Design of a multi-process multi-product wafer fab

Design of a multi-process multi-product wafer fab

Edgar van Campen

Cover: Maike van Rooij.

Print: Océ Facilities Service, Eindhoven.

/ stan ackermans institute, center for technological design

The work in this thesis has been carried out under the auspices of the research school EM (Engineering Mechanics).

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Campen, Edgar J.J. van

Design of a multi-process multi-product wafer fab / by Edgar J.J. van Campen.
Eindhoven : Technische Universiteit Eindhoven, 2001.
Proefontwerp. - ISBN 90-386-2752-1
NUGI 841

Subject headings: IC wafers; production design / semiconductor industry;
production design / production control / flow-line balancing /
production control; scheduling

Design of a multi-process multi-product wafer fab

PROEFONTWERP

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. M. Rem, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen op
dinsdag 24 april 2001 om 16.00 uur

door

Edgar van Campen

geboren te Pijnacker

Dit proefontwerp is goedgekeurd door de promotoren:

prof.dr.ir. J.E. Rooda

en

prof.dr. E.H.L. Aarts

Copromotor:

dr.ir. L.F.P. Etman

Voor Christel

en

voor mijn ouders

Preface

This thesis presents the results of almost five years work in the fields of wafer fab design and control of wafer fabs. Early 1995 I joined MOS4YOU – founded January 1995 – to conduct the final phase of the Post Graduate Designers Course on Logistic Control Systems. After those six months, I was offered the opportunity to combine the work with the preparation of a Ph.D. thesis, which I gratefully accepted. It was a great experience seeing MOS4YOU grow from 10 to 1000 people and from an empty building to a fab with four production floors in three adjacent buildings.

I am grateful to the management of Philips Semiconductors and MOS4YOU, in particular to Stuart McIntosh – former COO of Philips Semiconductors – for giving me the opportunity to carry out this study while working for the MOS4YOU wafer fab, Jan Smits – former department manager of MOS4YOU – who sought cooperation with the academic world, Ajit Manocha – former general manager of MOS4YOU; John Schmitz – general manager of MOS4YOU – for always stimulating people to go one step further, Gerard de Groot – department manager – who founded the production organization, and Leon Tjalkens – former logistics and production control department manager.

Special thanks to Aubin Wilkens for the numerous illuminations and the critical and stimulating discussions. Being my mentor and demonstrating active interest in the topic, he provided a constant source of inspiration. I also would like to thank my colleagues Hubert Rulkens, with whom I worked closely and had many differing discussions, Frans Brouwers, who never stops asking and because of that made manufacturing control in our fab to what it is now, and my room mates for their pleasant company.

Of the university, I would like to thank professor Koos Rooda for stimulating my interest in doing systematic and structured work on the design of wafer fabs, and for providing the opportunity to do so. His never-ending positive coaching was very motivating. Many thanks to my co-promotor Pascal Etman, our enlightening discussions have taught me that the quality of the solution always depends upon the magnitude of the design space.

I also want to thank the former graduate students who helped with collecting valuable material for the design. In particular I am very grateful to Koen Eijsvogels, Johan Jacobs, and Bart Lemmen for their explicit contributions on dynamic simulations of (parts of) the fab and on gaining insight in the concept of variability. Sharing experiences with Jeroen Fey, who also wrote his Ph.D. thesis while working in industry, was most enlightening. I am surely going to miss our Friday afternoon discussions on comparing findings from the fruit juice industry with those from the semiconductor industry and our competitiveness on the progress of the writing.

For graciously undertaking the task of correcting the English of the thesis, I would like to thank Joke Wilkens - van Herk. Her corrections and recommendations have certainly contributed, and brought the thesis to a higher level.

Last, but not least, I would like to mention my wife Christel: thank you for your love and patience.

Summary

Semiconductor wafer fabrication is one of the most complex manufacturing processes. The complexity is brought about by the high-tech processes being applied, their constant development, the multitude of process steps, and their reoccurring nature. Hence, comprehensive control strategies are required. Nowadays, due to this complexity, building a wafer fab costs over one billion dollars, whereas operating costs run into hundreds of millions of dollars a year. As designing has a great influence on costs and expenditures, a good fab design provides the manufacturer an advantage over his competitors.

The product of a wafer fab is a finished wafer, containing hundreds to thousands of integrated circuits (ICs) on its surface. Semiconductor industry distinguishes manufacturers of high-volume production of ICs, such as microprocessors and memories, and low-volume production of customer specific ICs, also called multi-process multi-product (MP2) wafer fabs. High-volume ICs are generally made to stock, whereas MP2 fabs usually are made in small batches on customer orders. Contrary to high-volume manufacturers, who are merely interested in maximizing productivity, MP2 manufacturers also aim for reliable and short cycle times. The contradiction of running a fab economically, thus at high productivity, and achieving reliable and low cycle times simultaneously is the challenge in designing wafer fabs.

Although a lot of research has been done on specific elements concerning the design of wafer fabs, like layout optimization or scheduling performance, no systematic approach is known that describes the activities and tools needed to design an MP2 wafer fab. The objective of this thesis is twofold. First to structure the design process of an MP2 wafer fab and to describe the accompanying methods and tools needed during the design process. Second to apply the proposed approach to the design of Philips' MOS4YOU wafer fab in Nijmegen.

Designing an MP2 wafer fab is structured into four design activities: determination of objectives and constraints, design of the architecture, design of capacity and layout, and design of the operations. Each successive activity is an extension of the previous one, as it reveals another level of detail. After determining the objectives and con-

straints, the architecture of both the material flow as well as the control system are determined. Design of the architecture is essential to achieve a well-balanced design of the wafer fab. Next, the resources and their location are determined. Finally, during the design of the operations the detailed micro layout, operating procedures, and control policies are determined. Continuous improvement activities are part of the design of operations

Following are the relevant design choices that have been implemented in MOS4YOU. Objectives, constraints, and characteristics of the fab led to a hybrid functional architecture. The layout of the fab is divided into main processing areas, each containing a specific process technology, for example, metal deposition, lithography, etc. Each sub-process area or bay is equipped with the appropriate process-equipment and optimized for its purpose. A partially automated material handling and storage system is applied to obtain controllable and cost effective wafer transport and storage, while flexibility is guaranteed. The production control concept consists of a lot release strategy and a lot sequencing strategy. Lots are to be released at the desired productivity level, however, the amount of work-in-progress triggers exception handling. Once on the shop floor, lots are processed according to a specific sequence of recipes and lots are dispatched for process steps in an order aiming for maximum flow line balance. Improvement studies were carried out, considering specific areas in isolation. During these studies detailed analysis was performed, taking many sources of variability into consideration which resulted in improved floor layouts, improved working procedures and refined sequencing rules.

Contents

Chapter 1

Introduction

Fabrication of integrated circuit wafers is a complex concatenation of manufacturing processes. The number of process steps, the re-entrant nature of the process flow, and the advanced process technologies are, amongst others, the causes of this complexity. Therefore, the design of a semiconductor wafer fabrication facility is of the utmost importance. It determines to a great extent the cost of expenditures, of building as well as operating a fab. A good design provides the manufacturer an advantage over its competitors.

This thesis describes a method for designing IC wafer fabs (integrated circuit wafer fabrication facilities). The current chapter starts with an overview of IC industry. Then the issue of fab design is described which leads to the objective of the thesis. Finally, the structure of the thesis is outlined.

IC industry

In IC industry wafers are manufactured. A picture of a wafer is shown in Figure 1.1. A wafer having completed the manufacturing sequence may contain one hundred to one thousand ICs. By definition, an IC integrates a large number of isolated tiny components into one die, or chip, to be cut from one silicon wafer. The components are the result of photographic patterning and subsequent etching or dope implantation steps on the silicon. They are interconnected through conductive poly-silicon or metal-layers. Manufacturing wafers is a process that takes more than 400 operations.

As fabrication technology becomes more advanced, the size of components in the IC decreases. Consequently, more components can be put in one chip. Combining

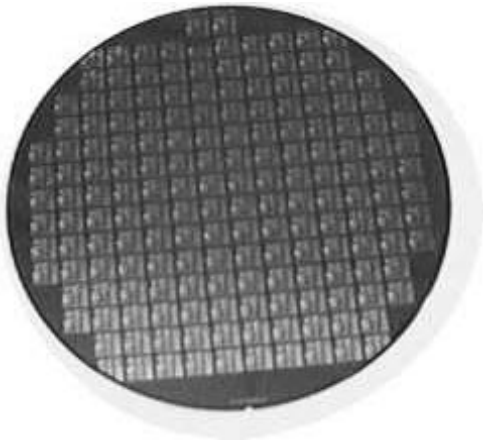

Figure 1.1: Photograph of an eight-inch silicon wafer.

different components leads to cost savings and faster manufacturing times. For example, the one-chip-TV combines all its components in one chip. In the past, separate components were used. Already in 1965 Moore observed that 'semiconductor manufacturers had been doubling density of components per integrated circuit at regular intervals, and they would continue to do so as far as the eye could see.' [Schaller, 1997]. This remarkable statement, called Moore's Law after its inventor, has predicted past growth in IC complexity for more than 30 years with great accuracy. Component integration is the major reason for the enormous market growth of ICs. ICs are the basic elements of calculators, wrist watches, telecommunications, robotics, personal computers, and many other established business and personal applications.

The IC world market can be divided into two segments: volume IC production and ASIC (application specific IC) production. Memory ICs, that is, DRAM ICs, and microprocessors are examples of volume products. Their production is characterized by low product diversity, very high production volumes, and manufacturing to stock. ASICs are made for specific applications and most of the time even on customer order only. Their production is characterized by high product diversity and low production volumes per product type, therefore the type of fab is also called multi-process multi-product. This thesis is confined to the design of MP2 (multi-process multi-product) wafer fabs.

In the MP2 semiconductor industry the following trends can be identified: in the first place the variety of product types is increasing, because the number of IC applications are growing by the day. Furthermore, the relative production volume per product type will decrease. Customers require smaller batches of ICs, because the life cycle of their end product is decreasing. As a result of these trends, the MP2 wafer production

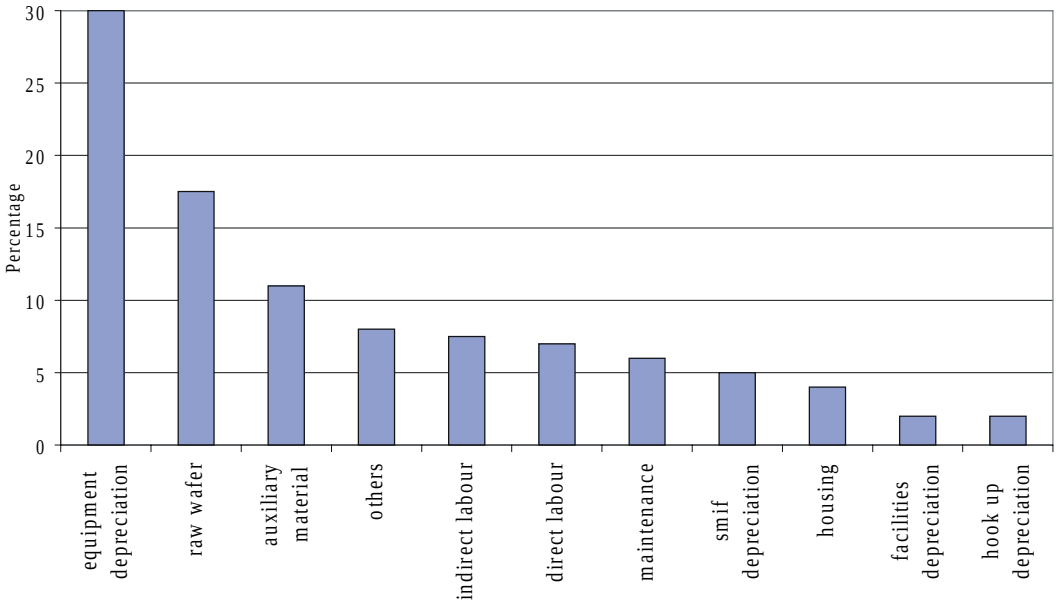

Figure 1.2: Breakdown of wafer production costs.

should be highly flexible and able to produce a large number of different ICs in one fab, whereas time to market should be minimal. For future MP2 wafer fabs this implies a strong focus on the design process of the production system.

Customers demand an ever-improving delivery performance – that is, more reliable delivery dates, shorter cycle times, and higher quality ICs. Short cycle times allow for fast feedback of process quality, and fast feedback allows for fast quality control. It has been shown that a decrease in cycle time results in an increase of product quality. As yield in IC industry is easily impacted, focus on quality is particularly important in order to be cost effective.

To be cost effective in the MP2 semiconductor industry, maximization of machine throughput is required, though this interferes with the goal of shortening cycle time. Maximizing machine throughput is desirable in order to increase return on assets by decreasing machine investment costs and depreciation costs. These costs form about 30% of the production cost price, as can be seen in the breakdown of production costs in Figure 1.2. The challenge in IC wafer fab design lies in the required flexibility and short cycle times, while simultaneously the machines have to be used in the most efficient way. Depending on the business decisions emphasis is either put on the throughput side or on the cycle time.

The IC market has grown enormously over the past decades and is expected to grow even more in the coming years – on average 15% a year. Figure 1.3 shows the expected IC demand for the coming years. None of the IC manufacturers is able to meet this demand. This is why most of the IC manufacturers are expanding their production capacity. To satisfy market needs, IC industry must invest 200 billion US-dollars in new fabs in the years 2000 to 2004.

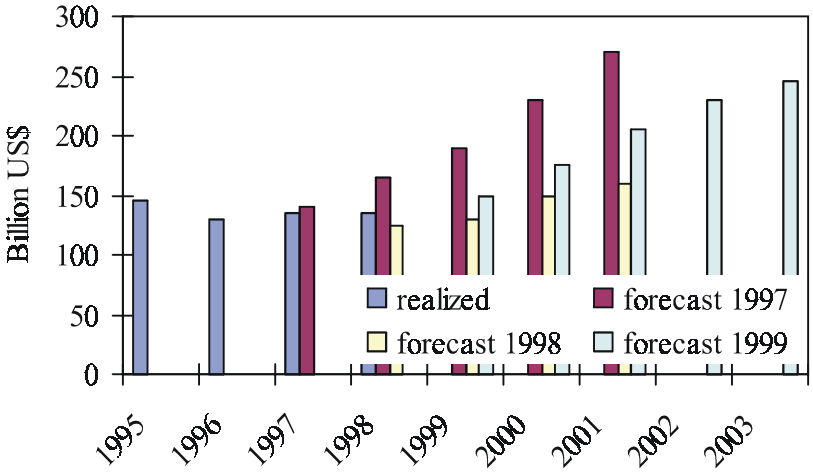

Figure 1.3: Illustration of difference between forecasted and realized IC market volume, [Dataquest, 1999].

Like many semiconductor manufacturers, Philips is expanding its production capacity. One of these expansions is the new wafer fab built on the Nijmegen plant. The design approach described in this thesis has been applied to the MOS4YOU wafer fab.

Fab design

The ultimate goal of manufacturing IC wafers is to generate earnings and return on investment, therefore, fab designers must create high productivity fabs. Productivity is expressed in good wafers out per invested dollar. The line yield of a wafer fab is measured by the number of good ICs produced per wafer. To be cost effective, a high line yield is required – in the order of 96% or more. Insights in operational behavior of the wafer fab can determine the success of that fab.

The basis for any fab design is the production process chosen. The complexity of wafer fabrication comes from the high-tech processes that are applied, the multitude

of process steps and their recurrence, and the comprehensive control required. Figure 1.4 depicts the recurrent behavior of the semiconductor manufacturing process. Subsequent layers are fabricated on the wafer surface, thus constituting the actual ICs. Common process technologies consist of 20 to 30 layers. Often, different layers are fabricated using the same set of machines.

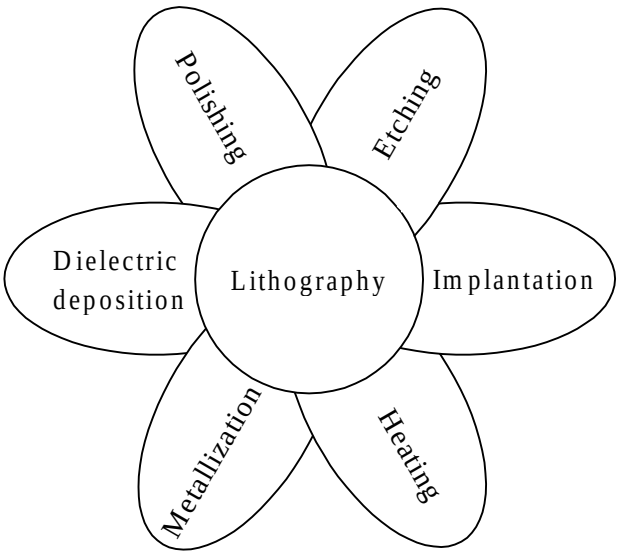

Figure 1.4: Recurrence behaviour of semiconductor manufacturing.

Both capital and operation costs of wafer fabs are increasing significantly. In 1980, building a new wafer fab required an investment of 100 million US-dollars. In 1998, roughly speaking, one new eight-inch wafer fab took an investment of over 1000 million US-dollars. This tremendous increase in cost can be explained by the ongoing advances that are being made in developing new technologies. As a consequence, machines that produce wafers are becoming more and more expensive. The main drive to overcome increased wafer costs is the reduction of the feature size of the transistors and wires in the IC. Smaller feature sizes enable more ICs to be made on one wafer. Furthermore, there is a drive for integrating IC functionality.

In spite of the huge investments required to build a wafer fab, the manufacturing costs for a single chip can range from only one to up to twenty dollars, which is considered to be low.

Initiation of a wafer fab starts with a top-level management decision that capacity is required and that in-house production is the way to achieve this. The business

plan identifies which types of products are to be made and in what quantities. The business plan also contains evaluation of alternatives and a rentability study. Before the project is started, it must be known if it is economically interesting. Outcomes of the business plan are the required resources (machines and people), target cycle times, and production specifications. The budget allocation for construction of the facility must then be determined in order to establish the boundary conditions that determine the profitability of the new fab. Once the business issues are resolved, a list of requirements can be reviewed. The facility requirements can be translated into various architectural and engineering solutions by designers.

From a project management point of view, the following phases can be identified in the constitution of a wafer fab: ground breaking, which ends with a weather proof building; clean room building; machine move in and qualification; start of the first silicon and first silicon out of the fab; full process qualification; and production ramp to full capacity. Advancing through these stages, the design of the wafer fab must become more specific. As can be seen from the description of these phases of a fab, the level of abstraction is lower as time advances. Before ground breaking can start, a structure of the fab needs to be available. Before machine installation starts, the layout of the fab needs to be clear. That is to say, a first version of that layout. As not all machines are placed at the same time in the fab, layout can and will change due to altered circumstances in the fab or in the fab's environment. When nearly all machines have been placed in the fab, optimization cycles are required for continuous improvement.

There is a trend towards automation of the wafer fabrication process. Ergonomic reasons form one of the major arguments for this trend. In mechanical sense this means that the wafer transport is automated. In logistical sense this means that the wafer flow control, that is, planning and scheduling, and equipment control are automated. Consequences of mechanization and automation are that a degree of freedom is taken away. Once specified and built, these systems provide few opportunities to adapt easily to changing requirements. The importance of fab design – illustrated by the diminishing flexibility – is opposed to the required flexibility as demanded by the market.

The above shows requirements for the design process. It illustrates the need for a design method and for design tools that support the design process during all phases. The design implications of many critical decisions made early in the design stages of a new semiconductor wafer fab can and do have a significant impact on both the initial and operating costs. The ability to influence the project costs decreases dramatically as the project gets underway.

Engineering consultants have prepared useful draft designs for fab layouts, based on process and machine characteristics and on their experience. Refining these drafts,

however, is always necessary, because process technologies will develop and new insights on how to run the fab will develop. Only the semiconductor manufacturer has the required in-depth knowledge, and therefore must be the generator of the organization and consequently the logistics of the fab.

Literature shows little references on the design of IC wafer fabs. Methods for layout determination are known since the 1960s, Apple [1963], Burbidge [1971], and Muther [1973] present general design methods. However, they are straightforward and are focussed on the goods flow. The control system is not considered in these design approaches. Recently, more sophisticated optimization methods for layout design have been explored. These methods still focus on the layout and the primary goods flow only. They take the conceptual architecture of the wafer fab as the starting point for their optimization, while it is this starting point that should be questioned.

The control of wafer fabs can be divided into planning and scheduling. The planning level concerns the interaction with the environment, the translation of customer orders to fab jobs, and capacity planning. The scheduling is primarily focussed on controlling the shop floor. The literature review on semiconductor planning by Uzsoy, Lee & Martin-Vega [1992] shows that it is the planning level that gets almost no attention in literature. The control of wafer fabs is extensively discussed in the terms of scheduling.

Objective

This thesis is concerned with the design of MP2 wafer fabs. MP2 wafer fabs exhibit specific characteristics urging the need for good design method: multiple process flows, often changing process mix, and several hundreds of different products. In short, high flexibility is required.

Although a lot of research is done on specific parts, like layout optimization or scheduling, no procedure is known that describes the design tools that are needed in the succeeding phases of the design process. Most known design approaches finish a complete design of a fab (at least the primary process of it) before the building of it starts. No attention is paid to the control and logistics part. There is no literature that describes a complete design approach of an MP2 wafer fab.

The objective of this thesis is to develop the design strategy with corresponding design tools to design an MP2 wafer fab. The proposed design strategy covers the complete design process, from the point where there is nothing but an idea to build a fab, via the first drafts, the empty skeletal, to the continuous improvement of a fully ramped fab. In each phase activities are performed with the use of design tools. The

tools that are needed to perform the design activities are designed and described in this thesis. Examples of tools are the investment model and discrete event dynamic simulation models. The design strategy and design tools are demonstrated using the MOS4YOU wafer fab as a case.

The thesis is restricted to MP2 semiconductor industry. However, the design concept could also be applied in other industries. Due to the specific characteristics, the design tools focus on semiconductor industry. The design approach assumes process technologies and machine characteristics as boundary conditions. Thus, the design subjects concern the wafer flow through the fab, both physically and from a control point of view.

The result of the work described in this thesis is twofold. First of all, the design tools are described. These tools can be used to make other designs. In the second place, the tools are applied to the design of an actual MP2 wafer fab, MOS4YOU. This leads to the executed design. For each activity, the design decisions, the rationale, and the design tools are described.

Structure

The proposed design method divides the designing into phases. These phases are: the description of the objectives and the constraints, the design of the architecture, the design of the capacity and layout, and the design of the operations. Each phase goes into more detail and builds upon the previous phase. The architecture of a fab describes the structure of the production process and is highly related to the control strategy and is a prerequisite for the layout. Thinking about architecture first, and not immediately in physical materials, provides a better basis for the final design than starting with the layout process right away. Operation design can only take place when the resources and their allocation have been determined. The structure of this thesis is depicted in an abstract way in Figure 1.5.

In Chapter 2 the processes that are needed to manufacture ICs are introduced. The design requirements following from these processes are also discussed. In Chapter 3 the design strategy is formulated. Four design activities can be identified, each consisting of several steps. The first activity comprises specifying the business plan, which identifies and characterizes the main manufacturing processes that are needed to produce the desired ICs, and formulating the objectives and constraints for the wafer fab design. The application of the first activity is also described in Chapter 3.

The second activity consists of designing the architecture of the manufacturing system and the production control system. In Chapter 4 the second activity is described

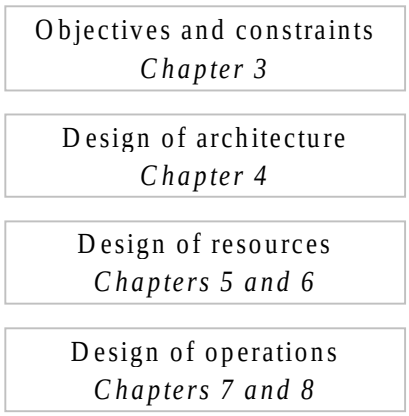

Figure 1.5: Structure of the thesis.

and the appropriate architecture for the IC wafer production is determined.

The third activity is the design of the resources and is concerned with assigning capacity to the manufacturing process steps and with determining the actual location of the machines. In Chapter 5 the capacity and layout are discussed. The design and operation of a billion dollar fab requires insight into the dynamic behavior of the equipment, the material flow, and the people operating it. In Chapter 6 this behavior is investigated using discrete event simulations.

The last activity, the design of the operations, designs the control concept to guide the wafers through the fab and improves the operational system. In Chapter 7 the designing of the overall control strategy for the production system is discussed. An example of a detailed improvement study is presented in Chapter 8, as the detailed design of the metal area is described.

The thesis is completed in Chapter 9 and Chapter 10, where conclusions and recommendations are presented respectively.

Chapter 2

Wafer fabrication

The complexity of processes, product routings, and equipment involved, make semiconductor wafer fabrication a very challenging manufacturing process. Extensive descriptions on IC wafer fabrication can be found in literature. A very thorough description of IC fabrication processes is given by Wolf & Tauber [1986], while an excellent introduction on IC manufacturing is described by Groover [1996]. This chapter starts with an overview of semiconductor manufacturing. The most comprehensive part of IC manufacturing – IC wafer fabrication – is dealt with in more detail. A discussion of the processes is a prerequisite to understand the manufacturing equipment that performs the various types of process steps. Both the separate processes as well as the integration of those processes are described. Finally, the challenges in designing IC wafer fabs and a set of design criteria are presented. The chapter concludes with the discussion.

IC history begins with the invention of the transistor [Groover, 1996]. Bell Telephone Laboratories were interested in developing electronic switching systems that were more reliable than electro-mechanical relays and vacuum tubes. As a result of their research, in 1947 Bell Labs produced the first transistor. In 1959 Fairchild Semiconductor Corporation described the planar process for fabricating transistors. Almost simultaneously, Texas Instruments presented multiple electronic devices and their interconnection on a single piece of semiconductor material. Both inventions formed the foundation of IC fabrication. The fact that electronic devices could be processed planarly, that is, on a flat surface, provided the opportunity for integration and miniaturization. Shortly afterwards, in 1960, Texas Instruments fabricated the first commercial ICs.

By definition, an IC (integrated circuit) contains millions of small electronic devices – such as transistors, resistors, and diodes – that are electrically interconnected. ICs

are fabricated on a thin piece of semiconductor material. Silicon is the most widely used semiconductor material, because of its good electric properties and low cost.

Since the fabrication of the first IC, there have been continuous efforts towards miniaturization and increased integration of multiple devices onto one chip. Both trends are mainly caused due to of economic reasons. Miniaturization is expressed in decreased feature size of devices. Figure 2.1a shows the trend of the characteristic feature size, also called Moore's law [Schaller, 1997]. Moore stated that, at equal costs, the characteristic feature size of chips will decrease by a factor two every two years. A reduction of characteristic feature size by a factor two results in an increased gain in dies on one wafer by a factor four (2^2). Figure 2.1b shows the trend for integration of electronic components onto one chip for microprocessors. The same trend can be shown for ICs made in the multi process - multi product wafer fab. The increasing density can be used to integrate the functionality of ICs. Increasing the functionality of ICs allows one chip to perform the functions that used to be performed by numerous single chips.

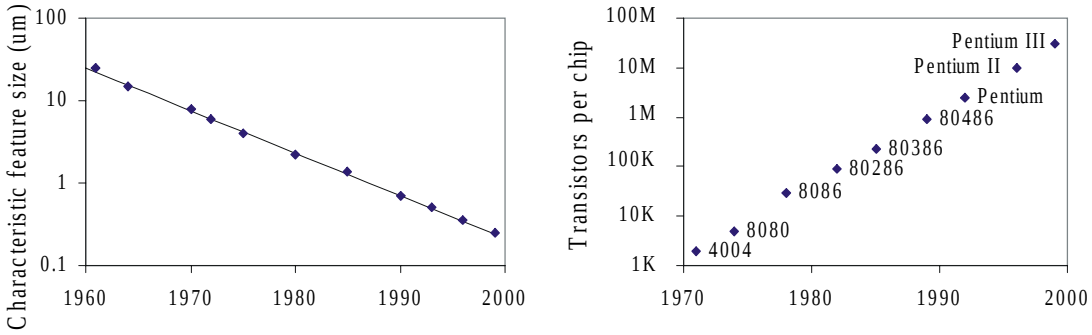

Figure 2.1: Development of a. feature size b. complexity.

2.1 Introduction

Millions of microscopic electronic devices are present on an IC. The individual components are formed by combining separate layers and isolated regions with different electrical properties. These components are electrically connected to obtain the electronic functionality of the IC. The final die – a 0.5 mm thin rectangular flat plate of 25 to 400 mm^2 – is attached to a lead frame and packaged as a component to be soldered into a printed circuit board and used for audio, video, and communication applications.

Basically, the processing of a silicon-based chip consists of three steps: raw silicon processing, IC wafer patterning, and IC packaging. In Figure 2.2 the transition of a wafer into an IC is schematically shown. Raw silicon processing deals with the production of pure silicon wafers. In the second stage ICs are defined onto these wafers. Finally, in the third stage, the individual chips are cut out of the wafer and each separate IC is assembled. The three stages are briefly described below.

