Inventory Management: Modelling Real-life Supply Chains and Empirical Validity

Ton de Kok

Ton de Kok
School of Industrial Engineering, Eindhoven University of Technology, e-mail: a.g.d.kok@tue.nl

Contents

Inventory Management: Modelling Real-life Supply Chains and Empirical Validity 1
Ton de Kok
1 Introduction 4
2 Modelling inventory systems 8
2.1 A brief history of inventory management research 8
2.2 Empirical validity of inventory models 9
2.3 The practice of inventory management: human intervention and correlations 10
2.4 Intervention frequency and inventory system performance 11
2.5 Modelling time between order release and order receipt . . 12
2.6 Conclusion on modelling inventory system 14
3 Single-item single-echelon inventory models 14
3.1 Deterministic demand 15
3.2 Stochastic demand 17
4 Uncapacitated multi-item multi-echelon models 27
4.1 Material availability and stochastic lead times 28
4.2 The example supply chain 29
4.3 Modelling multi-echelon inventory systems 30
4.4 Feasibility of order release quantities 32
4.5 Synchronization and allocation 35
4.6 Decision node structure for the case example 42
4.7 Control policies for divergent MIME systems 47
4.8 Generalized Newsvendor equations for divergent MIME systems 49
4.9 Performance of SBS policies 51
4.10 Empirical validity of SBS policies 54
4.11 Positioning inventory in the supply chain 57
5 Capacitated inventory systems 59
5.1 Feasibility of order release quantities 60
5.2 Comparison of rolling scheduling concepts 62

5.3 Optimal policies for serial MIME systems 62
5.4 Implicit modelling of finite capacity 63
6 Conclusion 64
7 Acknowledgements 65
References 66

1 Introduction

Inventory management has been a core topic of Operations Research since the 1950s. Inventory can be seen as a means to create efficiency in production and distribution: it enables scale by allowing to accumulate demand until a batch quantity can be released that can be produced and shipped efficiently. This role of inventory is of great importance in process industries, where set-up times are considerable. Inventory can be seen as a means to ensure sufficient customer service: as demand is unpredictable we must hold inventory just-in-case, to protect against unexpected surges in demand. This role of inventory is of great importance in retail, as we expect a product to be available off-the-shelf or at our doorstep within 24 hours.

Inventory can also be seen as a symptom of bad management, as waste of capital. Reduction of inventory capital has been high on the priority lists of CEO's over the last four decades. Early 1980's the Just In Time (JIT) philosophy proclaimed zero inventory as the key objective to ensure continuous improvement of processes, leading to less process variability, shorter processing time, smaller production and transportation batches, and higher product yield. In many businesses inventory is a forbidden word. Euphemisms for inventory were introduced, such as buffers, and supermarkets. Despite the continuous efforts to reduce process durations and volatility, zero inventory will remain a mirage, as fundamental uncertainty in demand and supply cannot be eliminated and trading-off efficiency, quality, customer service and cost of inventory capital inevitably yields the need for inventory at various places in global and local supply chains, acting as the lubricant.

The trade-offs to be made have been studied extensively in the inventory management literature. This has lead to optimal inventory control policies for various supply chain structures under various cost assumptions. Clearly, most results are known for the simplest inventory management situation, i.e. a single product at a single location. But both the qualitative and quantitative understanding of this simple inventory management situation is a building block for understanding inventory management in practice, where we have to deal with multiple items at multiple locations.

Thus, inventory control policies are implemented in every ERP (Enterprise Resource Planning) system, e.g. SAP and Oracle, and used at almost every company. ERP systems are the transaction backbone systems of enterprises in which product and process data are stored, and each customer order, production order, and purchase order is tracked and traced. Over the course of a few decades ERP systems have been enriched with planning and control modules that support inventory management, production management, and sales. Despite the availability and use

of inventory control policies in ERP systems, we observe that most of the control-policy-based replenishment proposals are overwritten by manual decisions. Indeed, being an inventory manager or planner, you want to manage and plan, and you can do better than the inventory management system. Unfortunately, it is shown again and again that proper use of inventory management systems yields higher service and lower costs at the same time. We observe that inventory managers have difficulties with the interpretation of unexpected events regarding demand and supply, i.e. distinguishing noise from signal. At the same time we observe that inventory managers have access to relevant information that an ERP system's inventory control module cannot exploit. This calls for the design of an inventory management approach that combines the strength of mathematically rigorously determined inventory control policies and tacit knowledge of human decision makers. This paper is motivated by these observations and builds on 32 years of working in (8 years) and with (the next 24 years) industry, applying and implementing inventory control models.

It is our intention to write a different overview of inventory models, from single item single echelon models to multi-item multi-echelon models, then is mostly provided in text books on Operations Management(e.g. Nahmias & Olsen (2015) and Silver et al. (2016)). We hope that this paper provides complementary knowledge. Instead of starting with inventory models that are tractable from a mathematical point of view, we start from the inventory management problem and the modelling challenges to be faced.

The first section of this paper is devoted to modelling inventory systems, such that these models are empirically valid by proper calibration. Inventory models are abstractions that cannot capture all possible actions to balance supply and demand. But with proper measurement of inventory management performance we can set the parameters in such a way that the customer service is consistently at the right level. We hypothesize that it is better to use mathematically tractable models and appropriately chosen performance measures, then to identify all possible actions under specific circumstances and model these explicitly. We found that many specific actions are focussed on preventing stockouts. Typically, such actions either postpone customer demand or expedite production orders released earlier. Herewith we create correlation between occurrences of high demands and arrivals of production orders that satisfy them. Ignoring this correlation yields considerable underestimates of customer service, while modelling this correlation is mostly mathematically intractable. Thus we propose to measure performance *before specific actions are taken*, which yields the notion of Intervention Independent Performance (IIP) indicators. Clearly, a company must measure the effectiveness of the specific actions taken as well, which yields the notion of Intervention Dependent Performance (IDP) actions. Applying IIP indicators in combination with inventory models in research projects provided an empirical basis for the validity of this approach: in both single-item single-echelon (SISE) situations and multi-item multi-echelon inventory systems we could explain the *quantitative* relationship between capital invested in item inventories and end-item customer service. One should not underestimate the importance of this finding: it provides a scientific basis for the use of inventory mod-

els as studied in OR literature. Here we take the position that mathematical models and their analysis are not science without empirical data supporting the causalities embodied by the model.

The second section discusses SISE models. We show that under linear holding and penalty costs the Newsvendor equation holds for virtually any sensible control policy. The Newsvendor equation states that the non-stockout probability at an arbitrary point in time equals the quotient of penalty cost rate and the sum of holding cost rate and penalty cost rate. We show that inventory management performance is primarily determined by average inventory and order frequency. In our view there should be more emphasis in inventory management education on average inventory levels instead of safety stocks. After all, we pay for the capital tied up in average inventories, not in safety stocks. As capital is tied up in inventory, it is relevant to consider trade-offs from a Return On Investment point of view. We discuss the impact of the change from cost minimization to ROI maximization using the Economic Order Quantity model. We discuss the prerequisites for empirical validity of the basic inventory models. One lesson should stand out here: mathematical analysis must be rigorous. Otherwise it is likely that the resulting control policies do not make any sense to inventory planners, and they are right in that case.

The third section extensively discusses multi-item multi-echelon inventory systems. This discussion is not aiming at a complete overview of the state-of-the-art on multi-echelon inventory system research. Having worked on the subject for over 25 years, we conclude that the emphasis in the scientific literature has primarily been on optimal policies under specific assumptions on the structure of multi-item multi-echelon systems, such as serial, divergent or convergent, (cf. Axsäter (2003) and Song & Zipkin (2003)) and much less on the underlying complexity of general multi-item multi-echelon (MIME) systems. There are no serial systems in practice! At best they are divergent (i.e. each item has a single upstream predecessor, or child) in the form of retail and spare parts distribution networks. Convergent MIME systems, i.e. systems in which each item has at most one parent, are rare, as most companies sell more than one product. In literature, convergent MIME systems are also referred to as (pure) assembly systems. So most of the time supply chains are networks with both embedded divergence and convergence (i.e. an item may have multiple children upstream and multiple parents downstream. Under uncertainty you are continually confronted with the dilemma to allocate item availability among parent items, i.e. the items that use the item under consideration. Allocating less to a particular parent item implies that less is needed of other child items used by this parent item, whereby these child items can be used for other parent items, but then we need other items as well, etc. We assume that orders released to the shopfloor can be executed with 100% due date reliability, provided material (and resource) constraints are taken into account. This implies that we model general MIME systems with constant flow times, i.e. constant times between order release and order receipt in inventory. In order to create a benchmark for control policies for general MIME systems, we formulate necessary conditions for a control policy to yield feasible solutions. Herewith we bridge the gap between mathematical programming formulations of supply chain planning problems that concern the problem to be solved

today, and the stochastic dynamic programming formulations that focus on control policy structures that generate optimal policies, and resulting solutions, over a relevant period of time.

The most-often used planning logic to plan and manage multi-item multi-echelon inventory systems in practice is called Material Requirements Planning, which is abbreviated as MRP I. The main principles of MRP I logic are lead-time offsetting and dependent demand. Starting from the constraint to maintain a safety stock at the end of each future period, and known future (gross) requirements for an item, and outstanding orders, inventory balance equations are used to determine the replenishment quantities in future period. By offsetting the replenishment quantities by the item lead time we obtain planned order quantities. These planned order quantities are translated into so-called dependent demand for child items by multiplying the order quantities by the number of child items needed to make one item. By proper administration we can determine the dependent demand for each item and derive the planned order for each item. For further details on the logic we refer to subsection 4.4. Initially Material Requirements Planning was abbreviated as MRP, but in the 1980s the MRP logic was embedded in an overall framework for planning and control called Manufacturing Resource Planning, which, having the same three-letter-abbreviation, was denoted as MRP II (cf. Vollmann et al. (2005)). MRP I was introduced as a ”killer app” for IBM mainframes in the early 1960s, and promoted by the American Production and Inventory Control Society (APICS) from 1970 onwards. For a historic perspective on MRP I we refer to Wilson (2016). We find that MRP I logic does not pass the test of adhering to material availability constraints. This finding cannot be emphasized often enough, as it explains symptoms like nervousness and expediting. On my return to academia early 1990s I set myself the research objective to determine safety stocks in MRP I. Pursuing this objective, I found that my quest would be in vain, because the MRP I logic is not mathematically sound. MRP I logic turned out to be a logic that generates requirements, but it is not a logic for planning. Planning involves the balancing of demand and supply, knowing that you must take decisions on supply before demand is known. That is why in general MIME systems there is a continual misalignment between demand and supply that is resolved by keeping inventory. But inventory does not always resolve the misalignment, and that is where scarce child item material availability must be allocated among multiple parent items, with the consequences sketched above: a problem mess, a Gordian knot. The concept of Synchronized Base Stock (SBS) policies for operational control of general multi-item multi-echelon inventory systems, is cutting this Gordian knot at the expense of suboptimality (though SBS policies are optimal for divergent systems, and convergent systems). The SBS concept generates a deep insight into the natural decision hierarchy embedded in any general multi-item multi-echelon system. In-depth case studies in the context of MSc thesis projects at companies indicate that the assumption of SBS policies yields empirically valid results, even though none of these companies used SBS policies. The only explanation for this result is that also in multi-item multi-echelon inventory systems inventory performance is driven by average inventories and order frequencies.

The fourth section briefly discusses the additional issues that come with taking into account resource constraints. While for single echelon systems finite capacity is (relatively) easy to do this, this is not the case for multi-echelon systems. I consider the results for serial systems in Janakiraman & Muckstadt (2009) as a milestone in the analysis of capacitated multi-item multi-echelon systems, and at the same time as a clear indication of the challenges ahead of us when trying to tackle this problem for general structures.

Inventory management is a challenging research subject due to its structural complexity, represented by general networks of interacting stockpoints, and the complexity induced by demand and supply uncertainty. The curses of dimensionality prohibit the calculation of optimal policies. I hope that this fact is a reason to pursue more research with great practical relevance. Admittedly, when allowing yourself to write down that something on the left hand side of an "equation" is approximately equal to something on the right hand side, you may be overwhelmed by the possible alternative routes that can be taken towards policies and algorithms. But at the end of the day, applied science should be about reality and reality happens to be complex.

2 Modelling inventory systems

In this section we focus on modelling of inventory systems as opposed to analysis of inventory models. We first provide a brief overview of the inventory management research since 1888. In subsection 2.2 we explain our view on empirical validity of inventory models (and stochastic models in general). We discuss how human interventions have a great impact on the empirical validity of inventory model in subsection 2.3. In subsection 2.4 we propose to develop performance measures that cannot be impacted by human interventions, so that they can be used to validate and calibrate inventory models. In subsection 2.5 we discuss the definition and measurement of time between order release and order completion and receipt, as proper modelling of this time interval is quintessential for the validity of inventory models.

2.1 A brief history of inventory management research

Quantitative models for Inventory Management has been an important subject in the context of Operations Management for about 130 years. The first publication on the subject is by Edgeworth (1888), who studies the Newsvendor problem in the context of a cash balance problem. The derivation of the Economic Order Quantity has been claimed by several different authors from different language backgrounds, but Harris (1913) seems to be the first source. Interesting to note here that the most basic stochastic inventory system has been studied before the most basic deterministic inventory system. The start of Inventory Management research as we know it today

can be attributed to Whitin (1953). The analysis of SISE systems is extensive and seems rather complete. It is a standard subject in IE, OM, and OR curricula, and algorithms to determine control policies have been embedded in inventory management systems, either as standalone applications or as modules of ERP systems. The analysis of multi-item multi-echelon systems has progressed substantially since 1990, but stochastic demand and multiple items to be managed in some coordinated and cost-effective manner implies that we face the curses of dimensionality. There is no hope to find optimal control policies for real-world multi-item multi-echelon systems. But the progress made on the analysis of these complex inventory systems is such that we conjecture that state-of-the-art inventory systems research can support the quantitative analysis of real-world supply chains under demand and supply uncertainty. The emergence of supply chain optimization software over the last two decades, and the extensive documentation of its successful application in literature, support this conjecture. This supply chain optimization software is seen as an example of so-called Advanced Planning and Scheduling (APS) systems. These APS systems are using optimization methods and heuristics to solve various kinds of deterministic planning problems. APS systems are used to solve facility location problems, aggregate planning problems with a horizon of 2-5 years, mid-term planning problems with a horizon of 1-2 years, but also shopfloor scheduling problems with a horizon of a few weeks or days, even. Thus the outcome of most APS systems is a plan. Inventory research based supply chain optimization software typically produces inventory control parameters, explicitly taking randomness in supply and demand into account. These inventory control parameters are used in the APS planning software to guide decision making. Inventory control parameters are safety stocks, safety times, lot sizes, and review periods. For an extensive overview on APS systems and their application, we refer to Kilger et al. (2015).

2.2 Empirical validity of inventory models

Empirical validity is a central concept in science. We adopt here the concept as used in Physics, as inventory models are models describing causality between inputs and outputs quantitatively, similar to Bohr's model of the atom describes emission of quanta of energy as a consequence of energy added, and Newton's model of gravity describes the speed of an object falling towards the earth. In Physics it is generally assumed that measurement errors are normally distributed, so that a confidence interval around the point estimate of the variable under consideration can be determined. A model is empirically valid if the value predicted by the model is within the confidence interval. Experiments are repeated as many times as needed under the same conditions to produce a supposedly identical outcome.

Validation of inventory models cannot be done this way. Firstly, inventory management takes place in reality, not in a lab. Conditions under which the control policies derived from the inventory model are executed cannot be controlled. Secondly, as demand and supply are stochastic by nature, a single experiment concerns

a item's inventory process over a period of time, typically at least half a year. Over this period data are collected from which probability distributions are determined to describe the demand and supply processes. As a probability distribution is fundamentally a concept relating to repeating an experiment infinitely many times under the same conditions, a model error is inevitable. The probability distributions derived have nothing to do with measurement errors, they are elements of the model itself. Thirdly, we cannot repeat experiments, as product life cycles are finite.

In our empirical studies we adopted the following approach: we collected data over a some period of time for a (large) number of different items. We assumed stationary stochastic processes and derived from the data estimates for the mean and standard deviation of the random variables involved. In the context of inventory management models, these random variables are demand per time unit for the situation of periodic review of inventory, order interarrival times and order sizes for the situation of continuous review of inventory, and replenishment order lead time. Furthermore we computed average inventories and average order sizes for each item over the same period of time. Under some control policy, we determine the policy parameters that yield the average inventory measured. Given those policy parameters we compute the customer service according to the model and compare against the actual customer service over the data collection period. So we execute the same experiment for different items.

This approach can be compared with throwing dices 52 times, say, and computing the probability that the resulting outcome is at most 5. According to probability theory the outcome should be $5/6$, but in reality this is not the case. If we execute the experiment with a large number of different dices, we may hope that our model provides an accurate aggregate outcome. When considering multiple different items we weighted the outcomes with the financial turnover of each item. This may be considered adding apples and oranges, but we want to assess the validity of a generic model that can be applied to many different inventory control situations.

When we claim empirical validity of models in the sequel of this paper, it should be understood in the way described above. Clearly, more research is needed to develop a rigorous methodology to validate stochastic models.

2.3 The practice of inventory management: human intervention and correlations

In this section we provide an overview of the state-of-the-art of quantitative inventory management research from a professional perspective. The professional perspective is that of an inventory manager who needs to create a setting in which inventory planners, supply planners, purchasers and expediters can operate effectively. Supply must be balanced with demand to the extent possible and the joint efforts of manager, planners, and expediters represent a giant array of possible means to that, even when demand for products comes unexpectedly. Our starting point is that we cannot model this array of options to respond to unexpected demand (and unexpect-

edly delayed supplies). We need to develop models that *support* the professionals involved in inventory management to exploit their array of options. In this introduction we provide arguments that support the application of the inventory theory developed over more than a century.

When applying inventory models in practice we often found that application of the textbook formulas for safety stocks and customer service provided in e.g. Fogarty & Hoffmann (1983) shows the results are empirically invalid. In many cases this could be explained by the invalidity of the assumptions underlying these textbook formulas and the invalidity could be resolved by relaxing assumptions and rigorous mathematics (cf. De Kok (1991*a*) and Section 2). However, in a substantial number of cases the invalidity could not be resolved. We found that the common denominator of these cases was *human intervention*: planners schedule-in orders when needed, expediters expedite these orders and reallocate materials to production orders to prevent belt stops. In general, human interventions create correlation between supply and demand: if demand is high, lead times are shorter. If demand is high, demand is aligned with supply by negotiating a later delivery date than requested. We stated above that we cannot model all response options. On top of that, taking into account correlation between stochastic demand and stochastic supply is mathematically inhibitive in most cases.

2.4 Intervention frequency and inventory system performance

Based on our extensive involvement in development of models for inventory systems in practice we conclude that, though we cannot model all responses to uncertainty, and we cannot model correlation between demand and supply, we can model *the triggers for non-modelled responses to uncertainty and effect of responses to uncertainty*. The main trigger for non-modelled responses to uncertainty are *projected shortages within the supply lead time*, where projected refers to both as derived from an algorithm, and to tacit knowledge of the professional. The projected shortages can be both immediate shortages and future shortages within the lead time. Assuming we know the events that trigger non-modelled response options, we can link these trigger events to events in the inventory model. In particular we may consider stockout events and customer backlog events as trigger events. Assuming that before such events occur, the inventory system is controlled according to the control policy modelled, we can use the frequency of trigger events in reality to the frequency of stockouts or the probability a customer must wait, respectively, in our model. Assuming that the trigger frequency does not change in the future, we can use this frequency as a performance target in our inventory model to derive the parameters of the inventory control policy.

