^{r-1}(\rho+\beta)}{1+\gamma\rho}\right). \tag{3.10}$$

Proof See the Appendix at the end of this chapter.

In order to solve $\{Y_r(t), t \geq 0\}$ we need to derive $p_0^{(r)}$ from the equilibrium equation (3.4) and the normalization equation (3.7). Using the explicit expression for the rate matrix and the matrix-geometric form, the equilibrium equation (3.4) is rewritten as

$$\lambda p_0^{(r)} T_R^{(r)} + \mu p_0^{(r)} T_L^{(r)} - (\lambda + \mu) p_0^{(r)} + \mu p^{(r)}(0,0) \left(e_0^{(r)}\right)^{\mathrm{T}} +$$
$$\lambda p^{(r)}(0, r-1) \left(\gamma e_{r-2}^{(r)} + (1-\gamma) e_{r-1}^{(r)}\right)^{\mathrm{T}} = \mathbf{0}^{(r)}.$$

Solving this system of linear equations we get

$$\begin{aligned} p^{(r)}(0,j) &= p^{(r)}(0,0)\rho^j, \ j = 0, \ldots, r-2, && (3.11)\\ p^{(r)}(0,r-1) &= p^{(r)}(0,0) \cdot \frac{\rho^{r-1}}{1+\gamma\rho}. && (3.12)\end{aligned}$$

Let us rewrite the term $\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)}$ in the normalization equation (3.7) as

$$\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = p_0^{(r)} \left(I^{(r)} + R^{(r)} + \left(R^{(r)}\right)^2 + \cdots\right) e^{(r)}.$$

Since the rate matrix $R^{(r)}$ has rows of zero except the last one, the power matrices of $R^{(r)}$ can be expressed as:

$$\begin{aligned} \left(R^{(r)}\right)^l &= \left(R_{r-1,r-1}^{(r)}\right)^{l-1} R^{(r)},\\ R_{r-1,r-1}^{(r)} &= \frac{\rho^{r-1}(\rho+\beta)}{1+\gamma\rho}.\end{aligned}$$

The lower diagonal structure of $R^{(r)}$ allows us to directly see that its largest eigenvalue is $\eta = R_{r-1,r-1}^{(r)}$. The stability of the QBD process directly implies that $\eta < 1$. Then,

$$\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = p_0^{(r)} \left(I^{(r)} + \frac{R^{(r)}}{1-\eta}\right) e^{(r)}, \tag{3.13}$$

where the term $\left(I^{(r)} + \frac{R^{(r)}}{1-\eta}\right) e^{(r)}$ is an r-dimensional column vector of ones except the last row being equal to $1 + \frac{R_{r-1}^{(r)} e^{(r)}}{1-\eta}$. Employing the explicit expression of $R_{r-1}^{(r)}$ provided in Equation (3.10),

$$1 + \frac{R_{r-1}^{(r)} e^{(r)}}{1-\eta} = \frac{(1+\beta\rho) - \rho^{r-1}(\rho+\beta)}{(1-\rho)(1-\eta)}. \tag{3.14}$$

This result, together with our findings in Equations (3.11) and (3.12) to represent $p_0^{(r)}$ in Equation (3.13), provides us with

$$\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = \frac{p^{(r)}(0,0)}{1-\rho}\left(1 + \frac{\rho^r(\beta-\gamma)}{(1+\gamma\rho)(1-\eta)}\right).$$

Then, the normalization equation (3.7) is rewritten as

$$\frac{p^{(r)}(0,0)}{1-\rho}\left(\frac{1}{\rho^{\bar{w}}} + \frac{\rho^r(\beta-\gamma)}{(1+\gamma\rho)(1-\eta)}\right) = 1, \tag{3.15}$$

which provides interesting insights into the utilization of the production unit depending on how the lead time is updated. For this purpose, let us define

$$\kappa = \frac{\rho^r(\beta-\gamma)}{(1+\gamma\rho) - \rho^{r-1}(\rho+\beta)}$$

as the *utilization effect* due to updating the lead time. Then, by finding $p^{(r)}(0,0)$ from Equation (3.15) and using it in the coupling relation defined in Equation (3.6), the utilization of the production unit with the dynamic lead time is given by

$$\nu = \rho \cdot \frac{1+\rho^{\bar{w}-1}\kappa}{1+\rho^{\bar{w}}\kappa}. \tag{3.16}$$

From the stability condition, the denominator of κ is always positive. Thus, the sign of $\beta - \gamma$ determines the sign of κ. This means that the utilization effect, κ, indicates the way that the planning system reacts to changes in the planned lead time, which affects the utilization level in the dynamic case. From Equation (3.16), it is obvious that for a symmetric response function, $\beta = \gamma$ or $\kappa = 0$, the utilization level in the static case is retained, $\nu = \rho$. This result is consistent with the stability condition presented in the previous section. Furthermore, if there is a positive utilization effect, $\kappa > 0$ or $\beta > \gamma$, then the production unit is utilized more than it is in the static case, yielding $\nu > \rho$. This means that the output rate in the dynamic case is bigger than the output rate in the static case. From a practical perspective, the planning system is kept more responsive to an increase in the planned lead time, and the workload in the production unit is kept at higher levels in order to safeguard the pipeline against increased delivery duration. However, this generates excess production, which can be tackled in different ways such as secondary markets, salvage, and etc. If there is a negative utilization effect, $\kappa < 0$ or $\beta < \gamma$, then the production unit is less busy than it is in the static case, yielding $\nu < \rho$. This means that the output rate in the dynamic case is less than the output rate in the static case. From a practical perspective, the planning system is kept more responsive to a decrease in the planned lead time, and the strategy is to avoid unnecessary workload with respect to the anticipated delivery duration. This strategy causes production lacking behind the demand, which can be tackled by supply channels external to the production unit. Considering external supply or secondary markets adds further dimensions to our discussion about the lead time syndrome, and is an interesting future research direction.

The design variables β, γ, and r are related to the characteristics of the dynamic case in a way that they model the frequency with which the lead time is updated and the sensitivity of the planning system to this change. The utilization level in the dynamic case behaves differently depending on the changes made in β, γ or r. The following proposition summarizes the relationships between the utilization level and these variables.

Proposition 3.1 *Keeping all other variables fixed, the utilization of the production unit in the dynamic case changes depending on the variables β, γ, and r as follows:*

- *ν increases as β is increased.*
- *ν decreases as γ is increased.*
- *For $\beta > \gamma$, ν increases for increasing frequencies of updating.*
- *For $\gamma > \beta$, ν decreases for increasing frequencies of updating.*

Proof See the Appendix at the end of this chapter.

Proposition 3.1 provides some of the intuitive insights on the effects of updating the planned lead time. As the planning system becomes more inclined to add a job in response to an increase in the lead time, through higher β values, then the utilization level increases. On the other hand, as the planning system becomes more inclined to cancel a job in response to a decrease in the lead time, through higher γ values, then the utilization level decreases. The effect of update frequency on the utilization level depends on which type of inclination is stronger over the other. Our results here are complementary to our previous discussion about the utilization effect κ. When the system operates with a positive utilization effect, $\beta > \gamma$, then the congestion effect of updating the lead time increases with the update frequency. When the system operates with a negative utilization effect, $\gamma > \beta$, updating the lead time lessens the congestion in the production unit with a higher degree as the update frequency increases.

In the following section, we provide further analytical results about the relationship between the update frequency and the performance of the production unit. Throughout the rest of this chapter we assume $\beta = \gamma$. Thus, the static case utilization is retained in the dynamic case. We also replace γ with β for brevity in the formulations.

3.4 Performance Evaluation

The key performance indicators of the production unit are related to the cost performance, the delivery performance, and the nervousness created in the planning system. Analytical results are provided mainly concentrated on the frequency of updating the lead time and the level of being reactive to the changes in the lead time. To avoid intricacy in the notations and the analysis, we should note that the parameter r refers to the reciprocal of the update frequency α.

It has already been mentioned that having $\beta = \gamma$, the dynamic utilization level is kept equal to the static utilization level. Therefore, keeping large number of jobs as workload introduces inefficiency, and increased costs are incurred due to material handling, inventory holding and etc. Maintaining a utilization level of ρ implies that the average number of jobs in process are the same both in the static and the dynamic situations. Thus, our attention is on the average number of jobs in the backlog. Let $B(r)$ denote the random variable for the backlog level with the probability mass function

$$\Pr\{B(r) = n\} = \begin{cases} \sum_{j=0}^{\bar{w}-1} z_j + p^{(r)}(0,0), & n = 0 \\ p^{(r)}(0,j), & n = j,\, j = 1,2,\ldots,r-1 \\ p^{(r)}(l,j), & n = rl + j,\, l = 1,2,\ldots, \\ & \qquad\qquad j = 0,1,\ldots,r-1 \end{cases}$$

The larger the backlog level is, the higher the costs of the production unit are. Using the explicit derivations that are provided in the previous section we formulate the relationship between the static and the dynamic case average backlog levels as in the following proposition:

Proposition 3.2 *The average number of jobs in the production backlog in the dynamic case is always larger than it is in the static case, and it is given by*

$$E\left(B(r)\right) = E\left(B(\infty)\right)(1+\theta), \tag{3.17}$$

where the average backlog level in the static case is $E\left(B(\infty)\right) = \frac{\rho^{\bar{w}+1}}{1-\rho}$, *and*

$$\theta = \frac{2\beta\rho^{r-1}(1-\rho)}{1+\beta\rho-\rho^{r-1}(\rho+\beta)}$$

is a monotonically increasing function of the update frequency $\alpha = 1/r$ and the response probability β.

Proof See the Appendix at the end of this chapter.

Proposition 3.2 has an intuitively appealing interpretation. We identify that updating the lead time increases the long-term average backlog level by a multiplicative term, which is a function of the utilization level, the update frequency and the response probability. When the lead time is increased, additional jobs are released increasing the backlog status until the lead time is decreased. In the long run, the number of jobs added due to the lead time increase is equal to the number of jobs canceled due to the lead time decrease. On average, the number of jobs kept in the backlog queue increases although the utilization of the production unit does not change. It is a fundamental intuition from Hopp and Spearman (2000) that inefficient increase in congestion is undesirable due to the increase in order processing, material handling, and inventory holding costs. Considering these costs that are related to keeping workload in the production unit and ignoring all other cost terms such as those related to due-date performance of orders, we can state that the static lead time policy or the non-reactive case yields the lowest cost situation. As the system becomes more reactive or sensitive to the changes in the workload level through higher response probability or update frequency, more workload is generated.

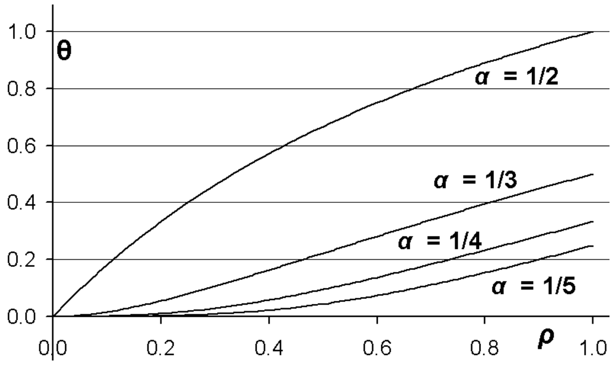


Figure 3.6: θ vs. ρ for $r = 2, 3, 4, 5$ and $\beta = 1$.

Figure 3.6 and 3.7 illustrate the level of increase in the average number of jobs in the backlog for different update frequencies as a function of utilization and response probability respectively. For the highest update frequency, $\alpha = 1/2$, and the highest response probability, $\beta = 1$, when the utilization level is close

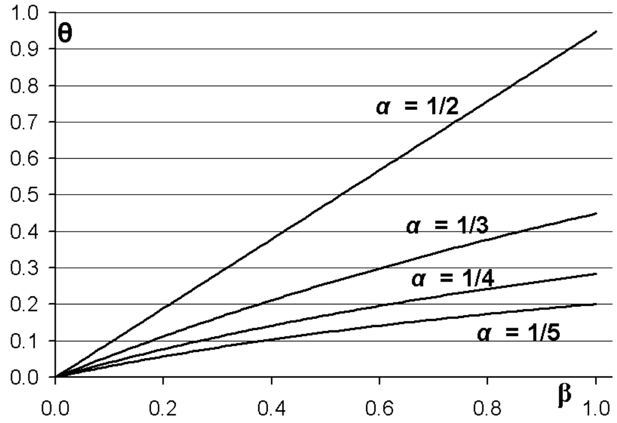


Figure 3.7: θ vs. β for $r = 2, 3, 4, 5$ and $\rho = 0.90$.

to 1, the backlog level becomes twice as much as it is when the lead time is fixed. As the update frequency decreases, the level of increase in the average production backlog significantly decreases. A similar relationship also holds between the backlog level and the response probability. As long as there is a response to the change in the lead time with $\beta > 0$, then there is always an increase in the average backlog level. In addition, the value of θ is more sensitive to the value of β when the lead time is updated more frequently.

In addition, the delivery performances of the jobs are evaluated by considering the actual durations of time that the processed jobs spend between the moment that they are released and the moment that they are completed. The random variable for the job flow time is denoted by $C(r)$, and it is the flow time of a finished job (not canceled). As the average workload level increases in a dynamic case, we would also expect to see on average longer flow times. For the full-reactive case, $\beta = 1$, the result is even stronger as provided in the following proposition:

Proposition 3.3 *Given that only the last job in the backlog queue can be canceled, the flow time of a processed job in the dynamic case is stochastically greater than or equal to the flow time of a job in the static case.*

$$Pr\{C(r) \geq l\} \geq Pr\{C(\infty) \geq l\},$$

where $C(\infty)$ denotes the random variable for the flow time of a job in the static case.

Proof See the Appendix at the end of this chapter.

This result is related to the increased variability in the release pattern of jobs created by updating the lead time. When the lead time is updated, jobs are released earlier than they would be in the static case. Some of the jobs are canceled. However, the net input rate does not differ from the static case, and on average, jobs spend more time in the system causing worse delivery performance for the downstream customers/processes. Further, Proposition 3.3 implies that a certain (i.e. 90^{th}) percentile of the flow time distribution of the completed jobs in the dynamic case is greater than or equal to it is in the static case. Thus, the service level in terms of the percentage of jobs completed in a certain time frame will never improve once the planned lead time is updated.

In Chapter 2, it has already been presented that updating the lead times generates nervousness in planning decisions upstream through the supply chain (see Equations (2.10) and (2.12)). Therefore, it is worthwhile to see how the distribution characteristics of the dynamic planned lead time changes. For this purpose, $L(r)$ is defined as the random variable for the dynamic planned lead time with the probability mass function

$$\Pr\{L(r) = L_{\min} + n\} = \begin{cases} \sum_{j=0}^{\bar{w}-1} z_j + \sum_{j=0}^{r-1} p^{(r)}(0,j), & n = 0 \\ \sum_{j=0}^{r-1} p^{(r)}(l,j), & n = l,\ l = 1, 2, \ldots, \end{cases}$$

A long and highly variable planned lead time implies increased nervousness in planning and large pipeline inventories. The expected planned lead time and its coefficient of variation are derived based on the updating characteristics as follows:

Proposition 3.4 *The average and the coefficient of variation of the dynamic planned lead time are respectively given by*

$$E\left(L(r)\right) = L_{min} + \xi, \tag{3.18}$$

$$CV\left(L(r)\right) = \frac{\sqrt{\xi\left(\frac{1+\beta\rho+\rho^{r-1}(\rho+\beta)}{1+\beta\rho-\rho^{r-1}(\rho+\beta)} - \xi\right)}}{L_{min} + \xi}, \tag{3.19}$$

where

$$\xi = \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho-\rho^{r-1}(\rho+\beta)}$$

is a monotonically increasing function of the update frequency $\alpha = 1/r$ and the response probability β.

Proof See the Appendix at the end of this chapter.

As a result, the average planned lead time monotonically increases with the update frequency and the response probability. It should be noted that for very large $L_{\min}$, updating the lead time does not have significant impacts on $E\left(L(r)\right)$ or $CV\left(L(r)\right)$.

The variation in the planned lead time depends on the variables such as $L_{\min}$ and $\bar{w}$. For a specific situation with $L_{\min} = 1$ and $\bar{w} = 1$, $CV\left(L(r)\right)$ is depicted with respect to ρ and β respectively in Figures 3.8 and 3.9 for different update frequencies.

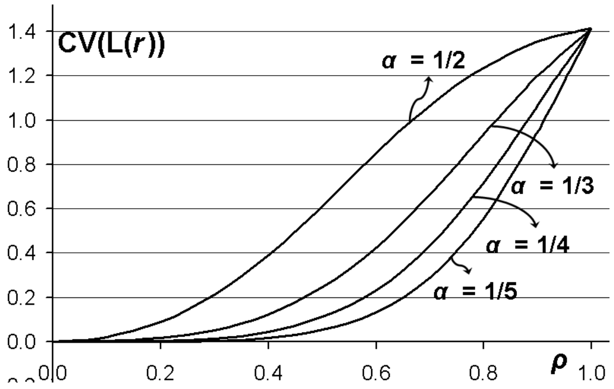


Figure 3.8: $CV\left(L(r)\right)$ vs. ρ for $r = 2, 3, 4, 5$ and $\beta = 1$.

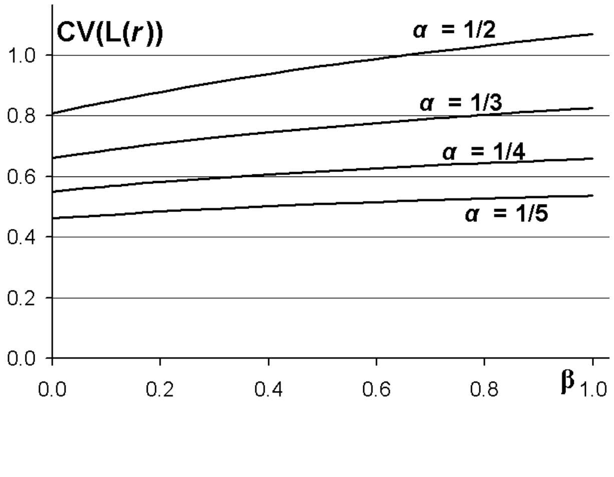


Figure 3.9: $CV\left(L(r)\right)$ vs. β for $r = 2, 3, 4, 5$ and $\rho = 0.75$.

Figure 3.8 shows that, for $\beta = 1$, $L_{\min} = 1$, $\bar{w} = 1$, and for a given update frequency, $CV\,(L(r))$ increases as ρ increases, and asymptotically reaches the value of $\sqrt{2}$ as ρ approaches to 1. For a given ρ, $CV\,(L(r))$ increases as the lead time is updated more frequently. However the relative effect of the update frequency depends on the value of ρ. The update frequency is much influential for moderate values of ρ, i.e. $0.5 \leq \rho \leq 0.8$. One may be interested in identifying the range of utilization levels for which there is a high level of variability, i.e. $CV\,(L(r)) \geq 1$, for a given update frequency and response probability. In that respect, for the full reactive case, $\beta = 1$, the high variability range is $0.66 < \rho < 1.00$ for $\alpha = 1/2$, $0.82 < \rho < 1.00$ for $\alpha = 1/3$, $0.88 < \rho < 1.00$ for $\alpha = 1/4$, and $0.91 < \rho < 1.00$ for $\alpha = 1/5$. Thus, in case one prefers to operate with dynamic planned lead times but at the same time keep $CV\,(L(r))$ less than one, when for example $\rho = 0.85$, then the update frequency must be kept less than 1/3.

Figure 3.9 implies that, for a given update frequency, $CV\,(L(r))$ increases as the response probability increases. In addition, $CV\,(L(r))$ is more sensitive to the response probability as the update frequency increases. For example, when $\rho = 0.75$, $CV\,(L(r))$ ranges between 0.81 and 1.07 for $\alpha = 1/2$, and between 0.46 and 0.54 for $\alpha = 1/5$.

3.5 Conclusion

In this chapter, we have modeled a situation known as the lead time syndrome using constructs from queueing theory. This chapter has been motivated by the lack of analytical analysis about a production system facing erratic order releases in response to changing planned lead times. Explicit results on the stability condition and the performance evaluation of such a system have been provided, and the following insights have been derived:

- If the reactions to the increase and to the decrease in the planned lead time are identical, then the static utilization level is retained in the dynamic case, and the stability condition is $\rho < 1$ irrespective of the update frequency.

- The term utilization effect, κ, has been defined to represent the effects of response probabilities β and γ on the utilization level. The utilization increases as β becomes higher, and decreases as γ becomes higher.

Given that the production unit have the same utilization level both in the static and the dynamic cases,

- Updating the lead time increases the average number of jobs waiting in the production backlog. The level of increase is bigger for higher utilization level or higher update frequency.
- Updating the lead time causes the processed jobs to spend longer durations of time in the system.
- The variation in the dynamic planned lead time increases with the update frequency except at the utilization boundaries, the lead time variation is insensitive to the update frequency.
- Higher response probability yields higher variation in the planned lead time.

The strength of our analysis lies in the modeling of the update frequency and the response function, which are relevant design parameters for dynamic, adaptive planning systems. Some of the results of this chapter support some of the conclusions of Chapter 2 such as erratic order release pattern, and its increasing effects with increasing frequency of updating. Differently, due-date performances of the dynamic planned lead time are not considered in this chapter.

The analysis and the results provided in this chapter promote further improvements in modeling and understanding dynamic, adaptive systems. The performance of different update policies can be evaluated under different assumptions on ordering behavior. In addition, the analysis for non-stationary demand conditions is an interesting research direction. It is also promising to extend the analysis considering multiple-stages of the supply chain including the downstream/upstream stock points and production units.

Appendix to Chapter 3

Proof of Theorem 3.1

For $r = 2$, the flow from state 0 to state 1 is equal to the flow from state 1 to state 0.

$$\pi_0^{(2)} (1 + \rho - \gamma) = \pi_1^{(2)} (1 + \rho - \beta\rho) . \tag{3.20}$$

For $r = 3$, the balance equations between states 0 and 1, and between states 1 and 2 yield

$$\begin{aligned} (1+\rho)\pi_0^{(3)} &= (1-\beta)\rho\pi_2^{(3)} + \pi_1^{(3)}, \\ (1+\rho)\pi_2^{(3)} &= \rho\pi_1^{(3)} + (1-\gamma)\pi_0^{(3)}. \end{aligned}$$

Then,

$$\pi_0^{(3)} \left(1 + \rho + \rho^2 - \gamma\right) = \pi_2^{(3)} \left(1 + \rho + \rho^2 - \beta\rho^2\right) . \tag{3.21}$$

For $r = 4, \dots$ we obtain, by balancing the flow out and into the set $\{0, \dots, k\}$, $k = 0, \dots, r-2$,

$$\begin{aligned} (1+\rho)\pi_0^{(r)} &= (1-\beta)\rho\pi_{r-1}^{(r)} + \pi_1^{(r)}, \\ (1+\rho)\pi_{r-1}^{(r)} &= \rho\pi_{r-2}^{(r)} + (1-\gamma)\pi_0^{(r)}, \\ \pi_k^{(r)} &= \pi_1^{(r)}\rho^{k-1} + \pi_0^{(r)}\sum_{j=0}^{k-2}\rho^j - \pi_{r-1}^{(r)}\sum_{j=1}^{k-1}\rho^j, \; k = 2, \dots, r-2. \end{aligned}$$

Hence,

$$\pi_0^{(r)} \left(\sum_{j=0}^{r-1}\rho^j - \gamma\right) = \pi_{r-1}^{(r)} \left(\sum_{j=0}^{r-1}\rho^j - \beta\rho^{r-1}\right) . \tag{3.22}$$

Equations (3.20), (3.21) and (3.22) together with Condition (3.2) show that the system is stable for every $r = 2, 3, \dots$ as long as the following condition holds:

$$\rho^r + \beta\rho^{r-1} - \gamma\rho < 1.$$

This completes the proof of Theorem 3.1.

Proof of Corollary 3.1

The monotonicity properties of the stability function with respect to ρ should be investigated. The first and the second order partial derivatives with respect to ρ are

$$\begin{aligned}\frac{\partial}{\partial(\rho)}s(\rho,\, r,\, \beta,\, \gamma) &= r\rho^{r-1} + (r-1)\beta\rho^{r-2} - \gamma,\\ \frac{\partial^2}{\partial(\rho)^2}s(\rho,\, r,\, \beta,\, \gamma) &= r(r-1)\rho^{r-2} + (r-1)(r-2)\beta\rho^{r-3} > 0,\end{aligned}$$

which imply that the stability function is strictly convex in ρ, and the first order partial derivative of the stability function with respect to ρ is monotonically increasing. In detail, we are interested in the domain $\rho > 0$. From the definitions of r, β and γ, it follows that

$$\begin{aligned} s(0,\, r,\, \beta,\, \gamma) &< 0,\\ \frac{\partial}{\partial(\rho)}s(\rho,\, r,\, \beta,\, \gamma)\Big|_{\rho=0} &\leq 0,\\ \frac{\partial}{\partial(\rho)}s(\rho,\, r,\, \beta,\, \gamma)\Big|_{\rho=1} &> 0.\end{aligned}$$

Together with the convexity of $s(\rho,\, r,\, \beta,\, \gamma)$ with respect to ρ, this set of inequalities implies that the stability function $s(\rho,\, r,\, \beta,\, \gamma)$ has a single root ρ^* that is greater than zero, and for $0 < \rho < \rho^*$, $s(\rho) < 0$.

The partial derivatives of $s(\rho,\, r,\, \beta,\, \gamma)$ with respect to β, γ, and r are respectively given by

$$\begin{aligned}\frac{\partial}{\partial(\beta)}s(\rho,\, r,\, \beta,\, \gamma) &= \rho^{r-1},\\ \frac{\partial}{\partial(\gamma)}s(\rho,\, r,\, \beta,\, \gamma) &= -\rho,\\ \frac{\partial}{\partial(r)}s(\rho,\, r,\, \beta,\, \gamma) &= \rho^{r-1}(\rho+\beta)\ln\rho.\end{aligned}$$

The stability function is monotonically increasing with β and decreasing with γ, which implies that, due to the convex stability function, ρ^* is monotonically decreasing with β and increasing with γ. $s(\rho,\, r,\, \beta,\, \gamma)$ is also decreasing with r for $\rho < 1$, and increasing with r for $\rho > 1$. As a result, the minimum level of ρ^* is achieved when $\beta = 1$, $\gamma = 0$ and $r = 2$ by solving $\rho^2 + \rho - 1 = 0$. Then,

$$\rho^* \geq \frac{\sqrt{5}-1}{2}. \tag{3.23}$$

Similarly, the maximum level of ρ^* is found when $\beta = 0$, $\gamma = 1$ and $r = 2$ by solving $\rho^2 - \rho - 1 = 0$. Then,

$$\rho^* \leq \frac{1+\sqrt{5}}{2}. \tag{3.24}$$

Knowing that $\Phi = \frac{1+\sqrt{5}}{2}$, Conditions (3.23) and (3.24) prove Corollary 3.1.

Proof of Theorem 3.2

We apply the results of Ramaswami and Latouche (1986), and derive explicit solutions using matrix algebra. Given the Markov process with generator $Q^{(r)}$ is ergodic, the rate matrix $R^{(r)}$ that exactly solves the matrix quadratic equation (3.8) is given by

$$R^{(r)} = -A_0^{(r)} \left(A_1^{(r)} + A_0^{(r)} e^{(r)} \left(\gamma e_{r-2}^{(r)} + (1-\gamma) e_{r-1}^{(r)} \right)^{\mathrm{T}} \right)^{-1}. \tag{3.25}$$

The structure of $A_0^{(r)}$ allows us to describe the rate matrix in more detail. Firstly, let us define the r-dimensional square matrix

$$A_3^{(r)} = A_1^{(r)} + A_0^{(r)} e^{(r)} \left(\gamma e_{r-2}^{(r)} + (1-\gamma) e_{r-1}^{(r)} \right)^{\mathrm{T}}.$$

Then, $R^{(r)}$ has rows of zero except the last one, and its last row can be derived from a linear combination of the first and the second rows of $\left(A_3^{(r)}\right)^{-1}$. Let us define $\left(a_{3(0)}^{(r)}\right)^{-1}$ and $\left(a_{3(1)}^{(r)}\right)^{-1}$ as the first and the second rows of the matrix $\left(A_3^{(r)}\right)^{-1}$ respectively. Similarly, $R_{r-1}^{(r)}$ is the last row of the rate matrix $R^{(r)}$. Then,

$$R_{r-1}^{(r)} = -\lambda \left((1-\beta) \left(a_{3(0)}^{(r)}\right)^{-1} + \beta \left(a_{3(1)}^{(r)}\right)^{-1} \right). \tag{3.26}$$

The explicit structure of $A_3^{(r)}$ can be written as follows:

$$\begin{aligned}
\left(A_3^{(r)}\right)_{i,i} &= -(\lambda+\mu),\ i = 0,\ldots,r-2, \\
\left(A_3^{(r)}\right)_{i,i+1} &= \lambda,\ i = 0,\ldots,r-2, \\
\left(A_3^{(r)}\right)_{i,i-1} &= \mu,\ i = 1,\ldots,r-2, \\
\left(A_3^{(r)}\right)_{r-1,r-2} &= (\gamma\lambda+\mu), \\
\left(A_3^{(r)}\right)_{r-1,r-1} &= -\left(\gamma\lambda+\mu\right),
\end{aligned}$$

and all other elements of $A_3^{(r)}$ are zero. $\left(a_{3(0)}^{(r)}\right)^{-1}$ and $\left(a_{3(1)}^{(r)}\right)^{-1}$ solve the following matrix equations:

$$\begin{aligned}\left(A_3^{(r)}\right)^{\mathrm{T}}\left(\left(a_{3(0)}^{(r)}\right)^{-1}\right)^{\mathrm{T}} &= e_0^{(r)},\\ \left(A_3^{(r)}\right)^{\mathrm{T}}\left(\left(a_{3(1)}^{(r)}\right)^{-1}\right)^{\mathrm{T}} &= e_1^{(r)}.\end{aligned}$$

By exploiting the tri-diagonal structure of $A_3^{(r)}$ we find

$$\begin{aligned}\left(a_{3(0)}^{(r)}\right)^{-1} &= -\left(\frac{1}{\mu}, \frac{\rho}{\mu}, \frac{\rho^2}{\mu}, \dots, \frac{\rho^{r-2}}{\mu}, \frac{\rho^{r-1}}{\gamma\lambda+\mu}\right),\\ \left(a_{3(1)}^{(r)}\right)^{-1} &= -\left(\frac{1}{\mu}, \frac{(1+\rho)}{\mu}, \frac{\rho(1+\rho)}{\mu}, \dots, \frac{\rho^{r-3}(1+\rho)}{\mu}, \frac{\rho^{r-2}(1+\rho)}{\gamma\lambda+\mu}\right).\end{aligned}$$

Consequently, together with Equation (3.26), this derivation yields

$$R_{r-1}^{(r)} = \left(\rho,\ \rho(\rho+\beta),\ \rho^2(\rho+\beta),\ \dots,\ \rho^{r-2}(\rho+\beta),\ \frac{\rho^{r-1}(\rho+\beta)}{1+\rho\gamma}\right).$$

This completes the proof of Theorem 3.2.