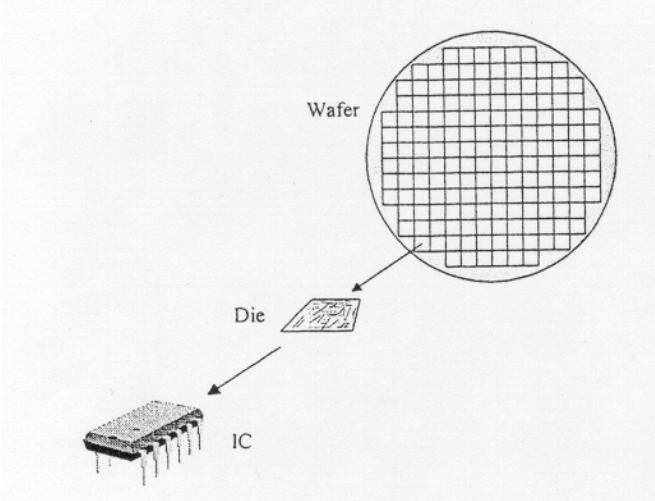

Figure 2.2: Transition from wafer to IC [Eijsvogels, 1998].

Raw silicon processing

Semiconductor material is the basis for the solid state electronic industry. Silicon is the most commonly used semiconductor material, 95% of all chips are produced on silicon. The reasons for this are the right properties and low cost of silicon (silicon is second in abundance only to oxygen). Fabrication of silicon-based chips starts with the processing of silicon of very high purity. Preparation of silicon substrate comprehends the following three steps: production of electronic grade silicon, crystal growth, and shaping silicon into slices.

The starting material is sand which is refined to form electronic grade silicon. Electronic grade silicon is polycrystalline silicon of ultra high purity, which means that the impurities range in the parts per billion, whereas sand has an impurity level that is eight orders of magnitude higher.

For flawless semiconducting material properties over a large area, the silicon substrate must be of single crystal structure. The Czochralski process [Wolf & Tauber, 1986]

is the most commonly used to obtain this structure. The process involves pulling a single crystal ingot upward from a pool of molten silicon, see Figure 2.3. The result of this process is a massive silicon bar of single crystal structure, the rod.

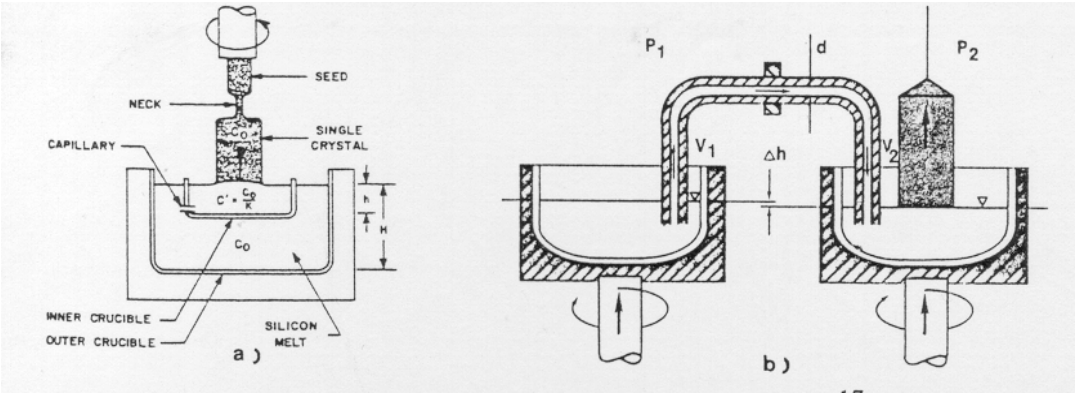

Figure 2.3: Formation of raw silicon using the Czochralski process [Wolf & Tauber, 1986].

Shaping the silicon rod into wafers starts with the preparation of the ingot. The rod is ground cylindrically. Then, wafers are sliced from the rod using a very thin saw blade with diamond grits. Finally, the wafers are prepared by rounding the rims of the wafer, removing surface damage by chemical etching, and polishing to provide the smooth surface needed to accept photo resist.

IC fabrication

IC wafer fabrication consists of several types of processes, all of them being repeated many times. Objective of these processes is to add, alter, or remove a layer of material in selected regions of the wafer surface. Processes to add layers are thin film deposition (physical vapor deposition and chemical vapor deposition) and thermal oxidation. Processes to alter layers are diffusion and ion implantation. Etching is the process that removes material. A procedure involving photo-lithography is applied to distinguish which regions will be affected in each processing step. In Section 2.2 the processes used for IC wafer fabrication are described in more detail.

IC packaging

Good chips, or dies, cut from a wafer are valuable. They break easily and atmospherical influences will corrode the interconnect system. Dies that pass the final inspection

are cut from the wafer and sealed in a package. Wire bonding is the process in which electrical connections between the contact pads on the chip surface and the package leads are made, using small-diameter wires of aluminium or gold. The sealed package can be a ceramic box or a plastic housing.

The package sealing is made of plastic, transfer molded around the assembled chip and lead frame. The packaged chip is finally tested to determine if the chip has been damaged during packaging and to measure performance characteristics of each device. Upon approval, the chip is shipped to the customer.

This thesis focuses on part of the wafer fabrication, starting with the raw wafer up to the pre-testing of the patterned wafer.

2.2 Processes

Making a chip on a wafer comprehends adding, altering, or removing regions on a silicon substrate. The regions constitute insulating, semi-conducting, or conducting areas that form the components and their interconnections. Regions are fabricated by a sequence of steps, each sequence forming another layer. Table 2.1 summarizes the commonly used IC wafer fabrication process steps and their description. Layers are fabricated one at a time, each requiring a separate photo lithographic mask. Common process flows can hold up to 25 layers.

Table 2.1: Description of common process steps for IC wafer fabrication.

Process	**Description**
photo lithography	transposes a pattern on the wafer
etching	removes material from the wafer surface
ion implantation	implants dope material into the wafer
oxidation	grows silicon dioxide layer on the wafer
chemical-vapor deposition	deposits dielectric or metal layers
sputtering	deposits dielectric or metal layers
chemical-mechanical polishing	smoothens the wafer surface

Using photo resist processes, photosensitive layers are sprayed on the wafer surface. Photo lithography processes transfer patterns from a mask to the photosensitive layers, via exposure to ultra-violet light, as depicted in Figure 2.4. Thus the required geometric pattern for each layer is obtained. The uncovered areas of the wafer surface can now be subjected to different kinds of treatments. The photo lithographic process

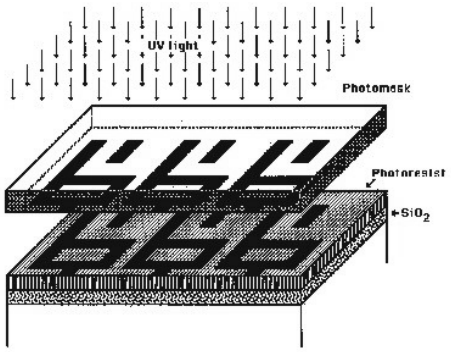

Figure 2.4: Basic concept of photo lithography [Maly, 1987].

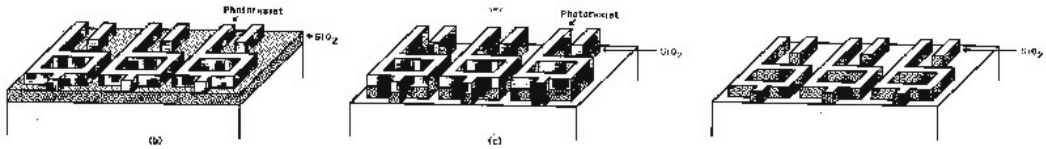

Figure 2.5: The process of etching and resist stripping [Maly, 1987].

usually is performed on a module, combining the applying, the exposing and the developing of the photosensitive layer. All reactors treat single wafers.

Etch processes are used to remove pre-selected parts from a wafer surface, whereas ion implantation processes insert dope material into the wafer. Figure 2.5 shows the wafer surface that is covered with resist, the etching of uncovered areas, and the removal of the resist respectively. By inserting dopants, for example p-n junctions are created. Before a new layer of photo resist can be applied using another mask, the remains of the previous layer have to be removed. This is done by the strip resist process.

Thermal oxidation is used to grow silicon dioxide on the wafer. SiO2 is used as insulator to separate devices and connection lines, and as protection for ion implantation processes. Oxidation is carried out in high temperature furnaces and requires a considerable time. Therefore, furnaces are equipped to treat numerous wafers at the same time.

Chemical-vapor deposition (CVD) processes are used to apply insulating or conducting layers on the wafer. Sputter processes are also used to deposit both insulating as well as conducting layers, see Figure 2.6. Both CVD and sputter processes are nowadays carried out in so called cluster tools, which contain several process reactors.

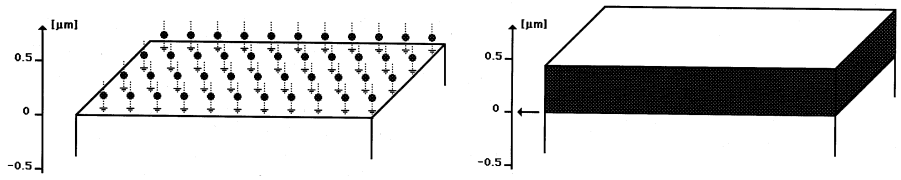

Figure 2.6: Thin film layer deposition [Maly, 1987].

Chemical-mechanical polishing processes are used to planarize layers of dielectrics or metal. Due to the partially mechanical character of the process, contaminants are produced. Therefore, these processes must be shielded from all other processes.

Inspection and defectivity analyses are performed frequently during the processing of ICs on wafers. Due to the sensitivity for failures and the high costs of producing wafers early failure detection is required. The ratio of inspection and defectivity processes versus actual forming processes ranges from 50% up to 100%.

2.3 Integration

In the previous section, the individual process steps have been briefly described. This section focuses on the sequence of steps that are performed to produce an IC. Each layer is patterned to form a structure that performs a specific function. The fabrication steps used to manufacture an IC must therefore be executed in a specific sequence, called a process flow. An IC wafer fabrication process flow is often divided into a front-end and back-end part. The front-end is concerned with building the components on the silicon, see for example Figure 2.7. The back-end deals with forming the interconnecting structures between the components. Except for the lithography processes, front-end and back-end use totally different operations. Lithography is required in both front-end and back-end of IC wafer fabrication. An example is used to schematically show a part of the process integration in IC fabrication.

The back-end of the process flow forms multiple interconnecting layers. These layers connect lower layers to another. Due to the enormous number of transistors and the complexity of the topology that forms the IC, multiple layers of interconnect are required. The process of building one interconnect layer is discussed below.

An interconnect layer is started by depositing a dielectric layer, which forms an insulator between two conductive layers. After this, a planarization takes place to

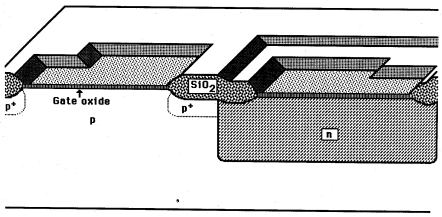

Figure 2.7: Formation process of n-well inverter [Maly, 1987].

smooth the surface of the wafer. Smoothening is required to enable deposition of more layers. It is often accomplished by CMP (chemical mechanical polishing).

Next, holes (called contact at the first level and vias at the other levels) are created in the dielectric layer, which are filled with metal to connect two successive layers. Holes are created using photo lithography and plasma etching. Holes are filled with tungsten, using a CVD (chemical vapor deposition) process. As tungsten does not adhere well to silicon dioxide, a titanium nitride layer is deposited first to enable adhesion. Titanium nitride is typically deposited using sputtering processes.

The excess tungsten, that was deposited to fill the holes, is removed and the surface is planarized using plasma etching. Now the complete wafer is deposited with metal, usually an aluminium alloy. The metal lines are formed using again a photo lithography and a plasma etching operation.

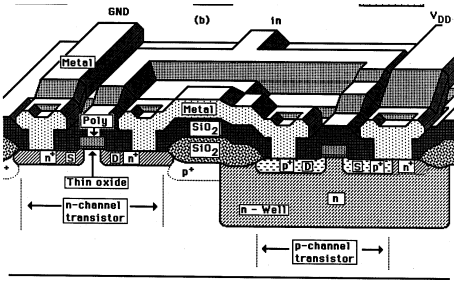

Figure 2.8: Cross section of CMOS inverter [Maly, 1987].

Even though the above described steps can easily be described at a macroscopic level, in fact they have an effect at the atomic level. A silicon-oxide layer, for example, has a thickness of $10\mu m$. These microscopic feature sizes require the production environment to be absolutely clean. There are two major sources of contamination, both in the order of 10^{-6} to 10^{-9} meter; the first source is generated in the process chambers during processing of the wafers. The second source originates from human

bacteria, viruses, hairs, and other particles. A clean room provides protection from contaminants. The air inside is constantly purified to remove (most) particles and people have to wear protective clothing to prevent (organic) particles from destroying production. Cleanliness is characterized by a standard classification system. For example, class 10 denotes a maximum of 10 particles of size 0.5 micro-meter in an area of 1 cubic feet. For human comfort and dictated by process conditions, the production environment is kept at a temperature of 21 degrees Celsius and a relative humidity of 45%.

2.4 Wafer fab design

The process technologies and manufacturing environment impose requirements on the design of a wafer fab. Therefore, their discussion is of importance for the further development of this thesis. Recent history has shown a development in clean room concepts: ballroom, tunnel, and mini-environment [Castrucci, 1995]. The ballroom concept assumes that the whole clean room has the same level of cleanliness. The tunnel concept divides the clean room in production areas and service areas. Production areas require a high level of cleanliness, whereas service areas may have a lower level. The tunnel concept has lower operational costs compared to the ballroom concept. The mini-environment concept is the latest development. In that environment the clean room has the cleanliness required at service level. Machines are placed in mini-environments, thus realizing the required cleanliness standards inside the machine. Wafers are transported in closed boxes, keeping them from being contaminated by particles. A mechanical interface is used to load and unload equipment, thus separating the human operator completely from the wafers and process chambers. Current process technologies require a cleanliness of class 1 or better, whereas a regular living room accommodates class 100000.

The dynamics of the market as well as the fast developing process technologies impose flexibility on the wafer fab. Although in the long term IC production is a growth market, short term demands vary heavily. Being able to deal with those variations is of vital importance for the wafer fab.

IC wafer fabrication processes contain relatively many operations. It is common for a 0.25 μm process flow to have more than 400 operations, which are performed on more than 75 different machine groups. This means that producing ICs on wafers requires several weeks. Furthermore, it is common to have several different process flows running in one fab. These flows to a great extent share the same machines. Both the numerous process steps and the multiple flows cause for great variability on the shop floor.

IC production flows exhibit a re-entrant character, the wafer returns to the same equipment numerous times, see Figure 2.9. The re-entrant character follows from the cyclic process structure, used to form ICs on wafers. As a consequence, lots compete for the same resource type, while being at different stages in the production process. This causes for great variations in both output and cycle time. Variations in production are also caused by machines not having the same batch size and by process times of successive operations that do not match.

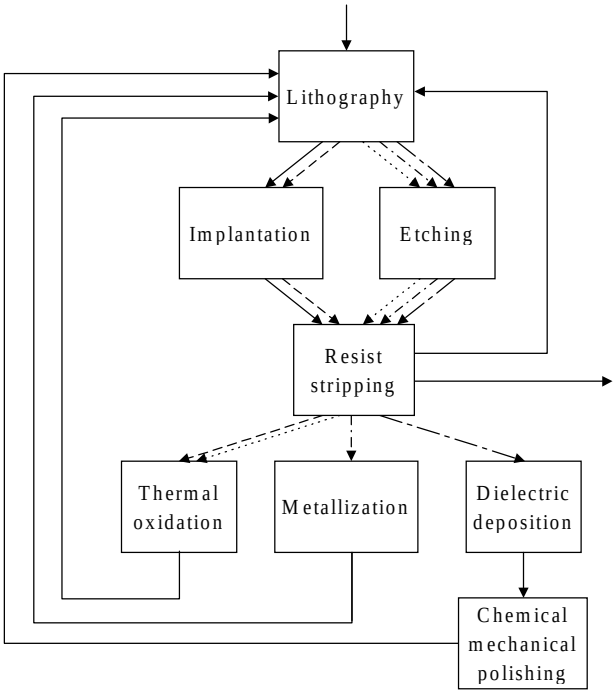

Figure 2.9: Re-entrant flow.

Building a new wafer fab brings about enormous investment costs. A major contribution to these cost are the huge investments required for semiconductor manufacturing equipment, the average price of one piece of equipment being two million dollars in 1998. This stresses the focus for high utilization to realize return on investment. Also the fab construction and clean room costs are extremely high. Therefore, compact factories are required to obtain an efficient use of space.

When starting the design of an IC wafer fab, design criteria have to be defined. A discussion on the priority of these design criteria is then needed to obtain the boundary conditions. Therefore, the mission statement of a company must be translated into specific manufacturing objectives, such as maximizing output and product quality, and minimizing cycle time and costs. Fab design criteria must cover aspects like

cost, capacity, quality, productivity, and flexibility. They serve the manufacturing objectives.

Costs can be divided into investments and operating costs. In the first place the investment in the building, the clean room, and installation should be kept as low as possible. Furthermore, operational costs should be kept low. Operational costs are to a certain extent already determined in the design phase.

The installed capacity must be used in the best possible way. This means that high output (or utilization) is needed to obtain return on investment. For example, if similar tools are placed near to each other, a better equipment utilization is obtained. Productivity of the installed capacity is expressed in output per installed equipment or output per square meter clean room. Also output per operator is often used, but operator costs are much less expensive than equipment, therefore equipment efficiency is more important than operator efficiency. Minimizing wafer transport distances contributes to improving both these efficiency parameters.

Process quality is expressed in yield. Major causes for bad yield are contamination and mis-processing. Contamination risk can be minimized by using a mini-environment construction. To track yield loss adequately, lots of inspections are needed.

2.5 Overview

In this chapter it was illustrated that wafer fabrication requires diverse and advanced process technologies. The interaction of these different processes and the way they are carried out impose a strong focus on the design of IC wafer fabrication facilities. The main characteristics and design criteria of IC wafer fabrication facilities are summarized below and are taken as starting point of the fab design.

- Growing market (20% per year),
- Market demand has cyclic character,
- Fast innovation of process technologies (one process introduction per year),
- Multiple process technologies in production (three different process generations and four process options per process generation),
- Multiple products in production (seven hundred products),
- Process flows experience re-entrant behavior (25 masklayers),

- Order based production,
- Customer requires small variance in cycle times (delivery performance above 99%),
- Return on investments require high productivity (utilization above 80%), and
- Mini-environments with isolated product carriers.

The strategy for designing IC wafer fabs is proposed in the next chapter.

Chapter 3

Design strategy

Designing production systems is a complicated and manifold task, therefore structuring the design process will lead to a better production system. In this chapter the design strategy used in the thesis is described. Basically the same strategy is also used by Fey [2000] to design a fruit juice blending and packaging plant. An IC wafer fab and a fruit juice factory are completely different and process characteristics hardly show any resemblance. A wafer fab is typically discrete, while a fruit juice plant combines both continuous and discrete characteristics, with the emphasis on the continuous side. Still, after close comparison, the design strategies used show great similarities. Section 3.1 and Section 3.3 have been written in co-operation.

Any industrial system is designed to serve specific objectives. With emerging new objectives or by changing old ones, new industrial systems arise and old ones disappear. The life cycle of an industrial system can be decomposed into five phases [Rooda, 2000], namely orientation, specification, realization, utilization, and elimination, see Figure 3.1. In the orientation phase the objectives of an industrial system are defined. After this phase, one is aware of the requirements to be fulfilled to make the industrial system successful. The functions that have to be performed to satisfy these requirements are defined in the specification phase. These functions are defined, together with the required resources. The definitions are presented in an abstract system or model. The orientation and specification phase are also called the design phase. In the subsequent realization phase the actual system is built and tested. The result is a functioning industrial system. The return on investments – made in the previous phases – must take place in the utilization phase. When the objectives change and the industrial system does not meet the original objectives anymore, it becomes obsolete. The last phase of the life-cycle concerns the elimination of the industrial system.

This thesis focuses on the design process. In this chapter the design strategy is

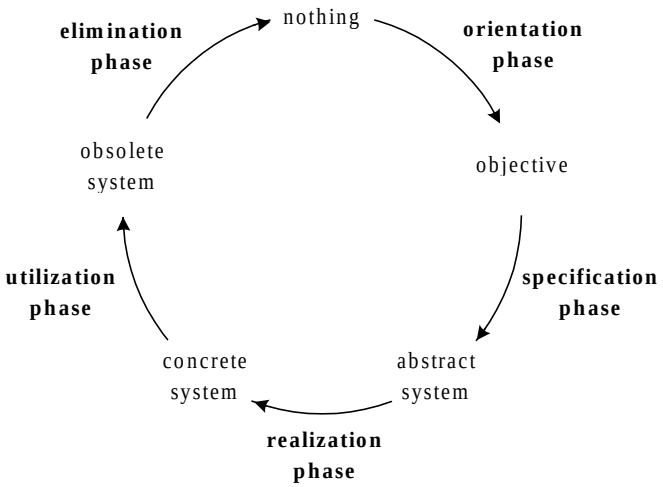

Figure 3.1: Life cycle of industrial systems.

outlined. First, a review of existing design methods is presented. The design strategy is deducted from these methods and described. The first step of the design strategy is to specify the objectives and constraints of the industrial system to be designed. Section 3.2 describes objectives and constraints for the MP2 wafer fab that was designed for this thesis. The last part of this chapter discusses the design activities that were followed to constitute the design. Section 3.4 concludes this chapter.

3.1 Structuring

Literature on designing industrial systems does not present one generally accepted methodology. It contains a large variety of detailed methods developed for a specific subset of industrial system designs. A large class of design methods focuses on one specific aspect of the relation between product and production system: Design for Assembly [Boothroyd & Dewhurst, 1983], Design for Quality [Nichols, 1992], and Design for Recycling are well known examples of these 'Design for X'-methods.

Another class approaches design problems from a production control perspective. The following methods can be named: load oriented manufacturing control [Wiendahl, 1995], group technology [Burbidge, 1971], socio-technics [Sitter, 1994], JIT (just in time) [Shingo, 1981], lean production [Womack, Jones & Roos, 1990], and OPT (optimized production technology) [Goldratt & Cox, 1984].

The business process re-engineering approach (BPR) [Hammer & Champy, 1993] is another approach that can be used to redesign an industrial system. BPR advocates to make revolutionary changes in existing systems by focusing on the core business of the system. It is not a process that preaches continuous improvement. Therefore, it is typically a method in which different design methods are used to improve an existing system, rather than to design a new one.

The review above shows that much knowledge has been formalized in numerous design methods. Brandts [1993] opposes that the use of these methods is hindered by the fact that the structure in how to use the methods is missing. A structured design method should point out what aspect to study at a certain moment and which design tools should then be deployed. Brandts proposes a five phase structuring of the industrial system design process: formalization of the objective definition, identification of basic sub-systems, phasing of the design processes of the various basic sub-systems, identification of the relevant attributes for every design phase, and selection or development of methods and techniques to support decision-making in the various design phases.

All design methods and structures mentioned above show a linear approach, such that the design process always advances the process of realizing the designed system. This omits the fact that optimizing and redesigning are essential tasks, not only for a new design but also for an operational production facility. While the design process advances, the realization process will be started. The level of detail in the design process increases as time proceeds. Design activities are not terminated the moment the specification phase is finished. On the contrary, the design process proceeds as long as the production facility is utilized.

The division of production systems in a manufacturing system and a manufacturing control system is often used to divide the design process in parts. Traditionally, both in research and in practice, production control has often been viewed in isolation from the manufacturing system design. Designing the control system starts as soon as the design of the manufacturing systems has been completed. As manufacturing system and control system show large interaction, it is a difficult task to first completely design the manufacturing system before starting the control system design. The relation between the design processes of the manufacturing system and the control system is so strong, that they should not be considered separately. Control issues have a direct impact on the manufacturing system design and vice versa. Johnston [1995] states that the effectiveness of control can be enhanced through re-engineering of the manufacturing system.

Proposed design strategy

The design strategy describes the sequence of activities that are carried out in the process of designing an industrial system. It provides a framework for intended activities. Within each activity design methods or design tools are used to obtain the desired result.

Development of a new design method is necessary for three reasons. First, literature provides a large amount of detailed design methods developed for specific problems rather than generally applicable methods for designing complete industrial systems. Second, literature disregards the fact that the majority of industrial systems evolve throughout their life cycle, and that optimization and redesign are essential tasks in the design process. Finally, although design of manufacturing systems and control systems show large interaction, literature treats them separately: designing the control system starts after the design of the manufacturing system has been completed. We propose a design strategy containing four activities, depicted in Figure 3.2.

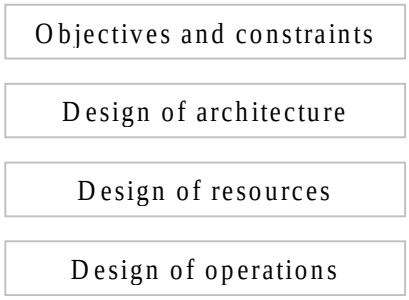

Figure 3.2: Design activities.

The strategy proposed consists of four activities: objective and constraints definition, design of architecture, design of resources, and design of operations. The activities follow each other in sequential order and the output of each activity is used as input for the next one. The first activity results in the objectives and constraints for the design. Using these results, the architecture of the manufacturing system and control system can be chosen. In the third activity it is determined how many resources are needed, how to locate them, and how to control the goods flow. The last activity takes the realized design as a starting point and focuses on improving performance by introducing for example working procedures or scheduling rules. The proposed design strategy incorporates the design of both the manufacturing system and the control system in all activities.

Designing an industrial system means designing a system that has to fulfil the objectives that have been imposed. Constraints define characteristics that are restrictive

to the design. Section 3.2 describes the objectives and constraints that concern the MOS4YOU MP2 wafer fab. In the remainder of this chapter the design activities are discussed.

3.2 Objectives and constraints

In general, the objectives for an industrial system are market dependent and deduced from business opportunities. In this thesis the design of a multi-process multi-product wafer fab is described and applied to Philips Semiconductors's wafer fab. Philips Semiconductors is a leading supplier of ICs and discrete semiconductors for application in consumer, telecommunication, multimedia, and automotive electronics. The division has a significant market position with chip sets for TV, audio, wired and wireless telephony, computer monitors desktop video, and PC peripherals, and moreover the division is world leader in one-chip-TV products.

As the IC market is growing with an average of 15% a year and because Philips plans to stay in the top ten of semiconductor manufacturers, wafer manufacturing capacity expansions need to be considered from time to time. The latest capacity expansion is MOS4YOU, located in Nijmegen in The Netherlands. MOS4YOU started production of ICs, on eight-inch wafers, in February 1996. In the name MOS4YOU, MOS stands for metal oxide semiconductor, indicating the type of ICs being produced; 4 stands for the fourth semiconductor fab which produces ICs of the MOS type; YOU stands for yield, output, and utilization, maximization of these three criteria is central to the fab's management philosophy.

The goal of an integrated circuit wafer fabrication facility is to make profit by producing integrated circuits that function. As machine investment costs form the major expenses, profitability is boosted by obtaining high equipment utilizations. Increasing the number of ICs requires a high line yield and a low defect density. By minimizing the cycle time, the time in which wafers might be affected by contamination is minimized. Furthermore, low cycle times result in fast learning cycles, providing the means to improve yield performance. Obtaining high machine utilization and low cycle times are two conflicting objectives. Therefore, finding the optimum between those two objectives is the real challenge for the designer of a wafer fab. For cycle times a benchmark value of two days per mask layer is common.

Dedicated products and low volume characterize the MP2 fabrication environment. Furthermore fast turn around times are required. An MP2 wafer fabrication facility typically runs a larger number of processes and process variations than a standard production facility. Also, many different products are being processed at the

same time for direct delivery. In some cases IC designs for several customers will be processed on the same wafer.

In an MP2 wafer fabrication facility, the larger part of the total wafer operations will typically be devoted to in-process and end-of-process monitors to ensure that each individual piece of process equipment and the overall processes are functioning properly in terms of parametric distributions and defect densities. First, this is needed because the customer's product wafers may have design flaws, which may prevent them from providing any useful information to the fabrication engineers. Second, the process engineers must convince the customer that the quality of fabrication meets all applicable acceptance criteria.

In Chapter 2 it was already stated that process technology advances rapidly. The reduction of feature sizes, the increase of component densities, and the increase of operation speeds illustrate this. A typical wafer fab is designed for two generations of ICs and will in practice be used for three generations. This indicates that not only products will change (e.g. new developments in telephones and multi media applications), but also processes. In many cases the second generation of processes will already be introduced before the fab is fully built. So, a complicating factor on the design process is that a fab design must comprehend flexibility for the future.

A wafer fab cannot exist without the help of a number of suppliers. First of all machine suppliers form an important group. Machine depreciation accounts for over 40% of the wafer cost price. Often, service contracts are agreed upon. Furthermore, mask making is outsourced. With the given specification, the mask is produced and then sent to the wafer fab. Mask-making lead times can be critical for wafer production. Finally, raw wafers are obtained from a raw wafer producer. Quality and delivery reliability are of importance.

Environmental care issues are taken into consideration by using processes that have a minimal effect on the quality of the environment. For example new production lines require fewer rinsing stages. More and more furnaces are cooled by air or closed-loop systems. Measures such as these are cutting water consumption by up to 45%, even though manufacturing output rises.

3.3 Activities

Architecture

Design of architecture is the second activity in the design process. As starting point, the dictionary is used to clarify the term architecture. Webster's collegiate dictionary

defines *architecture* as: 'having or conceived as having a single unified overall design, form, or structure'. The term denotes striving for coherent structures, something that the design process must achieve. In this thesis architecture is defined as the manner in which the components of a specific system are organized and integrated.

