Chapter 2
Semiconductor Manufacturing Process Description

In this chapter, we provide a process description of semiconductor manufacturing. Therefore, we describe the front-end and back-end areas in some detail. We introduce the notion of base system, base process, control system, control process, planning system, planning process, and finally of the information system from systems theory. Then we discuss important wafer fabrication operations including a description of the most important characteristics of the semiconductor manufacturing process. We also introduce the notion of complex job shops because this is the way wafer fabs are organized. Finally, we discuss the production planning and control (PPC) hierarchy in semiconductor manufacturing.

From an operations management point of view, we differentiate between planning at the highest level, order release, scheduling, and finally dispatching at the lowest level. Each of these different functionalities is related to a certain horizon. Planning has a time horizon ranging from months to years. Order release takes place in a weekly or biweekly frequency. Scheduling is performed each shift or day. Finally, dispatching is carried out in a minute-by-minute manner depending on the speed of the material flow. We develop this PPC hierarchy because it forms the skeleton for the remainder of this monograph.

2.1 Semiconductor Manufacturing Overview

A semiconductor chip is a highly miniaturized, integrated electronic circuit consisting of thousands of components. Every semiconductor manufacturing process starts with raw wafers, thin discs made of silicon or gallium arsenide. Depending on the diameter of the wafer, up to several thousand identical chips can be made on each wafer by building up the electronic circuits layer-by-layer in a wafer fab. There are about 40 layers for the most advanced technologies. Next, the wafers are sent to sort or probe, where electrical tests identify the individual dies that are not likely to be good when packaged.

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_2,
© Springer Science+Business Media New York 2013

Historically, bad dies were physically marked so that they would not be put in a package. Today, this has been replaced by producing an electronic map to identify the bad dies. The probed wafers are sent to an assembly facility where the dies with a reasonable quality are put into an appropriate package. Finally, the packaged dies are sent to a test facility where they are tested in order to ensure that only good products are sent to customers. Wafer fab and sort are often called front-end, and assembly and test are often called back-end. While front-end operations are often performed in highly industrialized nations, back-end operations are typically carried out in countries where labor rates are cheaper.

Considering the scale of integration, the type of chip, the type of package, and customer specifications, the whole manufacturing process may require up to 700 single process steps and up to 3 months to produce. The four main stages of semiconductor manufacturing are shown in Fig. 2.1.

In the past, all that was necessary for a semiconductor company to make money was to design a good product. However, over the last decade, increased competition has required semiconductor companies to also be able to manufacture their products in an efficient and cost-effective manner.

Several performance measures are commonly used to describe and assess semiconductor manufacturing systems including machine utilization, production yield, throughput, cycle time, and on-time delivery performance-related measures. Machine utilization is extremely important because the machines account for around 70% of the cost of a new wafer fab, which can be as high as $5 billion US. In this context, cycle time is defined as the time needed for a lot of wafers, called a job, to travel through the semiconductor

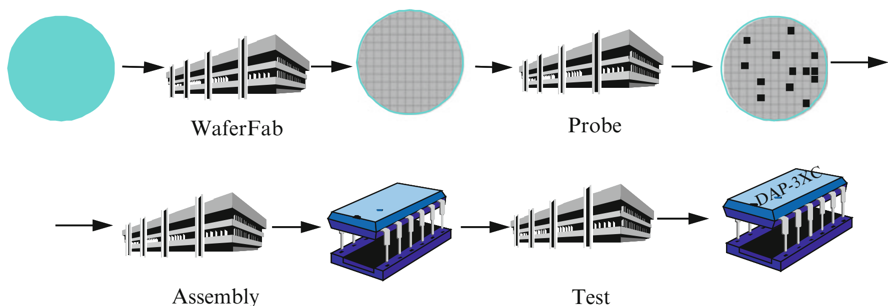

Figure 2.1: Stages of semiconductor manufacturing

manufacturing system including queue time, processing time, and transit time. Each job contains a fixed number of wafers. A high on-time delivery performance is important to satisfy customers. We also refer to Sect. 3.3 where these performance measures are introduced in a more formal way.

The competitiveness of a semiconductor manufacturer often depends on the ability to rapidly incorporate advanced technologies in electronic

products, continuous improvement of manufacturing processes, and the capability of meeting customer due dates. In a situation where prices as well as the state of technology have settled at a certain level, the capability of meeting due dates along with the reduction of cycle time has become the most decisive factor in the fierce competition in the global market place. Consequently, short and predictable cycle times are highly desirable.

Semiconductor companies have increasingly turned to data-intensive modeling and analysis tools and techniques because of their potential to significantly improve these performance measures, and hence the bottom line. The semiconductor manufacturing modeling and analysis community has been working over the last 20 years to modify general purpose manufacturing modeling tools and techniques to handle the intricacies and complexity of semiconductor manufacturing.

2.2 Front-End and Back-End Operations

In this section, we start by an overall framework for manufacturing systems. Then, we discuss the base system and finally the base process of semiconductor manufacturing.

2.2.1 Overall Framework for Manufacturing Systems

Before we describe front-end and back-end operations in detail, we present a framework that is used in the remainder of this monograph to discuss PPC problems in semiconductor manufacturing.