Proof of Proposition 3.1

From calculus it is straight forward that

$$\frac{\partial}{\partial(\beta)}\nu = \frac{\partial}{\partial(\kappa)}\nu \cdot \frac{\partial}{\partial(\beta)}\kappa, \tag{3.27}$$

and the same holds for partial derivatives of ν with respect to γ or r. Therefore, for any given ρ and $\bar{w}$, the utilization in the dynamic case can be characterized by looking at the behavior of the utilization effect κ and the first order partial derivative of ν with respect to κ, which is given by

$$\frac{\partial}{\partial(\kappa)}\nu = \frac{\rho^{\bar{w}}(1-\rho)}{(1+\rho^{\bar{w}}\kappa)^2}.$$

Then, the following properties of ν with respect to κ can be derived:

$$\frac{\partial}{\partial(\kappa)}\nu > 0 \quad \text{for}\,\rho < 1, \tag{3.28}$$

$$\frac{\partial}{\partial(\kappa)}\nu < 0 \quad \text{for}\,\rho > 1. \tag{3.29}$$

The behavior of κ with respect to β, γ and r can be formulated as follows:

$$\begin{aligned}
\frac{\partial}{\partial(\beta)}\kappa &= \frac{\rho^r(1-\rho^r)+\gamma\rho^{r+1}(1-\rho^{r-2})}{((1+\gamma\rho)-\rho^{r-1}(\rho+\beta))^2},\\
\frac{\partial}{\partial(\gamma)}\kappa &= \frac{-\rho^r(1-\rho^r)-\beta\rho^{r+1}(1-\rho^{r-2})}{((1+\gamma\rho)-\rho^{r-1}(\rho+\beta))^2},\\
\frac{\partial}{\partial(r)}\kappa &= \frac{(\beta-\gamma)\rho^r\ln(\rho)(1+\gamma\rho)}{((1+\gamma\rho)-\rho^{r-1}(\rho+\beta))^2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{\partial}{\partial(\beta)}\kappa > 0 \quad \text{for}\,\rho<1 \quad &, \quad \frac{\partial}{\partial(\beta)}\kappa < 0 \quad \text{for}\,\rho>1,\\
\frac{\partial}{\partial(\gamma)}\kappa < 0 \quad \text{for}\,\rho<1 \quad &, \quad \frac{\partial}{\partial(\gamma)}\kappa > 0 \quad \text{for}\,\rho>1,\\
\frac{\partial}{\partial(r)}\kappa > 0 \quad \text{for}\,\gamma>\beta\,\text{and}\,\rho<1 \quad &, \quad \frac{\partial}{\partial(r)}\kappa < 0 \quad \text{for}\,\gamma>\beta\,\text{and}\,\rho>1,\\
\frac{\partial}{\partial(r)}\kappa < 0 \quad \text{for}\,\beta>\gamma\,\text{and}\,\rho<1 \quad &, \quad \frac{\partial}{\partial(r)}\kappa > 0 \quad \text{for}\,\beta>\gamma\,\text{and}\,\rho>1.
\end{aligned}$$

Applying these results together with Properties (3.28) and (3.29) in Equation (3.27) yields

$$\frac{\partial}{\partial(\beta)}\nu > 0 \quad \text{for}\,\rho<1\,\text{or}\,\rho>1, \tag{3.30}$$

$$\frac{\partial}{\partial(\gamma)}\nu < 0 \quad \text{for}\,\rho<1\,\text{or}\,\rho>1, \tag{3.31}$$

$$\frac{\partial}{\partial(r)}\nu > 0 \quad \text{for}\,\gamma>\beta\,\text{and for}\,(\rho<1\,\text{or}\,\rho>1), \tag{3.32}$$

$$\frac{\partial}{\partial(r)}\nu < 0 \quad \text{for}\,\beta>\gamma\,\text{and for}\,(\rho<1\,\text{or}\,\rho>1). \tag{3.33}$$

For $\rho = 1$, we apply L'Hôpital's rule to Equation (3.27), and derive the following results:

$$\begin{aligned}
\lim_{\rho\to 1}\frac{\partial}{\partial(\beta)}\nu &= \frac{r+\gamma(r-2)}{\left(\left(\frac{\partial}{\partial(\rho)}\kappa-\bar{w}\right)(\gamma-\beta)\right)^2},\\
\lim_{\rho\to 1}\frac{\partial}{\partial(\gamma)}\nu &= \frac{-r-\beta(r-2)}{\left(\left(\frac{\partial}{\partial(\rho)}\kappa-\bar{w}\right)(\gamma-\beta)\right)^2},\\
\lim_{\rho\to 1}\frac{\partial}{\partial(r)}\nu &= \frac{(\gamma-\beta)(1+\gamma)}{\left(\left(\frac{\partial}{\partial(\rho)}\kappa-\bar{w}\right)(\gamma-\beta)\right)^2},
\end{aligned}$$

which yield

$$\begin{aligned}
\lim_{\rho\to 1} \frac{\partial}{\partial(\beta)}\nu &> 0, && (3.34)\\
\lim_{\rho\to 1} \frac{\partial}{\partial(\gamma)}\nu &< 0, && (3.35)\\
\lim_{\rho\to 1} \frac{\partial}{\partial(r)}\nu &> 0 \quad \text{for}\, \gamma > \beta, && (3.36)\\
\lim_{\rho\to 1} \frac{\partial}{\partial(r)}\nu &< 0 \quad \text{for}\, \beta > \gamma. && (3.37)
\end{aligned}$$

Properties (3.30) through (3.33) together with Properties (3.34) through (3.37) imply the following insights: (1) Utilization increases as β increases, (2) utilization decreases as γ increases, (3) For $\gamma > \beta$, utilization decreases as the lead time is updated more frequently, (4) For $\beta > \gamma$, utilization increases with the update frequency.

This completes the proof of Proposition 3.1.

Proof of Proposition 3.2

The expected number of jobs in the backlog is given by

$$\begin{aligned}
E\left(B(r)\right) &= \sum_{l=0}^{\infty}\sum_{j=0}^{r-1}(rl+j)p^{(r)}(l,\, j)\\
&= r\sum_{l=1}^{\infty} lp_l^{(r)}e^{(r)} + \sum_{j=1}^{r-1} j \sum_{l=0}^{\infty} p_l^{(r)}e_j^{(r)},
\end{aligned}$$

and using the derivation in Equation (3.13) together with the explicit form of $p_0^{(r)}$ given in Equations (3.11) and (3.12), $E\left(B(r)\right)$ can be further simplified as

$$E\left(B(r)\right) = \frac{rp_0^{(r)}R^{(r)}e^{(r)}}{(1-\eta)^2} + \sum_{j=1}^{r-1} jp_0^{(r)}\left(I^{(r)} + \frac{R^{(r)}}{1-\eta}\right)e_j^{(r)}. \qquad (3.38)$$

Let us denote the first and the second terms on the right-hand-side of Equation (3.38) by $B_r^{(1)}$ and $B_r^{(2)}$ respectively. We first write $B_r^{(1)}$ in its explicit form, and then, write $B_r^{(2)}$ similarly in order to provide an explicit expression for $E\left(B(r)\right)$.

Solving the normalization equation (3.15) for $p^{(r)}(0,\,0)$ and using it to express $p_0^{(r)}$ from Equations (3.11) and (3.12), we get

$$p_0^{(r)} = (1-\rho)\rho^{\bar{w}}\left(1,\, \rho,\, \rho^2, \ldots, \rho^{r-2},\, \frac{\rho^{r-1}}{1+\beta\rho}\right). \tag{3.39}$$

Besides, the expression for the rate matrix in Equations (3.9) and (3.10) implies that the term $\frac{R^{(r)}e^{(r)}}{1-\eta}$ is an r-dimensional column vector of zeros except the last row being equal to $\frac{(1+\beta)\rho}{1-\rho}$. This, together with the explicit form of $p_0^{(r)}$ in Equation (3.39), provides $B_r^{(1)}$ as

$$B_r^{(1)} = \frac{(1+\beta)r\rho^{\bar{w}+1}\rho^{r-1}}{1+\beta\rho-\rho^{r-1}(\rho+\beta)}. \tag{3.40}$$

Elaborating the explicit form of $p_0^{(r)}$ and the r-dimensional square matrix $I^{(r)} + \frac{R^{(r)}}{1-\eta}$, $B_r^{(2)}$ is rewritten as

$$B_r^{(2)} = (1-\rho)\rho^{\bar{w}}\left(\frac{1}{1-\eta}\sum_{j=1}^{r-2} j\rho^j + \frac{(r-1)\rho^{r-1}}{(1+\beta\rho)(1-\eta)}\right).$$

Further simplification yields

$$B_r(2) = \frac{\rho^{\bar{w}+1}(1+\beta\rho)(1-\rho^{r-1})}{(1-\rho)(1+\beta\rho-\rho^{r-1}(\rho+\beta))} - \frac{(1+\beta)(r-1)\rho^{\bar{w}+1}\rho^{r-1}}{1+\beta\rho-\rho^{r-1}(\rho+\beta)}. \tag{3.41}$$

Then, the average number of jobs in the backlog, given in Equation (3.38), can be written explicitly by

$$E\left(B(r)\right) = \frac{\rho^{\bar{w}+1}}{1-\rho}\left(1 + \frac{2\beta\rho^{r-1}(1-\rho)}{1+\beta\rho-\rho^{r-1}(\rho+\beta)}\right). \tag{3.42}$$

Let us define

$$\theta = \frac{2\beta\rho^{r-1}(1-\rho)}{1+\beta\rho-\rho^{r-1}(\rho+\beta)}.$$

In order to characterize how the average backlog level changes with the update frequency $\alpha = 1/r$ and the response probability β, it is sufficient that we take the partial derivatives of θ with respect to β and r.

$$\frac{\partial}{\partial(\beta)}\theta = \frac{2(1-\rho)\rho^{r-1}(1-\rho^r)}{\left(1+\beta\rho-\rho^{r-1}(\rho+\beta)\right)^2}, \tag{3.43}$$

$$\frac{\partial}{\partial(r)}\theta = \frac{2\beta(1-\rho)\rho^{r-1}\ln(\rho)(1+\beta\rho)}{\left(1+\beta\rho-\rho^{r-1}(\rho+\beta)\right)^2}. \tag{3.44}$$

From Equation (3.43) it is straightforward that θ is monotonically increasing with β for all $0 \leq \beta \leq 1$. Although the function θ is defined on the discrete domain of $r = 2, 3, \ldots$, the monotonicity property, on the continuous domain of r, implied by Equation (3.44), $\frac{\partial}{\partial(r)}\theta < 0$ for all r, is sufficient to state that θ is monotonically increasing as r decreases. This result directly implies that θ is monotonically increasing with the update frequency $\alpha = 1/r$. Together with Equation (3.42) this completes the proof of Proposition 3.2.

Proof of Proposition 3.3

We prove by keeping track of a sample path of arrivals and departures for both the static and the dynamic cases, and then, by applying induction. In the static case, assume that jobs arrive and are processed according to the following sequence:

$$\left(t_1^{(s)}, c_1^{(s)}\right), \left(t_2^{(s)}, c_2^{(s)}\right), \ldots, \left(t_k^{(s)}, c_k^{(s)}\right) \ldots,$$

which means k^{th} job arrives at time $t_k^{(s)}$ and is completed at time $c_k^{(s)}$ yielding a flow time $f_k^{(s)} = c_k^{(s)} - t_k^{(s)}$. In the dynamic case, $t_k^{(d)}$ refers to the arrival time of the k^{th} processed job, and similarly, $c_k^{(d)}$ and $f_k^{(d)}$ refer to the completion time and the flow time of the k^{th} processed job respectively. It is crucial to note that, in the dynamic case, some of the jobs are canceled and not processed, and the cancelation policy is always cancel the last job in the backlog queue.

From the transition rate diagram in Figure 3.2 it is obvious that along a sample path of $w(t) \leq \bar{w} + r - 1$ for all t, where the lead time is never updated, the dynamic and the static cases are identical.

$$C(r) = C(\infty) \quad \text{if} \quad w(t) \leq \bar{w} + r - 1 \quad \text{for all}\, t. \tag{3.45}$$

Let t_0 be any point in time, where there are p_0 jobs that are already processed in both cases. Assume there are k jobs in the static system, and the lead time of the dynamic system is $L_{\min} + l$. Then, there are $k + l$ jobs in the dynamic system, and at most l of these jobs can be canceled. Due to the sample path of arrivals, the arrival time of the last job in the static case is equal to the arrival time of the last job in the dynamic case. Because of the fact that jobs are sequenced in the queue according to the FCFS discipline in both cases, the following relationship holds for the j^{th} processed job after t_0, $j \leq k$:

$$t_{j+p_0}^{(d)} \leq t_{j+p_0}^{(s)}, \tag{3.46}$$

$$c_{j+p_0}^{(d)} = c_{j+p_0}^{(s)} = t_0 + \tau(p_0, j), \tag{3.47}$$

where $\tau(p_0, j)$ is the total processing time of j consecutive jobs, given that there are currently p_0 jobs already processed so far. Due to the sample path of processing times, this term is equal in both the dynamic and the static cases. Equations (3.46) and (3.47) directly imply the following fact about flow times of the $(j+p_0)^{th}$ processed jobs in the dynamic and the static cases:

$$f^{(d)}_{j+p_0} = \left(c^{(d)}_{j+p_0} - t^{(d)}_{j+p_0}\right) \geq \left(c^{(s)}_{j+p_0} - t^{(s)}_{j+p_0}\right) = f^{(s)}_{j+p_0}. \tag{3.48}$$

We can now prove by induction that if there exists p_0 number of jobs processed till time t_0 for which Proposition 3.3 is true, then, according to Equation (3.48), Proposition 3.3 does also hold for all the jobs processed after t_0. Assuming that the shop is initially empty, Equation (3.45) implies that the flow times of the processed jobs in the static case are identical to the flow times of the processed jobs in the dynamic case until the lead time is first updated at state $w(t) = \bar{w} + r - 1$. This provides a starting condition for p_0 in Equation (3.48). As a consequence, for any given sample path of arrivals and processing times of jobs, the flow time of the j^{th} processed job in the dynamic case is greater than or equal to the flow time of the j^{th} processed job in the static case, for any j. This yields

$$\Pr\{C(r) \geq l\} \geq \Pr\{C(\infty) \geq l\}.$$

This completes the proof of Proposition 3.3.

Proof of Proposition 3.4

Using the explicit derivation of the rate matrix in Equations (3.9) and (3.10), and $p_0^{(r)}$ given in Equations (3.11) and (3.12), the probability mass function of the planned lead time is given by

$$\Pr\{L(r) = L_{\min} + l\} = \begin{cases} 1 - \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho}, & l = 0 \\ \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho} \cdot (1-\eta)(\eta)^{l-1}, & l = 1, 2, \ldots. \end{cases}$$

Then, the dynamic planned lead time can be expressed as $L_{\min}$ plus a random variable of which the probability distribution is based on the update strategy and the system characteristics. $L(r)$ is rewritten in terms of its distribution characteristics as follows:

$$L(r) = L_{\min} + \begin{cases} 0, & \text{with probability } 1 - \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho} \\ 1 + \tilde{L}_r, & \text{with probability } \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho} \end{cases}$$

where the random variable $\tilde{L}_r$ has a geometric distribution with parameter η. It should be noted that η is the largest eigenvalue of the rate matrix $R^{(r)}$.

$$\Pr\left\{\tilde{L}_r = l\right\} = (1-\eta)\eta^l,\, l = 0, 1, \ldots.$$

Consequently, the average lead time is given by

$$\begin{aligned} E\left(L(r)\right) &= L_{\min} + \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho} \cdot \left(1 + E\left(\tilde{L}_r\right)\right) \\ &= L_{\min} + \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho - \rho^{r-1}(\rho+\beta)}. \end{aligned} \tag{3.49}$$

Similarly, using the moments of the geometric random variable $\tilde{L}_r$, we solve for the second moment of $L(r)$. From Equation (3.49), it is stated that the distribution of the dynamic planned lead time changes according to the term

$$\xi = \frac{(1+\beta)\rho^{r-1}\rho^{\bar{w}+1}}{1+\beta\rho - \rho^{r-1}(\rho+\beta)}.$$

Then,

$$\begin{aligned} E\left(L^2(r)\right) &= L_{\min}^2(1-\xi(1-\eta)) + \xi(1-\eta)E\left(\left(L_{\min} + 1 + \tilde{L}_r\right)^2\right) \\ &= L_{\min}^2 + (2L_{\min}+1)\xi + \frac{2\xi\eta}{1-\eta}, \end{aligned}$$

which simplifies to

$$E\left(L^2(r)\right) = L_{\min}^2 + \xi\left(2L_{\min} + \frac{1+\eta}{1-\eta}\right). \tag{3.50}$$

From the derivations of $E\left(L(r)\right)$ and $E\left(L^2(r)\right)$ in Equations (3.49) and (3.50), the coefficient of variation of the dynamic planned lead time is directly written by

$$CV\left(L(r)\right) = \frac{\sqrt{\xi\left(\frac{1+\beta\rho+\rho^{r-1}(\rho+\beta)}{1+\beta\rho-\rho^{r-1}(\rho+\beta)} - \xi\right)}}{L_{\min} + \xi}. \tag{3.51}$$

In addition,

$$\begin{aligned} \frac{\partial}{\partial(\beta)}\xi &= \frac{\rho^{\bar{w}+r}\left(1+\rho^{r-1}\right)(1-\rho)}{\left(1+\beta\rho - \rho^{r-1}(\rho+\beta)\right)^2}, \\ \frac{\partial}{\partial(r)}\xi &= \frac{(1+\beta)\rho^{\bar{w}+r}(1+\beta\rho)\ln(\rho)}{\left(1+\beta\rho - \rho^{r-1}(\rho+\beta)\right)^2} \end{aligned}$$

provide the fact that for $\rho < 1$, $\frac{\partial}{\partial(\beta)}\xi > 0$ for all β, and $\frac{\partial}{\partial(r)}\xi < 0$ for all r. Together with Equations (3.49) and (3.51), this completes the proof of Proposition 3.4.

Chapter 4

WIP Clearing in Supply Chain Operations Planning

In this chapter, we provide insights into the effectiveness of the clearing function concept in a hierarchical planning context. The clearing function is a mathematical representation of the relation between WIP and throughput of a production process, and it is used to anticipate the flow times of planned order releases, which are subject to uncertainties. The SCOP formulation of Chapter 2 is modified to include the WIP and the clearing function for a single-item produced in a production unit and kept in a stock point facing a stochastic non-stationary demand. At the SCOP level, the delivery schedules of the orders are determined through fixed planned lead times, and capacity loading decisions are separated from order release decisions in a way to plan for on-time deliveries. Early or late deliveries of the orders, which significantly affect the inventory costs, have not been considered explicitly by the previous studies on clearing functions. In this chapter, clearing functions arising from different modeling approaches are tested by simulation based on the quality of the plans in terms of the total cost and the level of operational consistency. The results indicate that modeling the clearing of WIP should be based on the short-term operational dynamics of the production unit.

4.1 Introduction

A model of a production process is embedded in each planning and scheduling tool. It is a mathematical representation of the set of technologically feasible operations within the production process. However, only a very narrow range of models of production has been considered in the context of SCOP. Traditional MP models assume fixed production capacity with zero or fixed flow times, and avoid considering the effect of WIP levels on shop congestion, throughput, and accordingly on order flow times (e.g., Billington *et al.* (1983), Shapiro (1993), and MRPII formulations in Voß and Woodruff (2003)). Recently, Spitter *et al.* (2005a) decomposed the production of orders over multiple periods during a fixed planned lead time, assuming no modeled relation between throughput and WIP levels. From queuing theory (Little's law), it is well known that the flow times and the throughput levels depend on the loading of the shop floor. It can therefore be argued that the production model used in SCOP should be based on actual queuing characteristics of the shop floor. This follows the line of reasoning proposed by Karmarkar (1987), and Hopp and Spearman (2000). These studies are based on the long-term relationships assuming stationary conditions, and do not explicitly model or recognize a higher-level order release mechanism. In this chapter, we are particularly interested in the interaction between the timings and the sizes of the release decisions and the operational execution in the shop floor, and the consequences of this interaction on the system performance. We concentrate on the consistency of generated plans with their actual executions. At the SCOP level, release decisions are given in anticipation of the short-term performance of the production processes in terms of the interaction between capacity loading, throughput and the flow times. This interaction is modeled through the concept of the clearing of WIP during each period. In a dynamic framework, the status information about the workload levels can be utilized through clearing functions at the SCOP level to anticipate dynamic flow times of order releases.

First initiated by Graves (1986), there has been a growing interest on the use of clearing functions in deterministic aggregate production planning models (see Pahl *et al.* (2005) for an overview of the literature on clearing functions). Most of the studies on clearing functions follow the line of reasoning of Hopp and Spearman (2000), and model the clearing functions based on the steady-state queueing constructs. We refer to Section 1.3.3 of this thesis for a brief discussion about a list of various studies on clearing functions. Although all of these studies have contributed to the development of the general framework for modeling production (see Hackman and Leachman (1989) for a detailed discussion), the concept of planned lead times was not considered in their models. The planned lead times are crucial in planning production of batches and inventory of items among various stages within a supply chain. Thus, it is worthwhile to do research on the use of clearing functions in coordinating the flow of material between a production unit and a stock point within a supply chain context.

In this chapter, we consider a two-level decision hierarchy, where orders and raw materials are released to the production unit by the SCOP level, and the released orders are scheduled both according to their planned lead times and the capacity restrictions by the operational scheduling level. Our objective is two-fold. Firstly, we want to provide a hierarchical framework for planning supply chain operations, where the concept of clearing is used to plan and control WIP levels in a way to achieve consistent plans in terms of the actual delivery schedule versus the planned schedule. Our study is different from the previous studies on clearing functions because, we explicitly consider the concept of planned lead time in the model so that the delivery performance of the released orders and its consequences can be measured effectively. Accordingly, our framework can be extended to model more complex supply chain situations. Separate production units in a supply chain are coordinated by the release of orders and their planned lead times. The level of consistency between the planning and the execution plays an important role in the performance of such systems. To our knowledge there is no literature on this aspect of consistency; some results based on experimental data have already been presented in Asmundsson *et al.* (2006) on the level of fit between the approximated clearing function and the actual clearing behavior, however, their study is not conducted in a hierarchical planning system with planned lead times. Secondly, it is not obvious what type of a model should form the basis of the clearing function. We develop an alternative approach in which we define the throughput quantity as a random variable with the probability distribution based on the available WIP during a period and the short-term probabilistic behavior of the shop floor. We test the performance of the established clearing functions (see Figure 1.5) and this alternative in various settings using the simulation of a single-stage, single-product supply chain.

4.2 Problem Definition and Modeling

We consider a production unit with uncertainty in its operations and a downstream stock point that sees a non-stationary stochastic demand for a single product. Unmet demand is satisfied through backorders. The orders are released in batches and sent to the stock point in batches, facilitating the need to keep finished items inventory by the production unit. This inventory is denoted by the term *finished WIP* to differentiate it both from the unfinished items waiting to be processed in the shop (WIP) and the finished products in the stock point. The order releases are aimed at the satisfaction of the forecasted demand with a time lag equal to the planned lead time. Separately, the capacity loading decisions, for a given clearing function, are responsible for planning periodic throughput levels such that the orders can be delivered on time. It is assumed that the production unit has ample stock of raw materials and the transportation times between the raw material stock point and the production unit and between the production unit and the finished item stock point are negligible. This is because, we want to concentrate on the analysis of the clearing functions that model the characteristics of the production process. The flow time of an order consists of the waiting time and the batch processing time that elapse in the production unit.

The planning is done in a periodic review setting for a certain planning horizon, and at the start of each period, the method of rolling horizons is applied to replan according to the random deviations due to stochastic production and demand processes. Between each consecutive planning the following sequence of events occurs: an order is released and scheduled among the previously released orders, and a planned quantity of work is loaded to the production unit as WIP at the start of the period, actual demand and throughput are realized, and finished orders are sent to the stock point at the end of the period. The status information of the production unit and the stock point is updated, and given these inputs, replanning is done starting from the next period.

The overall planning process consists of two hierarchical levels as illustrated in Figure 4.1. At the operational planning level, a multi-period single-product SCOP problem is solved for a predetermined number of periods. A linear programming formulation is used to determine optimal capacity loading and order release decisions. SCOP anticipates the expected throughput in a period based on the WIP level to be loaded to the shop floor at the start of that period. The size and the release time of each order can then be determined according to the planned lead time and the anticipated throughput quantities. At this point, for the sake of clarity, one should note that the production unit does not face a continuous arrival of work during each period, but all the WIP planned for a period is made available at the start of the period[1]. As outcome of the SCOP level, the order release for the first period in the planning horizon is given to the lower decision level for detailed scheduling together with the currently open orders. The capacity loading decision for the first period is

[1] This assumption on the release of raw materials is also utilized in Graves (1986).

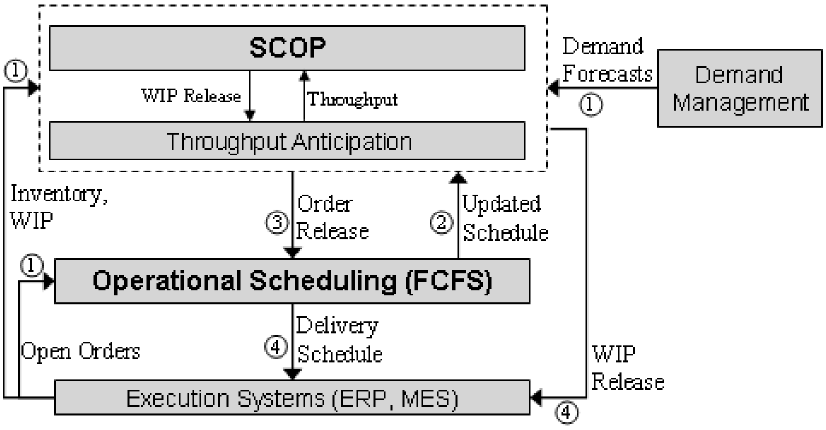


Figure 4.1: Planning system, information flow and decision hierarchy.

considered as part of the detailed production plan, and is directly instructed to the execution systems.

At the operational scheduling level, given the planned lead time and the release dates of the orders, the delivery schedule is planned according to the FCFS strategy. We provide a decomposition of the planning decisions on material flow and capacity from the execution decisions on finalized orders. Key argument for this hierarchical decomposition is that the lower level's actual execution generally differs from the planned outcome, which also motivates us to replan through rolling horizons. This aspect is ignored in the static models, and drives the need to adjust the current delivery schedule in response to the infeasible capacity requirements. The new schedule is then feedforward to the operational planning level.

The circled numbers in Figure 4.1 refer to the sequence of information flow in the planning system. Data set (1) refers to the input to the planning system regarding the statuses of the production unit and the stock point together with the demand forecasts. The operational scheduling level performs a check on the achievable delivery schedules of the open orders, and updates the current schedule if necessary. The updated schedule (given by the data set (2)) is feedforward to the SCOP level, and is used by the SCOP formulation in deciding the timings and the sizes of new order releases. Data set (3) refers to the first period's order release, which is then added to the end of the current delivery sequence by the lower level according to FCFS. The delivery schedule including the newly released order and the amount of WIP to be loaded to the shop are finalized decisions, and are given to the execution systems such as ERP, and MES within data set (4).

4.2.1 Clearing Functions

The clearing function is an approximate representation of the production process at an aggregate level of modeling. Although the underlying modeling assumptions may differ between different types of clearing functions, the clearing function gives the expected level of throughput to be realized during a period as a function of the available WIP during that period. In this chapter, there are four different clearing functions modeled and examined in terms of their relative effects on the cost and the delivery performance. These are:

1. The traditional model with the assumption that WIP can be cleared with the nominal production rate. We abbreviate it as *"Traditional Linear" (TL)* function, since it is used in the majority of production planning models.
2. The fixed lead time fixed capacity model where the clearing is proportional to the planned lead time, and is bounded by the nominal production rate. We abbreviate it as *"Capacitated Fixed Lead Time" (CFL)* function.
3. The long-term clearing function based on the steady state queueing constructs. It is abbreviated as *"Long-Term Nonlinear" (LTN)* function.
4. The short-term clearing function based on the short-term probabilistic behavior of the shop. We abbreviate it as *"Short-Term Nonlinear" (STN)* function.

In the following, the mathematical representations of these clearing functions are provided. Although the underlying modeling assumptions may be different, all clearing functions are defined in the discrete-time domain. Thus, a common notation is used irrespective of the basis that forms the clearing function. w is defined as the available WIP level during a period, and $p = f(w)$ is the total expected throughput quantity measured in number of items produced during that period with a nominal average value of μ items per period, where $f(w)$ denotes the clearing function. d refers to the demand rate per period, and in the long-run, throughput rate must be equal to the demand rate. The term period refers to some fixed duration of time within which the basic input/output parameters of the system are defined such as the demand and the throughput rates, and it is identical for all clearing functions.