It is important that a top-down approach is applied: the architecture of an industrial system has large influence on the specification of system components and their layout. Furthermore, a well designed architecture supports an industrial system in realizing manufacturing objectives. The need to excel at objectives such as high flexibility, short lead times, high throughput, and high effectivity has to be reflected in the systems architecture. For example, in a machine shop a large range of customer specific products are manufactured. Flexibility is often obtained by a process focussed architecture, which enables customer-specific routing of products.

Design of architecture is divided into two aspect systems: manufacturing structure and control architecture. The first and most abstract step is concerned with the structure of manufacturing or, in other words, how products are made. Technological developments play an important role at this step in the design process, the result of this step is a function model. The second step is concerned with the architecture of both the material flow as well as the control system. The result of this step is a process model. To illustrate the results of these two steps, simplified function and process models for the production of ICs are explained below.

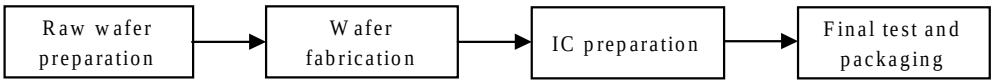

Figure 3.3: Function model.

Figure 3.3 depicts a function model for the production of ICs. As described in Chapter 2, IC production starts with raw wafer preparation. Subsequently, ICs are manufactured on these wafers, after which the ICs are prepared. Finally, wafers are tested and ICs are cut from the wafers and packaged. Each of these functions can be represented by one or more processes. Figure 3.4 depicts a process model for a part of the production of ICs on wafers. Wafers await processing in stocker S or in buffer B. To form a layer on the wafer, a sequence of processes is required, identified as $M1$, $M2$, $M3$, and $M4$. After the last operation, the wafer fabrication continues with the next layer.

The design decisions for the second activity are the functions that are used to make the products and the structure of the processes. The next step in the design process is to determine the necessary resources to perform the required functions, and satisfy the business objectives and constraints.

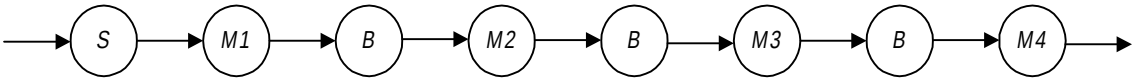

Figure 3.4: Process model.

Resources

Design of resources is the third activity in the design process of an industrial system. Resources are operators as well as machines. In this thesis, design of resources focuses on quantifying the architecture designed in the previous design activity. This is necessary to concretize the design and obtain a good approximation of the necessary equipment and factory layout.

Whereas the business objectives have a large influence on the design of the architecture, the architectural design, in its turn, has a large impact on the number of required resources and the factory layout. For example, a process oriented architecture in general may require less resources than a product oriented architecture. This is because the former has its resources grouped according to functional operation, while the latter is structured and laid out towards the product flow, which may therefore lead to several machines in different places that are not fully utilized.

The design of resources results in a list of resources and a factory layout. This result is accomplished in two successive steps: the first step is concerned with determining the resources, the second step is concerned with laying out these resources on the shop floor. Often the available information which is needed as input for the determination of resources is based on expectations and inaccurate. For example, sales volume can only be estimated, and the product range is adapted rapidly to changing customer demands. Therefore, the required resources may change over time. Layout has to be flexible to cope with these changes.

A quantitative analysis of the system to be designed is made to come to a set of resources, e.g. machines or operators. The main tool for handling the capacity analysis problem is modelling. A model is an abstraction of reality, which stresses some aspects of reality while ignoring irrelevant details. For example, consider a capacity analysis model which has the purpose of determining the necessary number of drilling machines in a machine shop. Relevant information which should be incorporated in the model is the capacity of one drilling machine, the number of available working hours and the total work load for the drilling section. Irrelevant aspects are for instance the size of the machines and its setup procedures. The art of modelling is to leave out as much information as possible without harming the relevance of the model. An important modelling activity when facing large and complex industrial

systems is decomposition. The industrial system is decomposed into several smaller parts which are modelled and evaluated independently. A model consists of some sort of formal specification, e.g. a set of mathematical equations, or a discrete event specification. The capacity analysis problem is solved by solving the set of mathematical equations or executing simulation experiments and analyzing the results.

The design decisions of the third activity are to find out what resources are needed, how many are needed, how they will be located, and how the goods flow that goes over these resources will be organized. Design of resources is critical to the success of a plant, as it affects important performance measures such as lead time, equipment utilization, and system throughput. Once the layout has been determined, attention must be paid to how the control strategy is translated into operations. This is done in the design of operations.

Operations

Design of operations is the last activity of the design strategy. Webster defines operations as: 'the way in which something works'. Design of operations is concerned with the design and optimization of the (detailed) way in which a fab works. The understanding of a plant's operative performance and how it is influenced by strategic and operative decisions, provides companies with a major advantage. For example, minimizing lead times, working with high equipment utilizations, and being able to adapt to changing product types are crucial to ensure profitability of the plant.

In the design of resources only a rough capacity analysis has been performed, due to the absence of accurate and detailed information on the operation of the manufacturing facility. During the process of designing a manufacturing facility, more and more detailed and accurate information becomes available. This enables zooming in on the operation of the facility, and retrieving a more accurate image of its behavior. The objective of the design of operations phase is twofold: validating the resource design, and optimizing the manufacturing operations. The first objective deals with checking whether the resources calculated in the previous design phase provide enough capacity to operate the factory successfully. Compared with the design of resources phase, more detailed information is taken into account. Furthermore, the dynamic interaction between processes is considered. The second objective is concerned with designing rules which support efficient use of manufacturing resources and designing tools with which these rules can be evaluated before implementing them. The production facility or part of the production facility is modelled, analyzed, and optimized.

The main objective of design of operations is to refine the design in such a way that

the operations of the plant can be performed in an efficient and effective way. The design decisions of the last activity are working procedures and scheduling rules. Through the evaluation of models, a set of resources and rules can be determined. Rules describe the way in which a fab is operated. Two examples are: first, where to store material on the shop floor when it is not in process and second, a set of scheduling rules. Determining scheduling rules and working procedures is here considered as optimizing. Implementing these rules leads to a better performance of the industrial system.

The process of specifying models, optimizing, and implementing rules does not take place just once, it is an iterative process that needs to be repeated often during the design stages and also when utilizing the industrial system. Figure 3.5 shows the circle of optimization [Rooda, 2000]. The start of the cycle is the existing system or the system to be designed. By making a specification, knowledge and characteristics of the system are obtained and design decisions are made clear. The formal specification can be used as a model, for example as a discrete event simulation model, with which different scenario's are evaluated. In the optimization phase, the optimal set of rules is determined. Eventually, these rules are implemented, leading to a better performing system.

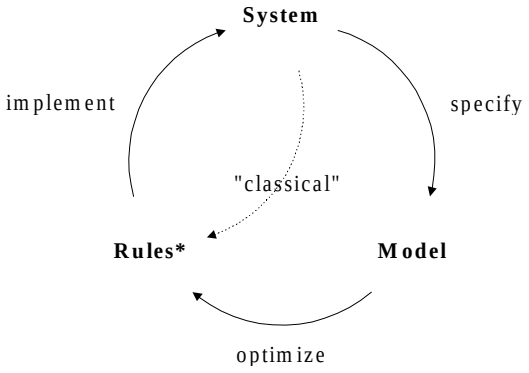

Figure 3.5: Circle of optimization.

The importance of optimization is illustrated by the following calculation. At a wafer price of 2500 US-dollar and a production level of 5000 wafers out per week, a turnover of 75.000 US-dollar is generated per hour. Improvement of the factory performance yields a significant profit.

3.4 Overview

In this chapter a strategy for designing a semiconductor wafer fab is proposed. Design literature provides a large amount of detailed design methods developed for specific design problems rather than generally applicable methods for designing complete industrial systems. The proposed design strategy provides a structure that can be used throughout the specification and utilization phase of an industrial system. It incorporates initial design as well as optimization of a utilized production facility. It consists of four activities: definition of objectives and constraints, design of architecture, design of resources, and design of operations. Every design starts with identifying design objectives and constraints. In the design of architecture, the structure of the material flow system and the control system is designed. The design of resources is concerned with determining the required resources and constituting a floor plan. In the design of operations the optimization of the industrial system is considered. The industrial system is modelled, analyzed, and optimized. Optimization leads to a valid capacity plan, and a set of rules which lead to a better production system when implemented. Design of operations is performed during the specification and utilization phase of the industrial system, and is therefore performed at different levels of detail. The lowest level of detail considers the whole fab, the highest level considers only parts of the fab.

The design process starts with identifying design objectives and constraints. The design objectives are deduced from the business objectives of a company. During the design process, these objectives and constraints remain subject of discussion. The business objective of the MOS4YOU wafer fab is to manufacture integrated circuits on wafers for the consumer industry. The multi-process multi-product wafer production can be characterized by:

- Fast technology development (one new process technology per year),
- Relatively short process life cycles (three years per process technology), and
- Different operations (a 0.35 μm process flow requires 400 operations).

The products can be characterized by:

- Short life cycles (six months per product type),
- Varying yields, and
- Relatively long cycle times (60 to 90 days).

The financial objective is to generate return on assets. As a result, the main design objectives for MOS4YOU are:

- High yield ($> 97\%$),
- High output (originally 20.000 wafers per month in a 0.5 μm process),
- High utilization ($> 80\%$),
- Short cycle times (< 2 days per mask layer), and
- High delivery performance ($> 99\%$).

In the remaining chapters of the thesis the design activities are described.

Chapter 4

Factory architecture

The description of the fab characteristics, in Chapter 2, illustrates the complexity of the fab and lists the parameters to be considered in the design. Structuring the design parameters results in a complete view of the system to be designed. An industrial system is a collection of products and a production system. The production system, in its turn, can be decomposed into four sub systems: the manufacturing system, the control system, the economical system, and the organizational system. The manufacturing system involves the transformation of materials into products. The control system involves the flow of information and contains the strategies that influence the behavior of the production system in the desired way. Activities performed in this subsystem are indicated as production planning and control. Production planning focuses on strategical and tactical decisions, such as demand management, aggregated - and detailed capacity planning. Production control monitors and influences the real-time flow of material through the fab. In this thesis, production control is subject of study. Other parts of the control system are slightly discussed. The economical system involves the flow of compensating materials and the organizational system involves the people in the production system. The architecture of manufacturing system, the control system, and the transport system will be distinguished in this chapter.

4.1 Manufacturing system

This section describes the classification of the manufacturing system of a multi-process multi-product wafer fabrication facility. First the characteristics are discussed and then the architecture is designed.

Multi-process multi-product wafer fabs produce a wide range of different products. Typically, at any moment in time over 200 different products are in process simultaneously. Therefore, individual lot performance is of key importance. Figure 4.1a shows a typical Pareto analysis of the product range of an MP2 wafer fab. It can be seen that there are some high volume runners and that the number of small volume products is enormous, 30% of the product range accounts for 80% of the fab load. However, it must be noted that the actual product range changes fast, as the life cycle of one product type is about two years.

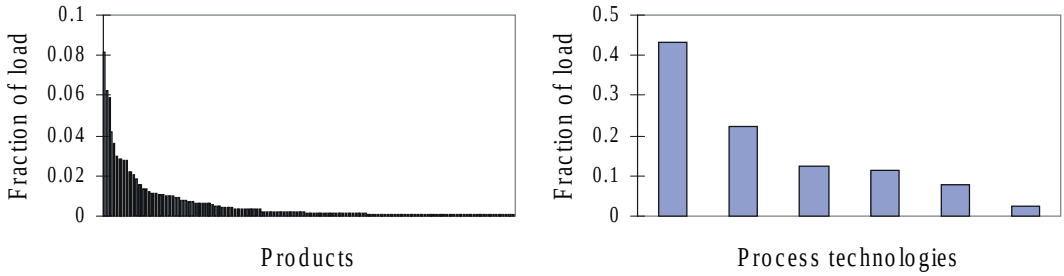

Figure 4.1: Pareto analysis of a. products and b. process technologies.

MOS4YOU uses six different process technologies to make the products. An example of such a process technology is the 0.35 μm CMOS process. A typical Pareto analysis of process technologies is presented in Figure 4.1b. The figure shows that three process technologies account for 80% of the fab load. The typical life cycle of a process technology is approximately three years, after that time the process will have become outdated and will be unattractive to design new products in. Existing products, however, will still be produced using the old process technologies.

Wafer fabs, in general, are designed to manufacture two to four generations of process technologies. After that a complete renewal of the fab is needed. The difference between two succeeding process technologies can be enormous. A new technology may require whole new operations for parts of the production process. New equipment is needed to satisfy the more advanced process characteristics. Building a new fab often provides better means for improving performance than the complete renewal of an existing fab would.

Within a process technology there are often several options. The differences between these options usually are usually quite small; the process routing will be the same to a large extent. The differences usually consist of two to six more mask layers, providing extra functionality to the integrated circuit, like a non-volatile option.

The number of process steps in the flow is high (e.g. 400 steps is very common for a 0.35 μm process) and a number of these steps take place on machines of the

same type. For example: a lot requires a lithography operation 25 times during its processing life cycle. Each time a different mask is used to define the next layer on the silicon.

Process yields are uncertain and vary due to environmental conditions or problems with machines. Also, the introduction of new processes will result in yield loss for well established products. Characteristics of machines used for producing integrated circuits vary widely. There are single wafer machines and there are batch machines that can process a number of lots simultaneously in one run. There can be time critical sequences between two succeeding steps. If a lot is not processed within a certain time interval on the second step, significant yield loss will occur.

There are about 75 different machine types necessary to produce silicon wafers. Machines require considerable preventive maintenance and calibration, and even then a lot of unforeseen down events occur. Machines that consume 20% of time for non-productive events are not exceptional. It is estimated that the main cause for performance disturbances in a wafer fabrication facility is due to unpredictable machine down time [Uzsoy, Lee & Martin-Vega, 1992].

The constant drive for new technologies asks for continuous development of new processes. Qualification of new processes is a time consuming activity. It takes about 6 months before a qualification traject is passed successfully. Often the same equipment is used for both production and qualification lots. Furthermore, development lots require a lot of engineering time relative to qualified standard lots. There may be conflicting goals between the production and engineering organization.

Machines in semiconductor industry are to a large extent automated; transport of lots and setup of machines are not, these are performed by operators. Maintenance and repair is performed by technicians. As labor costs are relatively small, about ten percent of overall costs, human capacity in the design is considered as not be a bottleneck factor. Not being automated completely provides an extra degree of freedom. With this extra degree, sources for variability are introduced due to unexpected absence, judgement errors, not fully efficient working procedures, or other factors.

Concluding, MP2 wafer fab processes are characterized by a fast technology development, relatively short life cycles, and multiple different operations that occur many times during the flow. The products are characterized by short life cycles, varying yields, and relatively long cycle times, that is, ten to fourteen weeks.

Architecture

The manufacturing system of industrial systems concerns the material flow. In that sense, it comprehends the transformation and transportation of material and its sole purpose is to add value. The structure of the primary system depends upon the products that are being made and the processes that are used to make them.

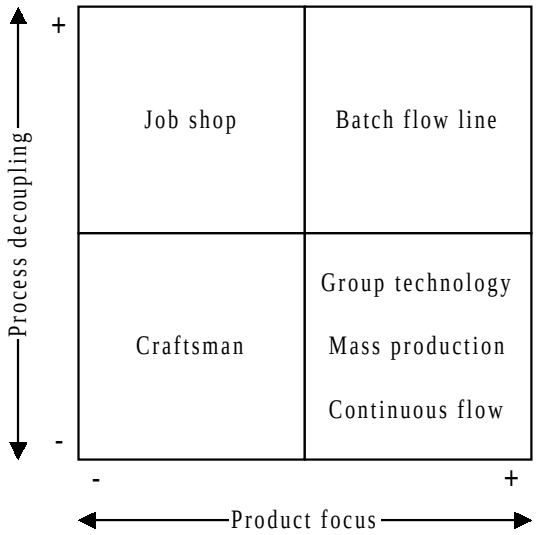

Figure 4.2: Classic structures [Browne, Harhen & Shivnan, 1996].

Browne, Harhen & Shivnan [1996] categorized production systems into four categories using by process- and product structure. Process structure is defined as the manner in which material moves through the plant. The degree of decoupling determines to what extent the production processes are divided into separate operations. Product focus expresses to what extent the production system is made unique for specific products. Figure 4.2 shows the corresponding structures. The four categories are discussed below.

The first category of process structures is the job shop. Job shops are process oriented production systems, that is, the architecture is focused on the effective use of machines. A job shop contains a variety of machines of different types. Similar machines are grouped together in so called workshops. Figure 4.3 shows four different types of work stations, $W1$ to $W4$, served via transporter T. A product enters the production system at transporter T. As long as the routing is not completed, the product stays in the system. After each operation, transporter T moves the product to the next workstation, where the product waits until it can be processed. A job shop structure is preferred when highly customized products are manufactured in

low volumes. This structure can handle a high variety of product routings; each custom order may involve a different routing and can revisit machines several times. Job shops provide maximum flexibility with respect to product changes.

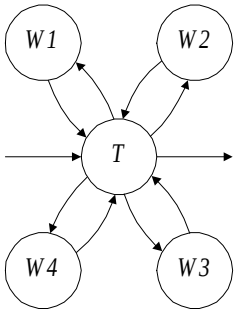

Figure 4.3: Process oriented architecture.

The second category is the batch flow line, that produces batches of products with a limited number of identifiable routings. By definition flow lines are product oriented, that is, the structure is focused on effective throughput. The product characteristics determine to what extent the production machines are connected. Products are manufactured in relatively low volume production. Individual machines are not connected, therefore, there will be a high amount of WIP (work in progress) between machines.

The third category is the mass production flow line. The automotive assembly line is a classic example of this architecture. A few major products are manufactured in high volumes. Diversions in product routings are small. Workstations are working at the same pace and connected by means of a material handling system, see Figure 4.4 for a schematic view. The figure shows four workstations, $W1$ to $W4$. A product enters the system at workstation $W1$. All workstations are visited subsequently to perform the needed operations on the product. After the last operation, the product leaves the system. A group technology structure can be seen as consisting of one or more flow lines that are dedicated to a group of similar products.

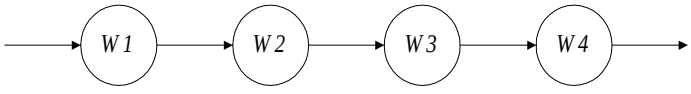

Figure 4.4: Product oriented architecture.

The final category is the craftsman, who manufactures a whole product all by himself. All operations are performed by one person and the number of different products can be large.

Some of the afore mentioned characteristics of an MP2 wafer fab fit the product oriented layout, that is, a flow shop structure. These characteristics are the low number of process technologies and the large similarities between these technologies. Furthermore, high output and short cycle times can best be realized with this architecture.

On the other hand a process oriented architecture would be appropriate if based upon the following product and process characteristics. Numerous products in the line, fast developing process technologies, highly specialized operations, and fluctuating machine reliability. Furthermore, a product oriented structure, where more than 400 machines are placed in line would not be cost effective. First of all, the throughput rates of the different equipment types differ significantly. Identical equipment would be required at several places in the flow, and there is a chance this will not be fully utilized. Second, the high intensive maintenance and relatively poor machine reliability would cause disturbances in the whole production line, as the equipment at that stage cannot be used for production.

A distinguishing characteristic of the wafer fabrication process is that lots re-visit machine types at several stages during their production process. This cyclical process is called re-entrant flow. The main consequence of a re-entrant nature is that wafers, at different stages of their production cycle, have to compete with each other for the same machines. The re-entrant architecture differs from the process oriented architecture, as the layout is based upon several main processing technologies. Each main technology is grouped and the auxiliary operations are divided over the fab and performed where needed.

Concluding, the wafer fab is designed as a number of re-entrant flow lines. Main technologies, like lithography, dry etch, implant, oxidation, and metallization, are each grouped by their own technology. Auxiliary technologies, like resist stripping or inspections are placed at all those places where they are needed. That way, a process oriented architecture with mini flow lines is obtained.

4.2 Control system

The objective of IC manufacturers is to maximize profitability by minimizing the production costs and by increasing the productivity. High volume standard product manufacturers have a policy to produce for stock. Doing so provides protection against fluctuations in customer demand in the wafer fab and furthermore it puts less strain on cycle time performance of the wafer fab because the number of products is limited. These characteristics and the enormous capital invested in a wafer fab have resulted in high attention for the cost effectiveness of throughput and machine

utilization, while on the other hand cycle times and inventory are (relatively) of less importance. For the high volume standard product manufacturers these objectives are the only ones, while MP2 IC manufacturers require more objectives.

In the MP2 IC market, where small lots of custom ICs are designed and produced on a low volume basis, each product is manufactured on customer order. Therefore, it is hard for the MP2 type of factory to produce for stock. Still, MP2 manufacturers need to have low cycle times and a high delivery performance. Reducing cycle times is important for several reasons. First of all, reducing the time to market is important for the customer. Time to market is defined as the time necessary to design and debug a new product. With product life times decreasing, a fast development of new products becomes of more and more importance. Second, improving responsiveness to customers is important. The shorter the cycle times, the shorter the feedback loops to customers, thereby enabling fast product debugging performance. Third, according to Little's law, a short cycle time results in a low WIP, which means less inventory costs. The above shows that cycle time reduction serves both external as well as internal objectives.

Concluding, the production control of wafer fabs should primarily be focussed on maximizing output and delivery performance under the constraint of achieving acceptable cycle times at minimum costs. In the remainder of this section, first the conventional architectures for production planning and control are presented. From that, the framework that is applicable for semiconductor wafer fabs is deducted. Discussions on push versus pull production, and scheduling versus sequencing are preliminary for the discussion of the chosen architecture.

Conventional architectures

Production planning and shop floor control are the processes used in adjusting the capacity to the demand. Production planning is defined as the planning on the medium term (six months to two years) of the manufacturing system's resources in order to fulfil (forecasted) demands. Short term planning (weeks to six months) considers orders, forecasts, and resource capacity to optimize the product mix and capacity planning. The planning process results in the capacity planning, the allocation of orders, and the due dates that go with each order.

Shop floor control is defined as the control on the short term (minutes to days) of the production activities on the manufacturing floor. Scheduling is considered to be a major part of the shop floor control activity. Scheduling determines which set of orders will be processed on which resources during a short-term period (typically a day or a week). The scheduling process is influenced by the planning process as

well as the operations. A non-realistic output of the planning process will result in a bad scheduling process, even if the design is perfect. Hereby, the importance of considering both planning as well as scheduling in one framework is demonstrated. The scheduling process provides the job assignment for the operations. On the other hand, operations influence the scheduling process.

Now an overview of existing production planning and control methods is presented as a starting point to describe the architecture that is applicable for semiconductor industry. As discussed in Section 4.1, conventional architectures can be divided into two extremes: the product oriented and the process oriented architecture. Product oriented manufacturing systems typically require CONWIP (constant work in progress) [Hopp & Spearman, 2000]. CONWIP is a manufacturing strategy with a very simple goal, that is, produce the required items, with the required quality, and in the required quantities, at the precise time they are wanted. The production planning concept of CONWIP focuses at balancing capacity against (forecasted) demand. It shows a large similarity to the MRP (materials requirements planning) production planning concept that is discussed further on in this section.

A system that executes CONWIP on the shop floor, the production control system, maintains a constant WIP. The production is triggered by demand, that is, the status and WIP at down stream machines triggers the production of up stream machines. The CONWIP principle is depicted in Figure 4.5. Material flows from left to right and requests for material flow from right to left. Workstation $W4$ is the last in line. When given authorization, it produces a part. The part that has just been used is replaced by sending an authorization signal to workstation $W1$. $W1$ is the first workstation in line and therefore it controls the release of work into the shop floor. To start an operation on a workstation, an operator needs material to produce and authorization to start producing. By thoroughly controlling the WIP levels in the system, a CONWIP control system assures controlled and stable throughput and cycle times.

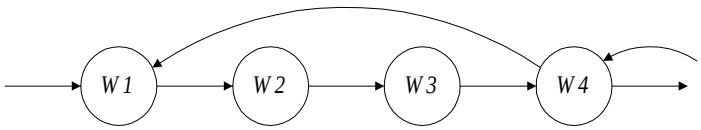

Figure 4.5: CONWIP principle.

Process oriented manufacturing systems are typically controlled using MRP (material requirements planning). MRP systems start with the master production schedule as input and apply a set of procedures to generate a schedule of net requirements for each component needed [Browne, Harhen & Shivnan, 1996]. Using predefined lead

times, MRP then calculates when to start manufacturing components to obtain a finished product at the right time.

Framework

In semiconductor industry, studies of production planning and scheduling have been considered apart from each other. The vast amount of shop-floor control research has not considered the interface between shop floor control (scheduling) and higher-level production planning decisions [Uzsoy, Lee & Martin-Vega, 1992]. In this thesis, a hierarchical framework is used, which contains planning as well as shop floor control (scheduling). The framework is inspired by [Pinedo, 1995] and [Browne, Harhen & Shivnan, 1996] and depicted in Figure 4.6.

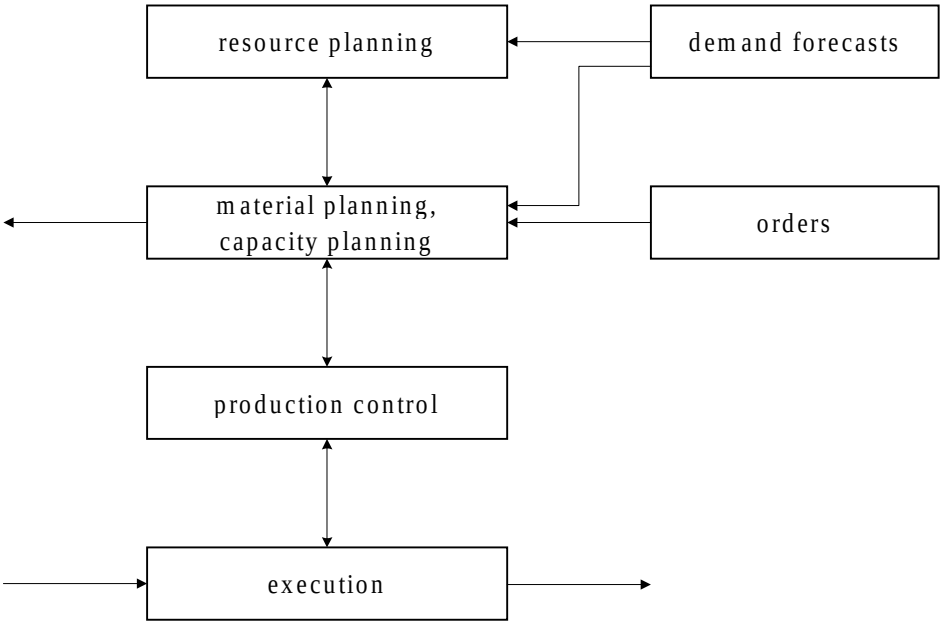

Figure 4.6: Production planning and control framework.

The relationship between capacity planning and production control is of major importance. The resource planning offers the objectives for the shop floor control and the behavior of the shop floor influences the capacity planning. Interfacing between resource planning and shop floor control takes place through the material and capacity planning. The result is presented by a certain product capability and machine capacity, and by the due date assignments to confirmed lots. Therefore, in linking production planning and shop floor control, a detailed machine capacity model is a

key factor. With this model machine capacity is geared to the (forecasted) demand. After making a capacity plan, due date assignment routines are required to give the customer feedback. On the basis of the installed capacity and the (forecasted) demand, order lead times are confirmed or orders are declined.

In the manufacturing system, accepted orders must be translated into jobs with routings and due dates. Jobs are produced on machines according to their routings. A detailed scheduling of the operations to be performed on the shop floor is necessary to ensure optimal performance of the system. This scheduling function can be affected by status changes from the shop floor. These will, amongst others, include machine breakdowns and operator absence. Wiendahl [1995] shows that work load control can be used by releasing work to the shop floor in a controlled manner. By doing so, the mean time and variance of time spent on the shop floor is decreased. Once lots are on the shop floor, priority rules are most commonly used for shop floor control policy. These rules are used to decide what lot is the next to be processed when a machine becomes idle. The popularity of dispatching rules in practice is due to their low computational requirements, their intuitive nature, and the ease of interfacing [Pinedo, 1995]. Due to their simplicity, the results of dispatching rules are not the best practice.

Push or pull

The choice between a product orientation or a process orientation is determined upon several grounds. A discussion that often occurs is the one between a push control system and a pull control system. In this thesis, the following definitions are used for push and pull. A push control system schedules the release of jobs based on demand. A pull control system authorizes the release of jobs based on system status [Hopp & Spearman, 2000]. Production control using CONWIP is considered to be a pull system, while MRP can be described in terms of a push system. Both systems are driven by a master production schedule, indicating the quantity and timing of products to be made. The difference is that for a push system jobs are released according to an external schedule, while for a pull system a schedule is prepared in advance, but an authorization depends on the status of the fab.

Estimating capacity is rather difficult. Because of this, a push system directly accommodates customer due dates, but has to be forced to respond to changes in the fab (for example, MRP must be regenerated). Similarly, a pull system directly responds to changes on the production floor, but must be forced to accommodate customer due dates (for example, by matching the production plan against demand and using overtime to ensure that the production rate is maintained).