As each trigger event enacts a response from the planner, expediter, or purchaser, we must link the response enacted to its effect, which should be prevention of a shortage or reduction of customer waiting time. In the case situation that has been the source of the above reasoning, the schedule-in actions resulted into 100% cus-

tomer service with one month of average inventory. No standard inventory model assuming normally or gamma distributed demand per period could explain this. However, by assuming that each schedule-in actions prevented a stockout, we could determine control policies from the number of stockouts per unit time. The empirical validation of the model supported the modelling approach (cf. De Kok (1998) and de Kok (2017)). In this case the effect of the response actions was unambiguous: 100% customer service. In other cases we may find an increase of customer service from 80% according to the inventory model based on historical data about the demand process, supply process and inventory process, to 98% in reality. In that case we need to make additional modelling assumptions to explain the 18% increase in service. Suppose we use the non-stockout probability at an arbitrary point in time as service measure. Then it should be that without exploiting the non-modelled response options, we would have found 80% service, while 90% of the responses was effective, i.e prevented a stockout. Assuming that this 90% measures the capability of the organization to prevent stockouts that cannot be prevented by the routine inventory management rules, and assuming that this capability is the same in the (near) future, we can use this capability estimate to derive the customer service target for the inventory model from the customer service target in reality.

The critical step in this approach is the correct determination of the response trigger frequency. Expediting an order may result in both expedition and postponement of other orders. Just determining which orders are produced or received earlier than planned from an ERP system is not enough, as this may seriously overstate the number of responses. But in most cases it is feasible to develop system support for correct identification of non-modelled responses.

We already mentioned that the response trigger events can be related to events in our inventory model. This leads to the concepts of *intervention-independent* and *intervention-dependent* performance. The intervention-independent performance can be predicted by mathematical analysis of inventory models, while this is impossible for the intervention-dependent performance without additional modelling assumptions as discussed above. Fortunately, we only need to be able to measure the actual intervention-independent performance, e.g. fill rate or non stockout probability, to validate and calibrate our inventory models, so that they can be effectively used to support inventory management professionals. For a detailed discussion of intervention-independent and intervention-dependent performance indicators and modelling manufacturing (and inventory) systems we refer to de Kok (2017). In the remainder of this paper we assume that the performance targets to be achieved relate to intervention-independent performance measures.

2.5 Modelling time between order release and order receipt

Before we conclude this discussion on modelling inventory systems and using data from information systems to effectively validate and calibrate inventory models, we would like to discuss modelling lead times. Over the course of time we find that

the concept lead time as defined in the inventory theory literature differs from the concept lead time in the *professional* literature, and in particular in the APICS literature. In the latter *body of knowledge* lead time is a norm for the flow time as used in planning systems, such as MRP II systems. The flow time of an order is the time that elapses between release of an order and completion of the order. In the *scientific* inventory theory literature the notion of lead time is used to describe the time elapsed between release and completion of an order. Thus the lead time in the scientific body of knowledge on inventory management is the same as the flow time in the professional body of knowledge on inventory management. In this text we use lead time as used in the scientific body of knowledge to denote the time elapsed between order release and order completion (or receipt). In practice inventory management is not just ordering at the right moment in time the right quantity, but also providing an outlook to the rest of the organization on future material availability. Such an outlook can be derived if we make an assumption on the time that elapses between release and completion of future orders. Thus a *norm* is introduced for the lead time of future orders, which is denoted as *nominal lead time* (cf. Graves & Willems (2003)). The nominal lead time can be derived as a sample mean of passed lead times, or as a sample percentile, to include protection against uncertainty in lead times.

In Zipkin (2000) exogenous and endogenous lead times are discussed. Exogenous lead times are modelled as random variables independent of the other stochastic processes. This is the most common way of modelling lead times in inventory theory. Endogenous lead times are lead times that result from the interaction between stochastic demand and finite resources. These are typically discussed in the context of queueing theory. Undoubtedly, lead times in real-life are endogenous. But we should be aware that in most cases the lead time of an order is not determined by its own workload, alone, but by the workload of many other orders. Queueing models typically apply to situations with one or multiple resources and multiple items that must be produced. From the perspective of a single item the lead time is exogenous. Subsequent lead times may be dependent, as in most cases orders do not overtake, but the lead time is not dependent on the individual item order size or order interarrival time. So we aggregate item orders into one or a few job types and apply queueing system analysis to determine an estimate of the lead time. This is similar to aggregating data about item order lead times into an empirical distribution of order lead times. When applying this approach in practice, we found that queueing system models provide a good estimate of the average lead time, but the standard deviation of order lead times is strongly overstated. The former could be explained by the work conservation law in queueing theory (cf. Kleinrock (1965)), the latter could be explained by similar arguments as above: shopfloor scheduling professionals constantly keep track of the progress of production orders, and intervene if the progress stalls for some reason, e.g. by assigning a high priority to delayed orders. We should be aware that in queueing systems lead times just happen as a consequence of the interaction between resource needs over time and resource availability over time. In reality each production order has a due date and this due date allows for feedback mechanisms. In reality not only can we reassign priorities

to orders, we can also reallocate resources to orders without affecting (too much) the due date adherence of other orders.

As stated above, in reality planning systems use nominal lead times to enable insight into the future evolution of inventory, and resource use. This insight is only of use if the nominal lead times are realized. Therefore the nominal lead time determines the due date of an order. And the due date provides a target that is used to ensure the order is completed in time. At the shopfloor equifinality is the name of the game: exploit all possible response options to ensure the order is finished in time. It may be that the nominal lead time is set as the sum of the average lead time as measured in the recent past (or according to a queueing system model) and a safety time that provides the slack to meet the due date without excessive costs. When modelling real-world supply chains with multi-item multi-echelon inventory systems and constant lead times taken from ERP systems, we found that this approach yields empirically valid results. In fact, this empirical validity fostered the above reasoning.

2.6 Conclusion on modelling inventory system

In summary, when modelling real-life systems, we should be aware that in reality inventory management and shopfloor management professionals have an array of response options to ensure timely delivery of customers and timely completion of production orders, respectively. Trying to model these options is a dead-end due to mathematical intractability, and from a conceptual point of view a fundamentally wrong approach. By identifying the response trigger events and linking them to events in the mathematical model, and defining and measuring the appropriate intervention-independent performance measures we can establish a correct link between model and reality. The above arguments and their empirical foundation lead to the conclusion that the inventory models developed over more than a century of inventory management research can be used in practice. In the next sections we discuss the results from inventory research that we think are most important for application in practice. The results selected and presented are based on highly subjective decisions and should not be seen as a definitive selection.

3 Single-item single-echelon inventory models

In this section we discuss the basic SISE inventory models. We first discuss the situation in which we assume that demand per unit time is constant and known. This is the situation discussed in Harris (1913). Instead of discussing the single-item model, we discuss a multi-item single-echelon situation, allowing us to introduce in a natural way another perspective on the problem than that of cost minimization: we assume the firm wants to maximize its return on investment. This leads to some new

insights that were first published by Trietsch (1995), but in our view did not receive sufficient attention. We also discuss the relation between EOQ and ABC classification. Next we discuss the Newsvendor equation and how it unifies optimization of SISE systems. We further show the robustness of stochastic inventory model performance under given average inventory level and ordering frequency. This prepares for the next section on multi-item multi-echelon systems.

3.1 Deterministic demand

In this subsection we consider the situation where a firm manages the inventory of N items under the assumption of constant item demand rate. We assume linear holding costs and fixed ordering costs for each item. We discuss the relationship between the optimal order quantities and the so-called ABC classification, which distinguishes between important and non-important items. We compare the optimal order quantities under the objective of cost minimization with the optimal order quantities under Return-On-Investment maximization.

Let us define

Table 1 Inventory system definitions, constant demand rate

Variable	Definition
r	yearly interest rate
D_i	yearly demand for item i, i=1,...,N
p_i	sales price per item i, i=1,...,N
v_i	cost price per item i, i=1,...,N
A_i	fixed ordering cost of item i, i=1,...,N
F	fixed assets
C	sum of holding and ordering costs
R	return on investment in fixed assets and inventory assets
${Q_i}^C$	optimal order quantity of item i under cost minimization, i=1,...,N
${f_i}^C$	optimal order frequency of item i under cost minimization, i=1,...,N
${Q_i}^R$	optimal order quantity of item i under ROI maximization, i=1,...,N

3.1.1 EOQ and ABC classification

Following Harris (1913) we find that the optimal order quantity under minimization of the sum of ordering and holding (capital) costs is given by

$$ {Q_i}^C = \sqrt{\frac{2A_iD_i}{v_ir}}, i = 1, ..., N \tag{1} $$

$$C = \sum_{i=1}^{N} \sqrt{2A_i v_i r D_i} \tag{2}$$

Though the assumption of constant demand is very restrictive, many authors have found that when extending the problem to stochastic demand and minimizing the sum of ordering costs, holding cost and penalty costs, the expression for ${Q_i}^C$ yields close to optimal results (cf. Silver et al. (1998)). It is also worth mentioning that it follows from the findings in Daganzo (2005) that the expression for C is an accurate approximation of the minimal yearly costs for the case of deterministic dynamic item demand, provided the volatility is not too high.

Another relevant observation emerges when we consider the optimal ordering frequency,

$$f_i^C = \frac{D_i}{{Q_i}^C} = \sqrt{D_i v_i}\sqrt{\frac{r}{2A_i}}, i = 1, ..., N \tag{3}$$

From equation (3) we observer that the optimal order frequency grows with the square route of the item turnover. The higher the item order frequency, the more often we must pay attention to the item, as it typically runs out of stock before arrival of an order. Then it follows that the higher the item turnover, the more attention we should pay to the item. This is fully in line with the so-called ABC classification, where A-item are items with high turnover and C-items are items with low turnover. The ABC classification assumes that we should pay most attention to A items. Thus we see that the professional perspective behind ABC classification can be formally supported by the Economic Order Quantity results of Harris (1913).

3.1.2 Cost minimization versus Return On Investment maximization

As mentioned above Trietsch (1995) takes a different perspective: maximization of ROI. In that case we can derive the following objective function,

$$R = \frac{\sum_{i=1}^{N}((p_i - v_i)D_i - \frac{Q_i v_i r}{2} - \frac{A_i D_i}{Q_i}) - Fr}{F + \sum_{i=1}^{N} \frac{Q_i v_i}{2}} \tag{4}$$

By taking partial derivatives we can solve for the optimal Q_i^R. It turns out to be most convenient for solving the set of equations to introduce a constant λ^*, which must satisfy the following set of equations,

$$Q_i^R = \sqrt{\frac{2A_i D_i}{v_i(\lambda + r)}}, i = 1, ..., N \tag{5}$$

$$\lambda = \frac{\sum_{i=1}^{N}((p_i - v_i)D_i - \frac{Q_i^R v_i r}{2} - \frac{A_i D_i}{Q_i^R}) - Fr}{F + \sum_{i=1}^{N} \frac{Q_i^R v_i}{2}} \tag{6}$$

The solution for the optimal λ is found by iteratively solving equations (5) and (6) as a fixed point problem. The convergence is very fast, typically within 6 iterations. Note that it follows from equations (4) and (6) that λ is the ROI itself! It follows from equation (5) that if ROI is positive, then the economic order quantity Q_i^C is greater than the order quantity under ROI maximization Q_i^R. It also follows from equation (5) that $\frac{Q_i^C}{Q_i^R}$ is the same for all i.

In Trietsch (1995) a remarkable finding is reported that also holds for the formulation of our ROI maximization problem that differs slightly from the one formulated in his paper: if the relative annual demand quantities are fixed, then the optimal Q_i^C does not depend on the total annual demand (this follows directly from the expressions for the partial derivatives, but is not obvious from equations (5) and (6)). Thus the behaviour of the optimal lot sizes under cost minimization is completely different from that under ROI maximization. In the former case the lot size grows as a square root of annual demand, while in the latter case it is independent of the annual demand. This implies that under ROI maximization the frequency of ordering increases linear in total annual demand, while under cost minimization the frequency of ordering grows with a square root of annual demand. These statements also hold when considering order frequency as a function of annual turnover. The substantial difference between optimal lot sizes under cost minimization and ROI maximization is illustrated in figure 1 and 2. We define the optimal ROI as R^* and $R(EOQ)$ as the ROI when using Q^C for each item. The figures show the optimal Return on Investment R^* (right axis), $\Delta_Q := \frac{Q_i^C}{Q_i^R}$, and $\Delta_{ROI} := \frac{R^*}{R(EOQ)}$ as a function of the fixed assets F.

The above findings are not widely reported in inventory management textbooks, while in our view they are important and provide a broader perspective on the lot sizing problem that is relevant to both students and professionals.

3.2 Stochastic demand

In this subsection we consider SISE models under stochastic demand. We assume that the lot size can be derived from the EOQ model. Thereby we focus our discussion on the control parameter that determines the amount of slack needed to cope with demand and supply uncertainty, so that customer requirements can be satisfied at minimal cost or at a sufficient level of customer service. Let us introduce some notation.

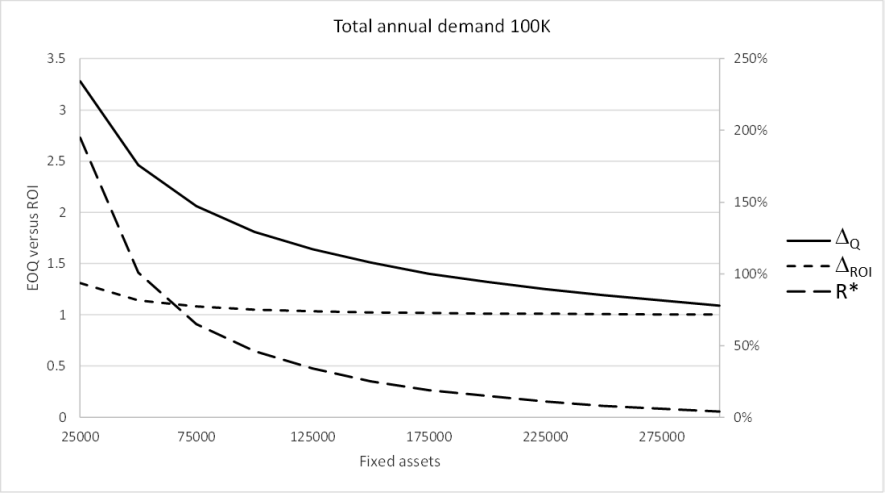

Fig. 1 Cost optimization versus ROI maximization

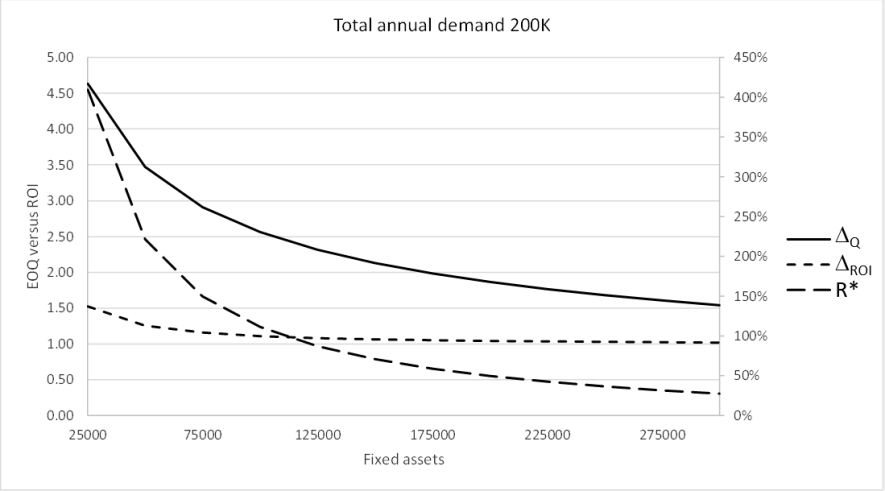

Fig. 2 Cost optimization versus ROI maximization

Note that the net stock of an item equals its physical stock minus its backlog, and the inventory position of an item equals the sum of its net stock and its outstanding orders.

Table 2 Inventory system definitions, stochastic demand

Variable	Definition
H	holding cost rate per item on stock
p	penalty cost rate per item short
ξ	parameter of the inventory control policy
$\mathscr{P}_\xi$	inventory control policy with parameter ξ
$X_\xi(t)$	net stock at time t under control policy $\mathscr{P}_\xi$
$Y_\xi(t)$	inventory position at time t under control policy $\mathscr{P}_\xi$
X_ξ	long-run stationary net stock under control policy $\mathscr{P}_\xi$
$C(\xi)$	long-run average costs under control policy $\mathscr{P}_\xi$
$C_1(\xi)$	long-run sum of holding and penalty costs under control policy $\mathscr{P}_\xi$
$C_2(\xi)$	long-run average costs minus the long-run sum of holding and penalty costs under policy $\mathscr{P}_\xi$
$D(0,t]$	demand until time t
$Q_\xi(0,t]$	cumulative amount replenishment orders received in stock until time t under policy $\mathscr{P}_\xi$

3.2.1 Newsvendor fractile

In this subsubsection we show that under most control policies used for SISE models the so-called Newsvendor equation holds under the optimal policy parameters. As stated in section 1, the Newsvendor equation states that the non-stockout probability at an arbitrary point in time equals the quotient of penalty cost rate and the sum of holding cost rate and penalty cost rate. This quotient is called the Newsvendor fractile and equals $\frac{p}{p+h}$. A theorem provides a set of sufficient conditions that the cost structure and control policy structure must satisfy to yield this result. These sufficient conditions are easy to verify.

The result holds for both continuous time and discrete time. In the continuous time case we incur costs at the relevant rate per time unit per unit short or in stock, while in the discrete time we incur costs at the relevant rate per item short or in stock at the end of a time unit. The policy $\mathscr{P}_\xi$ has a control parameter ξ and possibly other parameters. We assume in the theorem below that the other parameters are constant. For ease of reference we may think of $\mathscr{P}_\xi$ as an (s,nQ)-policy, where ξ is the reorder point s and the other parameter is the lot size Q. With the above notation we can formulate the following theorem that provides a set of conditions under which the optimal ξ, i.e. the value of ξ that minimizes the long-run average cost, satisfies the Newsvendor equation .

Theorem 1. *Assume that the inventory process $X_\xi(t), t \geq 0$, under $\mathscr{P}_\xi$ is ergodic $\forall \xi \in \mathbb{R}$. Assume that holding and penalty costs are the only costs affected by the inventory process $X_\xi(t), t \geq 0$, and all other costs are affected by the replenishment process $Q_\xi(0,t], t \geq 0$. Assume that $C(\xi)$ is continuously differentiable in ξ. Let ξ^* denote the cost optimal value of ξ.*

Assume that the following property holds $\forall \xi, \varepsilon \in \mathbb{R}$*,*

$$X_{\xi+\varepsilon}(t) = X_{\xi}(t) + \varepsilon, \forall t \geq 0 \qquad \textit{(Net stock translation property)}$$

Then the following properties hold:

(i) $Q_{\xi+\varepsilon}(0,t] = Q_{\xi}(0,t], \forall t \geq 0$
(ii) $C_2(\xi) = C_2(0), \forall \xi \geq 0$
(iii) $C_2^{'}(\xi) = HP\{X_{\xi} > 0\} - pP\{X_{\xi} \leq 0\}, \forall \xi \in \mathbb{R}$
(iv) $P\{X_{\xi^*} > 0\} = \frac{p}{p+H}$

Proof. By assumption we have that

$$X_{\xi+\varepsilon}(0) = X_{\xi}(0) + \varepsilon.$$

As the exogenous stochastic demand process $D(0,t]$ is not affected by the inventory control policy and the inventory balance equation states that net stock at time t equals the net stock at time 0 plus cumulative orders received minus cumulative demand, it follows that

$$\begin{aligned} Q_{\xi}(0,t] &= X_{\xi}(t) - X_{\xi}(0) + D(0,t] \\ &= (X_{\xi+\varepsilon}(t) - \varepsilon) - (X_{\xi+\varepsilon}(0) - \varepsilon) + D(0,t] \\ &= X_{\xi+\varepsilon}(t) - X_{\xi+\varepsilon}(0) + D(0,t] \\ &= Q_{\xi+\varepsilon}(0,t). \end{aligned}$$

This proves (i). Then (ii) follows directly from the assumption that the costs other than holding and penalty costs are affected only by the replenishment process. This leaves us to proof (iii) and (iv).

