Chapter 1
Introduction

The purpose of this chapter is to provide an overview of the book. Therefore, we start by discussing motivation for modeling and analysis of semiconductor manufacturing. Semiconductor manufacturing is an extreme environment for production planning and control, scheduling, and simulation models. The enormous size of the facilities and supply chains in the semiconductor industry, the permanent appearance of uncertainty, and rapid technological changes lead to an environment that brings approaches developed for other industries under stress (see Chien et al. [49] for a related discussion). The capital intensive nature of the semiconductor industry requires manufacturing systems to run consistently at high utilization levels, reentrant flows create complex competition for limited resource capacity, and the ever-increasing level of automation reduces the ability to rely solely on people for production planning and control. Models that are successful in the semiconductor industry will likely find reasonable applications in other areas. A second source of academic interest in modeling and analysis of semiconductor manufacturing is the insight that the semiconductor manufacturing environment initiates on the formulation of some problems that had not been widely studied in other industries (cf. Chien et al. [49]).

We obviously cannot give a complete account of modeling and analysis of semiconductor manufacturing in a book of only a few hundred pages. Hence, instead of attempting the impossible, we have chosen production planning and control of wafer fabs from the perspective of our own interests and research programs. In the second section of this chapter, we provide an outline of the content of the book.

1.1 Motivation

In the last decade, the electronics industry has become one of the largest industries in the world. At the heart of this industry is the manufacturing of integrated circuits (ICs or chips) on thin silicon discs (wafers). The fabrication

L. Mönch et al., *Production Planning and Control for Semiconductor Wafer Fabrication Facilities*, Operations Research/Computer Science Interfaces Series 52, DOI 10.1007/978-1-4614-4472-5_1,

of ICs on silicon wafers is arguably the most complex manufacturing process in existence [14, 110, 276, 306, 307]. This complexity is caused by many factors including multiple products, routes with several hundred process steps, and a large number of machines (tools). There are four basic steps in manufacturing ICs (cf. Uzsoy et al. [306]):

- Wafer fabrication
- Sort (or probe)
- Assembly
- Test

In wafer fabrication, the layers of the ICs are fabricated onto raw silicon wafers. Next, the completed wafers are sent to sort where electronic probes perform an electrical test on each IC to determine basic functionality. Then, the probed wafers are sent to assembly, where the wafers are cut into individual ICs and the functioning ones are put into a package that allows connection with higher level devices such as PCs, cell phones, etc. Finally, the packaged ICs are tested and labeled. Wafer fabrication is the most time-consuming and the most costly step and is the primary focus of this monograph. It is characterized by the following process conditions:

- Reentrant flow, i.e., a lot of wafers, called jobs to be consistent with the scheduling literature, may visit the same machine several times
- A mix of different process types, for example, batch processes, i.e., several jobs, can be processed simultaneously on the same machine vs. single wafer processes
- Unrelated parallel machines that are often highly unreliable or require considerable preventive maintenance to keep them reliable
- Sequence-dependent setup times that can in some cases take considerably longer than the time to process a job
- Variety of products with a changing product mix
- Customer due dates that are very aggressive

In addition, the machines used for processing jobs are extremely expensive, some as high as US$40 million, and thus are scarce resources. This is the main reason for reentrant flow of the jobs through the wafer fabrication facility (wafer fab). This type of flow causes problems related to production control of wafer fabs that are different than production control problems in classical job shops, for example, the occurrence of dynamic bottlenecks. The cost of today's fab, up to US$5 billion, leads to competition between production jobs and prototype (or engineering) jobs for processing time on the machines. Many companies do process development for the next generation of products in the same fab that produces the current generation products in high volume.

In the past, sources of reducing costs in semiconductor manufacturing were decreasing the size of the chips, increasing the wafer sizes, and improving the yield, simultaneously with efforts to improve operational processes inside the wafer fabs (cf. Schömig and Fowler [276]). While shrinking the size of the chips will likely continue to significantly reduce costs of semiconductor

manufacturing for the next several generations of products, the productivity gains from increasing wafer sizes and improving yields will not likely continue at historic levels. There will be wafer size increases, but the increased costs of the larger wafers will offset some of the productivity gains. In the case of yield, the gains will not be as large as in the past because yields are already high, which leaves less room for improvement. Currently, it seems that the improvement of operational processes creates the best opportunity to realize the necessary cost reductions. Therefore, the development of efficient planning and control strategies is highly desirable in the semiconductor wafer fabrication domain. In the course of the development of new planning and control algorithms, researchers and developers have to take into account the new opportunities in advanced software and hardware technologies.

