Chapter 1

Introduction

1.1 The Role of Scheduling

Scheduling is a decision-making process that is used on a regular basis in many manufacturing and services industries. It deals with the allocation of resources to tasks over given time periods and its goal is to optimize one or more objectives.

The resources and tasks in an organization can take many different forms. The resources may be machines in a workshop, runways at an airport, crews at a construction site, processing units in a computing environment, and so on. The tasks may be operations in a production process, take-offs and landings at an airport, stages in a construction project, executions of computer programs, and so on. Each task may have a certain priority level, an earliest possible starting time and a due date. The objectives can also take many different forms. One objective may be the minimization of the completion time of the last task and another may be the minimization of the number of tasks completed after their respective due dates.

Scheduling, as a decision-making process, plays an important role in most manufacturing and production systems as well as in most information processing environments. It is also important in transportation and distribution settings and in other types of service industries. The following examples illustrate the role of scheduling in a number of real world environments.

Example 1.1.1 (A Paper Bag Factory)

Consider a factory that produces paper bags for cement, charcoal, dog food, and so on. The basic raw material for such an operation are rolls of paper. The production process consists of three stages: the printing of the logo, the gluing of the side of the bag, and the sewing of one end or both ends of the

M. Pinedo, *Scheduling*, DOI: 10.1007/978-0-387-78935-4_1,

bag. Each stage consists of a number of machines which are not necessarily identical. The machines at a stage may differ slightly in the speed at which they operate, the number of colors they can print or the size of bag they can produce. Each production order indicates a given quantity of a specific bag that has to be produced and shipped by a committed shipping date or due date. The processing times for the different operations are proportional to the size of the order, i.e., the number of bags ordered.

A late delivery implies a penalty in the form of loss of goodwill and the magnitude of the penalty depends on the importance of the order or the client and the tardiness of the delivery. One of the objectives of the scheduling system is to minimize the sum of these penalties.

When a machine is switched over from one type of bag to another a setup is required. The length of the setup time on the machine depends on the similarities between the two consecutive orders (the number of colors in common, the differences in bag size and so on). An important objective of the scheduling system is the minimization of the total time spent on setups. ||

Example 1.1.2 (A Semiconductor Manufacturing Facility)

Semiconductors are manufactured in highly specialized facilities. This is the case with memory chips as well as with microprocessors. The production process in these facilities usually consists of four phases: wafer fabrication, wafer probe, assembly or packaging, and final testing.

Wafer fabrication is technologically the most complex phase. Layers of metal and wafer material are built up in patterns on wafers of silicon or gallium arsenide to produce the circuitry. Each layer requires a number of operations, which typically include: (i) cleaning, (ii) oxidation, deposition and metallization, (iii) lithography, (iv) etching, (v) ion implantation, (vi) photoresist stripping, and (vii) inspection and measurement. Because it consists of various layers, each wafer has to undergo these operations several times. Thus, there is a significant amount of recirculation in the process. Wafers move through the facility in lots of 24. Some machines may require setups to prepare them for incoming jobs; the setup time often depends on the configurations of the lot just completed and the lot about to start.

The number of orders in the production process is often in the hundreds and each has its own release date and a committed shipping or due date. The scheduler's objective is to meet as many of the committed shipping dates as possible, while maximizing throughput. The latter goal is achieved by maximizing equipment utilization, especially of the bottleneck machines, requiring thus a minimization of idle times and setup times. ||

Example 1.1.3 (Gate Assignments at an Airport)

Consider an airline terminal at a major airport. There are dozens of gates and hundreds of planes arriving and departing each day. The gates are not all identical and neither are the planes. Some of the gates are in locations with a lot of space where large planes (widebodies) can be accommodated

easily. Other gates are in locations where it is difficult to bring in the planes; certain planes may actually have to be towed to their gates.

Planes arrive and depart according to a certain schedule. However, the schedule is subject to a certain amount of randomness, which may be weather related or caused by unforeseen events at other airports. During the time that a plane occupies a gate the arriving passengers have to be deplaned, the plane has to be serviced and the departing passengers have to be boarded. The scheduled departure time can be viewed as a due date and the airline's performance is measured accordingly. However, if it is known in advance that the plane cannot land at the next airport because of anticipated congestion at its scheduled arrival time, then the plane does not take off (such a policy is followed to conserve fuel). If a plane is not allowed to take off, operating policies usually prescribe that passengers remain in the terminal rather than on the plane. If boarding is postponed, a plane may remain at a gate for an extended period of time, thus preventing other planes from using that gate.