We start in a general manner with systems. A system consists of a set of interacting components. Each single component of a system has a state. We introduce processes to deal with dynamic aspects of systems. A process is defined as a mapping between a partially ordered set of events E and actions A, i.e., exactly one action $a \in A$ is assigned to each event $e \in E$. The partial order of the elements of E might refer, for example, to the points of time where the events happen, but it can also represent precedence constraints among the events. Typically, we describe the system and at the same time the corresponding processes. The actions of the processes are performed on system components.

In this monograph, we study manufacturing systems. Manufacturing systems are systems that have the purpose to produce goods. A manufacturing system consists of a base system (BS) and an information system (IS). The BS is formed by system components that are used to transform raw materials and intermediate products into final products. It contains a job processing system (JS) and a material flow system (MS) as subsystems. The JS consists of all the system components that allow for value-added processing of working objects, i.e., jobs. The system components of the JS offer capacity for processing. Resources, like machines, operators, and auxiliary resources, form the JS. On the other hand, all the facilities that are

necessary to store, transport, and supply raw material, working objects, and auxiliary, also called secondary resources, form the MS.

The base process (BP) is resonsible for the usage of system components of the BS by working objects. Of course, we can differentiate between subprocesses related to the JS and the MS, respectively. The BP is specified by process flows, also called routes, and a given set of working objects. A process flow in semiconductor manufacturing is a sequence of process steps. A set of possible machines is assigned to each single process step within a process flow. A recipe is an execution program at a certain machine that is associated with a process step.

The IS is responsible for the control of the production of goods. It is given by the planning system (PS), the control system (CS), and the operational system (OS). The PS consists of a set of computers and software that are used to determine production planning instructions mp. Production planning results in quantities and points of time for releasing working objects into the BS. The production planning process (PP) determines when and under which circumstances certain production planning actions have to be performed. Similarly, the CS is given by a set of computers and software that are used to determine production control instructions mc that influence the BP. As a consequence, production control decisions only have impact on working objects that are already part of the BP. The corresponding control process (CP) determines when and in which situations a certain production control algorithm is used to determine production control instructions.

Finally, the OS is responsible for immediate control of the BP. The OS usually consists of hardware and software to represent the state of system components of the BS and working objects of the BP. It acts as a mirror of the BS and the BP. Usually, databases are used to implement the OS. The PS, the CS, the OS, and the human decision makers form the IS.

The OS, the CS, and the PS interact by instructions and feedback. A more detailed description of these interactions is provided in Sect. 2.3. The different subsystems and subprocesses of a manufacturing system and the corresponding manufacturing process are summarized in Fig. 2.2.

We will use the notation from Fig. 2.2 when we describe the manufacturing system and process for semiconductor manufacturing. In Sects. 2.2.2 and 2.2.3, we start with the BS and the BP, whereas the PS, the PP, the CS, and the CP are discussed in Sect. 2.3.

2.2.2 Description of the Base System

In this monograph, we will mainly focus on modeling and analyzing of wafer fab operations. These operations generally account for more than 75% of the total cycle time and are also the largest component of cost. However, for the sake of completeness, we will also briefly discuss back-end issues.

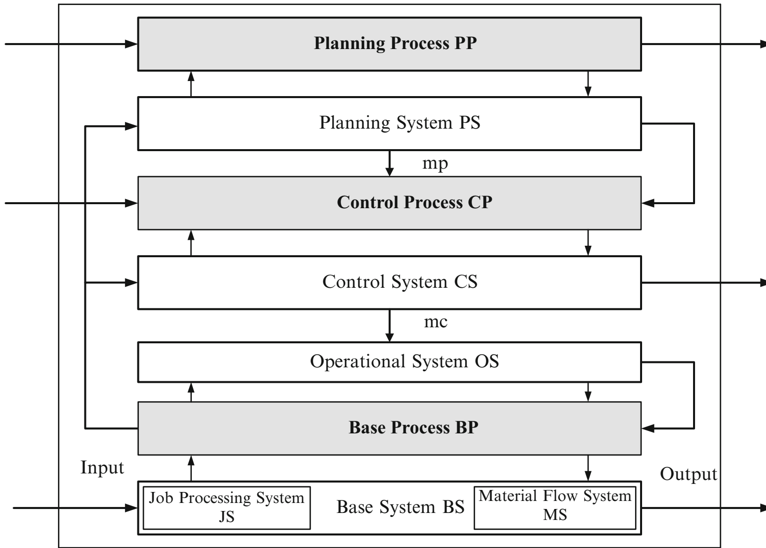

Figure 2.2: Subsystems and subprocesses of a manufacturing system and process

We start by describing the BS for semiconductor manufacturing. The entire enterprise may consist of several wafer fabs and back-end facilities, i.e., the BS of an enterprise is a collection of such factories.