When comparing the push with the pull control system, the most important characteristic is that a push system controls the output whereas a pull system controls the WIP level. Setting the release rate in a push system must be done with respect to capacity. If the rate is set too high, the system is choked with WIP. If too low, revenue will be lost due to insufficient throughput. A pull control system achieves less variation in cycle time than a push system, since cycle time increases with increasing WIP, and pull prevents WIP explosions, therefore pull also prevents cycle time explosions. A predictable cycle time is important in keeping a high delivery reliability. Therefore, a low cycle time variation is required. The underlying cause of the benefits of a pull system is that there is a limit on the maximum amount of inventory in the system. For a given level of throughput, a push system will have more WIP on average than an equivalent pull system [Hopp & Spearman, 2000]. Therefore, using Little's Law, for a given level of throughput a push system will have longer average cycle times than an equivalent pull system.

Scheduling or sequencing

A schedule is defined as the allocation of jobs to machines over time [Pinedo, 1995]. As a schedule provides the working plan for the next period, it requires a stable and predictable manufacturing environment. Unfortunately, the prerequisite of a stable and predictable environment is rarely true in a loaded wafer fab. The objective of scheduling is to maximize the use of machines so that the overall production objectives are met. To determine an optimal schedule a set of rules is required. The type of scheduling techniques that are used, depends upon the overall production objectives, the volume of orders in the system, the nature of the machines, and the complexity of the processes.

In practice, semiconductor wafer fabs are often corrupted by the influence of variability. Variability can be caused by the following factors: many jobs competing simultaneously for machines, machine breakdowns occurring at unexpected moments, operators being absent for some time, and many other factors. The influence of variability causes a certain optimal schedule to become readily far from optimal. The more variability there is in a system, the more often a renewed optimal schedule needs to be determined to deal with the rapidly changing environment.

Scheduling provides a basis for assigning jobs to work centres over time. Less strictly, sequencing only specifies the order in which jobs should be processed at each machine. Sequencing methods are referred to as priority rules. Dispatching of lots on machines is preferably done by choosing the lot with the highest priority from the sequence lists. Priority rules provide guidelines for the sequence in which the jobs should be

processed. They are used to optimize certain criteria, that are derived from the overall production objectives.

There are multiple reasons why sequencing is better suited to control a wafer fabrication facility than repeated calculation of optimal schedules. First of all, the calculation of an optimal schedule is done with the current situation and based upon expectations for the near future. In practice, the knowledge of the future is imperfect. For example, more new jobs can enter the system, or machines may break down unexpectedly. Therefore, often provisions have to be made to be prepared for the unexpected. Due to the changing environment, re-sequencing is necessary in practice. The existing schedule was constructed based upon certain assumptions. Unexpected events require either major or minor changes in the existing schedule.

Another complicating factor is that processing restrictions and constraints may vary over time. New processes are introduced, causing disturbances on lots that run in existing processes. Furthermore, in practice it often occurs that a job is scheduled on a given machine, but that for some reason there is a preference to schedule it on another machine. Not sticking to the calculated schedule will have a devastating impact and will turn the fab into chaos. Finally, in theoretical models that compute optimal schedules the job priorities are assumed to be fixed, i.e. they do not change over time. In practice the job priority may fluctuate over time, a low priority job may suddenly become a high priority job if seemingly, it tends to arrive late at the customer.

Concluding, it can be said that in the ideal situation an optimal schedule can be computed to obtain the best solution for the job allocation problem on individual machines. The more variability there is present in a system, the worse the realization of the optimal schedule will be, and the faster the correctness of the schedule over time will degrade. The more randomness there is in a system, the less advisable it is to employ very sophisticated optimization techniques [Pinedo, 1995]. In a situation with a lot of variability, calculation of an optimal schedule will not provide a good solution. A dispatch selection strategy (sequencing) will provide a better and more robust means. It is quite possible that, in some cases, calculating an optimal schedule results in a worse solution than simply using priority rules.

Architecture

Key part of the control architecture design is to divide the product routing into segments. These segments coincide with the miniflows within an area. A segment typically consists of two to six operations. Each segment has a main operation, which not necessarily needs to be the first operation in the segment. Now, like in

the CONWIP system, optimal WIP levels for each segment can be determined. A good WIP balance is achieved when all segments operate on their predetermined WIP levels. The control policy aims at doing this by prioritizing segments. For example, machine Y becomes available for processing a lot. There are three segments in which an operation has to be performed by machine Y. By looking at the WIP levels of these segments, the preceding segment and the next, it is determined which segment is going to be prioritized. Additionally, it is possible that there are more lots in this segment. Lot selection within the segment is performed on basis of the due dates. The lot with the least slack in reaching its due date will come first in line.

The architecture divides the wafer fab into main technologies. Auxiliary technologies are placed where needed. The control strategy must aim at maximizing the WIP balance in the fab. Maximizing the WIP balance, that is, focusing on a good distribution of work over all the machines in the fab, will result in high machine productivity and short cycle times. To assure individual lot delivery performance, additional steering on lot progress is required.

4.3 Transport system

The success of a wafer fab depends to a large extent upon the availability and performance of sophisticated and expensive process equipment. However, effective handling and storage of wafers is also crucial for world class performance, as proper wafer handling will reduce cycle time and increase yield.

Wafer handling is of importance because of ergonomical reasons. To enable cost effective manufacturing, wafer diameters are increasing. For example, an increase in wafer size diameter of 50% leads to an increased number of ICs per wafer of more than 100%, at constant chip size. As diameters increase, the weights will increase. A typical wafer carrier, containing twenty-five 200 mm wafers, weighs 9 kilogram. This drives semiconductor manufacturers to redesign jobs, and to rethink wafer handling and storage concepts to improve productivity and safety of the fab.

Basically, transport can be performed in two ways: manually or automated. Automated material handling and storage (AMHS) systems have been studied on many occasions. A detailed analysis of the needs and requirements of AMHS systems in wafer fabs is presented in [Pillai, 1990]. The article provides guidelines and design rules for designing AMHS systems. However, it is rather qualitative of nature. A more quantitative approach is found in [Pierce & Stafford, 1994]. Pierce & Stafford [1994] presents models and simulation experiments used to analyze and specify AMHS systems for wafer fabrication. Although the study delivers some good insights in the dynamic behavior of AMHS systems, the drawback is that it treats the AMHS system

as an isolated system, thereby neglecting the influence of variability originating in the rest of the wafer fab.

The architecture of a wafer fab design is heavily influenced by the degree of automation of the AMHS system. Up till recently, material handling was performed completely manually. In the past years a trend towards partial automation has emerged, the so called inter bay automation, where transport between work cells is automated but transport within a cell is performed manually. In the future, fabs are likely to run with fully automated material handling systems.

AMHS systems can be compared to operators on a number of characteristics. Here the major ones are mentioned. Automated transport improves equipment utilization, as it provides stable and reliable transport times. Tracking and tracing possibilities reduce retrieval times significantly, which will influence productivity positively too. AMHS systems improve yield in two ways. First, they reduce yield loss due to contamination as AMHS systems are designed for cleanliness and provide clean wafer transport. Second, it has been proven that AMHS systems cause less failures than manual transport does. An example to illustrate the effect: a scrapped lot with 25 200 mm wafers represents a selling price of 60.000 US-dollar. Besides lot scrap, automation of transport will have a positive effect on cost effectiveness because of the better use of floor space (transport can be performed using less floor space and lots can be stored in high stockers where a robot can retrieve them, thereby increasing the storage density) and because of the better controllability. On the other hand operators are much cheaper and much more flexible.

There is a tendency in literature to first design the factory and place all the process equipment on the floor, making material handling optimization a hard job, because there are a lot of fixed boundary conditions and therefore less choices to be influenced. As an example, a layout often is optimized by minimizing transport intensity (two work cells with high transport frequency are placed adjacent to each other). Taking transport and storage into account in the beginning of the design process, provides a better final design. Design choices regarding transport systems concern the transport intensity, the transport distance, the way of transport, whether or not buffers are applied, and the location of the buffers.

Concluding, it was chosen to equip the wafer fab with an automated inter bay transport and storage system. The inter bay transport forms roughly 60% of the total lot transport, so automation will relief operators significantly. The intra bay transport must be flexible, therefore manual transport is used to move lots.

4.4 Overview

In the design of architecture phase, the structure of the manufacturing system and the control system of MOS4YOU has been designed. The boundary conditions for this phase were reflected in the objectives and constraints. The re-entrant character of the process flows can best be captured in a hybrid functional architecture. A hybrid functional architecture groups main process technologies by their own technology and places auxiliary technologies at those locations where they are needed. The choice of combining a process focussed architecture with local flow lines allows for a proper control of the production system. The production control architecture is designed to be robust for randomness and variability. A balance strategy that balances work-in-progress fits this architecture best. Additionally, the architecture of the production system is influenced by the transport and storage architecture. To be able to cope with the many different sources of variability, the transport and storage architecture needs to be flexible. It was chosen to have a partially automated partially manual transport concept.

The main design choices for the second phase are:

- Re-entrant flow line architecture,
- Main process technologies grouped by their own technology,
- Auxiliary process technologies distributed where needed,
- Balancing of work-in-progress using sequencing rules,
- Automated transport between areas, and
- Manual transport within each area.

In the following chapter the resources and their location in the fab are determined.

Chapter 5

Equipment capacity

In the previous chapter, the architecture for both manufacturing system and control system was determined. In this chapter, a method and its application for calculating the number of required resources is presented. The following resources are considered: process machines, transport and storage machines, and people as they operate and maintain these machines. It is discussed how to obtain a layout for these resources. Design of resources is concerned with estimating how many machines are needed to process a certain product demand, and how many wafers a fab can produce with the installed machines.

The importance of resource design is illustrated by the high machine and operating costs, and by the fact that a well balanced capacity is a pre-requisite for good fab performance. Unbalanced machine capacity will lead to cost inefficiency. Furthermore, the ever changing product demand and processes require constant insight in and adaptation of capacity plans. Process machine capacity is usually determined in a rough way, using spreadsheet models [Witte, 1996]. Transport and storage capacity is often calculated using dynamic simulations, as these systems are too complex to be analyzed using static calculations. On operator capacity, almost no literature is available. Operator capacity is calculated in practice by using historical productivity numbers.

A fully ramped wafer fab (5000 wafer-starts per week) contains over 500 machines, that are divided in about 75 different machine groups. Typically, the manufacturing of a wafer takes more than 400 process steps. Due to the large number of process steps, the cycle time of a wafer exceeds eight weeks. The lots arriving at the different machines have to wait in queues before they get processed on the machines. Thus there are numerous queues at the various machines. During manufacturing, a wafer visits a specific machine group more than once. This re-entrant behavior causes wafers to compete at different stages of their manufacturing process for the same

machine group. Therefore, there are interactions between various machine queues. The arrival patterns of the lots depend on various stochastic parameters, one of which is the product demand. Furthermore, machines are subjected to foreseen and unforeseen failures. Together with these disturbances, the large cycle time and the re-entrant process flows may cause imbalance of work in progress, which in its turn will have a negative impact on machine capacity. Furthermore, an altering product mix may cause new bottleneck machines.

A capacity planning is needed to match the machine capacity and the product demand. IC industry is characterized by rapidly changing product demands. Furthermore, the process technology evolves very fast, causing new technologies to be introduced each year. Consequently, it is not unusual to have multiple process technologies in one fab. The wafer fab faces two kinds of planning challenges. First of all, an investment plan for the required machine capacity needs to be determined, using the expected product mix as guideline. This is called machine investment planning. Second, the most profitable product mix that can be manufactured needs to be determined, given the installed machines. This is called product mix planning. This step is needed, because the investment plan may not necessarily be carried out as scheduled. Product demands might have changed, causing the need for new capacity calculations.

Both machine investment planning and product mix planning are of importance for a multi-process multi-product wafer fab. The determination of capacity is a fundamental issue in these processes. Two types of methods for capacity determination are commonly used: static analytical models and discrete event dynamic simulations. Static analytical models are mostly based on queueing theory. queueing models are extensively described by Buzacott & Shanthikumar [1993] and Snowden & Ammons [1988]. Queueing models can be decoupled or linked. A decoupled queueing model considers machine groups as separate capacity sources. There is no relation with the actual process flow and the arrival pattern is specified stochastically. The second type of queueing models links the machine groups together. The buffers in front of the machine groups form the queues where the lots wait for processing. After being processed, a lot is directed to the next machine group according to the process flow. This is repeated until all operations have been completed. Static analytical models are widely used due to the ease of modelling and their relatively fast response times. However, as Govil & Fu [1999] notes, exact analytical models that incorporate capacity and blocking of resource are not very well developed. Several studies have focussed on improving the accuracy of static models. A good example of static models of manufacturing systems is described by Witte [1996]. Discrete event dynamic simulations deliver accurate results, provided that the model is detailed enough. But it usually takes considerable time to make a model and also to obtain results with the model. For wafer fab design, static models provide a fair level of accuracy. To

perform capacity analysis in a wafer fab, decoupled queueing models are preferred rather than dynamic simulation models.

Robinson, Fowler & Neacy [2000] describes a number of factors that contribute to capacity loss. Based upon that overview and upon own experiences, the main elements that need to be incorporated in static capacity modeling are machine availability and utilization, process yield, operation batch sizes, sharing of capacity between machines, and cycle time offset. Whenever machines are used, cycle time offset is used to deal with the position in the process flow. Because the lot cycle times are much larger than the single processing times, it is necessary to take this into account.

In recent literature, four descriptions of static capacity models, applied in semiconductor industry, can be found. Witte [1996] incorporates availability, utilization, and yield. Hsih & Wu [1998] and Wu, Lang & Liao [1998] incorporate cycle time offset and capacity sharing. In the latter approach, process constraints regarding dedication and backup are maintained a priori in a database. Work loads are manually assigned or shifted during the planning process. In general tool dedication and backups complicate capacity estimations. Neudorff [1999] focuses on batch efficiency. In the next sections a static machine calculation method and its application is presented. The method uses a decoupled queueing model, incorporating machine availability and utilization, operation batch sizes, sharing of capacity between machines, and cycle time offset.

5.1 Method

In this section a static calculation method is presented that can be used to determine machine capacity.

A machine group is defined as a set of identical machines. It is assumed that machines within the same group are able to perform exactly the same operations and are physically located in the same area. Machines that perform the same operations and that are not physically placed in the same area, are treated as two different machine groups. If required, they can be used as each other's backup. An example is the visual inspection machines; inspection machines are generally placed uniformly distributed over the shop floor. These machines are able to function as each other's backup capacity.

Figure 5.1 depicts the components of the capacity calculation method. These components are formed by a database, containing business-, process-, and machine related data, and by an engine consisting of investment module I and allocation module A.

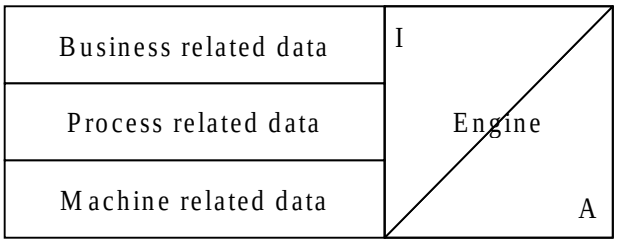

Figure 5.1: Structure of capacity planning tool.

Business related data

Business related data consist of product demand and cycle time objectives. Demand is denoted by d and is typically expressed in wafers out per month or wafer starts per week. The different process technologies are indicated by using index k. Doing this, the demand for process technology k is denoted by d_k. In the investment cycle module I demand is represented by actual orders and forecasted demand. The product cycle time, or lead-time, is an important performance measure. Delivery reliability is one of its derivations. Cycle time is denoted by φ. The cycle time factor, defined as the ratio of cycle time φ and theoretical process time φ_0, is denoted by Φ, so $\Phi = \varphi/\varphi_0$.

Illustration. In this chapter an illustration is used to demonstrate the capacity calculation method. There are two different process technologies. Demand is given by $d_1 = 1000$ wafers per week and $d_2 = 1500$ wafers per week, and customers require that $\Phi < 2.5$.

Process related data

The time that is required to manufacture a certain demand is calculated using process related data. These data consist of process flows, operation times, and operation batch sizes. A process flow consists of L operations and each consecutive operation is referred to with the index l. L_m contains a subset of operations, namely only those operations that are performed by machine group m. The theoretical process time of operation l is denoted by $\varphi_{0,l}$. The batch size of operation l is denoted by b_l. Current practice in industry is to use historical data of average batch sizes in the capacity calculations. You, Weng, Chou, Wu & Lu [1999] showed that approximation formulas, which are based upon traffic intensity, down time, and maximum batch size, provide a good estimation of average batch sizes. This provides good means to forecast batch sizes rather than using experience data.

The time required to manufacture d lots on machine group m can be calculated by:

$$T_m = d \sum_{l \in L_m} \frac{\varphi_{0,l}}{b_l}. \tag{5.1}$$

Similar, the number of required operations performed by machine group m can be determined by:

$$M_m = d \frac{length(L_m)}{b_l}. \tag{5.2}$$

For a situation with multiple process flows, the required time to manufacture a complete mix of products on machine group m is determined by:

$$T_m = \sum_k d_k \sum_{l \in L_{k,m}} \frac{\varphi_{0,l}}{b_l}, \tag{5.3}$$

where $L_{k,m}$ denotes the subset of operations from process flow k that has to be performed by machine group m. Again, the number of required operations can be determined by

$$M_m = \sum_k d_k \frac{length(L_{k,m})}{b_l}. \tag{5.4}$$

Illustration. For the illustration one machine family is considered, machine family 1. The first process technology requires 3 operations on machine family 1 and the second process technology requires 5 operations on this machine family. The accompanying processing times and batch sizes are presented in Table 5.1.

The required time to manufacture the demanded amount of wafers on machine family 1 is $T_1 = 1000 \cdot (90/25 + 60/25 + 105/25) + 1500 \cdot (80/25 + 60/25 + 90/25 + 105 + 75/25) = 34800$ minutes per week or 580 hours per week. The number of required operations is $M_1 = 1000 \cdot 3/25 + 1500 \cdot 5/25 = 420$ operations per week.

Table 5.1: Process times and batch sizes.

Operation	Process 1		Process 2	
	φ_0	b	φ_0	b
1	90	25	80	25
2	60	25	60	25
3	105	25	90	25
4	-	-	105	25
5	-	-	75	25

Machine related data

The machine capability and availability are captured in machine related data. The capability of machines influences the capacity, it indicates what recipes can be processed on which machines. A recipe capability matrix can be used to depict the possibilities and restrictions. Dedications, that is, recipes that can only be processed on a single machine, result in low and variable capacity. Besides the capability the availability of machines also impacts capacity. Machines in semiconductor industry are very sensitive towards disturbances. Hence, a lot of unexpected down time occurs, in spite of intensive preventive maintenance actions. The various machine stages that can occur have been standardized by Sematech [Dhudshia, 1997]. Figure 5.2 depicts a breakdown of the stages that can occur. T_t denotes the total time: 365 in a year. T_n denotes the non-scheduled time, for example the 25^{th} of December. T_o denotes the operations time, usually rounded up to 168 hours per week. T_d denotes the down time of a machine. It can be seen that a machine is not available for production when it is down, planned or unscheduled. The availability of a machine is defined by $A = T_a/T_t$, where T_a denotes the available time for production and T_t denotes the total time.

Illustration. For machine family 1 it holds that $T_n = 0$ and $T_d = 33$ hours per week. The resulting availability is then $A = 135/168 = 80.4\%$.

The available time, in its turn, can be decomposed into $T_a = T_p + T_i + T_e$, where T_p denotes the actual productive time, T_i the idle time or standby time, and T_e the engineering time. Utilization is defined as the quotient of productive time and the available time $u = T_p/T_a$. In the same example as above, a machine that was in use by engineers for 8 hours and idle for 10 hours a week, has a utilization of 86.6%

Cycle time is largely influenced by machine utilization. Other factors that influence cycle time are the number of identical machines m within a machine group and the coefficient of variation of both the arrivals and the processing of lots, c_a^2 and

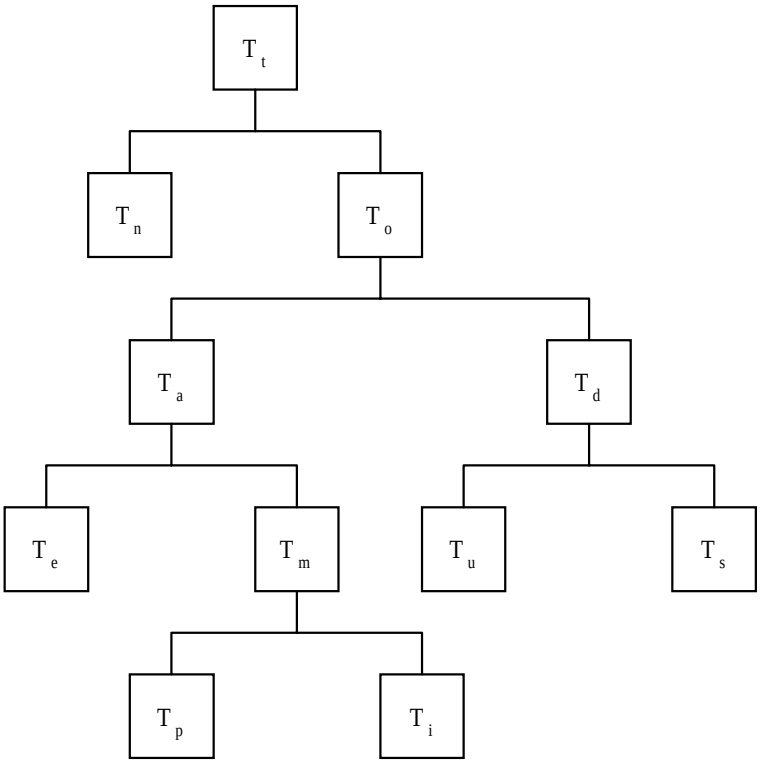

Figure 5.2: Summary of SEMI E10 states.

Table 5.2: Maximum allowed utilization and resulting available production time.

Number of machines	**1**	**2**	**3**	**4**	**5**
u_{max}	0.52	0.72	0.80	0.84	0.87
T_p	71	194	323	455	588

c_0^2 respectively. Hopp & Spearman [2000] have shown that for a situation with m identical machines, a general arrival pattern, and a general processing pattern, the relation for queueing time can be described by:

$$\varphi_q = \frac{u^{\sqrt{2(m+1)}-1}}{1-u} \frac{c_a^2 + c_0^2}{2m} \varphi_0. \tag{5.5}$$

Here φ_0 indicates the average theoretical processing time of the operations that are performed by the machine group. This formula has a product form of three terms: a traffic intensity factor, a variability factor, and a time-scale factor. The first factor indicates that at increasing utilization the queueing time will increase non-linearly. The second factor indicates that waiting time is linearly related to the variability both in the arrival times and the processing time: the larger the variation, the larger the queueing time. For a typical manufacturing situation $c_a^2 = 1.5$ and $c_0^2 = 1.25$ are common values. Finally, the waiting time depends upon the magnitude of processing time φ_0.

The available productive time for a machine group can be determined by $T_p = u_{\max} \cdot A \cdot T_t$. The above shows that the amount of available time for a machine group depends upon a number of parameters. These parameters are all machine related, except for the cycle time target which is a business parameter.

Illustration. In Table 5.2 the maximum allowed utilization $u_{\max}$ for machine family 1 is presented, depending upon the number of identical machines. The table also presents the accompanying available production time T_p.

Calculating investments

For any forecasted demand, the number of required resources can be determined by comparing required time with the productive time. For all machine groups m the number of machines n_m within the group can be determined. n_m equals the lowest

integer for which it holds that $T_p > T_m$. Obviously the productive time T_p must exceed the required time T_m.

Illustration. In order to have enough available capacity to manufacture the demanded amount of wafers, there should be 5 identical machines in machine family 1. For $n_m = 5$ it holds that $T_p = 588 > 580$ hours per week.

Calculating planning

On the other hand, for a given set of resources it can be determined what the demand allocation is. The full demand for a product set is d and the required time to manufacture this demand is T_m. The available ratio x_m indicates to what extent the demand can be met for machine group m:

$$x_m = \frac{T_p}{T_m}. \tag{5.6}$$

The lowest x_m forms the restricting capacity group and the resulting allocated demand is then $x_m \cdot d$. Note that the method above linearly cuts the demand in all products, until the level is such that it can be manufactured. The mix of products is not optimized here. Optimizing the mix is a process that is performed manually.

It is possible that one machine group can serve as backup for another, this will be used especially when the capacity of the latter is not sufficient. It might also be possible that the capacity of two machine groups are interchangeable. Then, the load that was originally projected for one machine group must be divided over the two in such a way that their resulting capacities are levelled. This is called capacity sharing. Suppose the load that is originally projected on machine group $m1$ is given by $T_{p,m1}$. However, this load exceeds the capacity, and part of this load can be manufactured by machine group $m2$. The share factor s_{12} indicates which percentage of the load is shifted from machine group $m1$ to $m2$. So, the new situation will be that:

$$T_{p,m1,new} = (1 - s_{12})T_{p,m1}, \tag{5.7}$$
$$T_{p,m2,new} = T_{p,m2} + s_{12}T_{p,m1}, \tag{5.8}$$

where s_{12} will be chosen in such a way that the difference between $T_{p,m1,new}$ and $T_{p,m2,new}$ is minimized.

Sharing capacity between two machine groups impacts the maximum allowed utilization. According to Equation 5.5, the more identical machines there are, the higher the utilization can be, whereas the cycle time still remains at a constant level. Thereby, the capacity for these machine groups will increase by sharing capacity. On the other hand, recipe dedications on machines will cause machine capacity to decrease. Dedication can be dealt with by creating a new machine group. Both the utilization of the new group and that of the existing one will decrease in order to keep meeting the cycle time targets.

Operator capacity

Machine costs form more than 50% of the wafer cost price (see Chapter 1), therefore machine usage is of primary concern. As discussed above, the two major performance characteristics for machines are utilization and availability. Operators and maintenance technicians play a significant role in achieving high machine productivity, see for example Pollitt [1998]. The classical approach in determining human capacity is ad hoc and usually based on experience. In this thesis, the subject of determining operator capacity is addressed shortly.

Static calculation methods which are based on the number of operations per operator are relatively easy to conduct. They show some significant drawbacks: the stochastic nature of machine operation and machine maintenance is not taken into account and the labor content is not accurate. However, as a first order estimation, the method is adequate.

The number of operations that must be performed on a machine group can be determined. Now standard operator productivity numbers can be used to calculate the required number of operators. To do so, assuming that a group of operators is responsible for servicing one or more machine groups, the number is known in advance. If it is known how many operations per day need to be performed, the required number of operators can be determined. Suppose a group of operators operate machines m_1 through m_{10}. The required number of operations can be determined using (5.4):

$$M_r = \sum_{m=1}^{10} M_m. \tag{5.9}$$

The number of operators O is then calculated using standard productivity measure q:

$$O = \frac{M_r}{q}. \tag{5.10}$$

The standard productivity denotes the number of operations that can be performed by one operator. This productivity measure includes non-productive operator time, like for example lunch breaks.

Illustration. Consider the 5 machines from machine family 1 located in one bay and operated by a group of operators. The required number of operations performed on these machines is $M_m = 420$ operations per week. For these types of operations a standard productivity of $q = 10$ operations per hour can be assumed, or 240 operations per day. Now the number of operators needed to perform these operations can be calculated as $O = 420/240/7 = 0.25$. So 1 operator is required (constantly) to operate the 5 machines and there will be time left to also operate other machines.

5.2 Application

The method for determining production capacity has been implemented in a decision support tool. This static machine capacity model has the objective to support decisions with respect to machine capacity. It enables to quickly evaluate decisions in the following fields:

- Determine the required number of machines to establish the actual and forecasted production volume,
- Estimate the production capacity given the number of installed machines as a boundary condition,
- Evaluate the impact of new process introductions, and
- Evaluate actual machine performance against target.

In this section the application of the static capacity model and its use for investment decisions is discussed. Such an investment analysis starts with a product demand. The demand will contain actual orders for the short term period, and sales forecasts for the long term. Figure 5.3 graphically depicts the forecasted development of the demand in the year 1998. The different colored bars indicate different process technologies or options. The bars indicate the product demand in wafers out per

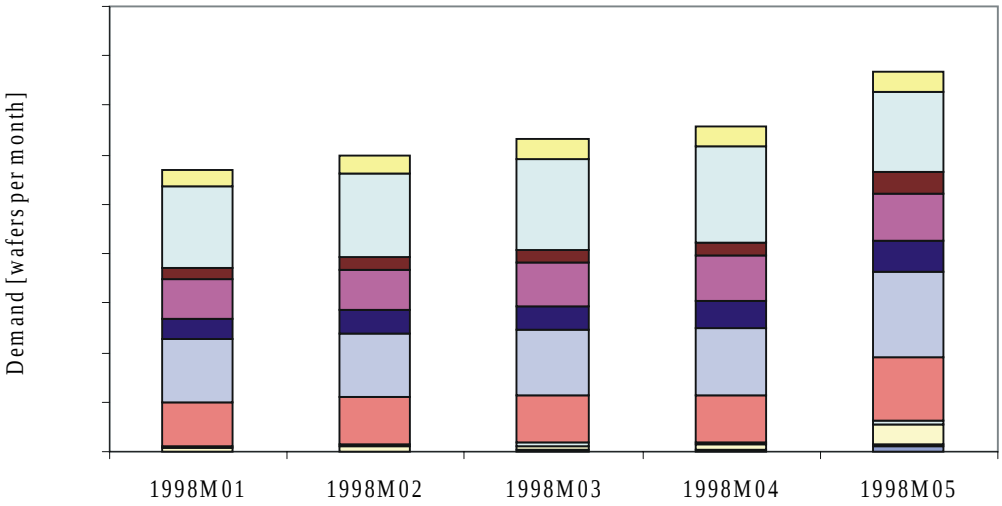

Figure 5.3: Development of forecasted demand.

month. It can be seen that the expected demand almost doubles within the year. The product demand is translated into requested wafer starts per week. This is done by taking into account the product cycle times and yields. Actual cycle time is influenced by the utilization of the fab. The lower the utilization the shorter the cycle time will be. This can also be put the other way around: if the cycle time must meet certain criteria, then it will impose a boundary for the maximum utilization.