The scheduler has to assign planes to gates in such a way that the assignment is physically feasible while optimizing a number of objectives. This implies that the scheduler has to assign planes to suitable gates that are available at the respective arrival times. The objectives include minimization of work for airline personnel and minimization of airplane delays.

In this scenario the gates are the resources and the handling and servicing of the planes are the tasks. The arrival of a plane at a gate represents the starting time of a task and the departure represents its completion time. ||

Example 1.1.4 (Scheduling Tasks in a Central Processing Unit (CPU))

One of the functions of a multi-tasking computer operating system is to schedule the time that the CPU devotes to the different programs that have to be executed. The exact processing times are usually not known in advance. However, the distribution of these random processing times may be known in advance, including their means and their variances. In addition, each task usually has a certain priority level (the operating system typically allows operators and users to specify the priority level or weight of each task). In such case, the objective is to minimize the expected sum of the weighted completion times of all tasks.

To avoid the situation where relatively short tasks remain in the system for a long time waiting for much longer tasks that have a higher priority, the operating system "slices" each task into little pieces. The operating system then rotates these slices on the CPU so that in any given time interval, the CPU spends some amount of time on each task. This way, if by chance the processing time of one of the tasks is very short, the task will be able to leave the system relatively quickly.

An interruption of the processing of a task is often referred to as a *preemption*. It is clear that the optimal policy in such an environment makes heavy use of preemptions. ||

It may not be immediately clear what impact schedules may have on objectives of interest. Does it make sense to invest time and effort searching for a good schedule rather than just choosing a schedule at random? In practice, it often turns out that the choice of schedule *does* have a significant impact on the system's performance and that it *does* make sense to spend some time and effort searching for a suitable schedule.

Scheduling can be difficult from a technical as well as from an implementation point of view. The type of difficulties encountered on the technical side are similar to the difficulties encountered in other forms of combinatorial optimization and stochastic modeling. The difficulties on the implementation side are of a completely different kind. They may depend on the accuracy of the model used for the analysis of the actual scheduling problem and on the reliability of the input data that are needed.

1.2 The Scheduling Function in an Enterprise

The scheduling function in a production system or service organization must interact with many other functions. These interactions are system-dependent and may differ substantially from one situation to another. They often take place within an enterprise-wide information system.

A modern factory or service organization often has an elaborate information system in place that includes a central computer and database. Local area networks of personal computers, workstations and data entry terminals, which are connected to this central computer, may be used either to retrieve data from the database or to enter new data. The software controlling such an elaborate information system is typically referred to as an Enterprise Resource Planning (ERP) system. A number of software companies specialize in the development of such systems, including SAP, J.D. Edwards, and PeopleSoft. Such an ERP system plays the role of an information highway that traverses the enterprise with, at all organizational levels, links to decision support systems.

Scheduling is often done interactively via a decision support system that is installed on a personal computer or workstation linked to the ERP system. Terminals at key locations connected to the ERP system can give departments throughout the enterprise access to all current scheduling information. These departments, in turn, can provide the scheduling system with up-to-date information concerning the statuses of jobs and machines.

There are, of course, still environments where the communication between the scheduling function and other decision making entities occurs in meetings or through memos.

Scheduling in Manufacturing Consider the following generic manufacturing environment and the role of its scheduling. Orders that are released in a manufacturing setting have to be translated into jobs with associated due