In a first step, we will discuss the JS of a single wafer fab. The JS of a wafer fab consists of several work areas that are used for wafer processing and sorting in a clean-room environment. Each work area is a collection of work centers that are closely related logically or due to their location. Work areas are also called bays when the relation among the work centers is based on the location. A work center is a collection of machines that provide similar processing capabilities. Work centers are also called tool groups in semiconductor manufacturing. A single machine is a non-human resource with a fixed location that is able to process jobs. It can have a buffer where jobs are stored before, during, or after processing. For 200-mm wafer fabs, these buffers are often assumed to be practically unlimited, whereas they usually have a limited capacity in 300-mm wafer fabs. The following machine states are possible according to SEMATECH [280]:

- Productive state: a period of time during which the machine is performing its intended function
- Standby state: a period of time during which the machine is not operated, although it is in a condition to perform its intended function, and the chemicals and facilities are available
- Engineering state: a period of time during which a machine is in a condition to perform its intended function but is operated for the purpose of conducting engineering experiments

- Scheduled downtime state: a period of time during which the machine is not available to perform its intended function because of planned downtime actions
- Unscheduled downtime state: a period of time during which the machine is not available to perform its intended function because of unplanned downtime actions
- Nonscheduled state: a period of time during which the machine is not scheduled to be utilized in production including off-line training, unworked shifts, weekends, and holidays

Some of the machines are able to process only one wafer or a job at a certain point of time, whereas some types of machines process two jobs in an overlapping manner, i.e., the second job is already started after the first one has been processed for a certain period of time. These machines are called pipeline tools. Another example of specific machines that can be encountered in wafer fabs are X-piece machines, in which X wafers of a job are loaded at a time in the machines and where X is smaller than the total number of wafers in a job, except for small jobs. Other types of machines can process entire batches. A batch is defined in semiconductor manufacturing as a collection of jobs that are processed at the same time on the same machine (cf. Mathirajan and Sivakumar [176]). Besides batch-processing machines (for short, batch machines in the remainder of this book), we also introduce cluster tools that are typical for many modern wafer fabs.

Cluster tools are special integrated tools for wafer processing in semiconductor manufacturing (cf. Lee [157]). They are used to maximize quality performance at the cost of very complex behavior. Since wafers with different types of process steps can circulate in a cluster tool simultaneously, it can be regarded as a fully automated machine environment. Cluster tools work under vacuum conditions inside the tool, which means very few particles could possibly contaminate wafers. As a consequence, the clean-room quality outside the cluster tool is allowed to be lower than in traditional wafer fabs. The basic components of a cluster tool are as follows:

- A vacuum mainframe with one or two wafer-handling robots
- Two (or more) load locks to pump to vacuum or vent to atmospheric conditions
- Several processing chambers, where some of them can be dedicated to identical processes and hence used in parallel
- Optional transfer chambers if there is more than one wafer-handling robot
- An equipment front-end module (EFEM), which is attached to the load locks, with an atmospheric wafer-handling robot and several load ports

These basic components of a cluster tool are shown in Fig. 2.3. The processing of wafers on cluster tools is included in the description of the semiconductor BP later in Sect. 2.2.3.

In wafer fabs, we typically have dozens of different work centers and several work areas. A more detailed description of the functionality of work

centers and work areas is also included in the BP description in Sect. 2.2.3. For each machine, there is a list of which products are allowed to be performed on the machine because of quality considerations, i.e., we find machine dedications. They are mainly influenced by the fact that qualifying a certain machine for a specific process is time-consuming and therefore expensive (see Johnzén [133]). The machines of a work center are heterogeneous because they tend to have a different age. A total of several hundred machines can be found in most wafer fabs. Machines are expensive, ranging in

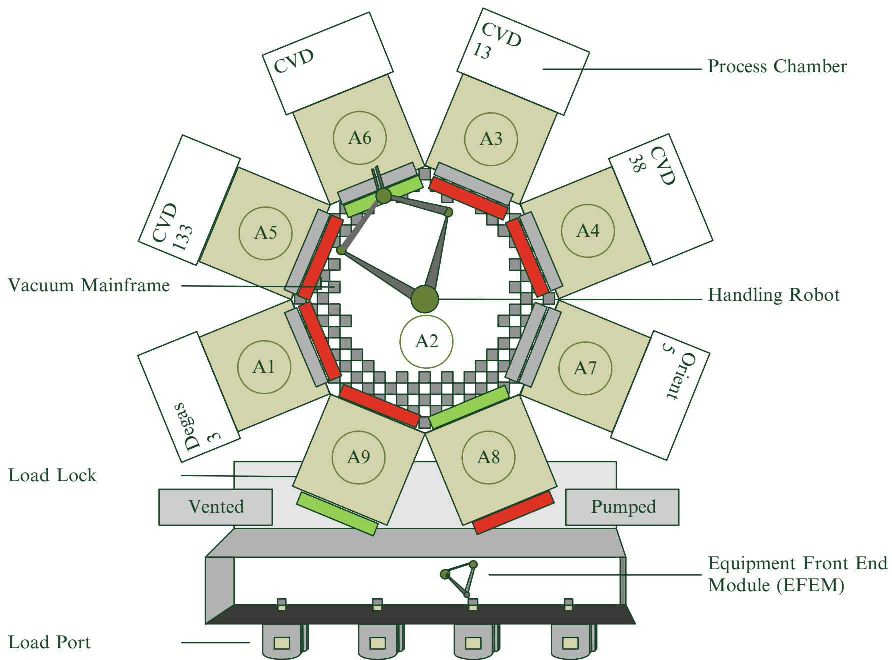

Figure 2.3: Basic components of a cluster tool

price from a couple of US$100 thousand to over US$40 million per machine. Note that in a few cases there are borderless wafer fabs, i.e., facilities that are in close geographical proximity (cf. Gan et al. [93]). In this situation, the different wafer fabs can share some of their machines to produce ICs.