Using the definition of C_2 we obtain

$$C_2(\xi^* + \varepsilon) - C_2(\xi^*) = (E[HX_{\xi+\varepsilon}^{+} + pX_{\xi+\varepsilon}^{-}]) - (E[HX_{\xi}^{-} + pX_{\xi}^{-}]). \qquad (7)$$

Let us consider an arbitrary point in time t. We distinguish between the events $\{X_{\xi}(t) > 0\}$, $\{-\varepsilon < X_{\xi}(t) \leq 0\}$, and $\{X_{\xi}(t) \leq -\varepsilon\}$. This yields

$$\begin{aligned}E[HX^+_{\xi+\varepsilon}(t)+pX^-_{\xi+\varepsilon}(t)] &= E[(HX^+_{\xi+\varepsilon}(t)+pX^-_{\xi+\varepsilon}(t))\mathbb{1}_{\{X_\xi(t)>0\}}]\\ &+E[(HX^+_{\xi+\varepsilon}(t)+pX^-_{\xi+\varepsilon}(t))\mathbb{1}_{\{-\varepsilon<X_\xi(t)\le 0\}}]\\ &+E[(HX^+_{\xi+\varepsilon}(t)+pX^-_{\xi+\varepsilon}(t))\mathbb{1}_{\{X_\xi(t)\le-\varepsilon\}}]\\ &= E[HX_{\xi+\varepsilon}(t)\mathbb{1}_{\{X_\xi(t)>0\}}]\\ &+E[(HX^+_{\xi+\varepsilon}(t)+pX^-_{\xi+\varepsilon}(t))\mathbb{1}_{\{-\varepsilon<X_\xi(t)\le 0\}}]\\ &+E[-pX^-_{\xi+\varepsilon}(t))\mathbb{1}_{\{X_\xi(t)\le-\varepsilon\}}]\end{aligned}$$

Now we use the sample path property in the theorem, which leads to the following equation.

$$\begin{aligned}E[HX^+_{\xi+\varepsilon}(t)+pX^-_{\xi+\varepsilon}(t)] &= E[H(X_\xi(t)+\varepsilon)\mathbb{1}_{\{X_\xi(t)>0\}}]\\ &+E[(HX^+_{\xi+\varepsilon}(t)+pX^-_{\xi+\varepsilon}(t))\mathbb{1}_{\{-\varepsilon<X_\xi(t)\le 0\}}]\\ &+E[-p(X_\xi(t)+\varepsilon)\mathbb{1}_{\{X_\xi(t)\le-\varepsilon\}}]\end{aligned} \tag{8}$$

In a similar way we can write

$$\begin{aligned}E[HX^+_{\xi}(t)+pX^-_{\xi}(t)] &= E[HX_{\xi}(t)\mathbb{1}_{\{X_\xi(t)>0\}}]\\ &+E[(HX^+_{\xi}(t)+pX^-_{\xi}(t))\mathbb{1}_{\{-\varepsilon<X_\xi(t)\le 0\}}]\\ &+E[-pX_{\xi}(t))\mathbb{1}_{\{X_\xi(t)\le-\varepsilon\}}].\end{aligned} \tag{9}$$

Combining equations (8) and (9), taking the limit $t\to\infty$, and using the definition of the indicator function, we obtain an expression for the righthand side of equation (7).

$$\begin{aligned}&(E[HX^+_{\xi+\varepsilon}+pX^-_{\xi+\varepsilon}])\\ &-(E[HX^-_{\xi}+pX^-_{\xi}]) \;= H\varepsilon P\{X_\xi>0\}-p\varepsilon P\{X_\xi\le-\varepsilon\}\\ &\qquad +E[(HX^+_{\xi+\varepsilon}+pX^-_{\xi+\varepsilon})\mathbb{1}_{\{-\varepsilon<X_\xi\le 0\}}]\\ &\qquad -E[(HX^+_{\xi}+pX^-_{\xi})\mathbb{1}_{\{-\varepsilon<X_\xi\le 0\}}].\end{aligned} \tag{10}$$

Rearranging terms of equation (10) and using equation (7), we obtain,

$$\begin{aligned}&\tfrac{C_2(\xi+\varepsilon)-C_2(\xi)}{\varepsilon}-(HP\{X_\xi>0\}-pP\{X_\xi\le-\varepsilon\}) =\\ &\tfrac{E[(HX^+_{\xi+\varepsilon}+pX^-_{\xi+\varepsilon})\mathbb{1}_{\{-\varepsilon<X_\xi\le 0\}}]-E[(HX^+_{\xi}+pX^-_{\xi})\mathbb{1}_{\{-\varepsilon<X_\xi\le 0\}}]}{\varepsilon.}\end{aligned} \tag{11}$$

Now note that under the condition that $X_\xi \leq -\varepsilon$ both X_ξ and $X_{\xi+\varepsilon}$ have an absolute value less than ε. Using this observation then this and equation (11) yields the following inequality,

$$|\frac{C_2(\xi+\varepsilon)-C_2(\xi)}{\varepsilon} - (HP\{X_\xi > 0\} - pP\{X_\xi \leq -\varepsilon\})| \leq 2(H+p)P\{-\varepsilon < X_\xi \leq 0\}. \tag{12}$$

As we assume that $C_2(\xi)$ is continuously differentiable, it must hold that X_ξ has a continuous distribution. Thus, taking the limit $\varepsilon \to 0$, the righthandside of equation (12) goes to 0. Then we find the following expression for the derivative of $C_2(\xi)$,

$$C_2^{'}(\xi) = HP\{X_\xi > 0\} - pP\{X_\xi \leq 0\}, \forall \xi \in \mathbb{R}, \tag{13}$$

which proves property (iii) of the theorem 1. In order to find the cost-optimal ξ^* we only need to minimize $C_2(\xi)$ with respect to ξ. As the KKT condition for ξ^* should hold, we have

$$HP\{X_\xi^* > 0\} - pP\{X_\xi^* \leq 0\} = 0 \iff P\{X_\xi^* > 0\} = \tfrac{p}{p+H}. \tag{14}$$

This concludes our proof. □

The theorem provides the underlying principle for the consistent emergence of the Newsvendor fractile when analyzing stochastic single-echelon inventory systems. For all commonly used inventory policies, such as MRP's time-phased order point (cf. section 4.4), (s,S), (s,nQ), (R,s,S), and (R,s,nQ), it is easy to verify that if the reorder point is increased by some ε, while keeping the lot size parameter constant, i.e Q and $S-s$, then the net stock also increases by ε. The aforementioned policies observe the inventory position $Y(t)$ and if $Y(t)$ is less than the reorder level s an order is placed at the supplier. For the case of an (s,nQ)-policy or an (R,s,nQ)-policy, as many times a quantity of size Q is ordered as needed to increase the inventory position above the reorder level s. For the case of an (s,S)-policy or an (R,s,S)-policy, the order quantity raises the inventory position to order-up-to-level S. Inventory control policies that only use information on outstanding orders, inventory and demand for the single item under consideration, are called installation stock policies.

An important consequence of the theorem is that it holds for any multi-item multi-echelon system, where the end-items are controlled according to (one of the earlier mentioned) installation stock policies for which the net stock translation property holds. I.e. for any end-item the Newsvendor fractile holds under that situation. The reasoning for this is as follows. Assume at time 0 some outstanding orders and assume that in system 1 the net stock equals $X_{\xi,k}(0)$ and in system 2 the net stock equals $X_{\xi,k}(0)+\varepsilon$ for each end-item k. Assume that the control parameter in

system 1 equals ξ_k and in system 2 $\xi_k + \varepsilon$. For the above-mentioned installation stock policies it is easy to show that the orders generated over time towards the upstream multi-echelon system are identical. As this upstream system's response is completely determined by the orders received from all end-items, it follows that this response, i.e. the replenishment process, is also identical for system 1 and system 2. It is easy to see that property (ii) is a necessary and sufficient condition for the net stock translation property to hold. Thus the optimal installation stock policy must satisfy the Newsvendor fractile.

It is worthwhile to note here that the theorem may also hold for the so-called echelon stock policies discussed in section 4, provided the end-items are controlled by installation stock policies. The echelon stock policies are using the echelon inventory position of an item as state variable. The echelon inventory position is defined as the physical stock of an item plus its outstanding orders plus the echelon inventory positions of its parent items, taking into account how many items are needed to make one parent item. It follows that for end-items the echelon inventory position is the same as the net stock, whereby end-item echelon stock policies are in fact installation stock policies. It suffices to check the set of conditions for the end-items only to see if the Newsvendor equations hold for end-items under echelon stock policies. In subsection 4.8 we show that the Newsvendor equation result can be extended to Newsvendor equations related to each item in the MIME system.

We also note here that we did not make any assumption on the demand process, other than assuming the demand process yields a continuous differentiable cost function $C_2(\xi)$. This implies that the Newsvendor fractile also holds for non-stationary demand, such as seasonal demand and autocorrelated demand, provided that the control policy used yields the net stock translation property. In case we assume that the policy is driven by a demand forecast, such as is the case for rolling scheduling policies, with a stationary additive forecast error, then a time-phased order point policy is appropriate and this policy ensures that the net stock translation property holds.

Finally we note here that for the case of discrete demand it can be shown that the net stock translation property is a sufficient condition for the *Newsvendor fractile inequality* to hold.

The natural emergence of the Newsvendor fractile points towards the importance of the non-stockout probability as a performance measure. A target non-stockout probability relates one-to-one to a penalty cost value. This provides insight into the implicitly assumed penalty costs per unit per unit time. Another advantage of the non-stockout probability as a service measure is the ease of determining its actual value over some time period. It suffices to keep track of the inventory level. The most-prominent service level in inventory management literature is the fill rate. In practice it is often used, too, but a major issue is that computation of the fill rate requires knowledge of the actual demand over the time period of interest. In most situations demand data are not known, only sales data. As sales is impacted by the inventory availability, the use of sales data may give a too optimistic picture of the situation.

3.2.2 Sensitivity analysis

In the introduction we discussed the difference between inventory model results and results in practice under the influence of planners and schedulers. One of the typical consequences is that we do not know which inventory model is applied. We may infer the model used from the demand data, inventory data and replenishment order data. This would be an interesting subject for further research. Another question that arises in this context is, whether it matters what inventory model is used to determine the control parameters to be used in practice. To answer this question we use the algorithms from De Kok (1991*a*) to determine inventory control parameters for the (R,S)-model, the (R,s,S)-model, and the (R,s,nQ)-model.

We set up a computational experiment as follows. Knowing that the main cost drivers of inventory management are the frequency of ordering and the average inventory level, we fix the average lot size and determine the average inventory needed under an (R,s,nQ)-policy with $R = 1$ to meet a target fill rate level P_2. The average lot size may be larger than Q, as at some instances multiple Q's may be ordered as a consequence of a large demand causing a high undershoot of the reorder level s. For an exact formula for the average order size in an (R,s,nQ)-model we refer to De Kok (1991*b*).

E[Q]	P2	L	c2	E[X]	(R,s,S)	(R,s,nQ)	(R,S)
200	0.9	5	0.5	271	0.91	0.90	0.90
200	0.9	5	0.75	336	0.91	0.90	0.90
200	0.9	5	1	394	0.91	0.90	0.90
200	0.9	5	1.5	500	0.91	0.90	0.90
200	0.9	10	0.5	352	0.91	0.90	0.90
200	0.9	10	0.75	436	0.91	0.90	0.90
200	0.9	10	1	511	0.91	0.90	0.90
200	0.9	10	1.5	643	0.91	0.90	0.90
200	0.95	5	0.5	348	0.95	0.95	0.95
200	0.95	5	0.75	434	0.95	0.95	0.95
200	0.95	5	1	512	0.95	0.95	0.95
200	0.95	5	1.5	655	0.95	0.95	0.95
200	0.95	10	0.5	449	0.95	0.95	0.95
200	0.95	10	0.75	560	0.95	0.95	0.95
200	0.95	10	1	658	0.95	0.95	0.95
200	0.95	10	1.5	833	0.95	0.95	0.95
500	0.9	5	0.5	327	0.91	0.90	0.88
500	0.9	5	0.75	383	0.91	0.90	0.88
500	0.9	5	1	435	0.91	0.90	0.88
500	0.9	5	1.5	533	0.91	0.90	0.88
500	0.9	10	0.5	396	0.91	0.90	0.89
500	0.9	10	0.75	473	0.91	0.90	0.89
500	0.9	10	1	543	0.91	0.90	0.89
500	0.9	10	1.5	669	0.91	0.90	0.89
500	0.95	5	0.5	411	0.95	0.95	0.93
500	0.95	5	0.75	487	0.95	0.95	0.93
500	0.95	5	1	558	0.95	0.95	0.93
500	0.95	5	1.5	691	0.95	0.95	0.93
500	0.95	10	0.5	470	0.94	0.95	0.93
500	0.95	10	0.75	602	0.95	0.95	0.94
500	0.95	10	1	694	0.95	0.95	0.94
500	0.95	10	1.5	862	0.95	0.95	0.94
1000	0.9	5	0.5	476	0.91	0.90	0.88
1000	0.9	5	0.75	516	0.91	0.90	0.87
1000	0.9	5	1	556	0.91	0.90	0.87
1000	0.9	5	1.5	635	0.91	0.90	0.87
1000	0.9	10	0.5	524	0.91	0.90	0.89
1000	0.9	10	0.75	584	0.91	0.90	0.88
1000	0.9	10	1	642	0.91	0.90	0.88
1000	0.9	10	1.5	751	0.91	0.90	0.88
1000	0.95	5	0.5	573	0.95	0.95	0.92
1000	0.95	5	0.75	632	0.95	0.95	0.92
1000	0.95	5	1	689	0.95	0.95	0.92
1000	0.95	5	1.5	802	0.95	0.95	0.92
1000	0.95	10	0.5	643	0.95	0.95	0.93
1000	0.95	10	0.75	727	0.95	0.95	0.93
1000	0.95	10	1	806	0.95	0.95	0.93

Table 3 Sensitivity analysis of inventory models

Then we determine the (R,s,S)-policy with $R = 1$ and the (R,S)-policy that have the same average inventory and average lot size. The first step in the procedure is to determine $(S-s)$ for the (R,s,S)-policy and R for the (R,S)-policy such that the average lot size equals the one for the (R,s,nQ)-policy applied to the same instance. Then we can compute s and S for the (R,s,S)-policy and (R,S)-policy, respectively, such that the average inventory equals the one for the (R,s,nQ)-policy applied to the same instance. For these policies we determine the actual fill rate P_2.

In Figure 3 we present the results. As expected, the fill rates of the (R,s,S)-policy are slightly better than that of the (R,s,nQ)-policy and in most cases identical. It is more striking that the fill rates under an (R,S)-policy are close to the (R,s,nQ)-policy fill rates and get lower as lot sizes get larger. The impact of the constant replenishment lead time L and squared coefficient of variation c_2 are negligible. Thus it can be argued that the choice of the inventory control model is not crucial and one can choose for the model that is mathematically more tractable. In particular the virtually identical results for the (R,s,S)-model and the (R,s,Q)-model favors the latter model as its inventory position distribution can exactly be determined (uniform distribution), while the inventory position distribution of the (R,s,S)-model involves the renewal function of the demand distribution, which is intractable in most cases. We would like to mention here that the findings presented here are confirmed by repeated experimentation in different real-life situations.

Another argument for being cautious to assume that we should always choose the cost-optimal (R,s,S)-model is that inventory models are abstractions from reality. Firstly, it may be that the production process requires fixed lot-sizes, which is captured by the (R,s,nQ)-policy. Secondly, production of replenishment orders involves equipment and human resources, which availability is planned to process multiple replenishment orders. For resource planning, timing of production order starts is key. For both the (R,s,nQ)-policy and the (R,s,S)-policy, timing of replenishment orders is stochastic, while timing is fixed under the (R,S)-policy. Given the robustness of the inventory model performance, in particular for items ordered at high frequencies, the (R,S)-policy is particularly suited for taking resource constraints into account, albeit implicitly.

3.2.3 Empirical validity

As with every model applied in practice, it is important to ascertain its empirical validity. After 30 years of extensive development and testing of these models in practice through numerous MSc thesis projects, it is safe to say that SISE have proven their empirical validity. We refer to the previous section and the introduction for modelling considerations, and in particular to the distinction between intervention-dependent and intervention independent performance measures. Only the latter are suited for validation of inventory models and for setting inventory model parameters in business information systems. The same MSc projects revealed that most ERP systems have mathematically incorrect algorithms for setting inventory model parameters. One of the most striking errors found is the application of the algorithm

for determining P_1, the non-stockout probability at the end of a replenishment cycle, to the situation with a P_2 (fill rate) performance target. Our only explanation for these incorrect formulas in ERP systems is the widespread (if not exclusive) use of the P_1-measure in basic textbooks on production and inventory management, combined with the extensive use of the normal demand assumption, which is rarely valid, due to the intrinsic high volatility of item demand.

Another typical mathematical anomaly in ERP system inventory models is the modification of the algorithm for items with high demand volatility. In that case the modified algorithm ensures that the inventory is lower than according to the correct mathematical analysis. Still, there is a relevant underlying tacit knowledge aspect to this type of modification. In this case we should understand what is the cause of high item demand volatility and how it manifests itself. High demand volatility can easily be understood when simulating a gamma distribution with a high coefficient of variation. One observes many small demands and now and then a demand which is many times the average demand. We should be aware that in practice demand is transient. We introduce an item, at which moment the demand process starts. Over time, demand increases, possibly stabilizes around some level during some period of time, and decreases to zero after that. This product life cycle can last from a few months (e.g. a fashion T-shirt) to many years (e.g. a white T-shirt). The high inventory associated with high volatility demand results from the recurring peak demands over an infinite time length. In practice we may face many small demands and only a few peak demands. And we may even face no peaks at all, as we derive demand volatility from past demand data. The peak demands are oftentimes due to special events, which may well be known far in advance, or in other cases customers understand it may take longer than normal to satisfy. These are reasons to ignore the peak demands, leaving us with many small demands. These constitute a demand process that lends itself perfectly for being managed with the inventory control models discuss above. The remaining volatility without the peak demands is much lower, whereby a lower average inventory is sufficient to satisfy these smaller demands at the target performance level. We refer to De Kok (1993) and Dekker et al. (1998) for further details.

We conclude that SISE inventory systems are empirically valid provided proper modelling considerations are taken into account (cf. section 2). The mathematical analysis of these models is available in the form of exact and accurate approximate algorithms, which can easily be implemented in large-scale inventory management systems. For extensive discussions of the mathematical analysis we refer to Zipkin (2000) and Axsäter (2015).

4 Uncapacitated multi-item multi-echelon models

In this section we discuss the modelling and analysis of MIME systems without restrictions on the resources that perform the transformation of child items into their parent item. We only consider the material constraints that are constituted by the

available child item availability at the moment of the order release of the parent item. In subsection 4.3 we present a generic MIME model under the assumption of constant lead times, for which we provide a motivation in subsection 4.1. We discuss feasibility conditions that control policies for MIME inventory system must satisfy in subsection 4.4. In subsection 4.5 we introduce two concepts that emerge when extending the control of SISE inventory systems to MIME inventory systems: synchronization and allocation. Synchronization ensures that orders are not released too early, allocation ensures that material availability constraints are respected. We first discuss these two concepts in the context of convergent MIME systems in subsubsection 4.5.1 and extend the concepts for application to general MIME systems in subsubsection 4.5.2. We show that the control of general MIME structures can be derived from associated divergent MIME structures, which we denote as decision node structures. The base stock policies for the divergent MIME systems applied to general MIME systems are called Synchronized Base Stock (SBS) policies. While for convergent MIME systems the concepts provide insight into the structure of optimal policies, for general systems this is no longer the case. We provide insight on this matter by discussing some structural properties of Bills of Material in subsubsection 4.5.3. In subsection 4.6 we use the case example presented below in subsection 4.2 to illustrate the derivation of decision node structures for general MIME systems. As we have reduced the control of general MIME systems to the control of divergent MIME systems, we discuss optimal policies for the latter systems in subsection 4.7 and close-to-optimal policies that are mathematically tractable and easy to implement in practice. The extension of the Newsvendor equation for SISE systems to divergent MIME systems is discussed in subsubsection 4.8. In subsection 4.9 we discuss the performance of SBS policies in comparison with rolling scheduling policies is commonly used in practice to decide on order releases in MIME systems. In subsection 4.10 we discuss the empirical validity of SBS policies as derived from extensive empirical research. The empirical validity implies that, likewise for SISE systems, average item inventories and average item order frequencies determine end-item customer service. Finally, subsection 4.11 is devoted to the strategic positioning of inventory capital across the supply chain.