The *TL* and *CFL* clearing functions are both based on a deterministic view over the production process. *TL* assumes immediate production of what is available in the shop bounded by the nominal production rate. In other words, a SCOP model without a decomposition of order release decisions from the capacity loading decisions and with *TL* used to model the production process becomes very similar to the MRP and MRPII formulations of Voß and Woodruff (2003) with the exception that the production output is partly smoothed through the nominal production rate. In an environment where the capacity is infinitely large relative to the sizes of the order releases, such a formulation indicates immediate production of order releases with an independent time-lag after the production till the order becomes ready for demand consumption. The formulation of *TL* is given by

$$\textit{TL:}\quad f(w) = \begin{cases} w & \text{if } w \leq \mu \\ \mu & \text{otherwise} \end{cases} \tag{4.1}$$

The *CFL* clearing function is similar to the *TL* clearing function in a way that a deterministic production rate is assumed that is linearly dependent on the WIP level, and is bounded by the nominal production rate μ. Differently, a smoother production output is considered based on the planned lead time in the same manner as in Graves (1986). The *CFL* formulation is

$$\textit{CFL}:\quad f(w) = \begin{cases} w/L & \text{if } w \leq \mu L \\ \mu & \text{otherwise} \end{cases} \tag{4.2}$$

When the formulation is carried over a continuous-time domain, for example in deriving LTN, w refers to the average WIP level and p refers to the average throughput level per period. As also noted in Hackman and Leachman (1989) for a general framework, in a discrete-time production model, each rate-based flow is a step function, constant in each period.

For *LTN*, the production process is modeled as an $M|G|1$ queue where orders arrive in batches of size x according to exponential inter-arrival times with mean $1/\lambda$ period. Items are processed in exponentially distributed processing times with rate μ yielding $Erlang(x,\, \mu)$ distributed processing time for each batch. It is important to note that the items arrive in batches but processed individually, and the clearing functions in this chapter express the throughput-WIP relationship in terms of items. Thus, the derivation of *LTN* in this chapter shows slight differences from the clearing functions in the previous studies utilizing long-term queueing constructs (e.g., Hwang and Uzsoy (2005)). We should also note that the exponential arrival process with fixed batch sizes is an approximation on the real arrival process, where batches of (possibly) different sizes are released according to a process that may (not) show an exponential pattern. Intuitively, the actual release pattern does depend on the specific clearing function used in the SCOP model, which in turn depends on some approximation of the actual release pattern. Analyzing this cyclic structure is a complex task. It is not in the scope of this chapter, but remains as an interesting direction for future research.

From Little's law we have the fundamental relationship, $p = w/F_{\text{item}}$, where F_{item} is the average flow time per item in the batch. Let us define the time it takes to process an item starting from the first in the batch as *batch-item processing time.* The batch-item processing time of an item depends on the position of that item in the batch, and, as mentioned in Lambrecht and Vandaele (1996), the average batch-item processing time in a batch of size x can be defined as $PT_{\text{item}} = (x+1)/2\mu$. From Polaczek-Khinchine mean-value formula for $M|G|1$ systems the average waiting time for each batch is given by

$$WT_{\text{batch}} = \frac{\lambda x(x+1)}{2\mu(\mu - \lambda x)}$$

The waiting time of an item in the batch before a batch starts to be processed is equal to the waiting time of the batch itself. Then,

$$F_{\text{item}} = WT_{\text{batch}} + PT_{\text{item}} = \frac{x+1}{2(\mu - \lambda x)}.$$

The arrival rate of the batches can be rewritten by $\lambda = p/x$. Then, from Little's law,

$$p = \frac{w}{F_{\text{item}}} = \frac{2w(\mu - p)}{x+1}$$

Solving this equation for p, and assuming a batch size equal to the average demand rate of d, the following formulation for the long-term nonlinear clearing function can be derived:

$$LTN\colon \quad f(w) = \frac{2\mu w}{2w + d + 1}. \tag{4.3}$$

This approximation can also be proved by solving an $M^{\lambda}|M^{\mu}|1$ system with bulk arrivals of size d. Given that the batch size is one, $d = 1$, then LTN becomes identical to the clearing function in Karmarkar (1989). Equation (4.3) directly implies that, for any given positive μ and d, LTN is increasing and concave for $w \geq 0$.

The *LTN* clearing function is insightful in modeling the nonlinearity in the production process. However, it generates some conceptual inconsistencies when used in an aggregate production planning problem involving a limited planning horizon and the planning periods that are mainly characterized by the starting and the ending statuses. The exact continuous-time formulation in Equation (4.3) is based on the long-term averages, and becomes an approximation in a periodic planning environment. The throughput is a step function in time. Its rate is determined based on the starting WIP level, and is considered as fixed during a single period. One may criticize this approximation by stating that *LTN* yields a throughput level that is greater than the WIP level for some cases such as when $w < \mu - (d+1)/2$, and the nominal throughput rate is only achieved with a very high WIP level causing the planning system to load the shop by a large amount of work during shortage periods. In an environment with expensive materials, the latter may be very costly. As an alternative to *LTN*, the clearing function may also be based on the short-term (periodic) probabilistic analysis of the throughput rate of the shop given the WIP level at the start of a planning period.

The *STN* clearing function refers to such a model of production emphasizing the short-term behavior of the production process between consecutive work release (arrival) opportunities, and it does not consider the average behavior driven by multiple number of work releases (arrivals) to the production system. The distribution of throughput probabilities in a period is dependent on the WIP level at the start of the period because, a throughput quantity greater than the total quantity of items in process is not possible. Given a WIP level of w available during a period, let us denote the random variable for the throughput level in that period as P_w. Following the previously defined terminology, the expected throughput in a period with a WIP level of w is

$$p = \sum_{k=1}^{w} k \Pr\left\{P_w = k\right\}.$$

Rewriting this equation yields

$$p = w - \sum_{k=0}^{w-1} (w-k) \Pr\left\{P_w = k\right\},$$

and with further simplification by replacing the term $1 - \sum_{j=0}^{k-1} \Pr\left\{P_w = j\right\}$ in the above formulation with $\Pr\left\{P_k = k\right\}$, for $k = 1, 2, \ldots, w$, the short-term nonlinear clearing function is derived as follows:

$$STN\colon \quad f(w) = \sum_{k=1}^{w} \Pr\left\{P_k = k\right\}. \tag{4.4}$$

Equation (4.4) is applicable for any assumption on the distribution of processing times. In this chapter, we assume exponential processing times with rate

μ. Then, the probability that all items are cleared in a single period is

$$\Pr\{P_w = w\} = 1 - \sum_{k=0}^{w-1} e^{-\mu} \cdot \frac{\mu^k}{k!}.$$

When $w \to \infty$, P_w becomes identical to a Poisson random variable with mean μ. So, both LTN and STN asymptotically approach to the nominal production rate as $w \to \infty$. However, STN approaches to μ more quickly then LTN does, and this makes an obvious distinction in the output of the SCOP model.

The clearing representation in Equation (4.4) is conceptually robust, since the linear clearing functions can also be modeled by this representation. Consider an uncapacitated situation with the assumption that what is put in the shop can immediately be produced, $\Pr\{P_w = w\} = 1$, then STN yields $f(w) = w$ for all w. In addition, the *CFL* clearing function can be modeled by setting $\Pr\{P_w = w\} = 1/L$ for $w \leq L\mu$ and $\Pr\{P_w = w\} = 0$ for $w > L\mu$. From Equation (4.4), it directly follows that STN is an increasing and concave function for $w \geq 0$.

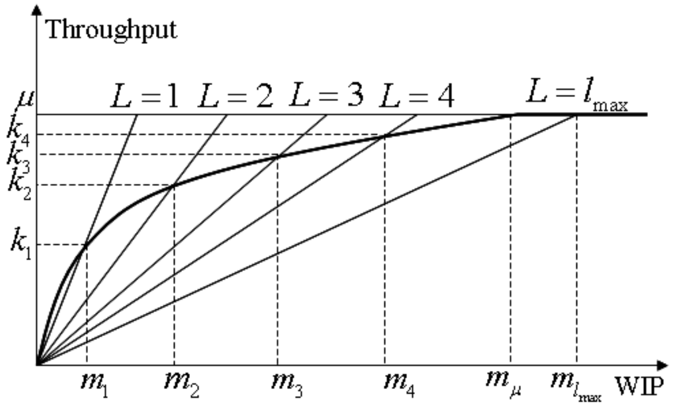


Figure 4.2: Lead time regions of a clearing function.

In the SCOP formulation, a piecewise-linear approximation of the clearing function is employed. For this purpose, we define *lead time regions* in the domain of the clearing function. An illustration is given in Figure 4.2 for an arbitrary nonlinear and concave clearing function, which is partitioned by $L = l$ lines that cut the clearing function at points (m_l, k_l), $k_l = m_l/l$. Hence, the clearing function is partitioned into different WIP ranges and corresponding throughput ranges. WIP levels up to m_1 can be cleared in a single period indicating $L = 1$ region. If the WIP level is between m_{l-1} and m_l, then the throughput level ranges between $k_{l-1} = m_{l-1}/(l-1)$ and $k_l = m_l/l$ implying a flow time of l periods. An intuitive approximation for a nonlinear clearing function is to assume that the clearing is linear between lead time shift points, and its slope differs between different lead time regions due to the concavity of the clearing function. The reason behind this intuition is that the clearing function reveals a fixed flow time independent of the WIP level once the production unit operates within a WIP range of a certain lead time region.

For *STN* and *LTN*, the nominal production rate, μ, may never be reached. We assume the clearing function achieves its nominal production rate when the slope of the piecewise-linear approximation falls at or below 0.01.

The piecewise-linear approximation for a nonlinear clearing function is formalized as in the following equation, assuming that the production unit reaches its nominal throughput rate at $w = m_\mu$, and the lead time regions up to $L = l_{\max}$ cover all clearing function.

$$f(w) \approx \begin{cases} w, & w \le m_1 \\ k_1 + \frac{k_2 - k_1}{m_2 - m_1} \cdot (w - m_1), & m_1 < w \le m_2 \\ \vdots & \vdots \\ k_{l_{\max}-1} + \frac{\mu - k_{l_{\max}-1}}{m_\mu - m_{l_{\max}-1}} \cdot (w - m_{l_{\max}-1}), & m_{l_{\max}-1} < w \le m_\mu \\ \mu, & m_\mu < w \end{cases} \tag{4.5}$$

4.2.2 SCOP

At the SCOP level, operational planning is done in terms of releasing the orders and planning the throughput levels to satisfy the planned order releases on time in such a way that the forecasted demand is met by holding minimum amount of material both in the production unit and in the stock point. In the formulations used in this section, $t = 0$ denotes the current planning period, and $f(\cdot)$ indicates a piecewise-linear clearing function. The static exogenous parameters of the SCOP model are defined as follows:

h_f = Per unit, per period cost of holding finished item inventory at the stock point.

h_w = Per unit, per period cost of holding WIP at the production unit.

$\hat{h}_f$ = Per unit, per period cost of holding finished WIP at the production unit.

M = Per unit, per period penalty cost for the inventory shortage at the stock point.

μ = Nominal production rate of the production unit.

L = Planned order lead time.

T = Planning horizon.

The dynamic exogenous inputs to the SCOP model are updated at every replanning opportunity. These constitute the demand forecasts, the status information about the production unit and the stock point, and a capacity feasible delivery schedule from the operational scheduling level. They are

$D(t)$ = Forecasted demand for period t, $t = 0, \ldots, T-1$.

$\widehat{Q}(t)$ = Total quantity, among the currently open orders, scheduled for receipt at the stock point at the start of period t, $t = 1, \ldots, L$.

$I^+(0)$ = Current on-hand inventory level at the stock point.

$I^-(0)$ = Current backorder level at the stock point.

$W(0)$ = Current WIP level at the production unit.

$\widehat{I}(0)$ = Current finished WIP level at the production unit.

The decision variables include the system variables such as the on-hand inventory level at the stock point, WIP and finished WIP levels at the production unit, and the planned throughput quantity. The variables that are not executed but used in the plan for evaluation and anticipation purposes are

$I^+(t)$ = Inventory on-hand at the stock point at the start of period t, just before the orders scheduled for period t are received, $t = 1, \ldots, T$.

$I^-(t)$ = Backorder level at the stock point at the start of period t, just before the orders scheduled for period t are received, $t = 1, \ldots, T$.

$W(t)$ = WIP level at the production unit at the start of period t, just before the release of additional WIP into the shop, $t = 1, \ldots, T-1$.

$\widehat{I}(t)$ = Finished WIP level at the production unit at the start of period t, just after the orders scheduled for period t are sent to the stock point, $t = 1, \ldots, T-1$.

$P(t)$ = Planned throughput level in period t, $t = 0, \ldots, T-2$.

The executable decisions are the order release and the capacity loading decisions, which are denoted by

$Q(t)$ = The size of the order to be released at the start of period t, $t = 0, \ldots, T-L-1$.

$R(t)$ = Amount of additional WIP to be loaded to the production unit at the start of period t, $t = 0, \ldots, T-2$.

The SCOP problem is modeled by the following linear programming formulation. It should be noted that all the constants and the variables in this formulation are nonnegative. For $t < L$, the order release variables $Q(t-L)$ are set to zero because, the decisions at time zero are limited to future order releases.

$$\text{Min.} \sum_{t=1}^{T} \big(h_f \cdot I^+(t) + M \cdot I^-(t)\big) + \sum_{t=1}^{T-1} \Big(h_w \cdot W(t) + \hat{h}_f \cdot \widehat{I}(t)\Big) \tag{4.6}$$

s.t.

$$I^+(t+1) - I^-(t+1) = I^+(t) - I^-(t) + Q(t-L) + \widehat{Q}(t) - D(t),$$
$$t = 0, \ldots, T-1 \tag{4.7}$$
$$W(t+1) = W(t) + R(t) - P(t),\ t = 0, \ldots, T-2 \tag{4.8}$$
$$P(t) \leq f\left(W(t) + R(t)\right),\ t = 0, \ldots, T-2 \tag{4.9}$$
$$\widehat{I}(t+1) = \widehat{I}(t) + P(t) - Q(t+1-L) - \widehat{Q}(t+1),\ t = 0, \ldots, T-2 \tag{4.10}$$

Constraint set (4.7) defines the balance of inventory at the stock point between consecutive planning periods using information on the current schedule and assuming that the order released at the start of period k will be available at the stock point at the start of period $k+L$. Constraint set (4.8) determines the WIP balance, where WIP is increased by the amount of work loaded to the shop and decreased by the throughput. Constraint set (4.9) provides the throughput and capacity loading relationship according to a piecewise-linear and concave clearing function. The finished WIP balance equations are modeled in constraint set (4.10), which implies that each order has to be finished and shipped to the stock point within its planned lead time. It also indicates, by the term $\widehat{Q}(t+1)$, that the schedule that is feedforward from the operational scheduling level must be met. It is important to note that this formulation for the SCOP is based on a safety stock level of zero. A safety stock adjustment procedure to guarantee a certain demand fill rate in our simulation models has been provided in the Appendix of Chapter 2.

If we ignore the finished item inventory holding cost and the shortage cost in the objective function and the constraint set (4.7), the resulting formulation is very similar to the aggregate production planning models presented in Karmarkar (1989), Asmundsson *et al.* (2003) and Asmundsson *et al.* (2004), where the term in constraint set (4.10), $Q(t+1-L) + \widehat{Q}(t+1)$, can be considered as an exogenous demand for period t.

As the concept of clearing implies, the throughput in a period is a function of the total WIP level at the start of that period. The capacity loading decisions are driven by the desired level of throughput quantities because, as formulated in constraint set (4.9), $R(t)$ determines the range in which $P(t)$ can be planned. It is obvious from the SCOP formulation that additional WIP loaded to the production unit each period increases the total costs. Thus, for a desired throughput level of $P(t) > f\left(W(t)\right)$, $R(t)$ is decided such that

$$P(t) = f\left(W(t) + R(t)\right), \, t = 0, \ldots, T-2,$$

and for a desired throughput level of $P(t) \leq f\left(W(t)\right)$, $R(t) = 0$. In other words, the SCOP model loads the capacity by the exact quantity that would increase the throughput to a desired level, and if the throughput needs not be increased, then no additional WIP is loaded to the production unit.

4.2.3 Scheduling and Rescheduling

At the start of the current planning period, $t = 0$, $Q(0)$ is released to the scheduling level from the SCOP level. At the scheduling level, decisions regarding the immediate execution of the past and the current order releases are provided. Thus the notation changes slightly in this section compared to the notation in the SCOP model. Given the current sequence of the open orders, $\left\{(q_1, \widehat{dd}_1), \ldots, (q_K, \widehat{dd}_K)\right\}$, as planned by the scheduling level where the order quantities, q_i's, are coupled with their planned delivery dates, $\widehat{dd}_i$'s, the newly released order is added at the end of the sequence, $(q_{K+1}, \widehat{dd}_{K+1}) = (Q(0), L)$. Then the new schedule is instructed to the execution systems. In a static framework, this is an implementation of the FCFS dispatching rule. However, in a dynamic framework that includes the plan-execute-(re)plan cycle, due to the stochastic nature of the production process, the actual throughput quantities may deviate from their planned values. As a result, there is a need to update the active schedules at each replanning according to the capacity considerations. A feasible delivery schedule has to be input to the top-level SCOP model from the bottom-level scheduling model at each replanning opportunity. For this purpose, a schedule update heuristic is applied.

Given that an arbitrary order is expected to be late, the heuristic finds the earliest period by which the order is expected to be finished and sent to the stock point. Let us define the current schedule of the open orders in the production unit, provided by the execution system to the scheduling level by the set $\widetilde{X} = \left\{(q_1, \widetilde{dd}_1), \ldots, (q_K, \widetilde{dd}_K)\right\}$. Due to the FCFS rule, $\widetilde{dd}_1 \leq \widetilde{dd}_2 \leq \ldots \leq \widetilde{dd}_K$. For brevity, assign $\widetilde{dd}_i = 1$ if the order $(q_i, \widetilde{dd}_i)$ is already late at the start of the current period $t = 0$. An arbitrary order $(q_i, \widetilde{dd}_i) \in \widetilde{X}$ is expected to be late if and only if

$$\widehat{I}(0) + \widetilde{dd}_i \mu - \sum_{j=1}^{i-1} q_j < q_i.$$

We denote the set of late orders as $\widetilde{X}_{\text{late}}$. The new schedule is determined according to the following rule for all $(q_i, \widetilde{dd}_i) \in \widetilde{X}_{\text{late}}$:

$$\widehat{dd}_i = \min_{s=0,1,\dots} \left\{ \widetilde{dd}_i + s : \widehat{I}(0) + (\widetilde{dd}_i + s)\mu - \sum_{j=1}^{i-1} q_j \geq q_i \right\}, \tag{4.11}$$

where $\widehat{dd}_i$ is the period by which order $(q_i, \widetilde{dd}_i)$ is scheduled for receipt in the stock point after updating the current schedule. For $(q_i, \widetilde{dd}_i) \in \widetilde{X} \setminus \widetilde{X}_{\text{late}}$, $\widehat{dd}_i = \widetilde{dd}_i$. The new schedule, $\widehat{X} = \left\{ (q_1, \widehat{dd}_1), \dots, (q_K, \widehat{dd}_K) \right\}$, is then embedded in the SCOP model as an input.

$$\widehat{Q}(t) = \sum_{j=1}^{k} 1(\widehat{dd}_j = t) q_j, \quad t = 1, 2, \dots, L,$$

where the indicator function $1(\widehat{dd}_i = t)$ is equal to 1 if $\widehat{dd}_j = t$, and 0 otherwise. The updated schedule has to be consistent with the FCFS scheduling discipline, and avoids order crossings as given in the following proposition:

Proposition 4.1 *The FCFS rule applies for any given pair of consecutive orders in $\widehat{X}$. Given $(q_i, \widehat{dd}_i) \in \widehat{X}$ and $(q_{i+1}, \widehat{dd}_{i+1}) \in \widehat{X}$,*

$$\widehat{dd}_i \leq \widehat{dd}_{i+1},$$

and the scheduled delivery duration does not exceed the planned lead time for any given order in $\widehat{X}$. Given $(q_i, \widehat{dd}_i) \in \widehat{X}$,

$$\widehat{dd}_i \leq L$$

Proof See the Appendix at the end of this chapter.

Proposition 4.1 is important for our dynamic hierarchical planning framework because, it reveals the fact that the SCOP models in successive replanning opportunities are conceptually consistent. This is achieved by feedforwarding the status information as seen by the operational scheduling level to the SCOP level.

4.3 Simulation Experiments

4.3.1 Setting

The production process is considered as a single entity with an exponentially distributed processing time per item. It may consist of a single machine or a complex manufacturing center where the transformation time of WIP into final products is exponentially distributed. Our motivation in aggregating a complex manufacturing center into a single entity stems from the fact that clearing functions are originally defined based on an aggregate approach to the production process. The processing rate is 20 items/period, $\mu = 20$. The demand forecasts, $D(t)$, $t = 0, 1, \ldots, T-1$, are generated from a *Gamma* distribution with mean d, and squared coefficient of variation 0.5, revealing an average utilization level of $\rho = d/\mu$. The planning horizon is $T = 10$ periods. The cost parameters are $h_f = 1.25$, $h_{fw} = 1.20$, and $h_w = 1.00$. "Today, direct labor constitutes less than 15% of the cost of most products" (cf. Hopp and Spearman (2000)), and material holding costs have the lion's share in this figure. In addition, holding high WIP requires large space in the shop floor, incorporating handling and spacing costs additional to the material costs. Therefore, the WIP holding costs and the final product holding costs are chosen close to each other. Each simulation run starts with initial conditions: $I^+(0) = Ld$, $I^-(0) = 0$, $W(0) = 0$, and $\widehat{I}(0) = 0$. That is, the shop is empty and there are enough items in the inventory for the demand during the planned lead time.

The *TL* and *CFL* clearing functions are modeled as defined in Equations (4.1) and (4.2). The piecewise-linear approximation of the *STN* clearing function is based on Equation (4.5), and is modeled as follows:

$$STN\colon \quad f(w) = \begin{cases} w, & w \leq 9 \\ 5.04 + 0.44w, & 9 \leq w \leq 34 \\ 20, & 34 \leq w \end{cases} \tag{4.12}$$

In *STN*, 100% productivity is disturbed starting from a workload level of 9 items, and the nominal throughput rate, 20 items/period, is reached when there are 34 items in process. For *LTN* we use $d = 17$ in Equation (4.3). It assumes higher (or equal) productivity than *STN* does for low levels of workload, $w \leq 16$, and quickly degrades in productivity as the shop congestion increases. The nominal throughput rate achieved by *LTN*, 18.71 items/period, is less than that of *STN*. The piecewise-linear approximation of the *LTN* clearing function is based on Equation (4.5), and is given by

$$\mathit{LTN}\colon \quad f(w) = \begin{cases} w, & w \leq 11 \\ 8.53 + 0.225w, & 11 \leq w \leq 31 \\ 13.18 + 0.075w, & 31 \leq w \leq 51 \\ 15.09 + 0.038w, & 51 \leq w \leq 71 \\ 16.15 + 0.023w, & 71 \leq w \leq 91 \\ 16.84 + 0.015w, & 91 \leq w \leq 111 \\ 17.31 + 0.011w, & 111 \leq w \leq 131 \\ 18.71, & 131 \leq w \end{cases} \tag{4.13}$$

The SCOP formulation together with the specific shape of the clearing function used in the formulation implies that WIP is bounded by the level

$$W_{\max} = \min\{w : f(w) = \mu'\}$$

because, increasing the WIP further does not add to the planned throughput quantities but increases the cost. Here, μ' is the nominal throughput rate indicated by the clearing function, which may be less than the theoretical maximum. For *TL*, $W_{\max} = 20$, for *CFL*, $W_{\max} = 20L$, for *LTN*, $W_{\max} = 131$, and for *STN*, $W_{\max} = 34$. One would expect the higher the $W_{\max}$ is, the higher the WIP costs are, but at the same time, the lower the $W_{\max}$ is, the more likely the shop becomes idle during a period increasing the possibility of late deliveries in the next periods.

4.3.2 Design

In addition to the type of the clearing function used in the SCOP model, we provide environmental factors such as demand uncertainty and utilization in the design of experiments. The demand uncertainty is modeled by a percent deviation from the forecasted demand. If the forecasted demand for a period is $\hat{d}$, the actual demand, for the same period, with 40% deviation is generated from a Uniform($0.60\hat{d}$, $1.40\hat{d}$) distribution, and with 80% deviation is generated from a Uniform($0.20\hat{d}$, $1.80\hat{d}$) distribution. The utilization is changed via changing the demand levels. Given the fixed value of $\mu = 20$, 80% utilization is achieved by setting $d = 16$, and 90% utilization is achieved by $d = 18$. Different planned lead times are also considered. The planned lead time can be either short, $L = 3$ periods, or long, $L = 5$ periods. These values are generated by rounding the mean flow time found from Polaczek-Khinchine mean value formula for an $M|G|1$ system with batch processing under utilization levels of 0.80 and 0.90 respectively. Table 4.1 provides the design of experiments.

Table 4.1: Experimental design.

Factors	Treatments	Number of Treatments
Clearing function	*CFL*, *LTN*, *STN*, *TL*	4
Demand uncertainty, U_D	40%, 80%	2
Utilization, ρ	0.80, 0.90	2
Planned lead time, L	3 periods, 5 periods	2

In total, there are 32 different combinations of experimental variables, and for each of them, we have a simulation run-length of 5460 periods. The duration that consists of the first 260 periods is used as the warm-up duration. Welch's procedure (see Law and Kelton (2000) for a complete description) is applied to approximate the warm-up duration for the output of each simulation. Each simulation with a given set of treatments is replicated 15 times using a different random number stream at each replication. Between different sets of replications the same set of random number streams are implemented. The experiments are performed using QUINTIQ 3.1.0.10 (see Quintiq (2007) for further detail) in simulating the production unit and the stock point, and CPLEX in solving the SCOP formulation.

4.3.3 Results

We are interested in two significant performance measures: the average periodic cost given that a target level of demand fill rate is satisfied, and the consistency of the planned schedule with the actual delivery of orders. The target fill rate is 98%. After each simulation run, with the safety stock equal to zero, the safety stock is adjusted according to the procedure described in the Appendix of Chapter 2, in a way to guarantee the target fill rate. Then, the simulation is repeated with the new safety stock.

In accordance with our objective in this chapter, the simulation results are interpreted by looking at the relative performance of different clearing functions. The lowest value of each performance criterion among different clearing functions is boldface, and pairwise comparisons are performed between different clearing functions keeping every other experimental factor fixed. The sign "†" refers to the absence of statistical difference between the results of different clearing functions with 95% confidence level.

Tables 4.2 and 4.3 provide the results related to the total average cost, respectively with the planned lead times of $L = 3$ and $L = 5$ periods. The abbreviations for the cost related performance measures are

TC = Total average cost per period.
I^+ = The average on-hand inventory at the stock point.
FW = The average finished WIP at the production unit.
W = The average WIP level.
SS = The safety stock level that satisfies 98% fill rate.

The total cost is calculated as follows:

$$TC = h_f I^+ + \hat{h}_f FW + h_w W.$$

Table 4.2: Cost performance of the clearing functions, $L = 3$ periods.

		$U_D = 40\%$				$U_D = 80\%$			
		CFL	*LTN*	*STN*	*TL*	*CFL*	*LTN*	*STN*	*TL*
$\rho = 0.80$	SS	41.0	**25.1**	46.7	94.1	70.7	**55.6**	76.7	142.8
	I^+	44.5†	**34.6**	45.4†	79.5	71.5†	**63.6**	72.4†	120.0
	FW	10.3	10.7	8.6	**8.0**	11.0	12.0	9.0	**8.4**
	W	15.1	23.5	5.3	**0.9**	17.4	33.3	5.9	**0.9**
	TC	83.0	79.5	**72.3**	109.8	120.0	127.3	**107.2**	161.0
$\rho = 0.90$	SS	97.3	**83.0**	103.0	348.6	155.1	**143.1**	163.3	453.3
	I^+	92.9†	**88.0**	93.5†	261.6	142.0†	**140.5**†	144.7†	327.9
	FW	11.2	12.2	10.0†	**9.9**†	11.6	13.3	10.2	**10.1**
	W	22.4	47.6	7.8	**1.2**	24.5	56.8	8.4	**1.2**
	TC	152.0	172.3	**136.6**	340.0	215.9	248.3	**201.5**	423.2

Table 4.3: Cost performance of the clearing functions, $L = 5$ periods.

		$U_D = 40\%$				$U_D = 80\%$			
		CFL	*LTN*	*STN*	*TL*	*CFL*	*LTN*	*STN*	*TL*
$\rho = 0.80$	SS	37.9	**25.3**	47.7	94.7	67.3	**57.1**	76.7	143.1
	I^+	46.8†	**35.2**	46.5†	79.4	75.3†	**66.0**	73.6†	120.6
	FW	10.7	10.1	8.7	**8.2**	11.7	10.9	9.3	**8.7**
	W	30.1	21.6	4.9	**0.9**	33.8	29.0	5.5	**0.9**
	TC	101.5	77.7	**73.5**	109.9	142.0	124.6	**108.6**	162.2
$\rho = 0.90$	SS	95.7	**85.6**	104.1	348.6	152.0	**144.1**	163.5	453.0
	I^+	96.3†	**90.9**	94.7†	260.7	146.0†	**142.7**†	146.2†	328.3
	FW	11.6	11.5	**10.1**	10.2	12.3	12.0	10.1†	**10.5**†
	W	44.2†	45.6†	7.4	**1.2**	47.8	53.5	8.0	**1.2**
	TC	178.5	173.1	**137.9**	339.3	245.1†	246.3†	**203.3**	424.1

These cost terms are more meaningful when considered together with the actual delivery performance of the released orders. It is usually difficult to explain the differences between different clearing functions if the delivery performance is ignored in the analysis. The results for the delivery performance under different clearing functions are provided in Tables 4.4 and 4.5 with the planned lead times of $L = 3$ and $L = 5$ periods respectively. The following abbreviations are used in these tables.

AF = The average actual order flow time.
CVF = The coefficient of variation in the flow times.
ΔL = The mean squared deviation of actual flow times from the planned lead time.
Π = The percentage of tardy orders.

Table 4.4: Delivery performance of the clearing functions, $L = 3$ periods.

