

10. Research Directions and References

The material that we have provided in the previous chapters has its foundations in existing literature, which will be referenced and put into perspective in this chapter. This also includes some historical remarks which should allow the reader to follow the evolution of supply chain planning up to today.

Furthermore, the models and methods that we have described in the previous chapters are nested in a number of important areas of ongoing research. In this chapter we also give the reader some pointers to the research literature. This is not intended to be exhaustive, but rather to provide the reader with connections to the literature. An exhaustive review is out of question because the respective literatures are so large and they are evolving all the time. However, the references cited here should provide a more than adequate entry point for further access to these research areas.

10.1 Supply Chain Management

A good definition of a supply chain was provided by Ganeshan and Harrison (1995):

A supply chain is a network of facilities and distribution options that performs the functions of procurement of materials, transformation of these materials into intermediate and finished products, and the distribution of these finished products to customers. Supply chains exist in both service and manufacturing organizations, although the complexity of the chain may vary greatly from industry to industry and firm to firm.

Here we give a brief discussion of those supply chain management issues that appear to have an influence on the overall management discussion when it comes to implementation of our models in practice. We start with a historical perspective on the evolution of logistics and end with a couple of thoughts concerning what might become important, especially for production planning within a supply chain.

Since our aim is not writing a general textbook on supply chain management we should mention that over the last couple of years a number of

good textbooks have been written that deal with tactical and strategic supply chain management issues; see, e.g., Handfield and Nichols (1999), Shapiro (2001b), Bowersox et al. (2002), Simchi-Levi et al. (2002), Chopra and Meindl (2003). Consideration of the tactical and operational level, however, can be primarily found in academic journals or in edited books with collections of papers such as Tayur et al. (1999), Klose et al. (2002), de Kok and Graves (2003), Dyckhoff et al. (2004) or Stadtler and Kilger (2005).

10.1.1 The Evolution of Logistics

Before the words “Supply Chain Management” became popular, many of the activities associated with these words were referred to as *logistics* and others as production planning. While we consider production planning later, let us start with the primary objective of logistics. It can be described as the delivery of the right product in the right place at the right time at the least costs; see, e.g., Bowersox (1974), Christopher (1986). Historically, most organizations had considered logistics to be deserving of modest priority.

Prior to 1950, logistics was treated on a fragmentary and often secondary basis. Bowersox (1974) sees two major factors for this neglect and subsequent development. First, prior to the time that computers emerged and before applied analytical tools were generally at the disposal of business, there was no reason to believe that an integrated attack on logistical activities would accomplish improved performance. Second, the prolonged profit squeeze of the early 1950s created an environment conducive to the development of new cost control systems. Integrated logistics provided a productive arena for new methods of cost reduction.

In the period from the mid 1950s to the mid 1960s the concept of integrated logistics crystallized. According to Bowersox (1974) the economic climate at that times was responsible for the “flurry of attention to logistical problems.” During the early 1960s, the horizons of emerging fields of integrated logistics began to expand. Emphasis began to shift towards a penetrating appraisal of the improved customer service capabilities enabled by a highly integrated logistics system.

The period of the 1970s was characterized by the integration of the intra-organizational logistics functions. Davis and Brown (1974) define *logistics management* “as the managerial responsibility of organizing, controlling, directing, staffing, and coordinating product flow from the point of initial procurement to the point of ultimate consumption.” This definition encompasses the activities of purchasing, inventory control, material handling, site determination, warehousing, packaging, order processing, and transportation in a company. Furthermore, it should bridge the gap between the inbound flow of raw materials and the distribution of finished products. Also, Bowersox (1974) emphasized the need for an integrated treatment of intra-organizational functions (refer to the logistics management process depicted in Figure 10.1). He defines the *logistical mission* as the development of “a system that meets the

stated corporate customer service at the lowest possible dollar expenditure.” Development of a satisfactory program requires two levels of adjustment: integration of the logistical system with other corporate systems like the production system, the marketing system, or the finance system and development of total cost balance between logistical system components such as facilities, communication, inventory, transportation, and material movement.

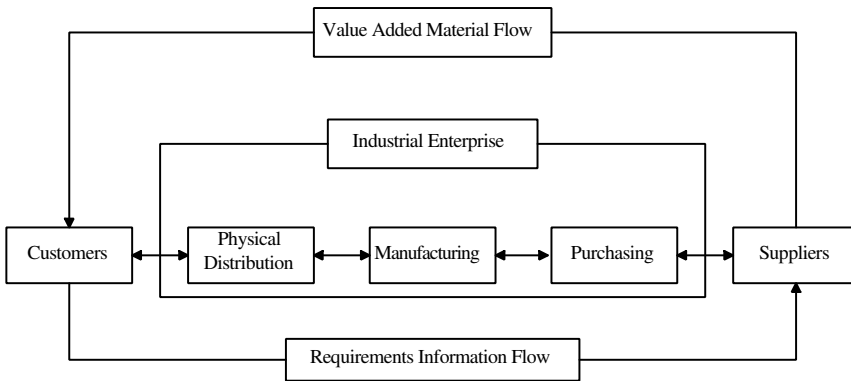


Fig. 10.1. Logistics Management Process from Bowersox (1974)

Since the early 1980s the integration of company-overlapping aspects in terms of logistics have become evident. In this connection the term *Supply Chain Management* was mentioned for the first time by Oliver and Webber (1982) (as noted by Christopher, 1999). Four aspects in which supply chain management differs significantly from classic materials and manufacturing control have been emphasized:

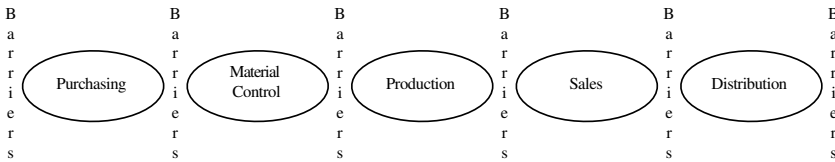
- Supply chain management views the supply chain as a single entity rather than relegating fragmented responsibility for various segments in the supply chain to functional areas.
- Supply chain management calls for strategic decision making. *Supply* is a shared objective of practically every function in the supply chain and is of particular strategic significance because of its impact on overall costs and market share.
- Supply chain management provides a different perspective on inventories which are used as a balancing mechanism of last, not first, resort.
- Supply chain management requires a new approach to systems: integration, not simply interfaces, is the key.

Over time the interest in inter-organizational integration has increased. According to Cooper et al. (1997) there is a definite need for the integration of business operations in the supply chain that goes beyond logistics. In addition to the internal functions of an organization (e.g., logistics, manufacturing, marketing, research) they see a need to integrate external organizations

(e.g., consumer, suppliers, customer) in the product process in order to reduce the time-to-market on new product introductions. The integration of business processes across the supply chain that adds value for customers is what they are calling supply chain management. This definition is similar to the original definition of the term.

The references investigated so far imply that supply chain management is to some extent a new slogan and not a completely new concept. According to its original definition supply chain management means the integration of independent organizations. A closer look at Figure 10.2 reveals that integration has taken place since the middle of the 1950s (stage two). Of course, every stage was regarded as different *independent units* which have been integrated. But at some level of abstraction it does not matter whether only business functions (e.g., procurement, production, sales) or entire companies are considered. The underlying principle of integration is the same in all cases. The difference is only the extent of the *Supply Chain*. This aspect has to be taken into account when the supply chain is defined; see, e.g., Stevens (1989), Christopher (1999), Lee and Billington (1993), Lamming (1996), Larson and Halldorsson (2004).

Stage One: Baseline



Stage Two: Functional Integration



Stage Three: Internal Integration



Stage Four: External Integration



Fig. 10.2. Stages in the Evolution of Logistics from Stevens (1989)

10.1.2 Closed Loop Supply Chains and Reverse Logistics

In Europe manufacturers and importers of various products are legally obliged to take-back and recover their products after use. In response, manufacturers have set up collection and recycling networks eventually including a network of regional storage centers where products that are collected via municipalities and retailers are sorted and consolidated and then shipped to some recycling subcontractors. Two different and yet closely related fields are emerging in this area, reverse logistics and closed loop supply chain management.

Reverse logistics deals with returning waste materials and used products to the producer. While the functionality of logistics as we discussed it above is often referred to as forward logistics, the collection and recovery of used products refers to reverse logistics. Once forward and reverse logistics are combined in the sense of reusing recovered and used products for remanufacturing and delivering those remanufactured products into customer markets again, we speak about closed loop supply chains. We should mention that, in the same spirit as having somewhat loosely coupled supply networks and not necessarily just pure chains (see the definition on page 187), closed loop supply chains are more of a network including cycles than just pure chains.

Closed loop supply chains assume product returns which may imply remanufacturing as well as disposal. Having our lead time discussion in mind, we have a situation where remanufacturing lead time functions can be significantly different from production lead times. Nevertheless, from a modeling standpoint our models may be used as a starting point for developing extended and useful models for reverse logistics and closed loop supply chain management.

The importance of reverse logistics as well as remanufacturing used products into new ones has been widely recognized in the literature and in practice. Good sources on various aspects of closed loop supply chains and reverse logistics can be found, e.g., in Fleischmann et al. (2001), Fleischmann (2001) as well as some of the contributions in Guide Jr. and van Wassenhove (2003), Dyckhoff et al. (2004). Savaskan et al. (2004), e.g., consider the problem of choosing an appropriate reverse channel structure for the collection of used products from customers. Options taken into account are collection by the manufacturer himself, by an existing retailer or by subcontracting. An empirical study within the automotive aftermarket industry was undertaken by Richey et al. (2005) again underlining the importance of this growing field.

10.1.3 The Importance of Information Technology

As mentioned in §10.1.1 prior to the time that computers emerged and before applied analytical tools were generally at the disposal of business, the overall process logistics was treated on a fragmentary basis. Gains in computing speed, coupled with improvements in communication and the flexibility of

data management software, have promoted a range of opportunities prevalent to supply chain management and supply chain planning. However, competitive advantage in supply chain management is gained not simply through faster and cheaper communication of data.

Shapiro (1999) points out that “to effectively apply information technology (IT) in managing its supply chain, a company must distinguish between the form and function of *Transactional IT* and *Analytical IT*.” Transactional IT comprises acquiring, processing, and communicating raw data about a company’s past and current supply chain operations, and the compilation and dissemination of reports summarizing these data. Analytical IT evaluates supply chain decisions based on models constructed from so-called supply chain databases. Usually these databases are derived from transactional data. Analytical IT is comprised of these supply chain decision databases, as well as modeling systems and communication networks linking corporate databases to them. It is concerned with analyzing decisions over short, medium, and long term futures.

According to Shapiro (1999, 2001a) inter-temporal coordination of supply chain decisions has received far less attention than functional coordination. Current efforts to improve supply chain management using IT and business process redesign have only focused on the operational and strategic levels with radically different time frames, planning concerns and organizational needs. Little effort has been made to link analytic tools and databases at the two extreme levels of planning.

Inter-temporal as well as functional integration can be achieved by the application of a suite of optimization modeling systems which take operational, tactical, and strategic aspects into account. These analytical IT systems are linked to overlapping supply chain databases created in large part from data provided by transactional IT systems. Figure 10.3 depicts a possible *Supply Chain System Hierarchy* comprised of optimization modeling systems and transactional systems responsible for inter-temporal and functional integration of supply chain activities in a manufacturing and distribution company with multiple plants and distribution centers.