1.2 Outline of the Book

In this monograph, we consider these problems. We show that from our point of view productivity improvements in the semiconductor industry will have to come through the implementation of operations research and industrial engineering tools and techniques and through application of state-of-the-art computing technologies.

In Chap. 2, we provide a detailed process description of semiconductor manufacturing. We use the notion of base system, base process, control system, control process, planning system, planning process, and finally of the information system. A manufacturing system consists of a base system that contains all the resources, i.e., tools, secondary resources, and operators. The corresponding base process is given by jobs that consume capacities of the resources during processing. The resource allocation process of the jobs is influenced by the production control process that is performed by using the production control system. The production control system consists of the computers and the software used to produce production control instructions, i.e., software with dispatching and scheduling capabilities. The production control process determines when and under which circumstances a certain control algorithm is used to determine production control instructions. The production planning system is given by a set of computers and software that is used to determine production planning instructions. The production control system and production planning system combined with human decision makers and the operational system form the information system. The production planning system determines when certain planning actions have to be performed. The main results of production planning are quantities and points of time for releasing orders into the base system. In this book, we are mainly interested in the design of the production control and production planning system and also in the production control and planning process. Less attention is paid to the design of the base system and process.

We introduce the notion of complex job shops (cf. [172, 223]). Complex job shops are a specific class of job shops that are characterized by unrelated parallel machines, batch machines, reentrant process flows, and process

variability. We describe the basic process steps of semiconductor manufacturing and introduce a production planning and control hierarchy that includes the enterprise level, the factory level, and finally the work center level from a resource point of view.

Most enterprises consist of several factories that are geographically distributed. The following questions are interesting on the enterprise level:

- How do we maximize enterprise capacity?
- What are the effects of product mix?
- How do we minimize enterprise costs?

However, in this monograph, we mainly concentrate on the factory and the work center level instead of the enterprise level. We are interested in the following types of questions for the factory level:

- What is the best dispatching strategy?
- Is it beneficial to use a scheduling system?
- What is the impact of different lot sizes?
- What is the impact of different order/job release strategies?
- Is it worthwhile to integrate machine-related and automated material handling system-related decisions?

A single factory consists of different work centers. The following issues will be discussed for the work center level:

- Is it necessary to use different dispatching strategies for different types of work centers?
- What is a good batching strategy?
- What is an effective way to manage reticles?
- How should cluster tools be scheduled?

A batch is a collection of jobs that are processed at the same time on the same machine. Reticles are secondary resources that contain the information for specific integrated chips. Finally, a cluster tool may perform several consecutive process steps of a certain job.

From an operations management point of view, we differentiate between planning at the highest level, order release, scheduling, and dispatching at the lowest level. Each of these different functionalities is related to a certain horizon. The decisions on the higher levels are generally made on a periodic, but infrequent basis. Each decision typically has a huge financial impact. The plans from the planning level are used to make order/job release decisions. Finally, priorities are assigned to each job in order to process the jobs on single machines.

In this book, we start by describing dispatching and end up with production planning. The resulting planning and control hierarchy consists of the following layers:

- Planning with a time horizon ranging from months to years
- Order release in a weekly or bi-weekly frequency

- Scheduling each shift/day
- Dispatching in a minute-by-minute manner depending on the speed of the material flow

Production plans are the output of the planning layer. These plans are then used to release jobs. These job starts are an input of the scheduling layer. A detailed description of activities at the machine level is the result of the scheduling layer. A schedule can be used to establish priorities for operations. These priorities are important for dispatching decisions.

In this monograph, of course, we cannot completely answer all questions. However, the aim of this monograph is to provide tools and techniques from operations research, industrial engineering, and computer science to model and control wafer fabs effectively. We hope that readers of this book are willing to accept the abstract modeling and operations research language.