		$U_D = 40\%$				$U_D = 80\%$			
		CFL	*LTN*	*STN*	*TL*	*CFL*	*LTN*	*STN*	*TL*
$\rho = 0.80$	AF	2.52	**2.29**	2.88	3.30	2.58	**2.28**	2.93	3.39
	CVF	0.30	0.34	0.20	**0.18**	0.31	0.36	0.21	**0.18**
	ΔL	0.80	1.12	**0.35**	0.44	0.80	1.20	**0.37**	0.51
	Π	4.87	**1.17**	9.16	35.72	7.62	**2.38**	12.18	43.09
$\rho = 0.90$	AF	2.64	**2.28**	2.92	3.57	2.76	**2.35**	2.98	3.62
	CVF	0.30	0.37	0.22	**0.16**	0.30	0.37	0.22	**0.15**
	ΔL	0.74	1.23	**0.43**	0.66	0.73	1.19	**0.43**	0.69
	Π	10.06	**3.48**	14.01	60.41	14.64	**5.24**	17.42	64.50

Table 4.5: Delivery performance of the clearing functions, $L = 5$ periods.

		$U_D = 40\%$				$U_D = 80\%$			
		CFL	*LTN*	*STN*	*TL*	*CFL*	*LTN*	*STN*	*TL*
$\rho = 0.80$	AF	**4.11**	4.26	4.88	5.37	**4.10**	4.24	4.89	5.47
	CVF	0.26	0.20	0.12	**0.12**	0.29	0.22	0.13	**0.12**
	ΔL	1.88	1.29	**0.38**	0.54	2.18	1.42	**0.43**	0.65
	Π	3.53	**1.00**	9.48	41.59	5.30	**1.75**	11.71	49.81
$\rho = 0.90$	AF	4.35	**4.23**	4.90	5.71	4.38	**4.27**	4.94	5.76
	CVF	0.24	0.21	0.14	**0.10**	0.26	0.23	0.14	**0.09**
	ΔL	1.47	1.42	**0.46**	0.82	1.65	1.48	**0.49**	0.87
	Π	8.32	**2.74**	13.98	72.11	10.52	**3.94**	16.44	76.76

The clearing function has a significant effect on both the external and the internal performance measures. About 55% improvement in the total periodic cost is possible by employing different clearing functions in the SCOP model, and the consistency of the planned schedule can be improved significantly. Recall from the previous sections that orders are released to the production unit and received by the stock point in batches of various sizes, and the demand is moderately variable. Thus one would expect that both ΔL and Π have significant impacts on *TC*.

The effect of *TL* especially on Π, and therefore on *SS* is very adverse. *TL* supports late production because it assumes that a high level of WIP can be cleared shortly. When the operational execution in the shop is not in line with this assumption, the number of tardy orders increases significantly, and the system has to hold a high *SS* to compensate for large backorders. As a result, we see a very high I^+, and consequently a high total cost. Especially under high utilization the effects are stronger. The average WIP level is the lowest under *TL* because, capacity loading is kept low due to the assumption of high productivity. The *TL* function implies that the amount of WIP to be loaded to the production unit at the start of a period is equal to the planned throughput increase in that period. Thus, random shortages in the production process significantly affect the actual delivery of released orders, and the stock point suffers from high backorders leading to a very high *SS*.

TL generates the longest average flow time in all cases, and it is greater than the planned lead time. On the other hand, the deviation from the planned schedule in terms of ΔL when *TL* is applied is smaller than ΔL of *CFL* or of *LTN*. However, this does not generate better cost performance due to the high number of late order deliveries. Under high utilization, over 60% of the orders are delivered late. So, the level of consistency should be considered as a joint effect of Π and ΔL.

From that point of view we can see a clear distinction between *STN* and *TL*, such that both Π and ΔL of *STN* are smaller than those of *TL*. As a result, *STN* outperforms *TL* with an improvement of 33% up to 62% in *TC*. *STN* assumes a lower productivity than *TL* does, and loads the capacity earlier resulting in decreased flow times. Having the flow times closer to the planned lead times and significantly decreased Π, *STN* provides a better coordination of the material flow between the production unit and the stock point than *TL* does.

CFL assumes a lower productivity than *STN* does, and provides lower Π. However, earlier than planned delivery of the production orders increases ΔL. Besides, the variation in the flow times is higher than that of *STN*. An appealing result is that the differences between I^+ of *CFL*, and I^+ of *STN* are insignificant in all cases. Although *STN* has a higher *SS* and Π, improved consistency and reliability in the planned lead times provide better planning of the final product inventory at the stock point. WIP levels can be decreased drastically, and a lower *TC* is achieved with *STN* in all cases.

Similar to *CFL*, *LTN* can be mainly characterized by the early delivery of the released orders. Since the utilization is kept at moderately high values (above 80%) *LTN* loads the shop early and in large quantities in order to achieve a certain planned throughput level. As a consequence, in most of the cases *LTN* provides the smallest *AF*. This results in the lowest *SS* and I^+ with *LTN* in all cases. However, with drastically decreased WIP levels, *STN* provides the lowest cost solution in all cases.

In addition, the performance of *LTN* is more sensitive to the environmental factors. *LTN* over-reacts to the increased uncertainty and the utilization causing increased *TC*. Under $\rho = 0.80$, (Table 4.2 and 4.3), I^+ of *LTN* almost doubles, and *W* of *LTN* increases by about 50% when the demand uncertainty increases from 40% to 80%. Under the same conditions, I^+ of *STN* increases by about 60%, and *W* of *STN* increases by about 10%. Similarly, when the utilization is increased from 0.80 to 0.90, I^+ of *LTN* increases by about 150%, and *W* of *LTN* increases by about 100%, while I^+ of *STN* increases by about 100%, and *W* of *STN* increases by about 50%. This relatively robust behavior of *STN* under changing environmental factors is due to the fact that *CVF* and ΔL of *STN* are always less than those of *LTN*. That is, both the variability and the unpredictability of the delivery schedules are less severe with *STN*. Thus, the system is less vulnerable to the environmental changes. For $\rho = 0.90$ and $U_D = 80\%$, both *STN* and *LTN* satisfy the same fill rate with insignificantly different I^+ levels, where, with its much lower WIP level, *STN* outperforms *LTN*. To sum up, *STN* ensures a better coordination between capacity loading and order release decisions. This results in improved cost performance especially under high demand uncertainty or high utilization levels.

Since the structure of *CFL* depends on the planned lead time, its performance differs significantly between $L = 3$ and $L = 5$. So, it becomes more crucial to choose the right planned lead time when *CFL* is deployed for the SCOP model. *CFL* with $L = 5$ provides I^+ and *FW* close to those of *CFL* with $L = 3$ but, it causes much higher WIP. It generates unnecessary build up of WIP that suggests us to model the production process using more optimistic clearing functions, i.e. *CFL* with $L = 3$ outperforms *CFL* with $L = 5$. However, very optimistic models such as *TL* (*CFL* with $L = 1$) also perform badly. We see that the *STN* clearing function provides a good reference point between the optimistic and the pessimistic production models. This can also be seen when one looks at the average order flow times; *AF* of *STN* is always larger than *AF* of *CFL* and smaller than *AF* of *TL*, and is closer to the planned lead time.

Since the throughput quantities per period are planned at the SCOP level according to the lead time, an increase in the lead time yields an increase in the actual order flow times. In addition, Tables 4.4 and 4.5 provide an interesting result that as the lead time increases then the production of the released orders are spread through a bigger number of periods. Thus, the coefficient of variation in the actual order flow times decreases due to the rolling horizons method applied at each replanning opportunity. In relation to the decrease in *CVF*, more orders are delivered early for a clearing function with a pessimistic approach such as *CFL* and *LTN*, and more orders are delivered late for a clearing function with an optimistic approach such as *TL*. In that respect, *STN* provides a relatively robust behavior such that ΔL and Π change very slightly under different planned lead times.

4.4 Conclusion

In this chapter, we have provided a planning framework for supply chain operations planning where the capacity loading decisions are decomposed from the order releases so that the throughput during the planned lead time can be determined to meet the planned delivery schedule. The throughput performance of the system is planned by modeling the production process through the clearing function. Our research question has been: what is the appropriate form of such a clearing function for the best performance of the SCOP model. We have shown that:

- The shape of the clearing function plays an important role in the completion time of orders. There is a tradeoff between loading the shop early with high WIP levels and forcing the system to deliver the orders earlier than planned and keeping low WIP in the shop but experiencing increased number of late deliveries.
- The consistency of the actual production schedule to the planned schedule improves the cost performance of the system.
- Deploying a production model based on the short-term probabilistic performance of the shop provides the best results in terms of the average periodic cost and robustness by achieving coordinated capacity loading and order release decisions.

Although, *STN* provides the best performance results in terms of the total cost and the lead time error, the tradeoff between *STN* and *LTN* with respect to the deviation of the actual flow times from the planned lead time and the percentage of orders that are tardy should be put into perspective. With *STN*, the squared deviation of the actual flow times from the planned lead time is the lowest while, the number of jobs that are tardy is the lowest with *LTN*. This tradeoff between tardy deliveries and consistent deliveries, especially considering general cost parameters, is an interesting subject for future research.

With *STN*, we would also expect to see an improved performance especially for multi-stage, multi-item production-distribution situations due to a better coordination of the material flow between the successive stages of the supply chain. The structure of the supply chain plays an important role. The performance evaluation for more complex structures is therefore an interesting subject for future research. The complexity can be further increased by considering detailed shop structures such as multi-resource flow shops or job shops. In these cases, modeling the clearing function becomes a technical challenge.

In addition to the performance related issues, especially through *STN*, more detailed discussions on modeling the clearing behavior can be initiated. For example, instead of a mean value approach, the approximated distribution characteristics of the throughput depending on the WIP level can be incorporated into the clearing function. In that respect, the *STN* clearing function can be elaborated in such a way that it relates the WIP level to a throughput level with a certain probability of occurrence.

In the next chapter, we will see how a clearing function may be used to update the planned lead times of a multi-stage serial production-inventory situation. Modeling the clearing function will be analyzed further based on a specific parameter that determines the piecewise-linear shape in detail.

Appendix to Chapter 4

Proof of Proposition 4.1

Let us define $\widehat{I}^*(\widetilde{dd}_i) = \widehat{I}(0) + \widetilde{dd}_i\mu - \sum_{j=1}^{i-1} q_j$. The schedule update heuristic partitions the set $\widetilde{X}$ into two disjoint subsets: the subset of orders that are rescheduled, $\widetilde{X}_{\text{late}}$, and the subset of orders that are not rescheduled, $\widetilde{X} \setminus \widetilde{X}_{\text{late}}$. We consider two cases:

Case 1: $(q_i, \widetilde{dd}_i) \in \widetilde{X}_{\text{late}}$ and $(q_{i+1}, \widetilde{dd}_{i+1}) \in \widetilde{X} \setminus \widetilde{X}_{\text{late}}$:

From the definition of the subset $\widetilde{X}_{\text{late}}$ the following conditions are determined: $\widehat{I}^*(\widetilde{dd}_i) < q_i$ and $\widehat{I}^*(\widetilde{dd}_{i+1}) \geq q_{i+1}$. Accordingly,

$$\begin{aligned}\widehat{I}^*(\widetilde{dd}_{i+1}) &= \widehat{I}^*(\widetilde{dd}_i) + (\widetilde{dd}_{i+1} - \widetilde{dd}_i)\mu - q_i \\ &\geq q_{i+1},\end{aligned}$$

which directly implies

$$\widehat{I}^*(\widetilde{dd}_i) + (\widetilde{dd}_{i+1} - \widetilde{dd}_i)\mu \geq q_i.$$

Then, the following relationship holds true due to Equation (4.11).

$$\widehat{dd}_i \leq \widetilde{dd}_{i+1} \tag{4.14}$$

Case 2: $(q_i, \widetilde{dd}_i) \in \widetilde{X}_{\text{late}}$ and $(q_{i+1}, \widetilde{dd}_{i+1}) \in \widetilde{X}_{\text{late}}$:

By the definition of the subset $\widetilde{X}_{\text{late}}$, $\widehat{I}^*(\widetilde{dd}_i) < q_i$. From Equation (4.11),

$$\widehat{I}(0) + \widehat{dd}_{i+1}\mu - \sum_{j=1}^{i} q_j \geq q_{i+1},$$

which implies

$$\widehat{I}^*(\widetilde{dd}_i) + (\widehat{dd}_{i+1} - \widetilde{dd}_i)\mu - q_i \geq q_{i+1}.$$

From Equation (4.11), this automatically satisfies the fact that

$$\widehat{dd}_i \leq \widehat{dd}_{i+1}, \tag{4.15}$$

For the case with $(q_i, \widetilde{dd}_i) \in \widetilde{X} \setminus \widetilde{X}_{\text{late}}$ and $(q_{i+1}, \widetilde{dd}_{i+1}) \in \widetilde{X}_{\text{late}}$, and the case with $(q_i, \widetilde{dd}_i) \in \widetilde{X} \setminus \widetilde{X}_{\text{late}}$ and $(q_{i+1}, \widetilde{dd}_{i+1}) \in \widetilde{X} \setminus \widetilde{X}_{\text{late}}$, Proposition 4.1 is satisfied directly from the definitions of $\widetilde{X}$ and $\widetilde{X}_{\text{late}}$.

The proof of our proposition about order crossovers is done by induction. From constraint sets (4.9) and (4.10) it is obvious that $\sum_{t=1}^{L} \widehat{Q}(t) \leq \widehat{I}(0) + L\mu$. Given that $\widetilde{X}_{\text{late}} = \emptyset$, this directly implies

$$\sum_{j=1}^{k} q_j \leq \widehat{I}(0) + L\mu. \tag{4.16}$$

During the current period the worst case throughput is zero, which causes, in the next replanning, some of the orders to become late. However, Equation (4.16) still holds, and from Equation (4.11), it is directly given that

$$\widehat{dd}_k \leq L. \tag{4.17}$$

The production unit is assumed to be initially empty, yielding $\widetilde{X}_{\text{late}} = \emptyset$ at the start of the simulation. Therefore, Condition (4.17) holds true during the rest of the simulation. Together with Conditions (4.14) and (4.15), this completes the proof of Proposition 4.1.

Chapter 5

An Effective Approach for Updating Lead Times

In this chapter, we provide insights into the effectiveness of updating the lead times of a supply chain in a hierarchical planning context using the clearing function concept. A two-stage serial supply chain is considered with each stage responsible to produce a single item. Orders are released by a SCOP model such that delivery schedules are determined through planned lead times. Capacity loading decisions are separated from order release decisions, and depend on the hierarchical coupling mechanism. The planning system is implemented in a rolling horizon setting such that the lead times are updated according to the current workload status and the anticipated future production requirements. Simulation experiments are performed for different demand conditions under changing clearing structure and hierarchical coupling. The results indicate that, in conjunction with the concept of clearing, updating the lead times provides the flexibility under fluctuating demand conditions, and generates less costly solutions. At the same time, the nervousness in the planning system due to the dynamic planned lead times is kept under control, and the lead time syndrome is avoided.

5.1 Introduction

In Chapters 2 and 3, it was shown from various perspectives that updating the lead times in planning and coordinating the flow of materials in a supply chain is a challenging task. Naive approaches based on exponential smoothing of realized order flow times, or simple workload dependent rules such as TWK and JIQ do not work for stationary situations with considerable uncertainty in the environment. In a hierarchical planning system employed in a dynamic setting, planned lead times that are updated at a higher-level and used to release the orders at a lower-level may cause erratic order releases and increased congestion in the production unit. This phenomenon, identified as the *lead time syndrome*, suggests the use of fixed planned lead times both in commercial advanced planning systems and in theoretical studies to manage production and plan inventory in a supply chain (e.g., Stadtler and Kilger (2000), Spitter *et al.* (2005a), and De Kok and Fransoo (2003)). On the other hand, "a major drawback of a fixed planned lead time is the ignorance of the correlation between actual workloads and the flow times that can be realized under a limited capacity flexibility" (Zijm and Buitenhek (1996)).

In this chapter, our objective is to challenge the fixed lead time assumption through developing a lead time update procedure that provides the flexibility to respond to structural changes in environmental conditions (e.g., non-stationary demand with seasonal fluctuations), and increases the efficiency of the material flow decisions among the supply chain. We aim to provide an answer to our second research question by also considering the issues raised by the fourth and the fifth research questions in the analysis. In particular, we want to show an effective way of updating planned lead times of a supply chain. In doing so, we want to provide performance evaluation results based on (1) the coupling mechanism between different decision levels of the planning hierarchy, (2) the detailed approach to anticipate throughput performance of the production units, and (3) the demand uncertainty.

The clearing function concept is used as the basis for updating the lead times and for controlling the workload at various stages in the supply chain. As discussed in the previous chapter, the clearing function provides an abstract representation of the relationship between the expected throughput during a period and the WIP available in that period. Our lead time update procedure basically associates throughput ranges for different lead times, and matches them with the anticipated future production requirements in the planning horizon. The idea is to have the planned lead times short during the periods of low demand, and long during the periods of high demand by explicitly considering the periods at which stock-outs are expected to occur. The low and high demand periods may both occur due to the stochastic behavior of the production processes and due to forecast errors. For example, production being faster than planned or actual demand being lower than forecasted demand decrease the production requirements for the next periods, and vice versa. In addition, new forecasted demand is included at the end of the planning horizon. Thus, both the demand seasonality and the uncertainty in the system provide motivation to vary the planned lead times.

In addition, we propose a detailed method for modeling the clearing function using a piecewise-linear representation. Such a detailed method is absent in the previous studies on clearing functions (e.g., Missbauer (2002), Asmundsson *et al.* (2003), Asmundsson *et al.* (2004), Armbruster *et al.* (2004), Hwang and Uzsoy (2005) and Asmundsson *et al.* (2006)). We identify a parameter to indicate a service measure for the production unit's throughput performance, and specify the clearing function based on that parameter. In general, the clearing of WIP is based on the short-term behavior of the production processes, and design choices on specific shape related issues are considered in relation to the dynamic lead times.

The planning process takes place in a hierarchical setting. The orders are released throughout the complete supply chain in a centralized approach according to the given planned lead times (set at the higher level) and the demand forecasts. At the lower planning level, detailed schedules and capacity loading decisions are determined for each production unit separately. In determining how the capacity is going to be loaded, the coupling mechanism between the SCOP and the operational scheduling levels plays a significant role. It indicates the level of aggregation considered at the SCOP level, and raises the question to which extent outcomes of the SCOP model should influence detailed capacity loading decisions. In De Kok and Fransoo (2003), it is mentioned that both order release and periodic production decisions are instructed to a lower level for execution in a purely deterministic environment. In this chapter, through simulation, we extend their analysis and provide insights into the effects of loose and tight coupling in a dynamic framework with stochastic demand and production processes.

A two-stage serial supply chain facing a stochastic final product demand is considered. The demand level changes dynamically. Each stage consists of a single production unit with stochastic processing times and a stock point to keep the finished items for downstream demand satisfaction (see Figure 1.1 for an illustration). We assume single item production at each stage, and there is an ample supply of raw materials at the upstream stage of the supply chain. The order release and capacity loading decisions are given periodically. Time-phased order releases are planned to meet future forecasted demand, and capacity is loaded with material to meet planned schedule of the released orders. In doing so, the clearing function concept plays an important role in anticipating the throughput performance of the production units.

5.2 Clearing Function

The periodic throughput quantity P_w is a positive random variable and depends on the available WIP level w. The clearing function, $f(w) = E\left(P_w\right)$, formulates the expected throughput in a period as a function of w. Let us define P_∞ as the random variable for the amount of throughput in a period under the assumption that the shop is loaded with an infinitely large quantity of WIP. We assume the information on the distribution characteristics of P_∞ is given. Say $g(\cdot)$ and $G(\cdot)$ respectively refer to the probability density function and the cumulative distribution function of P_∞, and $E\left(P_\infty\right) = \mu$ refers to the nominal throughput rate. Then, it follows from standard probability theory that $\Pr\left\{P_w = w\right\} = 1 - G(w)$. From the general representation of the *STN* clearing function in Equation (4.4), the clearing function we use in this chapter is written by

$$f(w) = \sum_{k=1}^{w} \left(1 - G(k)\right). \tag{5.1}$$

Theoretically, given that there does not exist a finite k such that $G(k) = 1$, a nonlinear clearing function as given in Equation (5.1) never reaches the nominal rate. In this chapter, as it has been done in Chapter 4, a finite WIP level for the nominal throughput rate is provided for experimental and practical purposes by rounding the value of the clearing function from the second decimal digit. Let us define $w^{(\mu)}$ as the WIP level where the clearing function approximately reaches the nominal rate μ.

$$w^{(\mu)} = \min\left\{w : f(w) \approx \mu\right\}.$$

The point $\left(w^{(\mu)},\, \mu\right)$ becomes one of the breakpoints in a piecewise-linear approximation of the clearing function.

Finding a piecewise-linear approximation of the clearing function is an interesting topic that requires a detailed discussion. Because, when implementing in a linear programming formulation, the clearing function is generally transformed into its piecewise-linear approximation (cf. Missbauer (2002), and Asmundsson *et al.* (2003)) as if it is a standard procedure, and alternative detailed aspects of this transformation have been ignored. In this chapter, we are concerned with modeling a piecewise-linear representation of the clearing function from a design perspective. We expect that the design choices in this approximation are very relevant in terms of determining the performance of the planning system. As a means of incorporating this aspect of the clearing function to our model, let us define a parameter ε such that $0 \leq \varepsilon \leq 1$. Associated with ε, a certain WIP level

$$w_\varepsilon = G^{-1}(\varepsilon)$$

is defined, implying that ε provides a service level measure for the clearing of the available WIP in the production unit. Given that the production unit is loaded by a WIP level $w \leq w_\varepsilon$, then the production unit is able to clear all the available WIP in the current period by at least $1-\varepsilon$ probability. A 2-slope approximation of the clearing function is derived with breakpoints at (0, 0), $(w_\varepsilon,\, w_\varepsilon)$, and $\left(w^{(\mu)},\, \mu\right)$. Accordingly, w_ε can be defined as the anticipated level of WIP above which the production unit is considered to operate with less than 100% productivity. By productivity we mean the level of increase in the expected throughput level relative to one unit of increase in the available WIP level. 100% productivity refers to a one-to-one correspondence. According to these definitions, the piecewise-linear approximation of the clearing function can be formulated as follows:

$$f(w) \approx \begin{cases} w, & w \leq w_\varepsilon \\ w_\varepsilon + \frac{\mu - w_\varepsilon}{w^{(\mu)} - w_\varepsilon} \cdot (w - w_\varepsilon), & w_\varepsilon < w \leq w^{(\mu)} \\ \mu, & w^{(\mu)} \leq w \end{cases} \tag{5.2}$$

Figure 5.1 illustrates an example of a nonlinear, concave clearing function together with its piecewise-linear approximations for different ε values. It is seen that ε is a means of modeling different approaches in approximating the clearing function.

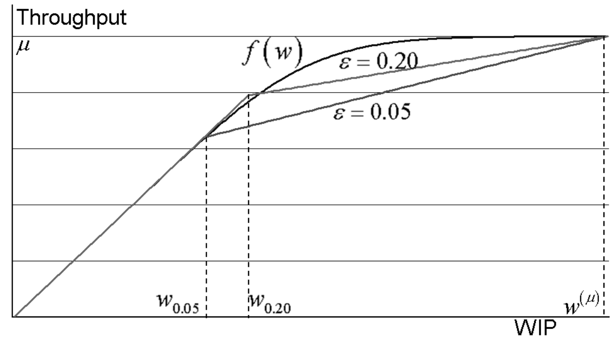


Figure 5.1: A nonlinear clearing function, and its piecewise-linear approximations for $\varepsilon = 0.05, 0.20$.

For small ε a tight service is required from the production unit, and for large ε a relaxed service is allowed. As ε is getting larger, a more optimistic approach is applied in modeling the production process such that the production unit is assumed to be capable of finishing a (relatively) large amount of WIP in a single period. As ε becomes smaller, a more conservative approach is implemented against the uncertainties in the production.

In the rest of this chapter, the term clearing function refers to the piecewise-linear representation. For brevity, we consider 2-slope representations like the ones in Figure 5.1. For STN clearing functions, a 2-slope representation is a reasonable approximation because, generally, the nominal throughput rate is achieved before the WIP level reaches to 2μ. For such cases, the piecewise-linear approximation given in Equation (4.5) becomes a special case (with a certain ε) of the one in Equation (5.2). One may need more slopes if the production process is highly variable, such that P_∞ has a coefficient of variation bigger than one.

5.3 Planning System

We decompose the planning system into three hierarchically coordinated decision levels: the tactical planning level, the SCOP level, and the operational scheduling level. The tactical level is responsible for setting the planned lead times. The planned lead times can be static (the lead times are not subject to change) or dynamic depending on the current workload status of the production units and the expected future production requirements. Clearing function for each production unit is used to determine the throughput ranges for different planned lead times. At the SCOP level a centralized approach is applied, where planning decisions on the flow of materials between consecutive production units and the stock points in the supply chain are made. In doing so, forecasted information on the final product demand levels, status information about the workloads of production units and stock levels are utilized, in addition to the planned lead times instructed by the tactical level. The decisions made at this level are periodic order release quantities and the production levels, which are determined based on the clearing function for each production unit. The method of rolling horizons is applied, where only the first period's decisions are given to the scheduling level. Given the order releases, their planned lead times, and the target throughput quantity for the current period, the planning is done in a decentralized manner at the operational scheduling level considering each production unit separately. The delivery schedules are determined according to the given planned lead times. The capacity loading levels for the current period are determined through the clearing function and the coupling mechanism between the SCOP and the operational scheduling levels. It should be noted that the same clearing function for a production unit is used at all levels of the planning hierarchy. Figure 5.2 illustrates the planning framework with flows of information in between the decision levels.

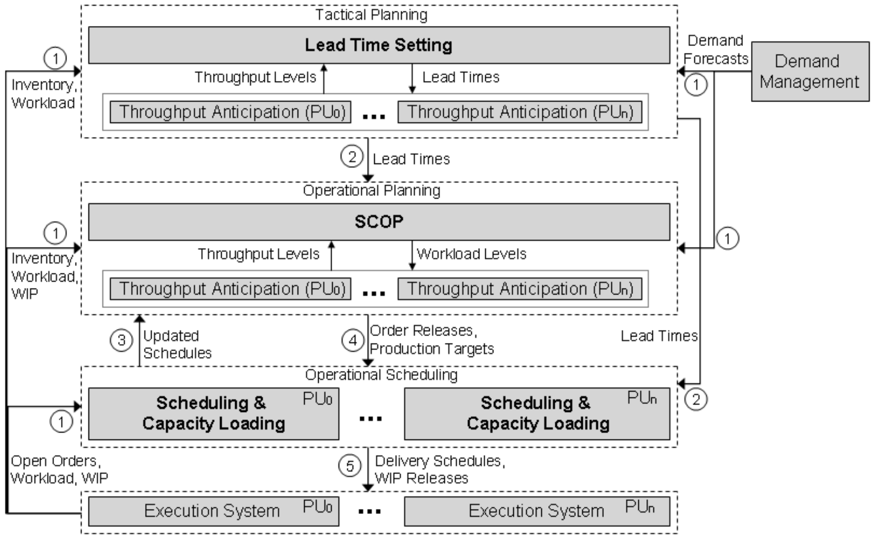


Figure 5.2: Hierarchical planning framework.

The circled numbers refer to the sequence of the information flow in the planning system. Data set (1) refers to the input data exogenous to the planning system, and includes exogenous demand information and the current status information of the production units and the stock points. In accordance with the hierarchical approach, a sequential decision making process starts from the upmost level. Data set (2) refers to the planned lead times set by the tactical level and instructed to the lower levels. Data set (3) provides the delivery schedule adjusted by the scheduling level in response to the capacity restrictions. The SCOP model provides the order releases and the production targets of each production unit for the current period in data set (4). Finally, data set (5) refers to the executable outcome of the planning system including the planned delivery schedules and the capacity loading levels for each production unit.

The index variables and the static parameters are determined a priori and do not change in time. They are

N = Set of items produced in the supply chain.
N_e = Set of final products, $N_e \subset N$
$i,\, j$ = Item indexes, $i,\, j = 1, \ldots, n$.
a_{ij} = Number of units of item i needed to produce one unit of item j.
μ_i = Nominal production rate for item $i \in N$.
$f_i(\cdot)$ = Clearing function for item $i \in N$.
T = Forecast horizon for the final products.
t = Period index, $t = 0, \ldots, T-1$.

The dynamic inputs are time-varying and depend on the previous decisions as well as on the uncertainty in the production and the demand processes. The current period is denoted as $t = 0$. The dynamic inputs to the planning system are

$D_i(t)$ = Forecast demand of the final product $i \in N_e$ for period t, $t = 0, \ldots, T-1$.
$I_i(0)$ = Current net inventory level, on-hand inventory minus backorders, of item $i \in N$.
$IP_i(0)$ = Current inventory position, net inventory plus pipeline inventory, of item $i \in N$.
$\widehat{W}_i(0)$ = Current workload level at the production unit of item $i \in N$. The total quantity of item i that has been ordered, but not yet produced.
$\widehat{I}_i(0)$ = Amount of finished item $i \in N$ currently waiting in the production unit, $\widehat{I}_i(0) = IP_i(0) - I_i(0) - \widehat{W}_i(0)$.

These static and the dynamic inputs do not change between different decision levels of the planning hierarchy. These are the constants and the variables that are part of the spatial and the temporal characteristics of the physical system, and are not considered as part of an aggregation-disaggregation scheme in the planning system.

The schedule for the open orders is not considered in the above list because, the same schedule update procedure as in Chapter 4 has been utilized in this chapter. Thus, discussion on updating the schedules has been omitted here. In the following subsections, detailed models for the decision levels of the hierarchy are provided. The inputs and the decision variables specific to each level are defined separately in each subsection.

5.3.1 Lead Time Setting

Lead times are determined based on the anticipated operational dynamics and the future production requirements of each production unit separately. First of all, let us denote L_i as the decision variable for the planned lead time of item $i \in N$. As shown in the planning hierarchy of Figure 5.2, the lead times are exogenous to the SCOP problem. This is in line with the discussion about planned lead times in De Kok and Fransoo (2003). Given the fact that flow times are related to resource utilization, De Kok and Fransoo (2003) point out that the actual choice of L_i should be consistent with the resource availability and resource requirements that can be derived from the *Bill-of-Processes* (BOP) and the independent demand characteristics. The procedure for updating the lead times presented in this section is a formalization of this idea in a dynamic framework.