As IT and supply chain management continue to improve and modeling applications expand, it is expected that more and more companies will implement versions of the entire system hierarchy in the near future. The subsystems of the hierarchy are:

- *Enterprise Resource Planning (ERP)*: managing the company’s transactional data on a continuous, real-time basis, i.e., standardizing data and information systems for order entry, financial accounting, purchasing, and many other functions, across multiple facilities and business units,
- *Materials Requirements Planning (mrp)*: developing net requirements of raw materials and intermediate products to be manufactured or ordered from vendors to meet demand for finished products,

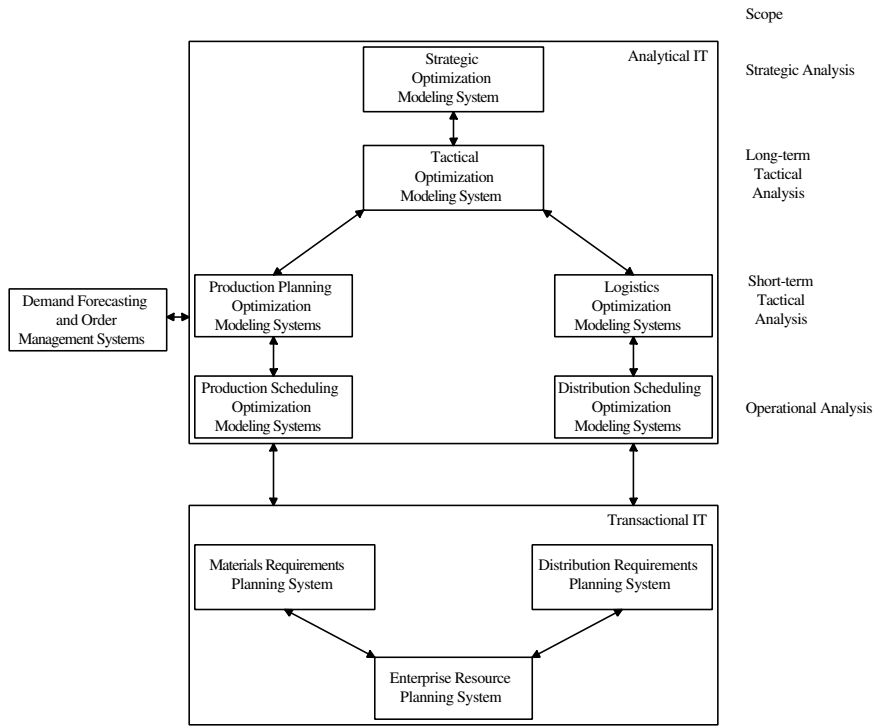


Fig. 10.3. Supply Chain System Hierarchy from Shapiro (1999)

- *Distribution Requirements Planning (DRP):* scheduling in-bound, inter-facility, and out-bound shipments through the company’s logistics network, taking into account a wide range of transportation factors such as vehicle loading and routing, consolidation, modal choice, channel selection, and carrier selection,
- *Demand Forecasting and Order Management:* combining data about current orders with historical data to produce requirements for finished products to be met by operational, tactical, and strategic plans,
- *Production Scheduling:* addressing operational decisions at each plant of a supply chain, e.g., the sequencing of orders on a machine, the timing of major and minor changeovers, or the management of WIP,
- *Distribution Scheduling:* determining vehicle schedules and deciding on a short-term basis which distribution center should serve each market based on inventory availability,

- *Production Planning*: determining multi-period and multi-stage master production plans of manufacturing, along with resource levels and resource allocations, that minimize manufacturing costs,
- *Logistics*: determining a logistics master plan for the entire supply chain that analyzes how demand for all finished products in all markets will be met over the next appropriate period, and assigning markets to distribution centers and other facilities responsible for sourcing them with the goal of minimizing controllable transportation, handling, warehousing, and inventory costs across the entire logistics network,
- *Tactical Optimization*: determining an integrated supply / manufacturing / distribution / inventory plan for the company's entire supply chain over the appropriate couple of periods with respect to minimizing total supply chain costs of meeting fixed demand but also incorporating estimated demands based on forecasts, or to maximize net revenues if product mix is allowed to vary,
- *Strategic Optimization*: analyzing resource acquisition and other strategic decisions faced by the company such as the construction of new manufacturing facilities, development of acquisitions, or the design of supply chains for new products.

The application of any optimization modeling system in the system hierarchy requires inputs from a supply chain database that is created by transforming transactional data found in the ERP, mrp, DRP as well as the forecasting and order management system.

Some of the principles for creating and exploiting decision supply chain databases are as follows; see Shapiro (1999):

- *Adapt Managerial Accounting Principles in Computing Costs*
For the purpose of decision making, the managerial accounting or modeling practitioner must develop relationships between direct and indirect costs, rather than point estimates of them.
- *Aggregation*
As it is not necessary or desirable to describe operations at the individual SKU level for the purpose of strategic or tactical planning, the modeling of supply chain operations should incorporate suitable aggregation of products, customers, and suppliers.
- *Incorporation of External Data Concerning Suppliers, Markets, and Economies*
Transactional data about the company's operations are not sufficient in scope for supply chain analysis of a strategic and tactical nature. An optimization model may require data about supplier costs and capacities, and market conditions for the company's products. Possibly, economic data about long-term prospects for the company's industry and national

economies in which the company operates its supply chain may also be required.

- *Forecast Development*

Analytical data in the supply chain databases help to address the company's future. These data must be based on historical, transactional data. The time horizon is longest for strategic planning, but some extrapolation may be needed even for scheduling purposes (e.g., short-term forecasting of demand from large customers).

- *Parameters of Management Policies*

The decision database must include data and structural inputs reflecting company policies and managerial judgments about risks. The decision variables and constraints that mechanize our optimization models have to be included.

- *Integration of Model Outputs with Model Inputs*

A supply chain decision database has to include output from optimization models (like those developed in this book) as well as the data used in generating the model. Here graphical displays of model inputs and outputs are necessary, too. (This feature is also valuable for comparing and contrasting plans for multiple scenarios.)

The models developed in the earlier chapters are part of initiatives to move down the hierarchy to develop and use optimization modeling systems. Sustainable competitive advantage can only be achieved if IT innovations are combined with complementary organizational and business initiatives as well as a proper linkage to optimization models for production planning.

In the same spirit as shown in Figure 10.3 efforts have been made to group the different tasks and items into a supply chain planning matrix (see Figures 10.4 and 10.5 taken from, and with detailed descriptions, in Fleischmann and Meyr (2003) and some of the contributions in Stadtler and Kilger (2005), Stadtler (2005)). The planning tasks are ordered according to the supply chain processes procurement, production, distribution, and sales in one dimension of the matrix and according to long-term, mid-term, and short-term decisions in the other dimension. Between the entries of the matrix there is a wealth of information and material flows that go along similar lines as we have seen in the system hierarchy above.

While Figure 10.5 refers to specific tasks to be undertaken in supply chain management, Figure 10.4 provides a close linkage to respective systems and can be seen as closely related to the supply chain system hierarchy described above (see also Figure 10.3). It should be noted that some of the planning functionalities described in the matrix in fact also include scheduling aspects. Moreover, it should be noted that the systems may be completely different depending on specific industries (so that some authors propose to have a third dimension for the matrix). Computerized planning tools as they can be deduced from the supply chain planning matrix or from the hierarchy

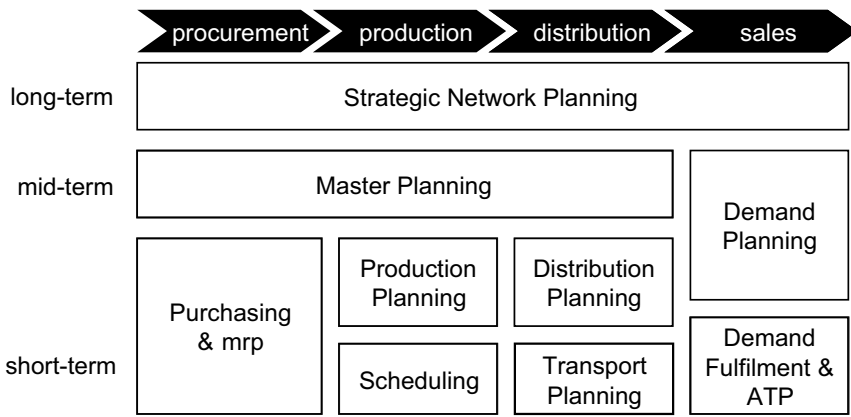


Fig. 10.4. Supply Chain Planning Matrix – System Hierarchy

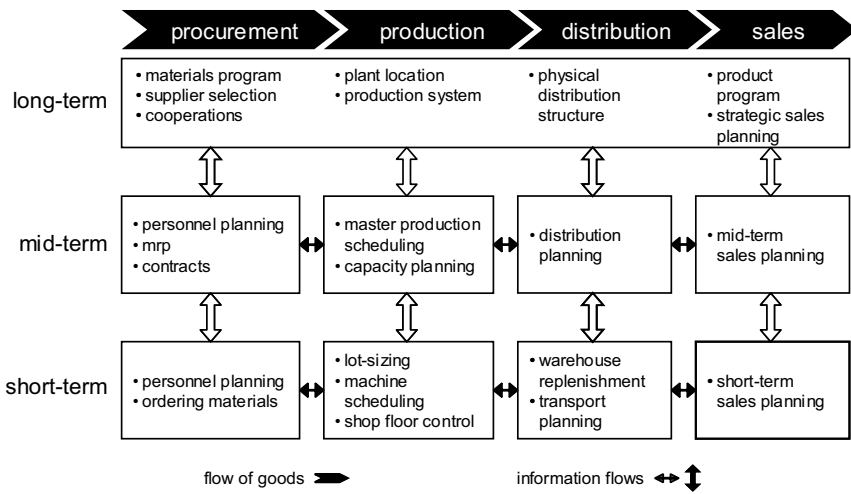


Fig. 10.5. Supply Chain Planning Matrix – Tasks

described above are often called advanced planning systems (APS). That is, APS are computerized planning tools aiming at supporting the various planning processes within supply chains.

Order management refers to management and control of customer orders from the very first customer inquiry to the finished product delivery. Once a customer order enters the systems of a company demand fulfillment is assumed to support taking care about it. Order promising refers to decisions

about the acceptance of orders and setting due dates for incoming orders. Matching demand and supply asks for appropriate allocation of already accepted but yet incomplete orders with respect to yet unassigned stock as well as projected supply. Existing inventory as well as projected production that are not yet assigned to a specific customer order may be used for demand fulfillment; we say that they are “available to promise” (ATP). Beyond ATP quantities one may also consider free or unused capacities of production resources; they are “capable to promise” (CTP). For an in-depth discussion of ATP and CTP see Fleischmann and Meyr (2004).

Many companies use modern information technology to help them gain competitive advantages in the marketplace. The rapid IT advancements have provided tools to enable supply chain partners to share information with each other. Yet, questions concerning the benefits that can be gained through the sharing of information are frequently raised. Several researchers have examined the impact of information sharing on business performance; see, e.g., Lee et al. (2000), Thonemann (2002), Zhao et al. (2002). A survey on the impacts of sharing information including classification of well over 100 references is provided by Huang et al. (2003), another comprehensive treatment is given by Chen (2003).

One of the most common of information sharing issues is the so-called bullwhip effect. It describes the effect that is observed when forecasts for intermediate SKUs within a supply chain are based only on the demand experienced for those SKUs: variability in the demand pattern is magnified. That is, the further “away” we are from customer demand the more volatility is observed. A mitigating solution is to treat intermediate SKUs as dependent items by basing their forecast on the forecast for end items. This is exactly what our models prescribe for intermediate items; however, it is challenging to deploy such models across enterprise boundaries. For references on the bullwhip effect see, e.g., Lee et al. (1997), Chen et al. (2000), Simchi-Levi et al. (2002). It is also closely related to inventory control; see, e.g., Gavirneni et al. (1999) and §10.3.2. Food for thought is provided by Daganzo (2003). He describes control methods for eliminating bullwhip related instabilities without increasing supplier costs and presents approximate cost formulas.

Various games and interactive simulations have become an important part of the pedagogy of supply chain management. They are used to help explain the material and information flows in production and distribution systems. For example, the bullwhip effect is aptly demonstrated by the so-called beer game. Based on these games and the delivery of exercises showing the details and complexity of production and distribution planning, an improved understanding of supply chain issues may be demonstrated; see, e.g., the contributions in Johnson and Pyke (2000). Moreover, the field of production and operations management up to supply chain management reveals many thought provoking issues related to developing teaching material and teaching cases; see, e.g., Kanet and Barut (2003).

While software for supply chain management and enterprise resource planning is still lacking some type of planning functionality the transactional data issues have greatly advanced over the last couple of years. Related hints can be found, e.g., in Knolmayer et al. (2002), Stadtler and Kilger (2005). Finally, we mention Geunes and Pardalos (2003), who provide an annotated bibliography on the extent to which network optimization approaches have contributed to the advancement of supply chain management and financial engineering research.

10.1.4 Supply Contracts

Supply chain integration refers to the connection of at least two parties within a supply chain or network. In addition to activities associated with supply chain management, integration refers to the fact that the parties have to agree upon the way they interact with each other. Especially in the economics literature there is a long history of analysis of the contractual treatment of relationships; see, e.g., Tirole (1988), Katz (1989). Within the modern supply chain discussion the importance of the topic has not diminished but is gaining even more prominence, especially when system-wide optimization is taking place and the incentives discussion as well as concepts like transfer pricing are considered. Supply chains are, by their very nature, based on partnerships. Products as well as the data necessary for an optimization must be exchanged by the partners. What a supply contract may add to this is the explicit specification of a relationship by articulating efficiency measures or metrics that are direct input to our models such as, e.g., lead times or capacity bounds.

A more normative examination of contracts is provided in the operations research literature. For example, Bassok and Anupindi (1997) examine optimal ordering policies for a buyer where there is a pre-specified annual minimum order quantity. Developing contracts also comes along with negotiation processes between supply chain partners. Consider a supplier and its retailer. Due to the supplier's commitments with other customers the negotiation could be, e.g., about the maximum order quantity the retailer can order at a certain price; see, e.g., Homburg and Schneeweiss (2000).

Since various aspects come to mind when dealing with contracts we provide a classification scheme for supply chain contracts following Tsay et al. (1999), Voß and Schneiderei (2002) as well as the literature given there. While the classification could be developed along the lines of timing, pricing, quantity, and quality we follow the idea of having contract clauses in the foreground.

- *Specification of Decision Rights*

Decision rights have to be defined in order to make a contract executable. Using different types of data and information, they constitute a determination of who is allowed to make decisions and within which range of action.