Back-end facilities also consist of work areas and work centers, but the number of machines is usually smaller than in wafer fabs. Furthermore, the clean-room conditions that are required are less strict than those for wafer fabs.

Operators as special human resources are necessary to run a wafer fab. In all but the most highly automated wafer fabs, humans load and unload wafers to machines, run the machines, and perform inspection steps. Often highly qualified and experienced operators are assigned to supervise work areas. In addition, there are generally several people responsible for the overall performance of the work areas; these people may spend the majority of their time outside the shop floor.

Typical auxiliary resources used in semiconductor manufacturing are reticles. A reticle is a photo mask that is a carrier of an IC pattern. They are used in the photolithography process. In back-end facilities, load boards are applied to load and place chips into burn-in ovens to subject them to thermal stress. The type of board required to test chips depends on the packaging of the circuits. Both reticles and load boards are quite expensive, and therefore their number is often limited. In a certain sense, operators might also be considered as an auxiliary resource that is necessary to process wafers on machines.

After the description of the JS, we discuss the MS of a wafer fab [82, 131]. In modern 300-mm wafer fabs, wafers and reticles are transported fully automatically in carriers called FOUPs, using an AMHS. A FOUP is a container that holds up to 25 wafer jobs of 300-mm wafers in an inert, nitrogen atmosphere. Automated material handling is always a critical operation in wafer fabs [3, 209]. In many 200-mm wafer fabs and in most back-end facilities, material handling is usually carried out manually. Note that machines are the main resources in the JS while vehicles take over this role in the MS. Jobs compete for the scarce machine and vehicle capacities.

We now need to differentiate between interbay systems and intrabay systems. Interbay systems are used to store and transport wafers or reticles between the various bays of a wafer fab. On the other hand, intrabay systems have the purpose to move carriers for wafers and reticles within a bay. In 200-mm wafer fabs automated interbay systems are common, but the transport within a bay is done manually. The automated transport in 300-mm wafer fabs also covers transport within a bay. This is caused by the increase in area and weight of the wafers. Furthermore, because of more advanced software and hardware solutions in 300-mm wafer fabs, a better integration and control of the interaction between machines and automated transportation systems is possible.

We start with the interbay situation. The interbay MS consists of carriers for wafers and reticles, stockers, and the transportation system itself. A stocker is an automated high-rack storage area where wafers and reticles can be stored before and after being processed. A robot moves the carrier into a shelf when it is inside the stocker. There are different

possibilities for stocker placement within a wafer fab. Often a single stocker is assigned to exactly one bay by locating it near the entrance of a bay. A discussion of different AMHS layouts can be found in Agrawal and Heragu [3]. Each stocker may contain several input/output stations, called load ports. Load ports are used for manual or automated load and unload operations of carriers by fab-operating personnel or the automated transportation system, respectively. If processing of wafers is done on different levels, i.e., floors, of a wafer fab, then interlevel lift systems are used to transport carriers between these multiple levels. The transport system transports carriers between different bays of a wafer fab. Overhead monorail-tracked vehicle-type systems are in widespread use in industry. When an overhead transport is not practical, then a floor running automated guided vehicle (AGV) system is typical for interbay transportation as shown by Foster and Pillai [82]. Note that interbay transportation takes place from stocker to stocker or from stocker to an interlevel lift system and vice versa.

In contrast, in the intrabay situation, the transportation is carried out between stockers and machines or directly between machines of a bay. Floor-based transport systems like AGVs and rail guided vehicles (RGVs) have been used in industry. In many 300-mm wafer fabs, ceiling-based overhead hoist vehicles (OHVs) are run instead of AGVs or RGVs. This type of transportation is called overhead hoist transportation (OHT). Stockers are equipped with intrabay input/output ports to allow carriers to enter and exit the intrabay system. Stockers are usually far away from a specific machine where a job is needed. This can lead to long delivery times and under track storage (UTS) is proposed to avoid this disadvantage (Fischmann et al. [80]). UTSs are single buffer storages that are mounted overhead but under the track system. They are passive shelves that do not require floor space in the clean room. OHVs can place carriers for temporary buffering in route to their destinations. They provide additional queue positions for high throughput machines (Foster and Pillai [82]). Load ports at the machines are the primary buffers besides stockers and UTSs. A single machine typically has three to four ports. The intrabay vehicles deliver unprocessed FOUPs to the ports and pick up finished wafers. A single OHV can transport a carrier directly from machine load port to machine load port.

Operators are often elements of the nonautomated part of the MS. They drive manual carts or personal guided vehicles (PGVs), especially in 200-mm wafer fabs or in ramp-up situations where the automated transport system is not running in 300-mm wafer fabs.

Two intrabay system configurations are in current use in industry. We differentiate between a unified transport configuration and a non-unified one. In the unified configuration, the track network for interbay and intrabay is connected directly, i.e., no load and unload at stockers are necessary to decouple the two transport loops. The track elevations of the interbay and intrabay system have to be the same.