The required number of machines can be calculated using process related data and machine related characteristics. A sample of process related data is presented in Table 5.3, where for the first eleven machine groups the required process time that is needed to process one lot is presented. The squared coefficient of variation of the arrival pattern and of the processing times, c_a^2 and c_0^2 respectively, can be determined using historical data. If no historical data are available, estimations can be made based upon the expected variations. Table 5.4 presents the resulting machine investment plan. For each month it is shown how many extra machines are required, again for the first eleven machine groups. The table indicates the number of machines needed to manufacture the complete demand. This means that in each month the indicated number of machines must be installed and qualified to perform production. A fractional value indicates that only part of the machine capacity is required in that specific month, for example, the increase of 0.25 in period 1998M06 for group ten indicates that only one of the four chambers of this cluster tool is required. In the next period another chamber is required.

Table 5.3: Process times.

Machine	**Process**				
group	**01**	**02**	**03**	**04**	...
01	15:30	18:36	12:24	15:30	...
02	00:00	00:00	01:45	01:45	...
03	15:13	16:19	11:43	14:08	...
04	04:36	03:40	05:30	05:30	...
05	01:26	00:49	00:00	00:00	...
06	14:30	05:18	05:06	05:06	...
07	04:10	02:30	03:20	03:20	...
08	03:45	01:55	01:55	01:55	...
09	09:56	12:53	15:15	14:35	...
10	01:14	02:54	03:44	03:24	...
11	04:20	05:12	03:28	04:20	...
...	...	...	...	...	...

Table 5.4: Resulting machine planning.

Machine	**1998**				
group	**M03**	**M04**	**M05**	**M06**	...
01	1	2	-	-	...
02	-	1	-	-	...
03	-	-	-	-	...
04	-	1	-	1	...
05	-	-	-	-	...
06	-	-	2	-	...
07	-	-	-	1	...
08	-	2	-	-	...
09	1	-	-	1	...
10	-	-	-	0.25	...
11	-	0.5	-	0.5	...
...	...	...	...	...	...

The evolution of the capacity results in a restricted fab capacity. This restriction is formed by the bottleneck of the fab. Information about the number of wafer starts per week, the occurrence of bottleneck machines, machine utilization, and so on, is presented and used in operations.

Using the static machine capacity model, a good estimation of the production capacity can be made. Experiences have shown that an accuracy in predicting wafer throughout of ninety percent or better is obtained. A prerequisite is that regular updates of the data are available. Therefore, it can be concluded that the model is fit to support capacity related decisions in wafer fab operations.

The model is static by nature and is therefore restricted to present stationary situations only. Effects caused by dynamic influences are considered as average values over longer time periods. Short term fluctuations are not taken into account. Therefore, the model is not suitable to perform day to day capacity analysis of the fab. To do this, a model containing more detailed information is needed. The model also does not provide accurate information about work in progress levels and cycle times. The conversion between starts and outs is performed by using reference cycle time data. These data are verified against actual performance.

5.3 Process equipment layout

The topic of wafer fab layout is discussed regularly in literature. The majority of literature emphasizes the building construction and the facilities installation, whereas only minimum consideration is given to production objectives, like high equipment productivity and short product cycle times. Nonetheless, fab layout influences fab productivity, throughput, and cycle times to a great extent. For example, an inappropriate floor layout can cause unbalanced WIP, long transport distances and times, and poor operator communication. Therefore, a layout design method should fit in the total concept of fab designing, focus on improving productivity, optimizing product flows, and improving work methods.

During manufacturing, wafers are moved from storage to machine, from machine to inspection, from inspection to assembly and inventory, and finally shipped to the customer. The wafers may either be moved manually or by some mechanical means. Time is required to transport them from one location to another. Material handling is defined as the transport, storage and control of materials in the total manufacturing life cycle of a product. Fab layout influences the flow of wafers throughout the manufacturing cycle to a large extent. The time and distances required for moving wafers should be minimized and storage areas should be organized accordingly. Decreasing the length of intensive transport routes will not only decrease the average

of the transport time, it will also have a positive effect on the variance in transport time. This leads to a better predictability and a lower variance in the arrival pattern of wafers at next operations.

Layout design is the physical arrangement of all resources involved in manufacturing products [Hayes & Wheelwright, 1984]. The required manufacturing resources are arranged in order to obtain an efficient wafer flow, characterized by short distances of high intensity flows and a minimum complexity of the flows. The wafer flow complexity is determined by the number of intersecting material flows. Besides the number of machines, their dimensions and location, other components have to be taken into account when designing the layout. Facilities like electricity, cooling water, process supplies (for example gases, liquids, and chemicals), and exhaust are necessary to perform production. These facilities typically comprise 10% of the total fab costs. Optimization of the facilities design to minimize cost and maximize flexibility can play a significant role in reducing both the initial construction cost, the maintenance cost, and the cost of future renovations. Therefore they have to be taken into account as a boundary condition in the overall layout design. For example, if the layout design follows the tunnel concept, where process tools and support tools are located on the same level, most of the facility requirements are supplied to the main level of the fab only. In a ballroom design, the main process tools are placed in the clean room, and the tool support equipment may be located on as many as three different levels.

There are many opportunities for either under- or over-designing the clean room and facilities using averaging or past experience. For example, with the incorporation of SMIF (standard mechanical interface) and FOUP (front opening unified pod) isolation technologies, the demand for a specific clean room classification for the fab has changed significantly. In clean rooms that are designed class 1 environments, the velocity of the air and number of air exchanges per hour provide more than adequate cooling of process tools. In an SMIF fab the clean room class is typically 100 or 1000, therefore the location and design of special filters must be engineered to handle the head load from the process and support tools.

Overview layout design methods

Various procedures for layout planning have been proposed and developed. Most of them use some sort of heuristic approach, since optimization is rather difficult for both process and layout planning. The traditional approaches in literature are systematic layout planning from Muther [1973], and production flow analysis from Burbidge [1971] and Burbidge [1989].

Systematic layout planning (SLP) is one of the most commonly used traditional approaches for layout design. Its objective is to minimize transport distances with a high transport frequency. To do so, four activities are defined to come to a detailed layout. First the area division is made. Then, the basic flow patterns are investigated. Next, the placement of machines and support devices is considered. Finally the installation takes place. SLP is used to design a functional layout in which machines of equal type are grouped. The method minimizes transport distances of material flows with high intensities. However, it often results in a layout with complex material flows.

Production flow analysis (PFA) is a technique used to plan the change in a factory from process organization to product organization and to plan the change from process layout to product layout. These two changes simplify the material flow in the factory. Therefore, PFA can also be defined as a technique for simplifying the material flow systems in a factory. In batch and jobbing production, PFA is defined as a technique for finding the families (sets of products) and groups (related sets of machines and other facilities) for group technology.

PFA consists of a succession of sub-techniques. It starts in large companies by simplifying the flow between factories or divisions using 'company flow analysis'. It then finds the best division of each factory into departments based on product organization and simplifies the material flow between them using 'factory flow analysis'. Next, it plans the division of the departments into groups with 'group analysis'. The flow of materials between the work centers in a group is then studied using 'line analysis'. Finally, 'tooling analysis' is used to find tooling families (sets of products which can all be made at the same setup using tools from the same set) to plan operation sequencing and to find sets of products suitable for automation. The method delivers a layout with less complex material flows. However, the areas are configured on the basis of the product routing. An important aspect such as transport intensity is not taken into account in this method.

Haagh, Wilkens, Van Campen, Rooda & Rulkens [1998] present a method that results in an area layout that has the advantages of group technology. Focussing on transport as well as process machine, this method results in a layout, which is characterized by high tolerance for machine down, manageable material flows, and short transport distances. The method has been used to design a rectangular shaped area. To design a complete fab, the method can be applied to every area. The analysis of the whole set of calculated optimized area lengths may then yield a motive to change the fabs outer dimensions or the size of compartments. Furthermore, it is possible to extrapolate the ideas of the method to other shaped areas. The method should benefit the design of recently proposed hexagonal area layout [Jansen, 1998].

Layout in practice: MOS4YOU

The requirements of a fab change during its useful lifetime. Productivity improvements, capacity expansions, and evaluation of process technology may cause changes in these requirements. The complexity of layout design lies in the fact that circumstances are changing very fast. A layout design is based upon best knowledge and predictions of the future. When time proceeds, new insights and new developments cause the layout design to be less from optimal. Therefore, an optimal layout design is calculated, but the layout will not be filled completely. The equipment that is needed to produce the demand is placed in the fab on time. Each time a new investment is needed, use can be made of the latest knowledge. Of course, the main objectives of the fab will not change that often. But since process technology develops very fast, the impact on equipment capacity can cause investments to change and therefore, the layout has to be adapted. Maintenance support equipment, MES terminals, convenience outlets, and other minor items often are let out. Though the utility requirements for each item are small, in total they do add up.

The above implies that the layout design method should deliver an initial layout. This initial layout should, however, be flexible towards the future, meaning that updates on the layout, based upon new knowledge or forecasts, will impact the layout. In practice, one can see that this may yield a non optimal situation. Whereas the initial layout design might be optimal, the changes over time might have caused the actual layout to be far from optimal. In the case of MOS4YOU, situations have changed so drastically, that it was decided to build an extension to the existing fab. One can imagine that this does not contribute to optimal fab layout and thereby to optimal fab performance. In order to satisfy maximum flexibility and to accommodate future changes, many facility systems are designed with excess capacity even though this drives up costs. This flexibility is then created by, for example, installing an abundance of pipes and valves, and an excess of clean room space. Practical experiences show that often the abundance installed facilities are not in the right place. Having an excess of clean room space is costly, especially if the clean room is of the ball room type. Although installation costs of ballroom and SMIF do not differ, the SMIF concept, with cleanroom conditions of class 1000, might contribute towards flexibility because it provides much cheaper costs of maintaining a clean production floor than the ballroom type, with cleanroom conditions of class 1 or better. Furthermore, mini-environments protect the wafers better and thereby contribute to higher product yields. For example, ammonia is a major source of contamination, against which mini-environments provide a better protection than is achieved in the ballroom. The concept consists of SMIF (standardized and isolated product carriers) and mini-environments. Flexibility is only provided if the concept is used consequently, thereby reducing the costs of keeping the production floor clean. Using WIP

racks, instead of automated stockers, also provides an extra degree of freedom. The automated stockers can be moved less easily.

For the layout design of MOS4YOU a combination of PFA and SLP was used. Machines with a key process function are placed close to each other, in the same area. Supporting equipment is distributed over the areas where ever they are needed. In this way, so called mini flows are obtained. Each equipment type has to be identified as being main (grouped according to function) or supportive (grouped according to PFA).

Boundary conditions that were imposed on the design of MOS4YOU are listed below.

- The building was already there, although it was empty.
- Facilities and use of chemicals were restricting the design. Supply of gasses and exhaust of polluted air had to be considered. For example, sulferic acid can not be transported vertically over more than 10 meters.
- Machines that use toxic or dangerous chemicals need to be placed in a separated space or in a place that can be separated from the rest of the fab easily: in case of emergencies it will not be necessary to evacuate the whole wafer fab, thereby reducing the impact on production.
- In the macro layout the location of equipments was determined. To operate a fab, however, a lot of auxiliary tools are required. Therefore, space for PCs, WIP racks, reticle storage, stockers, etc. needs to be reserved.

At the start of the design phase, MOS4YOU was planned to have one production floor. Introduction of the SMIF concept made it possible to build and use a second production floor. A first solution to the layout design would be to split front-end and back-end operations between the two floors. In this way the transport between the two floors would be minimized. However, because of technical constraints, it was decided to stick to the functional layout with auxiliary equipment to be placed in each functional area as needed. Since the lithography equipment forms the bottleneck of the fab, splitting it up would not achieve the highest utilization.

Table 5.5 shows the from-to matrix for the areas. It has already been mentioned that the lithography area was not to be divided over the front-end (implant and furnace) and the back-end (metal, dielectrics, and chemical mechanical polishing). The formation of areas is realized by manipulating the from-to matrix into a near-block diagonal form. In real life, the nature of the data set is such that a perfect decomposition is hardly ever obtained. In this situation, a near-perfect decomposition is obtained by considering the following objectives: minimize the number of zeros inside

the diagonal blocks and minimize the number of ones outside the diagonal blocks. Figure 5.4 shows the resulting layout for one of the production floors of MOS4YOU.

Table 5.5: Area from-to matrix, number of moves for producing one wafer.

		to									
		aux	**lit**	**dry**	**fur**	**imp**	**met**	**del**	**cmp**	**def**	**pcm**
from	**aux**	-	-	-	1	-	-	-	-	-	-
	lit	-	-	15	2	9	5	1	-	5	-
	dry	-	14	-	9	1	-	10	-	1	-
	fur	-	12	3	-	2	1	4	-	-	-
	imp	-	-	10	2	-	-	-	-	-	-
	met	-	2	3	1	-	-	-	-	11	-
	del	-	6	1	3	-	-	-	4	2	1
	cmp	-	-	-	-	-	4	-	-	-	-
	def	-	3	3	4	-	7	2	-	-	-
	pcm	-	-	-	-	-	-	-	-	-	-

5.4 Transport and storage

When specifying an internal transport and storage system, the following criteria are of importance: storage capacity, throughput, speed, flexibility, expendability, investment costs, cost of operation, reliability, time between failure, time for repair, cleanliness, engineering, and support. See also Lee [1997] for a discussion on performance analysis of automated storage and retrieval systems. Transport system capacity is expressed in number of moves per hour, thereby indicating the number of lots that can be transported in one hour. One move concerns the transportation of one lot from source to destination location. The cycle time of a lot consists of waiting time and transport time. The waiting time of a lot is the time between a request for transport and the actual starting time of the transport. The transportation time of a lot includes handling and actual transport.

The transport of wafers in pods is decomposed into three components. These consist of transport inside the areas, between the areas on one production floor, and between different production floors. Transport can be performed manually or by using an automated system. An automated system requires a significant investment cost and an average cost of operation. Manual transport requires the labor costs of operators that perform the transport. Over a longer time period, for example ten

Figure 5.4: Schematic layout of one production floor.

years, there is no economical difference between automated and manual transport. The automated system provides stable and predictable transport. The chance of failures (for example wafer breakage) is minimal and quick identification and retrieval of lots is guaranteed. A disadvantage of automated systems is their inflexibility. Once installed, the configuration can not be changed easily. Manual transport provides much more flexibility. But it has the disadvantages of being less reliable (both in performance and in quality).

MOS4YOU

The choice between automated or manual transport is made in the first place on reliability arguments. The economical arguments are of second order. For MOS4YOU the choice was made to automate the inter area transport and to perform manual transport within the areas. In this way, a stable transport system is obtained, while flexibility inside the areas is still provided. The inter area transport is performed by an automated transport system, running in a closed loop.

Taking the initial layout as a starting point, a rough estimation of the transport distance is made. It is assumed that wafers take the shortest route to the next bay, without storage. The result will therefore be an underestimation, but can be used to verify numerical results later on. The minimum transport distance of a fully produced wafer is 7.5 km. The transport distance of different alternatives is determined using computer simulation. The program uses the following data as input: the locations of the input and output places of the areas, characteristics of the transport loop, characteristics of the elevators, and the process flow.

To determine the transport capacity, the number of wafer moves has to be determined first. Here the inter area transport is considered. It is assumed that each area has a storage location that forms both input and output of the area. The objective of MOS4YOU is to produce 20.000 wafers per month. Besides production wafers, monitor and engineering wafers have to be transported also. Furthermore, as the transport system may never become the bottleneck of the fab, it should be over dimensioned. Wafers are produced and transported in quantities of 25. The wafers are placed in a cassette, which in its turn is placed in a so called SMIF pod, see Figure 5.5.

The activities a wafer undergoes once the process is started are the following. The lot starts in the machine family buffer. The shop floor control system signals which lot is going to be processed and the operator takes the lot from the buffer to the machine. Once the lot has been processed, the operator puts it into the next machine family buffer, if the next operation is within the same area, or on the output port of the area, if the next operation is within another area. In the latter case, the lot

Figure 5.5: SMIF pod.

is transported to the next area and placed in the machine family buffer of the next operation. The number of moves one lot has to make can be given in a so-called from-to matrix. Table 5.5 shows the from-to matrix containing all moves between the areas. This matrix can be transformed into a capacity requirement matrix by multiplying by the productivity level of the fab in number of moves per hour.

Storage capacity is expressed in number of storage locations. Practical experiences show that there must be enough storage locations to store 70% of the total work in progress in the fab. In the previous section, the relation between productivity, cycle time, and WIP has already been explained.

5.5 Overview

In the design of resources phase, the design decisions concerning the architecture were quantified. The resources can be divided into process equipment, metrology equipment, transport and storage equipment, facilities, and operating and maintenance people. In this chapter process, metrology, transport and storage equipment were considered, together with operating people. A decoupled queueing line model was developed and used to determine the number of these required resources. This model is an extension of existing approaches as relations from queueing theory are taken into account. The definitions used in the model follow the SEMI E10 standards, which enables internal discussions and external discussions with other fabs or suppliers. With the required number of resources the layout can be constituted.

The design criteria of this chapter are:

- Required number of resources, and

- Layout design plan.

The changing market environment and the fast development of new process technologies require that the design of resources is a repetitive activity and not something that is done only at the beginning of the design process. Therefore the initial design needs to be flexible with regards to future needs.

The model that was developed in this chapter is for use in investment and planning analysis. It is based upon average statistical values. However, a fab in operations experiences all kinds of variations. Therefore a production control system is needed to react on these disturbing influences. In the next chapter some important sources of variability and their effects are explored using discrete event dynamic simulations. This forms the basis for the production control concept that is described further on in the thesis.

Chapter 6

Factory dynamics

In the previous chapter a queueing model was used to determine the number of required resources. This approach can be classified as 'static'; the behavior is expressed using stochastic measures and always considers a steady state situation. There are, however, several non-steady state situations that need to be considered during the design process of a wafer fab. In this chapter four non-steady state situations are investigated using discrete event dynamic simulations. The objective is to gain insight in the dynamic behavior of the wafer fab and to derive design decisions. The situations are referred to as factory dynamics and concern the ramping, that is, filling, of an empty fab, batching machines, the down time behavior, and sequencing rules. First, ramping an empty fab can be done using several lot release policies. Determining the best performing policy is crucial for a fast fab start-up. Second, batching machines are not desirable because they disturb the wafer flow, however, to enable return on assets and productivity in some case batch machines are a must. Investigating the effect of batch machines on overall fab performance helps making the deliberation in whether or not batch machines are an appropriate solution. Third, obviously non-available machine time is a loss of capacity. However, the distribution of the non-available time also impacts fab behavior. Investigating the influence of the down time pattern can therefore be useful in determining variation reduction programs. Fourth, since sequencing rules influence the order of lot processing, their effect on fab performance should be investigated.

In this chapter, the four non-steady state situations are investigated with the purpose to obtain insight and to derive design rules, using a discrete event simulation model. First, performance measures are defined in order to evaluate the dynamic behavior. Then the discrete event dynamic simulation model is described, followed by the simulation results. An overview of design decisions completes the chapter. The results presented in this chapter were obtained by Eijsvogels [1998] and Jacobs [1998].

6.1 Little's law

To judge the performance of an industrial system, adequate measures are required. Measures reflect the customer requirements: a good product needs to be delivered at the right time at acceptable costs. On time delivery is obtained by realizing a short, reliable, and controlled cycle time. Acceptable costs are achieved when the throughput of the production system and the utilization of the machines are both high. Based upon the customer requirements, relevant performance measures should reflect financial -, productivity -, and yield measures. Financial measures consist of wafer cost and income from operation. Productivity measures are the quantity of wafers produced, cycle time, and machine utilization. Yield measures reflect the wafer yield and defect density. The above mentioned measures are commonly used in semiconductor industry. They provide good insight in the overall fab performance and there is extensive information available on benchmark numbers. However, for daily production performance measurements, more detailed measures are required.

The mean cycle time of lots in a production flow line can be analytically estimated using the queueing equations. For each operation the mean waiting time in the queue has to be computed. Adding these to the mean process times of the operations gives the resulting cycle time of the lots. In many cases we are not only interested in the cycle time but also in the number of lots that are in progress in the flow line. This is called work-in-progress, that is, the total number of lots either present in one of the buffers or being processed by one of the machines.

There is a simple relation between work-in-progress and cycle time. The more lots waiting are in a buffer the longer the waiting time will be. This is represented by Little's law, $w = \delta \cdot \varphi$, which states that the mean work-in-progress w equals the mean throughput δ times the mean cycle time φ [Little, 1961]. Here w is expressed in lots, δ in lots per time unit, and φ in time units. Little's law is valid in the steady-state situation and holds for all production systems. It is independent of the number of workstations and the amount of variability. The concept of Little's law is expanded in the following section.

6.2 Characteristic curves

In an ideal flow line, without any variability in arrival and processing times, products do not experience waiting times as long as the utilization is below 100%. At low work-in-progress levels, the cycle time has a minimum value which equals the theoretical process time plus the transport and handling time. As long as the output is smaller than the bottleneck capacity, the cycle time stays at its minimum value, without

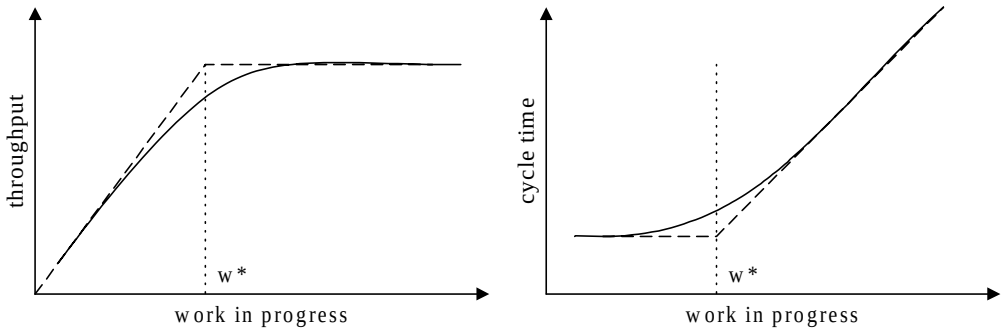

Figure 6.1: Characteristics curves.

any waiting times. The moment the throughput reaches its maximum value, waiting times start to occur. From this point on, an increase of the work-in-progress level results in an increase of cycle time, caused by increasing waiting times in front of machines. This theoretical situation is depicted by the dashed lines in Figure 6.1. w^* is the point where the throughput has reached its maximum value, while the cycle time is still at its minimum value. w^* is called the working point or critical work-in-progress, and can be expressed as:

$$w^* = \delta_{\max} \cdot \varphi_{\min}. \tag{6.1}$$

Thus, the theoretical throughput and cycle time in the zero-variability case becomes:

$$\delta_0(w) = \begin{cases} w/\varphi_{\min} & \text{if } w < w^* \\ \delta_{\max} & \text{if } w \geq w^* \end{cases}, \text{and} \tag{6.2}$$

$$\varphi_0(w) = \begin{cases} \varphi_{\min} & \text{if } w < w^* \\ w/\delta_{\max} & \text{if } w \geq w^* \end{cases}. \tag{6.3}$$

The ideal flow line provides the asymptotes for the non-ideal situation, where variability is present. Due to variable inter-arrival and processing times, throughput and cycle time curves resemble those in Figure 6.1. Increasing variability yields a smaller throughput and a higher cycle time, given a certain constant WIP. The minimum cycle time equals process time plus transport and handling time. At increasing work-in-progress levels, even below a utilization of 100%, waiting times will start to occur,

as expressed by the queueing equations. The waiting times are caused by variability in the production system. The non-ideal situation does not have a working point like the ideal situation does, it has a working range. Within this range, a choice can be made for maximum throughput with high cycle times or short cycle times with less than optimal throughput.

6.3 Model

In this section the discrete event dynamic model is briefly described. First, the assumptions are presented. Then, the model architecture is shown. Finally, the model verification is discussed. For a more detailed description the reader is referred to Appendix A, where the specification of the model is presented in more detail.

Assumptions

By definition a model always is a simplification of reality. Making a model of a real world system implies that some of the characteristics of that real world system are not implemented in the model and others are implemented in a more abstract way. The model described in this chapter is a simplified representation of the wafer fab. The assumptions and restrictions that were made are listed below.

Three different process flows are taken into account, characterized as 0.5, 0.4, and 0.35 μm. The processing times for each operation are assumed to be deterministic and known in advance.

Down times are determined from historical fab data as well as from specifications from the equipment owner. They are modelled using stochastic distributions. Furthermore, it was assumed that machines do not experience breakdown during processing. This is fairly realistic, as most of the unplanned maintenance is a reaction on mal-processing of wafers, which is only found during the inspections that find place afterwards.

The transport between the transporter, the buffers, and the machines is assumed to be timeless, the reason being that transport is designed to be not a bottleneck in the real-life system. For a process flow with 400 operations the actual transport time results in 24 hours, or 0.05 day per mask-layer. Since the total cycle time is in the order of 2 to 3 days per mask-layer, it is considered to be acceptable to neglect transportation times. Furthermore, it is assumed that scrap does not occur.

Batch behavior of machines is modelled on lot level, meaning that if a machine has a batch size of 12 wafers the batch size of that machine is assumed to be 0.5 lot. In

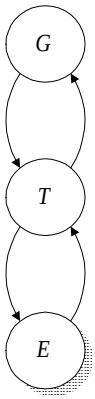

Figure 6.2: Model architecture, for explanation see text.

the same way, a machine with a batch size of 150 wafers is modelled by allowing a maximum of 6 lots. Furnaces are typical examples of batch machines.

Inspection machines are modelled to have abundant capacity, so they do not restrict fab capacity. Finally, operator behavior is not taken into account in the model.

Architecture

A discrete event dynamic simulation model has been built, using the specification language Chi [Arends, 1996], [Naumoski & Alberts, 1998], and [Rooda, 2000]. The result is a model architecture which consists of a lot generator and absorber G, a transporter T, and multiple machine families E. Figure 6.2 illustrates this model architecture. Process G forms the generator and the exit for the lots. At the start of the simulation, the empty fab is filled with products, using a specific lot release policy. Filling the fab with lots is called start-up. Later on in this chapter several start-up strategies are examined. The exit process is always able to receive lots that are completely finished. When the fab is fully loaded, the generator creates new lots on demand. The policy for creating new lots uses either a specific inter start time or WIP control.

Transporter T performs the transport of lots in and out the fab, and from one machine to another. Lots run in a specific process flow, thus the process routing is an attribute of each lot. Furthermore, the upcoming operation is kept with the lot. Using this information, the fab controller can determine the next destination of a lot.

Machine family E consists of a buffer B and a set of identical machines Eqt. The

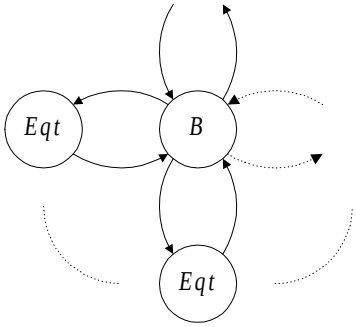

Figure 6.3: Decomposition of equipment family E.

decomposition of system E is presented in Figure 6.3. Buffer B can always receive lots from transporter T and from machines Eqt. When signalled by a machine, the buffer sends available lots to that machine for processing, using a sequencing rule. Depending upon the type of machine, one or more lots are sent. The processing of multiple lots with the same recipe in the same run is called batch processing. A machine family is modelled as a conveyor, see Rooda [2000]. The maximum capacity of the conveyor equals the number of machines m. A machine can be idle, processing, or down. Machine Eqt can receive a batch if there is at least one machine idle. If a machine is idle and there are lots to process, the machine status is changed from idle to processing. After a batch has been processed, it is sent back to the buffer, that updates the next destination of the lots. A machine that has finished processing a batch with lots can either go to the idle state or to the down state. The latter occurs when the machine is due for maintenance or when an unscheduled down event occurs.

Validation

The model was validated to check if its behavior is in correspondence with reality. Validation was performed by comparing the cycle times and utilizations of the model with that from the real fab. These comparisons are shown in Figures 6.4 and 6.5, and give confidence in the correctness of the model. The equipment utilization of the simulation model resembles reality. However, the cycle times of the model are too low with respect to reality, the difference can be up to 25%. This can be explained by the fact that the model lacks a number of variability sources, like detailed information about down times, operator availability and behavior, product holds, variable lot sizes, altering product mix, and the monitoring policy. The robustness of the model

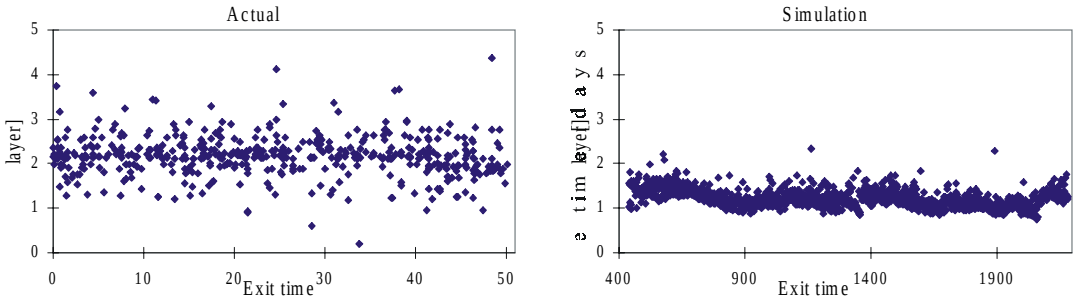

Figure 6.4: Cycle times: actual versus simulated.