Control mechanisms may be centralized or decentralized as well as global versus local.

- *Information*
Supply partners have to mutually agree about which data and information have to be exchanged at what time and through which channels.
- *Pricing (including Incentives)*
Pricing refers to the specification of financial terms of the supply partners. Commitments have to be made regarding most aspects of the contract such as production costs or retail prices but also the dynamic aspects of cost functions (e.g., allowing for discounts in certain cases, having modified pricing on different lead times based on alternate routings). We include also the implementation of mechanisms to divide profits based on cooperation as well as the arrangement for incentives.
- *Bounds on Purchase Commitments*
Upper and lower quantity bounds for the purchase of goods or products have to be specified. This also includes terms of flexibility, e.g., regarding early or late delivery as well as deviations from previously planned quantity estimates.
- *Allocation*
Defines mechanisms for allocating goods in cases of limited availability.
- *Timing (including Lead Times)*
The time of delivery of items or parts has to be specified. The lead times have to be determined and controlled. When linked with transportation clauses, this includes possible definition of push or pull mechanisms for the ordering processes.
- *Transport*
The contract can include clauses on how the delivery is performed (e.g., by using third party logistics provider) including the definition of penalties for modifications or late arrivals as well as items damaged during transport. This is related to the implementation of rules for having various transport possibilities enabling, e.g., expedited delivery in cases of necessary adjustments in the lead times as described in §6.1.
- *Quality*
Thresholds for the quality of the parts or items as well as allowable modifications like possibilities for upgrades or price reductions in case of unavailability of desired items are articulated in the contract.
- *Buybacks or Return Policies*
Responsibilities for unsold inventory or products with a different quality than agreed upon have to be determined.

All these factors may become part of supply contracts. And yet, soft factors that are beyond our focus on optimization are important issues not to be neglected. This includes language skills along multinational supply chains, cultural differences as well as legal matters. From the optimization perspective we might add complications such as supply chain structures that may be characterized as one-to-one, one-to-many, many-to-one, or many-to-many. For some references see the bibliography in Cachon (2003).

10.2 mrp, MRP II and Beyond

Starting with basic concepts and then extending, we have provided an incremental approach to building optimization models for production planning in supply chains. Somehow this goes along the evolution of software available in this field. While beginning with optimization “at your fingertips,” i.e., doing everything by hand, we now see a tremendous success of software vendors slowly entering the field of real planning functionality.

10.2.1 The Early Steps

Orlicky is widely credited with having “invented” mrp, or at least with popularizing it. The second edition of his seminal work on mrp is Orlicky (1975). This book explains the methods and associated record keeping needed for mrp. Bear in mind that mrp was a tremendous improvement over older management systems that were better suited to a make-to-stock environment. Shorter product life cycles and make-to-order environments require a planning system that anticipates the need for varying mixes of components.

To make better use of mrp, deeper understanding of the relationships between inventory and lead time was needed. Early work by Wight (such as Wight (1974)) helped make mrp successful. In fact, Wight is often credited with inventing MRP II as a way to make mrp logic work correctly.

The actual practice of MRP II was, and is, invented and reinvented by the software firms, consultants, planners and schedulers who make it work. A number of books and a large number of articles provide practical tips for implementing and using MRP II such as Wallace (1990) and Luscombe (1993). In such works, MRP II is referred to as a *closed loop* production planning system because the capacity check is followed by adjustments to the data followed by another execution of mrp and so forth.

A classic text by Vollmann et al. (1988) places mrp, MRP II and the associated planning tools in a broader perspective of planning and scheduling tools popularized in the 1970's and 80's. At the time of their second edition in 1988, mathematical programming approaches to production planning were considered to be an advanced concept. Simultaneous consideration of an objective function along with the materials requirements constraint and

the capacity constraint was treated as part of the human aided processing of MRP II. The book presents a number of sophisticated and practical methods for planning as well as scheduling and forecasting.

At about the same time, some of the shortcomings of the overall philosophy of MRP II were beginning to be discussed (see, e.g., Kanet (1988), Spearman et al. (1990)). This process is on-going (see, e.g., Drexl et al. (1994), Spearman and Hopp (1998)) in the academic literature. The task of bridging the gap between verbally describing the philosophy of mrp and MRP II and deriving mathematical models by means of simple objective functions and constraints has been undertaken by Voß and Woodruff (2000).

Some lines of research address fundamental issues that determine the difficulty of production planning and the successful execution of a plan. For example, there is a large literature concerning SMED, where the seminal work is Shingo (1985). Another vein of research follows Goldratt and Fox (1986) and explores issues related to identifying and managing bottlenecks in production facilities and business in general.

10.2.2 Supply Chain Planning

Based on the supply chain definition on page 187, even if we speak about a chain, we really have or could have a network in mind. Among the more strategic questions in this respect are those related to network design and location. An early reference is Cohen and Lee (1988). More recently, Santoso et al. (2005) consider large-scale supply chain network design problems under uncertainty and discuss a framework for identifying and testing a variety of candidate design solutions. In §10.4 we consider a location-allocation problem. For a review of integrated strategic and tactical models and design issues see Goetschalckx et al. (2002).

Extending simple requirements planning for mrp and beyond is an important topic in the literature. For instance, Graves et al. (1998) study models for requirements planning in multi-stage production inventory systems. They develop a mathematical model to capture many of the planning issues arising in common industrial settings. The idea is to have a planning model for a single stage system as a building block and to extend appropriately. Incorporating locational decisions into a supply chain planning model encounters cross-facility capacity management. For some problems a multi-commodity flow network formulation may be used as a modeling concept; see, e.g., Wu and Golbasi (2004).

Meanwhile, a literature surrounding new technologies from ERP and supply chain management software vendors is beginning to appear (see, e.g., Gumaer (1996) as well as the references in §10.1). One of the evolving products related to supply chain planning is the Advanced Planner & Optimizer from SAP; a detailed discussion and implementation details can be found in Knolmayer et al. (2002), Dickersbach (2004). Especially in these technologies we

believe that solvers based on heuristic search and constraint programming should play, and will play, a prominent role.

A final comment refers to the widespread use of ERP systems all over the world. While common understanding is that mathematics and respective models are universal, many other things like culture or language are not. A thought provoking question for software vendors refers to the markets and the possible use of ERP systems. As supply chains become global we continue to encounter boundaries that are literally beyond planning in our sense. Interesting entries into some literature, e.g., considering questions of the use of mrp, MRP II, and ERP systems in, say, China are Wang et al. (2005), Zhao et al. (2002).

10.3 Production Planning and Scheduling

The models that we have introduced are primarily oriented toward planning for production, but as we noted plans must be constructed with an eye toward the eventual creation of a corresponding schedule. There is a large body of academic literature concerning planning, scheduling and closely related topics. In the subsections that follow we provide some connections to this literature that can be used by the interested reader to gain access to these lines of research. There are a number of outstanding texts devoted to production planning and scheduling such as Johnson and Montgomery (1974), Hopp and Spearman (2000), Nahmias (2004) and, in German, Domschke et al. (1997). The citations in these books also provide a good entry point for further study of the academic literature.

10.3.1 Lot Sizing Models

A Classification Scheme. Lot sizing problems can be characterized by a variety of aspects and classification criteria. The most important distinction refers to deterministic versus stochastic models. While in deterministic models all data are known in advance, in stochastic models data are based on distributions or a measure of uncertainty.

Static models assume that parameter values do not change over the planning horizon (e.g., a continuous demand at the same rate in every period) while dynamic models allow for variation. The planning horizon can be assumed to be finite or infinite. Some of the most important data within lot sizing models are cost data. They may refer to various sorts of cost, such as, e.g., holding costs, setup costs, or production costs.

For the number of products we distinguish between models that consider exactly one and those which take multiple products into account. The latter may imply the difficulty of having to provide plans for these products on several scarce resources. In multi-stage models other than in single-stage models

one considers a given product structure based on given interdependencies between the products as we have used it when defining data $R(i, j)$ based on a bill of materials.

An important distinguishing characteristic of lot sizing formulations is capacity modeling. Capacitated models recognize that some resources are given in a limited number or amount so that planning and scheduling systems need to avoid overutilizing these resources. In situations where there is not enough capacity, one might consider producing or ordering goods after they are actually needed. In this respect one distinguishes between backorder, when this indeed is possible (while paying, e.g., some sort of delay costs), and lost sales (i.e., where the customer refuses to accept any produced items after a given due date).

While in reality we are facing some finite production times, academics often assume that they are able to produce infinitely fast. Depending on the objectives this simplification may make sense.

To exemplify concepts we discuss some modeling aspects regarding a specific lot sizing problem in more detail.

The Capacitated Lot Sizing Problem. The capacitated lot sizing problem (CLSP) in its original form is a simple to state and yet difficult to solve dynamic lot sizing problem that is very similar to our “better” MRP II model (see model **SCPC** in §5.5). To start with a simple version of the models that appear in the research literature, we assume that we have a set of P SKUs that are to be produced within T time buckets.

As in §3.4, we have decision variables, $x_{i,t}$, which specify the quantity of SKU i to be produced in period t . That is, these variables indicate the lot sizes which may change over time. Whenever production of an SKU i takes place in any period we have to pay a setup cost which will be denoted by $C(i)$ using the same cost data as in §5.1.1. In order to enforce the payment of the setup cost, we use an indicator variable, $\delta_{i,t}$, that will be one if any of SKU i will be produced in period t .

To simplify further we assume that there are no lead times, and there is no bill of materials, i.e., all $R(i, j)$ will be 0 and, therefore, omitted from the model. As an important part of the objective function we have to pay a holding cost $H(i)$ for every unit of SKU i that is kept in stock. Inventory of SKU i at period t will be denoted by $I_{i,t}(x, \delta)$ with $I_{i,0}(x, \delta) = I(i, 0)$ indicating the beginning inventory of SKU i and $D(i, t)$ the external demand for SKU i in period t . Then the demand and materials requirement constraints for all $t = 1, \dots, T$ and $i = 1, \dots, P$ read as follows:

$$\sum_{\tau=1}^t x_{i,\tau} + I_{i,t-1}(x, \delta) - \sum_{\tau=1}^t D(i, \tau) - I_{i,t}(x, \delta) \geq 0$$

The meaning of these constraints refers to the fact that in each period we need to have enough inventory from the previous period and enough made

in that period to fulfill the demand. Note that one of the assumptions of the CLSP is that lead times are not explicitly considered, i.e., any production $x_{i,t}$ is available within period t . Whatever remains goes over to the next period as inventory. Equivalently, we could have used the following set of constraints:

$$x_{i,t} + I_{i,t-1}(x, \delta) - D(i, t) - I_{i,t}(x, \delta) \geq 0 \quad i = 1, \dots, P, \quad t = 1, \dots, T$$

The modeling constraint for the production indicators is: $\delta_{i,t} \geq \frac{x_{i,t}}{M}$. As an art of modeling we ask ourselves how small M could be to do what it is supposed to do? One guess refers to the amount yet to be produced, i.e.,

$$M = \sum_{\tau=t}^T D(i, \tau).$$

Furthermore, we have the integer constraint for the production indicator $\delta_{i,t} \in \{0, 1\}$ and the non-negativity of the production $x_{i,t} \geq 0$.

As the CLSP is a capacitated problem, the last thing to take care of is the available capacity. The capacity constraints typically used in the literature can be rewritten to match those used in our models. Let $U(i, t)$ denote the fraction of available time needed to make one unit of SKU i . Then we have the ‘‘capacity constraint’’ $\sum_{i=1}^P U(i, t)x_{i,t} \leq 1$ for all time buckets t . However, authors in the CLSP literature often use slightly different notation and typically refer to time as the scarce resource. In other words, for them $U(i, t)$ represents the fraction of the time bucket consumed by one unit of SKU i .

To summarize, the CLSP is given in Figure 10.6

Minimize:

$$\sum_{t=1}^T \sum_{i=1}^P [H(i)I_{i,t}(x, \delta) + C(i)\delta_{i,t}]$$

subject to:

$$\begin{array}{ll} x_{i,t} + I_{i,t-1}(x, \delta) - D(i, t) - I_{i,t}(x, \delta) \geq 0 & i = 1, \dots, P, \quad t = 1, \dots, T \\ \delta_{i,t} - \frac{x_{i,t}}{M} \geq 0 & i = 1, \dots, P, \quad t = 1, \dots, T \\ \sum_{i=1}^P U(i, t)x_{i,t} \leq 1 & t = 1, \dots, T \\ \delta_{i,t} \in \{0, 1\} & i = 1, \dots, P, \quad t = 1, \dots, T \\ x_{i,t} \geq 0 & i = 1, \dots, P, \quad t = 1, \dots, T \end{array}$$

Fig. 10.6. CLSP Model

A Modification. The CLSP generally considers P products that are not linked by means of a BOM. As a matter of modeling variety we should note that there is a possibility to define sets regarding the BOM, i.e., $Pred(i)$ as

the set of predecessors of SKU i and $Succ(i)$ as the set of successors of i . Then

$$\sum_{j=1}^P R(i, j)x_{j, \tau} \quad \text{can be replaced by} \quad \sum_{j \in Succ(i)} R(i, j)x_{j, \tau}.$$

We will now consider a special situation where there is exactly one end-item and the product structure is strictly convergent, which means that each of the other SKUs is a component in only one subsequent SKU. Keeping the numbering of the SKUs in a low-level-coding we assume a BOM where every SKU but the first has exactly one successor (like in the simple example in Figure 3.1 in §3.1). In such a case we name this successor of i by $succ_i$. The product structure may be viewed as convergent because there is only a single end item, which is SKU 1. As data in this restricted case we have $R(i, succ_i)$ indicating the quantity of SKU i needed to make one $succ_i$. Consider the problem in Figure 10.7.