Therefore, after a description of the base process and base system in Chap. 2, we provide the necessary modeling and analysis tools in Chap. 3. Models are used within the production planning and control process for representing the base system and base process and for decision-making. They can be part of the production control and also the production planning system. We differentiate between dynamic and static, deterministic and stochastic, and descriptive and prescriptive models. Dynamic models contain a time dependency, while static models do not. Stochastic models include model attributes that are specified by probability distributions, while model attributes in deterministic models are not random. Descriptive models are used to describe how a system behaves. For the purpose of this book, descriptive models usually are given by queueing models and simulation models of the base system and process. Prescriptive models are generally used for the immediate derivation of planning and control instructions. These models are used to modify the future evolution of the system. Scheduling models are examples for this class of models. Furthermore, we also describe statistical models for the design of experiments (cf. Montgomery [208]). These models are important for the performance assessment of new production planning and control approaches and will be used in the subsequent chapters of the book.

We describe various decision methods including branch-and-bound techniques, linear and mixed integer programming, stochastic programming, dynamic programming, metaheuristics, queueing theory, and discrete-event simulation for the sake of completeness. These descriptions are simply meant to equip the readers with the necessary tools to understand the models and methods to solve specific problems in the remaining chapters of the monograph.

We also deal with basic questions of performance assessment and therefore introduce important performance measures used in this monograph. We describe simulation-based methods to assess the performance of the production planning and the production control system within a dynamic and stochastic environment due to Mönch [192].

In Chap. 4, we deal with dispatching rules. A dispatching rule selects the next job to be processed among the jobs that are waiting in front of a machine group (cf. [29, 116]). Dispatching rules are generally myopic in time and space, and it may be difficult to adapt them to different situations on the shop floor. However, their decision logic is easy to understand, and they can be implemented with less effort on the shop floor. We introduce several simple dispatching rules that are in common use in the semiconductor industry like earliest due date, critical ratio, and least slack (cf. Sarin et al. [274]). Simple dispatching rules answer only the question of which job should be processed on which machine next. In this monograph, we provide results of simulation experiments with different commonly used dispatching rules.

Chapter 4 continues by combining these simple rules into composite rules. We discuss several variants of the apparent tardiness cost rule (cf. Vepsalainen and Morton [311]). A third decision is important for batching machines in addition to assignment and sequencing decisions. In this situation, we have to decide which jobs should form the batch. This decision is typically made by batching rules, and we describe several of the most important of these rules. Finally, we also introduce look-ahead rules (cf. [84, 85, 101]) that take into account the situation of downstream work centers. Look-ahead rules are important in manufacturing systems with sequence-dependent setups and batching.

In contrast to dispatching, scheduling approaches consider a time horizon for decision-making and not only a discrete set of points of time. We discuss the use of scheduling techniques in semiconductor manufacturing in Chap. 5. Scheduling is defined as the process of allocation of scarce resources over time [34, 240]. The goal of scheduling is to optimize one or more objectives in a decision-making process. The two major categories in scheduling are deterministic and stochastic scheduling. Deterministic scheduling is characterized by processing times, setup times, and job priorities that are known in advance. They are not influenced by uncertainty. In contrast, stochastic scheduling problems do not assume the existence of deterministic values for processing times, setup times, or other quantities that are used within the scheduling model. The deterministic values are replaced by corresponding probability distributions. Deterministic scheduling problems can be further differentiated into static problems where all jobs to be scheduled are available at time $t = 0$. Dynamic scheduling problems relax this condition. In this situation, jobs are ready at different points in time, i.e., $t \geq 0$.

Simulation-based scheduling means that simulation is used to determine schedules with a horizon ranging from several hours to a day. Dispatching rules that are already part of the simulation engine are used to determine what will be processed next on each machine. The assignment and the sequencing of jobs observed in the simulation are used to produce a control instruction in the original production control system that is used to influence the base system. Simulation-based scheduling relies to a large extent upon the capability to produce simulation models that represent the base

system and the base process in a very detailed manner. Automated or semi-automated simulation model generation abilities based on data in operational systems like the manufacturing execution system (MES) are necessary in order to run a simulation-based scheduling system. The selection of a final schedule as production control instructions can be based on several criteria (cf. Sivakumar [284] for a more detailed description of this approach). Usually, all stochastic effects, for example, machine breakdowns, are turned off because of the short time horizon. Appropriate model initialization is a non-trivial issue in simulation-based scheduling. Simulation-based scheduling is somewhere between dispatching and more traditional scheduling.