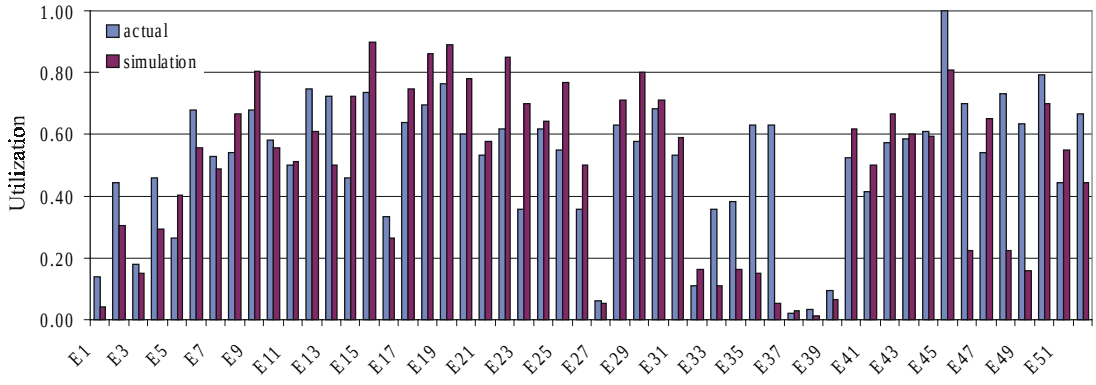

Figure 6.5: Utilization: actual versus simulated.

was investigated by varying the down time pattern, altering the process mix, and changing the batch sizes. The influence of the down time pattern and the batch sizes are subject of study of the remainder of this chapter. Changing the process mix has a significant influence on the performance of the fab. The difference in number of process steps between the three process flows is ten to twenty percent. Therefore, a changing process mix will have impact on capacity. This will, in its turn, have an impact on cycle times. The influence of other process flows was not investigated. However, the number of process steps is determining for the performance of the fab. Although the model could not be validated to the actual situation, interpretation of the simulation results do deliver reliable conclusions on factory dynamics. The impact of the effects that are found using simulation, will be even stronger in reality.

6.4 Application

In this section, the model that was described above is used to investigate four cases. The cases and the design decisions are described.

- Ramp of empty fab,
- Down time pattern,
- Batching machines, and
- Sequencing rules.

Ramp-up of empty fab

The start-up policy can influence the performance of a fab for quite some time. Start-up effects are for example encountered when production in a new empty wafer fab is started, when an existing fab is ramped (capacity expansion), or when lots are released from the wafer-bank (a logistic storage halfway the production process). The start-up behavior of a wafer fab is investigated using discrete event simulations. At the start of the simulation there are no lots in the fab. The fab is filled using some kind of start-up policy. The fab behavior is examined until the fab has reached a steady state. The steady state is characterized by a constant output level and by constant cycle times of the lots. Five different start-up policies were used. Each policy has a different input characteristic.

- Policy A releases the WIP into the fab at once. This means that the complete WIP is put in the buffer of the first machine family at the start of the simulation. After this, the release policy is constant WIP, meaning that a new lot may enter the fab when a finished one has left. Figure 6.6a depicts the lot start frequency and the number of lots in the fab as a function of time.
- Policy B releases the lots at a release rate that equals the desired output rate, depicted in Figure 6.6b. The WIP level in the fab gradually increases, but the burst of lots at the start of the simulation is avoided.
- For policy C the release rate is done in such a way that the fab reaches the desired WIP level at the same time as the first lot leaves the fab. This means that the lot release rate in the start-up stage will be higher than it is in the steady state situation, see Figure 6.6c.

- Policy D has an initial input rate that is higher than that of policy C. This input rate is kept constant until the first lot has finished processing. Therefore, the WIP level of the fab will be higher than the desired WIP level, see Figure 6.6d.
- Policy E has a high release rate, that is kept at a constant level until the desired WIP level of the fab is reached. Then constant WIP release policy will become active. Figure 6.6e depicts the release pattern and the WIP as a function of time.

The resulting cycle times for the five described release policies are shown in Figure 6.7. Policy A exhibits a large peak in cycle times of the lots, due to the fact that all lots are released at once at the start of the simulation. Policy D takes the longest time to reach the steady state situation. This is due to the burst of lots and the subsequent excess WIP level. It takes some time to decrease the WIP to the desired level. Policy B delivers the best result, in that no cycle time peaks are generated at all. For the fab, policy B delivers the most controllable situation. So the best release policy is to release the lots in the fab at the rate of the desired output level. There should always be WIP control as a valve to avoid excess WIP peaks. Although this conclusion might seem obvious, in practice it is not, because fab controllers often are tempted to quickly fill a fab with WIP.

Down time pattern

Machine down times play an important role in wafer fabrication industry. To examine the influence of the down time pattern on fab performance, deterministic down time distributions are compared to stochastic ones. Figure 6.8 shows the cycle time for both the deterministic and the stochastic case. Frequent and relatively short preventive maintenance can be characterized as deterministic, while unexpected downs cause a stochastic behavior. The increase in cycle time on the stochastic situation can be explained by the increase in waiting times in the buffers. This increase is due to the increase in variation. Using the queueing formula that was introduced in the previous chapter, this effect can be demonstrated. Due to the stochastic down times, the variance increases, thus having a negative effect on cycle time.

The above described effect occurs most at highly utilized machines. Therefore, a simulation was performed in which the top ten bottleneck machines have deterministic down times and the rest of the machines have stochastic down time patterns. Figure 6.8 shows the results of these simulations. Clearly, this simulation represents an ideal situation, as it is not possible to eliminate all elements that cause stochastic patterns. However, it can be concluded that reduction of the variations in down times

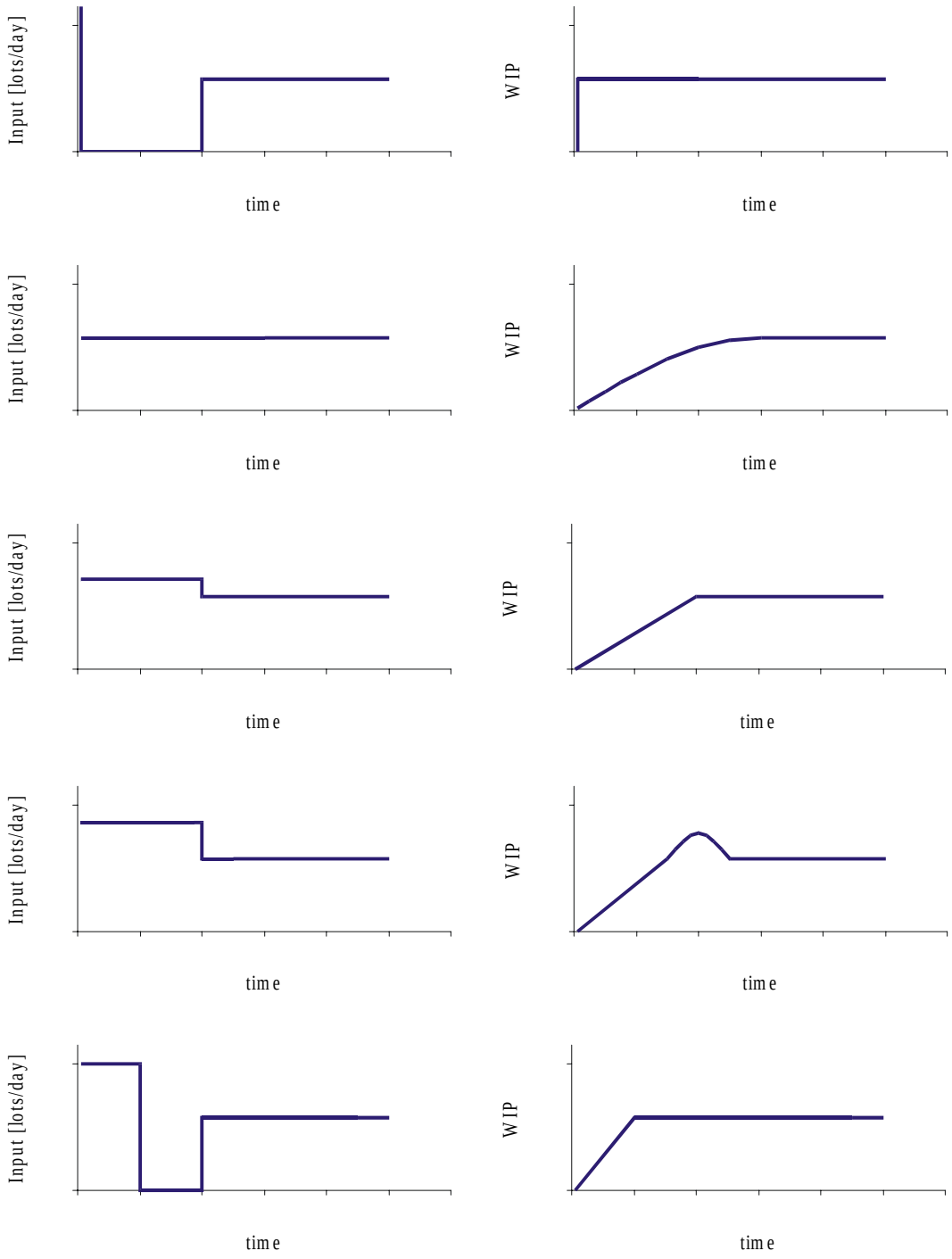

Figure 6.6: Release policies A to E, from top to bottom.

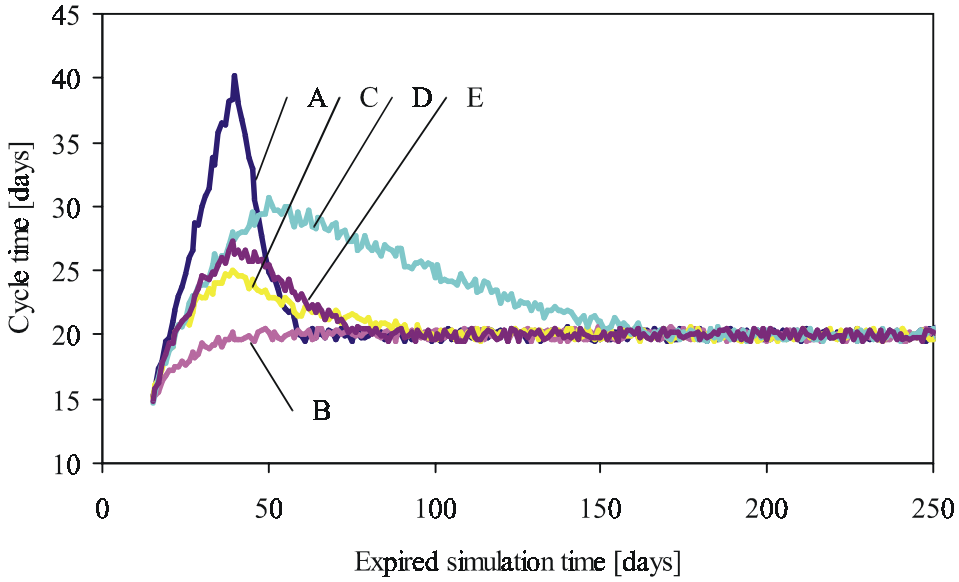

Figure 6.7: Results of release experiments.

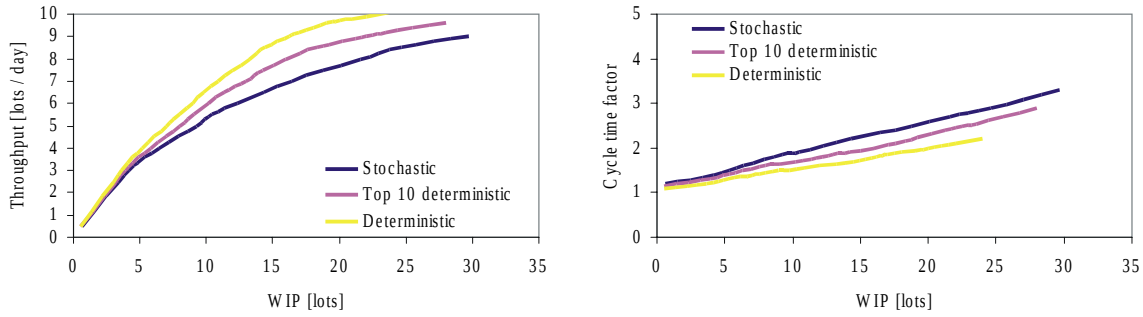

Figure 6.8: Influence of down time pattern.

(note that this does not necessarily means a reduction of down times) of bottleneck machines has significant impact on cycle times. Such a reduction can, for example, be achieved by performing more frequent preventive maintenance activities and reducing the time of the activities itself. For a single tool, the result of these actions can be calculated using the queueing equations. The added value of the simulation study described here is that all machines can be considered in interaction.

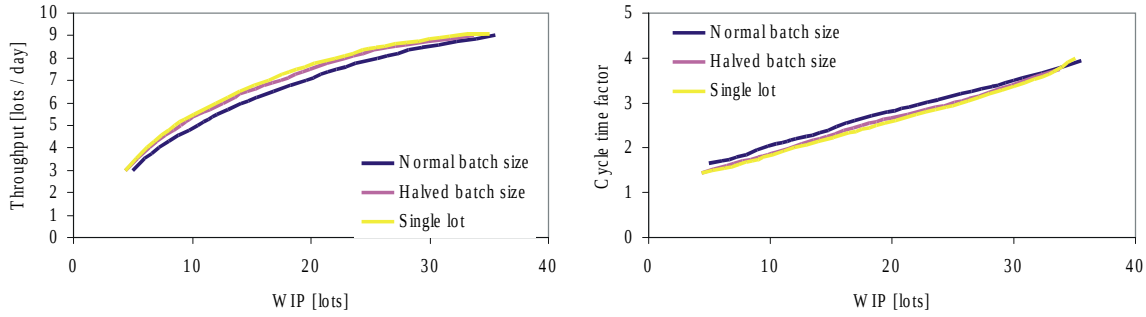

Figure 6.9: Influence of batching.

Batching

Several machine types are equipped to process lots in batches. A batch is defined as a group of lots that is processed on one machine at the same time. Furnaces are examples of batch machines. The process time is independent of the number of lots that are in the batch. Batches are formed in the buffer in front of a machine family. At the time a machine is ready to process the next batch, this batch is formed. Therefore, it can be that the number of lots in the batch does not equal the maximum batch size.

To examine the influence of batch machines on the overall fab performance, a simulation study was performed where batch machines were decomposed into machines with a smaller batch size. The normal batch situation was compared to a situation where the batch size was half the normal size and by a situation where the machines process single lots. For example, for the halved batch size experiment a furnace with a maximum batch size of six lots was replaced by two machines, each with a batch size of three lots.

Figure 6.9 compares the results of the simulations. It can be seen that the influence of batch machines on cycle time is present, especially at average levels of utilization. At higher levels of utilization and because there are six lots ready at the same time, the lots from the batch experiment will experience waiting time before the batch can be started, and waiting time at the next operation. At low levels of utilization, there is little difference, because the batches will contain one lot only. It is therefore concluded that batch machines must be avoided. If batch machines are required, for example because process times are exceptionally long, there should be abundance in capacity.

Sequencing

Given that there will always be a significant amount of WIP on the shop floor, decisions have to be made about which lot should go to which machine. Rules that are used to make these decisions are called sequencing rules. In this section two different sequencing rules are discussed. These rules follow from literature that was already discussed in Chapter 4. The rules are:

- EDD (earliest due date), and
- LCT (largest cycle time).

Each of the rules is compared to the FIFO (first in first out) rule. They are also compared with each other. One of the rules that is often mentioned in literature is EDD. Lots in the buffer are sorted by increasing due dates. The lot with the smallest amount of time left before its due date arrives will be processed first. EDD minimizes the maximum lateness, that is, the amount of time by which the completion time exceeds the due date. The due date of a lot is calculated at the start of the simulation and is set at 2.5 times the nominal processing time.

The LCT rule is used to determine the speed at which a lot has been processed sofar. Lots with a large cycle time have spent a considerable time waiting in queues. The cycle time can be compared to the target cycle time. In this way it can be calculated whether a lot is on schedule or not. The difference with EDD is that the number of passed operations is taken into account. To clarify the working of the LCT rule, an example is given using a re-entrant flow, see Figure 6.10. The figure shows a simplified re-entrant flow shop with two lots at different stages in their process cycle. The second lot has had more operations than the first lot. The current cycle time of the first lot is 2.5 days per mask layer and that of the second lot is 1.5 days per mask layer. Because the second lot is further in its process, it has a due date that is smaller than that of the first lot. The EDD rule therefore chooses lot 2 to be processed next. The LCT rule however, chooses lot 1, because it is behind on schedule. This example shows why the EDD rule should not be used in re-entrant situations.

Figure 6.11 shows the resulting throughputs and cycle times for the different sequencing rules. Both EDD and LCT show a smaller cycle time compared to the FIFO situation. The cycle time decreases with 5%. Furthermore, the standard deviation decreases also significantly. Figure 6.12 shows that the variance in cycle time for EDD and LCT decreases when compared to the FIFO situation. The difference between EDD and LCT is not significant.

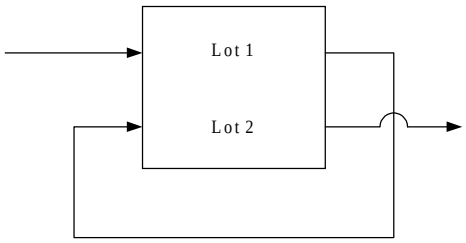

Figure 6.10: re-entrant system with two lots.

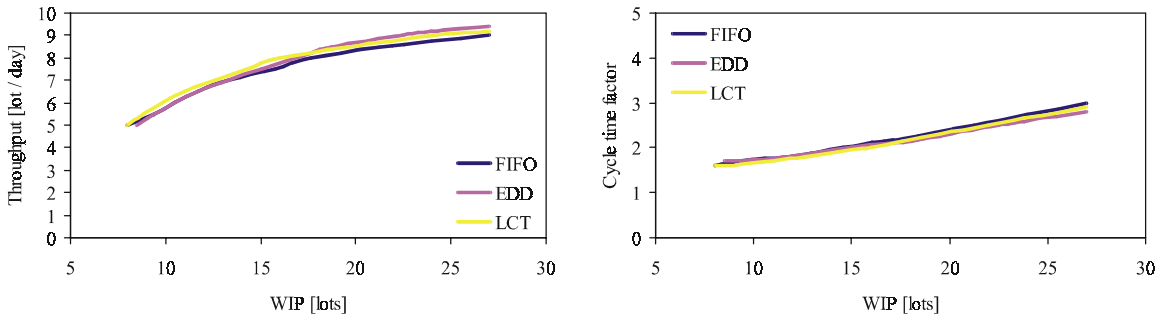

Figure 6.11: Comparison of different sequencing rules.

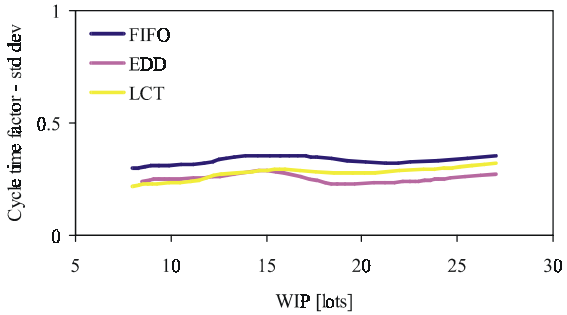

Figure 6.12: Comparing standard deviation of cycle times.

6.5 Overview

In this chapter factory dynamics were investigated. Four situations were subject of study: the ramp-up of an empty fab, batching machines, down time pattern, and simple sequencing rules.

The success of a new wafer fab is primarily determined by the speed of the start-up. Therefore, the lot release policy used to fill the empty fab is of importance. Releasing lots at the same rate as the desired output delivers the best results. In all other cases a peak in cycle time will occur first.

Besides the down time itself, the down time distribution influences both throughput and cycle time. Full reduction of variability in down time results in an improvement of more than 10%. Improvement activities should focus on machine groups that are capacity restrictive.

In cases of exceptionally long processing times, that is, longer than approximately four hours, batching machines provide a means to still fulfill productivity requirements. The introduction of batching machines results in an increase in cycle times and a decrease of productivity. This effect becomes more visible at high utilizations, therefore, batch machines should not be the restricting capacity of a fab.

Simple sequencing rules influence the order of lot processing at a machine group. Thereby they impact lot cycle times. For wafer fabrication a sequencing rule that takes into account the re-entrant behavior is desired. The largest cycle time rule considers the 'speed' of lots rather than their due dates. Therefore, the LCT rule is preferred. The LCT rule is part of the production control concept that is developed in the next chapter.

Chapter 7

Flow line balancing

In this chapter the production control system of a wafer fab is designed. The determination of required resources, discussed in Chapter 5, was based upon static (average) values. However, a wafer fab in operation suffers from many kinds of disturbances, causing variability in performance. Machine failures and operator availability are common examples of these disturbances. In order to cope with these variances a shop floor control system is required to perform fab scheduling. The rules that are needed to control the fab must be robust, to guarantee long term performance, and flexible in order to deal with short term fluctuations.

The objective of the wafer fab is to find an optimum between maximizing wafer output and controlling delivery performance. Wafer output can be directly translated into financial earnings. In times of scarce wafer production capacity, when the demand is higher than the installed capacity, wafer output comes first. However, a good delivery performance is also of importance, although it cannot be directly expressed financially. A good delivery performance is obtained by minimizing variance in cycle time, thereby enabling predictability of cycle time.

The interest in controlling wafer fabrication has increased significantly over the past ten years, something that is expressed in the number of articles found in recent literature. A lot of sophisticated approaches have been proposed and analyzed, often evaluated using discrete event simulation. However, actual implementation of advanced scheduling strategies is seldom discussed. In this chapter a survey on scheduling in semiconductor industry is presented, followed by the development of the shop floor scheduling strategy. This strategy consists of a lot release policy and of fab scheduling rules. A discussion concludes the chapter.

Survey

In production systems, the shop floor control activities can usually be divided into three hierarchical levels. The highest level is concerned with strategical planning and involves capital investments such as the development or acquisition of new production capacity. The middle level is concerned with the planning of production sales and forecasts on the basis of the available capacity specified by the strategic planning. The lowest level deals with detailed production scheduling. The wafer flow through the fab is monitored and controlled in order to realize the plans that were made on the middle level. At the lowest level, scheduling and sequencing can be distinguished. Scheduling can be described as the process of appointing orders to be processed on a specific machine at a fixed time. Scheduling is a global (fab wide) activity and results in a list of planned activities that have to be done and which also shows the intended times at which they will be done. Sequencing can be described as the process of deciding the correct order in which jobs should follow each other. Each time a new dispatch decision has to be made, the sequence of jobs is calculated based on the latest information.

Graves, Konopka & Milne [1995] present a review of material flow control mechanisms, such as just in time (JIT), kanban, optimized production technology (OPT), and materials requirement planning (MRP). These mechanisms were already named in Chapter 4. They emphasize the planning level and provide control mechanisms for the middle term.

Graves [1981] presents a review of production scheduling. An extensive listing of dispatching rules is discussed. These scheduling rules apply for a single machine situation only, more complex systems require apprehensive scheduling concepts. With semiconductor industry being the most emerging industry in the past years, the traditional classification of manufacturing systems needed to be expanded to consider the class of systems that is called re-entrant, see Chapter 2 and Kumar [1993]. In that class of systems, scheduling problems arise because several lots at different stages of processing may compete for the same machine. Uzsoy, Lee & Martin-Vega [1992] and Uzsoy, Lee & Martin-Vega [1994] present a review of production planning and shop floor scheduling policies in semiconductor industry. Common practice is to have due-date focussed scheduling rules, whereas little or no attention is paid to productivity focussed rules. However, according to Wein & Chevalier [1992], focussing on due-date alone does not result in an optimal fab performance.

Wein [1988] considers the impact that scheduling can have on the performance of semiconductor facilities. A variety of input control concepts and sequencing rules are evaluated using a simulation of a representative but fictitious wafer fab. The simulation results indicate that scheduling has a significant impact on average cycle time,

with larger improvements coming from input control, that is lot release strategies, than from lot sequencing. The reason that Wein [1988] experiences that scheduling rules have minimal impact on fab performance is twofold. First of all, the model assumptions neglect a large amount of variability, which, in reality, will have much more impact. Second, the scheduling rules are due-date focussed and aim at local optimization (that is, considering a single machine or a small group of machines) rather than at overall productivity improvement.

Lozinski & Glassey [1988] introduce a bottleneck starvation avoidance policy for controlling inventory in a semiconductor wafer fabrication. The bottleneck starvation indicators provide the mechanism for implementing such a policy in a manual decision-making environment. The fluctuation smoothing policy is expanded for multi-process flow systems by Sohl & Kumar [1995]. It aims at simultaneously reducing the burstiness in arrivals at all the stations in a system. Computational efforts are very high. The use of a super computer is mentioned and the method is therefore not usable in practice.

Bitran & Tirupati [1988a] and Bitran & Tirupati [1988b] discuss implementation issues of scheduling concepts together with the algorithms that form the basis of the scheduling system. Gathering information from the shop-floor is seen as one of the major issues in scheduling.

In the remainder of this chapter presents three subjects. First the lot release concept will be discussed, something which is not dealt with in literature. Then, the scheduling concept is described. It forms an extension to existing literature, as it has a global focus. Finally, its implementation for an IC wafer fab is discussed.

7.1 Lot starts

Shop floor control can be divided into two types of decisions: releasing lots on the shop floor and scheduling lots on machines. Customer orders have to be released on the shop floor and this should be done according to some sort of release policy. Scheduling decisions are carried out for every machine in the fab. Each time a machine is ready to commence processing another lot, the scheduling decision selects which lot to process.

The outcome of the middle level planning is a release rate, that is, an agreed number of lots per product type, to be used for the lowest level of control. The actual lot releases should be based upon the status of the fab and the management objectives. Simulation experiments described in Chapter 6 demonstrate that lots can best be released at a constant pace, thereby minimizing the variation in lot arrivals. As

discussed before, the WIP level is a parameter that influences output and cycle time. By defining a working range for the WIP, upper and lower limits for both output and cycle time are set. Releasing lots on the shop floor must be used to keep the WIP within the specified working range. The released number of lots and the WIP profile should be carefully monitored and controlled so that the WIP inventory level stays within the specified working range.

7.2 Fab-level rules

In this chapter, three types of balances are distinguished: fab balance, line balance, and flow balance. Fab balance considers balancing of equipment capacity towards the processes. The aim of improving fab balance is to become more cost effective. This can be illustrated by an example of a process mix that shifts from three-metal processes to five-metal processes. The latter processes require more back-end capacity (like metal deposition and metal etch equipment). Without changing the resources, the fab balance would be impacted. Fab balance is influenced in the design of resource phase, see Chapter 5, and is not considered in this chapter. Line balance considers the division of work-in-progress over the resources. The aim of improving line balance is to obtain a high overall productivity. Disturbances in line balance are caused by, for example, temporary bottleneck machines. Line balance can be influenced by scheduling rules. Flow balance considers the division of products over the process steps. The aim is to have products distributed uniformly over the subsequent process steps (or a group of process steps), to guarantee continuity in throughput. Flow balance is the result of production control activities (lot release and scheduling rules). In this chapter the term flow line balancing is used to indicate a combination of the latter two balances.

The balance for a single process step embraces the actual WIP and the target WIP. A good balance is obtained when the actual WIP equals the target WIP. Deviations from target, either upwards or downwards, result in a worse balance. At a certain time τ the balance for a single process step is defined as

$$\xi(\tau) = \left(\frac{w(\tau)}{w_t} - 1 \right), \tag{7.1}$$

where ξ denotes the balance for a process step, τ denotes the time, w denotes the actual WIP, and w_t denotes the target WIP in the process step. To smoothen positive and negative deviations the balance must cover a certain time window. For a certain time window $[\tau_1, \tau_2]$ the balance for a single process step is defined as

$$\xi = \int_{\tau_1}^{\tau_2} \left(\frac{w(t)}{w_t} - 1 \right) dt. \tag{7.2}$$

Using the theory of sample data, an estimation of ξ can be reconstructed by measuring a limited number of data points per time window.

The objective of scheduling is to make 'good' decisions regarding the processing of which lot and on what machine. These decisions immediately influence the performance of the fab. Therefore, focus cannot only be on long-term fab behavior, but must also be on dealing with short term practical issues. Output, or throughput, is one of the most important performance measures for a wafer fab: it generates income. The common opinion in literature is that, given an installed set of resources, output can hardly be influenced by scheduling [Uzsoy, Lee & Martin-Vega, 1994]. The reason for this is that only scheduling strategies are developed and evaluated that consider the queue immediately in front of an equipment. Especially for the reentrant flow factory, this view needs to be expanded in order to come to an optimal division of work over the fab. Balancing the whole wafer production will lead to a significant improvement in throughput and a smaller variation in cycle times.

From Little's law it can be seen that cycle time reduction is obtained by reducing work-in-progress. Decreasing work-in-progress while maintaining the output level can be accomplished by reducing variability ($c_a^2 + c_0^2$) or increasing installed capacity (causing utilization u to decrease). Scheduling affects variability in complex wafer fabs, because scheduling influences one of the major variability components: the arrival pattern. Although scheduling is a tool that can help improve the performance of an existing system, it can not change the basics of the system. For example, variability may originate from equipment behavior which is technologically limited. Therefore, changing the principle behavior of a system, such as required maintenance, has a far greater impact on variability than scheduling does [Hopp & Spearman, 2000].