Minimize:

$$\sum_{t=1}^T \sum_{i=1}^P [H(i)I_{i,t}(x, \delta) + C(i)\delta_{i,t}]$$

subject to:

$$\begin{aligned} x_{1,t} + I_{1,t-1}(x, \delta) - D(1, t) - I_{1,t}(x, \delta) &\geq 0 && t = 1, \dots, T \\ x_{i,t} + I_{i,t-1}(x, \delta) - D(i, t) - R(i, succ_i)x_{succ_i,t} - I_{i,t}(x, \delta) &\geq 0 && i = 2, \dots, P, \quad t = 1, \dots, T \\ \delta_{i,t} - \frac{x_{i,t}}{M} &\geq 0 && i = 1, \dots, P, \quad t = 1, \dots, T \\ \delta_{i,t} &\in \{0, 1\} && i = 1, \dots, P, \quad t = 1, \dots, T \\ x_{i,t} &\geq 0 && i = 1, \dots, P, \quad t = 1, \dots, T \end{aligned}$$

Fig. 10.7. A Lot Sizing Model with Convergent Product Structure

As an advanced exercise one could verify that this model is an example of the **SCPC** model on page 57 under very special conditions. This problem is interesting from a modeling standpoint as pointed out by Afentakis et al. (1984) and Domschke et al. (1997). For instance, we may use a general variable redefinition approach following Martin (1987). While staying with the same data as above as well as the variables $\delta_{i,t}$ as production indicator for production of SKU i in period t and $I_{i,t}(x, \delta)$ to denote the inventory of SKU i at period t some additional binary variables will be defined. We call them availability variables and they are denoted by $z_{i,\tau,t}$ indicating, if they take the value 1, that the demand for SKU i in period t is produced in some period from the first period up to period τ . With these binary variables we have a nice way of re-formulating our problem.

Let us discuss the availability variables in more detail. Whenever the demand $D(i, t)$ is produced in some period $\tau^* \in \{1, \dots, t\}$ then $z_{i,\tau,t} = 0$ for $\tau \in \{1, \dots, \tau^* - 1\}$ and $z_{i,\tau,t} = 1$ for $\tau \in \{\tau^*, \dots, t\}$. That is, $z_{i,\tau,t} = 1$ indicates the (systemwide) availability of that demand.

We have to ensure that the demand for any period t is available not later than in that period, i.e.:

$$z_{i,t,t} = 1 \quad i = 1, \dots, P, \quad t = 1, \dots, T$$

Furthermore, one has to guarantee that the inventory for an SKU, once available, does not get lost:

$$z_{i,\tau+1,t} - z_{i,\tau,t} \geq 0 \quad i = 1, \dots, P, \quad \tau = 1, \dots, t-1, \quad t = 1, \dots, T$$

Once the left hand side of this constraint is one, this indicates a change in the availability and hence the production indicator needs to be forced to 1:

$$z_{i,\tau+1,t} - z_{i,\tau,t} \leq \delta_{i,\tau+1} \quad i = 1, \dots, P, \quad \tau = 1, \dots, t-1, \quad t = 1, \dots, T$$

Finally, we have to consider the BOM which is a convergent product structure for this problem with a single end item, which is SKU 1:

$$z_{i,\tau,t} - z_{succ_i,\tau,t} \geq 0 \quad i = 1, \dots, P, \quad \tau = 1, \dots, t-1, \quad t = 1, \dots, T$$

The objective function has to consider all relevant costs over the planning horizon which are setup costs and holding costs.

$$\begin{aligned} \text{minimize: } & \sum_{t=1}^T \sum_{i=1}^P C(i) \delta_{i,t} + \sum_{t=1}^T \sum_{\tau=1}^{t-1} H(1) D(1, t) z_{1,\tau,t} \\ & + \sum_{i=2}^P \sum_{t=1}^T \sum_{\tau=1}^{t-1} H(i) D(i, t) [z_{i,\tau,t} - z_{succ_i,\tau,t}] \end{aligned}$$

The second term of the objective function considers holding costs for finished amounts of the end item. Whenever an SKU is available but has not yet been used for the production of its successor we have to account for the holding costs as is done in the third term of the objective function. Without loss of generality we can assume that there is a demand for all SKUs at the first time bucket.

With this we can summarize the model in Figure 10.8. The solution to this problem directly implies the values for $x_{i,t}$. The reformulated version of the problem is computationally attractive because the problem has only binary variables and binary constraint coefficients. The use of big M has been eliminated.

Minimize:

$$\begin{aligned} & \sum_{t=1}^T \sum_{i=1}^P C(i) \delta_{i,t} + \sum_{t=1}^T \sum_{\tau=1}^{t-1} H(1) D(1,t) z_{1,\tau,t} \\ & + \sum_{i=2}^P \sum_{t=1}^T \sum_{\tau=1}^{t-1} H(i) D(i,t) [z_{i,\tau,t} - z_{succ_i,\tau,t}] \end{aligned}$$

subject to:

$$\begin{aligned} z_{i,t,t} &= 1 & i &= 1, \dots, P, \quad t = 1, \dots, T \\ z_{i,\tau+1,t} - z_{i,\tau,t} &\geq 0 & i &= 1, \dots, P, \quad \tau = 1, \dots, t-1, \quad t = 1, \dots, T \\ z_{i,\tau+1,t} - z_{i,\tau,t} &\leq \delta_{i,\tau+1} & i &= 1, \dots, P, \quad \tau = 1, \dots, t-1, \quad t = 1, \dots, T \\ z_{i,\tau,t} - z_{succ_i,\tau,t} &\geq 0 & i &= 1, \dots, P, \quad \tau = 1, \dots, t-1, \quad t = 1, \dots, T \\ z_{i,\tau,t} &\in \{0, 1\} & i &= 1, \dots, P, \quad \tau = 1, \dots, t-1, \quad t = 1, \dots, T \\ \delta_{i,t} &\in \{0, 1\} & i &= 1, \dots, P, \quad t = 1, \dots, T \end{aligned}$$

Fig. 10.8. A Reformulation of the Lot Sizing Model from Figure 10.7

Further Dynamic Lot Sizing Models. A paper by Billington et al. (1983) is the first that we know of to propose a model like **SCPC**. They also provide some guidance for problem reduction based on the bottleneck (see §6.6). The problem is viewed directly from the MRP II perspective by Adenso-Díaz and Laguna (1996) where the model includes the possibility of overtime. The objective is to minimize the use of overtime to meet the demand requirements. Other works, such as Tardif and Spearman (1997) take a more algorithmic approach. This work focuses on capacity constrained mrp systems and provides a methodology for finding capacity feasible production plans.

Many of the planning and scheduling models proposed in the research literature are labeled as lot sizing. We have argued against lot sizing, but really we oppose only *static* lot sizes. The literature on dynamic lot sizing is primarily about scheduling, and this can be quite sensible.

An excellent example of this literature is a paper by Trigeiro et al. (1989), which addresses a variant of the CLSP. Their assumptions are consistent with the CLSP in that they ignore sequencing. They assume that production of all parts from a family in a period is done in one batch that requires a setup. The objective function considers holding costs (due to early completion), temporal variations in production costs, and setup costs. The model is very similar to the **SCPC** model given here except that they provide the modeling details necessary to allow production of an SKU to span two time buckets. They also provide details of a special purpose solution technique based on Lagrangian relaxation that proves to be very effective for this problem.

A great deal of effort has been put into various modeling and solution approaches for different modifications and generalizations of the CLSP; see, e.g., Ertogral and Wu (2000), Suerie and Stadtler (2003), Stadtler (2003). The CLSP may be extended in a variety of ways. Let us consider the case where time buckets are so small that in each (“micro”) time bucket only one SKU can be produced. In the literature this problem is referred to as *Discrete Lot Sizing and Scheduling Problem*. Accordingly, the proportional lot sizing and scheduling problem allows for the production of at most two SKUs in one time bucket. For a survey on these and related problems see Drexl and Kimms (1997).

Besides the problems already mentioned, the lot sizing literature is quite large; see Domschke et al. (1997) for a comprehensive survey as well as, e.g., Clark and Armentano (1995), Katok et al. (1998), Tempelmeier and Derstroff (1996), Stadtler (2000). Each of these models makes different assumptions that result in a different model. All of them are intended to determine production quantities. The lot sizing models typically assume a fixed lead time of one period and do not consider alternative routings. These models are similar to the basic model **SCPc** in that they are useful for production scheduling within a factory, but not appropriate for assigning production to factories within a supply chain. Alternatively, our work can also be seen as extending the category denoted as “multiple-stage production planning with limited resources” by Simpson and Erenguc (1996). In the lot sizing literature, if the work considers multiple levels in the bill of materials, the data $R(i, j)$ are often called *production coefficients*.

Naturally, most lot sizing problems cannot be solved without taking into account capacity constraints as we have seen, e.g., for the CLSP. The capacity constraints may refer to a single machine or to non-identical parallel production lines (heterogeneous machines), just to mention some possible complications. Recent interest in solving such problems also considers the application of meta-heuristics; see, e.g., Meyr (2002). Note that often the sequencing problems are treated separately from the lot sizing problems; see §10.3.3. Depending on specific industries this makes sense while for others a simultaneous lot sizing and sequencing seems more appropriate which is in line with our discussion regarding the supply chain planning matrix on page 195.

A prototype modeling and optimization system for lot sizing problems based on the branch and bound principle is provided by Belvaux and Wolsey (2000). The user needs to formulate the problem as a MIP using Xpress-MP (based on a modeling language, see §7.5) taking into account a reserved set of key words for specific lot sizing objects.

10.3.2 Planning and Inventory Control

Planning Horizon. Often data evolve over time. Research on horizon issues focuses on quantifying the diminishing effect of future data on initial

decisions. To formalize the different horizon concepts, we say that a problem has a finite (planning) horizon if a finite number of, say T , time buckets is considered for planning. Assume given numbers $1 \leq t_d \leq t_f < T$. If optimal decisions up to period t_d are independent of the data beyond t_f up to T then t_d is called a decision or planning horizon and t_f is called forecast horizon; see Bes and Sethi (1988). As issues related to planning horizons and forecasting are becoming more and more important in planning and supply chain management, a comprehensive collection of references in this area, as it is provided in Chand et al. (2002), is helpful.

Multiple Routings and Subcontractors. The presence of alternative routings and subcontractors is an important feature of supply chain planning, which has not received enough attention. Chandra and Tombak (1992) look at ways to evaluate the flexibility that is provided by alternate routings. This work is useful during the design of supply chains. A paper by Kamien and Li (1990) examines subcontracting at the aggregate level and discusses economic effects and the structure of the subcontracting relationship. They show that, under certain conditions, subcontracting reduces the variability in production and inventory. This paper, like van Mieghem (1999), provides insight into subcontracting policies but does not prescribe methods for production planning in the presence of multiple routing opportunities.

A paper more directly related to our model for multiple routings is one by Logendran and Ramakrishna (1997) who create a mathematical programming model for the problem of scheduling work in manufacturing cells that have duplicate machines available for bottleneck operations and/or subcontractors who can perform the bottleneck operations. They also give details necessary to use a general-purpose heuristic solution method described in §8.4.6.

Solution methodologies for single products with subcontracting are provided by Atamtürk and Hochbaum (2001). This work also considers the interaction between the operational decision to subcontract and tactical capacity decisions. The paper provides algorithms and insights for both aspects of the problem.

Inventory Control. The models that we have proposed in the preceding chapters are appropriate for plans based on the best information available at the time of the planning process as opposed to *policies* that are parameters for making decisions. For example, a class of inventory control policies are of the basic form Q,R where Q gives the quantity to order or to produce whenever the inventory level gets down to R. An interesting example from this literature is Sobel and Zhang (2001), where ordering policies are considered for a single product when some of the demand is best modeled as being deterministic and some is best modeled as stochastic. Although we might use the word “planning” when describing the process of setting policies, it is clearly not the same activity that we have been concerned with. However, the problem statements are similar.

A problem tangentially related to the one studied in §6.1 is the problem of setting inventory policies when there exist two supply modes with differing lead times. Whittemore and Saunders (1977) look at the problem of determining the appropriate reorder policies when there are two delivery options: one fast and expensive and the other slower and less expensive. They use a stochastic dynamic programming formulation to balance the cost of backlogging and order costs with the cost of holding inventory. Moinzadeh and Nahmias (1988) produce an extension to Q,R policies for a similar model in the continuous case, except that their model includes a cost per stockout incident rather than per unit time.