In releasing the orders, a lead time dependent workload control rule is applied for each production unit. The total workload of a production unit cannot exceed the maximum level that it can produce within the given planned lead time. Here, the workload is measured in terms of the number of items released but not yet produced. Given the lead time $L_i = l$, this workload limit is defined as $\vartheta_i(l)$, and yields the concept of *lead time dependent throughput rate* formulated by the clearing function,

$$\mu_i(l) = f_i\left(\vartheta_i(l)\right) = \vartheta_i(l)/l.$$

This concept is one of the fundamental constructs that we use in setting the lead times dynamically.

In addition, the latest stock-out period is defined for each planned lead time of a certain item. It is the latest period within the planning horizon, T, at which a stock-out for that item is anticipated given the current net stock, demand forecasts and the planned lead time. It provides the time range within the planning horizon during which the inventory of that item is prone to stock-outs due to cumulative throughput (based on $\mu_i(l)$) lacking behind cumulative requirements. An illustration of the latest stock-out period is given in Figure 5.3, where it is expected to occur at period 6. There may be stock-out or non-stock-out periods before the latest stock-out period as seen in Figure 5.3. The latest stock-out period reveals the duration of imbalance between the cumulative production and the cumulative requirement for an item. For algorithmic purposes, if no stock-out is expected to occur during the planning horizon, the latest stock-out period is set to zero.

The objective of the lead time setting procedure is to decide on a planned lead time for each item that minimizes the backorder level at the minimum latest stock-out period within the planning horizon. The logic behind this procedure stems from the fact that the earlier is the latest stock-out period, the smaller is the imbalance between supply and demand. Given that the stock-out cannot be avoided, then it is best to have as low backorders as possible. In a capacitated production-inventory situation, anticipated future stock-outs can be avoided by increasing the throughput levels. On the other

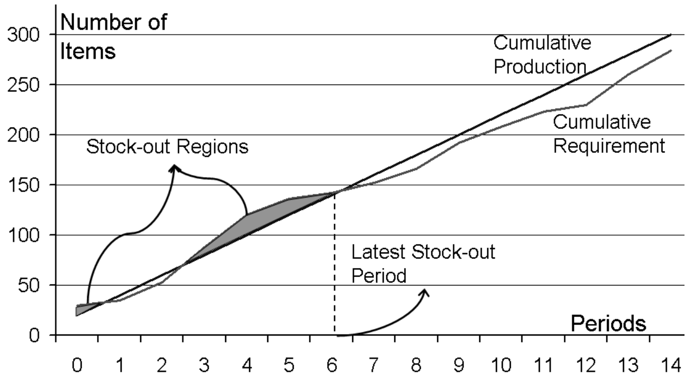


Figure 5.3: An example for the latest stock-out period of an item.

hand, the higher the throughput is, the higher the workload level is. Thus, it is favorable, in terms of decreasing the material holding costs, to hold low workload in the production units and produce less in case of no backorders and excess inventories.

Backorders for the final products are clearly defined. However, there is no record of backorders for the intermediate items. In this chapter, similar to our analysis of the DRP model in Chapter 2, a backorder for an intermediate item can be defined as the quantity demanded by a downstream production unit but cannot be satisfied and hence, carried onto the next periods' requirements hidden within the SCOP model. A similar argument has been raised in analyzing MRP systems by Buzacott *et al.* (1992).

The lead time setting procedure utilizes the tradeoff between short and long planned lead times that can be observed in a batch production system under limited capacity. Increasing the lead times may help avoid future stock-outs through increasing the workload levels and thus increasing the throughput levels but at the same time, it causes postponement in the delivery of the next order to be released, which may increase backorders in the early periods of the planning horizon. This relationship is formulated during the rest of this section considering the dependency between the consecutive stages in the supply chain together with a lower limit on the planned lead times depending on the current workload status at each production unit.

Based on $\widehat{W_i}(0)$ the minimum acceptable lead time for item i is defined because, the lead time should be long enough to clear the current backlog of orders on time, and the lead time is always greater than or equal to one period. The lower bound for the lead time of item i is denoted as $L_i^{\min}$, and is given by

$$L_i^{\min} = \min\left\{l; \widehat{W}_i(0) \leq \vartheta_i(l)\right\}.$$

Let $\widehat{D}_i(t)$ denote the net requirement of item i in period t. It is the quantity of item i required to be available at the start of period t that cannot be satisfied from the initial available stock of finished items. Since the effects of the past occurrences have already been included in the current state information, we assume that $\widehat{D}_i(-1) = 0$. For $i \in N_e$, the net requirements are determined through independent demand forecasts as follows:

$$\widehat{D}_i(t) = \begin{cases} \max\left\{0, \sum_{k=0}^{t} D_i(k) - \left(IP_i(0) - \widehat{W}_i(0)\right)\right\}, & \text{if} \quad \widehat{D}_i(t-1) = 0 \\ D_i(t), & \text{otherwise} \end{cases} \tag{5.3}$$

For the intermediate items, $i \in N \backslash N_e$, the production requirements are driven by the orders released by the downstream stages. As a result, their planning horizons change due to the changing planned lead times of their downstream items. If component i is only consumed for the production of item j, then $T_j - L_j - 1$ is the last period where the dependent demand information for item i is available because, the order releases of item j further into the future are not visible yet. As a result, the planning horizons are defined by

$$T_i = \begin{cases} \max_{j \in N}\left\{T_j - L_j;\, a_{ij} > 0\right\}, & i \in N \backslash N_e \\ T, & i \in N_e \end{cases}$$

The logic in computing T_i's is motivated by the time-phased order point approach in DRP systems. The release time of an order for a specific item is offset according to the planned lead time of that item, and that order is accounted for the gross requirements of its intermediate items at that release period. Exploiting the DRP approach further, the lead times in successive stages are determined sequentially, starting from the final products and then proceeding to the intermediate items.

Let us denote $t_{s,i}(l)$ as the latest stock-out period under the lead time l for item i. It is determined through the current inventory position and the lead time dependent throughput rate of item i.

$$t_{s,i}(l) = \max\left\{\begin{array}{c} \max\left\{t;\, \widehat{W}_i(0) < \sum_{k=0}^{t} \widehat{D}_i(k),\, t = L_i^{\min}, \ldots, l-1\right\}, \\ \max\left\{t;\, t \cdot \mu_i(l) < \sum_{k=0}^{t} \widehat{D}_i(k),\, t = l, \ldots, T_i - 1\right\} \end{array}\right\} \tag{5.4}$$

As we pointed out previously, in case no stock-out is detected within the planning horizon, then $t_{s,i}(l) = 0$. The first term within the outer brackets in Equation (5.4) considers the anticipated stock-outs before the lead time. It is a consequence of the previously occurred random deviations in demand forecasts or throughput quantities. The current workload, $\widehat{W}_i(0)$, is planned to be processed by the start of period $L_i^{\min}$. If $\widehat{W}_i(0) < \sum_{k=0}^{L_i^{\min}} \widehat{D}_i(k)$, then a stock-out at the end of period $L_i^{\min}$ is anticipated. There are no deliveries scheduled between the end of period $L_i^{\min}$ and the start of period $l-1$. Therefore, if there is a stock-out planned before the lead time, the latest stock-out occurs at the end of period $l-1$. The second term within the outer brackets in Equation (5.4) considers the anticipated stock-outs after the lead time. It is a consequence of the cumulative planned throughput lacking behind the cumulative net requirements. Let us formulate the minimum latest stock-out period for item i as

$$t_{s,i}^* = \min_{L_i^{\min} \leq l} \{t_{s,i}(l)\},$$

and the set of lead times that yields $t_{s,i}^*$ as the latest stock-out period is given by $S_i^* = \left\{l;\, t_{s,i}(l) = t_{s,i}^*\right\}$.

Let us denote $b_i\left(t_{s,i}^*, l\right)$ as the backorder level of item i at the end of period $t_{s,i}^*$ under lead time l. Then, L_i is given by

$$L_i = \arg\min_{l \in S_i^*} \left\{b_i\left(t_{s,i}^*, l\right)\right\}, \tag{5.5}$$

Given that $t_{s,i}^* \geq 1$,

$$b_i\left(t_{s,i}^*, l\right) = \begin{cases} \sum_{k=0}^{t_{s,i}^*} \widehat{D}_i(k) - \widehat{W}_i(0), & \text{if} \quad t_{s,i}^* < l \\ \sum_{k=0}^{t_{s,i}^*} \widehat{D}_i(k) - t_{s,i}^* \cdot \mu_i(l), & \text{otherwise} \end{cases} \tag{5.6}$$

and for $t_{s,i}^* = 0$, $b_i\left(t_{s,i}^*, l\right) = 0$. It should be noted that the backorders in Equation 5.6 are computed according to the throughput levels and not according to time-phased order releases. The ties in Equation (5.5) are broken by setting $L_i = \min\{l;\, l \in S_i^*\}$. In such a case, operating with shorter planned lead times decreases the total workload and the inventory in the system.

The set S_i^* can be decomposed into two mutually exclusive subsets. These are $S_i^*\left(> t_{s,i}^*\right) = \left\{l;\, l > t_{s,i}^*\right\}$ and $S_i^*\left(\leq t_{s,i}^*\right) = \left\{l;\, l \leq t_{s,i}^*\right\}$. This decomposition can be used to characterize L_i as defined in Equation (5.5). Using the definition for the latest stock-out period and the backorder formulation in Equation (5.6), we can state the following:

$$L_i = \begin{cases} \min\{l;\, l \in S_i^*\}, & \text{if} \quad S_i^*\left(\leq t_{s,i}^*\right) = \emptyset \\ \max\{l;\, l \in S_i^*\}, & \text{if} \quad S_i^*\left(> t_{s,i}^*\right) = \emptyset \\ \max\left\{l;\, l \in S_i^*\left(\leq t_{s,i}^*\right)\right\}, & \text{if} \quad S_i^*\left(\leq t_{s,i}^*\right) \neq \emptyset \,\text{and}\, S_i^*\left(> t_{s,i}^*\right) \neq \emptyset \end{cases}$$

The first property implies the case that the latest stock-out is expected to occur at a period earlier than the planned lead times. Then the backorder level at the end of the latest stock-out period cannot be influenced by changing the planned lead time. Ties occur in solving Equation (5.5), which are broken by choosing the minimum lead time in the set. The second property emphasizes the other side of the coin. If the latest stock-out is expected to happen later than the lead times, then the backorder level can be minimized by operating with longer lead times and thus larger throughput rates. The last property stems from the fact that for $t \geq l$ and $l \geq L_i^{\min}$, $t \cdot \mu_i(l) \geq \widehat{W}_i(0)$ and from the concavity of the clearing function.

In a multi-stage production-distribution situation, lead times are first determined for the most downstream stage. The planned lead time for an item provides information about the expected order releases for that item within the current planning horizon. This information is utilized to determine the lead time of an intermediate item at an immediate upstream stage in the supply chain.

Let us denote $R_j(t)$ as the expected order release quantity for item $j \in N_e$ in period t, $t = 0, \ldots, T_j - L_j - 1$. $R_j(t)$ depends on the latest stock-out period for item j such that

if $t_{s,j}^* < L_j$, then

$$\begin{aligned} R_j(0) &= \sum_{k=0}^{L_j} \widehat{D}_j(k) - \widehat{W}_j(0), \\ R_j(t) &= \widehat{D}_j(t + L_j),\ t = 1, \ldots, T_j - L_j - 1, \end{aligned}$$

if $t_{s,j}^* \geq L_j$, then

$$\begin{aligned} R_j(0) &= \vartheta_j\left(L_j\right) - \widehat{W}_j(0), \\ R_j(t) &= \mu_j\left(L_j\right),\ t = 1, \ldots, t_{s,j}^* - L_j, \\ R_j\left(t_{s,j}^* - L_j + 1\right) &= \widehat{D}_j\left(t_{s,j}^* + 1\right) + b_j\left(t_{s,j}^*,\, L_j\right), \\ R_j(t) &= \widehat{D}_j(t + L_j),\ t = t_{s,j}^* - L_j + 2, \ldots, T_j - L_j - 1, \end{aligned}$$

Then, for each $i \in N \backslash N_e$, the net requirements per period are computed by having $D_i(t) = \sum_{j \in N_e} a_{ij} R_j(t)$ in Equation (5.3), and their planned lead times are determined accordingly.

An initial condition that has to be satisfied in determining the planned lead times is that the planning horizon T must always be greater than the cumulative supply chain lead time. An upper bound for the planned lead times may be necessary to decide on a feasible length of the planning horizon. The lead time setting algorithm in Equation (5.5) yields an upper bound for the planned lead times depending on the clearing function. Given that the nominal throughput rate is reached, further increasing the lead time does not yield an improvement in the periodic throughput but increases the backorder levels due to postponed deliveries, and also the workload levels are increased due to early release of orders. The upper bound for the lead time is given by

$$L_i^{\max} = \min \left\{ l; \mu_i(l) = \mu_i \right\}.$$

5.3.2 SCOP

In the SCOP model, inventory levels are planned to minimize total material holding and penalty costs subject to capacity and workload constraints. The costs are incurred according to the amount of material waiting in the production units or in the stock points of the supply chain. In addition, penalty costs for the final product shortages are incurred, which are set so high that the system always targets a nonnegative net inventory level for these items. This is also equivalent to operating under zero safety stock. The cost coefficients are described as follows:

$h_i^{(f)}$	=	Per unit, per period cost of holding finished item $i \in N$ either in the stock point or in the production unit.
$h_i^{(w)}$	=	Per unit, per period cost of holding unfinished item $i \in N$ either being processed in the shop or waiting to be processed in front of the shop.
M_i	=	Unit penalty cost for the final product inventory shortage, $i \in N_e$.

The schedule for the previously released orders is feedforward by the scheduling system to the SCOP model. $\widehat{Q}_i(t)$ indicates the total quantity of item i previously released and scheduled to be received at its stock point at the start of period t. The decision variables for the stock and the workload levels in the SCOP model are

$I_i^+(t)$ = Inventory on-hand of item $i \in N$ at the start of period t, just before the order scheduled for period t is received, $t = 1, \ldots, T_i$.

$I_i^-(t)$ = Backorder level of final product $i \in N_e$ at the start of period t, just before the order scheduled for period t is received, $t = 1, \ldots, T_i$.

$\widehat{W}_i(t)$ = Total workload level of item $i \in N$ at the start of period t, just before the release of orders, $t = 1, \ldots, T_i - 1$.

$\widehat{I}_i^+(t)$ = Level of finished item $i \in N$ at the start of period t waiting in the production unit to be sent to its stock point, $t = 1, \ldots, T_i - 1$.

The decision variables for the flow of material between successive stages and for the utilization of each production unit are

$Q_i(t)$ = The size of the order to be released for item $i \in N$ at the start of period t, $t = 0, \ldots, T_i - L_i - 1$.

$P_i(t)$ = Planned production quantity of item $i \in N$ in period t, $t = 0, \ldots, T_i - 2$.

All the variables and the parameters defined above are nonnegative. The SCOP problem is modeled by the following linear programming formulation:

$$\text{Min.} \quad \sum_{i \in N} \sum_{t=1}^{T_i} \left[h_i^{(f)} \cdot I_i^+(t) + M_i \cdot I_i^-(t) \right] + \sum_{i \in N} \sum_{t=1}^{T_i - 1} \left[h_i^{(w)} \cdot \widehat{W}_i(t) + h_i^{(f)} \cdot \widehat{I}_i^+(t) \right]$$

subject to

$$I_i^+(t+1) - I_i^-(t+1) = I_i(0) + \sum_{k=L_i}^{t} Q_i(k - L_i) + \sum_{k=1}^{t} \widehat{Q}_i(k) - \sum_{k=0}^{t} D_i(k) - \sum_{k=0}^{t} a_{ij} Q_j(k), \quad i \in N,\, t = 0, \ldots, T_i - 1 \tag{5.7}$$

$$\widehat{W}_i(t+1) = \widehat{W}_i(0) + \sum_{k=0}^{t} Q_i(k) - \sum_{k=0}^{t} P_i(k), \quad i \in N,\, t = 0, \ldots, T_i - 2 \tag{5.8}$$

$$\widehat{W}_i(0) + \sum_{k=0}^{t} Q_i(k) \leq \sum_{k=0}^{t+L_i-1} P_i(k), \quad i \in N,\, t = 0, \ldots, T_i - L_i - 1 \tag{5.9}$$

$$P_i(t) \leq f_i \left(\widehat{W}_i(t) + Q_i(t) \right), \quad i \in N,\, t = 0, \ldots, T_i - L_i - 1 \tag{5.10}$$

$$P_i(t) \leq \mu_i \left(L_i \right), \quad i \in N,\, t = T_i - L_i, \ldots, T_i - 2 \tag{5.11}$$

$$\widehat{I}_i^+(t+1) = \widehat{I}_i(0) + \sum_{k=0}^{t} P_i(k) - \sum_{k=L_i}^{t+1} Q_i(k - L_i) - \sum_{k=1}^{t+1} \widehat{Q}_i(k), \quad i \in N,\, t = 0, \ldots, T_i - 2 \tag{5.12}$$

Constraint set (5.7) stipulates that for each item, the net inventory at the start of period $t+1$ is equal to the current net inventory plus the cumulative delivered orders by period t minus the cumulative exogenous and endogenous demand up to and including period t. Constraint set (5.8) indicates that the total workload at a production unit increases with released orders and decreases with produced items. Constraint set (5.9) provides the workload limitation for each production unit according to the planned lead times. The total workload at the start of a period in a production unit cannot be greater than the total amount that the production unit can clear within the given planned lead time. Constraint set (5.10) stipulates that the production process is modeled through a piecewise-linear and concave clearing function. Since the order releases in periods beyond $T_i - L_i - 1$ are not visible yet at $t = 0$, the planned production quantities are limited by $\mu_i\,(L_i)$ in these periods. This is given by constraint set (5.11). Constraint set (5.12) indicates that the finished items have to wait in the production unit before the whole released order has been processed and sent to its stock point.

Due to inherent uncertainties in the production processes, a certain amount of workload may be carried over the next replanning. In addition, material shortages may occur causing insufficient loading of the production unit and a decline in the throughput levels. This yields the constraint set (5.12) to become infeasible. In that case, the SCOP formulation is relaxed by replacing $\widehat{I}_i^{+}(t{+}1)$ on the left-hand-side of the constraint set (5.12) by $\widehat{I}_i^{+}(t+1) - \widehat{I}_i^{-}(t+1)$, $\widehat{I}_i^{-}(t+1) \geq 0$. In addition, the term $\sum_{i\in N}\sum_{t=1}^{T-1} M_i \cdot \widehat{I}_i^{-}(t)$ is added to the objective function in order to limit the situation to be temporary. Thus, at the SCOP level, infeasible material flows are temporarily allowed till the feasibility is reaffirmed in the next replanning opportunities. We expect this adjustment does not significantly affect the simulation results because, the system strives to return to feasible conditions as soon as possible due to very large penalty costs in the objective function.

For each planning cycle in a rolling horizon, only the first period's planning decisions, $Q_i(0)$ and $P_i(0)$ for every $i \in N$, are given to the operational scheduling level. At this level, the planning is done in a decentralized manner separately for each production unit.

5.3.3 Capacity Loading

The finalized production orders are processed according to the FCFS discipline at each production unit. The operational scheduling level is responsible for capacity loading decisions (the amount of WIP that is going to be loaded to the shop) and the planned delivery schedule according to the given planned lead times and the capacity restrictions. The rescheduling for the released orders of each production unit is done in the same manner as described in Section 4.2.3 with the term μ replaced by the term $\mu_i\left(L_i\right)$.

Let us define $W_i(0)$ as the current WIP for item i in its production unit, and $V_i(0)$ as the decision variable for the amount of additional WIP to be released to the production unit in the current period for the production of item i. Only the first planning period's capacity loading decisions are considered because, the order releases and the production targets for the following periods are not instructed from the SCOP level yet, due to the application of rolling horizons. A feasible $V_i(0)$ must satisfy the following conditions:

$$\begin{aligned} W_i(0) + V_i(0) &\leq \widehat{W}_i(0) + Q_i(0) && (5.13) \\ P_i(0) &\leq f_i\left(W_i(0) + V_i(0)\right) && (5.14) \end{aligned}$$

The decision variables $P_i(t)$ of the SCOP model may be interpreted differently depending on the level of aggregation assumed at the operational planning level. One interpretation is that these production decisions are parts of the detailed production plan, and the instruction $P_i(0)$ should be transferred to the lower level as a targeted quantity. Another interpretation is that these production decisions are parts of aggregate capacity check in determining achievable order releases at the SCOP level. Consequently, they should be ignored at a lower level operational scheduling and capacity loading model. This discussion on the level of aggregation at the SCOP level yields two types of coupling mechanism distinguished between the operational scheduling and the SCOP levels. These are *tight coupling* and *loose coupling*. The capacity loading decisions, $V_i(0)$, are given based on the coupling mechanism. In the following, we will first derive $V_i(0)$ under tight coupling.

At each production unit, the local costs are considered such as the cost of producing plus the end-of-period costs of holding unfinished items as workload and holding finished items due to excess production over the target quantity, $P_i(0)$. The expected production quantity is computed from the clearing function. The unit production cost is denoted by c_i. The costs of holding unfinished items as workload and finished items as excess production are determined based on the given probability distribution for P_∞. The objective of the operational scheduling and capacity loading model is to minimize the total cost function,

$$\begin{aligned}TC = &c_i \cdot f_i\left(W_i(0)+V_i(0)\right)+\\ &h_i^{(w)} \cdot \int_0^{W_i(0)+V_i(0)} \left(\widehat{W}_i(0)+Q_i(0)-x\right) g(x)dx+\\ &h_i^{(w)} \cdot \left(\widehat{W}_i(0)+Q_i(0)-W_i(0)-V_i(0)\right) \cdot \left(1-G\left(W_i(0)+V_i(0)\right)\right)+\\ &h_i^{(f)} \cdot \int_{P_i(0)}^{W_i(0)+V_i(0)} \left(x-P_i(0)\right) g(x)dx+\\ &h_i^{(f)} \cdot \left(W_i(0)+V_i(0)-P_i(0)\right) \cdot \left(1-G\left(W_i(0)+V_i(0)\right)\right) \qquad (5.15)\end{aligned}$$

subject to (5.13) and (5.14). The first term in Equation (5.15) represents the total cost of production during the current period. The second and the third terms are related to holding workload at the end of the current period after the production is realized. The forth and the fifth terms are associated with the cost of holding finished items due to excess production over the target quantity.

The first order partial derivative of TC with respect to $V_i(0)$ is

$$\begin{aligned}\frac{\partial}{\partial\left(V_i(0)\right)} TC = &c_i \cdot \frac{\partial}{\partial\left(V_i(0)\right)} f_i\left(W_i(0)+V_i(0)\right)+\\ &\left(h_i^{(f)}-h_i^{(w)}\right) \cdot \left(1-G\left(W_i(0)+V_i(0)\right)\right).\end{aligned}$$

One should note that the clearing function $f_i(\cdot)$ is a monotonically increasing and concave function, and $h_i^{(f)} \geq h_i^{(w)}$. As a result, the total cost function monotonically decreases as $V_i(0)$ decreases, $\frac{\partial}{\partial(V_i(0))}TC \geq 0$ for all $V_i(0) \geq 0$. Let us denote $V_i^*(0)$ as the optimum capacity loading decision for item i in the current period, and $f_i^{-1}(\cdot)$ as the inverse of the clearing function for item i. Then, from the monotonicity property of the total cost function and from Condition (5.14), the optimal capacity loading decision for item i in the current period under tight coupling is given by

$$V_i^*(0) = \max\left\{f_i^{-1}\left(P_i(0)\right) - W_i(0),\, 0\right\}. \qquad (5.16)$$

Another approach in loading the capacity ignores the production decisions made at a higher operational planning level. This refers to *loose* coupling; the released orders are immediately loaded to the shop floor regardless of the target production decisions made by the SCOP level. That is,

$$V_i^*(0) = Q_i(0). \qquad (5.17)$$

In such a case, the WIP level in a production unit is always equal to the workload of that production unit.

The planned delivery schedule of the released orders and $V_i^*(0)$ are the final outcomes of the planning system and are given to the execution systems such as ERP or MES.

5.4 Preliminary Analysis under Deterministic Assumptions

In this section, we provide insights into updating the lead times in a dynamic and deterministic environment. Our purpose here is to explicitly see the effects of updating the lead times on the release pattern within the supply chain exempt from some random events in the system. We consider a two-stage serial supply chain structure with identical production units. Each production unit operates with a nominal production rate of 10 units/period for a single item. The clearing function is provided with $\varepsilon = 0.05$ under the assumption that P_∞ follows from a Gamma distribution with coefficient of variation 0.25. It is assumed that the models of the production processes are exact; the quantity implied by the clearing behavior is actually produced at each production unit. Final product demand fluctuates in a seasonal manner between levels 6 units/period and 9 units/period as seen in Figure 5.4, and each season lasts for 25 periods. The illustrations in this section are provided for a duration of 100 periods of simulation considering two consecutive seasonal cycles. The fixed lead times are set to two periods for both stages. The following abbreviations are used with $k = 1, 2$ respectively indicating the downstream and upstream stages of the supply chain:

VL_k = Dynamic planned lead time for stage k of the supply chain.

I_k = Net inventory level at stock point of stage k under static lead times.

$I_k^{(VL)}$ = Net inventory level at stock point of stage k under dynamic lead times.

$\widehat{W}_k$ = Workload level at production unit of stage k under static lead times.

$\widehat{W}_k^{(VL)}$ = Workload level at production unit of stage k under dynamic lead times.

Figure 5.4 illustrates the fluctuating demand and the flow times of the released orders for the final product in a static planning framework with loose hierarchical coupling. The figure simply tells us that, although the planned lead times are fixed, the flow times are actually fluctuating with the changing market conditions. The purpose of updating the lead times should be keeping the

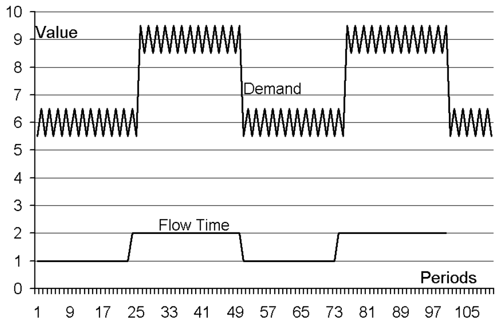
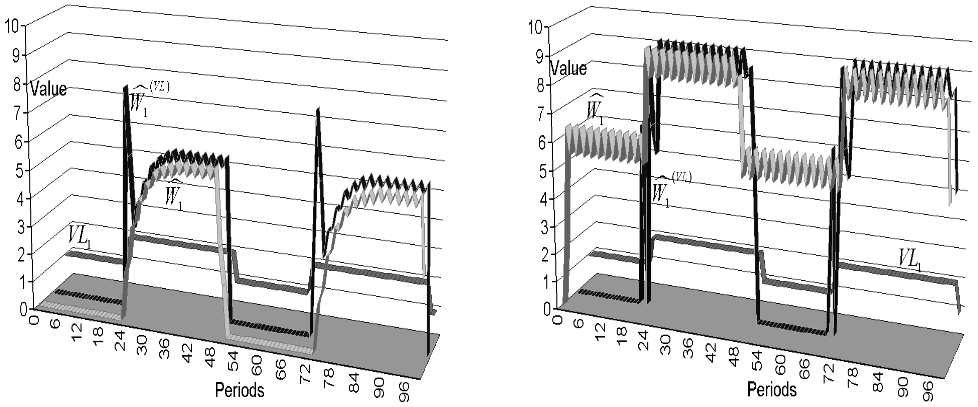


Figure 5.4: Fluctuating demand and flow times of the final product in the static case.

planning system responsive to the dynamic conditions such that the dynamic flow times are represented at the planning level. However, being responsive at the same time yields increased variability.



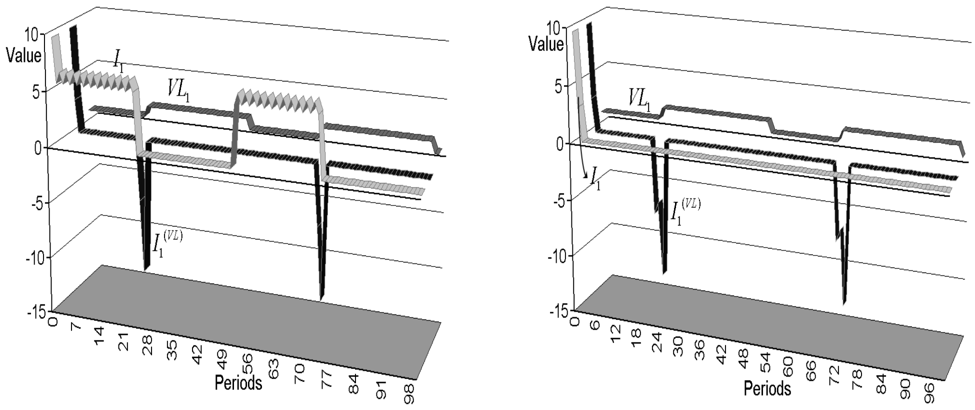
(a) Loose coupling.

(b) Tight coupling.

Figure 5.5: Total workload levels at the downstream production unit for static and dynamic cases.

Figures 5.5(a) and 5.5(b) show the variation in the workload level of the downstream production unit in the static and the dynamic cases (the figures for the upstream production unit are very similar to these figures). A large increase in the workload is experienced when the lead time is increased, and vice versa.

This is due to the fact that as the lead time is increased the system strives to balance the pipeline inventory with the increased demand during the planned lead time, and large orders are released at both stages. Figure 5.5(a) insights that the dynamic case responds more quickly to the changing demand levels (the workload is increased earlier, and the steady state is reached earlier) with the expense of increased costs of holding larger amount of workload in the production unit. Under tight coupling the capacity loading decisions are given in a way to realize the planned delivery schedules. That is, if the workload is small and the orders have long lead times, then the production can be delayed in order to decrease the finished item holding costs at the SCOP level. Therefore, as Figure 5.5(b) implies, the system with tight coupling has to hold larger workload in the production unit in both the dynamic and the static cases as compared to the cases with loose coupling. The effect is stronger for static lead times. Especially during low-demand season, the system holds larger workload due to early order release and postponed production when the lead time is static. In the dynamic case, the system responses to low demand by decreasing the planned lead times. Thus, the workload is decreased.