Line balance indicates how good load (WIP) is divided over all machines in the fab. Also the division of load over the successive operations is of importance, this is called flow line balancing. Trines [1997] applies fab balance theory to a flow line and a job shop. The goal was to determine the number of resources and to divide them over the fab in such a way that the optimal capacity balance is obtained.

A perfectly balanced flow line is kept in balance if all lots move at a constant pace. Hereby it is achieved that all lots make the same progress, that is, lots that are in the beginning of their process undergo exactly the same number of operations as lots that are in the middle of their process. This is specifically important for reentrant production systems, where lots in different stages of the process are waiting

in the same queue. Flow line balancing is achieved when all machines also operate at the same pace. In this way, WIP is shifted uniformly through the fab and all machines will keep on working. Therefore, the first goal in flow line balancing is to have lots processed at a constant pace. One could make the analogy with a serpent (kind of elastic band) that has to be pulled through a narrow spiral shaped cylinder. To prevent the serpent from breaking, it has to be moved uniformly over its whole length.

Since no fab is perfectly balanced and is constantly liable to disturbances, something has to be done to try to restore the balance. Scheduling can reduce variability when it concerns machine to machine variability. At a given equipment group, the order of processing lots affects the variability of the output of that equipment group. As the output of one equipment group determines the input of others, the order of processing affects the variability at the other equipment groups. In the ideal case, the inter departure times of an equipment group equal the inter arrival times. In other words, the inter departure time of the sending equipment group should equal the inter processing time of the receiving equipment group.

A good measure relating the arrival pattern and the processing pattern is work-in-progress. The work-in-progress for the whole fab can be divided into segments. The work-in-progress for a segment not only indicates the current state, but also reflects the history of the states. Work-in-progress can be measured by comparing the actual work-in-progress w with a target value w_t. Since all machine groups have to perform at the same pace and their maximum capacity can differ, the target work-in-progress can also differ for each machine group. The target work-in-progress depends upon the machine utilization and the variability. The target work-in-progress values can be estimated using Little's Law and the queueing equation. These values represent the minimum inventory that is required.

Flow line balancing must be performed by taking WIP levels of both sending and receiving operations into account. First of all, the amount of WIP in the queue of the receiving operations is compared to the target levels. The lower the WIP at the receiving operation, the higher the priority of the lots will be. Sending WIP to operations with low WIP will prevent them from running idle, while sending WIP to operations with high WIP does not improve the cycle time. Second, the amount of WIP waiting in the queue of the sending operation is compared to the target levels. Now the opposite holds: the WIP surplus at the sending operation is given priority. In this way, large amounts of WIP are prevented. Table 7.1 indicates the order of priority at the sending operation, based upon the WIP levels of the sending as well as the receiving operations. The highest priority is denoted by 1 and the lowest by 4.

The flow line balance concept is illustrated by an example with four operations (or process steps). The four operations are at different places in the process flow and

Table 7.1: Flow line balance priority.

		receiver	
		$w \geq w_t$	$w < w_t$
sender	$w \geq w_t$	3	1
	$w < w_t$	4	2

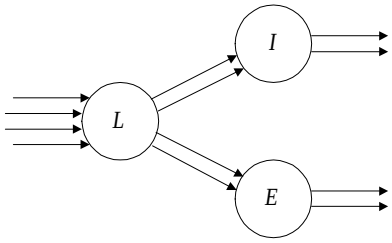

Figure 7.1: Part of a wafer flow with one sender and two receivers.

they must be performed on the same equipment group: lithography machines L. The first two operations have the same subsequent equipment group: implant machines I. The latter two also have the same subsequent equipment, but different from the first two: etch machines E. This situation is schematically depicted in Figure 7.1. Now, for each of the four operations, the WIP at the sending equipment can be determined as 4, 8, 10, 3 respectively. The WIP at the receiving equipment is given by 10 and 14 at I, and 3 and 5 at E. The flow line balance priority can be determined using Table 7.1. The resulting priorities are presented in Table 7.2, assuming that all operations have an equal WIP target of $w_t = 7$. The latter will be the case when there is only one process flow active. In this example, lots that are in operation 3 have the highest priority for being processed on machine group L, before operations 4, 2, and 1 respectively.

Once the flow line balance priorities for the operations have been determined, a selection has to be made of the lots that are within the same operation. Here, use is made of results obtained in Chapter 6. Lots that are within the same operation are prioritized on largest cycle time factor. The largest cycle time factor rule has the objective to minimize due date tardiness and thereby maximize delivery performance.

Table 7.2: Flow line balance priority example.

Operation	wip at L		wip at I		wip at E		Priority
	actual	target	actual	target	actual	target	at L
1	4	7	10	7	-	-	4
2	8	7	14	7	-	-	3
3	10	7	-	-	3	7	1
4	3	7	-	-	5	7	2

7.3 Area-level rules

The previous section discussed the real-time reactive scheduling strategy that is used to control the fab on flow line balance. It is a reactive strategy because it reacts on disturbances and aims at returning to the optimal situation. A proactive approach must be used to improve equipment productivity. This proactive element concerns the decision of what recipe can be best processed on what machine. The need for proactive scheduling can be illustrated using the setup of cluster tools. Assume a cluster tool with four chambers, which are all different. This machine is able to process four different operations. Assume another cluster tool, also with four chambers, but with two identical chambers. The latter machine is only able to process two operations, but does it in half the time. When determining the lot sequence for the first machine and taking into account the capability and throughput of the second overall performance will increase.

The recipe capability indicates which recipes can be processed on what machine and what the accompanying throughputs are. This first element, determines the capacity of the machine group. Besides not having enough capacity, the amount of excess capacity will impact cycle times. As a rule of thumb, in order to meet disturbances such as machine downs or unbalanced process mix, it is necessary that at least three machines are capable of processing each recipe. The second element determines the productivity of the machine group. Taking the expected processing time into account when making the scheduling decision, results in improved performance.

Taking the recipe capability into account (in the scheduling rules) will lead to a higher throughput, because machine utilization can be increased and variations will be lowered. For example, the machines with low capability need to be used for processing first in order to have the machines with high capability available to offer flexibility with respect to recipe capability. MOS4YOU has implemented the recipe capability matrix (RCM) and takes this into account in the sequencing (by using an equipment preference indicator). This has led to better controllability, a better line balance, and higher equipment utilization.

7.4 Implementation

The principal thought in the production control concept is that operators dispatch lots, based upon the information they obtain from the scheduling system. Thus, the operator can use his own insights to differ from the proposed lot sequence. There should, however, only be urging reasons to differ from this sequence. Reasons to differ can be: one of the storage systems for reticles is down unscheduled, making it impossible to retrieve the reticle of the first lot on the list. So the reasons to differ are disturbances not taken into account in the scheduling rules. The freedom for operators to differ from the proposed sequence is of importance, because it offers the flexibility to cope with actual disturbances. Variability elements can never be fully integrated into the scheduling decision.

Engineers and management of MOS4YOU recognized the immediate need for effective scheduling rules on the shop floor. In the absence of simulations results, the approach chosen was to implement the above discussed scheduling concept in a home-grown prototype scheduling tool. The scheduling tool was based on Excel and used many different data sources. It was an off-line tool, which was updated every ten to fifteen minutes. Consequently operators had to consult two computers (one to select and one to dispatch a lot) and the information that was presented might be outdated and therefore incorrect.

The tool was introduced in one area as a pilot. Using those experiences, both the rules and the tool were adapted and rolled out in the rest of the fab. The tool was not integrated in the manufacturing execution system, thus providing flexibility to the operator to use his or her own insights and to overrule the advice provided by the scheduling tool. During this learning process the rules were improved during production. Impact on fab performance was significant. Figure 7.2 shows a trend graph of the cycle time over six months. These experiences provided confidence in the shop floor control concept.

The prototype tool however, was not reliable enough to support production without restrictions. The large variation in products and recipes, the heavy load development, and the fluctuating availability of machines indicated the need for a more flexible tool. Besides that, the restrictions of a spreadsheet based application limited the use of complex rules, which are needed to refine the scheduling concept on the area level. Flexibility is required to account for super-hot lots, hot lots, monitor-, dummy-, and test-wafer lots as well as hibernated and hold lots. Therefore, a new tool was required which enabled real time sequencing, using the developed rules, in fast response times.

Requirements for a scheduling system include that it must be an off-the-shelf package, able to support the sequencing strategy that was described before, and provided

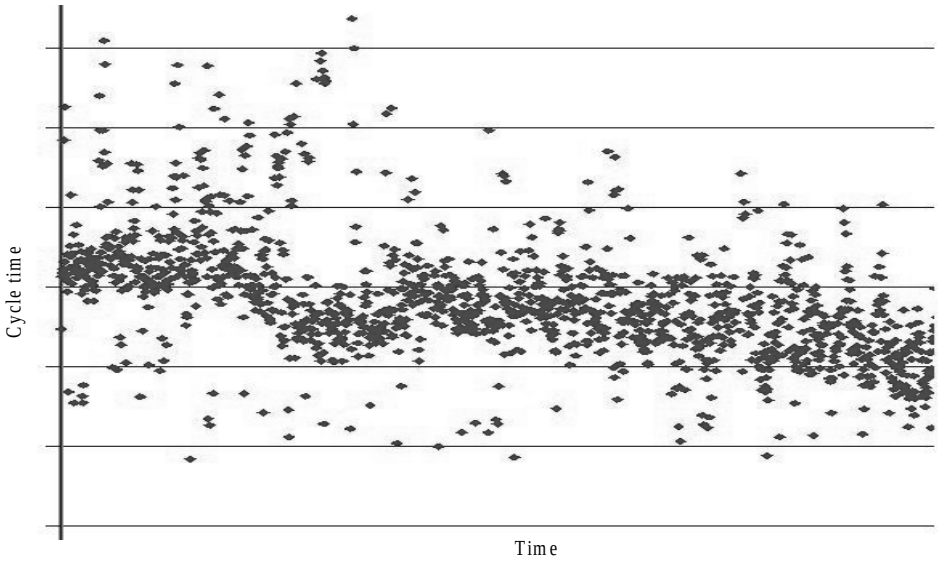

Figure 7.2: Actual cycle time trend.

by a supplier with excellent references in semiconductor industry. After a thorough definition and selection phase, at the end of 1998 MOS4YOU operations team decided to obtain APF-RTD (Autosched Productivity Family - Real Time Dispatcher) and integrate it with the manufacturing execution system (MES). APF-RTD is made accessible to the operators using a graphical user interface, called LOT4YOU, see Figure 7.3. LOT4YOU is located on the MES computer, next to the GUI that operators use to dispatch lots, where it places dispatch requests on demand and displays the resulting candidate lot list. Eventually, this functionality should be integrated in the MES GUI.

7.5 Overview

Wafer fabs are bothered by lots of disturbances. There is a constant drive to achieve excellent machine utilization [McIntosh, 1997]. Instead of randomly selecting WIP to proceed, operators have to carefully select the proper WIP to optimize the distribution of the profile of WIP. In this chapter, a control concept for MP2 wafer fabs was described, with the objective to come to an optimal WIP profile. This is done by controlling the wafer flow on the activity level and also by reacting on disturbances in the WIP division. Only after flow line balance priorities have been met, attention is paid to individual lot cycle time performance.

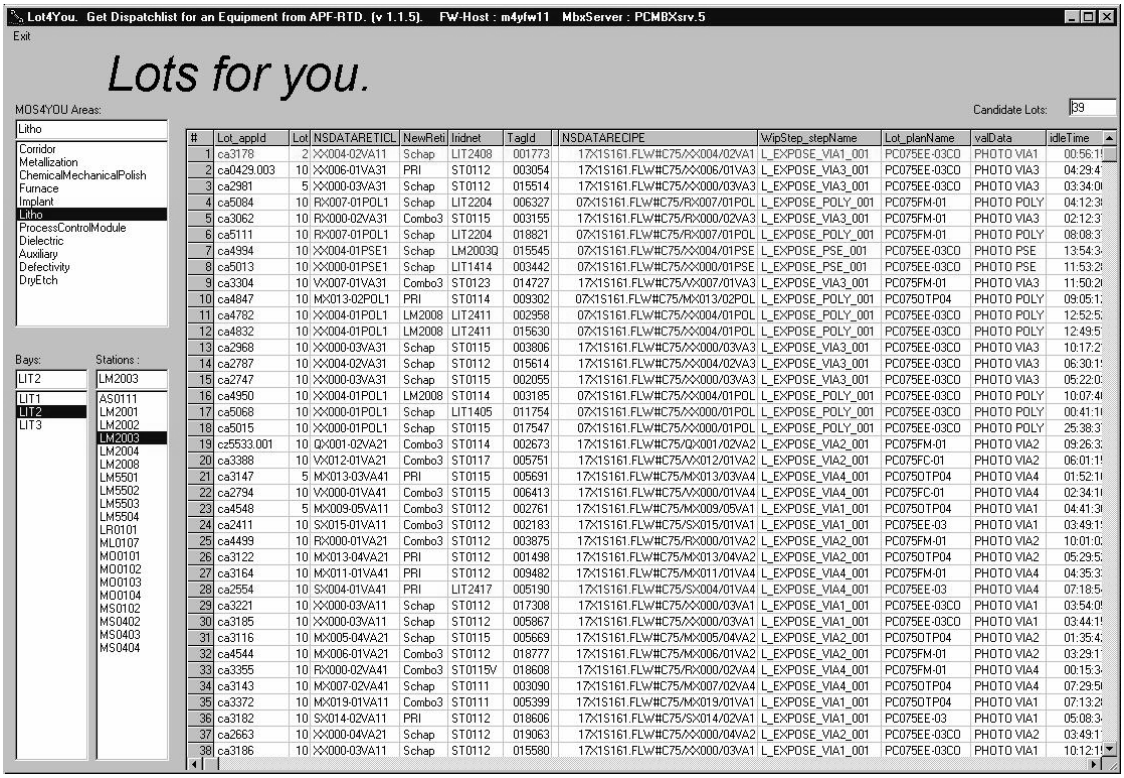

Figure 7.3: User interface to the scheduling tool, prepared by [Ledder, 1999].

The design decisions of this chapter are:

- Lots are started at a constant release rate, derived from the middle planning level,
- The rate of lot starts are corrected depending on the actual WIP status,
- Operators dispatch lots according to a preferred lot sequence, and
- The lot sequence is determined by flow line balance criteria, recipe capability characteristics, and least slack.

Controlling wafer fabrication by focusing on flow line balance, as described in the above, leads to reduction of variability and thereby reduction in cycle times, while throughput remains at constant level. It is a paradox that better cycle time performance can be achieved by focusing on efficient machine performance and ignoring due-dates when sequencing on the shop-floor. The reason for this is that focus on flow line balance maximizes the utilization of all machines, especially that of the bottleneck machines. Thus, while maintaining output, WIP can be reduced and thereby cycle times. Cycle time based dispatching policies, on the other hand, will prevent the bottleneck machines from being utilized effectively in the long run. A larger amount of WIP is required to maintain the output level, resulting in increasing cycle times. The use of due-dates in sequencing is counterproductive in the long run, and will lead to inferior due-date performance [Wein & Chevalier, 1992].

To maximize the utilization of production machines, fab operators have to monitor the status of machines closely to prevent the machines from running idle. A short time before a machine finishes processing, operators must select the next lot to be processed. But the current operator user interface cannot provide such information. Therefore, a user interface that presents the machine progress is needed. Furthermore this interface must provide access to the scheduling tool in order to obtain sorted dispatch lists according to the scheduling concept. Only then can the implementation of the scheduling tool be regarded as being completed.

Chapter 8

Area line balancing

Design of operations is the last activity in the design process. During this activity the fab is optimized at the area level. Each area contains a specific main process technology, with its own specific characteristics. Therefore, it is of importance to consider each area apart from the others, to obtain the final optimization. Several studies have been conducted considering the area level. The litho area by Rulkens [1997] and Leijsen [1998], the furnace area by Rulkens, Van Campen, Van Herk & Rooda [1998], the dielectric area by Balleggi [1998], and the dry etch area by Noben [2000]. These studies resulted in implementation of improved micro lay-out, improved working methods, and new or improved scheduling rules. This chapter presents the improvement study that was carried out for the metal area. Parts of this chapter haven been published by Lemmen, Van Campen, Roede & Rooda [1999].

The objective of this chapter is to improve the performance of the metal area using scheduling rules. Discrete event dynamic simulation models are used to investigate the influence of scheduling rules on the performance. Preceding the specification of the model, a detailed analysis of the area was performed. MOS4YOU is divided into hybrid functional areas, see Chapter 4, one of which is the metal area. The metal area is characterized by recirculation of lots, that is, lots visit the area various times for processing. Using scheduling rules in the metal area has a larger effect on area and fab performance than it would have in an area without recirculation, like for example the implant area. More specific, the optimization approach is applied to the equipment family of cluster tools which experience the recirculation of lots.

The main characteristic for the metal area is that the operations are mostly performed on so called cluster tools. Individual operations, which used to be processed on discrete machines, are grouped or clustered together, and are now performed on integrated tools. Clustering several operations in one tool offers important economic

advantages, like saving cycle time and inventory as well as manufacturing simplification and yield improvement. As a result of the increased complexity of integrated machines, the cost of semiconductor equipment has increased much faster than other factory related costs, and therefore performance analysis for cluster tools is much more important than it used to be.

Optimized semiconductor wafer fabs with high tool utilization and low cycle time are required in order to compete in today's multi-process multi-product market (MP2). Due to their specific characteristics, cluster tools turn out to have a significant influence on the fab performance. Cluster tools combine multiple processes in one machine - like small factories - having the advantage of low operational costs per wafer, see Chapter 2 and [Atherton & Atherton, 1995]. Integrating processes in one machine, however, introduces more complexity in controlling the wafer flow. The tendency of the semiconductor industry to use cluster tools more and more, increases their impact on overall fab performance. Analyzing and optimizing the interaction of the chambers and the transport robot of a single cluster tool (also called cluster tool balancing) is therefore a basic necessity for a well performing fab. Previous research focused on this aspect of cluster tool optimization [Wood, Tripathi & Moghadam, 1994].

Additional to cluster tool balancing, fab performance will increase by optimizing the interaction between multiple cluster tools, especially if a cluster tool family is characterized by cluster tools with different chamber setups. Different chamber setups decrease recipe availability of equipment; that is, not every cluster tool in a family is capable of processing all process steps. Recipe availability also decreases due to chamber maintenance and failure, as the number of possible recipes to be processed on the cluster tools decreases during these events. As the number of different products and processes in MP2 wafer-fabs grow, the number of different recipes in the fab will do so too, increasing the effect of recipe availability at cluster tools. Figure 8.1 shows, as an example, the impact two different scheduling rules, rule R1 and rule R2, can have on cycle time performance. Rule R2 results in a lower cycle time, while utilization (and output) remain on a constant level. A lower average cycle time obviously results in a lower variance in cycle time.

In this chapter, a dynamic simulation model is used to analyze cluster tools and their interaction behavior. The focus is on performance improvement of cluster tools by using scheduling rules. Studies by Panwalker & Iskander [1977], Haupt [1989], and Uzsoy, Lee & Martin-Vega [1994] show that scheduling has a significant effect on fab performance.

The approach starts by investigating the metal area, focussing on cluster tools. Second, a discrete event dynamic simulation model of the area is developed, in which the cluster tools are specified at a high detail level. Third, scheduling rules for lot

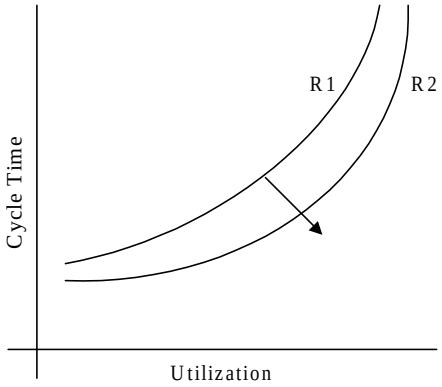

Figure 8.1: Effect of two different scheduling rules.

dispatching at cluster tools are defined and implemented in the model. Fourth, simulation experiments are designed to obtain a structured analysis of the effects of the defined rules. Finally, the results of the experiments are investigated and conclusions are drawn.

8.1 Analysis

The three components of the area - that is, product, equipment, and control - were thoroughly analyzed. Three different product-technologies have been distinguished, each with its own critical dimension for wafer production, that is 0.5 μm, 0.35 μm, and 0.25 μm. With the first two technologies, ICs are manufactured using a three metal layer process, resulting in a wafer re-entering the metal area three times. The third technology is a five metal layer process. A salicide layer is formed on the wafer after the first entry into the area. On top of this layer, three or five layers of barrier (TiTiN) followed by metal (AlTiN) films are deposited on the wafer. Tungsten plugs (W) inter-connect the conducting metal films. Figure 8.2 depicts the forming of the Tu plugs and the construction of the metal layer. The table on the right defines the processes used for each step.

Five different equipment types are used in the metal area and each type is functionally grouped within its own type or family. The equipment types perform the following operations: sputter-deposition of the AlTiN and TiTiN layers (TS), annealing of the layers (TA), deposition of the W-plugs (TD), etch-back of the plugs (TE), and cleaning of the wafers (TC). It is assumed that the capacity of each individual process was calculated properly. The machine capacity for each equipment type was determined

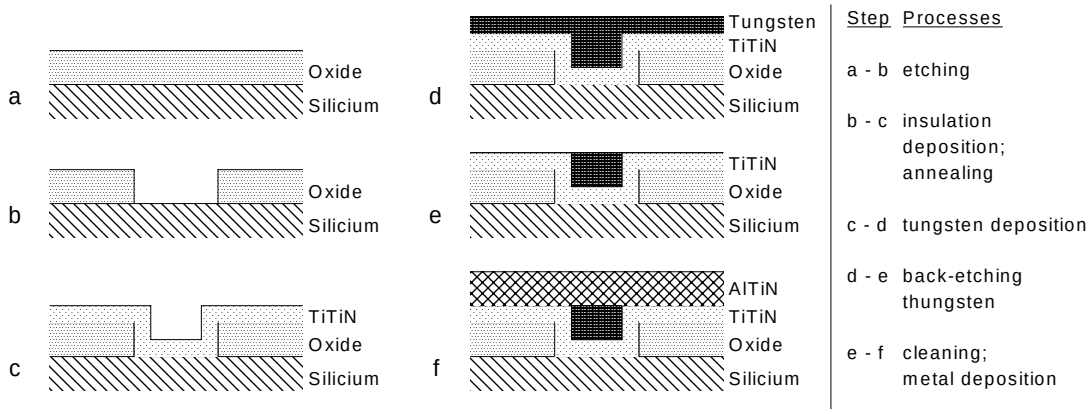

Figure 8.2: Forming of the metal layer.

using the method presented in Chapter 5. In this chapter no changes are made to the equipment capacity, but scheduling aims at better utilization of the equipment.

In this chapter the primary focus is on the TS family, because that is the equipment type which is the element of recirculation. Those parallel processing radial cluster tools [Atherton & Atherton, 1995], as depicted in Figure 8.3, can be divided into three classes depending upon their chamber configuration. A class can deposit either AlTiN, or TiTiN, or both film types on wafers. The classes are denoted by TS-MD, TS-BD, and TS-ND respectively (where MD stands for metal dedicated, BD stands for barrier dedicated, and ND stands for no dedication). All machines are maintained on a periodic time schedule as well as on their usage. An example of a periodic time schedule is a quarterly preventive maintenance. An example of usage based maintenance is cleaning after 100 Å have been deposited. Furthermore, the TS family needs to be reconditioned after having been idle for some hours and before switching from TiTiN to AlTiN. Last but not least machines have to be repaired whenever unexpected down events occur. The etch-chamber E is needed only for processing barrier layers.

Lots, consisting of 25 wafers, can be stored in a buffer in front of equipment families. Fab-level rules are used for lot dispatching in those buffers, see Chapter 6 and Chapter 7. Table 8.1 shows the four parts of product routings that can be distinguished. It can be seen that the last three routings recirculate the area once.

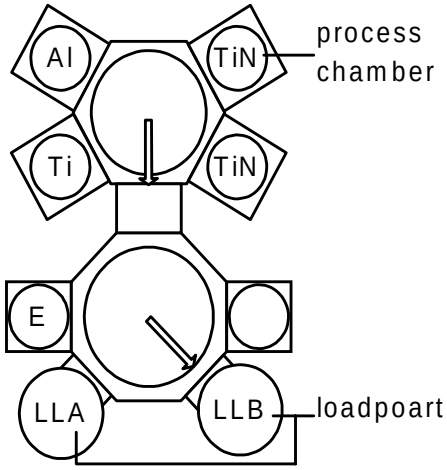

Figure 8.3: Cluster tool configuration.

Table 8.1: Process routings for the metal area.

Equipment sequence	**Process specification**
TS TA	salicide
TS TA TD	metal 1a
TS TA TD TE TC TS	metal 1b
TS TD TE TC TS	metal 2, 3, 4, 5

8.2 Modeling

A discrete event dynamic model is an appropriate and powerful aid to investigate the effects of cluster tool performance using different scheduling rules. In this section, a model of the metal area is built using the information presented in the previous section.

An architecture defines in which way components of a specific system are organized and integrated. The metal area can be viewed as a separate area in the fab, which is visited one or more times by the wafers. The wafers undergo one or more operations in the area and then move on to undergo operations elsewhere in the fab. Figure 8.4 presents the interaction between the metal area and the rest of the fab.

Analysis of Section 8.1 shows that routing TS-TA-TD-TE-TC-TS is the longest product-routing. All other routings can be derived from this routing by skipping one or more equipment families. The metal area can be described as a flow shop with

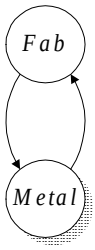

Figure 8.4: Model architecture of a wafer fab.

skipping and re-circulation. The re-circulation is needed because, within the same mask layer, wafers need to be processed twice on the TS. Compared to a job shop, it has the advantage of a relatively small number of possible routings. This advantage simplifies the control of the system. Figure 8.5 depicts the architecture used to specify the model of the metal area. In this figure, circles and arrows denote processes and product-flow channels, respectively. The shadowed circles depict the equipment families, G generates lots, I and R recirculate layers and X is the exit. The wafer flow is re-entrant, meaning that wafers return to the metal area after they have been processed in one or more of the other areas. This is indicated by the waferflow from process X to process G.

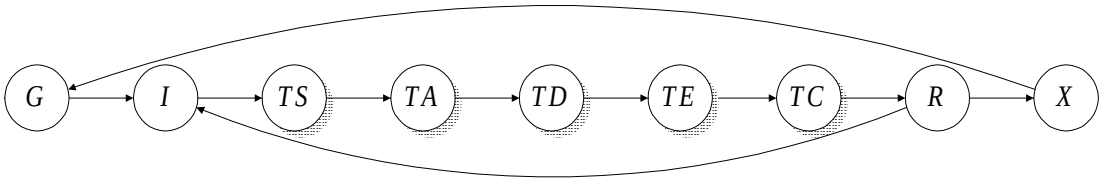

Figure 8.5: Model architecture of the metal area.

A formal specification is defined, based on the architecture of Figure 8.5. At each family it is decided whether the lot needs to be processed or passed to the next equipment family. Each equipment family consists of one or more similar machines. For each machine the recipe capability is specified, something which is expressed by the chamber setup. This means that although the machines within a family are similar, they do not necessarily need to have the same capability. A chamber can process one wafer at a time and process times are dependent upon the recipe of the operation. Preventive maintenance, for example triggered when a limit for the amount of processing hours is exceeded, is taken into account. Machine availability is modelled using stochastic down time behavior. MOS4YOU aspects with a computed relatively small effect on the area's performance, as determined by Lemmen, Van

Campen, Roede & Rooda [1999] - for example transport time, wafer scrap, equipment monitoring, operator availability, measuring equipment - are not modeled in full detail. The discrete event specification language Chi [Arends, 1996] is used for the specification of the model. This specification can then directly be used to simulate the dynamic behavior of the area using the Chi engine [Naumoski & Alberts, 1998].

The model proved to predict the effects correctly during the verification of both its components and the entire system, that is the combination of all the components. Validation tests for a specific set of workload levels were conducted to determine and analyze the differences between the model and the fab. The cycle time of a lot and the throughput of the area were monitored during those simulations and compared to data of the actual fab, as depicted in Figure 8.6. Simulations show that the model has a 20% higher throughput. As some aspects of the metal area - which do disturb the fab performance - are modeled with less detail or omitted in the model, the model defines the upper boundary for the performance of the metal area.

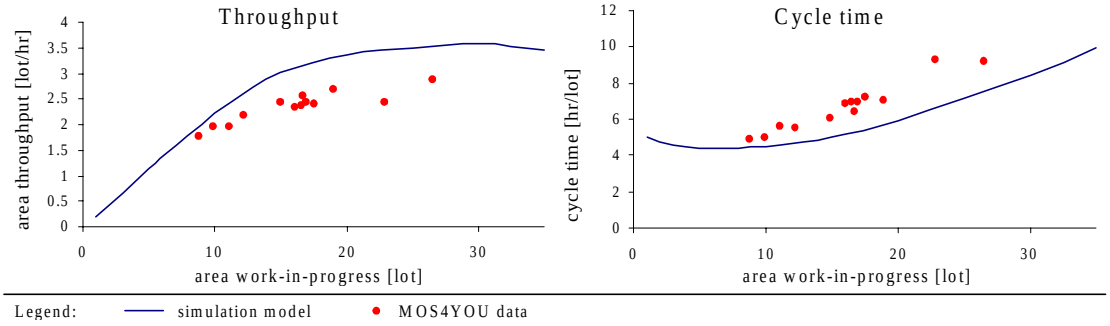

Figure 8.6: Characteristic curves validation.