There is also a considerable literature on setting inventory control policies for an entire bill of materials simultaneously. The multi-echelon inventory and related literature provides methods for setting policies to control inventory levels for a complete production/distribution system under a variety of conditions; see, e.g., Chen (1998), Hwang and Singh (1998), Minner (2000), Roundy (1986). In the simplest case, inventory policies set reorder points that imply minimum planning levels for inventory (i.e., *safety stock*) as we mentioned in §6.3. Formulas for setting safety stock levels are contained in most operations management texts; see, e.g., Martinich (1997). Sethi et al. (2005) have recently published a book that has planning and control models for inventory and supply chain planning with uncertain demand.

One of the newer ideas related to inventory control is that of vendor managed inventory (VMI). In fact this is about partnering. Assuming a supplier and a retailer in a supply chain then VMI is about the supplier taking control over the inventory policies of the retailer as well as the decisions influencing production, distribution and shipment. An interesting study by Disney and Towill (2003) investigates the impact of VMI on the bullwhip effect. The analysis shows that with VMI implementation two sources of the bullwhip effect may be completely eliminated.

Deterioration and Perishability. Deterioration may be regarded as the process of decay, damage or spoilage of products such that they cannot be used for their original purpose anymore, i.e., they gradually undergo a change in storage and lose their utility at least partially. This is in contrast with perishable items that, at some point in time, lose all of their value. Deterioration (and perishability) need to concern supply chain managers when thinking about inventory control as well as when undertaking production planning with products eventually waiting in front of a resource in order to be processed. While classical inventory models assume that inventory can be stored indefinitely in order to meet future demands, this is not realistic for products or items subject to deterioration or perishability. Early literature on deterioration and perishability is chronicled in a survey by Nahmias (1982). A more recent collection of references is compiled by Goyal and Giri (2001).

A few additional examples of this recently expanding literature are, e.g., Benkherouf et al. (2003), Balkhi and Benkherouf (2004). Quality alteration

of products can be affected by random changes in the ambient environment resulting in possible deterioration in some periods while there is no deterioration in others. Aggoun and Benkherouf (2002) consider inventory control approaches in such an environment where, additionally, prices of the items of different quality are allowed to change from period to period in a random fashion.

Hsu (2000) represents an economic lot size model for perishable inventory where stock deteriorating rates depend on the stock age as well as on their production periods. The latter seems realistic as deteriorating items may decay with variable speed at different points in time. Dye and Chang (2003) discuss an economic order quantity system that includes time varying demand and deteriorating items with conditions of permissible delay in payments. Sana et al. (2004) investigate a production-inventory model for a deteriorating item over a finite planning horizon with a linear time varying demand, finite production rate and shortages. Inderfurth et al. (2005) provide analytical insights into optimal lot sizing in a hybrid production/rework environment when both switching from production to rework and vice versa is associated with perceptible cost and time.

10.3.3 Machine Scheduling

Problems in machine scheduling are closely related to lot sizing problems, but consider a higher level of detail than is used in a planning model. The solutions to the planning models developed in the first part of this book are often used as inputs to scheduling problems. Consequently, it is sometimes necessary to solve the scheduling problems as part of the planning process in order to verify that the induced problems have a solution. A simple example of this is given in §8.4.3 where the capacity constraint is replaced with a scheduling problem. A wide variety of scheduling problems have been considered in the research literature.

Textbooks on the topic of scheduling include Baker (1974), Blazewicz et al. (2001) and Brucker (2004). Related material with a focus on manufacturing has been collected in Pinedo (2005). A survey of research concerning MIP modeling of changeovers in production planning and scheduling is provided by Wolsey (1997). A survey of research on sequencing with earliness and tardiness penalties is provided by Baker and Scudder (1990).

An important topic in machine scheduling is the creation of production sequences when there are significant setups, but only a single resource to schedule. As introduced in §8.4.2 we speak of jobs which resume a collection of one or more of the same SKU to be produced. If the jobs to be sequenced are given (perhaps by the planning process) and the time to change from one job to another depends on both jobs, then the problem of finding the fastest sequence can be modeled as the *traveling salesman problem* (TSP). The TSP is a classic problem where one is given a list of cities and the distances between them and asked to find the shortest route that visits all cities. Replacing the

cities by jobs and the distances by production and changeover times, we see that the model applies to sequencing problems as well.

There are a number of other models that are more general than the single machine scheduling problem just introduced. A problem with multiple resources in series where all jobs make use of all resources one after the other is called a *flow shop problem*. To make it more practical, all sorts of modifications are treated in the literature, such as setup times or no-wait constraints. For a survey on the literature in the first case see Cheng et al. (2000).

The second case requires that an SKU or a job once finished on one resource has to be processed immediately after that on the next resource without any interruption until it has left the last resource. For those so-called continuous flow shop scheduling problems the processing of each job has to be continuous, i.e., there must not be any waiting times between the processing of any consecutive tasks regarding this job. To allow processing of a job without interruption on all resources, the order in which the jobs are processed on a resource is the same for all of them (assuming non-zero processing times). If the objective is to finish all jobs as fast as possible this modification reduces again to the TSP. If we strive for minimizing the sum of the completion times of all jobs it relates to a generalization of the TSP which is called time-dependent TSP; see Gouveia and Voß (1995), Fink and Voß (2003).

When there are multiple resources in parallel, such models no longer apply and the scheduling problems become much harder; see, e.g., Belouadah and Potts (1994), Monma and Potts (1993). Another complication is when there are multiple resources needed for each job in some order which varies from job to job. This is referred to as the job shop scheduling problem; see, e.g., Vaessens et al. (1996), Pezzella and Merelli (2000), Meloni et al. (2004). In the fully general case, these two problems are combined to create the hybrid job shop problem; see, e.g., Imaizumi et al. (1998), Gupta et al. (1997).

Both the flow shop and the job shop scheduling problem generalize to the resource constrained project scheduling problem which is concerned with scheduling a set of jobs (or activities) subject to constraints on the availability of several shared resources; see, e.g., Klein (2000). Naturally, one may incorporate temporal constraints allowing the specification of minimal and maximal time lags between two activities; see, e.g., Dorndorf et al. (2000) who consider the minimization of the maximum of the completion times of all activities. Using the resource constrained project scheduling problem together with a basic mrp model allows for a reasonable augmentation leading to options for incorporating capacity constraints and variable lead times as has been investigated, e.g., by Rom et al. (2002).

10.3.4 Aggregation and Part Families

An important form of abstraction in the planning process is the consideration of aggregated parts and SKUs as developed in §6.6. This concept has appli-

cations outside the planning process as well. Given its importance, a variety of research has been conducted concerning this topic.

Some of the earliest work on systematic formation of part families was done by Soviet engineers in the later 1950's to support application of cellular manufacturing. The collection of a group of machines into a cell to process a family of similar parts remains an important topic. There are a number of algorithms available for dividing SKUs into groups to support manufacture by different cells; see, e.g., Miltenburg and Zhang (1991), Venugopal (1999). Another important application of part families is computer assisted process planning (see, e.g., Koenig (1994)). This results in a hierarchy of families driven by a part classification scheme. The classifications are done so that it is possible to reuse process plans for similar parts when a new part is designed. Such classification schemes, therefore, are based on manufacturing characteristics and can be very useful for family formation for the purposes described in §6.6.

Our application of part family formation was for the problem of aggregate planning. This has also been the subject of a significant research literature. In the 1970's researchers at MIT collectively developed a planning system that they referred to as hierarchical production planning (HPP). The HPP system is designed to translate aggregate forecasts into part production requirements. The overall HPP approach is described and justified in a paper by Hax and Meal (1975).

The hierarchy is based on scheduling parts in increasing levels of disaggregation. At the lowest level, parts are scheduled. Before that, families of parts are scheduled (intra-family changes do not require a setup, but inter-family changes do). At the highest level are *types* which are part families grouped according to similarities in production and demand. The scheduling of types is referred to as aggregate planning and the decomposition into family and part schedules is referred to as disaggregation. Bitran and Hax (1977) suggested optimization of sub-problems for the various levels and a rigorous means of linking them. We will now discuss simplified versions of the optimization sub-problems to make their methodology more concrete.

Aggregate planning is a cost minimization problem with respect to temporal variations for production to meet demand forecasts. There are constraints to assure that demand is met, periods are correctly linked via inventory, and capacity is not exceeded. Disaggregation to families is done with the objective of minimizing the total setup *cost*. A survey of disaggregation procedures is contained in Bitran et al. (1981). The disaggregation to parts is done with the objective of minimizing setup cost subject to capacity feasibility, producing the quantities specified in the family disaggregation and keeping inventories within safety stock and overstock limits. Bitran et al. (1982) have proposed a two-stage version of the model which also schedules production of components.

Although this approach is demonstrably better than mrp in some circumstances (see Hax and Candea (1984)), it is not universally applicable. It does not consider due dates. Due dates can be an important factor as flow times drop and competitive pressures increase. Their work is perhaps most appropriate when production of components can and must be begun in response to forecast demand. A more severe problem is that the minimization of setup “costs” is not generally appropriate. In some cases there are actual costs associated with a setup such as materials and supplies. But more generally, the costs are due to lost capacity and in fact depend on the schedule and part mix; see, e.g., Karmarkar (1987).

A hierarchical planning system was also proposed by Spearman et al. (1989) for a specific type of control system known as CONWIP (see Spearman et al. (1990)), where CONWIP stands for “constant work in process.” These systems proposed the use of a hierarchical system as a way of dividing the problem along sensible lines to improve the ability to solve the resulting problems, which was our goal in §6.6 as well. A good reference for hierarchical planning is Schneeweiss (1999). A somewhat extended view of aggregation can also be found in Leisten (1998). A hierarchical planning approach in the light of lot-sizing and scheduling allowing only one setup per period, i.e., at most two products are allowed to be produced in each period, is discussed by Rohde (2004).

10.3.5 Load Dependent Lead Times

Lead times are an important attribute of a product. Consequently, lead times are the subject of research into their causes and effects; see, e.g., Bartezzaghi et al. (1994), Ben-Daya and Raouf (1994), Hopp et al. (1990), de Kok and Fransoo (2003), Lambrecht et al. (1998), Lee et al. (1989), Ornek and Collier (1988), Vendemia et al. (1995). The management of lead times at the control level can be accomplished using control strategies such as CONWIP (see Spearman et al. (1990), Spearman and Hopp (1998)) or Kanban (see Hall (1983), Krajewski et al. (1987), Kimura and Terada (1981), Schonberger (1986)). Use of these methods reduces variations in realized flow times for physical parts due to congestion. But unless the planning systems take lead times into account, the effect will be to increase the waiting time for parts to enter production but not the overall flow time from order release to completion. Clearly, it is better for parts to suffer congestion delays before they have started production rather than after (they can be rerouted much more easily, for one thing) so use of CONWIP and Kanban have significant benefits. These benefits can be extended when coupled with a planning system that is lead time sensitive.

It is useful to consider these systems under the classification scheme of so-called *push* versus *pull* systems. Whereas in a pull system production is initiated as a reaction to present demand, in a push system production is

performed in anticipation of future demand (see, e.g., Karmarkar (1987)). CONWIP and Kanban may be referred to as pull systems.

Our primary interest is in including lead time effects in planning models. Since congestion phenomena go along with bottlenecks causing load dependent lead times it is useful to start with a queuing model in order to obtain some approximations for the key parameters of the capacity constraint formulation or objective functions to be implemented in an aggregate planning model; see, e.g., Buzacott and Shantikumar (1993). As a side-remark we should mention that congestion may be related to heavy loading or heavy traffic, i.e., situations where the average fraction of time at which a server or processor is free is small, or where a machine has little spare capacity. For some mathematical background see, e.g., Kushner (2001).

In order to capture the relationship between system loading and waiting times some authors discuss planning models with *clearing functions*. The idea of clearing functions was introduced by Graves (1986) and further developed by Karmarkar (1987) and Srinivasan et al. (1988). Recently Asmundsson et al. (2002, 2003) employ a clearing function with the aim of modeling the non-linear dependency between lead times and WIP workload.

Zijm and Buitenhok (1996) develop a scheduling model with load dependent lead times that could be extended to a planning model but their most important contribution to our work is that they provide guidance on constructing functions for the waiting time given a loading. That is, they develop a manufacturing planning and control framework for a machine shop that includes workload oriented lead time estimates in order to account for the necessity to consider both lead time and capacity management in a management planning tool. For that purpose they suggest a method that determines the earliest possible completion times of arriving jobs with the restriction that the delivery performance of any other job in the system will not be adversely affected, i.e., that every job can be completed and delivered in time. The goal is to determine reliable planned lead times based on the workload that results in due dates for jobs that can be met and that can be implemented at a capacity planning level, serving there as an input for a final detailed capacity scheduling procedure that also takes into account additional resources, job batching decisions as well as machine setup characteristics. They use queuing network techniques to determine the mean and variance of lead times dependent on lot sizes, production mix and expected annual production volume. Their framework is partly based on the work of Karmarkar (1993a) and Karmarkar et al. (1985) as they employ for each network service station a queuing model with multiple part types.