4.1 Material availability and stochastic lead times

Though we did not explicitly discuss the role of lead times in the analysis of SISE systems, all performance characteristics of these systems depend on the distribution of the replenishment lead time (cf. Axsäter (2015) and Zipkin (2000)). If we assume the replenishment lead time is constant, this can be seen as modelling the situation where the supplier of the item ordered always can deliver the order according to a pre-specified nominal lead time. This may be due to ample inventory at the supplier or the capability of the supplier to produce and ship the item according to the nominal lead time. If we learn from replenishment lead time data that the actual lead time is stochastic, then apparently the supplier is not always capable of adhering to

the nominal lead time. This incapability may be due to lack of available item inventory to ship from stock, when this is necessary to meet the nominal lead time. It could also be due to lack of material availability, whereby production is delayed and subsequently replenishment order shipment is delayed, which eventually leads to exceeding the nominal lead time. Next to lack of inventory availability, insufficient resource availability in transportation, warehousing or production can cause delays that imply exceeding the nominal lead time. Modelling explicitly finite resource availability is discussed in Section 5. In this section we concentrate on the explicit modelling of inventory availability over time at stock locations supplying other stock locations. We note here that finite resource availability can be modelled *implicitly* by assuming a nominal production lead time that enables the production department to smooth resource requirements within this lead time so that both capacity constraints are met and due date requirements derived from the order arrivals and the nominal production lead times are met with high probability. In that way the nominal production lead time decouples the inventory management across multiple stockpoints from shopfloor management at each production location. For an extensive discussion on this way of hierarchically decomposing planning and control in supply chains we refer to De Kok & Fransoo (2003). For the remainder of this section we thus assume that each item order is delivered according to its nominal lead time, *provided that the child items of the item are sufficiently available in stock to release the order to production upon arrival*.

4.2 The example supply chain

Throughput this section we use an example to illustrate the problems emerging when developing control policies for general MIME systems. We consider a 2-echelon supply chain consisting of 4 end-items, numbered 1,2,3, and 4, and three components, numbered 5,6, and 7. End-item 1 is assembled from one item 5 and one item 6 with a nominal lead time of 2 periods. End-item 2 is assembled from one item 5, two items 6, and two items 7, with a nominal lead time of 4 periods. End-item 3 is assembled from one item 6 and two items 7 with a nominal lead time of 2 periods. End-item 4 is assembled from one item 6 and one item 7 with a nominal lead time of 4 periods. The component items 5, 6, and 7 have a nominal lead time of 2, 10, and 6 periods, respectively. This structural information about the supply chain is depicted in figure 3. The average demand per period for end-items 1,2,3, and 4 equals 100.

The example supply chain represents a general structure as we see that each component can be associated with an embedded divergent supply chain, and each end-item can be associated with an embedded convergent supply chain. Note that items 6 and 7 have both one and two as quantity per in relation to the end-items. As we will see different quantities per of child items in their parent items creates additional complexity on top of the inherent complexity of interactions between child item and parent items as a consequence of limited child item availability.

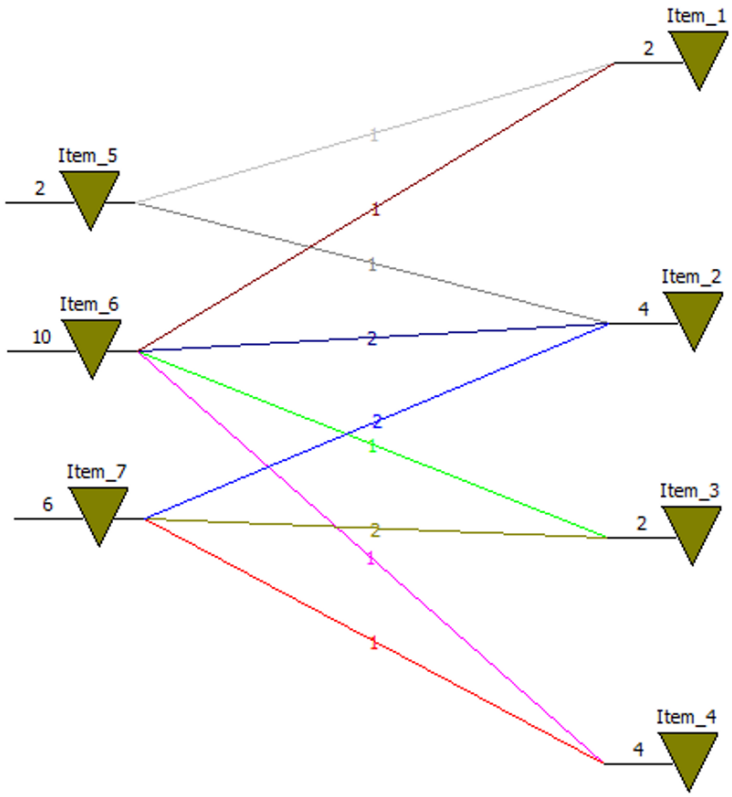

Fig. 3 The example supply chain

4.3 Modelling multi-echelon inventory systems

The challenge of multi-echelon inventory system modelling is to capture the interactions between item orders across the inventory system. These interactions are quite complex due to the stochasticity of demand over time (*uncertainty-driven complexity*) and the *structural complexity* of the Bill-Of-Material (BoM). The Bill-Of-Material describes what child items are needed and in what quantity, to produce one particular item. In this section we aim to provide fundamental insight into the relationship between order release decisions over time in real-life supply chains, i.e. general MIME inventory systems. The emphasis is on constraints imposed by order release decisions in the past on order release decisions now and in the future. We provide necessary conditions for feasible order release decisions, and sufficient conditions that enable transparency of the order release decisions over time. These sufficient conditions are induced by MIME inventory control policies, for which we can find optimal policies under an assumption that may not always hold, but whose

performance is sufficiently accurate for practical purposes. Furthermore these policies allow for highly efficient calculation of operational order release decisions and allow for efficient computation of close-to-optimal policies. Thus this section centers around a particular perspective on controlling general MIME inventory systems, while other approaches have been proposed in literature. We refer to these other approaches where appropriate, so that we can articulate the main differences with the MIME inventory control concept discussed below. In general we can state that the other approaches concentrate on calculating safety stock norms without dealing with the operational details of order release decisions over time in MIME inventory systems, or only consider the operational details of order release decisions without dealing with the question how safety stock norms must be set.

Let us define the BOM and some other relevant structural aspects of multi-echelon inventory systems.

Table 4 Defining a multi-echelon inventory system structure

Variable	Definition
$a_{i,j}$	number of items i required to produce one item of its parent item j
V_i	set of immediate successors of item i
U_i	set of all predecessors of item i
W_i	set of all successors of item i and item i itself, the echelon of i
E	set of all end-items
E_i	set of all end-items in W_i

The matrix $(a_{i,j})$ is the BoM matrix. In practice the BoM matrix is sparse, as the natural built-up of a product from materials into subassemblies and final assembly creates only parent-child relations between materials and subassemblies, and between subassemblies and final assembly. In ERP systems only the BoM *relations* are stored, which ensures efficient storage-and-retrieval. We assume that only end items face customer demand. If customers also demand non-end-items, such as in the case of spare part demand, then it is easy to modify the network structure by introducing dummy end-items for each upstream item with customer demand. The set V_i is the set of parent items of item i. The set U_i contains all the items from which item i is assembled (or in general, produced). The set W_i is the set of items in the echelon of i.

Case example
For our case example we have $E = \{1,2,3,4\}$, $W_1 = E_1 = \{1\}$, $W_2 = E_2 = \{2\}$, $W_3 = E_3 = \{3\}$, $W_4 = E_4 = \{4\}$, $E_5 = \{1,2\}$, $E_6 = \{1,2,3,4\}$, $E_7 = \{2,3,4\}$, $W_5 = \{1,2,5\}$, $W_6 = \{1,2,3,4,6\}$, $W_7 = \{2,3,4,7\}$, $U_1 = \{5,6\}$, $U_2 = \{5,6,7\}$, $U_3 = \{6,7\}$, $U_4 = \{6,7\}$.

In reality company supply chains consists of three or more echelons. In Business to Consumer (B2C) supply chains we identify materials sourced from outside sup-

pliers, finished goods at the factory and finished goods in the distribution network. In Business to Business (B2B) supply chains we identify parts and subassemblies at 1st tier suppliers and finished goods at the Original Equipment Manufacturer (OEM). Furthermore many items have both multiple children and multiple parents. For such an item i we have $|\{a_{j,i} > 0\}| > 1$ and $|\{a_{i,j} > 0\}| > 1$. Clearly, for an end-item k, $|\{a_{k,j} > 0\}| = 0$. The distribution network is a network where we have $|\{a_{j,i} > 0\}| = 1$ for all items, and $|\{a_{i,j} > 0\}| >= 1$ for all items but the end-items k. As we are interested in finding optimal control policies, it is important to be aware that this is not possible for divergent MIME systems, as follows from Eppen & Schrage (1981), Diks & De Kok (1998), and Doğru et al. (2009), amongst others. This implies that for any realistic supply chain structure there is no hope that we can find optimal control structures. This is an important observation that should be taken as a starting point for studying multi-echelon inventory systems under stochastic demand. If it is impossible to find optimal policies, the infinite space of heuristic policies opens up as a relevant space, similar to the situation of NP-hard combinatorial optimization problems. We should be aware that many mathematical properties only hold under optimality, such as the KKT conditions, and the Bellman equations. Though the Bellman equations cannot be used due to the *curses of dimensionality*, the KKT conditions can still be applied for finding optimal policies within *a class of policies*, by formulating the optimization problem as a function of the policy *parameters*. A nice example of such an approach for general systems can be found in Ettl et al. (2000). Before we discuss the optimization of given policies, we discuss some concepts that are foundational for controlling multi-echelon systems under uncertainty in demand.

4.4 Feasibility of order release quantities

Likewise any production and distribution planning problem, we want to balance supply and demand. In principle supply should be such that all demand can be satisfied according to the requirements of the customers. As demand is uncertain, this is not possible, as decisions on supply, i.e. order releases, precede the revelation of demand into customer orders. This also implies that order release decisions must be based on demand forecasts. As it matters, an analysis of SISE systems that explicitly incorporates the demand forecasting process into the inventory control policy leads to mathematical intractability. Thus most papers assume stationary demand processes, where it is implicitly assumed that we have an *optimal demand forecast*, as we assume to know the mean. In reality this is not the case.

Despite the mathematical intractability of SISE (and MIME) systems under demand forecasting, it is possible to formulate necessary conditions that an inventory control policy for MIME systems should satisfy. We emphasize here that these necessary conditions concern the inventory control *model*. We start from the premiss that the model as an abstraction of reality should be internally consistent. An important concept regarding internal consistence is feasibility of a decision: a decision

should be such that all constraints formulated are satisfied. As the model is an abstraction, it can be the case that decisions regarded as infeasible according to the model, are feasible in reality. However, such a solution is assumed to surface out of the interaction between a decision support system that proposes a solution and the decision maker who sees possibilities to improve the solution by relaxing constraints that are violated according to the model. This is common practice in planning and scheduling, although it is not always considered as explicitly as formulated here. In order to formulate the necessary conditions for feasible order release decisions in MIME systems given in De Kok & Fransoo (2003), we assume that we must decide on order release decisions at time 0. We define the following variables.

Table 5 Defining a multi-echelon inventory system operational characteristics

Variable	Definition
L_i	nominal lead time of item i
$L^c_{i,k}$	cumulative nominal lead time of item i with respect to end-item k
$X_i(t)$	net stock of item i at time t
ν_i	safety stock of item i
$Y_i(t)$	echelon inventory position of item i at time t, immediately after ordering
$Q_i(t)$	order quantity of item i that replenishes inventory of item i at time t
$P_i(t)$	quantity of item i produced in period $(t-1,t]$
$Z_i(t)$	echelon stock of item i at time t
$D_k(s,t]$	independent demand for end-item k during the interval $(s,t]$
$D_i(s,t)$	sum of independent demand for end-items $k \in E_i$ during the interval $(s,t]$
$G_i(s,t]$	dependent demand for end-item i during the interval $(s,t]$
$r_i(t)$	quantity released of item i at time t
$B_{i,k}(t,s)$	item i coverage of future demand of end-item k from time t until time $t+s$

The cumulative lead time $L^c_{i,k}$ is the sum of all nominal lead times over all items on the path from item i to end-item k. Furthermore, the dynamic variables defined above can refer to the instance of the planning problem to be solved at time 0, or to the actual realizations over time. As we assume that the model should be internally consistent, we assume that the dynamic variables refer to the solution of the planning problem at time 0, unless stated otherwise.

The necessary conditions for uncapacitated MIME systems concern material availability.

$$G_i(t-1,t] = \sum_{j=1}^{N} a_{ij} r_j(t) \leq X_i(t), \forall i \tag{15}$$

Noting that $G_i(t-1,t]$ denotes the dependent demand at time t, inequality (15) states that the amount of item i needed to release the orders of its parent items cannot exceed item i's availability. As the system is uncapacitated we assume that the orders released at time t arrive according to their nominal lead time, i.e.

$$Q_i(t+L_i] = r_i(t), \forall i \tag{16}$$

Inequality (15) is the Achilles heel of most inventory control policies proposed in literature for general supply chains. If SISE control policies are assumed in a multi-echelon context, then either one must compute explicitly the upstream delays due to lack of child item availability (cf. Ettl et al. (2000) and Kiesmüller et al. (2004)) or ignore delays by setting high target service levels at upstream stockpoints (cf. Graves & Willems (2000). The latter typically yields MIME policies that are far from optimal (cf. subsection 4.11). The former approach assumes FCFS allocation, as SISE inventory control policies do. FCFS allocation is not optimal in a MIME setting, where there is mutual dependency between child item availability, especially when there is a high commonality degree, i.e. many child items are used in the same parent items. As mutual dependency is mathematically complicated, most papers on MIME systems concern serial systems and divergent systems, where there is only one child item for each parent item.

Inequality (15) is also the Achilles heel of MRP I logic. Though this logic is by far most used in practice, it does not mean it is a correct MIME logic. Though MRP I stands for Material Requirements *Planning*, a better description of the logic would be Material Requirements *Generation*. The explosion process translates the Master Production Schedule (MPS), the end-item order release (or make) plan over time, into requirements for each upstream item, by using the BOM *gozinto* quantities (a_{ij}), the nominal lead time $\{L_i\}$, safety stocks, and lot sizing policies (cf. Orlicky (1975)).

The MPS can be interpreted as the scheduled receipts for an end-item within its nominal lead time, and planned receipts for the end-item k outside its nominal lead time. The explosion process uses the nominal lead times $L_.$ to offset planned receipts (replenishments of inventory) into planned orders, which constitute the dependent demand for child items to be able to start production of the orders released over time. The planned receipts are derived from the so-called time phased order point logic and lot sizing rules. Ignoring the lot sizing constraints yields the following ordering logic,

$$G_i(t-1,t) = \sum_{j \in V_i} a_{ij} r_j(t), \forall i$$

$$Q_i(t) = G_i(0,t+1) + v_i - \sum_{s=1}^{t-1} Q_k(s), \forall i, \forall t \geq L_i \tag{17}$$

The above equations indeed generate requirements aimed at each level to have the planned net stock equal to v_i from time $t+L_i$ onwards. But generating requirements is something completely different than planning! The logic generates (gross) requirements and planned orders, assuming that all order releases can be realized according to the generated timing. After generation of the gross requirements, the inconsistencies, or infeasibilities, regarding item availability over time are signalled as *past due* and *schedule-in*. As lot sizes have decreased structurally over the last

four decades, demand volatility causes regularly that the safety stock target cannot be met taking into account gross requirements and scheduled receipts. In that case *past due* of the first planned order is reported, which effectively is a signal that the immediate planned order cannot be executed according to the "plan". In that case we face the issue discussed above: ad hoc coordination of order release across different links in the supply chains. Interestingly, past due immediate orders occur irrespective of the amount by which the planned on-hand is below the safety stock target. If planned on-hand is well above zero at the time of arrival of the immediate planned order, this order could be reduced in size to take into account child item availability. In that case there is no need for ad hoc coordination across different links. However this is a manual process without much support from the MRP system. Organizationally, material planners are often dedicated to a subset of suppliers delivering different items. The mutual dependency of child items through their parents implies that past due orders require a complex coordination process among multiple child items, and multiple parent items, executed by multiple planners, without adequate system support. The complex coordination problem is discussed in more detail in the next subsection.

As a final remark, it is intriguing that the MRP I logic dates back to 1963, while George Dantzig introduced Linear Programming (LP) in 1947. Both MRP I and LP formulate the planning problem as a *deterministic* problem. As inequality (15) is linear in the decision variables, it can be incorporated in the planning problem formulation (cf. De Kok & Fransoo (2003)). To date MRP I has not been replaced by LP, even though running a large scale LP is feasible. Our hypothesis is that using LP as a substitute for MRP I would greatly improve material **planning**, and reduce the need for ad hoc coordination across links in the supply chain.

4.5 Synchronization and allocation

In single-item single-echelon (SISE) inventory systems balancing supply and demand implies covering average demand during the lead time plus safety stock with orders in the pipeline and on-hand stock. In a sense this is not different in balancing supply and demand in multi-item multi-echelon (MIME) inventory systems. The inventory position in SISE inventory systems represents the coverage of demand over the lead time plus the review period. The classical control policies mentioned in Section 3 use the inventory position as the state variable that determines whether to order or not, and if so, how much. The echelon inventory position in MIME systems plays a similar role as a state variable, but it is harder to identify what demand is covered over what period. To see this, let us first define the echelon inventory position and echelon stock for an item i at an ordering epoch t, immediately after ordering,

$$\begin{aligned} Z_k(t) &= X_k(t), \forall k \in E \\ Y_k(t) &= X_k(t) + \sum_{s=1}^{L_k} Q_k(t,s), \forall k \in E \\ Z_i(t) &= X_i(t) + \sum_{j \in V_i} Y_j(t), \forall i \notin E \\ Y_i(t) &= Z_i(t) + \sum_{s=1}^{L_i} Q_i(t,s), \forall i \notin E \end{aligned} \tag{18}$$

4.5.1 Convergent MIME systems

Now let us first consider a convergent system. In that case the supply chain produces a single product. This allows to express item units in end-item units, so that w.l.o.g. we can assume that $a_{i,j} \in 0,1$. Then the item i echelon inventory position $Y_i(t)$ must cover the end-item demand over item i's cumulative lead time plus review period (which we assume equal to 1), implying that the demand for end-item k satisfied directly from on-hand end-item inventory over time period $(t, t+L^c_{i,k}+1]$ is at most $Y_i(t)$. Given the coverage $Y_i(t)$ at time t, it follows that at time $t+s$, after a demand $D_k(t,t+s]$ has occurred, the remaining coverage of demand over time period $(t+s, t+L^c_{i,k}+1]$ equals $Y_i(t) - D_k(t,t+s]$. Thus we find

$$B_{i,k}(t+s, L^c_{i,k}+1-s) = Y_i(t) - D_k(t,t+s], \forall i \tag{19}$$

Now suppose that at time $t+s$ some item j must be ordered to cover demand for end-item k over its cumulative lead time $L_{j,k}$ plus review period. This implies that $s = L_{i,k} - L_{j,k}$. As it does not make sense to create a coverage of demand for end-item k exceeding the coverage from item i it follows that

$$Y_j(t+s) \leq B_{i,k}(t+s, L^c_{j,k}+1) = Y_i(t) - D_k(t,t+s], \forall (i,j) \in \{L^c_{j,k} \leq L^c_{i,k}\} \tag{20}$$

Thus the natural order created by the cumulative lead times of items in the supply chain implies a natural set of constraints that the item coverages of end-item demand should satisfy. Each item order implies a series of constraints, determined by the immediate coverage created at the time of ordering and the realizations of demand over time, for the ordering decisions for items with shorter cumulative lead times. This formalizes the obvious fact that decisions taken in the past constrain decision now. This set of constraints is the basis for the results in Rosling (1989), who proves that echelon base stock policies are optimal for convergent systems and each convergent system has an equivalent serial system. The echelon base stock policies for serial system operate as follows: each point in time t the inventory position of item j is increased to the minimum of its base stock level S_j and the sum of its current eche-

lon inventory position plus the on-hand inventory at its predecessor. For convergent systems the echelon base stock policies operate as follows: each point in time t the inventory position of item j is increased to the minimum of its base stock level S_j and the minimum of the coverages $B_{i,k}(t,t+L_{j,k}+1]$. Because the coverages over time constrain future coverages, the minimum of the coverages equals the coverage of the item with the next longer cumulative lead time. So once the base stock levels are known, the execution of the echelon base stock policies is highly efficient, if not trivial.