According to the suggested planning and control hierarchy, we consider single machine-related and work center-related scheduling problems. These scheduling problems may be the result of decomposition techniques that divide the overall, full factory scheduling problem into scheduling problems for single machines or work centers. Hence, single machine or parallel machine scheduling problems are the building blocks of full factory scheduling problems. On the other hand, this type of scheduling problem may arise independently for the processing of jobs on bottleneck machines. Many scheduling problems are known to be NP-hard (cf. Brucker [34]). Therefore, we often resort to using efficient heuristics. In this monograph, we describe mainly dispatching rule-based techniques and approaches based on genetic algorithms. We consider scheduling problems for single and parallel batch machines and for parallel machines with sequence-dependent setup times.

Furthermore, we also consider cluster tools as special mini fabs. We discuss modeling issues for cluster tools. The scheduling of jobs on cluster tools is challenging because cluster tools consist of parallel chambers that do not allow for a straightforward estimation of processing times of a job on these machines. Simulations and neural networks are used to perform this task. We discuss the scheduling of single and parallel cluster tools.

After the discussion of single and parallel machine scheduling models, we consider full factory scheduling problems. Until recently, full factory scheduling methods seemed to be too costly in comparison to dispatching methods. However, with the recent dramatic increase in computer efficiency, full fab scheduling methods have become more competitive. Because we can model a wafer fab as a complex job shop, we also have to deal with scheduling approaches for large-scale job shops.

We describe the shifting bottleneck heuristic by Adams et al. [1] as an important representative of scheduling heuristics for complex job shops. The main idea of the shifting bottleneck heuristic consists in using disjunctive graphs to model the dependency of job processing on different machines. Based on the calculation of longest paths within the disjunctive graphs, the overall scheduling problem is decomposed into smaller, more tractable scheduling problems for single or parallel machines. After the solution of these subproblems, the structure of the graph has to be updated to incorporate the scheduling decisions that were made.

We then describe modifications of the shifting bottleneck heuristic suggested by Mason et al. [172, 175]. These modifications take batching machines and reentrant flows into account with the goal of minimizing weighted total tardiness of the jobs. While subproblem solution techniques are often based on dispatching rules, we also present more sophisticated approaches based on genetic algorithms (cf. Mönch et al. [206]). We then describe simulation experiments that allow for the application of the shifting bottleneck heuristic in a rolling horizon manner in a dynamic environment (cf. Mönch et al. [202]).

Because of the reduction of solution complexity, distributed solution heuristics for production control problems seem to provide some advantage. Usually, it is difficult and time-consuming to collect in one place all the required data for centralized algorithms in real-world manufacturing systems. From this point of view, distributed algorithms working on local data provide a highly desirable approach. Therefore, we also discuss a distributed variant of the shifting bottleneck heuristic (cf. Mönch and Driessel [193]). This heuristic is based on a two-level hierarchical approach. The upper level determines expected start dates and completion dates for the jobs with respect to a certain work area, i.e., a collection of work centers, in a first step. Then in a second step, we use this information in order to apply the shifting bottleneck heuristic for the jobs in each work area. The schedules for the single work areas can be improved by using an iterative improvement technique. The distributed shifting bottleneck heuristic requires less memory and can be distributed on several computers.

We also describe an extension of the shifting bottleneck heuristic from the single-objective case to the multiobjective case (cf. Pfund et al. [235]) via a desirability function approach. By using this approach, we can model preferences for certain objectives by appropriate weight settings.

Dispatching and scheduling systems assume jobs have already been started in the base system. In Chap. 6, we discuss order release approaches (cf. Fowler et al. [87]). After a brief overview of the general push and pull philosophies, we describe the starvation avoidance approach (cf. Glassey and Resende [100]) and the workload regulation approach (cf. Wein [318]). We then discuss the use of CONWIP-like (cf. Spearman et al. [290]) order release strategies in semiconductor manufacturing. We also present work that investigates the interaction of order release schemes with the full factory scheduling approaches presented in Chap. 5. The main results of a simulation study to find appropriate order release schemes are also discussed. An optimization formulation is presented that supports order release decisions in wafer fabs.

One prerequisite for order release is (operational) capacity planning. Therefore, we consider capacity planning approaches in semiconductor manufacturing in Chap. 7. The basic problem consists of allocating production capacity to alternative products over time in the occurrence of forecasted demands to optimize some performance measure of interest.