Our model in §9.1.2 could be seen as making use of a piecewise linear *clearing function* for the tradeoff between loading and waiting time as envisioned by Karmarkar (1989b) with extensions to include multiple routings or subcontractors. Figure 10.9 depicts some possible clearing functions where the constant level clearing function corresponds to an upper bound for capacity

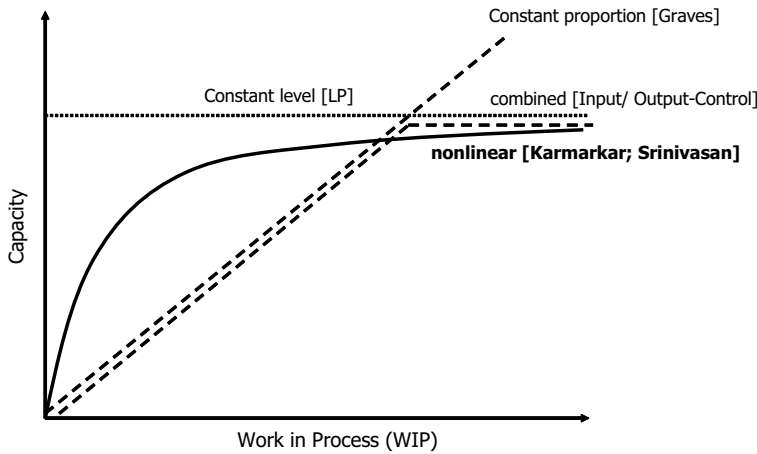


Fig. 10.9. Clearing Functions

as mainly employed by linear programming models. This implies instantaneous production without lead time constraints since production takes place independently of WIP in the production system. The constant proportion clearing function represents a control rule given by Graves (1986) which implies infinite capacity and hence allows for unlimited output. In contrast to the non-linear clearing function of Karmarkar (1987) and Srinivasan et al. (1988), the combined clearing function in some region underestimates and in others overestimates capacity. Moreover, the non-linear clearing function relates WIP levels to output and lead times to WIP levels which are influenced by the behavior of load dependent lead times. Additionally, the slope of the clearing function represents the inventory turn with lead times given by the inverse of the slope (see Karmarkar (1989a)). Special properties of the clearing function allow for formulating a linear programming version in order to develop a model which remains numerically tractable and, therefore, can deal with the problem of combinatorial explosion of company or supply chain size. The clearing function model reflects the characteristics and capabilities of the production system better than models using fixed planned lead times (like mrp).

Lautenschläger and Stadtler (1998) have a well developed method for incorporating load dependent lead times in a capacitated lot sizing model. They base their model on the standard CLSP assumption of a lead time of one period, but their model automatically delays the planned completion if the capacity is heavily utilized. Their work includes guidance on constructing functions for the waiting time given a loading.

Jagannathan and Juang (1998) describe a model where the lead times depend on lot sizes ordered. In this paper, the production level of each SKU

is considered separately for the purpose of determining lead times rather than modeling shared resources and routings. They report on a solution method specifically designed for their formulation. Enns (2001) describes simulation experiments that link static lot sizes, lead times and the performance of an mrp system.

Voß and Woodruff (2004) use a piecewise linear clearing function as suggested by Karmarkar (1989a) in order to model the dependency of lead (or waiting) times in their tactical planning model including multiple routing and subcontractors and highlight the fact that lead times are dependent on the decisions considering the utilization of production resources. Consequently, in order to create realistic, feasible and robust production plans, it is imperative to integrate the effects of load dependent lead times into the tactical planning models. Research issues still remain on how high the utilization of resources could be before being considered a bottleneck and on finding more sophisticated approximations for the clearing function.

We will now give an overview of approaches to formulate clearing functions in order to capture the non-linear relationship between lead times and workload of the production system. A more comprehensive survey is compiled in Pahl et al. (2005).

Graves (1986) studied the extent to which the job flow time (or WIP inventory) depends on the utilization of each resource of a job shop or production stage. The production system is modeled as a network of queues with multiple routing and planned lead times as the decision variable. The clearing function serves as a release control rule at each resource which determines the amount of work performed during a time period which is a fixed portion of the queue of work remaining at the start of the period. As this formulation implies infinite capacity other functional formulations are suggested by, e.g., Karmarkar (1989a,b, 1993b), Srinivasan et al. (1988), Zijm and Buitenhek (1996), Missbauer (2002), Lautenschläger and Stadtler (1998), Asmundsson et al. (2002, 2003), Hwang and Uzsoy (2004), Caramanis and Anli (1999) and Mendoza (2003).

As opposed to Graves (1986), Karmarkar (1989a) and Srinivasan et al. (1988) model the non-linear relationship deriving a clearing function of the following form:

$$\text{Capacity} = \alpha(WIP) \cdot WIP$$

Here, the clearing factor α specifies the fraction of the actual WIP which can be completed, i.e., “cleared” by a resource in a given time period. Missbauer (2002) refers to this factor as the “utilization factor.” In order to give an idea how the clearing function “works” in an aggregate production planning model we refer to the model of Asmundsson et al. (2002) as a reference.

The mathematical programming approach of Asmundsson et al. (2002) models the non-linear dependency between lead times and WIP (workload) by employing a clearing function, too. Special properties of the clearing function allow for formulating a linear programming version in order to develop a

model which remains numerically tractable. In accordance with the procedure of Karmarkar (1989a), Asmundsson et al. (2002) define the performance of a resource (work center) as dependent on the workload. For that reason they use a queuing model including variation coefficients for the service time and the arrival time. Moreover, the utilization of a resource is formulated as a function of the WIP. Also batching and lot sizing have an effect on lead times especially when small batches give rise to frequent setup changes leading to time losses for production, lower throughput and eventual starvation of resources on further production stages.

In order to develop the clearing function, there are two methods available in the literature to date, where the first is the analytical derivation from queuing network models and the second an empirical approximation using a functional form which can be fitted to empirical data. Because of the large amount of details in practical systems the complete identification of the clearing function will not be possible, so we have to work with approximations. Asmundsson et al. (2002) integrate the estimated clearing function in a mathematical programming model where the framework is based on the production model of Hackman and Leachman (1989) with an objective function that minimizes the overall costs. It is assumed that backorders do not occur and that all demand must be met on time. In contrast to Ettl et al. (2000) the non-linear dynamic is incorporated in the clearing function in the constraints and thus not included in the objective function. For more detail see Asmundsson et al. (2003).

Modifications of batch sizes can be a good instrument to control the workload in the system since workload, WIP, safety stocks and lead times (or flow times) are dependent on the choice of the batch size; see, e.g., Zipkin (1986), Karmarkar (1989a), or Karmarkar et al. (1985).

Karmarkar (1989a) develops a capacity and release planning model which explicitly takes into account WIP costs and lead time consequences caused by the production system workload. For that purpose it is based on order releases and batching and applies the traditional capacity planning methodology that combines release planning, master scheduling issues and seasonal planning. Additionally, it aims at surmounting the shortcomings of aggregate production planning models like those of Graves (1986), Kekre and Kekre (1985) and the limitations of input/output - control models. Like Srinivasan et al. (1988), Karmarkar (1989a) uses the non-linear "clearing function" to represent the output as a function of the average WIP in the production system. The form of the curve is also valid for synchronous deterministic flow lines with batched flows. In order to keep things simple, Karmarkar (1989a) considers a discrete period model for a single product production system to proceed to the dynamic reformulation of the model.

Hwang and Uzsoy (2004) combine the work of Karmarkar (1987) and Asmundsson et al. (2002) and add lot sizing in order to show how small or large lot sizes influence the resulting production plans. For that purpose

they present a single-product dynamic lot sizing model which takes into account WIP and congestion using queuing models such as those presented by Karmarkar (1987, 1993b) and develop a clearing function which captures the dependency of the expected throughput of a single-stage production system closely related to the classical Wagner-Whitin model (see Wagner and Whitin (1958)) including setups, expected WIP levels and lot sizes which is then integrated in a dynamic lot sizing model. Results demonstrate that their proposed model provides significantly more realistic performance and, therefore, production plans than models ignoring the relationships between lead times, workload, throughput, production mix and lot sizes (setups).

Before closing this section we should mention that research on workload control has had a wealth of interest especially in the semiconductor industry. This area has developed and considered various workload control concepts investigating general dispatching and order release methods with a focus on wafer fabrication, i.e., the combination of lot release and dispatching strategies used to control the flow of lots through a semiconductor wafer fabrication facility. Uzsoy et al. (1994) provide a general survey and review of production planning and scheduling models in semiconductor manufacturing. A more recent survey with a clear focus on workload control is provided by Fowler et al. (2002). Moreover, we like to point to some references that we feel provide some innovative or thought provoking ideas in one sense or another: Hackman and Leachman (1989), Hung and Leachman (1996), Leachman (1993), Schoemig (1999).

10.4 Transportation

Product movement and transportation is an important part of supply chain management. Transportation modeling may be concerned with routing SKUs between different machines (see §10.3.3) or between a variety of locations. We have proposed extensions to our planning models for some aspects of transportation planning, see §6.4. Within supply networks we are also concerned about shipping, which in our models was assumed to be included in the lead time without careful consideration of how to control it. Most related problems are those “above” our models, i.e., building up a transportation system, and those “below,” i.e., up to now the real movement had been left out on purpose. Much has been written about pure transportation problems as well as on various generalizations; see, e.g., the surveys and collections of Laporte and Gendreau (1995), Cordeau et al. (1998), Kwon et al. (1998), Hall (2003).

A family of models has been developed for optimization of the transportation of goods when production and demand quantities as well as the locations of factories, distribution centers and customers are known in advance. This is a classic optimization model as the basis for teaching and understanding the application of optimization models in addition to being a useful transportation model. One form of the transportation problem is given in Figure 10.10

where the decision variables $t_{i,j}$ are the quantity of SKUs to transport or ship from i to j and the data are the cost to transport or ship from i to j , $T(i, j)$, the demand at j , $D(j)$, and the capacity at i , $C(i)$. Often all i 's are called origins and all j 's destinations. The number of origins is L and the number of destinations is N .

Minimize:

$$\sum_{i=1}^L \sum_{j=1}^N T(i, j)t_{i,j}$$

subject to:

$$\sum_{i=1}^L t_{i,j} = D(j) \quad j = 1, \dots, N$$

$$\sum_{j=1}^N t_{i,j} \leq C(i) \quad i = 1, \dots, L$$

Fig. 10.10. Transportation Problem

While minimizing the cost of all shipped quantities, we have two sets of constraints that guarantee that all the demands are fulfilled and that all the capacity is not overused. This problem and many of its variants have been thoroughly studied over the years and can be solved efficiently. See, e.g., Clarke and Wright (1964), Hall and Racer (1995), Solomon (1987).

This model has been the basis for a plethora of more realistic transportation models. For example, these include the binary transportation problem or the more general design of transportation systems; see, e.g., Belenky (1998), Bhaskaran and Turnquist (1990), Fleischmann (1998), Fleischmann et al. (2001).

These transportation models find optimal quantities, but do not seek to specify operational details such as delivery routes. Shipment cost data are typically based on averages. When we consider models for optimal route planning, we once again encounter the TSP, which is a classic both within transportation planning as well as optimization in general (see §10.3.3).

Because the TSP is a hard problem there has been a lot of work for almost every exact as well as every heuristic method or principle applied to this problem (for excellent surveys on this see, e.g., the books by Lawler et al. (1985), Reinelt (1994), Gutin and Punnen (2002)). Local search approaches are very effective especially regarding real-world and large scale instances of the TSP (see, e.g., Johnson and McGeoch (1997)).

The TSP offers lessons in the art of modeling linear problems. It can be modeled, e.g., with a number of constraints that is linear in the problem size, but a cubic number of variables or a linear number of variables and exponential growth in the number of constraints; see, e.g., Gouveia and Voß

(1995), Padberg and Sung (1991). From a teaching perspective one may learn a lot once the question has been answered, “what makes a TSP a TSP?” (see Sniedovich and Voß (2005)).

The TSP was also used as a very good starting point for various modifications and extensions such as the time constrained TSP or the time-dependent TSP. An important literature concerns the vehicle routing problem where one salesman is replaced (conceptually) by many vehicles perhaps not starting at one and the same depot but at more than one depot; see, e.g., the collection in Toth and Vigo (2002). Extensions may be considered in the same spirit as we have seen it above for the TSP (e.g., with time windows). Finally, we arrive at vehicle scheduling. Another research area is the problem of simultaneously planning production and transportation as proposed by Daskin (1985). Examples from this literature include Blumenfeld et al. (1991), van Buer et al. (1999) and van Roy (1989).

Transportation issues are often linked with locational decisions as well. Usually the production and distribution locations are assumed to have been optimized by a decision process that operates on a longer time scale than the models considered in this book. Nevertheless, as locational decisions and building a distribution network greatly influences transportation costs, these problems are closely related to simultaneous production planning and supply chain design. As an example we borrow a model from Domschke and Voß (1990).