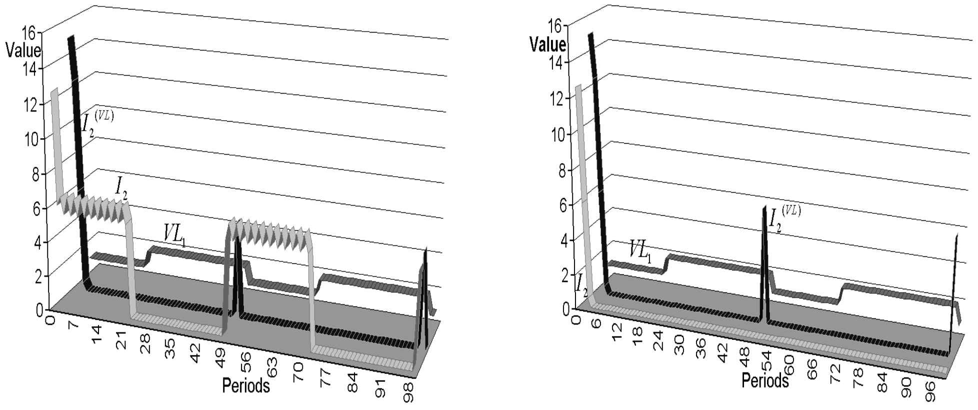


(a) Loose coupling.

(b) Tight coupling.

Figure 5.6: Downstream inventory levels for the static and the dynamic cases.

Figures 5.6(a) and 5.6(b) show the net inventory levels of the downstream stage under loose and tight coupling respectively. The dynamic lead times follow a seasonal cycle in accordance with the demand, and the inventory levels follow a repeating, characteristic pattern for each cycle. For example, when the demand level is low, static lead times generate larger inventory levels due to early delivery of orders under loose coupling. This pattern is



(a) Loose coupling.

(b) Tight coupling.

Figure 5.7: Upstream inventory levels for the static and the dynamic cases.

changed to a consistent inventory level of zero under tight coupling because, the delivery schedules planned by the SCOP level are executed exactly by the lower level. In the dynamic case, backorders are realized when the lead times are increased because, the demand for the early period(s) cannot be satisfied due to the extended delivery dates. This situation lasts only for a short duration after which the inventory level returns to its steady state. A relatively longer stock-out duration is seen under tight coupling as compared to the one under loose coupling.

Similarly, Figures 5.7(a) and 5.7(b) illustrate the inventory levels of the upstream stage respectively for situations with loose and tight coupling. In the static case, the upstream inventory follows a pattern similar to that of downstream inventory. In the dynamic case, an increase in the inventory level of the upstream stock point is seen when the lead time of the downstream stage is decreased. This occurs because, a decrease in the downstream planned lead time generates the effect that some of the intermediate items kept in stock are no longer necessary. The amount of intermediate items that is previously planned to be used for production in the downstream stage is decreased causing a temporary excess inventory in the upstream stock point.

To sum up, the basic insights from this preliminary analysis are:

1. Updating the lead times keeps the system responsive to structural changes in the demand level.

2. The type of coupling between the SCOP and the operational scheduling

levels significantly affects the supply chain performance.

3. An increase in the lead times causes temporary backorders, and the effect is stronger under tight coupling.

In the following section, we conduct simulation experiments under uncertain demand and production process conditions. We provide a more extensive and statistical analysis on the issues discussed in this section. Furthermore, we investigate the other interesting aspects of the planning system such as the effects of the clearing function parameter ε.

5.5 Simulation Experiments

5.5.1 Setting

The following system characteristics are not subject to change in our simulation experiments.

- Each production unit is assigned to produce a single item, and demands for the intermediate items originate only from the downstream order releases.
- The upstream and the downstream production units are assumed to have independent identically distributed processing times. Thus, the performance of the upstream and the downstream stages for changing design parameters can be evaluated independent of the specific process characteristics but based on the interaction between each other.
- The clearing representations for both production units are identical, and are derived from the assumption that P_∞ is Gamma distributed with mean 100 items/period and coefficient of variation 0.25. Accordingly, $L_i^{\max} = 2$ periods for $i = 1, 2$.
- The demand for the final product is seasonal with low season represented by a mean level of 60 items/period, and the high season is represented by a mean level of 90 items/period. Thus, in the long run, the production units are subject to 75% utilization.
- The final product demand forecasts are sampled from a Gamma distribution with coefficient of variation 0.50. Distribution mean changes between high and low seasons.
- Each season lasts for 25 periods.

- The forecast horizon for the final product demand is $T = 15$ periods. Given $L_i^{\max} = 2$ periods for $i = 1, 2$, the planning horizon is longer than the cumulative lead time.

- The target fill rate for the final product demand is 98%.

- The cost parameters are: $h_2^{(w)} = 1.0$ per unit per period, $h_2^{(f)} = 1.2$ per unit per period, $h_1^{(w)} = 1.5$ per unit per period, $h_1^{(f)} = 1.8$ per unit per period.

5.5.2 Design

There are three important design factors that we wish to evaluate in this study. These are related to the hierarchical planning structure, the clearing function model, and the level of responsiveness to changing market and shop conditions. We model these issues respectively through the coupling mechanism between the SCOP and the operational scheduling levels, the choice of ε in modeling the clearing behavior of the production processes, and whether or not the lead times are updated in response to changing inputs to the planning system. First of all, the coupling mechanism indicates the level of aggregation assumed at different levels of the planning hierarchy. We believe the performance of a hierarchical planning system depends on the extent with which the decision outcome of a higher level model is influential on a lower level model. Secondly, clearing function is a means of anticipating the operational characteristics of each production unit at a higher operational planning level. It is important to discuss the design related issues based on a well-defined parameter such as ε, to exclusively evaluate the impact of different anticipation approaches. These two issues are related to our fourth research question. Finally, we would like to investigate how good is our lead time setting procedure, which is related to the first and the second research questions.

The coupling mechanism can be tight or loose depending on how the capacity loading decisions are given at the operational scheduling level. Tight coupling indicates that the capacity loading decisions are given in accordance with the production targets set by the SCOP level so that the actual delivery schedule is kept close to the planned schedule. Loose coupling refers to an approach that immediately introduces the released orders into the shop, regardless of their planned schedules. Tight and loose coupling are respectively formulated in Equations (5.16)and (5.17).

Recent studies on clearing functions have mainly focused on the general structure of clearing whether through long-term steady state or short-term analysis. However, given the general structure, the design issues related to the detailed modeling of a clearing function have not been considered thus far. For this purpose, we introduce the parameter ε in modeling a piecewise-linear clearing function such that ε refers to an *internal service requirement* for the production process. For small ε a tight service is assumed such that the aim is to avoid late delivery of orders, and for large ε an optimistic approach is employed such that the aim is to avoid early deliveries. ε also plays a significant role in our lead time setting procedure, such that large ε supports shorter planned lead times and vice versa.

In addition to such design related issues, demand uncertainty is also considered. It is the percentage of deviation of the realized demand from the forecasted demand, and low and high uncertainty levels are respectively modeled through 0% and 50% deviation. 0% deviation refers to the deterministic demand case, and 50% deviation refers to the case that the actual demand in a period is generated from a Uniform($0.50\hat{d}$, $1.50\hat{d}$) distribution where $\hat{d}$ refers to the forecasted demand for that period. Thus, the actual demand values do not follow a stationary pattern between consecutive periods.

Table 5.1: Experimental design.

Factors	Treatments	Number of Treatments
Hierarchical coupling	Loose, Tight	2
Lead time strategy	Static ($L = 2$), Dynamic ($L = 1, 2$)	2
Clearing parameter, ε	0.05, 0.20	2
Demand uncertainty, U_D	0%, 50%	2

The list of all design factors are provided in Table 5.1. There are in total 16 different treatments, and the simulation for each treatment is performed with 15 replications using different random number streams between each replication. Between different treatments the same set of random number streams (see Law and Kelton (2000)) are implemented. Each simulation is performed for 7750 periods where the first 250 periods are considered as warm-up duration. The experiments are performed using QUINTIQ 3.1.0.10 (see Quintiq (2007) for detail) together with the CPLEX software used to solve the SCOP formulation.

Based on the intuition gathered from the preliminary analysis in Section 5.4 the following hypotheses on the performance of various design factors are developed:
Hypothesis 1 *Updating the lead times increases the safety stock required to guarantee the target final product demand fill rate as compared to the case with*

static lead times.

Updating the lead times generates nervousness by changing the delivery decisions made in successive epochs. When the lead time is increased from one period to two periods, the previously planned order to be released now and delivered at the end of the current period has to be postponed. This may yield increased backorders and thus, increased safety stock is needed to avoid this.

The logic behind updating the lead times is to plan the order releases on time so that the forecasted demand can be met by keeping the lowest amount of material both in the stock points and in the production units. Although the safety stock is increased, the total amount of material kept in the supply chain is expected to be decreased by better management of total workload levels and items in stock.

Hypothesis 2 *Updating the lead times decreases the total amount of material kept in the supply chain.*

Given that the planned lead time is greater than or equal to one period, tight coupling tends to keep the throughput levels relatively low by postponing production of items scheduled to be delivered in future periods. As implied by Figures 5.5(a) and 5.5(b), tight coupling is expected to increase the workload levels throughout the supply chain. In addition, with stochastic production processes, tight coupling may generate severe increases in the percentage of tardy orders. Thus, one would expect increased safety stock levels under tight coupling.

Hypothesis 3 *Tight hierarchical coupling increases the total amount of material kept in the supply chain.*

Modeling the clearing function with a large ε generates planned throughput quantities that may not be easily satisfied, which may cause late deliveries of the released orders. In addition, our lead time update procedure implies that a large ε may cause shorter lead times be planned at the tactical planning level.

Hypothesis 4 *Modeling the clearing function with a large ε value increases the percentage of orders that are tardy.*

In the following subsection, Hypotheses (1)-(4) are tested through statistical analysis of the simulation results.

5.5.3 Results

The measures of performance we are interested in are basically related to the coordination of material flow within the supply chain that generates costs (external performance), and the level of consistency between the outcomes of different hierarchical decision levels (internal performance). The following abbreviations are used in this section to represent various simulation outputs. $k = 1$ and $k = 2$ respectively refers to the downstream and the upstream stages of the supply chain.

SS = Safety stock level that satisfies the desired 98% fill rate for the final product demand.
I_k^+ = Average amount of finished items kept in the stock point or in the production unit of stage k.
$\widehat{W}_k$ = Average workload level in the production unit of stage k.
TC_k = Total inventory holding cost at stage k of the supply chain, $TC_k = h_k^{(f)} I_k^+ + h_k^{(w)} \widehat{W}_k$.
TC = Total inventory holding cost of the supply chain, $TC = \sum_{k=1}^{2} TC_k$.
EI = Average highest-echelon inventory position of the supply chain, $EI = \sum_{k=1}^{2} \left(I_k^+ + \widehat{W}_k \right)$.

For each different simulation, an initial run is taken with safety stock equal to zero. Then, SS is found by applying the safety stock adjustment procedure described in the Appendix of Chapter 2, and the simulation is repeated with SS as the new safety stock. EI is the average amount of material periodically kept in the supply chain, either as a workload or as a finished item, and relates to the total cost of the supply chain.

The other performance measures are related to the delivery of released orders. They are

L_k = Average lead time planned for the delivery of orders at stage k of the supply chain.
F_k = Average flow time of the orders at stage k of the supply chain.
ΔL_k = Average squared deviation of the flow times from the planned lead times at stage k of the supply chain.
Π_k = Average percentage of orders that are tardy at stage k of the supply chain.

We define the term ΔL_k as the *lead time error*, since it reflects the level of inconsistency between the outcome of the SCOP model and the result of the capacity loading decisions at the operational scheduling level.

Summarized simulation outputs are presented in Tables 5.3 and 5.4 in the Appendix of this chapter. Table 5.3 presents the performance measures, and their 95% confidence intervals, related to the total cost of the supply chain. Table 5.4 presents the performance measures, and their 95% confidence intervals, related to the delivery of the released orders. The results for the statistical tests on these sets of simulation outputs are provided in Tables 5.5, 5.6, and 5.7 in the Appendix of this chapter. Pairwise comparisons between different simulation outputs are performed based on a t-distribution test (see Law and Kelton (2000) for detail). Superscripts of "$\star$" and "$\diamond$" respectively indicate a 95% and a 99% confidence level in the rejection of the hypothesis that the outputs have the same distribution mean. In order to perform a reliable statistical significance test, the corresponding sets of data should be closely correlated with each other. The following abbreviations are used in the statistical analysis of the simulation outputs:

cc = Correlation coefficient between the data sets that are going to be compared to each other.

t-Stat = The corresponding t statistics value as a result of the t-test.

$\Delta\%_{VL}$ = The percentage of decrease in the corresponding performance measure due to updating the lead times.

$\Delta\%_{\text{loose}}$ = The percentage of decrease in the corresponding performance measure due to loose coupling.

$\Delta\%_{\varepsilon}$ = The percentage of decrease in the corresponding performance measure due to modeling the clearing function with $\varepsilon = 0.05$.

The greater the t-Stat is, the larger the confidence level is that the two data sets are coming from a distribution with different means.

Table 5.2: Coefficients of correlation between EI and TC, and Π_1 and TC.

		$U_D = 0\%$			
		Loose		Tight	
		$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$
$L = 2$	$EI - TC$	0.99	0.99	0.99	0.99
	$\Pi_1 - TC$	0.16	0.22	0.00	0.01
$L = 1, 2$	$EI - TC$	0.99	0.99	0.99	0.99
	$\Pi_1 - TC$	0.15	0.32	−0.20	0.09
		$U_D = 50\%$			
		Loose		Tight	
		$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$
$L = 2$	$EI - TC$	0.99	0.99	0.99	0.99
	$\Pi_1 - TC$	−0.15	−0.12	−0.21	−0.37
$L = 1, 2$	$EI - TC$	0.99	0.99	0.99	0.99
	$\Pi_1 - TC$	0.07	−0.16	−0.21	0.33

It is important to note that there is a very high positive correlation between the highest-echelon inventory position and the total periodic cost incurred in the supply chain (see Table 5.2). This means, in order to achieve an improvement in the total cost one should concentrate on keeping less workload and less finished items in the supply chain without sacrificing from the service level.

This also implies that the relative behavior of the total cost term between different cases is not very sensitive to changes in the unit cost parameters. In addition, Table 5.2 shows that there is not an obvious correlation between the percentage of orders that are tardy and the total cost of the supply chain. This reveals the idea that the cost increasing effects of tardy orders are compensated by other factors such as improved coordination of supply chain inventories in accordance with changing demand.

Updating the lead times increases the safety stock, but only when a large ε is used to model the clearing function. That is, Hypothesis 1 holds for large ε. The t-Stat values for $\varepsilon = 0.05$ (see Table 5.5 in the Appendix of this chapter) are all very close to zero yielding the argument that we cannot reject the idea that the dynamic and the static case SS values have identical means. Under $\varepsilon = 0.20$, we can state that the lead time update increases the safety stock. The increase ranges from 5% to 15%. As the demand uncertainty increases or as the hierarchical coupling becomes tight, the percentage of increase becomes (relatively) less. An illustration of the SS values (in terms of the percentage deviation from the minimum) is given in Figure 5.8 for $U_D = 0\%$. The minimum SS value is achieved with loose coupling, $\varepsilon = 0.05$, and with fixed lead times. A maximum of about 50% increase in the safety stock is seen when coupling is tight, $\varepsilon = 0.20$, and the lead times are dynamic.

As shown in Table 5.5 in the Appendix of this chapter, Hypothesis 2 holds for all cases. For the cases with tight coupling and $\varepsilon = 0.20$, the differences in TC are not large but still at a reasonable level. The highest improvement of the dynamic system is achieved for $\varepsilon = 0.05$ under loose coupling with $U_D = 0\%$ (see also Figure 5.8). As the coupling mechanism becomes tight the percentage of improvement in TC by updating the lead times decreases. Tight coupling apparently increases the deteriorating effects of schedule changes on the backorders in the dynamic case.

Another interesting result revealed by Figure 5.8 is that the decrease in TC is still maintained when SS increases significantly with dynamic lead times, i.e. the case with $\varepsilon = 0.20$. It gives the insight that although the nervousness in the schedules and thus the variability in the final product stock levels are increased, updating the lead times may still decrease the costs by better managing the workload levels and keeping the planned delivery schedules realistic. A further analysis on this issue should take into account the value adding structure among the supply chain, which we believe is an interesting topic for future research.

Under loose coupling, orders are introduced to the shop immediately after they are released, whereas under tight coupling, capacity loading decisions

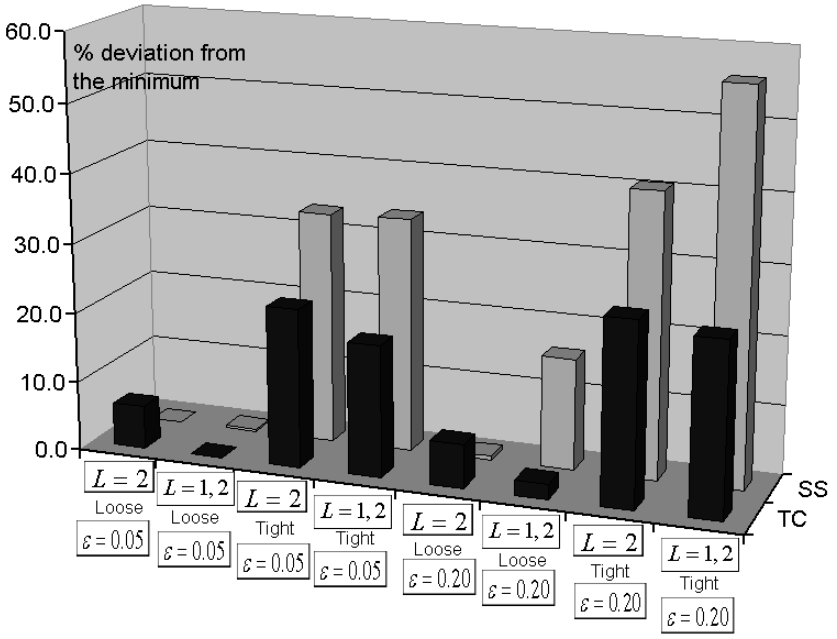


Figure 5.8: SS and TC values with respect to their minimums, $U_D = 0\%$.

are given in a way to realize the planned deliveries. Although tight coupling generates consistent planning outcomes, through reduced lead time errors, the deterministic point-of-view of the SCOP model is carried onto the actual execution of the orders through lower level capacity loading decisions. As a result, the number of orders that are tardy increases substantially, and higher final product inventory is kept due to significantly increased safety stock levels. Despite the decrease in the upstream item inventory level, we can state that the total amount of material kept in the supply chain increases, and the total cost increases under tight coupling as illustrated by Figure 5.8. Thus, our claim in Hypothesis 3 is supported. Table 5.6 in the Appendix of this chapter provides the levels of improvements in EI and TC achieved due to loose coupling between the SCOP and the operational scheduling levels. The difference is greater in the dynamic case than it is in the static case, and it is smaller under higher demand uncertainty. Loose coupling generates a safety factor against the random deviations in the stock levels by releasing and producing the orders earlier, which may become more important for the dynamic case by smoothing the effects of schedule changes due to dynamically changing lead times. As it is expected, the choice of ε plays a crucial role in determining the relative performance of tight and loose coupling. The tight coupling generates higher costs as compared to the loose coupling when a larger ε is applied for the clearing function. Loose coupling is most preferable when the lead times are updated under $U_D = 0\%$ and $\varepsilon = 0.20$. It is the situation where the effect of schedule nervousness can be smoothed mostly through early production and delivery of orders.

The value of ε indicates a manufacturing service level whether a target production quantity in a period can be met or not. Table 5.7 in the Appendix of this chapter provides the statistical results for the relative effect of different ε values. A low ε value causes easily achievable periodic production levels be planned at the SCOP level. Hypothesis 4 holds for almost all cases, except the case with loose coupling and fixed lead times, where we expect ε does not play any significant role at the operational scheduling level. Decreasing ε decreases the percentage of tardy orders by up to 7.73%. In the static case, ε is mainly effective in determining the periodic production quantities. Thus, the choice of ε in the static case is more important when there is tight coupling as compared to the situation with loose coupling situation. SS and TC values for different ε's in Figure 5.8 supports this intuition. In the dynamic case, ε plays a more important role by affecting the average planned lead times in addition to the capacity loading decisions. The planned lead times decrease as ε gets larger, and vice versa. The distribution of planned lead times is most effective under loose coupling in determining the delivery performance. Thus, low ε yields the biggest improvement on Π_1 under dynamic lead times in a planning system with loose coupling. In a dynamic system, the improvement by low ε on Π_1 decreases as the coupling becomes tight and as the demand uncertainty increases.

As we can see in Table 5.2, the percentage of tardy deliveries is by itself not a significant factor in determining the total cost of the supply chain. Thus, in assessing the cost effects of a design variable, we should look at how the workload and the finished items inventory changes. We see that the value of ε does not play any significant role on TC in the static case under loose coupling. This is intuitive from the fact that in such a case, the effects of ε on the planned lead times and on the capacity loading levels are not transferred to the executable decisions. Under tight coupling, large ε supports the production be postponed to later periods for orders with long lead times causing an increase in the average workload levels. As a result, in the static case with tight coupling, a low ε value is preferable in terms of the total supply chain costs. In the dynamic case, the highest improvement with low ε is 4.73%, and it is achieved under tight coupling with $U_D = 0\%$. As the coupling becomes loose or the demand uncertainty increases, the level of improvement in TC by low

ε decreases. For $U_D = 50\%$ and under loose coupling, the choice of ε fails to generate a significant difference between the TC performances.

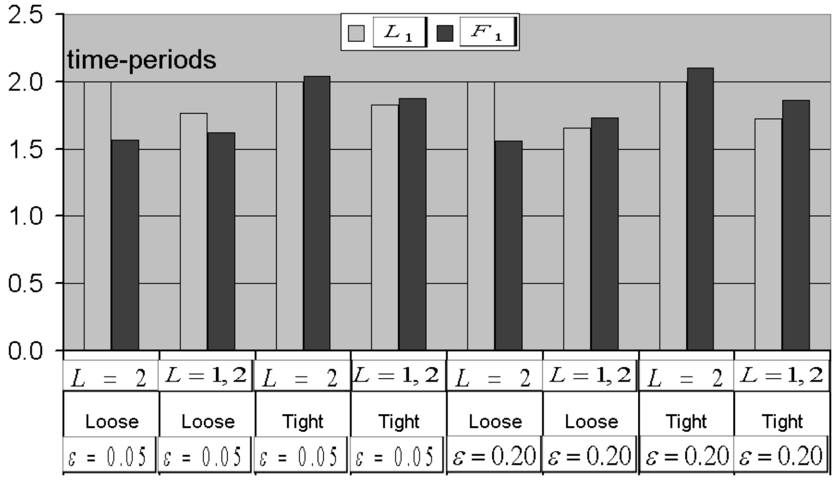


Figure 5.9: L_1 and F_1 values under $U_D = 0\%$.

One of the main characteristics of the dynamic case is that the coupling between the SCOP and the operational scheduling levels is already maintained by fine-tuning demand and supply through updating the lead times at a higher tactical level. By means of updating the lead times, the planned delivery schedule of order releases is kept responsive to the anticipated production requirements. If the requirements are smaller than a certain level, then shorter lead times are planned for smaller order sizes, which are expected to be delivered shortly. Similarly, longer lead times are planned for larger orders. Thus, a considerable improvement in the consistency between the planned lead times and the flow times is achieved. This can be seen in Figure 5.9 where we illustrate the average planned lead times and the average order flow times of the final product for the deterministic demand situation. As one applies tight coupling in the dynamic case, then the lead times and the flow times are expected to become longer. This is due to the increased workload levels under tight coupling. When the lead times are fixed, tight coupling is more effective in terms of decreasing ΔL_k as compared to its effect in the dynamic case. This is done by increasing the order flow times, since the planned lead times are fixed. The use of dynamic planned lead times generates increased flow times as compared to the static case when the coupling is loose. This is due to the increased variability in the order sizes caused by the dynamic lead times.

However, under tight coupling, the dynamic lead times have a reverse effect that the flow times are shortened due to the decreased planned lead times as compared to the static case.

In addition, tight coupling diminishes the difference between the static and the dynamic cases in terms of the lead time error. Apparently, tight coupling aims at realizing the delivery schedules as determined by the SCOP level whether the lead times are fixed or not. The sensitivity of the lead time error on the choice of the clearing parameter ε increases as one moves up in the supply chain (see Table 5.4 in the Appendix of this chapter). This is due to the smoothing effect of the upstream inventory on the downstream order releases and due to the increased variability in the upstream order releases in response to updating the downstream lead times. These insights are also supported by the results provided in Section 2.2.2.

5.6 Conclusion

The contributions of this chapter can be discussed in various directions. First of all, we have provided an effective procedure to set the planned lead times of a supply chain operations planning system. The procedure has been described in a general framework, and applied for a two-stage, two-item serial structure. The concept of clearing function has been used as a basis in updating the planned lead times. Thus, our second contribution is related to this stream of literature. We have introduced the clearing parameter ε in approximating the clearing function. The results show various insights related to high or low values of ε. Finally, we have contributed to the hierarchical planning literature by modeling loose and tight coupling between the SCOP and the operational scheduling levels. The results indicate that for situations with stochastic production processes and with stochastic demand, the loose coupling is favored over the tight coupling. The deterministic point of view at a higher SCOP level should not be carried onto the execution related decisions at a lower scheduling level. Our results in this chapter provide an extensive discussion on the scope and the range of instructions from a higher decision level to a lower decision level in the planning hierarchy.

The lead time update procedure described in this section can be applied to general supply chain settings including more than two stages and complex (convergent, divergent or mixed) product structures. The main idea is to start from the top of the BOM and consider each level separately by making the necessary substitutions as we have described in this chapter. For divergent situations, the procedure may be applied without structural changes since each downstream item is associated to a single upstream item and the demand explosion upstream through the supply chain can be easily traced. For convergent situations, there is a positive correlation between the production requirements of the upstream items that are used by the same downstream item. Thus, a procedure for updating the lead times of a convergent supply chain may additionally have to consider this important aspect of demand explosion.

Therefore, an immediate direction for further research is to update the planned lead times for more complex production-distribution structures. The demand seasonality remains to be an important system characteristic for such an attempt, and the derivation of the clearing behavior for multi-item production is still a complex task. In terms of modeling the clearing behavior, the relationship between the clearing parameter ε and the delivery reliability should be further analyzed by concentrating on the interaction between the clearing function and the order release pattern.

Appendix to Chapter 5

Table 5.3: Cost performance.

		Loose Coupling			
		$\varepsilon = 0.05$		$\varepsilon = 0.20$	
		$L = 2$	$L = 1,2$	$L = 2$	$L = 1,2$
U_D = 0%	SS	232.5 ± 5.6	233.2 ± 5.5	233.9 ± 6.1	269.8 ± 5.8
	I_1^+	268.4 ± 5.0	253.1 ± 5.1	268.6 ± 5.6	255.5 ± 4.9
	$\widehat{W}_1$	27.5 ± 0.5	29.8 ± 05	27.1 ± 0.5	36.3 ± 0.5
	TC_1	524.3 ± 9.2	500.2 ± 9.3	524.2 ± 10.2	514.4 ± 9.2
	I_2^+	61.9 ± 02	50.2 ± 0.4	62.6 ± 0.2	46.8 ± 0.4
	$\widehat{W}_2$	22.3 ± 03	24.0 ± 0.3	21.9 ± 0.3	26.6 ± 0.3
	TC_2	96.6 ± 03	84.3 ± 0.4	97.1 ± 0.4	82.7 ± 0.5
	EI	380.1 ± 5.2	357.1 ± 5.4	380.3 ± 5.8	365.2 ± 5.3
	TC	620.9 ± 9.3	584.5 ± 9.5	621.3 ± 10.3	597.1 ± 9.4
U_D = 50%	SS	341.6 ± 9.6	342.5 ± 9.7	345.8 ± 10.5	372.3 ± 10.0
	I_1^+	359.6 ± 9.0	343.6 ± 9.1	361.9 ± 9.9	339.8 ± 9.3
	$\widehat{W}_1$	30.6 ± 0.5	34.2 ± 0.5	30.2 ± 0.5	39.9 ± 0.4
	TC_1	693.2 ± 16.2	669.8 ± 16.4	696.8 ± 17.9	671.6 ± 16.8
	I_2^+	65.3 ± 0.2	60.7 ± 0.3	66.1 ± 0.3	56.8 ± 0.5
	$\widehat{W}_2$	24.5 ± 0.4	28.3 ± 0.4	24.3 ± 0.4	30.3 ± 0.4
	TC_2	102.9 ± 0.4	101.2 ± 0.5	103.6 ± 0.4	98.5 ± 0.7
	EI	480.0 ± 9.1	466.9 ± 9.2	482.5 ± 9.9	466.9 ± 9.4
	TC	796.1 ± 16.3	771.1 ± 16.5	800.4 ± 17.9	770.1 ± 16.9
		Tight Coupling			
		$\varepsilon = 0.05$		$\varepsilon = 0.20$	
		$L = 2$	$L = 1,2$	$L = 2$	$L = 1,2$
U_D = 0%	SS	310.3 ± 8.7	310.9 ± 8.1	327.2 ± 8.8	362.7 ± 8.3
	I_1^+	299.7 ± 7.9	297.0 ± 7.4	305.3 ± 8.0	315.9 ± 7.3
	$\widehat{W}_1$	48.8 ± 0.2	41.3 ± 0.3	53.4 ± 0.3	41.5 ± 0.4
	TC_1	612.6 ± 14.0	596.4 ± 13.3	629.6 ± 14.3	630.9 ± 13.3
	I_2^+	35.6 ± 0.2	37.7 ± 0.3	34.5 ± 0.2	37.9 ± 0.2
	$\widehat{W}_2$	63.0 ± 0.3	54.1 ± 0.3	67.5 ± 0.2	53.9 ± 0.5
	TC_2	105.7 ± 0.4	99.3 ± 0.5	109.0 ± 0.4	99.3 ± 0.7
	EI	447.0 ± 7.9	430.0 ± 7.6	460.7 ± 8.0	449.2 ± 7.7
	TC	718.3 ± 14.1	695.8 ± 13.5	738.6 ± 14.4	730.3 ± 13.6
U_D = 50%	SS	425.7 ± 11.6	426.3 ± 11.1	444.8 ± 10.4	468.8 ± 11.9
	I_1^+	394.4 ± 10.5	390.3 ± 10.2	401.2 ± 9.6	402.8 ± 10.8
	$\widehat{W}_1$	48.3 ± 0.2	44.3 ± 0.3	52.9 ± 0.2	44.3 ± 0.4
	TC_1	782.5 ± 18.8	769.0 ± 18.3	801.6 ± 17.2	791.5 ± 19.3
	I_2^+	43.0 ± 0.2	46.3 ± 0.2	41.8 ± 0.2	45.3 ± 0.3
	$\widehat{W}_2$	64.0 ± 0.3	59.3 ± 0.4	68.5 ± 0.3	59.2 ± 0.5
	TC_2	115.6 ± 0.5	114.9 ± 0.6	118.6 ± 0.4	113.5 ± 0.6
	EI	549.8 ± 10.5	540.2 ± 10.2	564.4 ± 9.6	551.6 ± 10.8
	TC	898.1 ± 18.9	883.9 ± 18.4	920.2 ± 17.3	905.0 ± 19.3

Table 5.4: Delivery performance.