The above may be somewhat technical. Conceptually, inequalities (20) constitute the *synchronization of orders* in convergent supply chains. Synchronization can be interpreted as the horizontal coordination of orders for components and subassemblies that together enable the assembly of the final product to be sold to the market. Synchronization takes into account constraints set by earlier ordering decisions, implying that it may be suboptimal to execute the item's ordering policy based only on its own state, i.e. its $Y_(t)$. The consequence of ignoring the constraints expressed in inequalities (20) are stocks on hand that serve no purpose. We call these stocks *dead stocks*, as it can be shown that these stocks do not contribute to the service of the supply chain to the market. Below we discuss the calculation of these dead stocks.

At this stage it is of interest to return to the MRP I concept, i.e. Materials Requirements Planning (cf. Orlicky (1975)). MRP I is a logic that generates orders in a MIME inventory systems. The synchronization equations (20) are not respected by the *explosion equations* (17). This implies that items are ordered too early and are held in stock while waiting several time units for other items to arrive. The information to prevent this is available in the inventory database of the ERP system, but not used.

Not respecting the synchronization constraints (20) and thereby ordering to early is one thing, but ordering a quantity that is impossible to fulfill is another. We must be aware that for each item we can distinguish between synchronization equations that express that an order should not exceed coverage constraints to prevent ordering too early, and synchronization equations that express that an order cannot be released by lack of upstream materials. The latter subset is given by

$$Y_j(t+s) \leq B_{i,k}(t+s, L^c_{j,k}+1), \forall i \in \{m | a_{m,j} > 0\}. \tag{21}$$

The inequalities (21) for convergent MIME systems are equivalent to inequalities (15), the feasibility constraints for order release quantities for any BoM structure. Inequalities (21) represent the constraints set by the child items of an item j. They are an example of so-called *allocation constraints*. Note that inequalities (21) represent *explicit* relations between order release decisions for different items over time, while inequalities (15) represent *implicit* relations between order release decisions over time. Such implicit relationships are typical for mathematical programming formulations. While this can be seen as the strength of mathematical programming: decomposing a problem in sets of constraints, for analysis of systems under uncertainty we need explicit relationships in order to develop probabilistic expressions

for the performance of MIME systems, similar to those for SISE systems under uncertainty.

4.5.2 General MIME systems

So far we restricted ourselves to convergent systems, for which the synchronization inequalities (20) are unambiguously defined. For general BoM structures this is no longer the case. To see this, suppose that item i has multiple parent items and these parent items have multiple child items. Then there are many alternative ways to allocate child item availability among these parent items. Assuming that some allocation decisions have been taken before deciding on the allocation of item i availability, we can formulate a set of synchronization equations related to the child item allocation decisions already taken, but not for those child items which availability is allocated after the allocation decision for item i is taken. So the order in which allocation decisions are taken determines the synchronization inequalities to be respected, and the allocation decisions have many feasible solutions, likewise inequalities (20) allow for many feasible solutions. In a way we should take these decisions simultaneously, such that e.g. long-run costs average costs are minimized. Without going into detail here, we note that formulation of the Bellman equations for the optimal control problem related to MIME inventory system control shows that we are faced with the *curses of dimensionality*, implying there is no hope for ever finding an optimal solution to this problem. The interaction between allocation and synchronization decisions to be taken simultaneously could be seen as a chicken-and-egg situation or, from a somewhat more pessimistic perspective, as a catch-22 situation.

In De Kok & Visschers (1999) a possible way out of this situation is proposed, which we conceptually discuss below in detail as it provides a deeper insight in the fundamental complexities of inventory management in general MIME inventory systems. The cutting of the *Gordian knot* of the general MIME control problem can be formulated as: *allocation before synchronization*. The control policy proposed, denoted as Synchronized Base Stock (SBS) policies in De Kok & Fransoo (2003), is based on two structural characteristics of the MIME system: the BoM that defines how item coverages relate to each other through the end-items that contain them, and the item cumulative lead times that determines which order release decision precedes the other. This structural foundation translates into a mapping from the original BoM and associated lead times to a set of divergent *decision node structures*. These divergent decision node structures can be considered as the generalization of the serial system structures that are equivalent to convergent systems as shown by Rosling (1989). Each decision node (C, E_C) represents a unique combination of a set of items C and a set of end-items E_C that use these items and a point in time that one or more items not in C are ordered. At this moment it is decided if the coverage constraints of the items in C will be binding for the orders of the items not in C, or not. If the latter is the case, an immediate overage of availability of items in C is created that translates in physical inventory not used at the moment of receipt

of these items in C in their stockpoints. This immediate overage is represented as an inventory level in the decision node associated with C and E_C.

The leaves of a divergent decision node structure are 1-1 related to end-items, which together constitute a subset of E. Each end-item is a leaf in exactly one divergent decision node structure, but there can be multiple divergent decision node structures. The root node of a divergent decision node structure relates to one or more items that are released at some time t and no preceding order release decisions have been made that relate to the end-items that contain the items released at time t. Thus, the root node of the decision node structure contains the longest lead time items of all end-items in the leaves of the divergent structure. More precisely, if items i_1 and i_2 are in the root node of a decision node structure defined by the item set C_0 and end-item set E_{C_0}, then we have

$$L^c_{i_1,k} = L^c_{i_2,k} \forall k \in E_{C_0} \tag{22}$$

This implies that all items associated with the root node of a decision node structure have the same cumulative lead time $L_i k^c$ for all end-items associated with the structure. Divergence in a decision node structure occurs when at some time $t+s$ an item is ordered that is contained in a strict subset of the end-items for which items have been ordered before time $t+s$. The divergent decision node structure unambiguously determines end-item demand coverages by items based on their cumulative lead times. If at time $t+s$ an item is ordered that causes divergence of the decision node structure it belongs to, then the coverage of demand for end-items during the cumulative lead time from time $t+s$ until the moment of replenishment of the end-items is first allocated between the subset of end-items that emerge due to the divergence, and after that the relevant demand coverage constraints can be calculated, similar to above for convergent structures. We refer to De Kok & Visschers (1999) and De Kok & Fransoo (2003) for further details.

Under the modelling assumptions discussed above we can derive the order release decisions for each item in the MIME system at each point in time from the ordering decisions in the decision node structures. If an item is represented in multiple divergent decision node structures, we simply sum the order release quantities from each divergent structure. The SBS policies assume that echelon base stock policies are used for each decision node, and an allocation policy is defined that determines the amount released to each successor in case an item has insufficient availability to satisfy the parent items' orders. The latter case occurs if the echelon stock at (C, E_C) is smaller than the sum of the echelon base stock levels of its successors. If all orders can be satisfied, than the physical stock remaining at (C, E_C) represents the immediate overage created in the pipeline or on-hand of the items in C in the original MIME system.

The decision node structures provide fundamental insight in the mutual dependency of order release decisions taken over time in a MIME inventory system. Each divergent tree represents a strict hierarchy in decision making. Whereas in the original MIME inventory systems we are confronted with the above-mentioned catch-22

of mutual dependencies, under the SBS policies based on the principle of allocation before synchronization, we have disentangled the problem mess, caused by mutual dependencies of decisions in general MIME systems. This disentangling process is demonstrated in section 4.6.

4.5.3 Commonality of items

In De Kok & Visschers (1999) it is assumed that the BoM matrix $(a_{i,j})$ satisfies $a_{i,j} = 0,1, \forall i,j$. If this is not the case then the above described procedure can still be applied, but some preprocessing is required. As stated above we represent the assembly structure as a set of divergent decision node structures, and each node is represented by item set C and end-item set E_C. Let us assume that $|E_C| > 1$ and define $b_{i,k}$ as

$b_{i,k}$:= number of items i required to produce one item of end-item k, $k \in E$

The coefficients $b_{i,k}$ are often referred to as the *Flat BoM* coefficients.

As the divergent structure also presents the way the end-item demand propagates upstream and a decision node is represented only by the item set C, demand propagation must be the same for all items in C, up to a multiplicative factor, i.e.

$$\frac{b_{i_1,k}}{b_{i_2,k}} = c_{i_1,i_2}, \forall k \in E_C, \forall i_1, i_2 \in C \tag{23}$$

Clearly, this condition is not always satisfied in practice. In that case we cannot combine items i_1 and i_2 in a single decision node. Assuming that i_1 has the longest lead time, this implies that upon ordering i_2, the decision node structure diverges from the decision node containing i_1, defined by the sets C and E_C, into multiple decisions nodes containing i_1 and i_2 that satisfy equation (23). So even if i_2 is used by exactly the same end-items as i_1, we do not exploit the commonality in demand for i_1 and i_2, using the portfolio effect from adding demands for all end-items in E_C. Let us argue that we cannot exploit commonality like we might expect in case equation (23) does not hold.

Towards this end let us first assume that equation (23) holds. In that case it is easy to see that the propagated demand for i_1 and i_2 over time interval $(s,t]$ satisfy

$$D_{i_1}(s,t] = c_{i_1,i_2} D_{i_2}(s,t] \tag{24}$$

This implies that, up to a multiplicative constant, the demand processes for i_1 and i_2 are identical. Thus we can use the same control policies for both items over the uncertainty period they have in common, defined by their cumulative lead times. However, if equation (24) does not hold for i_1 and i_2 then the demand processes are clearly not identical and thereby we cannot use the same inventory control policy.

The divergence of the decision node structure at the ordering of i_2 is a *structural expression* of this fact.

	k				i	
	1	2	3	4	5	6
E[D]	50	30	15	5	100	175
σ(D)	25	25	20	10	41.83	91.24
c_D	0.50	0.83	1.33	2.00	0.42	0.52
$b_{5,k}$	1	1	1	1		
$b_{6,k}$	1	2	3	4		

Table 6 Impact of heterogeneity in flat BoM on upstream demand heterogeneity

In Table 6 we illustrate the above by a numerical example. We consider two items 5 and 6 that are common to end-items 1-4. Item 5 has $b_{5,k} = a_{5,k} = 1$ while item 6 has $b_{6,k} = a_{6,k} = k$. We see that the demand for item 5 exhibits the portfolio effect: its coefficient of variation c_D is lower than that of the end-items. However, the demand for item 6 has a coefficient of variation exceeding that of the fast mover, end-item 1. This difference in demand behaviour must be reflected in different control policies for items 5 and 6. While a demand of 50 for item 1 and 30 for item 2 results in the same demand for item 5 as in the situation with a demand of 30 for item 1 and 50 for item 2, this is not true for item 6. Item 6 is sensitive to demand mix changes.

Even though the control policy resulting from the control policy for the divergent decision node structures is not optimal (commonality cannot be exploited if equation (23) is not satisfied, items are allocated to covering future demand for specific end-items too early), it does reflect the complexity induced by BoM structures not satisfying equation (23). The author was first confronted with this phenomenon when studying the commonality structure of TV's at Philips Consumer Electronics in 1990, finding that 100% common items exhibited high coefficients of variation in demand, where at first sight low ones were expected. Looking at the $b_{i,k}$ coefficients for these items revealed that these were different for different end-items k. Our mental point of reference is typically $b_{i,k} = 1$, from which we mentally add up demands for end-items, which yields the expected commonality. However, heterogeneity of the flat BoM coefficients $b_{i,k}$ drives demand volatility of upstream items.

Another common feature of BoM structures is that the same item, i say, is used multiple times at different levels in the BoM. This implies that for item i at different points in time an order release decision must be taken, creating coverage for future demand of end-items over different cumulative lead times. In the MRP logic this is dealt with by low level coding, where gross requirements for the same period are consolidated before an order release decision is taken. By consolidating these decisions, the fundamental relationship between the timing of orders for item i and what part of an order covers which cumulative lead time demand of which end-items is

blurred. In the SBS policy approach this relationship is *exposed* by considering item i at different levels of the BoM as different items. In that way we create a new BoM structure, which is again translated into an associated decision node structure, from which we derive feasible order release decisions, and from which we can directly identify the relationship between order release decisions for different items over time.

When assuming periodic review echelon order-up-to-policies for the divergent decision node structures, we can consolidate the ordering decisions for each item into base stock policies for the original general multi-item multi-echelon inventory system. These base stock policies satisfy the material availability constraints (15). However, in case of fixed lot sizes, this is no longer obvious. Though we can take into account lot sizing constraints for the ordering policies in the divergent decision node structures, adding the resulting order releases may violate the lot sizing constraints in the original system. The decision node structures allow for different review periods for different items, yet at this moment in time it is even unclear how to analyze divergent systems with different review periods. Though nested policies have been studied for serial systems (cf. van Houtum et al. (2007)), it is argued in Karaarslan et al. (2013) that nested policies for the decision node structures are inappropriate in most cases: there is a strong positive correlation between the cost of an item and its lead time. As high variable costs imply small lot sizes and thus frequent ordering, optimal policies for controlling general multi-item multi-echelon systems are unlikely to be nested.

Having said this, some of the issues that come with non-nestedness and lot-sizing can be overcome by *smart modelling*. Firstly, large lot sizes are typical for relatively cheap items. A large lot size of an item decouples the supply chain producing the item from the supply chain using the item. This may imply that we can assume that the item is always available without the need for excessive inventory. Secondly, we can synchronize end-item coverages of short-lead-time items with large lot sizes with the known end-item coverages of long-lead-time expensive items, thereby ensuring that the short-lead-time-items are always available. This idea has been explored in Karaarslan et al. (2013) for a two-component-one-end-item supply chain. Clearly more research is needed to test these hypotheses. In the remainder of this section on uncapacitated MIME systems we assume that all items are controlled by echelon-order-up-to policies with the same review period. For an extensive discussion on modelling MIME systems we refer to Willems (2008). In section 4.10 we discuss the modelling of MIME systems to test empirical validity of the quantitative results derived from applying SBS policies.

4.6 Decision node structure for the case example

We use our case example to explain the creation of the decision node structure associated with an assembly network. In the first step we create the decision node structure associated with each end-item. This yields serial structures in accordance with

the results in Rosling (1989). We use our decision node structure notation where above each decision node (triangle) we show the associated end-item set and below each decision node we show:

- its component set,
- the multiplicity of each item in the component set
- its added cost.

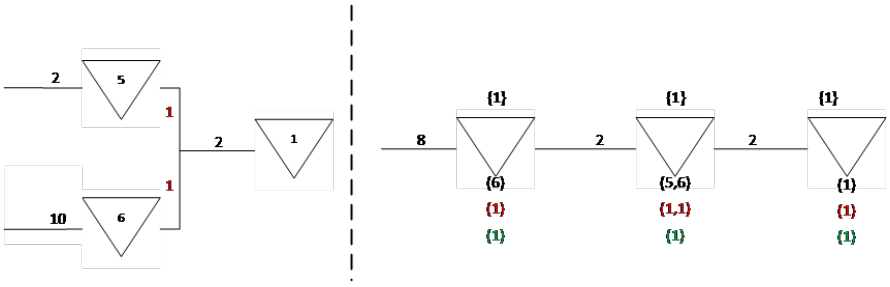

Fig. 4a Item 1: BOM and equivalent serial structure

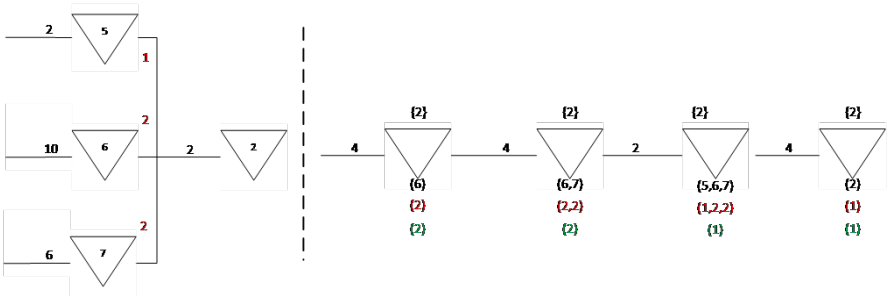

Fig. 4b Item 2: BOM and equivalent serial structure

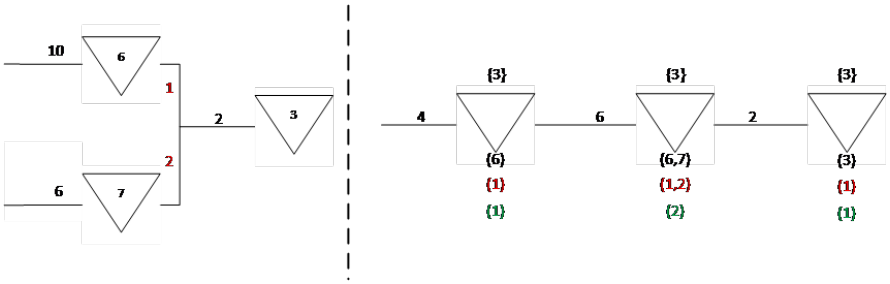

Fig. 4c Item 3: BOM and equivalent serial structure

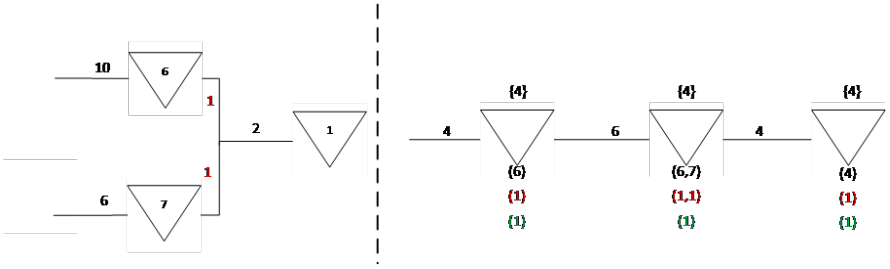

Fig. 4d Item 4: BOM and equivalent serial structure

In the next step we consider the serial structures of items 1 and 2. In both structures item 6 is ordered first, at time 0, say. At time 4 item 7 is ordered, which is used in end-item 2, but not in end-item 1. If we coordinate the control of the supply chains of end-items 1 and 2, this implies that after 4 periods we must allocate the coverage of future demand between item 1 and item 2, such that the coverage of item 7 can be synchronized with the allocated coverage of item 6. At time 8 item 5 is ordered, upon which item 5's coverage of end-item 1 demand is synchronized with the coverage of item 6, and in parallel item 5 coverage of end-item 2 demand is created by synchronization with the joint coverage of item 6 and 7 (as a result of synchronization of these items' coverages created at time 4). At time 10 both item 1 and 2 are ordered, implying that the synchronized coverages of their child items can be used to release orders satisfying the material constraints defined by equations (21).

Another issue must be dealt with before we arrive at the combined structure. Root item 6 is used once in item 1, but twice in item 2. In order to create a consistent combined decision node structure, we normalize the usage of root item 6 to 1 in both end-items. This can be realized by redefining the unit of demand of item 2, which relates to the observation leading to equation (24). The new unit of demand for item 2 is half of the original unit of demand for item 2. This implies that we must multiply the end-item 2 demand per time unit by two. As item 7 is used twice in the original item 2, it is used once in the modified item 2. Concerning item 5, it is used once in item 1, but it is used only 0.5 times in terms of the new demand unit of item 2. Also the multiplicity of the end-item itself is now 0.5, as one root item 6 is consumed by half of the new demand unit of end-item 2. Thus we find the multiplicities for each decision node in figure 5. From the multiplicities we can derive the cost added in each decision node, E.g. when ordering item 5 and synchronizing its coverage with that of items 6 and 7, 0.5 units of item 5 are added to the units of 6 and 7, which implies an added cost of 0.5. Similarly, the added cost associated with the decision node defined by item 2 equals 0.5, too.

We note here that the modification of the demand unit for item 2 is not depicted in 5. The original demand unit can be derived from $b_{6,2} = 2$, i.e. the multiplicity of the root item 6 in end-item 2.