In this model we assume an enterprise which produces P products or SKUs which are used at N different markets. External demand for SKU k at market j is assumed to be $D(j, k)$. While in the transportation model presented above we had decision variables $t_{i,j}$ indicating the quantity of a homogenous good to ship from i to j , these may be modified for handling various products $k = 1, \dots, P$ by adding one more index: $t_{i,j,k}$. Cost values are then given by $T(i, j, k)$, correspondingly. This leads to the following set of constraints guaranteeing that all demands are fulfilled.

$$\sum_{i=1}^L t_{i,j,k} = D(j, k) \quad j = 1, \dots, N, \quad k = 1, \dots, P$$

In this model we further apply some nice way of modeling non-linearities by using, other than SOS2, some piecewise linear functions. Assuming production within certain boundaries we consider a linear function that takes into account (similar to our discussion of marginal transportation discounts in §6.4.4) economies of scale. More specifically let us define $\lambda(i, k)$ and $B(q, i, k)$ indicating a minimum and a maximum amount of production allowed at facility i for SKU k . The values $\lambda(i, k)$ can be thought of as strategic lower limits according to our first abstract optimization model on page 3. Within these boundaries there are piecewise linear functions indicating production costs. To illustrate the concept we first assume that we have $q = 2$ cost functions; see Figure 10.11. The change between functions happens at some amount of

production, say $B(1, i, k)$. That is, any amount of SKUs produced between $\lambda(i, k)$ and $B(1, i, k)$ has a cost of $C(1, i, k)$ per unit while any additional SKU above $B(1, i, k)$ is produced at a different cost rate between $B(1, i, k)$ and $B(2, i, k)$ and accounts for a per unit cost of $C(2, i, k)$. We may view this as different cost rates of production with corresponding cost functions. Naturally, this may be generalized to the case where we have more than one intermediate boundary and a correspondingly enlarged number of piecewise linear functions for production costs.

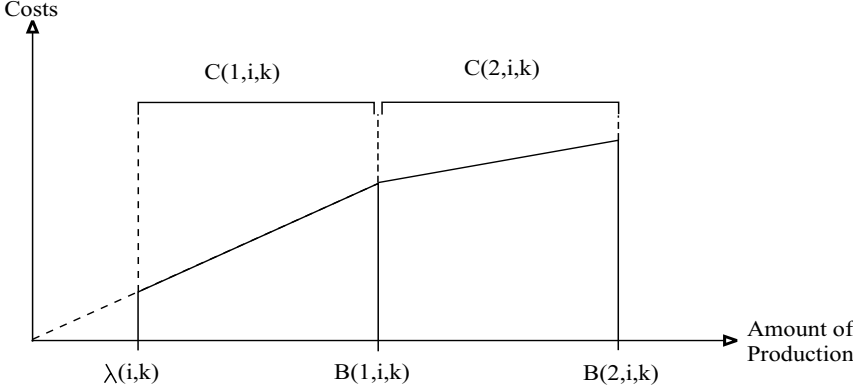


Fig. 10.11. Cost Function

The production of an SKU is automatically allocated to the cost rates. Variables $x_{i,k}^s$ indicate the number of SKUs k produced at location i at cost rate s . Hence, to compute the production quantity for an SKU, the $x_{i,k}^s$ must be summed over all rates. Binary variables $z_{i,k}^s$ indicate whether production of k at i is allowed at cost rate s . We have the following constraints that set these variables correctly depending on the overall number of SKUs produced. If production takes place, it has to be at least $\lambda(i, k)$, and with $B(0, i, k) = 0$ production variables are set as follows.

$$\begin{aligned} \lambda(i, k)z_{i,k}^1 &\leq x_{i,k}^1 && \forall i, \forall k \\ (B(s-1, i, k) - B(s-2, i, k))z_{i,k}^s &\leq x_{i,k}^{s-1} && \forall i, \forall k, s = 2, \dots, q \\ x_{i,k}^s &\leq (B(s, i, k) - B(s-1, i, k))z_{i,k}^s && \forall i, \forall k, s = 1, \dots, q \end{aligned}$$

Production is going to take place at any of at most L locations depending on whether we use these locations or not. Whenever we open up a facility at a specific location, say i , then this implies some fixed costs of $F(i)$. Using variables y_i indicating whether to open a facility at location i or not, these location variables may also be used to allow the initialization of production indicator variables:

$$z_{i,k}^1 \leq y_i \quad i = 1, \dots, m, \quad k = 1, \dots, P$$

P	Number of SKUs
L	Number of possible locations
N	Number of markets
q	Number of production cost rates
$D(j, k)$	External demand for SKU j at market k
$F(i)$	Fixed cost for opening a facility at location i
$C(s, i, k)$	Production cost rate s for SKU k at location i
$B(s, i, k)$	Cost function boundary for rate s and SKU k at location i
$B(0, i, k)$	$= 0$ (dummy cost function boundary)
$\lambda(i, k)$	Lower bound for useful production of SKU k at location i
$T(i, j, k)$	Transportation costs for SKU k between location i and market j

Table 10.1. Data for the Location-Allocation Model

y_i	Location variable
$z_{i,k}^s$	Production indicator variable for rate s and SKU k at location i
$x_{i,k}^s$	Production quantity at cost rate s for SKU k at location i
$t_{i,j,k}$	Transportation variable for SKU k between location i and market j

Table 10.2. Variables for the Location-Allocation Model

In the objective function

$$\text{minimize: } \sum_{i=1}^L F(i)y_i + \sum_{i=1}^L \sum_{j=1}^N \sum_{k=1}^P T(i, j, k)t_{i,j,k} + \sum_{s=1}^q \sum_{i=1}^L \sum_{k=1}^P C(s, i, k)x_{i,k}^s$$

we consider the fixed costs for opening facilities, $\sum_{i=1}^L F(i)y_i$, the sum of all transportation costs, $\sum_{i=1}^L \sum_{j=1}^N \sum_{k=1}^P T(i, j, k)t_{i,j,k}$, as well as the production costs, $\sum_{s=1}^q \sum_{i=1}^L \sum_{k=1}^P C(s, i, k)x_{i,k}^s$. Using data and variables as shown in Tables 10.1 and 10.2 we can now summarize our model in Figure 10.12.

10.5 Optimization

A wealth of mathematics literature is devoted to one or the other aspect of optimization. The research literature typically divides optimization problems along a number of lines.

- We may distinguish between deterministic and stochastic models based on the characteristics of the data that we are provided with.
- Up to this point we have considered only models with one objective function but frequently *multi-criteria* models are considered. These are models with more than one objective function. If we want to find an optimal solution to a model with more than one objective, then we might have to provide data concerning the relative importance of the objectives.
- We may distinguish between linear and non-linear models based on the characteristics of the constraints and the objective function.

Minimize:

$$\sum_{i=1}^L F(i)y_i + \sum_{i=1}^L \sum_{j=1}^N \sum_{k=1}^P T(i, j, k)t_{i,j,k} + \sum_{s=1}^q \sum_{i=1}^L \sum_{k=1}^P C(s, i, k)x_{i,k}^s$$

subject to:

$$\begin{aligned} \sum_{i=1}^n t_{i,j,k} &= D(j, k) && \forall j, \forall k \\ \sum_{j=1}^m t_{i,j,k} &= \sum_{s=1}^q x_{i,k}^s && \forall i, \forall k \\ z_{i,k}^1 &\leq y_i && \forall i, \forall k \\ \lambda(i, k)z_{i,k}^1 &\leq x_{i,k}^1 && \forall i, \forall k \\ x_{i,k}^{s-1} &\geq (B(s-1, i, k) - B(s-2, i, k))z_{i,k}^s && \forall i, \forall k, s = 2, \dots, q \\ x_{i,k}^s &\leq (B(s, i, k) - B(s-1, i, k))z_{i,k}^s && \forall i, \forall k, s = 1, \dots, q \\ t_{i,j,k} &\geq 0 && \forall i, \forall j, \forall k \\ x_{i,k}^s &\geq 0 && \forall i, \forall k, s = 1, \dots, q \\ z_{i,k}^s &\in \{0, 1\} && \forall i, \forall k, s = 1, \dots, q \\ y_i &\in \{0, 1\} && \forall i \end{aligned}$$

Fig. 10.12. Location-Allocation Model

- Further, we may make distinctions between models with only real variables, only binary variables, only integer variables or models that have mixtures of two or more types of variables.
- Regarding input data, offline models assume all input data of a problem instance as known in advance. On the other hand, there are many real-world (decision) problems where one can not assume that all input data is known beforehand. Online models cope with new data that become available dynamically, e.g., when the problem instance or the given constraints change.
- We may also make distinctions concerning the effort required theoretically in the worst case to find an optimal solution for a model. This, however, is beyond the scope of our book; see Garey and Johnson (1979) for a comprehensive treatment.

10.5.1 Exact Methods

The literature on exact methods is far too large for us to consider. Consequently, we restrict our attention only to those areas that were explicitly discussed in Chapter 8.

For our purposes the term solver describes readily available software for the solution of problems or models with certain properties. That is, there are solvers for linear programming problems (LP solver), those for mixed integer programs (MIP solver, which make special use of LP solver combined with branch and bound), and constraint programming solver.

Branch and bound is a very old idea and can also be applied to problems other than MIPs. A good discussion of early applications of branch and bound for MIPs is Geoffrion and Marsten (1972). Beale and Tomlin's proposal for SOS facilities reportedly appeared first in Beale and Tomlin (1970). And as in many other fields, branch and bound research is on-going (see, e.g., Liao (1994), Belvaux and Wolsey (2000)).

An important issue once a MIP has been solved is referred to as *sensitivity analysis*. A significant line of research has been devoted to providing methods that determine the effect on the optimal solution of changes to the input data. A related area of research concerns determining what changes to the data would change a problem from being infeasible to feasible. An extensive survey of this literature is provided by Greenberg (1998).

Constraint (logic) programming has origins in artificial intelligence; see Robinson (1965), Laurière (1978). Early applications were to problems in scheduling, e.g., by Fox and Smith (1984). Cooperation between CP and methods developed for MIPs is a relatively new and promising research area; see Hooker (1998), McAloon et al. (1998), Milano (2004).

10.5.2 Heuristic Search Methods

Heuristics for many optimization problems in production planning and supply chain management are based on the notion of greediness as introduced in §8.4.1. Especially in production planning many of these heuristics are called scheduling rules. More specifically, we may speak of a priority rule to represent the technique by which a number (a priority) is assigned to each job that has to be processed. Then jobs are sequenced according to these numbers, e.g., in increasing order. A simple example is the earliest due date rule where priority is given to jobs with an earlier due date over those with a later due date. Priority rules may be characterized as being static or dynamic. They are static, if they do not change the given priority once assigned. If some information is included into the rule that might change the priority in due course then it is called dynamic. An example is the nearest neighbor routine for solving the TSP. Starting with a single city, any as yet unvisited city can get a priority based on the distance to reach it. Then in every iteration the city with the best priority (in this case the smallest value, the nearest

neighbor) among all cities not yet visited is chosen. From that city priorities are again given to all unvisited cities until a route through all cities has been found.

Priority rules for machine scheduling can be found, e.g., in Haupt (1989). An overall good starting point into the area of heuristic search is the book of Pearl (1984).

Much of the research on heuristic search literature focuses on meta-heuristics, which have been defined as follows: “A meta-heuristic is an iterative master process that guides and modifies the operations of subordinate heuristics to efficiently produce high-quality solutions. It may manipulate a complete (or incomplete) single solution or a collection of solutions at each iteration.” (Voß et al. (1999), p. ix) These methods include simulated annealing, tabu search, genetic algorithms and many others. Recent surveys and collections can be found in Blum and Roli (2003), Glover and Kochenberger (2003), Ibaraki et al. (2005), Rego and Alidaee (2005), Ribeiro and Hansen (2002), Voß (2001).

One of the key aspects regarding metaheuristics in general is the interplay between intensification (concentrating the search into a specific subset of all possible solutions; one can think in terms of a region of the search space) and diversification (elaborating various diverse regions within the search space). That is, it has very often been appropriate, on one hand, to explore promising regions of the search space in a detailed manner (intensification) and, on the other hand, to lead the search into new and yet unexplored regions of the search space (diversification). Within intelligent search including the relationship between these two significant mechanisms the exploration of memory plays a most important role in ongoing research; see, e.g., Greistorfer and Voß (2005).

Simulated annealing traces its origins to computational simulation of the cooling of metals; see Kirkpatrick et al. (1983), Metropolis et al. (1953). Using certain cooling schedules, simulated annealing algorithms can be shown to converge to optimal solutions; see, e.g., Lundy and Mees (1986) or Hajek (1988). Cooling schedules used in practice usually differ significantly from those that are theoretically best, since the latter would result in impractically large computing times. For a more detailed discussion of simulated annealing and the effects of the cooling schedule see, e.g., Johnson et al. (1989). Based on experience using the algorithm as described in §8.3, we recommend the parameter settings `INITPROB = 0.4`, `TEMPFACTOR = 0.8145`, `SIZEFACTOR = 4`, and `MINPERCENT = 2`, which are slightly different from those recommended by Johnson et al. (1989). Cooling schedules that may behave superior to those offered by Johnson et al. are used in the Adaptive Simulated Annealing developed by Ingber (1993).