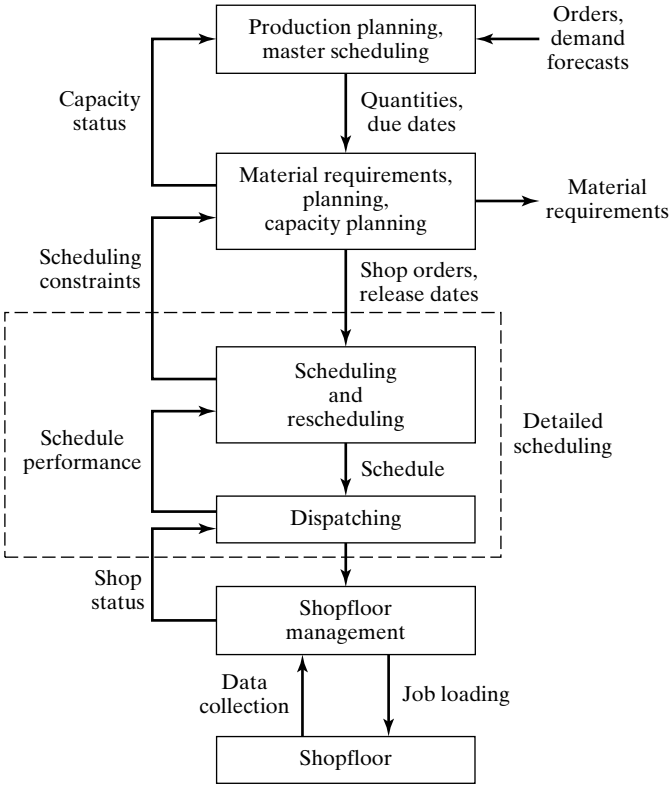

Fig. 1.1 Information flow diagram in a manufacturing system

dates. These jobs often have to be processed on the machines in a workcenter in a given order or sequence. The processing of jobs may sometimes be delayed if certain machines are busy and preemptions may occur when high priority jobs arrive at machines that are busy. Unforeseen events on the shop floor, such as machine breakdowns or longer-than-expected processing times, also have to be taken into account, since they may have a major impact on the schedules. In such an environment, the development of a detailed task schedule helps maintain efficiency and control of operations.

The shop floor is not the only part of the organization that impacts the scheduling process. It is also affected by the production planning process that handles medium- to long-term planning for the entire organization. This process attempts to optimize the firm's overall product mix and long-term resource allocation based on its inventory levels, demand forecasts and resource requirements. Decisions made at this higher planning level may impact the scheduling process directly. Figure 1.1 depicts a diagram of the information flow in a manufacturing system.

In a manufacturing environment, the scheduling function has to interact with other decision making functions. One popular system that is widely used is the Material Requirements Planning (MRP) system. After a schedule has been generated it is necessary that all raw materials and resources are available at the specified times. The ready dates of all jobs have to be determined jointly by the production planning/scheduling system and the MRP system.

MRP systems are normally fairly elaborate. Each job has a Bill Of Materials (BOM) itemizing the parts required for production. The MRP system keeps track of the inventory of each part. Furthermore, it determines the timing of the purchases of each one of the materials. In doing so, it uses techniques such as lot sizing and lot scheduling that are similar to those used in scheduling systems. There are many commercial MRP software packages available and, as a result, there are many manufacturing facilities with MRP systems. In the cases where the facility does not have a scheduling system, the MRP system may be used for production planning purposes. However, in complex settings it is not easy for an MRP system to do the detailed scheduling satisfactorily.

Scheduling in Services Describing a generic service organization and a typical scheduling system is not as easy as describing a generic manufacturing organization. The scheduling function in a service organization may face a variety of problems. It may have to deal with the reservation of resources, e.g., the assignment of planes to gates (see Example 1.1.3), or the reservation of meeting rooms or other facilities. The models used are at times somewhat different from those used in manufacturing settings. Scheduling in a service environment must be coordinated with other decision making functions, usually within elaborate information systems, much in the same way as the scheduling function in a manufacturing setting. These information systems usually rely on extensive databases that contain all the relevant information with regard to availability of resources and (potential) customers. The scheduling system interacts often with forecasting and yield management modules. Figure 1.2 depicts the information flow in a service organization such as a car rental agency. In contrast to manufacturing settings, there is usually no MRP system in a service environment.

1.3 Outline of the Book

This book focuses on both the theory and the applications of scheduling. The theoretical side deals with the detailed sequencing and scheduling of jobs. Given a collection of jobs requiring processing in a certain machine environment, the problem is to sequence these jobs, subject to given constraints, in such a way that one or more performance criteria are optimized. The scheduler may have to deal with various forms of uncertainties, such as random job processing times, machines subject to breakdowns, rush orders, and so on.