We describe simple (static) spreadsheet-type models for this production planning problem (cf. Ozturk et al. [224]). Furthermore, we also discuss the

use of simulation models that take the dynamics of the wafer fab into account better than the spreadsheet models do. Besides short-term capacity planning schemes, we also discuss models used for medium- and long-term capacity planning. Linear and stochastic programming are used to solve these kind of problems. We discuss the problem of modeling load-dependent cycle times. Therefore, we consider iterative simulation techniques due to Hung and Leachman [120], provide a methodology to efficiently generate cycle time throughput curves (cf. Fowler et al. [86]), and introduce clearing functions due to Srinivasan et al. [292], Karmarkar [136], and Asmundsson et al. [12].

Agrawal and Heragu [3] indicate that semiconductor wafer fabs are highly automated manufacturing systems. Compared to other industries, accurate and up-to-date data are available. Therefore, we have a rather good starting position for establishing more advanced production planning and control approaches in wafer fabs. On the other hand, we also have to deal with the operational information systems on the shop floor. We have to analyze the current state of these systems. Furthermore, based on the presented methodological framework in the monograph, we derive future needs for production planning and control systems in Chap. 8.

An MES is an operational information system that is between the enterprise resource planning (ERP) systems and the base process. We describe the core functionality of an MES, which consists of providing correct information about the process flows, the machine set, the status of machines, and also the status of jobs (cf. McClellan [178]). An MES also sometimes supports the implementation of dispatching and scheduling algorithms and decisions. Scheduling and dispatching heuristics have to use the data from the base process that is contained in the MES. Furthermore, an MES is used to provide maintenance and quality assessment functionality. These information systems work together with higher level operational systems like ERP systems and with different databases. Because most of the commercial MESs have difficulties with the integration of more sophisticated dispatching and scheduling approaches, we suggest the use of scheduling systems in a plug-and-play manner via an object-oriented data layer. The data layer acts as a mirror of the base process, and its objects are updated in an event-driven manner (cf. Mönch et al. [190]).

Usually, an MES does not provide adequate dispatching and scheduling functionality. Therefore, we discuss several extensions of MESs for this purpose. These extensions are typically software systems on their own. We start with dispatching systems that are quite common in the semiconductor industry (cf. Pfund et al. [234]). We describe the main architecture of such a system and its interaction with both the MES and ERP systems. Scheduling systems are also discussed.

Software agents allow for the implementation of distributed planning and control algorithms. The agents are able to act autonomously; on the other hand, their communication abilities ensure a cooperative behavior and the fulfillment of global system goals. Furthermore, agent-based systems

facilitate maintenance and further development tasks of the software [319, 324]. In this book, we present the design and the architecture of a multi-agent-system for wafer fabs called FABMAS. The FABMAS system is aimed at production control of wafer fabs. We differentiate between decision-making agents and staff agents. Staff agents encapsulate scheduling logic. They support the decision-making agents in the course of their decision-making. The suggested architecture is used to implement the distributed version of the shifting bottleneck heuristic on a cluster of computers.

Manual material handling is typical for 200-mm wafer fabs. A 300-mm wafer usually visits several hundred machines to perform hundreds of different process steps. Modern 300-mm wafer fabs consist of several bays. To transport wafers, front-opening unified pods (FOUPs) are used as wafer carriers. The FOUPs have to be transported not only within a bay but also from one bay to another. Material control systems (MCSs) are used to initiate and coordinate concurrent movements of carriers within the automated material handling system (AMHS). Therefore, material handling is a critical issue in wafer fabs. An AMHS is an important tool to achieve the goal of reducing cycle time and improving yield rates (cf. Agrawal and Heragu [3]). Running and controlling an AMHS is challenging. Advanced software is required to run an AMHS that performs all the material handling requirements. We discuss the functionality of such MCSs for an AMHS and their interaction with the MES.

Wafer fabs, i.e., the front-ends, are geographically distributed over North America, Europe, and Asia, but most assembly and test sites, the back-end, are in the Pacific Rim. Thus, production orders have to be coordinated between the front-end and back-end facilities, and the management and coordination of the entire supply chain is an important issue. Supply chain management functionality is usually provided by advanced planning systems (APSs). We describe the main functionality of such systems. Furthermore, we also discuss the interaction of an APS with ERP systems and with the MES.