Genetic algorithms originated in work by Holland and others; see, e.g., Holland (1975), Fogel (1998). Readers interested in mathematical characterizations of early GA's should refer to Liepins and Vose (1992). Attempts to

characterize the theoretical behavior continue. See, for example, the work of Salomon (1996) or Aytug and Koehler (1996).

The simple GA from Vose (1999) that we presented to ease exposition, as well as the stylized versions used for theoretical analysis, can be improved substantially for use in practice. For example, we recommend steady-state replacement without duplicates (see Syswerda (1989) or Davis (1991)) rather than generational replacement. The simple selection technique we gave is dominated by others such as a linear normal ranking scheme for all parents; see Whitley (1989). Many modern GAs perform a descent from each new population member (i.e., they combine the ideas of GA and local search). Genetic and evolutionary algorithms are large areas of ongoing research with many new, partially tested ideas; see, e.g., Smith et al. (1998), Bäck (1997), Reeves and Rowe (2003).

There exist several libraries for genetic algorithms. In principle, an advantage of using classic genetic algorithm libraries such as Genitor (2005) or GALib (2005) is that no neighborhood must be specified. If the built-in genomes of a genetic algorithm library adequately represent one's problem, a user-specified objective function may be the only problem-specific code that must be written. Unfortunately, genetic algorithms without a local search component have not generally proven to be very effective. For a comprehensive overview of genetic algorithm libraries the reader is referred to Pain and Reeves (2002).

GAs are closely related to *evolutionary strategies*. Whereas the mutation operator in a GA serves to protect the search from premature loss of information, evolutionary strategies may incorporate some sort of local search procedure with self adapting parameters involved in the procedure. For some interesting insights on evolutionary algorithms the reader is referred to Hertz and Kobler (2000).

Tabu search was originally developed by Glover (1986) and has been extended in many directions as described in Glover and Laguna (1997). The flexibility and general applicability of TS has caused it to be used in conjunction with other heuristic search methods and much of the development work in TS is done as part of more general heuristic search efforts; see, e.g., Voß et al. (1999).

Recently, *scatter search* ideas established a link between early ideas from various sides – evolutionary strategies, TS and GAs. As an evolutionary approach, scatter search originated from strategies for creating composite decision rules and surrogate constraints. Scatter search is designed to operate on a set of points, called reference points, that constitute good solutions obtained from previous solution efforts. The approach systematically generates linear combinations of the reference points to create new points, each of which is mapped into an associated point that yields integer values for discrete variables. For a very comprehensive treatment of scatter search see Laguna and Martí (2003).

GRASP is usually composed of the following components: A greedy construction phase combined with a probabilistic component and a local search procedure. An adaptive mechanism is used to modify the greedy construction after each iteration. The basic concept goes back to ideas from Hart and Shogan (1987). Resende and Festa (2005) present a general bibliography of *GRASP*.

The research literature is full of comparisons of different heuristic search methods for various problems and it is difficult to declare one or the other method as clear winner. Nevertheless, based on our own research we believe that more intelligent approaches have advantages (e.g., advanced TS implementations over SA; see, among others, Voß (1996), Fink and Voß (1999a), Woodruff and Spearman (1992)).

One of the important research topics over the last couple of years is the development of *class libraries* and *frameworks*; see Voß and Woodruff (2002), Fink et al. (2003). The crucial problem of local search based meta-heuristics libraries is a generic implementation of heuristic approaches as reusable software components, which must operate on arbitrary solution spaces and neighborhood structures. The drawback is that the user must, in general, provide some kind of a problem/solution definition and a neighborhood structure, which is usually done using sophisticated computer languages such as C++.

An early C++ class library for heuristic optimization by Woodruff (1997) included both local search based methods and genetic algorithms. This library raised issues that illustrate both the promise and the drawbacks to the adaptable component approach. From a research perspective such libraries can be thought of as providing a concrete taxonomy for heuristic search. So concrete, in fact, that they can be compiled into machine code. This taxonomy sheds some light on the relationships between heuristic search methods for optimization and on ways in which they can be combined. Furthermore, the library facilitates such combinations as the classes in the library can be extended and/or combined to produce new search strategies.

From a practical and empirical perspective, these types of libraries provide a vehicle for using and testing heuristic search optimization. A user of the library must provide the definition of the problem specific abstractions and may systematically vary and exchange heuristic strategies and corresponding components.

We briefly mention one example from several heuristic optimization libraries from the research field, which differ, e.g., in the design concept, the chosen balance between “ease-of-use” and flexibility and efficiency, and the overall scope. All of these approaches are based on the concepts of object-oriented programming.

HOTFRAME, a Heuristic OpTimization FRAMEwork implemented in C++, provides both adaptable components that incorporate different meta-heuristics and an architectural description of the collaboration among these components and problem-specific complements. All typical application-speci-

fic concepts are treated as objects or classes: problems, solutions, neighbors, solution and move attributes. On the other side, meta-heuristics concepts such as different methods and their building-blocks such as tabu criteria and diversification strategies are also treated as objects. HOTFRAME uses genericity as the primary mechanism to make these objects adaptable. That is, common behavior of meta-heuristics is factored out and grouped in generic classes, applying static type variation. Meta-heuristics template classes are parameterized by corresponding aspects such as solution spaces and neighborhood structures.

All heuristics such as TS, SA and GA are implemented in a consistent way, which facilitates an easy embedding of arbitrary methods into application systems or as parts of more advanced/hybrid methods. Both new meta-heuristics and new applications can be added to the framework. For example, the *pilot method* of Duin and Voß (1999) is a technique based on lookahead that was readily implemented and added to HOTFRAME. Starting with a simple greedy algorithm such as a construction heuristic the pilot method builds primarily on the idea to look ahead for each possible local choice (by computing a so-called “pilot” solution), memorizing the best result, and performing the according move. The look ahead mechanism of the pilot method is related to increased neighborhood depths as it exploits the evaluation of neighbors at larger depths to guide the neighbor selection at depth one; see also Voß et al. (2005).

HOTFRAME includes built-in support for solution spaces representable by binary vectors or permutations, in connection with corresponding standard neighborhood structures, solution and move attributes, and recombination operators. Otherwise, the user may derive specialized classes from suitable built-in classes or implement corresponding classes from scratch according to a defined interface. For further information about HOTFRAME see Fink and Voß (1999b, 2002).

10.5.3 Progressive Hedging

We limit our discussion of the stochastic programming literature to those articles related to progressive hedging or multi-stage mixed integer problems. Readers interested in more general treatment should see Kall and Wallace (1994) or Birge and Louveaux (1997).

Progressive hedging is not the only method that has been proposed for multi-stage stochastic MIPs. Klein Haneveld and van der Vlerk (1999) provide descriptions of general formulations and solution methods for integer stochastic programs. Carøe and Tind (1997, 1998) describe two different methods for stochastic MIPs. Schultz et al. (1998) have developed a mathematically sophisticated method of finding provably optimal solutions to classes of stochastic MIPs. Jonsbråten et al. (1998) describe a class of stochastic MIPs where decisions affect the timing of information discovery along with a solution method.

Progressive hedging has been used in a number of applications reported in the literature. Mulvey and Vladimirou (1991, 1992) have reported success solving network problems. Helgason and Wallace (1991), Wallace and Helgason (1991) have reported success solving fishery problems and have suggested the use of tree based data structures for managing data of the PH progressive hedging algorithm.

Birge et al. (1995) report on the use of progressive hedging for power system optimization (although they use a linear, rather than a quadratic, penalty term). Carøe and Schultz (1999) propose the use of a relaxation that is similar to progressive hedging, but also uses a linear penalty. Both papers report good computational results.

The progressive hedging algorithm as described in §9.2.3 developed by Løkketangen and Woodruff (1996) is based on a more general algorithm proposed by Rockafellar and Wets (1991). For some basics regarding an interpretation as dual prices for the implementability constraints see, e.g., Wets (1989). For move evaluation functions and respective tabu search mechanisms associated with solving general stochastic MIPs see also the work described in detail by Løkketangen and Glover (1996). For an application of this algorithm to a classic single machine lot sizing problem see Haugen et al. (2001). The notion of integer convergence for progressive hedging is introduced by Løkketangen and Woodruff (1996). Related to the topics raised in this book we investigate the progressive hedging algorithm in Woodruff and Voß (2006). Based on the **SCPC** model we consider the case when an actor in the supply chain is faced with the potential for a major disruption. The progressive hedging algorithm is combined with a GRASP aiming at a realistic chance to solve models that explicitly consider the possibility of a “big bang” in the supply chain.

10.5.4 Simulation

Owing to its inherent modeling flexibility, simulation is often regarded as a proper means for supporting decision making, e.g., on supply chain design. Especially, discrete event simulation has been used to analyze and improve operations in logistic systems for more than two decades by now. Typical simulation tasks are to verify whether a system is able to produce the demanded output per time unit, to determine buffer-sizes, to identify bottlenecks, or to optimize control policies especially in cases when analytical tools are not at hand or somewhat not applicable (e.g, due to computation times). In discrete event simulation the state of a model changes at only a discrete, but possibly random, set of simulated points in time.

For a good textbook on simulation we refer to Law and Kelton (2000). Moreover, optimization in stochastic systems incorporating parametric (static) as well as control (dynamic) optimization asks for simulation and has been investigated to some extent; see, e.g., Gosavi (2003). A survey on available simulation software is conducted by Swain (2003).

As one example for supply chain simulation we mention van der Zee and van der Vorst (2005), who provide a brief literature survey with the aim of listing simulation model qualities essential for supporting successful decision making on supply chain design. Based on this the authors propose an object-oriented modeling framework that facilitates supply chain simulation.

10.6 Modeling

Modeling is a very broad and important topic. We have focused on the creation of mathematical models for optimization, but there are numerous alternative model forms, some of which we have briefly mentioned. The work of Pidd (2003) provides consideration of a wider view of modeling.

For a discussion of the art and science of MIP and LP modeling, the work of Williams (2000) is arguably the best. This book covers a large number of modeling concepts and considers the implications for solvability. The book is very comprehensive. For a more gentle introduction, operations research and management science textbooks such as the work of Hillier and Lieberman (2004) or Moore and Weatherford (2001) are useful and these books contain information about other facets of operations research modeling as well.

Muhanna (1993), Muhanna and Pick (1994) advocate object based approaches to the creation and management of mathematical programming models in a fashion similar in spirit to the structured methods proposed by Geoffrion (1992). Although not commercially available, the idea is compelling. By creating object classes to correspond to model components, models can be constructed more quickly and maintained more efficiently. Their work draws on concepts developed in the Object Oriented Modeling (OOM) literature (see, e.g., Booch et al. (1998)). A related idea is the merging of OOM techniques and modeling languages, particularly in the area of constraint logic programming and combinations with local search. For example, Michel and van Hentenryck (2001) and Laburthe and Caseau (1998) explore these notions.

In order to apply optimization methods to a new type of problem, corresponding models and algorithms have to be “coded” so that they are accessible to a computer. One way to achieve this is the use of a *modeling language*. Over the years substantial progress has been made in developing tools to simplify the design and implementation of models and algorithms. One of the research achievements is a considerable reduction in development time while preserving most of the efficiency of specialized software.

Modeling languages are being extended into new domains such as complementarity problems (see Ferris et al. (1999)) and stochastic linear programming (see Buchanan et al. (2002)). Extension of the modeling language domain to include combinatorial optimization problems is also the topic of on-going research. Such problems can often be specified more naturally as constraint programs than as integer programs, which can be exploited by

modeling languages that have constraint programming capabilities (Fourer (1998), van Hentenryck (1999), van Hentenryck and Michel (2002)). These approaches have been quite successfully applied to problems with a significant number of logical constraints (for example, special scheduling and assignment problems). Sometimes modeling language support for a mixture of CP and MIP capabilities is the most effective means of addressing a particular problem (see, e.g., Jain and Grossmann (2001)).

Modeling languages provide very high-level algebraic and set notations to concisely express mathematical problems that can then be solved using state-of-the-art solvers. Because these modeling languages do not require specific programming skills they are readily used by a wide audience. In Chapter 7 we provided implementations of **mrp**, **MRPII**, and **SCPc** in some popular modeling languages, namely AMPL (*Algebraic Modeling Language for Mathematical Programming*; see, e.g., Fourer et al. (2002), Fourer (1998)), GAMS (*General Algebraic Modeling System*; see Bisschop and Meerhaus (1982), Brooke et al. (1992)), MPL (*Mathematical Programming Language*), OPL (*Optimization Programming Language*), and Mosel (see, e.g., Colombani and Heipcke (2002), Guéret et al. (2002), Begain et al. (2001)). As we have shown, the modeling languages provide the power and flexibility to express well-known production planning models as optimization opportunities and support their extension to enterprise planning models.