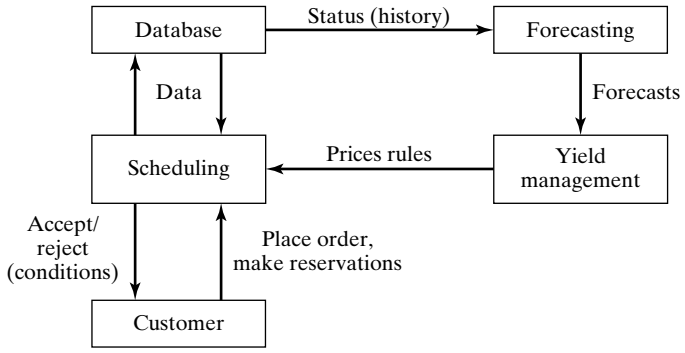

Fig. 1.2 Information flow diagram in a service system

Thousands of scheduling problems and models have been studied and analyzed in the past. Obviously, only a limited number are considered in this book; the selection is based on the insight they provide, the methodology needed for their analysis and their importance in applications.

Although the applications driving the models in this book come mainly from manufacturing and production environments, it is clear from the examples in Section 1.1 that scheduling plays a role in a wide variety of situations. The models and concepts considered in this book are applicable in other settings as well.

This book is divided into three parts. Part I (Chapters 2 to 8) deals with deterministic scheduling models. In these chapters it is assumed that there are a finite number of jobs that have to be scheduled with one or more objectives to be minimized. Emphasis is placed on the analysis of relatively simple priority or dispatching rules. Chapter 2 discusses the notation and gives an overview of the models that are considered in the subsequent chapters. Chapters 3 to 8 consider the various machine environments. Chapters 3 and 4 deal with the single machine, Chapter 5 with machines in parallel, Chapter 6 with machines in series and Chapter 7 with the more complicated job shop models. Chapter 8 focuses on open shops in which there are no restrictions on the routings of the jobs in the shop.

Part II (Chapters 9 to 13) deals with stochastic scheduling models. These chapters, in most cases, also assume that a given (finite) number of jobs have to be scheduled. The job data, such as processing times, release dates and due dates may not be exactly known in advance; only their distributions are known in advance. The actual processing times, release dates and due dates become known only at the *completion* of the processing or at the actual occurrence of the release or due date. In these models a single objective has to be minimized, usually in expectation. Again, an emphasis is placed on the analysis of relatively simple priority or dispatching rules. Chapter 9 contains preliminary material. Chapter 10 covers the single machine environment. Chapter 11 also covers the

single machine, but in this chapter it is assumed that the jobs are released at different points in time. This chapter establishes the relationship between stochastic scheduling and the theory of priority queues. Chapter 12 focuses on machines in parallel and Chapter 13 describes the more complicated flow shop, job shop, and open shop models.

Part III (Chapters 14 to 20) deals with applications and implementation issues. Algorithms are described for a number of real world scheduling problems. Design issues for scheduling systems are discussed and some examples of scheduling systems are given. Chapters 14 and 15 describe various general purpose procedures that have proven to be useful in industrial scheduling systems. Chapter 16 describes a number of real world scheduling problems and how they have been dealt with in practice. Chapter 17 focuses on the basic issues concerning the design, the development and the implementation of scheduling systems, and Chapter 18 discusses the more advanced concepts in the design and implementation of scheduling systems. Chapter 19 gives some examples of actual implementations. Chapter 20 ponders on what lies ahead in scheduling.

Appendices A, B, C, and D present short overviews of some of the basic methodologies, namely mathematical programming, dynamic programming, constraint programming, and complexity theory. Appendix E contains a complexity classification of the deterministic scheduling problems, while Appendix F presents an overview of the stochastic scheduling problems. Appendix G lists a number of scheduling systems that have been developed in industry and academia. Appendix H provides some guidelines for using the LEKIN scheduling system. The LEKIN system is included on the CD-ROM that comes with the book.

This book is designed for either a masters level course or a beginning PhD level course in Production Scheduling. When used for a senior level course, the topics most likely covered are from Parts I and III. Such a course can be given without getting into complexity theory: one can go through the chapters of Part I skipping all complexity proofs without loss of continuity. A masters level course may cover topics from Part II as well. Even though all three parts are fairly self-contained, it is helpful to go through Chapter 2 before venturing into Part II.

Prerequisite knowledge for this book is an elementary course in Operations Research on the level of Hillier and Lieberman's *Introduction to Operations Research* and an elementary course in stochastic processes on the level of Ross's *Introduction to Probability Models*.

Comments and References

During the last four decades many books have appeared that focus on sequencing and scheduling. These books range from the elementary to the more advanced.