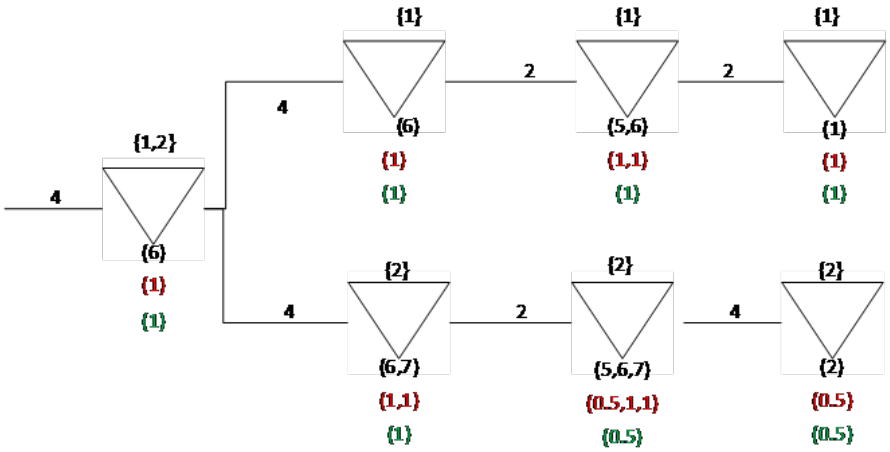

Fig. 5 Decision node structure item 1 and 2

The next step is to combine the serial decision node structure of end-item 3 with that of end-items 1 and 2. In the latter structure we see divergence at time 4 due to the ordering of item 7 that is not used in end-item 5. Item 7 is used in item 3 together with item 6, however not in the same ratio as item 6 and 7 are in item 2. Item 7 and item 6 are used twice in the original demand unit of item 2, but they are used once and twice, respectively, in item 3. Thus item 3 cannot be added to the existing decision node defined by end-item set $\{1,2\}$ and component set $\{6,7\}$. A new decision node is created, defined by end-item 3 and component set $\{6,7\}$, where the multiplicity of 7 equals 2. At time 10 item 3 is ordered and uses the synchronized coverage of items 6 and 7. As the multiplicity of 7 equals 2, the added cost of the decision node associated with end-item set $\{3\}$ and component set $\{6,7\}$ equals 2. The added cost of the decision node associated with end-item set $\{3\}$ and component set $\{3\}$ equals 1.

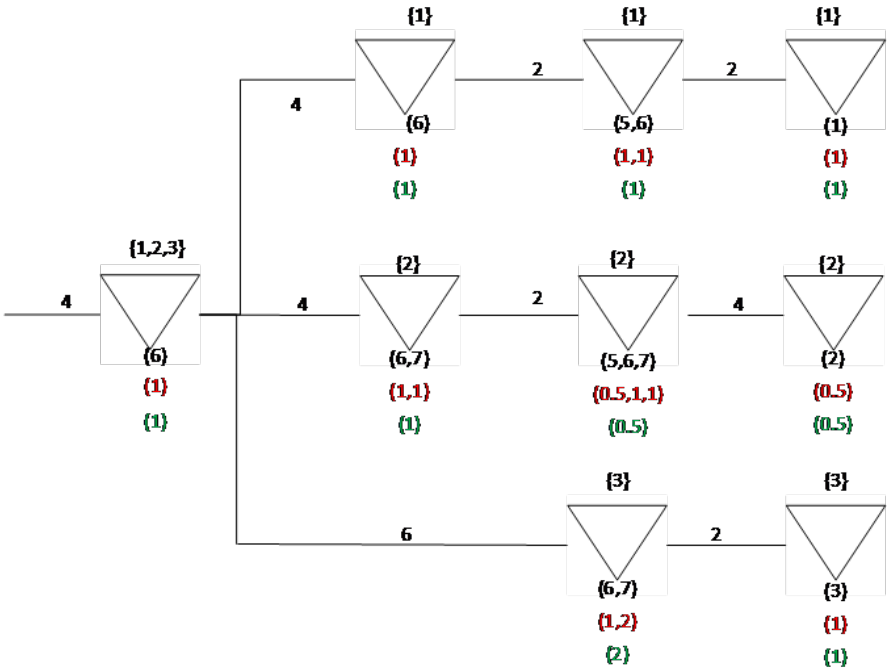

Fig. 6 Decision node structure item 1, 2, and 3

In the final step we combine the serial decision node structure of item 4 with that of items 1, 2, and 3. Item 4 uses items 6 and 7 in a 1-1 ratio. This implies that item 4 can be added to the decision node defined by end-item set $\{2\}$ and component set $\{6,7\}$. As item 4 does not use item 5, at time 8 we allocate the synchronized coverage of items 6 and 7 for items 2 and 4 among these two items, so that item 5 can be synchronized with the allocated coverage for item 2. Thus a new decision node is created that is defined by the end-item set $\{4\}$ and the component set $\{6,7\}$. As the predecessor of this decision node has the same component set, the added value of this node equals 0. At time 10 item 4 is ordered such that it respects the synchronized coverage constraints set by items 6 and 7. Figure 7 depicts the decision node structure for our case example. We have obtained a single divergent decision node structure, as the root item of the structure is in all end-items.

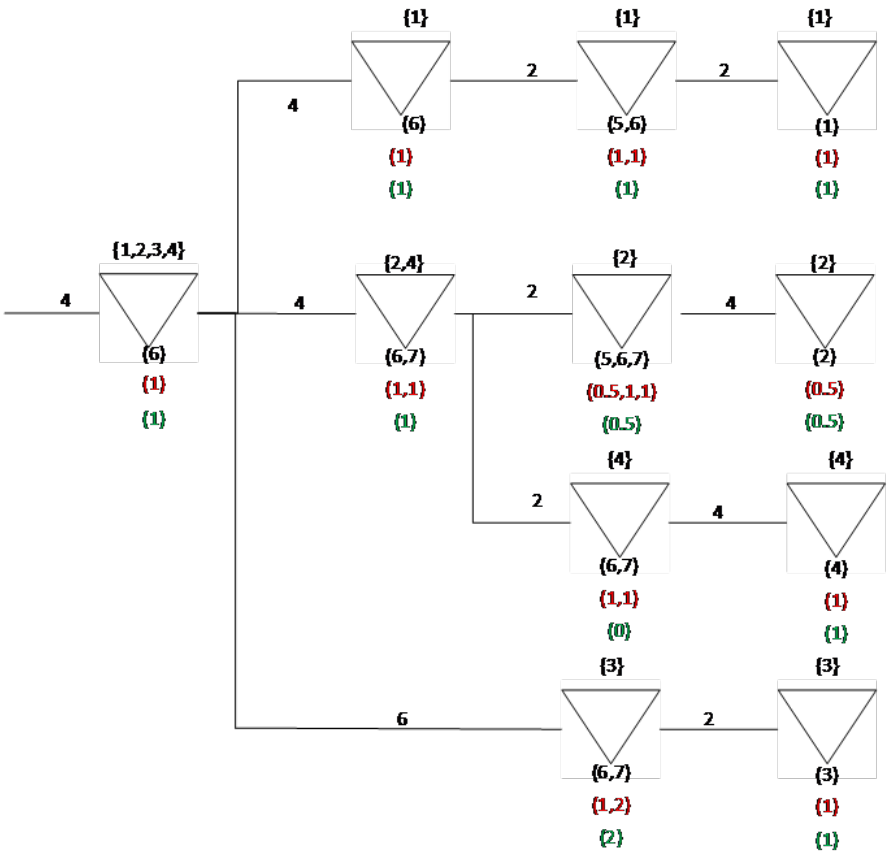

Fig. 7 Decision node structure item 1, 2, 3, and 4

4.7 Control policies for divergent MIME systems

As SBS policies yield divergent decision node structures, finding an optimal SBS policy for a general MIME systems reduces to finding an optimal policy for divergent MIME systems. In Diks & De Kok (1998) the results of Clark & Scarf (1960) for serial systems with linear holding and penalty costs have been generalized to divergent MIME systems:

- Echelon base stock policies are optimal
- Optimal echelon base stock levels can be recursively determined, solving single-variable equations

The results in Diks & De Kok (1998) only hold under the so-called balance assumption defined by Eppen & Schrage (1981), which in turn is equivalent (for our optimization problem) with allowing for negative allocations. If we do not allow for

negative allocations, then finding the optimal policy for divergent MIME systems is as complex and computationally inhibitive as solving the Bellman equations for any stochastic dynamic programming problem (cf. Doğru et al. (2009)). The cost performance of a divergent MIME system computed under the balance assumption is a lower bound of the cost performance under the true optimal policy. Several numerical studies on divergent MIME systems (e.g. Federgruen & Zipkin (1984), Diks & De Kok (1999), and Doğru et al. (2009)) have shown that the policies found under the balance assumption, corrected for negative allocations in a discrete event simulation setting, have a cost performance close to this analytically computed lower bound performance when the following conditions hold:

- End-item coefficient of variation of demand is below 1
- End-item mean demands are similar

The above conditions have deliberately been formulated informally, as the various numerical studies show that it is impossible to derive a precise range in which the balance assumption is effectively non-binding.

In their proof of the optimality of echelon base stock policies for divergent systems, Diks & De Kok (1998) show that the results only hold under optimal allocation policies. However, these optimal allocation policies are intractable. This motivated Van der Heijden et al. (1997) to study so-called linear allocation rules. These are defined by allocation fractions q_{ij} with $j \in V_i$ which sum up to 1. In case the cumulative successor item order release quantity exceeds item i available stock, the excess is allocated according q_{ij} and the allocated excess is subtracted from the original successor item order release quantity. This leads to the following simple control rule for divergent systems.

$$Y_j(t) = S_j - q_{ij}(\sum_{m \in V_i} S_m - Z_i(t))^+, \forall j \in V_i \tag{25}$$

The linear allocation policy ensures feasibility of the order releases, but suffers from the same problem as the "optimal" allocation policy under the balance assumption: it may result into negative order release quantities. To address this issue, Van der Heijden (1997) proposed to determine allocation fractions that minimize (a proxy of) the probability that a negative allocation occurs. In De Kok & Fransoo (2003) the closed-form expression proposed by Van der Heijden (1997) was corrected for a minor error, yielding the following expression for q_{ij}.

$$q_{ij} = \frac{\sigma(D_j)^2}{2\sum_{m \in V_i} \sigma(D_m)^2} + \frac{E[D_j]^2}{2\sum_{m \in V_i} E[D_m]^2} \tag{26}$$

Thus ensuring a low probability of negative allocations requires to allocate a larger part of the excess to successor items with high average demand and high demand volatility. Though Van der Heijden et al. (1997) show that the so-called Balanced Stock rationing improves the quality of the approximations of MIME system

performance measures, such as average inventories and fill rates, it does not remedy the deterioration of performance in high demand volatility scenarios. Here it seems that the only remedy is to keep more stock of items to reduce the occurrence of excesses of demand over item availability. The computational study of Doğru et al. (2009) shows that for high demand volatility scenarios the true optimal policy keeps more inventory upstream than the policy derived under the balance assumption. It is to date unclear how to compute echelon base-stock levels that circumvent the inherent mathematical flaw implied by the balance assumption.

4.8 Generalized Newsvendor equations for divergent MIME systems

Efficient computation of optimal control policies for divergent systems under the balance assumption is based on the results from Diks & De Kok (1998). They show that under the optimal policy so-called generalized Newsvendor equations hold. In order to formulate these equations, we explicitly define the cost structure of a MIME inventory system, which is in line with the cost structure for the SISE inventory system in Section 3.

Table 7 Cost structure of a MIME system

Variable	Definition
H_i	holding cost rate per item i on stock
h_i	echelon holding cost rate per item i on stock
p_k	penalty cost rate per end-item k short
$I_{ki}(t)$	net stock of end-item k in the subsystem defined by W_i
S_i	echelon base stock level of item i

Then Diks & De Kok (1998) show that the optimal policy satisfies the following optimality equations.

$$P\{I_{ki} \geq 0\} = \frac{p_k+H_i-h_i}{p_k+H_k}, \forall k \in E_i. \tag{27}$$

Recall that $I_k i$ denotes the net stock at end-item k in the MIME system W_i of which item i is the root node. From the divergent structure of the subsystems it follows that equations (27) can be solved recursively. Suppose we number the nodes (stockpoints) in the divergent system by low level coding, similar to low level coding in MRP. I.e. the end-nodes of the divergent system have low level code 1, the predecessors low level code 2, etc., and if a node has successors with different low level codes, we number this node as the maximum low level code of the successors plus 1. We start with solving equations (27) for the end-nodes with low-level code

1. These are in fact the same equations as for SISE systems, and provide the optimal (echelon) base stock levels for each end-node (end-item). Then we solve equations (27) for the nodes with low level code 2, etc. Finally we compute the base stock level for the root node of the divergent system, which can be written as

$$P\{X_k \geq 0\} = \frac{p_k}{p_k + H_k}, \forall k \in E. \tag{28}$$

In Diks & De Kok (1998) is it proven that under the optimal policies the solution to the $|E_i|$ equations (27) is a single echelon base stock level for each item (node) i. And equations (28) show that under optimal base stock levels and optimal allocation functions the Newsvendor fractile holds for the end-items. Remember that this holds for MIME systems where the end-items are controlled by an installation stock policy, too, through Theorem 1.

As the optimal allocation policies are intractable, Diks & De Kok (1999) propose to use the linear allocation rules defined above. Although these policies do not satisfy equations (25), as we have $|E_i|$ equations for a single variable S_i, the heuristic based on averaging the $|E_i|$ solutions for the echelon base-stock level performs well.

The recursive solution of equations (27) under the assumption of linear allocation rules as described in equations (25) can be solved by bisection, which ensures that the computation time is linear in the number of nodes in the decision node structure. As the maximum number of decision nodes arises when all end-item decision node structures cannot be combined, the maximum number of bisection equations is bounded by the product of the number of end items $|E|$ and the total number of items. The expressions for the non-stockout probabilities in optimality equations (27) are mathematically intractable, but can be accurately approximated by two-moment fits using mixed-Erlang distributions (cf. De Kok (2003)). The two-moment fits are recursively applied to the shortfalls $Z_i(t)$ that emerge from the linear allocation policy equations (25). Thus, it follows that the computational effort involved in finding close-to-optimal SBS policies is of the same order of magnitude as solving a SISE model for each item in the MIME system. The numerical procedure sketched here is used to determine the parameters of the SBS policies of which we present the performance in the next section.

Before closing this subsection on divergent MIME systems, it should be pointed out that there is an extensive literature on divergent MIME systems, which developed between 1980 and 2000. We refer to Axsäter (2003) for an excellent overview of this literature. We chose to limit our discussion to divergent MIME systems under periodic review, as our aim is to describe the operational control problems that come with MIME inventory systems. Most other approaches are limited to two-echelon systems and focus on the computation of control policies, and do not pay much attention to the allocation problem by assuming e.g. FCFS of replenishment orders from parent items.

4.9 Performance of SBS policies

The SBS policies have a structural foundation in the divergent decision node structures, and a control policy foundation in the base stock policies that are optimal for serial MIME systems according to Clark & Scarf (1960), for convergent MIME systems according to Rosling (1989) and for divergent MIME systems under the so-called balance assumption according to Eppen & Schrage (1981), Federgruen & Zipkin (1984), and Diks & De Kok (1998). All these results have been derived under linear holding and penalty costs. In De Kok & Fransoo (2003) an example is given that shows that SBS policies are non-optimal in general, yet their numerical comparison against LP-based rolling schedule policies shows that SBS policies yield substantially lower long-run average costs than a policy that in practice is believed to be effective for real-life large scale systems (i.e. LP under rolling scheduling).

The main findings in De Kok & Fransoo (2003) were as follows:

- Even though safety stocks for components were set to zero, LP-based control yielded too high average inventories upstream. This is due to the fact that upstream items are cheaper than end-items, so that LP solutions favor to hold stock in upstream items after periods of (coincidentally) low demand.
- End-items with identical BOM structures (including item lead times), identical cost built-up over time and identical demand processes could have widely different safety stocks. This is mainly due to tie-breaking rules as under such a situation the LP problem has infinitely many optimal solutions. This result implies that under identical safety stocks similar end-items would have widely different service levels. This phenomenon has been observed in practice in the context of order release plans proposed by LP-based Advanced Planning and Scheduling (APS) system engines. This could be a reason why APS system planning proposals are often overwritten by planners.
- Similar to the last finding, in case an end-item has slightly lower holding costs than others, it needs much higher safety stocks than the other items. This is due to the extremal solutions generated by LP. In case of overages pushed downstream, the largest part (if not all) of the overage is allocated to the cheapest end-item, and in case of underages the cheapest end-item is likely to get no child item availability allocated at all. This yields a lumpy supply process for the cheapest end-items, which effectively implies that end-item order releases are delayed substantially or not. To compensate for the high end-item lead time volatility, high safety stocks are needed to ensure the required service level.

From the above we concluded that LP tends to use fixed priorities in its allocation mechanism, which were shown to be ineffective in De Kok & Visschers (1999). In their simulation study they explored the impact of the *allocation before synchronization*. Assuming that it is always better to allocate as late as possible, they applied simple allocation rules based on fixed priority, random priority and run-out times. Under these rules, first the priority of end-items is determined and second the available child items are allocated end-item by end-item according to this priority. They found that fixed priorities were performing worst, and the run-out time rule per-

formed best. But the main finding was that the postponement of allocation to the very last moment, i.e. at the moment of order release, using the run-out time rule, hardly improved the customer service levels of the end-items. This finding suggested that the synchronization of item order releases over time is key for supply chain performance. The results in De Kok & Fransoo (2003) confirmed this finding as LP-based control allows for postponement of allocation to the last possible moment.

In Spitter (2005) the comparison of SBS policies against standard rolling scheduling was further explored. With the findings above in mind the rolling schedule heuristics proposed aimed at preventing the extremal allocations that make LP perform badly under demand uncertainty. Two approaches were identified as more effective than LP-based rolling scheduling:

- Adding the linear allocation rules defined by equations (24) as constraints
- Replacing the linear objective function by a quadratic objective function

In the sequel we refer to the rolling schedule policy for which at the start of each period we solve the standard LP model taking into account the feasibility constraints (15) and (16) as the LP_{st}-policy, the rolling schedule policy for which we add the linear allocation constraints (24) to the standard LP model as LP_{alloc}-policy, and the rolling schedule policy based on the model with the feasibility constraints (15) and (16) and a quadratic objective function as QP-policy. The linear objective function takes into account the sum of holding and penalty costs over the planning horizon. When exploring the appropriate formulation of the quadratic objective function, we found that it sufficed to use the original holding and penalty costs as coefficients of the squared physical inventory and squared backlog, respectively, to obtain an effective Quadratic Programming (QP) formulation.

Spitter (2005) uses the same experimental setting as De Kok & Fransoo (2003). The general MIME system under consideration consists of 11 items and is depicted in figure 8. The 11 item product structure consists of 4 end-items. All end-items contain common component 11. End-items 1 and 2 share component 9, while end-items 3 and 4 share component 10. And each end-item contains a specific component.