In the second configuration, stockers are used to exchange the load of the vehicles between interbay and intrabay segments. The track elevations of the interbay and intrabay systems can be different. Within each bay, intrabay vehicles are in place to transport carriers from the bay stocker to the storage facilities of the bay. On the other hand, interbay vehicles transport carriers to the destination bay stockers. We finish this brief description of the MS by making the comment that often the MS is as complex and difficult to control as the JS and a source of problems in many wafer fabs.

2.2.3 Description of the Base Process

We continue by describing the BP of a wafer fab. Many authors have discussed the difficulties of the semiconductor manufacturing process (cf. Wein [318], Uzsoy et al. [306], Atherton and Atherton [14], Ovacik and Uzsoy [223], Sze [294], Sarin et al. [274], among others). Up to now, work areas are identified as an important component in the JS hierarchy. Now we describe the basic process steps, i.e., the operations, that can be performed in different work areas. The following process steps have to be performed in a wafer fab after starting the raw wafer [14, 122]:

1. Oxidation/diffusion: A layer of material is grown or deposited on the surface of a cleaned wafer. Oxidation aims at growing a dioxide layer on a wafer. Diffusion is a high temperature process that disperses material on the wafer surface. Diffusion furnaces and rapid thermal processing equipment are in place at the oxidation/diffusion work area. The furnaces are typical batch machines.
2. Film deposition: Deposition is used to deposit films onto wafers. The corresponding steps deposit dielectric or metal layers. There can be a dozen or more such deposition layers in an advanced circuit. Deposition can be executed by different processes, such as physical vapor deposition (PVD) or chemical vapor deposition (CVD), epitaxy, or metalization.
3. Photolithography: Coating, exposure, developing, and process control are the main steps of the photolithography process. In the first step, the wafer is coated with a thin film of a photosensitive polymer, called photoresist strip. Accurate and precise three-dimensional patterns are produced on the silicon wafer's surface when an IC pattern is transferred via a photo mask, i.e., reticle, onto the photosensitive polymer, which replicates the pattern in the underlying layer. Exposure tools, called steppers, transfer the pattern onto the wafer by projecting light through the reticle to expose the wafer using ultraviolet light. The exposed wafer is then developed by removing polymerized sections of photoresist from the wafer. Every wafer passes through the photolithography area up to 40 times because the circuits are made up of layers. The photolithography work area is a typical example of a bottleneck in a wafer fab because steppers are very expensive machines.

4. Etch: This step is responsible for removing material from the wafer surface. The wafers are partially covered by photoresist strip after the photolithography step. Areas on the wafer that are not covered are then removed from the wafer. We differentiate between wet and dry etching. In the first case, liquids are used, whereas gases are necessary for the latter case.
5. Ion implantation: Dopant ions are selectively deposited on the surface of the wafer. Doping material is deposited where parts of the wafer have been etched. Ion implanters are used for between four and eight applications for most modern ICs.
6. Planarization: This step cleans and levels the wafer surface. It is called chemical-mechanical polishing (CMP). A chemical slurry is applied to a wafer and the surface is equalized. This results in the thickness of the wafers being diminished before adding a new layer.

Before the wafers are entered into the oxidation/deposition/diffusion work area, a cleaning step is performed. Several inspection and measurement steps are necessary to control the processes within and between work areas. Inspection machines can be found in all work areas.

At certain process steps, it can happen that jobs, wafers, or dies are processed in a way that they become damaged. In some situations, rework is possible to repair the wafer. When rework is not allowed, the useless wafers are called scrapped material. The yield is the percentage of dies that meet their electrical specifications.

Wafer fabrication has a number of unusual facets that are described below. In a typical wafer fab, there often are dozens of process flows. Products that follow the same basic process flow are often said to be of the same technology. Typically, the only differences among products in the same technology are the photolithography reticles used. Individual products are often referred to as devices. Depending on the number of different products, or product variety, wafer fabs can be classified as low- or high-mix wafer fabs. In low-mix wafer fabs, machines can be dedicated to products, whereas, in high-mix wafer fabs, the same machine can be shared by many products of various technologies, i.e., requiring different setup and processing times. Hence, production control is more complex in high-mix fabs, and efficient production control is usually more critical.

Each process flow contains 300–700 process steps on more than one hundred machines. The economic necessity to reduce capital spending dictates that such expensive machines be shared by all jobs requiring the particular processing capabilities provided by the machine, even though they may be at different stages of their manufacturing cycle. This results in a manufacturing environment that is different in several ways from both traditional flow shops as well as traditional job shops. A job shop is characterized by an individual process flow of each single product using different machines, whereas a flow shop is described by the fact that all products have a fixed machine sequence, i.e., the jobs are processed on the same sequence of machines or work centers. The main consequence of the reentrant flow nature is that wafers at different

stages in their manufacturing cycle have to compete with each other for the same machines.

A simplification of the typical reentrant flow of a wafer fab is shown in Fig. 2.4 where the different work areas within a wafer fab are also summarized.