		Loose Coupling			
		$\varepsilon = 0.05$		$\varepsilon = 0.20$	
		$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$
U_D = 0%	L_1	2.00 ± 0.00	1.77 ± 0.00	2.00 ± 0.00	1.66 ± 0.01
	F_1	1.56 ± 0.01	1.62 ± 0.01	1.56 ± 0.01	1.73 ± 0.01
	ΔL_1	0.63 ± 0.00	0.38 ± 0.00	0.64 ± 0.00	0.31 ± 0.00
	Π_1	9.50 ± 0.34	11.54 ± 0.33	9.61 ± 0.33	19.27 ± 0.37
	L_2	2.00 ± 0.00	1.71 ± 0.00	2.00 ± 0.00	1.54 ± 0.01
	F_2	1.51 ± 0.01	1.54 ± 0.00	1.50 ± 0.01	1.59 ± 0.00
	ΔL_2	0.61 ± 0.00	0.32 ± 0.00	0.62 ± 0.00	0.24 ± 0.00
	Π_2	5.81 ± 0.16	7.66 ± 0.15	5.91 ± 0.16	14.19 ± 0.25
U_D = 50%	L_1	2.00 ± 0.00	1.86 ± 0.00	2.00 ± 0.00	1.74 ± 0.00
	F_1	1.63 ± 0.01	1.72 ± 0.01	1.63 ± 0.01	1.80 ± 0.01
	ΔL_1	0.59 ± 0.00	0.42 ± 0.00	0.60 ± 0.00	0.34 ± 0.00
	Π_1	10.97 ± 0.32	13.23 ± 0.30	11.21 ± 0.34	19.87 ± 0.28
	L_2	2.00 ± 0.00	1.80 ± 0.00	2.00 ± 0.00	1.63 ± 0.01
	F_2	1.55 ± 0.01	1.64 ± 0.01	1.55 ± 0.01	1.66 ± 0.01
	ΔL_2	0.58 ± 0.00	0.34 ± 0.00	0.59 ± 0.00	0.26 ± 0.00
	Π_2	6.79 ± 0.21	8.85 ± 0.21	7.07 ± 0.20	14.21 ± 0.25
		Tight Coupling			
		$\varepsilon = 0.05$		$\varepsilon = 0.20$	
		$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$
U_D = 0%	L_1	2.00 ± 0.00	1.83 ± 0.00	2.00 ± 0.00	1.72 ± 0.00
	F_1	2.04 ± 0.00	1.87 ± 0.00	2.10 ± 0.00	1.86 ± 0.01
	ΔL_1	0.22 ± 0.00	0.22 ± 0.00	0.23 ± 0.00	0.25 ± 0.00
	Π_1	12.80 ± 0.20	13.36 ± 0.26	16.35 ± 0.31	18.94 ± 0.30
	L_2	2.00 ± 0.00	1.80 ± 0.00	2.00 ± 0.00	1.67 ± 0.01
	F_2	2.17 ± 0.00	1.99 ± 0.00	2.21 ± 0.00	1.95 ± 0.01
	ΔL_2	0.23 ± 0.00	0.23 ± 0.00	0.27 ± 0.00	0.31 ± 0.00
	Π_2	19.60 ± 0.20	20.88 ± 0.22	23.72 ± 0.26	29.18 ± 0.29
U_D = 50%	L_1	2.00 ± 0.00	1.91 ± 0.00	2.00 ± 0.00	1.79 ± 0.01
	F_1	2.03 ± 0.00	1.95 ± 0.00	2.10 ± 0.00	1.92 ± 0.01
	ΔL_1	0.24 ± 0.00	0.25 ± 0.00	0.24 ± 0.00	0.26 ± 0.00
	Π_1	13.65 ± 0.23	14.41 ± 0.22	17.01 ± 0.30	19.29 ± 0.28
	L_2	2.00 ± 0.00	1.88 ± 0.00	2.00 ± 0.00	1.75 ± 0.01
	F_2	2.19 ± 0.00	2.09 ± 0.01	2.24 ± 0.00	2.05 ± 0.01
	ΔL_2	0.25 ± 0.00	0.27 ± 0.00	0.29 ± 0.00	0.33 ± 0.00
	Π_2	21.92 ± 0.35	23.69 ± 0.35	26.15 ± 0.29	30.76 ± 0.32

Table 5.5: Statistical significance tests for the static and the dynamic case SS, EI and TC values.

	SS							
	$U_D = 0\%$				$U_D = 50\%$			
	Loose		Tight		Loose		Tight	
	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$
cc	0.96	0.73	0.95	0.91	0.97	0.92	0.97	0.91
$\Delta\%_{VL}$	-0.32	$-15.33^\diamond$	-0.19	$-10.84^\diamond$	-0.25	$-7.65^\diamond$	-0.13	$-5.40^\diamond$
t-Stat	0.79	13.82	0.37	16.37	0.63	10.80	0.35	8.03
	EI							
	$U_D = 0\%$				$U_D = 50\%$			
	Loose		Tight		Loose		Tight	
	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$
cc	0.96	0.77	0.95	0.91	0.97	0.92	0.97	0.89
$\Delta\%_{VL}$	$6.04^\diamond$	$3.97^\diamond$	$3.81^\diamond$	$2.50^\diamond$	$2.74^\diamond$	$3.24^\diamond$	$1.75^\diamond$	$2.27^\diamond$
t-Stat	25.40	6.73	11.41	5.89	10.10	6.77	6.76	4.40
	TC							
	$U_D = 0\%$				$U_D = 50\%$			
	Loose		Tight		Loose		Tight	
	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.20$
cc	0.96	0.76	0.95	0.91	0.97	0.92	0.97	0.89
$\Delta\%_{VL}$	$5.86^\diamond$	$3.89^\diamond$	$3.14^\diamond$	$1.13^\star$	$3.15^\diamond$	$3.79^\diamond$	$1.59^\diamond$	$1.65^\star$
t-Stat	22.77	5.99	8.54	2.38	10.67	7.33	5.58	2.93

Table 5.6: Statistical significance tests between loose and tight coupling for EI and TC values.

		$U_D = 0\%$				$U_D = 50\%$			
		$\varepsilon = 0.05$		$\varepsilon = 0.20$		$\varepsilon = 0.05$		$\varepsilon = 0.20$	
		$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$
EI	cc	0.82	0.89	0.76	0.83	0.94	0.95	0.93	0.79
	$\Delta\%_{\text{loose}}$	$14.97^\diamond$	$16.95^\diamond$	$17.46^\diamond$	$18.70^\diamond$	$12.69^\diamond$	$13.57^\diamond$	$14.51^\diamond$	$15.36^\diamond$
	t-Stat	24.33	33.44	26.16	32.34	31.55	38.96	37.98	21.77
TC	cc	0.82	0.88	0.76	0.83	0.94	0.95	0.93	0.79
	$\Delta\%_{\text{loose}}$	$13.55^\diamond$	$15.98^\diamond$	$15.88^\diamond$	$18.24^\diamond$	$11.36^\diamond$	$12.76^\diamond$	$13.02^\diamond$	$14.91^\diamond$
	t-Stat	19.77	28.19	21.23	28.71	25.67	33.62	30.48	19.39

Table 5.7: Statistical significance tests between low and high ε for Π_1 and TC values.

		$U_D = 0\%$				$U_D = 50\%$			
		Loose		Tight		Loose		Tight	
		$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$	$L = 2$	$L = 1, 2$
Π_1	cc	0.94	0.67	0.68	0.71	0.93	0.77	0.81	0.68
	$\Delta\%_\varepsilon$	0.11	$7.73^\diamond$	$3.55^\diamond$	$5.58^\diamond$	$0.24^\diamond$	$6.64^\diamond$	$3.36^\diamond$	$4.88^\diamond$
	t-Stat	1.72	45.22	26.49	43.66	3.22	56.55	32.43	39.68
TC	cc	0.94	0.86	0.89	0.91	0.97	0.89	0.86	0.89
	$\Delta\%_\varepsilon$	0.05	$2.10^\diamond$	$2.75^\diamond$	$4.73^\diamond$	0.53	-0.13	$2.40^\diamond$	$2.34^\diamond$
	t-Stat	0.16	4.25	5.05	10.41	1.54	0.22	3.84	4.08

Chapter 6

Conclusion and Future Research

The main topic of this thesis has been to evaluate the dynamic performance of hierarchical planning systems with the emphasis on modeling and evaluation of dynamic planned lead times. The method of rolling horizons is at the core of our applied methodology because, it is very relevant from a real life perspective, and it is a common planning methodology to deal with stochastic events in a periodic planning environment. We have used it in fine-tuning the outcomes of different planning levels and for implementing the relevant update activities.

Discussion on the limitations of the traditional HPP methodology has been carried over and comprehensive frameworks have been introduced in a conceptual manner by a few studies (e.g., Bertrand *et al.* (1990), and Schneeweiss (1999)). In this thesis, we have extended those discussions, to a certain degree, from a conceptual framework towards a more explicit framework considering issues related to implementation such as application of rolling horizons, feedback of status information and updating of parameters, and the coupling mechanism between different levels.

We have related our models and methodology to the design and implementation of commercial APS software. Thus, the insights we have delivered as part of the contributions of this thesis are relevant inputs in configuring real-life APS applications. These systems are supported by an advanced information technology backbone, and it is claimed that any new information generated by a planning module within APS can be automatically transferred to other modules as an input. We wanted to extend the discussion further by posing the question to what extent all the available information should be used to update plans and planning parameters.

The research conducted in this thesis can be separated into two related interests, which are also associated respectively to our first and second research questions. Firstly, we were interested in describing the performance consequences of updating the planned lead times through experimental and analytical findings. Secondly, motivated by the common intuition that uncertainty and workload levels are important determinants of actual flow times, we were interested in developing an effective way of updating the lead times. In conducting these studies, important design factors such as the frequency of updating, the coupling mechanism, and the level of anticipation have to be considered together with various environmental factors such as demand uncertainty and utilization levels.

In the following, we briefly present our findings and discuss about further research directions that would be initiated from this thesis.

6.1 Order Release Variability and Dynamic Lead Times

Our attempts in understanding and describing the operational dynamics of updating the planned lead times have resulted in a simulation study and an analytical study that is based on queueing theory. We have employed naive update procedures such as determining the lead times based on exponential smoothing of order flow times, or based on the total number of orders waiting in the production backlog. Through simulation we have shown that

- Common methods of updating the lead times generate erratic order releases and large variation in inventory levels.
- More frequent updates or ignoring capacity constraints increase the deteriorating effect of dynamic planned lead times.

Although improved forecast accuracy is achieved by frequently updating the lead times, the inherent variability in the order generating process is further amplified through dynamic planned lead times. This phenomenon has conceptually been defined as the lead time syndrome in the literature. We wanted to further analyze this phenomenon because, it has revealed an interesting cyclic interaction between separate decision levels that can only be observed in a dynamic setting through plan-execute-feedback-(re)plan cycles. For this purpose, we have modeled a simplified situation by concentrating on the status of a single production unit. We have used analytic queueing constructs, and derived the following insights:

- Updating the lead times increases the size of the production backlog, and order flow times get longer.
- For a symmetric response function the utilization level in the static case is retained in the dynamic case.
- The effect of updating the lead times increases with the update frequency or with the utilization level.

We have shown that a multi-dimensional queueing model can be used to analyze planning systems with dynamic planned lead times, with one dimension being the planned lead time and the other dimension being the backlog level that changes depending on the lead time and on the stochastic behavior of the system itself. This methodology can be extended to general settings to analyze planning systems in a plan-execute-feedback-(re)plan framework, where the control parameter and the dependent variable may be observed simultaneously. However, one should overcome the challenge of finding an explicit rate matrix for the matrix geometric representation in order to achieve closed form solutions. The current state of the literature presents working techniques for a limited class of QBD processes (see Ramaswami and Latouche (1986), and Van Leeuwaarden and Winands (2005) for some examples).

In the studies we have conducted within the context of erratic orders releases caused by dynamic lead times, the modeled situation is characterized by stationary demand and production processes where updating the planned lead times is performed in response to random deviations in the system status.

Steady-state analysis of stationary situations are useful in characterizing the performance consequences of various changes in the configuration of the planning system. In that respect, since we are generally interested in planning supply chains, there are some interesting extensions to our models. First of all, we have limited the explicit analysis of the lead time syndrome to the workload level of the production unit. Our queueing model may be extended to include inventory levels, and possibly relating the analysis to safety stock calculations. In this way, the phenomenon of lead time syndrome can be described in terms of its effects on the safety stocks.

Furthermore, considering that the consistency between the planned and the actual delivery of orders can be improved by updating the lead times, there occurs a tradeoff between improved service to downstream stages and increased variability in upstream operations. It is an interesting future research direction to perform an in-depth analysis of this tradeoff (or to see if it really exists), and interesting research questions may be raised such as at what stage of the supply chain it is most effective to update the planned lead times. Especially, the supply chain structure (convergent, divergent, or serial) or the value adding structure within the supply chain play crucial roles in improving the discussion on these issues.

6.2 Clearing Functions and Dynamic Lead Times

Zijm and Buitenhek (1996) pointed out an important limitation of static systems by saying that "a major drawback of a fixed planned lead time is the ignorance of the correlation between the actual workloads and the flow times that can be realized under a limited capacity flexibility". Considering different levels within the planning hierarchy, we have discussed about how to model the relation between workloads and flow times at an aggregate operational planning level. The concept of clearing function has been defined in the literature for this purpose, mainly concentrating on the analysis of single-stage manufacturing situations. We have extended the analysis to consider supply chain situations with emphasis on the concept of planned lead times and the associated order flow times. In addition, we have introduced a clearing function that is based on the short-term probabilistic analysis of the production unit. We have argued that the clearing behavior, as an anticipation on the operational dynamics of the production processes, modeled at a higher planning level should be based on a probability distribution determined jointly by that level's decision outcomes and the operational characteristics of lower level execution systems. Through simulation we have tested and compared the short-term clearing function with the established clearing functions in the literature. We have shown that

- The shape of the clearing function is an important factor that affects the consistency of the planned schedules with their executions.
- Applying a clearing function that is based on the short-term probabilistic analysis improves the coordination of capacity loading and order release decisions.

Our analysis on different clearing functions has revealed an interesting result that there is a tradeoff between loading the production unit early with high WIP levels causing early delivery of orders and keeping low WIP in the shop causing increased number of late deliveries. It is an interesting future research direction to further analyze this tradeoff by considering the value adding structure of the supply chain. In conducting such an analysis it may be fruitful to discuss the impact of the detailed clearing function shape instead of the underlying modeling assumptions of the clearing function.

In addition, we have ignored the delay between the time that the production unit is loaded with newly released raw materials and the time its effect on the throughput level is actually realized. We have assumed that the capacity loading decisions are effectuated within the period that they are given. An extension to the established models may include a positive time-delay between the moment that the production unit is loaded with raw materials and the moment that its effect on the production output is realized. Such an analysis would require considering detailed shop floor structures such as dedicated flow lines with a number of bottleneck stations or dynamic job shops. Models of clearing functions incorporating such detailed manufacturing situations have not been available in the literature yet.

In developing an effective procedure to update the planned lead times, we have utilized the idea that the clearing function is a useful tool in establishing the correlation between workloads and flow times at a tactical planning level. Our procedure is mainly based on the concept of lead time dependent throughput rates (derived from a clearing function) and the concept of latest stock-out period within the planning horizon. Simulation experiments have been performed under non-stationary seasonal demand conditions. We have shown that the proposed model is an effective way of responding to structural changes in environmental conditions, and the performance depends on design choices in configuring the planning hierarchy such as the specific shape of the clearing function (through clearing parameter ε), and the coupling mechanism between the operational planning and the capacity loading levels. We have found the following:

- Updating the planned lead times by our procedure decreases the total amount of material kept in the system. The dynamic lead times are less effective when $\varepsilon = 0.20$ as compared to the case when $\varepsilon = 0.05$.
- Modeling the clearing function with $\varepsilon = 0.20$ increases the percentage of tardy orders and decreases the average planned lead time as compared to the case with $\varepsilon = 0.05$.
- Loose coupling is more effective than tight coupling in terms of decreased safety stock levels and decreased tardy deliveries.
- Dynamic lead times are more effective under loose hierarchical coupling.

The concept of hierarchical coupling plays a crucial role in designing planning hierarchies because, it indicates the level of aggregation assumed at different levels of the hierarchy. Hax and Meal (1975) modeled it by predefined aggregation levels under deterministic assumptions. Since then, there have not been more explicit discussions on this issue, especially for situations with stochastic demand and production processes. In this thesis, we have included a certain degree of flexibility in modeling the coupling mechanism between the operational planning and the capacity loading levels. We have shown that the effect of hierarchical coupling on the system performance is not trivial, and further formal analysis of this concept needs to be conducted. Traditional HPP models based on static performance evaluation (the result of the hierarchical planning system is compared to the result of the corresponding monolithic model) suggest maintaining tight coupling. Our results suggest that loose coupling should be applied when the performance is evaluated in a dynamic setting, and when one assumes that demand and production processes are stochastic. A combination of loose coupling and dynamically adjusting the planned lead times performs best.

We have to indicate that the lead time setting procedure we have established may not be the best possible technique, and there are future research opportunities to challenge our procedure. One alternative maybe to anticipate the future workload levels through the probability distribution of demand and update the probability distribution of demand periodically, and accordingly set the planned lead times that best fit the given demand distribution and the anticipated future workload levels.

Our discussion on the clearing parameter ε adds a new dimension to the current literature on clearing functions that the problem of modeling the clearing behavior is not only an issue related to the representation but also, it can be considered as a design problem based on a given choice of parameter ε. In this way, we can further enhance the discussion about "realistic" clearing functions towards "optimal" clearing functions. In specific, as a future research question we may ask, what is the value of the clearing parameter ε that provides the lowest cost solution to a supply chain operations planning problem with given cost parameters, lead times, and demand distribution. There are available techniques such as simulation based optimization that may be used to tackle such questions.

To sum up, this thesis contributes to the development of planning hierarchies evaluated and modeled in a dynamic framework. Our emphasis has been on periodically updated planned lead times in a rolling horizon setting. In this thesis, we have provided numerical insights both from analytical models and from simulation experiments to better understand the performance consequences of dynamic planned lead times and to pursue effective ways of updating them. One may consider the content of this thesis as a set of initial studies about an interesting, yet evolving and a challenging research topic. This is about managing supply chain operations through dynamic and adaptive decision tools.

Planning and controlling supply chains is a complex task, and the development of APS supported by advanced IT infrastructures has been directed to help the practitioners make their decisions. There is a continuous change in market requirements and operating conditions, and one has the possibility to adapt APS using the data available from ERP systems. This brings forward new challenges and opportunities in the design and implementation of APS in a dynamic framework. This thesis provides formal results that are helpful in the evaluation and development of new techniques to be used in configuring and implementing adaptive APS in a rolling horizon setting.

References

Anthony, R.N. 1965. *Planning and control systems: A framework for analysis.* Boston: Harvard Business School Press.

Ari, E.A., and Axsäter, S. 1988. Disaggregation under uncertainty in hierarchical production planning. *European Journal of Operational Research*, **35**, 182–188.

Armbruster, D., Ringhofer, C., and Jo, T.C. 2004. Continuous models for production flows. *Pages 4589–4594 of: Proceeding of the 2004 American Control Conference.*

Asmundsson, J.M., Rardin, R.L., and Uzsoy, R. 2003. *Tractable nonlinear capacity models for production planning part I: Modeling and formulations.* Research Report, Laboratory for Extended Enterprises at Purdue, School of Industrial Engineering, Purdue University.

Asmundsson, J.M., Rardin, R.L., and Uzsoy, R. 2004. *Tractable nonlinear capacity models for production planning part II: Implementation and computational experiments.* Research Report, Laboratory for Extended Enterprises at Purdue, School of Industrial Engineering, Purdue University.

Asmundsson, J.M., Rardin, R.L., and Uzsoy, R. 2006. Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. *IEEE Transactions on Semiconductor Maufacturing*, **19**(1), 95–111.

Axsäter, S. 1981. Aggregation of product data for hierarchical production planning. *Operations Research*, **29**(4), 744–756.

Bechte, W. 1988. Theory and practice of load-oriented manufacturing control. *International Journal of Production Research*, **26**(3), 375–395.

Bertrand, J.W.M. 1983. The effect of workload dependent due-dates on job shop performance. *Management Science*, **29**(7), 799–816.

BERTRAND, J.W.M., AND WIJNGAARD, J. 1985. The structuring of production control systems. *International Journal of Operations and Production Management*, **6**(2), 5–20.

BERTRAND, J.W.M., AND WORTMANN, J.C. 1981. *Production control and information systems for component-manufacturing shops.* Amsterdam: Elsevier.

BERTRAND, J.W.M., WORTMANN, J.C., AND WIJNGAARD, J. 1990. *Production control: A structural and design oriented approach.* Amsterdam: Elsevier.

BILLINGTON, P.J., MCCLAIN, J.O., AND THOMAS, L.J. 1983. Mathematical programming approaches to capacity-constrained MRP systems: Review, formulation and problem reduction. *Management Science*, **29**, 1126–1141.

BITRAN, G.R., AND HAX, A.C. 1977. On the design of hierarchical production planning systems. *Decision Sciences*, **8**, 28–55.

BREITHAUPT, J.W., LAND, M., AND NYHUIS, P. 2002. The workload control concept: Theory and practical extensions of load oriented order release. *Production Planning and Control*, **13**(7), 625–638.

BUZACOTT, J.A., PRICE, S.M., AND SHANTHIKUMAR, J.G. 1992. Service level in multistage MRP and base stock controlled production systems. *Pages 445–463 of:* FANDEL, G., GULLEDGE, T., AND JONES, A. (eds), *New Directions for Operations Research in Manufacturing.* New York: Springer-Verlag.

CHANG, F.C.R. 1994. A study of factors affecting due-date predictability in a simulated dynamic job shop. *Journal of Manufacturing Systems*, **13**(6), 389–406.

CHUNG, C.H., AND KRAJEWSKI, L.J. 1984. Planning horizons for master production scheduling. *Journal of Operations Management*, **4**(4), 389–406.

CHUNG, C.H., AND KRAJEWSKI, L.J. 1986. Replanning frequencies for master production schedules. *Decision Sciences*, **7**, 263–273.

CHUNG, C.H., CHEN, I.J., AND CHENG, G.L.Y. 1988. Planning horizons for multi-item hierarchical production scheduling problems: A heuristic search procedure. *European Journal of Operational Research*, **37**, 368–377.

CLARK, A.J., AND SCARF, H. 1960. Optimal policies for a multi-echelon inventory problem. *Management Science*, **50**(12), 1782–1790.

COASE, R.H., AND FOWLER, R.F. 1937. The pig-cycle in Great Britain: An explanation. *Economica*, **4**(13), 55–82.

CONWAY, R.W., MAXWELL, W.L., AND MILLER, L.W. 1967. *Theory of scheduling*. London: Addison-Wesley.

DANTZIG, G.B., AND WOLFE, P. 1963. Decomposition principle for linear programs. *Operations Research*, **8**, 101–111.

DE KOK, A.G., AND FRANSOO, J.C. 2003. Planning supply chain operations: Definition and comparison of planning concepts. *Pages 597–675 of:* DE KOK, A.G., AND GRAVES, S.C. (eds), *Handbook in Operations Research and Management Science, Volume 11: Design and Analysis of Supply Chains*. Amsterdam: Elsevier.

DIKS, E.B., AND DE KOK, A.G. 1998. Optimal control of a divergent multi-echelon inventory system. *European Journal of Operational Research*, **111**(1), 75–97.

ENNS, S.T. 1995. A dynamic forecasting model for job shop flowtime prediction and tardiness control. *International Journal of Production Research*, **33**(5), 2045–2057.

ENNS, S.T. 2001. MRP performance effects due to lot size and planned lead time settings. *International Journal of Production Research*, **39**(3), 461–480.

ENNS, S.T., AND SUWANRUJI, P. 2004. Workload responsive adjustment of planned lead times. *Journal of Manufacturing Technology Management*, **15**(1), 90–100.

FORRESTER, J.W. 1980. *Industrial dynamics*. 10^{th} edn. Massachusetts: The M.I.T Press.

FRANSOO, J.C., SRIDHARAN, V., AND BERTRAND, J.W.M. 1995. A hierarchical approach for capacity coordination in multiple products single-machine production systems with stationary stochastic demands. *European Journal of Operational Research*, **86**, 57–72.

GARTNER. 2006. *Magic quadrant for supply chain planning in process manufacturing industries,* $1H06$. www.gartner.com, last seen on 30.04.2006.

GRAVES, S.C. 1982. Using lagrangean techniques to solve hierarchical production planning problems. *Management Science*, **28**(3), 260–275.

GRAVES, S.C. 1986. A tactical planning model for a job shop. *Operations Research*, **34**(4), 522–533.

Hackman, S.T., and Leachman, R.C. 1989. A general framework for modeling production. *Management Science*, **35**(4), 478–495.

Hax, A.C., and Meal, H.C. 1975. Hierarchical integration of production planning and scheduling. *Pages 53–69 of:* Geisler, M.A. (ed), *Logistics.* Amsterdam: Elsevier.

Hopp, W.J., and Spearman, M.L. 2000. *Factory physics: Foundations of manufacturing management.* 2^{nd} edn. New York: McGraw-Hill.

Hopp, W.J., and Sturgis, M.L.R. 2000. Quoting manufacturing due dates subject to a service level constraint. *IIE Transactions*, **32**, 771–784.

Hoyt, J. 1978. Dynamic lead times that fit today's dynamic planning (Q.U.O.A.T lead times). *Production and Inventory Management*, **19**, 63–72.

Hwang, S., and Uzsoy, R. 2005. *A single stage multi-product dynamic lot sizing model with work in process and congestion.* Research Report, Laboratory for Extended Enterprises at Purdue, School of Industrial Engineering, Purdue University.

IBM. 1972. Manufacturing activity planning. *In: Communications Oriented Production Information and Control System, Volume 5.* White Plains, New York: IBM Corp.

Kanet, J.J. 1982. Towards understanding lead times in MRP systems. *Production and Inventory Management*, **3**rd **Quarter**, 1–14.

Kanet, J.J. 1986. Towards a better understanding of lead times in MRP systems. *Journal of Operations Management*, **6**, 305–316.

Karmarkar, U.S. 1987. Lot sizes, lead times and in-process inventories. *Management Science*, **33**(3), 409–418.

Karmarkar, U.S. 1989. Capacity loading and release planning with work-in-progress (WIP) and leadtimes. *Journal of Manufacturing Operations Management*, **2**, 105–123.

Karmarkar, U.S. 1993. Manufacturing lead times, order release and capacity loading. *Pages 287–329 of:* Graves, S.C., Rinnooy Kan, A.H.G., and Zipkin, P.H. (eds), *Handbook in Operations Research and Management Science, Volume 4: Logistics of Production and Inventory.* Amsterdam: Elsevier.

Karmarkar, U.S., Kekre, S., and Kekre, S. 1985. Lot sizing in multi-item multi-machine job shops. *IIE Transactions*, **17**, 290–298.

KINGSMAN, B.G., TATSIOPOULOS, I.P., AND HENDRY, L.C. 1989. A structural methodology for managing manufacturing lead times in make-to-order companies. *European Journal of Operational Research*, **40**, 196–209.

KLEIJNEN, J., AND VAN GROENENDAAL, W. 1992. *Simulation: A statistical perspective.* Chichester: Wiley.

KOHLER-GUDUM, C.K., AND DE KOK, A.G. 2001. *A safety stock adjustment procedure to enable target service levels in simulation of generic inventory systems.* Tech. rept. BETA Working Paper 71. Technische Universiteit Eindhoven, The Netherlands.

LAMBRECHT, M.R., AND VANDAELE, N.J. 1996. A general approximation for the single product lot sizing model with queueing delays. *European Journal of Operational Research*, **95**, 73–88.

LAMBRECHT, M.R., MUCKSTADT, J.A., AND LUYTEN, R. 1984. Protective stocks in multi-stage production systems. *International Journal of Production Research*, **22**(6), 1001–1025.

LAND, M.J., AND GAALMAN, G.J.C. 1998. The performance of workload control concepts in job shops: Improving the release method. *International Journal of Production Economics*, **56/57**, 347–364.

LAW, A.M., AND KELTON, W.D. 2000. *Simulation modelling and analysis.* 3^{rd} edn. New York: McGraw-Hill.

LEONG, G.K., OLIFF, M.D., AND MARKLAND, R.E. 1989. Improved hierarchical production planning model. *Journal of Operations Management*, **8**(2), 90–114.