We assume that the demand for the end-items is stationary. More precisely, demand for end-item i in consecutive periods is i.i.d. We also assume that the demand processes for different end-items are uncorrelated. The mean demand is 100 for all end items. We vary the squared coefficient of variation cv_i^2 for each end-item i as 0.25, 0.5, 1 and 2. The costs structure is as follows

$h_i = h_f = 100$ inventory costs end-items, i=1,2,3,4
$h_i = h_s = 10$ inventory costs specific components, i=5,6,7,8
$h_i = h_{sc} = 30$ inventory costs semi-common components, i=9,10
$h_i = h_c = 50$ inventory costs common components, i=11

For the planned lead times we have analogously to the costs structure the following variables

$L_k = L_f$ nominal lead time end-items, k=1,2,3,4

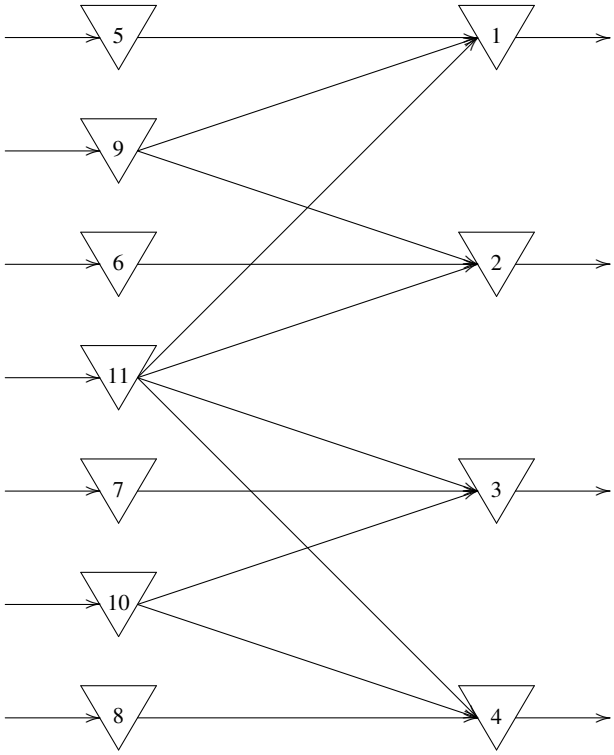

Fig. 8 *Schematic representation of 11 item model.*

$L_i = L_s$ nominal lead time specific components, i=5,6,7,8
$L_i = L_{sc}$ nominal lead time semi-common components, i=9,10
$L_i = L_c$ nominal lead time common components, i=11

We vary the planned lead times (L_s,L_{sc},L_c) as follows (1,2,4), (4,2,1) and (1,4,2). The safety stocks are chosen such that we obtain a non-stockout probability of 95% for each of the control policies. Here we use the safety stock adjustment procedure developed in Køhler-Gudum & De Kok (2002), which exploits the translation property that is the cornerstone of 1. The results of the comparison between the SBS concept and the rolling schedule policies are given in table 8, where we present the inventory costs of each policy and the relative difference between the cost of a rolling schedule policy and the SBS policy (indicated as $\Delta_.$).

Our experiment shows that the QP rolling scheduling policy is superior to both LP policies and comes close to the performance of SBS policies, especially for the cases where the specific components have the longest lead time, i.e. where SBS controls each end-item's supply chain in isolation. The insertion of the linear allocation constraints makes the LP formulation more robust. But we may conclude that the

		Supply chain inventory cost						
cv_i^2	(L_s, L_{sc}, L_c)	*SBS*	LP_{st}	Δ_{LP}	LP_{alloc}	Δ_{alloc}	*QP*	Δ_{QP}
0.25	(1,2,4)	71682	79477	10.9%	73458	2.5%	73332	2.3%
0.25	(4,2,1)	76476	78133	2.2%	78853	3.1%	76765	0.4%
0.25	(1,4,2)	73550	80620	9.6%	74197	0.9%	74172	0.8%
0.5	(1,2,4)	104448	115227	10.3%	106923	2.4%	106679	2.1%
0.5	(4,2,1)	112316	114659	2.1%	115696	3.0%	113381	0.9%
0.5	(1,4,2)	107616	115386	7.2%	108376	0.7%	108362	0.7%
1	(1,2,4)	152203	168533	10.7%	158101	3.9%	157752	3.6%
1	(4,2,1)	165328	169122	2.3%	170590	3.2%	167180	1.1%
1	(1,4,2)	157034	169030	7.6%	159588	1.6%	159722	1.7%
2	(1,2,4)	218551	247578	13.3%	233044	6.6%	232392	6.3%
2	(4,2,1)	245998	249849	1.6%	250075	1.7%	247705	0.7%
2	(1,4,2)	228789	247571	8.2%	235761	3.0%	235694	3.0%

Table 8 *Inventory costs of different SCOP functions.*

allocation before synchronization concept underlying the SBS policies yields the best performance in all cases.

Limitations of the above experiment are two-fold:

- Drawing conclusions from the analysis of a single supply chain structure may not be justified. We developed the 11-item case to represent the key elements of a general MIME system.In De Kok (2001) a real-world case is reported that confirms the findings, but clearly further research is needed.
- We have not been able to optimize the control parameters for items $i \notin E$ under the rolling schedule policies. As we had to use discrete event simulations with a run length of 25000 periods or more to ensure accurate point estimates, where running each case took at least 15 minutes, we considered optimization of control parameters for items $i \notin E$ prohibitive. Given the currently available CPU's, it should be possible now to set up an experiment to optimize the control parameters for all items under rolling schedule policies.

Below further evidence is provided for the effectiveness of SBS policies in comparison to other methods to determine control policies for general MIME systems. We also discuss our empirical findings with the application of SBS policies in real-life situations.

4.10 Empirical validity of SBS policies

The empirical validity of modelling MIME inventory systems assuming SBS policies has been extensively studied in a series of MSc theses (Camp (2002), Janssen (2004), Roose (2007), Bisschop (2007), Uquillas Andrade (2010), Hernandez Wesche (2012), Van Wanrooij (2012), Radstok (2013), Van Pelt (2015), and Van Cruchten (2016)). Though it is unknown what policies are used in practice (cf. Section 2), we

found that the calculated fill rate or ready rate under SBS policies was close to the actual measured ones in several distinct case studies. As inputs we used historical demand data to calculate mean and variance of end-item demand, historical inventory data to calculate average inventories for all items, historical lot size data to calculate the average lot size for each item, BoM data and nominal lead time data. In the software used we converted average lot sizes to nested power-of-2 review periods. Though this yields a heuristic analysis of the MIME systems considered, the favorable outcomes suggest a similar policy independent behaviour of MIME systems as we found for SISE systems (cf. the results in Table 3).

We note here that the nominal lead times should not include safety times for child items being late. The multi-echelon model explicitly takes into account such delays. In many cases we found that the main reason for lateness of orders is late starts due to lack of child item availability. We argued above that under MRP I logic orders are released without formal check on child item availability. Fortunately, in most MRP systems nominal lead times are distinguished from safety lead times. The fact that we found empirical evidence for the validity of multi-echelon models under SBS policies shows that nominal lead times can be adhered to in practice with a high probability.

When testing the empirical validity of the SBS policies we identified a major distinction between MIME systems and SISE systems. In the latter systems, assuming some inventory control policy, there is a 1-1 relationship between average inventory and customer service. In fact, as just mentioned, it seems that there is a 1-1 relationship between average inventory and customer service, irrespective of the inventory control policy. An higher average inventory of an item yields a higher customer service level. This is no longer the case for MIME systems, but in a different sense than may be expected. Our empirical validation does suggest that *average inventories and order frequencies of all items determine end-items' customer service*. However, we found that increasing average inventories of items does not always yield higher end-item service levels. The SBS policies provide a formal argument for this phenomenon.

We mentioned in section 4.5.1 that in convergent MIME systems dead stocks may emerge as a consequence of not respecting the synchronization inequalities (20). Dead stocks are defined as the part of an item inventory that does not add to end-item customer service levels. As under SBS policies we also synchronize (after allocation), dead stocks cannot emerge under SBS policies. This implies that if an SBS policy yields a higher average inventory for some item, while all other average item inventories remain the same, than for at least one end-item customer service level increases, while all other end-item service levels are not reduced. When starting from actual average inventories, we found that dead stocks naturally emerge. To provide further understanding we need to define the inventory levels in the decision node structures, from which we can determine the inventory levels in the underlying general MIME system.

Variable	Definition
$X_C^{E_C}$	stationary physical stock in decision node determined by component set C and end-item set E_C
$\bar{X}_i^{act}$	actual average physical stock of item i
X_i^{dead}	dead stock of item i

With the above definition we can write the average physical stock of an item as the sum of the average physical stock in decision nodes of which the component set contains item i.

$$E[X_i] = \sum_{\{(C,E_C)|i\in C\}} E[X_C^{E_C}], \forall i. \tag{29}$$

But as in real-life dead stocks may emerge as a consequence of violation of equations (20), we can write the following equations from which we must derive the average physical stocks in each decision node.

$$\bar{X}_i^{act} = \sum_{\{(C,E_C)|i\in C\}} E[X_C^{E_C}] + X_i^{dead}, \forall i. \tag{30}$$

As illustrated in the case example in section 4.6 in general MIME systems items are part of component sets in multiple decision nodes. This implies that the set of linear equations is underdetermined, i.e we have more variables than equations. To resolve this we need an objective function. An obvious one would be the system's customer service level and maximize this. This would yield a nonlinear objective function and linear constraints, which may lead to prohibitive computation times. Another approach could be to minimize the value of the dead stocks, which implies solving an LP. Based on our finding on inventory stock positioning, discussed below in section 4.11, we developed a simple greedy heuristic aimed at creating as much as possible inventory downstream in the decision node structures, while respecting the linear constraints. Once the average physical stocks of the decision nodes are known, it is rather straightforward to determine the base stock levels that yield these physical stocks. Given the base stock levels end-item customer service can be calculated. We concluded that it is easier to find optimal policies for general MIME systems under SBS policies than to compute the performance of a general MIME system from historical data.

It should be noted that empirical validation of stochastic models is not trivial. We already explained that it is important to identify the appropriate performance indicators for customer service in section 2 for comparison against model fill rate or model ready rate. We also discussed the impact of high demand volatility in section 3.2.3. In a particular case in a high volume supply chain we found that without removing the end-items with coefficient of variation greater than 1.5 we estimated an aggregate ready rate of 80%, while the actual aggregate ready rate was 97%. After removing the high coefficient of variation end-items, we estimated an aggregate

ready rate of 96.9%, while the actual aggregate ready rate remained 97%. Clearly, aggregation over end-item ready rates is adding apples and oranges, but estimation of individual end-item aggregate ready rates is similar to throwing dices a few times. The underlying idea is that applying a consistent modelling approach causes random variations around the actual targets that cancel out after aggregation. Developing some theoretical background for this statement is highly relevant. In the above-mentioned MSc projects at various companies with various types of supply chains, including process industry, FMCG, pharma, and high-tech we observed that the aggregate model service levels were close to the aggregate real-life service levels.

A generic question concerning empirical validity is how many cases are needed to support empirical validity. We may argue that given the complexity of the multi-dimensional functions involved in computing long-run average service levels and inventory on-hand, it would be very coincidental if the empirical outcomes are in line with the model outcomes, if the model is incorrect. It would be valuable is formal statistical methods can be applied to ensure sufficient rigour.

As has been demonstrated in de Kok et al. (2005), SBS policies can be used as the basis for real-life planning systems that enable to generate material-feasible order releases in large-scale MIME systems within a split second. To-date, SBS policies are the only policies that have this capability. Clearly there have been many other approaches proposed for MIME inventory systems, such as the bounded-demand model (a.k.a. the guaranteed service model, GSM) proposed by Graves & Willems (2000) and the stochastic service model (SSM) proposed by Ettl et al. (2000), but neither of these approaches provide an explicit allocation mechanism that can be used to ensure material-feasible order releases. The GSM and SSM models have been developed to compute safety stocks, not to provide a control policy that can be executed in real-life situations. Below we discuss the consequences of different modelling assumptions for the same MIME inventory system in more detail.

4.11 Positioning inventory in the supply chain

One of the main contributions of the optimization of MIME inventory systems has been the insight into the optimal positioning of on-hand inventory capital in the supply chain subject to end-item service level constraints. The extensive simulation study in Whybark & Yang (1996) is conclusive: minimizing total inventory in the supply chain implies putting 90% or more of the on-hand inventory at end-item level. When minimizing inventory capital across the supply chain subject to end-item service level constraints, the numerical study in De Kok & Fransoo (2003) shows that by far the most part of the supply chain buffers in time and quantity should be concentrated at end-item levels. Other inventory capital is allocated to child items that have parent items that add substantial marginal costs, and in particular when these child items have long lead times. In most of the supply chains studied in the MSc projects mentioned in De Kok (2015), the optimal policies do

not hold any stock for many items, implying that the optimal policies create a flow from purchase items until the end-items. Though flow is strongly advocated by "lean thinkers", we found that most supply chain professionals consider our findings counterintuitive.

These findings are based on assuming SBS policies for inventory control in MIME inventory systems. As mentioned above the GSM and SSM modelling approaches are based on different assumptions:

- Bounded demand assumption
  The bounded demand assumption, also known as the guaranteed service assumption, was originally coined by Graves & Willems (2000) and their paper along with Magnanti et al. (2006) has lead to a substantial number of papers that allow to deal with general structures and various other modelling considerations. The bounded demand assumption allows for setting installation base stock levels that guarantee all demand is satisfied. The positioning of inventory is optimized by setting so-called service times. An item's outbound service time is the time its parent items have to wait for replenishment in addition to the nominal lead time for shipment or production. An item's inbound service time is the additional delay the item itself experiences on top of its nominal lead time. For every choice of the service times the installation base stock level is chosen to cover for demand over the sum of nominal lead time and inbound service time minus the outbound service time.
- Stochastic service assumption
  In Graves & Willems (2003), the stochastic service assumption is described as the assumption that orders are released without a guarantee that they are delivered within the nominal lead time, i.e., after a constant delay. In particular, Graves & Willems (2003) discuss the paper by Ettl et al. (2000) that assumes installation base stock policies and FCFS as allocation policy. Under this assumption, an approximate analysis is possible.

Note that the main assumption underlying the analysis and optimization of SBS policies is the balance assumption.

The three approximations can be applied to the same systems, in particular assuming constant delays and assuming bounded demand is derived from some percentile of the normal or gamma distribution. In Graves & Willems (2003), the bounded demand assumption (GSM) model is compared with the stochastic service (SSM) model for two stylized cases, each derived from a real-world case. We added the solution from assuming SBS policies to this comparison. We present the results of the extended comparison in tables 9 and 10.

The results in tables 9 and 10 show the impact of the modelling assumptions. Not only do we see different total supply chain costs, but more importantly, the allocation of costs across the supply chain is quite different. This reveals that different strategic decisions are proposed under different modelling assumptions. In our view it is important that further research provides deeper insight into the cause of these differences. It also shows the importance of empirical validation of the models to assess whether a modelling assumption is justified. This implies a strong need for

Method	*Safety stock costs*					
	Local DC's	*Central DC*	*Packaging*	*Manufacturer*	*Components*	**Total**
GSM	347,080	383,567	91,827	0	30,524	852,998
SSM	369,664	399,192	43,557	54,467	60,217	927,097
SBS	493,387	3,069	64,301	5,201	43,442	609,399

Table 9 Cost division across battery supply chain

Method	*Safety stock costs*				
	Final assembly	*Main Modules*	*Submodules*	*Other*	**Total**
GSM	607,969	0	0	24,751	632,720
SSM	299,472	187,922	129,613	104,870	721,877
SBS	364,734	104,472	55,403	67,086	591,695

Table 10 Cost division across bulldozer supply chain

- more comparisons of different model paradigms on an identical set of benchmark problems not necessarily meeting all technical assumptions of the models and
- more field studies where a user specifies what he prefers to see as a result of the application of inventory management models, e.g. optimal solutions, robustness of solutions.

5 Capacitated inventory systems

So far we assumed infinite resource capacity. In this section we discuss the implications of taking into account finite capacity resources when modelling and analyzing MIME systems. In subsection 5.1 we discuss the feasibility constraints to be respected when developing control policies for capacitated MIME systems. In subsection 5.2 we briefly discuss our findings when comparing alternative rolling schedule policies. In subsection 5.3 we discuss the main findings from literature on capacitated serial MIME systems. As currently no results on optimal policies for capacitated MIME systems other than for serial systems, we discuss ways to implicitly model finite capacity through nominal lead times.

Taking into account finite resource availability when releasing orders to production or transportation is by no means obvious for three reasons:

- Resources process items some time after the order release decision is taken, and in between events occur that are unknown at the moment of order release
- Resources are mostly used to process more than one item, and the overall processing rate depends on the mix of items to be processed.

- Processing rates can be modified when needed, i.e. resources are flexible, but when what kind of resource flexibility is needed is not known at the moment of order release.

The vast production planning literature may suggest that resources can be modelled easily in the form of constraints on number of units processed per time unit, but we should be aware that most models proposed are deterministic: a planning *instance* is solved. In a real-life situation the immediate order releases from the solution are implemented and some time later a new planning instance is solved, which concerns for a large part the same period of time, e.g. a month or a year. In between the planning instance solutions a myriad of initially unknown events have realized themselves different from the assumptions made in the model. This makes it unclear what objective is appropriate when solving subsequent *deterministic* instances. In section 4.9 we discussed the finding that under demand uncertainty SBS policies outperform LP-based rolling schedule policies for uncapacitated MIME systems (cf. De Kok & Fransoo (2003). Unfortunately, it is not clear how to extend the concept of SBS policies for finite resource capacity. To date, we do not know the optimal policy for capacitated serial MIME systems, unless we impose specific conditions on the capacity constraints. To our knowledge no results are available for divergent capacitated MIME systems.

This section starts with the formulation of necessary conditions for feasible order release decisions, extending inequalities (15) to capacitated MIME systems. Note that we again adopt the modelling approach that we assume nominal lead times for each item. These nominal lead times represent the flow time of released orders to the shop floor as a consequence of stochastic demand, stochastic processing times and finite capacity. In our view the finite capacity cannot substitute for the nominal lead times, as is often claimed by authors with a deterministic point of view. Flow times are not endogenous to the models formulated, but endogenous to the vastly more complex reality, and thereby exogenous to the models formulated. We discuss the findings from a numerical study in Spitter (2005). We conclude this section with a discussion of recent results for serial capacitated MIME systems, which can be seen as building blocks for finding effective policies for general capacitated MIME systems.

5.1 Feasibility of order release quantities

Let us first add the variables that determine the resource structure to take into account. Similar as in De Kok & Fransoo (2003) and Spitter (2005) we assume that each item is processed at a single resource type.

When deriving order release decisions from a quantitative model that is an abstraction of a capacitated MIME inventory system, within the model it should respect the following constraints:

- material availability

Table 11 Defining a multi-echelon inventory system operational characteristics

Variable	Definition
$C_{m,t}$	Amount of capacity available in units of time of resource m in period t
c_i	Time required to process one unit of item i
K_m	set of items that can be processed on resource m

- resource availability
- due date targets based on the nominal lead times

Let us assume we are at time 0 and want to create a feasible production plan respecting the above constraints. Using the notation from Tables 5 and 11 we formulate the following necessary constraints.

$$G_i(t-1,t] = \sum_{i=1}^{N} a_{ij} r_j(t) \leq X_i(t), \forall i, \forall t \geq 0 \tag{31}$$

$$\sum_{s=0}^{t} r_i(s) \geq \sum_{s=1}^{t+1} P_i(s), \forall i, \forall t \geq 0 \tag{32}$$

$$\sum_{m=0}^{t} r_i(m) \leq \sum_{s=1}^{t+L_i} P_i(s), \forall i, \forall t \geq 0 \tag{33}$$

$$\sum_{i \in K_m} c_i P_j(t) \leq C_{m,t)}, \forall m, \forall t \geq 1 \tag{34}$$

Inequality (35) states that the amount of item i needed to release the orders of its parent items cannot exceed item i's availability. Inequality (32) states that the cumulative number of items i received in stock until (and including) time $t+s$ cannot exceed the cumulative number of item i released until time $t+s$. Inequality (33) states that the cumulative number of item i released until time t must have been produced before time $t+L_i$. Under FIFO this ensures that the orders satisfy their due dates based on the nominal lead times. Finally, equation (34) states that the total production time spent on resource type m in period t cannot exceed the available capacity resource type m. Under these conditions we can apply the nominal lead time assumption likewise in the uncapacitated case.

$$Q_i(t+L_i] = r_i(t), \forall i \tag{35}$$

5.2 Comparison of rolling scheduling concepts

Building on the promising results of the Quadratic Programming formulation in a rolling scheduling setting for general MIME systems, Spitter (2005) designed an experiment similar to the one discussed in section 4.9 with a W-structured BOM and various assignments of items to resources. On average the QP formulation approach outperformed the LP formulation approach by 25% lower costs. For robustness of the QP solution it appeared to be important to choose appropriate weights of the quadratic physical inventories and backlogs. Choosing these weights consistent with the allocation fractions defined in equation (26) ensured that in all cases the QP formulation outperformed the LP formulation. Interestingly, under high resource utilization scenarios, LP performed very well. This may be due to the fact that under high utilization long delays for customer orders are inevitable. As nominal lead times are respected for each item, and nominal lead times are much shorter than the customer lead times under high utilization, there is perfect knowledge about customer demand. In that case the allocation problem to be solved is deterministic and linear by nature, given the linear holding and penalty costs.