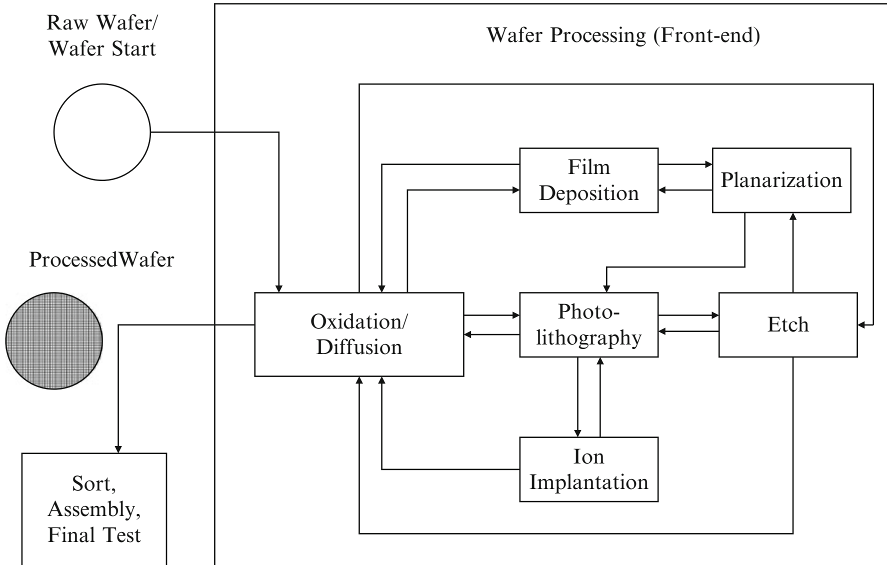

Figure 2.4: Operations in a wafer fab

Furthermore, the nature and duration of the various process steps in a process flow differ significantly. Some process steps require 15 min or less to process a job, while others may require over 12 h.

Many of these long operations involve batch processes. In reality, it is not uncommon for one-third of the wafer fab operations to be batch operations. Batch machines tend to off-load multiple jobs, 1 to 12, onto machines that are capable of processing only one job at a time. This leads to the formation of long queues in front of these serial machines and ultimately to a nonlinear flow of products in the wafer fabs. The diffusion furnaces in wafer fabs are an example of batch machines. Here, the jobs are assigned to incompatible job families. While several jobs can be processed at the same time, jobs of different families cannot be processed together due to the chemical nature of the processes. The processing time of all jobs within a family is the same.

The combination of decreased line widths and more area per wafer in 300-mm wafer fabs results in fewer wafers being needed to fill an IC order of a customer. Each wafer fab will have only a limited number of FOUPs as they are expensive. A large number of FOUPs have the potential to cause the AMHS to become overloaded. In addition, some machines have the same processing times regardless of the number of wafers in the batch. Thus, it

is not reasonable to assign an individual FOUP to each order. Therefore, 300-mm manufacturers often have the need and the incentive to group orders from different customers into one or more FOUPs to form production jobs. We refer to this as the multiple orders per job problem (cf. Mönch et al. [207]).

Historically, equipment reliability has been a major source of uncertainty in wafer fabs. The continual drive to reduce the size of ICs has led to the adoption of machines that have innovative processing modes but have been built by manufacturers without a long history of developing manufacturing equipment. The failure of equipment or processes is often not a hard failure in the sense that something obviously breaks or goes wrong but rather a soft failure in which the equipment begins to produce out of the tolerance region. Due to the nature of the product and process, one may not detect this fact for some time. Therefore, many inspection steps are added to the process flow. The large downtime of some wafer fabrication machines, for example, ion implanters may be down 30–40% of the time, has significant impact on the production control function. The probabilistic occurrence of long machine failures results in large variability in the time a job spends in process. High variability in cycle times prevents accurate prediction of production cycle times, resulting in longer lead-time commitments.

Preventive maintenance operations are used to reduce the number and the duration of machine failures. But at the same time, they reduce machine capacity. There is also a competition between production jobs and prototype jobs for processing times on the machines. Many prototype/engineering jobs are necessary because of the difficulty of the technological processes. Prototype jobs also consume machine capacities.

Often, certain jobs in a wafer fab are more important than others. Then these jobs will be expedited to meet their due dates. They are called hot or rocket jobs. Because of these hot jobs, the congestion of many wafer fabs is further increased.

Time constraints between consecutive process steps are another important restriction. For example, there is often a time restriction between operations in the etch work area and oxidation/diffusion work area (cf. Scholl and Domaschke [275]). Time windows are installed by the process engineering department to respect the time constraints. This is important to prevent native oxidation and contamination effects on the wafer surface. More than two consecutive process steps might be involved, and these time constraints might be nested. Jobs with a violation of the recommended time windows often have to be scrapped since rework is generally not allowed.

Sequence-dependent setup times occur in some work areas and are related to changing temperature, gas pressure, metal composition, etc. It is not only important which product is going to be processed next but also which was the last product processed before the current one. For example, in the ion implantation work area, dopants have to be changed frequently. The effort to do this change depends on the predecessor dopant. If the setups are not

treated correctly by the production control staff, the corresponding machines can become bottlenecks.

Finally, some process steps require an auxiliary resource in order to process the job. For example, reticles are required in photolithography to process wafers on steppers. Therefore, the challenge is to ensure that the machine and the auxiliary resource are available at the same time.