A volume edited by Muth and Thompson (1963) contains a collection of papers focusing primarily on computational aspects of scheduling. One of the better known textbooks is the one by Conway, Maxwell and Miller (1967) (which, even though slightly out of date, is still very interesting); this book also deals with some of the stochastic aspects and with priority queues. A more recent text by Baker (1974) gives an excellent overview of the many aspects of deterministic scheduling. However, this book does not deal with computational complexity issues since it appeared just before research in computational complexity started to become popular. The book by Coffman (1976) is a compendium of papers on deterministic scheduling; it does cover computational complexity. An introductory textbook by French (1982) covers most of the techniques that are used in deterministic scheduling. The proceedings of a NATO workshop, edited by Dempster, Lenstra and Rinnooy Kan (1982), contains a number of advanced papers on deterministic as well as on stochastic scheduling. The relatively advanced book by Blazewicz, Cellary, Slowinski and Weglarz (1986) focuses mainly on resource constraints and multi-objective deterministic scheduling. The book by Blazewicz, Ecker, Schmidt and Weglarz (1993) is somewhat advanced and deals primarily with the computational aspects of deterministic scheduling models and their applications to manufacturing. The more applied text by Morton and Pentico (1993) presents a detailed analysis of a large number of scheduling heuristics that are useful for practitioners. The monograph by Dauzère-Pérès and Lasserre (1994) focuses primarily on job shop scheduling. A collection of papers, edited by Zweben and Fox (1994), describes a number of scheduling systems and their actual implementations. Another collection of papers, edited by Brown and Scherer (1995) also describe various scheduling systems and their implementation. The proceedings of a workshop edited by Chrétienne, Coffman, Lenstra and Liu (1995) contain a set of interesting papers concerning primarily deterministic scheduling. The textbook by Baker (1995) is very useful for an introductory course in sequencing and scheduling. Brucker (1995) presents, in the first edition of his book, a very detailed algorithmic analysis of the many deterministic scheduling models. Parker (1995) gives a similar overview and tends to focus on problems with precedence constraints or other graph-theoretic issues. Sule (1996) is a more applied text with a discussion of some interesting real world problems. Blazewicz, Ecker, Pesch, Schmidt and Weglarz (1996) is an extended edition of the earlier work by Blazewicz, Ecker, Schmidt and Weglarz (1993). The monograph by Ovacik and Uzsoy (1997) is entirely dedicated to decomposition methods for complex job shops. The two volumes edited by Lee and Lei (1997) contain many interesting theoretical as well as applied papers. The book by Pinedo and Chao (1999) is more application oriented and describes a number of different scheduling models for problems arising in manufacturing and in services. The monograph by Baptiste, LePape and Nuijten (2001) covers applications of constraint programming techniques to job shop scheduling. The volume edited by Nareyek (2001) contains papers on local search applied to job shop scheduling. T'kindt and Billaut (2002, 2006) provide an excellent treatise of multicriteria scheduling. Brucker (2004) is an expanded version of the orig-

inal first edition that appeared in 1995. The *Handbook of Scheduling*, edited by Leung (2004), contains numerous papers on all aspects of scheduling. The text by Pinedo (2005) is a modified and extended version of the earlier one by Pinedo and Chao (1999). Dawande, Geismar, Sethi and Sriskandarajah (2007) focus in their more advanced text on the scheduling of robotic cells; these manufacturing settings are, in a sense, extensions of flow shops.

Besides these books a number of survey articles have appeared, each one with a large number of references. The articles by Graves (1981) and Rodammer and White (1988) review production scheduling. Atabakhsh (1991) presents a survey of constraint based scheduling systems that use artificial intelligence techniques and Noronha and Sarma (1991) review knowledge-based approaches for scheduling problems. Smith (1992) focuses in his survey on the development and implementation of scheduling systems. Lawler, Lenstra, Rinnooy Kan and Shmoys (1993) give a detailed overview of deterministic sequencing and scheduling and Righter (1994) does the same for stochastic scheduling. Queyranne and Schulz (1994) provide an in depth analysis of polyhedral approaches to non-preemptive machine scheduling problems. Chen, Potts and Woeginger (1998) review computational complexity, algorithms and approximability in deterministic scheduling. Sgall (1998) and Pruhs, Sgall and Torng (2004) present surveys of an area within deterministic scheduling referred to as online scheduling. Even though online scheduling is often considered a part of deterministic scheduling, the theorems obtained may at times provide interesting new insights into certain stochastic scheduling models.