LIVIO, M. 2002. *The golden ratio: The story of Phi, the world's most astonishing number.* New York: Broadway Books.

MATHER, H., AND PLOSSL, G.W. 1978. Priority fixation versus throughput planning. *Production and Inventory Management*, **3^{rd} Quarter**, 27–51.

MCKAY, K.N., SAFAYENI, F.R., AND BUZACOTT, J.A. 1995. A review of hierarchical production planning and its applicability for modern manufacturing. *Production Planning and Control*, **6**(5), 384–394.

MEYR, H., ROHDE, J., AND WAGNER, M. 2000a. Architecture of selected APS. *Pages 241–249 of:* STADTLER, H., AND KILGER, C. (eds), *Supply chain management and advanced planning: Concepts, models, software and case studies.* Heidelberg: Springer-Verlag.

MEYR, H., WAGNER, M., AND ROHDE, J. 2000b. Structure of advanced planning systems. *Pages 75–77 of:* STADTLER, H., AND KILGER, C. (eds), *Supply chain management and advanced planning: Concepts, models, software and case studies.* Heidelberg: Springer-Verlag.

MISSBAUER, H. 2002. Aggregate order release planning for time-varying demand. *International Journal of Production Research*, **40**(3), 699–718.

MISSBAUER, H. 2006. *Models of the transient behaviour of production units to optimize the aggregate material flow.* Working paper, University of Innsbruck, Austria.

MOLINDER, A. 1997. Joint optimization of lot-sizes, safety stocks and safety lead times in an MRP system. *International Journal of Production Research*, **35**(4), 983–994.

NEGENMAN, E.G. 2000. *Material coordination under capacity constraints.* Ph.D. thesis, Technische Universiteit Eindhoven.

NEUTS, M.F. 1981. *Matrix-geometric solutions in stochastic models: An algorithmic approach.* Baltimore: The Johns Hopkins University Press.

PAHL, J., VOSS, S., AND WOODRUFF, D.L. 2005. Production planning with load dependent lead times. *4OR*, **3**, 257–302.

PLOSSL, G.W. 1988. Throughput time control. *International Journal of Production Research*, **26**(3), 493–499.

QUINTIQ. 2007. *http://www.quintiq.com.* last seen on January 12, 2007.

RAAYMAKERS, W.H.M., BERTRAND, J.W.M., AND FRANSOO, J.C. 2000. The performance of workload rules for order acceptance in batch chemical manufacturing. *Journal of Intelligent Manufacturing*, **11**, 217–228.

RAMASWAMI, V., AND LATOUCHE, G. 1986. A general class of Markov processes with explicit matrix-geometric solutions. *OR Spectrum*, **8**, 209–218.

RIAÑO, G. 2002. *Transient behavior of stochastic networks: Application to production planning with load-dependent lead times.* Ph.D. thesis, Georgia Institute of Technology.

ROHDE, J., AND WAGNER, M. 2000. Master Planning. *Pages 117–134 of:* STADTLER, H., AND KILGER, C. (eds), *Supply chain management and advanced planning: Concepts, models, software and case studies.* Berlin: Springer-Verlag.

RUEFLI, T.W. 1971. A generalized goal decomposition model. *Management Science*, **17**(8), 505–518.

SCHNEEWEISS, C. 1995. Hierarchical structures in organizations: A conceptual framework. *European Journal of Operational Research*, **86**, 4–31.

SCHNEEWEISS, C. 1999. *Hierarchies in distributed decision making.* Berlin: Springer-Verlag.

SCHNEEWEISS, C., AND SCHRÖDER, H. 1992. Planning and scheduling the repair shops of the Deutsche Lufthansa AG: A hierarchical approach. *Production and Operations Management*, **1**(1), 4–31.

SELÇUK, B., FRANSOO, J.C., AND DE KOK, A.G. 2006a. The effect of updating lead times on the performance of hierarchical planning systems. *International Journal of Production Economics*, **104**(2), 427–440.

SELÇUK, B., ADAN, I.J.B.F, DE KOK, A.G., AND FRANSOO, J.C. 2006b. *An explicit analysis of the lead time syndrome: Stability condition and performance evaluation.* Working paper, Technische Universiteit Eindhoven, The Netherlands.

SELÇUK, B., FRANSOO, J.C., AND DE KOK, A.G. 2006c. *Supply chain operations planning with load dependent planned lead times.* Working paper, Technische Universiteit Eindhoven, The Netherlands.

SELÇUK, B., FRANSOO, J.C., AND DE KOK, A.G. 2006d. *Work-in-process clearing in supply chain operations planning.* To appear in IIE Transactions.

SHAPIRO, J.F. 1993. Mathematical programming models and methods for production planning and scheduling. *Pages 371–443 of:* GRAVES, S.C., RINNOOY KAN, A.H.G., AND ZIPKIN, P.H. (eds), *Handbook in Operations Research and Management Science, Volume 4: Logistics of Production and Inventory.* Amsterdam: Elsevier.

SIMCHI-LEVI, D., KAMINSKY, P., AND SIMCHI-LEVI, E. 2003. *Designing and managing the supply chain: Concepts, strategies, and case studies.* 2^{nd} edn. New York: McGraw-Hill.

SPITTER, J.M., HURKENS, C.A.J., DE KOK, A.G., NEGENMAN, E.G., AND LENSTRA, J.K. 2005a. Linear programming models with planned lead times. *European Journal of Operational Research*, **163**, 706–720.

SPITTER, J.M., DE KOK, A.G., AND DELLAERT, N.P. 2005b. Timing production in LP models in a rolling schedule. *International Journal of Production Economics*, **93-94**, 319–329.

Stadtler, H., and Kilger, C. 2000. *Supply chain management and advanced planning: Concepts, models, software and case studies.* Heidelberg: Springer-Verlag.

Tatsiopoulos, I.P., and Kingsman, B.G. 1983. Lead time management. *European Journal of Operational Research*, **14**, 351–358.

Van Leeuwaarden, J.S.H., and Winands, E.M.M. 2005. *Quasi-birth-and-death processes with an explicit rate matrix.* To appear in Stochastic Models.

Van Ooijen, H.P.G., and Bertrand, J.W.M. 2001. Economic due-date setting in job-shops based on routing and workload dependent flow time distribution functions. *International Journal of Production Economics*, **74**(1), 261–268.

Vig, M.M., and Dooley, K.J. 1991. Dynamic rules for due date assignment. *International Journal of Production Research*, **29**(7), 1361–1377.

Vollmann, T.E., Berry, W.L., and Whybark, D.C. 1997. *Manufacturing planning and control systems.* 4^{th} edn. New York: McGraw-Hill.

Voss, S., and Woodruff, D.L. 2003. *Introduction to computational optimization models for production planning in a supply chain.* Heidelberg: Springer-Verlag.

Wein, L.M. 1991. Due-date setting and priority sequencing in a multiclass M/G/1 queue. *Management Science*, **37**(7), 834–850.

Wight, O. 1970. Input/Output control: A real handle on lead time. *Production and Inventory Management*, **3^{rd} Quarter**, 9–31.

WolframMathWorld. 2007. *http://mathworld.wolfram.com/GoldenRatio.html.* last seen on January 12, 2007.

Yano, C.A. 1987. Setting planned lead times in serial production systems with tardiness costs. *Management Science*, **33**(1), 95–106.

Zäpfel, G. 1996. Production planning in the case of uncertain individual demand: Extension for an MRPII concept. *International Journal of Production Economics*, **46/47**, 153–164.

Zäpfel, G., and Missbauer, H. 1993. New concepts for production planning and control. *European Journal of Operational Research*, **67**, 297–320.

ZIJM, W.H.M. 2000. Towards intelligent manufacturing planning and control systems. *OR Spectrum*, **22**, 313–345.

ZIJM, W.H.M., AND BUITENHEK, R. 1996. Capacity planning and lead time management. *International Journal of Production Economics*, **46/47**, 165–179.

ZORYK-SCHALLA, A.J. 2001. *Modeling of decision making processes in supply chain planning software: A theoretical and empirical assessment of i2 tradematrix.* Ph.D. thesis, Technische Universiteit Eindhoven.

Glossary

APS	:	Advanced Planning Systems
ATP	:	Available to Promise
BOM	:	Bill of Materials
BOP	:	Bill of Processes
CFL	:	Capacitated Fixed Lead Time
CON	:	Constant Allowance
DRP	:	Distribution Requirements Planning
DSS	:	Decision Support Systems
ERP	:	Enterprise Resources Planning
FCFS	:	First Come First Serve
GFC	:	Goods Flow Control
HP	:	Hierarchical Planning
HPP	:	Hierarchical Production Planning
JIQ	:	Jobs in Queue
JIT	:	Just in Time
LTN	:	Long-Term Nonlinear
MES	:	Manufacturing Execution Systems
MIP	:	Mixed-Integer Programming
MP	:	Mathematical Programming
MRP	:	Material Requirements Planning
MRPII	:	Manufacturing Resources Planning
NOP	:	Number of Operations
QBD	:	Quasi Birth and Death
PPW	:	Processing Plus Waiting Time
RKM	:	Regular Knapsack Method
SCM	:	Supply Chain Management
SCOP	:	Supply Chain Operations Planning
SCP	:	Supply Chain Planning
STN	:	Short-Term Nonlinear
TL	:	Traditional Linear
TWK	:	Total Work Content

Summary

Dynamic Performance of Hierarchical Planning Systems:
Modeling and Evaluation with Dynamic Planned Lead Times

Within the last few decades, supply chain practitioners have faced major challenges in planning manufacturing and logistics operations that are confronted with a high degree of uncertainty as a result of highly dynamic market conditions. At the same time, there has been a growing awareness on the potential applications of rapid advances in data processing and communication technology. As a result, the attention for new concepts and solution methodologies has increased not only in business management but also in scientific community. Integration of various business units and improved coordination of material and information flows along the supply chain have become essential for the companies to stay competitive in the market. In response to these conceptual requirements, advanced planning systems have been introduced, which are constructed along the principles of hierarchical planning.

Hierarchical planning has been a predominant mode for production planning both in academic research and in industrial practice. It is a management philosophy that is based on the decomposition of a large complex planning problem into small and manageable subproblems. Since the late seventies, the research on hierarchical planning has evolved in various directions mainly concentrating on perfect aggregation and disaggregation issues, and the efficiency of hierarchical decomposition with respect to monolithic models in static (mostly deterministic) settings. There have been recent conceptual advances emphasizing the coordination of different production units and different decision functions both from a material flow and from an information flow perspective. However, the research on the dynamic performance of hierarchical planning systems is quite scarce.

The word *dynamic* refers to *plan-execute-feedback-(re)plan* cycle. That is, the planning process is described as a series of decisions taken consecutively in time, and additionally, the planning parameters are subject to changes during

the course of time. Accordingly, the performance evaluation is conducted in a dynamic setting. The performance of the planning decisions given for a single problem instance is not only evaluated based on the status information at the time of decision, but based on their effects on the actual system status changing through time, about which exact information is not available at the time of decision. This aspect becomes essential if there is considerable variability and uncertainty in the demand and production processes. From a practical point of view, rolling horizons are used to keep the system status up-to-date. For an APS that is based on data from an ERP system, all the relevant information e.g. stocks, work-in-process, etc. that is needed to update the plan is continuously available.

There is a positive duration of time between the moment that an order is released to its production unit and that order is available in its stock point, which is referred to as *flow time*. In planning the release of orders, the flow time is represented by a parameter, which is referred to as *planned lead time*. In coordinating the flow of materials in a supply chain, planned lead times are indispensable. The traditional approach is to consider the planned lead times as fixed inputs, exogenous to the planning system. However, flow times are generally not fixed, and depend on various factors such as the level of process uncertainty, the workloads in the production units, capacity flexibility, and the sizes of the released orders. It can therefore be argued that status feedback can be used to anticipate the flow times of order releases. It is interesting to consider updating the planned lead times regularly in order to represent the dynamic characteristics of flow times in the planning system.

In this thesis, our objective is to shed some formal light on the dynamic performance of hierarchical planning systems, where our focus is on the coordination of the flow of materials in a supply chain using dynamic planned lead times. The hierarchical planning systems constructed in this thesis are mainly composed of three different decision levels:

- *Tactical Planning*: Planned lead times for each item are determined at this level as integer multiples of a period.
- *Operational Planning*: An MP formulation is provided to decide on periodic order releases, production quantities, and stock levels for each item to satisfy periodic demand forecasts. The objective is to minimize material holding costs given that a target service level (e.g., demand fill rate) is achieved.
- *Operational Scheduling*: Detailed, execution related decisions such as schedule of released orders at each production unit based on the given planned

lead times, and quantities of materials to be loaded to the shop floor at each production unit are given in a decentralized manner.

The performance is measured along two lines:

- *External*: Expressed in terms of the average (periodic) costs such that a predefined customer service level is met.
- *Internal*: Measures the level of consistency between the higher and the lower level planning outcomes.

Related to the use of status feedback and the integration of different decision levels within the planning hierarchy, three different factors that determine the dynamic performance are considered. These are:

- The *frequency* of updating the planned lead times.
- *Anticipation* on the characteristics of the production processes, which are controlled by lower level planning decisions.
- The type of *coupling* between the higher and the lower level decision models in the planning hierarchy.

In the thesis, we first describe the performance consequences of updating the planned lead times. Then, we model dynamic planned lead times such that the relationship between workload levels, throughput quantities, and flow times is considered to realize effectiveness in the flow of materials within a supply chain.

As a first step, using simulation, we show the effects of updating the planned lead times in a multi-stage production-inventory situation. A two-stage serial supply chain is considered, where only the final product planned lead times are updated based on exponentially smoothed averages of the history of actual flow times. Exponential smoothing is considered as a relevant method in estimating the flow times due to a high level of correlation between the flow times of consecutive orders especially when the scheduling discipline is FCFS. Our results indicate that frequently updating the planned lead times leads to erratic order releases with large variation in inventory levels and very long planned lead times. This phenomenon has conceptually been defined as *lead time syndrome* in the literature, and we provide a formal analysis concentrating on the update frequency and the anticipation on production capacity.

We enhance the discussion on the lead time syndrome by providing an analytical evaluation of the phenomenon. A single-stage, single-item produce-to-order situation is considered with the order releases sensitive to the planned lead time, which is determined according to the number of jobs present in the system. The situation is modeled by a two-dimensional Markov process that is solved by using the matrix-geometric methods. Analytical results on the utilization level and the variability in the system are presented in relation to various design parameters such as the update frequency, and the degree with which the planning system responds to changes in the planned lead time. We have achieved closed form solutions yielding insights that the static utilization level is retained in the dynamic case irrespective of the update frequency when the response function is symmetric, the average backlog size is bigger for higher update frequencies, and on average, jobs spend more time in the system in the dynamic case.

It is shown that updating the planned lead times in planning and coordinating the flow of materials in a supply chain is a challenging task. Naive approaches based on exponential smoothing of realized order flow times, or simple workload dependent rules do not work. In a hierarchical planning system employed in a dynamic setting, the planned lead times that are updated at a higher-level and used to release the orders at a lower-level may cause erratic order releases and increased congestion in the production unit. Thus, there is need for developing more advanced tools to consider the dynamic behavior of production processes in releasing the orders. For this purpose, we use the concept of clearing function to anticipate the flow times of planned order releases, and determine appropriate production quantities at the operational planning level. We first provide a detailed understanding on the relative effects of different clearing functions when the planned lead times are fixed. A single-item produced in a production unit and kept in a stock point facing a stochastic non-stationary demand is considered. Using simulation, clearing functions arising from different modeling approaches are tested based on the internal and the external performance measures. The results indicate that modeling the clearing of WIP should be based on the short-term operational dynamics of the production unit.

Following, we provide insights into the effectiveness of updating the planned lead times of a supply chain in a hierarchical planning context using the clearing function concept. A two-stage serial supply chain is considered, where the capacity loading decisions are separated from the order release decisions, and depend on the hierarchical coupling mechanism. Final product demand is non-stationary, and follows a seasonal pattern. A parameter ε is introduced in modeling a piecewise-linear approximation of the clearing function. Through ε, various anticipation approaches on the production processes can be implemented. Simulation experiments are performed for dynamic and fixed planned lead times under changing demand uncertainty, clearing structure and hierarchical coupling. The results indicate that, in conjunction with the concept of clearing, updating the planned lead times provides the flexibility under fluctuating demand conditions, and generates less costly solutions.

The results presented throughout the thesis indicate the need for further research about planning supply chain operations through dynamic and adaptive decision tools. We suggest some ideas for future research topics such as extending the analysis on the lead time syndrome to multi-stage production-inventory situations, enhancing the discussion on realistic clearing functions towards optimal clearing functions by further elaboration of the parameter ε, and developing more efficient updating procedures for planned lead times.

Samenvatting

Dynamische Prestatie van Hiërarchische Planning Systemen: *Modellering en Evaluatie met Dynamische Geplande Doorlooptijden*

In de laatste decennia worden eigenaren van logistieke ketens geconfronteerd met grote uitdagingen in productieplanning. Logistieke operaties worden in toenemende mate gekarakteriseerd door een hoge mate van onzekerheid die het gevolg is van dynamische marktcondities. Tegelijkertijd is er een groeiend bewustzijn van de potentile toepassing in de logistiek van de snelle vooruitgang in dataverwerking en communicatietechnologie. Het gevolg is dat de aandacht voor nieuwe planningsconcepten en oplossingsmethodieken sterk zijn toegenomen, niet alleen in het bedrijfsleven, maar ook in de wetenschappelijke wereld. Integratie van verschillende business units en verbeterde materiaalcoördinatie en informatiestromen langs de logistieke keten is voor bedrijven essentieel geworden om in de markt concurrerend te blijven. Als antwoord op deze conceptuele eisen zijn geavanceerde planningssystemen (*Advanced Planning Systems* - APS) ontwikkeld op basis van de principes van hiërarchische productieplanning.

Hiërarchische productieplanning is een planningsconcept dat zowel in het wetenschappelijk onderzoek als in de bedrijfspraktijk op grote schaal wordt toegepast. De gedachte is dat grote complexe planningsproblemen worden gedecomponeerd in kleinere beheersbare subproblemen. Sinds het eind van de jaren zeventig heeft het onderzoek over hiërarchische productieplanning zich met name op technische zaken geconcentreerd, zoals het perfect aggregeren en disaggregeren, en de efficiëntie van gedecomponeerde modellen ten opzichte van monolithische modellen in statische (en meestal deterministische) situaties. Meer recent is er vooruitgang bereikt in vraagstukken die de coördinatie van verschillende productie-eenheden en verschillende beslissingsfuncties bestuderen, zowel vanuit het perspectief van de goederenstroom als de informatiestroom.

Het woord *dynamisch* refereert naar een *plan-uitvoer-terugkoppeling-(her)plan* cyclus. Daarbij is het planningproces beschreven als een serie, in tijd achtereen-

volgende, beslissingen. De planningsparameters zijn onderhevig aan veranderingen. Vandaar dat de prestatie-evaluatie in een dynamische setting dient plaats te vinden. De prestatie van planningsbeslissingen voor een enkel probleemgeval is niet alleen geëvalueerd op basis van de toestandsinformatie op het moment van de beslissing, maar ook op basis van het effect op de feitelijke systeemtoestand die continu verandert en waarover geen exacte informatie beschikbaar is op het beslissingsmoment. Dit aspect is met namen belangrijk is situaties waarin de vraag en de productieprocessen gekarakteriseerd worden door onzekerheid. Wanneer er onderzekerheid in de omgeving is, wordt er in de bedrijfspraktijk meestal een rollende horizon gebruikt om de systeemtoestand up-to-date te houden. Voor een *Advanced Planning Systeem* dat gekoppeld is aan de data uit een ERP-systeem is alle relevante informatie (bijv. voorraden, onderhanden werk, etc.) die nodig zijn om het plan te herzien min of meer continu beschikbaar.

Er is een positieve tijdsduur tussen het moment dat een order is vrijgegeven naar de productie-eenheid en het moment waarop die order beschikbaar is in het voorraadpunt; deze tijd noemen we *flow time*. Bij het plannen van ordervrijgave wordt de *flow time* weergegeven als een besturingsparameter; deze besturingsparameter noemen we *planned lead time* (geplande doorlooptijd). Bij het coördineren van de materiaalstroom in een logistieke keten zijn geplande doorlooptijden onmisbaar. Traditioneel worden de geplande doorlooptijden gezien als vaste invoerwaarden, exogeen aan het planningssysteem. Echter, *flow times* hebben geen vaste waarde en zijn afhankelijk van verschillende factoren zoals de mate van procesonzekerheid, de werklast in de productie-eenheden en de grootte van de vrijgegeven orders. Er kunnen dus argumenten bestaan om status feedback te gebruiken om te anticiperen wat de *flow time* zal zijn van nog vrij te geven orders. Daarnaast is het bovendien interessant om te bezien of de geplande doorlooptijden regelmatig kunnen worden herzien teneinde de dynamische eigenschappen van de *flow time* in het planningssysteem te modelleren.

De doelstelling van dit proefschrift is om een formele analyse te maken van de dynamische prestatie van hiërarchische planningssystemen. Daarbij ligt de nadruk op de coördinatie van de materiaalstroom in een logistieke keten, gebruik makend van dynamische geplande doorlooptijden. De hiërarchische planningssystemen die in dit proefschrift geanalyseerd worden, zijn opgebouwd uit drie verschillende beslissingsniveaus:

- *Tactische planning*: Geplande doorlooptijden worden op dit beslissingsniveau voor elk item bepaald (in een geheeltallig aantal tijdsperioden).

- *Operationele planning*: Een mathematisch-programmeringsformulering is opgesteld om te beslissen over periodieke ordervrijgaven, productiehoeveelheden en voorraadniveaus voor elk item om aan periodieke voorspellingen van de vraag te voldoen. De doelstelling is om voorraadkosten te minimaliseren gegeven dat een bepaalde service niveau behaald wordt.

- *Operationele scheduling*: Gedetailleerde, uitvoeringsgerelateerde beslissingen, zoals het plannen van vrijgegeven orders bij elke productie-eenheid; dit is gebaseerd is op de reeds op het tactische niveau vastgestelde geplande doorlooptijden en de op operationele planningsniveau vastgestelde materiaalhoeveelheden die vrijgegeven worden naar de productievloer. Deze vrijgave gebeurt gedecentraliseerd voor elke productie-eenheid.

De prestatie wordt volgens twee lijnen gemeten:

- *Extern*: Uitgedrukt in termen van de gemiddelde (periodieke) kosten zodanig dat een van te voren vastgestelde leverbetrouwbaarheid wordt behaald.
- *Intern*: De mate van consistentie tussen het hogere en lagere planningsniveau resultaat.

In verband met het gebruik van terugkoppeling van toestandsinformatie en de integratie van verschillende beslissingsniveaus binnen de planningshiërarchie worden drie verschillende factoren in beschouwing genomen die de dynamische prestatie benvloeden. Deze factoren zijn:

- De *frequentie* van het herzien van de doorlooptijden.
- *Anticipatie* op de karakteristieken van de productieprocessen die beheerst worden door planningsbeslissingen op een lager niveau.
- Het type van *koppeling* tussen het hogere en lagere niveau beslissingsmodellen in de planningshiërarchie.

In het proefschrift beschrijven we eerst de gevolgen van het updaten van de geplande doorlooptijden. Daarna modelleren we dynamische geplande doorlooptijden zodanig dat de relatie tussen de bezettingsgraden, doorzethoeveelheden en de *flow times* wordt meegenomen om tot een betere prestatie te komen van de materiaalstroom.

Als een eerste stap, gebruik makend van simulatie, laten we de effecten zien van het updaten van de geplande doorlooptijden in een meerniveau productie-voorraadsituatie. Een twee-niveau seriële logistieke keten wordt in beschouwing genomen, waarbij alleen de geplande doorlooptijden van het eindproduct worden herzien op basis van exponentieel gedempte gemiddeldes van historische gerealiseerde *flow times*. We gebruiken *exponential smoothing* om de *flow times* te schatten omdat er een sterke correlatie bestaat tussen de *flow times* van opeenvolgende orders, vooral indien de orders volgens FCFS worden gepland. Onze resultaten wijzen erop dat frequente herziening van geplande doorlooptijden leidt tot onregelmatige ordervrijgaven met hoge variatie van voorraadniveaus en hele lange geplande doorlooptijden. Dit fenomeen wordt in de literatuur conceptueel gedefinieerd als *doorlooptijdsyndroom*, en we geven een formele analyse die zich concentreert op de frequentie van herziening en het anticiperen op productiecapaciteit.

We breiden de discussie over het doorlooptijdsyndroom uit met een analytische evaluatie van het verschijnsel. Een één-niveau, één-item productie op order situatie is beschouwd waarbij de ordervrijgave gevoelig is voor de geplande doorlooptijd. De situatie is gemodelleerd aan de hand van een twee-dimensionaal Markov-proces dat opgelost is door gebruik te maken van matrix-geometrische methoden. Analytische resultaten van de bezettingsgraad en de variabiliteit in het systeem worden gepresenteerd in relatie tot verschillende ontwerpparameters zoals de herzieningsfrequentie en de mate waarin het planningssysteem reageert op veranderingen in de geplande doorlooptijd. We bereiken oplossingen in gesloten vorm die laten zien dat de bezettingsgraad wordt gehandhaafd in een dynamische situatie onafhankelijk van de herzieningsfrequentie wanneer de responsfunctie symmetrisch is, dat de gemiddelde orderachterstand groter wordt voor hogere herzieningsfrequenties, en dat gemiddeld orders langer in het systeem verblijven in de dynamische situatie.

We laten zien dat het updaten van de geplande doorlooptijden in het plannen en coordineren van de materiaalstroom in een logistieke keten een lastige taak is. Naïeve benaderingen die zijn gebaseerd op het exponentieel dempen van gerealiseerde order *flow times*, of eenvoudige werklastafhankelijke regels werken niet. In een hiërarchisch planningssysteem onder dynamische omstandigheden kunnen de geplande doorlooptijden die op een hoger niveau worden gewijzigd en worden gebruikt om op een lager niveau de vrijgave te regelen, leiden tot zeer onregelmatige ordervrijgaves en toenemende congestie op de werkvloer. Er is dus een noodzaak tot het ontwikkelen van meer geavanceerde hulpmiddelen die het dynamische gedrag van productieprocessen meenemen bij de ordervrijgavebeslissing. Hiertoe gebruiken we het *clearing function* concept om op de *flow times* van de geplande ordervrijgaves te anticiperen, en om de goede productiehoeveelheden op het operationele niveau vast te stellen. Hierbij geven we initieel inzicht in het effect van verschillende *clearing functions* bij vaste geplande doorlooptijden. We beschouwen een enkel item geproduceerd in een productie-eenheid en op voorraad gehouden in een voorraadpunt dat onderhavig is aan stochastische niet-stationaire vraag. Gebruikmakend van simulaties zijn *clearing functions* getest op basis van de interne en externe prestatiematen die verschijnen vanuit verschillende modelleringmethodieken. De resultaten wijzen erop dat modellering van de *clearing* van onderhanden werk gebaseerd moet zijn op de korte termijn operationele dynamiek van de productie-eenheid.

Aansluitend verschaffen we inzichten in de effectiviteit van herziening van de geplande doorlooptijden van een logistieke keten in een hiërarchische planningscontext, daarbij gebruikmakend van het *clearing function* concept. Een twee-niveau seriële logistieke keten is in beschouwing genomen waar de capaciteitsbeladingsbeslissingen onderscheiden worden van ordervrijgavebeslissingen en afhankelijk zijn van het hiërarchische koppelingsmechanisme. Eindproductvraag is niet stationair en heeft een seizoenspatroon. De parameter ε is geïntroduceerd in het modelleren van een piecewise lineaire benadering van de clearing function. Door ε kunnen verschillende anticipatie-methodieken voor productieprocessen worden geïmplementeerd. Simulatie-experimenten zijn uitgevoerd voor dynamische en vaste geplande doorlooptijden onder veranderende vraagonzekerheid, *clearing* structuur en mate van hiërarchische koppeling. De resultaten wijzen erop dat in combinatie met het concept van *clearing*, herziening van de geplande doorlooptijden flexibiliteit verschaft onder fluctuerende vraagcondities en minder kostbare oplossingen genereert.

Onze resultaten geven de noodzaak aan van additioneel onderzoek over operationele planning van logistieke ketens door dynamische en adaptieve beslissingsgereedschappen. We suggereren enige ideeën voor verder onderzoek zoals de uitbreiding van de analyse van het doorlooptijdsyndroom naar meer-niveau productie-voorraad situaties, versterking van de discussie over realistische *clearing functions* naar optimale *clearing functions* door verdere uitwerking van de parameter ε en ontwikkeling van herzieningsprocedures.

About the Author

Barış Selçuk was born in Niğde, Turkey (TR), on June 28, 1978. He graduated from Science High School in Mersin (TR), in 1995, and he continued his university education in the Department of Industrial Engineering at Bilkent University, Ankara (TR). In year 2000, he obtained the degree of Bachelor of Science in Industrial Engineering, and started to work as Business Analyst at Corbuss in Istanbul (TR), where he gained experience and knowledge on high-tech service industry. After more than a year, in September 2001, his growing interests on supply chain management made him pursue the degree of Master of Science in Industrial Engineering at Bilkent University, Ankara (TR). He also worked as Teaching Assistant in the same department till he completed his master thesis, entitled "Facility Location Decisions under Vehicle Routing Considerations", in December 2002.

Thereafter, Barış was enrolled to the PhD program of Beta Research School at Technische Universiteit Eindhoven, Eindhoven, The Netherlands (NL). Since then, he has been working as a Research Assistant in the Department of Technology Management at the same university, and he has conducted research on dynamic performance of hierarchical planning systems under the supervision of prof.dr. A.G. De Kok and prof.dr.ir. J.C. Fransoo. This PhD dissertation is an output of the research activities conducted between January 2003 and February 2007. His research interests include hierarchical planning systems, lead time management, supply chain planning, and queueing theory.

On February 22, 2007, Barış will defend his PhD dissertation at Technische Universiteit Eindhoven. As of February 2007, he will be working as Business Intelligence Analyst at ASML Netherlands B.V. in Veldhoven (NL).