Next, we consider cluster tools as they cause specific processing restrictions. Each wafer of a job has to undergo the same process steps in the cluster tool. This sequence of process steps is usually referred to as a recipe. A typical product flow in a cluster tool starts with loading jobs into one of the load ports. After that, single wafers are consecutively transferred from the load port to the load lock by an atmospheric robot. Then the load lock will pump to achieve vacuum conditions. Next, the mainframe robot can transfer the wafer to its destination chamber where it will be processed. The next step depends on whether the wafer shall leave the system or is required to be processed in another chamber according to its recipe. After the last process step, the wafer will be guided through the load lock back to the load port. With more than one load port occupied, the controller of the cluster tool will always process jobs of the same recipe sequentially one after another and jobs of different recipes in parallel. Usually, the mainframe robot is a dual blade robot with the two blades either on the same side or opposite to each other. Advantages compared to single blade robots are reduced wafer transfer times and, with regard to multiple product flows, a reduced amount of possible deadlocks as well.

Cluster tools can be seen as a collective term for a certain type of machine with a wide range of varying configurations. The number of load ports varies usually between two and four, an EFEM may not be given, and loading stations are directly attached to the load locks. Some cluster tools have two mainframes with a transport robot and transfer chambers to connect them.

Cluster tools can process a certain number of jobs in parallel, which is determined by the number of given load ports at a certain tool. The logic of job processing requires jobs of different recipes to be processed in parallel and jobs of the same recipe to be processed sequentially one after another. Through parallel job processing, resource conflicts arise that result from a shared use of the handling robots, load locks, and process chambers. The conflicts lead to a job processing time that will be increased compared to its stand-alone processing time.

Another effect that needs to be considered is called pipelining, which implies that two jobs of the same recipe need to be processed sequentially. The recipe is such that each wafer needs to go through more than one process chamber. If this is the case, the first wafer of the second job may already start processing in the first chamber if the last wafer of the first job is finished and

no wafer of this job will return to the chamber. Hence, there will be a certain overlap time between the start of the second job and the completion of the first job.

In summary, concluding the description of BS and BP, wafer fabs can be considered as complex job shops [172, 223]. A job shop is called flexible when the same processing capabilities are offered by more than one machine, i.e., machines are run in parallel in the work centers. According to Mason et al. [172], a flexible job shop is called complex when the following processing conditions appear:

- Unequal release dates of the jobs
- Sequence-dependent setup times
- Prescribed due dates of the jobs
- Reentrant flows of the jobs
- Different types of processes, for example, single job vs. batch processing
- Frequent machine breakdowns and other types of disturbances

We do not describe the part of the BP that is related to the MS because it is simpler and it will not be covered in detail in the remainder of this book.

Finally, we briefly consider sort operations to complete the description of the BP of the front-end part of semiconductor manufacturing. The sort operations are performed within the inspection work area. Inspections for defects, film composition, critical measurements, and wafer profiling are performed on automatic inspection stations.

Next, we briefly discuss back-end operations. We differentiate between assembly and final test operations that are performed in the corresponding facilities (Hutcheson [122]):

1. Assembly: In the main assembly work area, typically dicing saw, die attach, wire bonding, and optical inspection operations are performed. Packaging, molding, lid sealing, and environmental testing are usually carried out in work areas that need less strict clean-room conditions.
2. Final test: A series of electrical tests similar to those in wafer sort is performed for the individual ICs to make sure that they meet complex specifications. A heat-stress test of ICs is performed in burn-in ovens. Before the ICs can be processed in the burn-in ovens, they have to be inserted onto a load board. They are kept at a specific temperature for a certain period of time. Then they are packed into tubes and delivered to customers.

There are usually more product types being made in an assembly factory than in a wafer fab, but each product type requires 10–30 steps instead of 300–700 in wafer fabs. One difficulty of these operations is the fact that a job is often divided into subjobs with each subjob being sent to the next machine when it completes an operation. Thus, one job may be being processed across several machines at different steps at the same time.

Another difficulty is that there is often a very significant amount of setup required to change over from one product type to another. Yet another difficulty is that a single type of wafer can become one of several different products based on tested levels of performance, for example, speed. This process is called binning. Finally, batch machines are also often present in assembly factories.

Test operations have several problems that are difficult to treat. First, the sequence of test operations and the test times are not always fixed. These can be changed based on recent product yields or maturity of a product. Second, there are two major types of equipment used in test operations. These are the test system itself, called the tester, and the loading mechanism, called the handler. The tester may have a single or multiple test heads connected to it. The interactions between the tester, the test heads, and the handler can be quite complex. There can be significant sequence-dependent changeover times. The burn-in operations are another example for batch processes in semiconductor manufacturing. In contrast to the diffusion furnaces in wafer fabs, the processing time associated with such a batch is determined by the longest processing time of one of the jobs that form the batch.

In the remainder of this monograph, we will primarily restrict ourselves on the PPC problems found in a single wafer fab. However, in a few situations, we will also discuss PPC problems related to the back-end stage. When we discuss planning approaches in Chap. 7, we consider also the case of simultaneous planning approaches for several wafer fabs or back-end facilities, i.e., we work also on the enterprise level.

2.3 Production Planning and Control Hierarchy

In this section, we discuss the PS and the CS of wafer fabs and also make some comments on the PP and the CP (cf. Sect. 2.2.1 for this notation). The resulting hierarchy forms the starting point for the remaining chapters of this monograph and determines their sequence to a certain degree.