There is no benchmark available to judge the quality of the QP rolling schedule formulation for general MIME systems. Existing results for serial MIME systems cannot be used, as there is no allocation problem to be solved: each stage in the serial system has its own resource. As mentioned above, increasing computer power may allow for solving for small problems, like the W-system, the value iteration scheme that comes with an SDP formulation of the problem.

5.3 Optimal policies for serial MIME systems

Before discussing results for serial capacitated MIME systems, it is appropriate to mention that for SISE systems without fixed costs is has been shown in Federgruen & Zipkin (1986*a*) and Federgruen & Zipkin (1986*b*) that the optimal policy is a modified base stock policy. I.e. each time the system orders, it orders up to a base stock level, unless the capacity constraint does not allow for that. In that case the order equals the capacity available expressed in units. Given this result for SISE systems, it is not difficult to show that for both serial and divergent systems where only the most upstream stage is capacitated, the optimal policy is a modified echelon base stock policy for the most upstream stage and a base stock policy for all other stages.

In their seminal paper on two-stage capacitated serial MIME systems Parker & Kapuscinski (2004) characterize the optimal policy under the assumption that the most downstream stage has the tightest capacity. Again they find a modified echelon base stock policy. But they also show that if the capacity is most tight at the upstream stage, then a modified echelon base stock policy is no longer optimal. They extend their results to N-echelon serial MIME systems by assuming that the two most upstream stages are capacitated, the most upstream stage has more ca-

pacity than the one but most downstream stage and all other stages have infinite capacity.

Janakiraman & Muckstadt (2009) extend the results of Parker & Kapuscinski (2004) to the case of N-echelon capacitated MIME systems where all stages have the same capacity. Assuming integer demand they find that so-called multitier base stock policies are optimal: each stage has multiple base stock levels that are targeted dependent on the state of the system. This extends the finding of Parker & Kapuscinski (2004) who mentioned a hidden base stock level in case the most upstream stage had the tightest capacity. This multidimensionality of the optimal policy suggests that it cannot be expected that computationally attractive optimal policies can be found for general capacitated MIME systems. Remember that even for uncapacitated divergent systems we need the balance assumption to ensure echelon base stock policies are optimal that can efficiently be computed recursively.

Recently Huh & Janakiraman (2010), Huh et al. (2010), and Huh et al. (2016) explore the properties of serial capacitated MIME systems under base stock policies. The results seem to be promising, as under high end-item service level regimes base stock policies are asymptotically optimal and can be calculated straightforwardly. Such asymptotic arguments may be extended to more complicated MIME systems. This seems to be a promising route to go.

In light of this recent work it is important to mention the contributions of Glasserman & Tayur (1994) and Glasserman & Tayur (1996), who use Infinitesimal Perturbation Analysis (IPA) to find optimal base stock policies for capacitated serial MIME systems. This simulation-based approach seems to be a promising candidate to find optimal base stock policies for general MIME systems, as IPA is a generic method to exploit cost gradient information from discrete event simulation sample paths. On the other hand, the fact that M/M/1 queues need at least 3 million customer arrivals before simulation results converge to long-run results from mathematical analysis, seems to suggest that capacitated MIME systems need long-run times before ensuring that results are representative. One may (rightfully) argue that such a number of jobs or orders is not realistic, but then we should be aware that short-run simulation outcomes are very sensitive to the coincidental inputs regarding customer demands and item processing times. There seems to be a methodological issue here: what to compare simulation outcomes with, if comparing against long-run optimal policies is impossible.

5.4 Implicit modelling of finite capacity

Given the unsolved problem of finding close-to-optimal policies for general MIME systems, while in reality we are faced with finite resource availability, we may want to follow a heuristic route. In our extensive discussion on human behaviour and resource flexibility in section 2 we provided arguments for modelling MIME systems with nominal lead times. Our empirical findings reported in section 4.10 suggest that such modelling yields valid results. That opens the route for decomposition of

the problem of capturing the impact of finite capacity by determining a nominal lead time for each resource or resource-item combination that can be realized with a high probability. An estimate for such a nominal lead time could be derived from queueing models, assuming that the average throughput time from the model is a good estimate of the average flow time of a production order (cf. our discussion in section 2). Though this approach is feasible, given the knowledge on queuing models for production departments (cf. Hopp & Spearman (2011)), it will not be easy to test in simulation environments. This relates to our arguments in section 2 that in practice there is always more flexibility than we model for. Given the state of the art of research on general capacitated MIME systems, it is clearly worthwhile to explore this alternative, albeit less rigourous, perspective.

6 Conclusion

In this paper we discussed various aspects of inventory modelling and analysis. Inventory modelling is about abstracting from real-world details, maintaining only the essence of the inventory management process. This essence is about ordering in time in the right quantity at the right frequency. However, in practice a myriad of feed-back mechanisms and responses are available that have a substantial impact on the performance of inventory management. Because this array of interventions and their impact cannot be captured in mathematical models, we propose to distinguish between inventory management performance before (human) interventions that deviate from the ones assumed in the model and the inventory performance after interventions. This leads to the concept of Intervention Independent Performance (IIP) and Intervention Dependent Performance (IDP). The IIP indicators can be used to derive inventory control parameters, the IDP indicators measure the real performance. Our assumption is that we can identify a relationship between IIP performance and IIP performance. We report the feasibility of this approach implicitly, when discussing the empirical validity of inventory models for both SISE systems and MIME systems. This empirical validity is based on consultancy projects executed by the authors and many MSc thesis projects executed by IE students from Eindhoven University of Technology, of which for some references are provided.

Over almost 6 decades the objective of inventory modelling has been minimization of discounted or long-run average costs. The underlying idea is that we should identify the relevant costs to take into account. However, on the stock market (expected) Return On Investment is the objective. Using the simplest possible inventory model, the EOQ model of Harris (1913), we discuss the findings in Trietsch (1995), who showed that under the ROI maximization objective optimal order quantities depend differently on model parameters than under the cost minimization objective. In short, the classical EOQ is always greater than the ROI-minimizing order quantity R^*. In fact, EOQ equals R^* in a limiting sense, as fixed investments go to infinity.

Probably the most important result discussed is that average inventory and average order frequency determine the end-item customer service levels in both SISE

and MIME inventory systems. This is of prime importance for future research. Baring in mind that optimal policies cannot be found for general MIME systems, it suffices to develop results for mathematically tractable policies that provide full freedom in setting average inventories and average order frequencies for individual items. In that case we can exploit generic optimization techniques to find optimal policy parameters, and possibly exploit specific properties of the relevant objective function to find fast optimization algorithms.

A potential building block for such fast optimization techniques is the generic validity of the Newsvendor fractile under optimal policies within a class of policies. The *net stock translation property* holds for most known SISE inventory control policies. Combination of this result with the Safety Stock Adjustment Procedure of Køhler-Gudum & De Kok (2002) enables efficient simulation-based optimization of the end-item control policy parameters in general MIME systems. Open question is of this result can be extended to optimization of control policy parameters of non-end-items.

The extension of this result holds under Synchronized Base Stock policies. We show that SBS policies can control general MIME systems under the assumption of periodic review and nominal lead times. SBS policies reveal the hierarchical item order release decision structure embedded in general MIME systems. Simulation experiments provide evidence of the cost-effectiveness of SBS policies against mathematical-programming-based rolling schedule policies. Close-to-optimal SBS policies can be determined efficiently by recursively solving one-dimensional generalized Newsvendor equations. Case studies show that modelling real-life supply chains with the SBS policy framework yields empirically valid results.

Major challenges regarding the analysis and optimization of control policies of general MIME systems remain. We mention here the need to develop results for MIME systems with fixed setup and ordering costs that naturally lead to lot sizing constraints. And our short discussion on capacitated MIME systems makes clear that, even though base stock policies seem to be effective for serial MIME systems, it is yet unclear how to analyze general capacitated MIME systems under base stock policies. Presumably, there is another Gordian knot to be cut.

7 Acknowledgements

I am indebted to Henk Tijms who taught me to appreciate development of accurate approximations instead of making assumptions for the convenience of tractability, which could rule out application of results in practice.

I am indebted to Ed Silver, for his book with Rein Peterson that combines rigour and deep insights into what inventory management (and many more subjects in Operations Management) in practice is about. The 1985 edition is my inventory management Bible. But even more am I thankful to Ed for his open mind, when I started sending him letters, questioning some modelling assumptions in this edition.

And I am indebted to Will Bertrand, who provided the conceptual production and inventory control framework within which I could embed my work on multi-echelon inventory systems and hierarchical planning to ensure that the models I worked on could be implemented in practice and at least would provide relevant insights to practitioners.

References

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

Axsäter, S. (2015), *Inventory Control*, International Series in Operations Research & Management Science, Springer International Publishing.
URL: *https://books.google.nl/books?id=v9YjCgAAQBAJ*

Bisschop, J. (2007), Supply chain performance evaluation : application of the synchronised base stock policy in a high-tech complex equipment supply chain with contract manufacturers, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/47041546/631639-1.pdf*

Camp, B. (2002), Startrek supply chain planning : modeling, optimization and generalization, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/46793392/561053-1.pdf*

Clark, A. J. & Scarf, H. (1960), 'Optimal policies for a multi-echelon inventory problem', *Management science* **6**(4), 475–490.

Daganzo, C. F. (2005), *Logistics systems analysis*, Springer.

De Kok, A. (1991*a*), Basics of inventory management, Technical Report 510, 521-525, Tilburg University.

De Kok, A. (1991*b*), Basics of inventory management (part 5): The (r,b,q)-model, Technical report.

De Kok, A. (1993), Demand management in a multi-stage distribution chain, Technical Report 93-35, Eindhoven University of Technology.

De Kok, A. (1998), Inventory control with manufacturing lead time flexibility, Technical Report 345, Eindhoven University of Technology.

De Kok, A. (2001), Comparison of supply chain planning concepts for general multi-item, multi-echelon systems, Technical report, Research Report TUE/TM/LBS/01-03. Eindhoven: Technische Universiteit Eindhoven.
URL: *https://pure.tue.nl/ws/files/3577889/552034.pdf*

De Kok, A. & Fransoo, J. (2003), Planning supply chain operations: Definition and comparison of planning concepts, *in* 'Supply Chain Management: Design, Coordination and Operation', Vol. 11 of *Handbooks in Operations Research and Management Science*, Elsevier, pp. 597–675.
URL: *http://www.sciencedirect.com/science/article/pii/S0927050703110122*

De Kok, T. (2015), Buffering against uncertainty in high-tech supply chains, *in* 'Proceedings of the 2015 Winter Simulation Conference', IEEE Press, pp. 2991–3000.

De Kok, T. G. (2003), 'Ruin probabilities with compounding assets for discrete time finite horizon problems, independent period claim sizes and general premium structure', *Insurance: Mathematics and Economics* **33**(3), 645–658.

de Kok, T. G. (2017), 'Modelling short-term manufacturing flexibility by human intervention and its impact on performance', *accepted for publication in International Journal of Production Research* .

De Kok, T. G. & Visschers, J. W. (1999), 'Analysis of assembly systems with service level constraints', *International Journal of Production Economics* **59**(1), 313–326.

de Kok, T., Janssen, F., Van Doremalen, J., Van Wachem, E., Clerkx, M. & Peeters, W. (2005), 'Philips electronics synchronizes its supply chain to end the bullwhip effect', *Interfaces* **35**(1), 37–48.

Dekker, R., Kleijn, M. & De Kok, A. (1998), The break quantity rule's effect on inventory costs in a 1-warehouse, n-retailers distribution system, Technical Report 7.

Diks, E. B. & De Kok, A. (1998), 'Optimal control of a divergent multi-echelon inventory system', *European journal of operational research* **111**(1), 75–97.

Diks, E. & De Kok, A. (1999), 'Computational results for the control of a divergent n-echelon inventory system', *International Journal of Production Economics* **59**(1), 327–336.

Doğru, M. K., De Kok, A. & Van Houtum, G. (2009), 'A numerical study on the effect of the balance assumption in one-warehouse multi-retailer inventory systems', *Flexible services and manufacturing journal* **21**(3-4), 114–147.

Edgeworth, F. (1888), 'The mathematical theory of banking', *Journal of the Royal Statistical Society* **51**, 113–127.

Eppen, G. & Schrage, L. (1981), 'Centralized ordering policies in a multi-warehouse system with lead times and random demand', *Multi-level production/inventory control systems: Theory and practice* **16**, 51–67.

Ettl, M., Feigin, G. E., Lin, G. Y. & Yao, D. D. (2000), 'A supply network model with base-stock control and service requirements', *Operations Research* **48**(2), 216–232.

Federgruen, A. & Zipkin, P. (1984), 'Computational issues in an infinite-horizon, multiechelon inventory model', *Operations Research* **32**(4), 818–836.

Federgruen, A. & Zipkin, P. (1986*a*), 'An inventory model with limited production capacity and uncertain demands i. the average-cost criterion', *Mathematics of Operations Research* **11**(2), 193–207.

Federgruen, A. & Zipkin, P. (1986*b*), 'An inventory model with limited production capacity and uncertain demands ii. the discounted-cost criterion', *Mathematics of Operations Research* **11**(2), 208–215.

Fogarty, D. & Hoffmann, T. (1983), *Production and Inventory management*, South-Western Publishing Co., Cincinnati, Ohio.

Glasserman, P. & Tayur, S. (1994), 'The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy', *Operations Research* **42**(5), 913–925.

Glasserman, P. & Tayur, S. (1996), 'A simple approximation for a multistage capacitated production-inventory system', *Naval Research Logistics* **43**(1), 41–58.

Graves, S. C. & Willems, S. P. (2000), 'Optimizing strategic safety stock placement in supply chains', *Manufacturing & Service Operations Management* **2**(1), 68–83.

Graves, S. C. & Willems, S. P. (2003), 'Supply chain design: safety stock placement and supply chain configuration', *Handbooks in operations research and management science* **11**, 95–132.

Harris, F. (1913), 'How many parts to make at once', *Factory, The Magazine of Management* **10**(2), 135–136.

Hernandez Wesche, E. (2012), Impacts op implementing a retailer cross-dock on the western europe procter & gamble supply chain, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/46910223/739767-1.pdf*

Hopp, W. J. & Spearman, M. L. (2011), *Factory physics*, Waveland Press.

Huh, W. T. & Janakiraman, G. (2010), 'Base-stock policies in capacitated assembly systems: Convexity properties', *Naval Research Logistics (NRL)* **57**(2), 109–118.

Huh, W. T., Janakiraman, G. & Nagarajan, M. (2010), 'Capacitated serial inventory systems: sample path and stability properties under base-stock policies', *Operations Research* **58**(4-part-1), 1017–1022.

Huh, W. T., Janakiraman, G. & Nagarajan, M. (2016), 'Capacitated multiechelon inventory systems: Policies and bounds', *Manufacturing & Service Operations Management* **18**(4), 570–584.

Janakiraman, G. & Muckstadt, J. A. (2009), 'A decomposition approach for a class of capacitated serial systems', *Operations Research* **57**(6), 1384–1393.

Janssen, F. (2004), Voorraadverlaging door scm bij diosynth, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/46832256/577518-1.pdf*

Karaarslan, A., Kiesmüller, G. & De Kok, A. (2013), 'Analysis of an assemble-to-order system with different review periods', *International Journal of Production Economics* **143**(2), 335–341.

Kiesmüller, G. P., de Kok, T. G., Smits, S. R. & van Laarhoven, P. J. (2004), 'Evaluation of divergent n-echelon (s, nq)-policies under compound renewal demand', *OR Spectrum* **26**(4), 547–577.

Kilger, C., Meyr, H. & Stadtler, H. (2015), 'Supply chain management and advanced planning: concepts, models, software, and case studies'.

Kleinrock, L. (1965), 'A conservation law for a wide class of queueing disciplines', *Naval Research Logistics Quarterly* **12**(2), 181–192.

Køhler-Gudum, C. & De Kok, A. (2002), A safety stock adjustment procedure to enable target service levels in simulation of generic inventory systems, Technical report.
URL: *https://pure.tue.nl/ws/files/3596094/554365.pdf*

Magnanti, T. L., Shen, Z.-J. M., Shu, J., Simchi-Levi, D. & Teo, C.-P. (2006), 'Inventory placement in acyclic supply chain networks', *Operations Research Letters* **34**(2), 228–238.

Nahmias, S. & Olsen, T. L. (2015), *Production and operations analysis*, Waveland Press.

Orlicky, J. (1975), *Material requirements planning*.

Parker, R. P. & Kapuscinski, R. (2004), 'Optimal policies for a capacitated two-echelon inventory system', *Operations Research* **52**(5), 739–755.

Radstok, K. (2013), Fast & slow freight distribution in the fast moving consumer goods industry, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/46922783/754756-1.pdf*

Roose, S. (2007), Rethinking inbound operations management at procter & gamble mechelen : multi-echelon inventory management applied in process industry, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/47021144/627272-1.pdf*

Rosling, K. (1989), 'Optimal inventory policies for assembly systems under random demands', *Operations Research* **37**(4), 565–579.

Silver, E. A., Pyke, D. F. & Thomas, D. J. (2016), *Inventory and Production Management in Supply Chains*, CRC Press.

Silver, E., Pyke, D. & Peterson, R. (1998), *Inventory Management and Production Planning and Scheduling*, Wiley, New York.

Song, J.-S. & Zipkin, P. (2003), 'Supply chain operations: Assemble-to-order systems', *Handbooks in operations research and management science* **11**, 561–596.

Spitter, J. M. (2005), *Rolling schedule approaches for supply chain operations planning*, Technische Universiteit Eindhoven Eindhoven.
URL: *https://pure.tue.nl/ws/files/2089632/200511140.pdf*

Trietsch, D. (1995), 'Revisiting roq: Eoq for company-wide roi maximization', *The Journal of the Operational Research Society* **46**, 507–515.

Uquillas Andrade, R. (2010), An integral supply chain operations planning system for a global pharmaceutical company, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/46966985/668984-1.pdf*

Van Cruchten, A. (2016), Multi-echelon safety stock optimization under supply, process and demand uncertainties as a part of operational risk management, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/46933553/845023-1.pdf*

Van der Heijden, M. (1997), 'Supply rationing in multi-echelon divergent systems', *European Journal of Operational Research* **101**(3), 532–549.

Van der Heijden, M., Diks, E. & De Kok, A. (1997), 'Stock allocation in general multi-echelon distribution systems with (r, s) order-up-to-policies', *International Journal of Production Economics* **49**(2), 157–174.

van Houtum, G.-J., Scheller-Wolf, A. & Yi, J. (2007), 'Optimal control of serial inventory systems with fixed replenishment intervals', *Operations Research* **55**(4), 674–687.

Van Pelt, T. (2015), Multi-echelon inventory management at sligro food group n.v., Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/47009988/789280-1.pdf*

Van Wanrooij, M. (2012), Strategic supply chain planning in a multi-echelon environment : identification of the codp location constrained by controllability and service requirements, Master's thesis, School of Industrial Engineering, Eindhoven University of Technology.
URL: *https://pure.tue.nl/ws/files/46910380/739787-1.pdf*

Vollmann, T., Berry, W., Whybark, D. & Jacobs, F. (2005), *Manufacturing Planning and Control for Supply Chain Management*.

Whitin, T. M. (1953), *The Theory of Inventory Management*, Princeton.

Whybark, D. C. & Yang, S. (1996), 'Positioning inventory in distribution systems', *International Journal of Production Economics* **45**(1-3), 271–278.

Willems, S. P. (2008), 'Data setreal-world multiechelon supply chains used for inventory optimization', *Manufacturing & Service Operations Management* **10**(1), 19–23.

Wilson, J. M. (2016), 'The origin of material requirements planning in frederick w. taylors planning office', *International Journal of Production Research* **54**(5), 1535–1553.

Zipkin, P. (2000), *Foundations of Inventory Management*, McGraw-Hill, Boston.