We start by describing the typical decisions that have to be made by the PS and the CS. Planning is performed with a time horizon ranging from months to years. Anticipated demand is an important input for any production planning approach. Planning usually assumes that the time horizon is divided into time buckets with a length of a week or a month. All the planning decisions are related to these time buckets. The results of a typical planning decision are the quantities that have to be released or completed within a certain bucket in such a way that certain performance measures are optimized and the finite capacity of the manufacturing system at an aggregated level is taken into account. Typically, the revenue is considered as a performance measure on the planning level. Certain cycle time assumptions are also the basis for planning decisions. In semiconductor manufacturing, we differentiate between long-term capacity planning that is more strategic and master planning, also called supply network planning, that is more operational. While

capacity planning usually determines the quantities and the product mix for the next years on the enterprise level, master planning has a horizon of several months and assigns quantities to time buckets and also to specific facilities (cf. Vieira [313]).

The PP can be performed in a time- or event-driven manner. Hybrids between these two extrema are also possible. In the time-driven situation, plans are basically determined in a rolling horizon setting. Each planning decision is made for $h := \tau_\Delta + \tau_{ah}$ time buckets. But the planning decisions will be implemented only for the next τ_Δ time buckets in the BS and the BP. After τ_Δ time buckets, a new plan with time horizon h will be determined taking the current state of the BS and the BP into account. The quantity τ_Δ is called the planning interval, whereas τ_{ah} is the additional planning horizon. The event-driven approach initiates the determination of new plans with time horizon h as a consequence of certain changes of the BS or events within the BP.

Order release is at the interface between the planning and the control level. It refines the decisions made on the planning level by disaggregating the quantities in time and space. A typical order release decision results in a set of jobs that have to be launched into a wafer fab at a certain point in time. Weekly or biweekly order release schemes are very common in semiconductor manufacturing. The order release scheme definitely affects the load and consequently the cycle times in wafer fabs. Therefore, order release also influences the planning decisions.

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. Scheduling can be performed for jobs on single machines, work centers, work areas, and finally for all machines of the shop floor, i.e., for the BS. At the same time, scheduling decisions are also made for the vehicles in the MS. The result is a schedule, i.e., an assignment for each job to at least one time interval on the different resources, i.e., machines or vehicles. Scheduling is usually done with a horizon of one shift or one day. Scheduling is only performed for jobs that have been already released into the BS. As in case of the PP, the CP is basically given by the used rescheduling policy. Similar to the planning case, we can differentiate between time- and event-driven rescheduling schemes (Vieira et al. [314]).

Dispatching is the activity to assign the next job to be processed from a set of jobs awaiting service on an available machine in the JS or a free vehicle in the MS [29, 116, 274]. To select the next job, each job is assigned a priority. These priorities can be determined using a schedule. When there is no feasible schedule available, the priorities can be determined by dispatching rules. The priority for each waiting job is calculated by taking different job and resource attributes into account. Dispatching is on the lowest level in the PPC hierarchy. It is performed in a minute-by-minute manner. We show the described PPC hierarchy in Fig. 2.5.

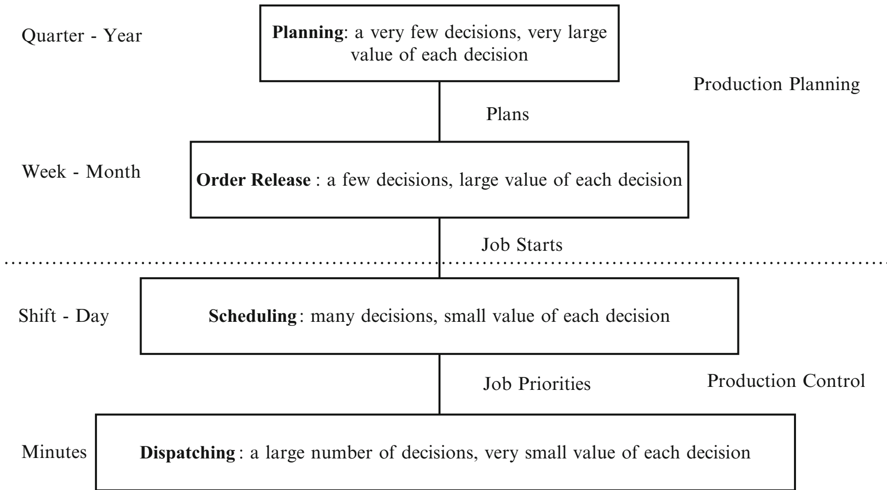

Figure 2.5: Production planning and control hierarchy

In Fig. 2.5, we do not depict the interactions between the different levels in detail. As described in Sect. 2.2, the automated parts of the PS and the CS consist of software that is used to implement the decision-making algorithms and hardware. For the sake of completeness, we will briefly mention the different application systems that are related to PPC decisions in semiconductor manufacturing.

The OS is given by transactional systems like the ERP system or the MES. They are used to gather data from the BS and BP. The ERP system is sometimes also used to support planning decisions. Because of the known shortfalls of ERP systems with respect to planning, often APS or other specialized software systems are in place to support planning decisions of managers. The MES typically contains more fine-grained data. Hence, it can be used to support production control decisions related to the JS. Besides the MES, MCSs are used to control the AMHS and support MS-related decisions. Again, because of the shortfalls of MESs, often more specialized software solutions are applied in wafer fabs [234].