The shapes of the throughput and cycle time curves of the simulation model closely resemble the fab data. To investigate the effect of different scheduling rules the model is considered a valid representation of the metal area, as far as the goal of this chapter is concerned.

8.3 Scheduling rules

Fab-level scheduling creates a balanced product flow through the fab, see Chapter 6. Fab-level scheduling rules already exist and are applied to lot dispatching in the equipment family buffers. Fab-level rules are based on lot-priority, due date, work-in-progress (WIP) balancing, and First In First Out (FIFO). Fab-level scheduling rules generate a candidate lot list for dispatching.

Area-level scheduling takes care of a balanced use of equipment within the area. In 1998, MOS4YOU used intuitive area-level scheduling rules. Several new rules have been proposed. As the fab-scheduling results in a general order for lot-processing, there remains a freedom in selecting (FS) lots. Area-level scheduling rules can be used on top of global scheduling rules for more detailed lot selection in order to further optimize the fab performance. FS equals the number of TS machines, however, the freedom in applying area-level rules is restricted to the first five lots generated by the fab-level scheduling algorithm. Below the most relevant of the proposed area-level rules are presented.

1. Deposit thick layers on high-speed TS, that is, with two Ti or Al chambers available, and deposit thin layers on low speed TS ,that is, with one available Ti or Al chamber.

2. Prevent conditioning of TS-ND machines by processing identical layers as long as possible.

3. After conditioning a TS-ND

   (a) if the difference betweem the queue lengths of AlTiN and TiTiN lots exceeds 0.5· FS, give preference to processing a lot from the largest queue first,

   (b) otherwise, give preference to processing the lot type - barrier/salicide or metal - with the smallest available equipment capacity.

4. If there is a TS-ND with the E-chamber down and

   (a) a TS-ND is needed for salicide processing, prefer to process salicide on the one with the E-chamber down.

   (b) a TS-ND is needed for metal processing, prefer to process metal on the one with the E-chamber down.

5. Give preference to processing salicide lots on TS-BD with the E-chamber down.

To evaluate the effect of the scheduling rules on the performance, the rules are implemented in the model specified in Section 8.2. Scheduling is conducted in the buffer at each equipment family, just before a lot is dispatched. At that moment, the scheduler has the highest possible amount of lots to select from and the most recent information about the equipment. The rules are verified to ensure a correct behavior.

Table 8.2: Results of the scheduling rules experiments.

Rule	Main [-]	Interaction [-]	Criterion [-]	Accept
1	1.9	1.0	2.9	Yes
2	7.9	-2.2	5.7	Yes
3	2.0	-0.6	1.4	Yes
4a	1.6	-2.1	-0.5	No
4b	-0.1	0.1	0.0	No
5	1.2	1.5	2.7	Yes

8.4 Design of experiments

The defined rules need to be evaluated on their effects on the performance of the fab and the area. The strategy used for this evaluation is based on Full Factorial Analysis (FFA) [Montgomery, 1997]. Every rule can be switched on or off within the scheduling algorithm. A complete FFA results in 2^n simulations, where n is the number of scheduling rules to investigate. However, for a first evaluation, the interaction effects between all rules need not be known, but only the interaction for the one observed compared to the set of all the other rules. As an example, the interaction effect between rule X and Y is not the most important for evaluation of rule X, but the interaction between X and the set containing the rest of the rules is. This FFA reduction results in $2(n+1)$ simulations.

The reduced FFA calculates a normalized main- and interaction effect for every rule and for every performance indicator. The evaluated indicators are area throughput, layer cycle time, and layer cycle time deviation. Each indicator for a simulated set of rules is normalized to its unscheduled value. A rule is accepted for scheduling if the total sum of its effect on those indicators is positive. The reduced FFA is conducted by simulating a set of arrival patterns at the metal area with the model. Additionally, the set of accepted rules is simulated under extreme circumstances, for example an increased workload and a disturbed arrival pattern.

8.5 Experimental results and discussions

The experiments proposed in Section 8.4 produce the results as shown in Table 8.2. This table presents the sum of the main- and interaction effects summed over the observed area variables for the rules presented in Section 8.3.

As Table 8.2 shows, Rules 4a and 4b are not accepted for scheduling by the evaluation criterion. Those rules have no positive effects on the area's performance. All other presented rules are accepted for scheduling.

The experiment on the evaluation of the scheduling rules leads to a set of accepted rules. Figure 8.7 depicts the effects of the rules on the cycle time of the three layers deposited on the TS. It shows that the different configured TS - the TS-ND types - are better utilized due to these scheduling rules. The rules decrease the negative influence of recipe non-availability by decreasing the difference in utilization of the three different cluster tools configurations. Scheduling increases the maximum allowed lot start level from 17 to 18 lots per day.

The simulations show that not only the performance of the TS, but also that of the area increases. For the average arrival pattern, the accepted rules result in a decrease of:

- 9% of the average layer cycle time,
- 8% of the deviation of the layer cycle time,
- 7% of the area's workload,
- 10% of the deviation of the area's workload, and
- 8% of layer-change maintenance parts on the TS.

Using these results, the rules have been analyzed to understand their individual behavior. In extreme situations, specific lot types may remain in the buffer too long. Fine-tuning of the rules is needed to create a sort of safety net for those situations. Expanding the scheduling algorithm with the requirement that the first advised lot on the candidate lot list may not be overruled more than five times - equal to 0.5 FS - solves this.

The effective rules - rule 1, 2, 3, and 5 - and the adjustments for extreme situations have been implemented in the metal area of MOS4YOU and have led to a substantial improvement of both area and fab performance. The observed area improvements are in correspondence with the simulation results.

8.6 Overview

The study presented in this chapter shows that performance in semiconductor fabs can be improved significantly using area-level scheduling rules in addition to fab-level rules, especially if cluster tools are subjected to restricted recipe availability.

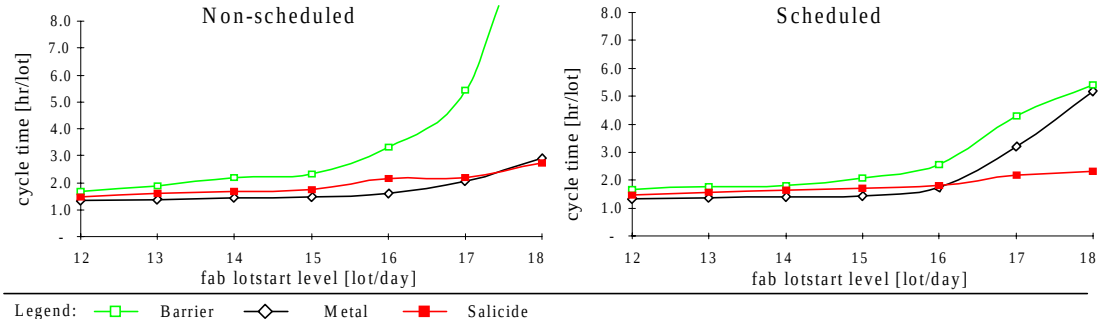

Figure 8.7: Results of simulation.

Machine dedication, that is restricted recipe availability, is corruptive for the performance of not only the area, but also the whole wafer fab. Therefore, attention should be paid to restricting the number of recipes in a wafer fab. Furthermore, recipes should be made available for as many machines as possible.

The positive effects of the scheduling rules can be used to decrease the cycle time of lots as well as to increase the throughput of the fab. If the realized product cycle time is satisfactory, meaning that there is no need to decrease cycle time any further, then the number of outs of the fab can be increased by 4%.

Operational area design uses discrete event simulation as a tool to improve the performance of a specific area, in order to come to a better overall fab performance. As each area in the wafer fab contains a specific main process technology, operational area design provides the ability to consider the specific characteristics of that area and improve the performance.

Chapter 9

Conclusions

In this thesis a method has been presented that supports the design process of a multi-process multi-product wafer fab. The method was further developed in order to support the design of the MOS4YOU wafer fab. Two elements indicate the need for a structured design approach. First, the market requires a larger variety of product types while the life cycle of the product types becomes shorter. To manufacture a large number of different integrated circuits in one fab with a short time to market, high flexibility is required. Flexibility is obtained when the wafer fab is able to adapt quickly to the introduction of new products and process technologies. However, from a cost perspective view, flexibility should be restricted. Cost effective production is a must in order to cope with the technology developments and is achieved by maximizing yield, throughput, and utilization. Second, literature does not provide an approach that adequately supports the design of a multi-process multi-product wafer fab. The MOS4YOU wafer fab was designed using the proposed design approach.

Design approach

Design literature provides a large amount of detailed design methods developed for specific design problems rather than generally applicable methods for designing production systems. Furthermore, designing production systems is treated as a static problem and the design process is finished before the actual building starts. In practice, production systems evolve throughout their life cycle due to ever changing market demands. As a result the design and redesign process should continue as long as the production system is used. Finally, in research as well as in practice, production planning and control has been viewed in isolation from the manufacturing system design. However, as manufacturing and control show large interaction, the quality of the production system design can be enhanced by considering them

simultaneously. Future wafer fabs cannot afford to start defining their production planning and control concept after the fab has been built and started up.

The proposed design method can be used throughout the specification and utilization phase of the production system. It supports initial design as well as redesign of a utilized production facility and regards the design of both manufacturing system and control system, where it affects the manufacturing system. The design method consists of an approach with an accompanying set of tools. The approach defines which decisions have to be made and at what time they have to be made. The structure of the approach is generally applicable for a large class of production systems. The method consists of four phases: (1) definition of objectives and constraints, and (2) design of architecture, (3) resources, and (4) operations.

The design tools support the process of decision making. The set of tools used in this thesis, makes the method fit for the design of multi-process multi-product wafer fabs only.

Objectives and constraints

The design process starts with identifying design objectives and constraints. The design objectives are deduced from the business objectives of a company. During the design process, these objectives and constraints remain subject of discussion. The business objective of the MOS4YOU wafer fab is to manufacture integrated circuits on wafers for the consumer industry. The multi-process multi-product wafer production can be characterized by:

- Fast technology development (one new process technology per year),
- Relatively short process life cycles (three years per process technology), and
- Different operations (a 0.35 μm process flow requires 400 operations).

The products can be characterized by:

- Short life cycles (six months per product type),
- Varying yields, and
- Relatively long cycle times (60 to 90 days).

The financial objective is to generate return on assets. As a result, the main design objectives for MOS4YOU are:

- High yield ($> 97\%$),
- High output (originally 20.000 wafers per month in a 0.5 μm process),
- High utilization ($> 80\%$),
- Short cycle times (< 2 days per mask layer), and
- High delivery performance ($> 99\%$).

Design of architecture

In the design of architecture phase, the structure of the manufacturing system and the control system of MOS4YOU was designed. The boundary conditions for this phase were reflected in the objectives and constraints. The re-entrant character of the process flows can best be captured in a hybrid functional architecture. A hybrid functional architecture groups main process technologies by their own technology and places auxiliary technologies at those locations where they are needed. The choice of combining a process focussed architecture with local flow lines allows for a proper control of the production system. The production control architecture is designed to be robust for randomness and variability. A balance strategy that balances work-in-progress fits this architecture best. Additionally, the architecture of the production system is influenced by the transport and storage architecture. To be able to cope with the many different sources of variability, the transport and storage architecture must be flexible. A transport concept which is partially manual and partially automated was chosen.

The main design decisions for the second phase are:

- Re-entrant flow line architecture,
- Main process technologies grouped by their own technology,
- Auxiliary process technologies distributed where needed,
- Balancing of work-in-progress using sequencing rules,
- Automated transport between areas, and
- Manual transport within each area.

Designing a production architecture is an essential phase in the design process, as it strongly influences the controllability of the manufacturing system, the structure of the control system, the number and type of resources, and the layout. The investments and cost of operation can be influenced to a large extent in the beginning of the design process. The further the design process has advanced, the less the investments and cost of operation can be influenced. A proper design of architecture is required to design a cost-effective wafer fab.

Design of resources

In the design of resources phase, the design decisions regarding the architecture were quantified. The following resources can be identified: process equipment, metrology equipment, transport and storage equipment, facilities, operators and maintenance people. A decoupled queueing line model was developed and used to determine the number of required resources. Taking into account queueing theory relations in the model is a new element for static capacity determination methods. Given the required number of resources, the layout can be constituted. Due to the ever changing market demands and process technologies, static capacity determination is an activity that is performed continuously, and not just at the beginning of the design process. This imposes flexibility on the layout.

The design decisions of the third phase are:

- Required number of resources, and
- Layout design plan.

Design of operations

In the design of operations phase, the resource design was refined. This last phase considers development and improvement of the operating procedures.

The production control concept for MOS4YOU was developed and a tool to support this concept was brought in place. To be able to cope with the relatively high level of variability, the production control concept consists of a release policy and a set of sequencing rules. Using capacity calculations, boundaries for the work in progress are determined. These boundaries specify the lower and upper limit for the work in progress. Within those boundaries, wafers are released at a constant rate. If the boundaries are exceeded, the release rate is adjusted until an acceptable situation has again been reached. Once on the shop floor, wafers are processed according to

a specific sequence of recipes and lots are dispatched for process steps in an order aiming for maximum flow line balance. The recipe capability indicates two things: what recipes can be processed on what machines and what are the preferences in case a recipe can be processed on multiple machines. Flow line balance sequencing results in an optimal division of work in progress over the machines. Here, sequencing is preferred above scheduling, because the amount of controlled and uncontrolled variability is substantial. Recipe capability together with flow line balancing result in maximum throughput and maximum utilization.

The design decisions of the fourth phase are:

- Releasing wafers at a constant rate, but adjusting release rate if predefined boundaries are exceeded,
- Process wafers according to the recipe capability, and
- Sequencing rules to maximize flow line balance.

The final step in the design of operations concerns improvement of the working wafer fab. In this last step, a detailed analysis was made of a part of the MOS4YOU wafer fab, using discrete event dynamic simulations. The detailed analysis takes into account all major sources of variability. The improvements focussed on the separate areas, since the different areas each have their own characteristics. Refinement of the operating procedures or sequencing rules can be done by taking the specific characteristics of an area into account.

Summarizing

Summarizing it can be said that this thesis has shown that use of a structured design approach and use of the right tools has helped to create an initial design and to improve the final design of a wafer fab. Although it is hard to give exact figures on the contribution of the described approach, handling complexities of this kind only can be obtained by working in a systematic manner.

Chapter 10

Recommendations

In the previous chapters the approach and the tools to design an IC wafer fab have been described. Furthermore, the relevant design decisions for the MOS4YOU case were presented. In this final chapter recommendations and reflections are discussed.

The design presented in this thesis was used to support the design of the MOS4YOU wafer fab. One of the difficulties in designing a wafer fab is the uncertainty of future developments. MOS4YOU was originally designed to produce 5000 0.5 μm wafers per week. Currently, the process mix has changed to 0.35 μm and 0.25 μm, while the output should exceed the 5000 wafers per week. To cope with the more complex process mix and the increase in output, extra production floor space was created. This required a drastic layout change, causing a complete area to be moved. Furthermore, the concept of physically grouping machines that share main process technologies had to be violated. Layout and transportation systems should be flexible towards changing requirements.

Designing in practice was an iterative process; although the structure of the thesis suggests a chronological order, the different design phases have impacted each other both top-down and bottom-up. The described method presents a way for balancing the wafer fab as a whole and for improving areas in detail. The design of architecture and design of resources phase have evolved completely. The approach and the accompanying design tools have proven their use during the design of MOS4YOU. Also the improvement approach of the design of operations has been applied with success to different areas of MOS4YOU and resulted in improved floor layouts, improved working procedures, and improved sequencing rules. The part of design of operations that focuses on production control, however, shows perspectives for further development. In practice, there are wafer fabs that consider a global production control strategy, however, this is often not quantified and only implemented using manual intervention. In this thesis a global approach was chosen. The concept of balancing work in

progress was investigated and implemented in a very rudimentary way. The reason being that although semiconductor industry is advanced in its process technologies, it is conservative in other fields. Using a prototype sequencing tool, the concept was evaluated and adapted. After that, it had to be integrated into the manufacturing execution system. Theory which balances the wafer fab be refined in order to deal with the variability sources. The part of design of operations, where the production control concept is specified, should be considered as a starting point for this refinement. Global production control concepts, that consider the whole wafer fab, are seldom discussed in literature. New production control concepts should be simple, effective, understandable, and explainable. Taking the described control concept of this thesis as a starting point, research is being done on the use of standard control theory to improve the production control further by Lefeber [2001].

Besides the production planning control concept, attention should also be paid to the supply chain management strategy. Customers require functionally correct dies. During the wafer manufacturing process dies are produced on wafers. Wafer fabrication is characterized by highly uncertain yields. Process errors, machine failures, and operator mistakes are examples that have an impact on the yield. Managing the supply chain is crucial to the success of both the end customer and the manufacturer of the dies. Supply chain management starts with planning which fab will manufacture which wafers, and from monitoring the ordering process up to delivery and financial compensation of the finished dies. Supply chain management concepts for ICs have already been subject of studies for many years, but have not lead to industry standards. With the emerging generic ERP packages, more and more attention is focussed on generic supply chain management solutions. However, due to the specific characteristics, concepts from semiconductor industry can not easily be implemented in those generic solutions. Therefore, several suppliers have developed semiconductor specific solutions, such as Hadavi & Shahraray [1998] and I2 [2001]. The commercial available semiconductor specific solutions are mostly MRP based [Bhatnagar, Chandra & Goyal, 1992]. Development of these solutions should be preceded by a thorough study of the desired planning concept. The lack of these studies causes that suggested solutions still require lots of customization before they can be implemented. However, until recently little attention was paid to it in literature. Approaches were always practical, but no theoretical foundation was available. Therefore, more research towards supply chain management is recommended.

Machines are subject to many sources of variability. Well known sources of variability are machine failures, rework, setups, and operator availability. Especially variability of high-utilized machines causes long queueing times. Since queueing time is the most important cycle time loss, minimizing it is of great interest. Without changing utilization, reducing variability can decrease queueing time. Therefore, identification and reduction of the main sources of variability is required. In semiconductor

industry, no measures of variability in operations are used. Recently, the overall equipment efficiency (OEE) has been introduced by SEMI. This measure is based on mean values with respect to availability, productivity, and yield. However, the OEE does not cover the variability of the operations. Both means and variances should be included to make the correct conclusions on how well machines are performing. Sturm, Silvi, Frauenhoffer & Treiber [1999] observed that it is impossible to measure each individual source of variability. An approach is suggested to measure all effects of variability as a whole in one single cycle time distribution. Hopp & Spearman [2000] introduce the effective process time (EPT) define it as the time seen by lots from a logistical point of view. However, both studies do not give a practical usable approach to actually *measure* the EPT. Basically the EPT is the total amount of time a lot could have been or actually was being processed on a machine. Jacobs, Etman, Van Campen & Rooda [2001] have proposed a new method to actually compute effective process times from fab data for single lot machines. This method proved to be valuable during an initial case study using data from MOS4YOU. A generalization of this new concept towards different types of machines would provide a very powerful new measure for cycle time monitoring and reduction in semiconductor manufacturing.

The design of wafer fabs in the (near) future will remain challenging, because internal transport of wafers will be automated more and more. This will have an impact on the architecture and layout of the fabs. The ongoing automation also requires the production control concept to be more and more refined. Nowadays, variability, that seems inevitable for semiconductor industry, is often 'buffered' by operators. For example, if a problem with one of the reticle stockers causes some types of reticles not to be available for production, operators choose those lots of which the reticles are available to be processed. Because automation restricts flexibility, the design of resources and the design of operations should be thought over even more carefully. Simply installing excessive capacity is not sufficient to make up for this decreasing flexibility, because then the cost effectiveness would be jeopardized.

Bibliography

Apple, J.M. [1963], *Plant layout and material handling*, Wiley, New York.

Arends, N.W.A. [1996], *A systems engineering specification formalism*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

Atherton, L.F., and R.W. Atherton [1995], *Wafer fabrication: factory performance analysis*, Kluwer Academic, Boston.

Balleggi, M. [1998], Analysis and optimization of the MOS4YOU dielectric area, Master's thesis, Eindhoven University of Technology, The Netherlands, SE420194.

Bhatnagar, R., P. Chandra, and S.K. Goyal [1992], Models for multi-plant coordination, *European Journal of Operational Research* **41**, 141–160.

Bitran, G.R., and D. Tirupati [1988a], Development and implementation of a scheduling system for a wafer fabrication facility, *Operations Research* **36**, 377–395.

Bitran, G.R., and D. Tirupati [1988b], Planning and scheduling for epitaxial wafer production facilities, *Operations Research* **36**, 34–49.

Boothroyd, G., and P. Dewhurst [1983], *Design for assembly. A designer's handbook*, University of Massachusetts.

Brandts, L.E.M.W. [1993], *Design of industrial systems*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

Browne, J., J. Harhen, and J. Shivnan [1996], *Production management systems: an integrated perspective* (second ed.), Addison-Wesley, Harlow.

Burbidge, J.L. [1971], Production flow analysis, *The Production Engineer* **4**, 139–152.

Burbidge, J.L. [1989], *Production flow analysis: for planning group technology*, Oxford series on advanced manufacturing, Clarendon Press, Oxford.

Buzacott, J.A., and J.G. Shanthikumar [1993], *Stochastic models of manufacturing systems*, Prentice Hall, Englewood Cliffs.

Castrucci, P. [1995], The future fab: changing the paradigm, *Solid State Technology* **1**, 49–56.

Dhudshia, V.H. [1997], SEMI E10 - equipment reliability, availability and maintainability, *Semiconductor International* **6**, 167–174.

Eijsvogels, K.J. [1998], Dynamics and scheduling of the MOS4YOU wafer fab, Master's thesis, Eindhoven University of Technology, The Netherlands, SE420174.

Fey, J.J.H. [2000], *Design of a fruit juice blending and packaging plant: Riedel's production facility*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

Goldratt, E.M., and J. Cox [1984], *The goal: a process of ongoing improvement*, North River Press, New York.

Govil, M.K., and M.C. Fu [1999], Queueing in manufacturing: a review, *Journal of Manufacturing Systems* **18**, 214–240.

Graves, R.J., J.M. Konopka, and R.J. Milne [1995], Literature review of material flow control mechanisms, *Production Planning and Control* **6**, 395–403.

Graves, S.C. [1981], A review of production scheduling, *Mathematics Operations Research* **29**, 646–675.

Groover, M.P. [1996], *Fundamentals of modern manufacturing: materials, processes, and systems*, Prentice Hall, Englewood Cliff.

Haagh, P.A.M., A.U. Wilkens, E.J.J. van Campen, J.E. Rooda, and H.J.A. Rulkens [1998], Application of a layout design method to a dielectrics deposition area in a 300 mm wafer-fab, *International Symposium on Semiconductor Manufacturing*, Tokyo, 69–72.

Hadavi, K.C., and M. S. Shahraray [1998], *Factory order release control and evaluation*, Technical report, Paragon Management Systems, http://www.paragonms.com.

Hammer, M., and J. Champy [1993], *Re-engineering the corporation: a manifest for business revolution*, Harper Business, New York.

Haupt, R. [1989], A survey of priority rule-based scheduling, *OR Spektrum* **11**, 3–16.

Hayes, R., and S. Wheelwright [1984], *Restoring our competitive edge: competing through manufacturing*, John Wiley and Sons, New York.

Hopp, W.J., and M.L. Spearman [2000], *Factory physics: foundations of manufacturing management* (second ed.), Irwin, Boston.

Hsih, H.W., and H.C. Wu [1998], Equipment loading dynamic forecasting system, *International Symposium on Semiconductor Manufacturing*, Tokyo, 83–86.

i2 [2001], *Product documentation*, Technical report, i2, http://www.i2.com/.

Jacobs, J.H. [1998], *Modeling of the wafer fabrication at MOS4YOU*, Research report, Eindhoven University of Technology, The Netherlands, SE420214.

Jacobs, J.H., L.F.P. Etman, E.J.J. van Campen, and J.E. Rooda [2001], Quatifying operational time variability: the missing parameter for cycle time reduction, *IEEE/SEMI Advanced Semiconductor Manufacturing Conferenence*, Munich, (accepted as invited paper).

Jansen, R. [1998], The Crystal Cell FAB - A design innovation for semiconductor manufacturing facilities, *Semiconductor Fabtech*, 93–98.

Johnston, R.B. [1995], Making manufacturing practices tacit: a case study of computer-aided production management and lean production, *Journal of the Operational Research Society* **46**, 1174–1183.

Kumar, P.R. [1993], Re-entrant lines, *Queueing Systems* **13**, 87–110.

Ledder, J. [1999], *Definition of an integrated software architecture for wafer fabs*, Internal report, Philips Semiconductors, The Netherlands.

Lee, H.F. [1997], Performance analysis for automated storage and retrieval systems, *Industrial Engineering* **29**, 15–28.

Lefeber, A.A.J. [2001], *Control of industrial systems*, Technical report, Eindhoven University of Technology, The Netherlands, http://se.wtb.tue.nl/research/control.php3.

Leijsen, R.J. van [1998], Optimization of throughput and cycle time in the lithography area of MOS4YOU, Master's thesis, Eindhoven University of Technology, The Netherlands, SE420150.

Lemmen, B., E.J.J. van Campen, H. Roede, and J.E. Rooda [1999], Clustertool optimization through scheduling rules, *International Symposium on Semiconductor Manufacturing*, Santa Clara, 99–102.

Little, J.D.C. [1961], A proof for the queueing formula $l = \lambda w$, *Operations Research* **9**, 383–387.

Lozinski, C., and C.R. Glassey [1988], Bottleneck starvation indicators for shop floor control, *IEEE Transactions on Semiconductor Manufacturing* **1**, 147–153.

Maly, W. [1987], *Atlas of IC technology: an introduction to VLSI processes*, The Benjamin / Cunnings Publishing Company, Menlo Park.

McIntosh, S. [1997], MOS4YOU - a record breaking fab, *Future Fab International* **3**, 123–128.

Montgomery, D.C. [1997], *Design and analysis of experiments*, John Wiley & Sons, New York.

Montgomery, D.C., and G.C. Runger [1994], *Applied statistics and probability for engineers*, John Wiley & Sons, New York.

Muther, R. [1973], *Systematic layout planning*, CBI Publishing, Boston.

Naumoski, G., and W. Alberts [1998], *A discrete-event simulator for systems engineering*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

Neudorff, J. [1999], Static capacity analysis using microsoft visual basic, *International Conference on Semiconductor Manufacturing Operational Modeling and Simulation*, San Fransisco, 207–212.

Nichols, K. [1992], Design for quality and reliability, *Journal of Engineering Design* **3**, 139–148.

Noben, R. [2000], Optimization of throughput and cycle time in the dry etch

area of MOS4YOU, Master's thesis, Eindhoven University of Technology, The Netherlands, SE420174.

Panwalker, S.S., and W. Iskander [1977], A survey of scheduling rules, *Operations Research* **25**, 45–61.

Pierce, N.G., and R. Stafford [1994], Modeling and simulation of material handling for semiconductor wafer fabrication, *Proceedings of the 1994 winter simulation conference*, 900–906.

Pillai, D. [1990], Material handling automation for wafer fabrication facilities, *IEEE/CPMT International Electronics Manufacturing Technology Symposium*, 277–286.

Pinedo, M. [1995], *Scheduling: Theory, algorithms, and systems*, Prentice Hall, Englewood Cliffs.

Pollitt, C. [1998], Quantifying capacity loss associated with stafuture fab internationalng in a semiconductor manufacturing line, *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, Boston, 125–128.

Robinson, J., J. Fowler, and E. Neacy [2000], Capacity loss factors in semiconductor manufacturing, *Semiconductor Capacity Planning Process Paper*, http://www.fabtime.com/bibliogr.htm.

Rooda, J.E. [2000], *Modelling industrial systems*, Lecture notes, Eindhoven University of Technology, The Netherlands, http://se.wtb.tue.nl/documentation/.

Rulkens, H.J.A. [1997], *Analysis and optimization of the lithography area of MOS4YOU*, Post-master's thesis, Eindhoven University of Technology, The Netherlands.

Rulkens, H.J.A., E.J.J. van Campen, J. van Herk, and J.E. Rooda [1998], Batch size optimization of a furnace and pre-clean area by using dynamic simulations, *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, Boston, 439–444.

Schaller, R.R. [1997], Moore's law: past, present, and future, *IEEE Spectrum* **6**, 53–59.

Shingo, S. [1981], *Study of "Toyota" Production System: From an Industrial Engineering Viewpoint*, Japan Management Association, Tokyo.

Sitter, L.U. de [1994], *Synergetisch produceren*, Van Gorcum, Assen, In Dutch.

Snowden, L.S., and J.C Ammons [1988], A survey of queueing network packages for the analysis of manufacturing systems, *Manufacturing Review* **1**, 14–25.

Sohl, D.L., and P.R. Kumar [1995], Fluctuation smoothing scheduling policies for multiple process flow fabrication plants, *IEEE/CPMT International Electronics Manufacturing Technology Symposium*, 190–198.

Sturm, R., T. Silvi, F. Frauenhoffer, and T. Treiber [1999], A simulation model for advanced release and order planning, *Future Fab International* **6**, 71–74.

Trines, W.G. [1997], Assembly line balancing: theory and applications, Master's

thesis, Eindhoven University of Technology, The Netherlands, SE420141.

Uzsoy, R., C. Lee, and L.A. Martin-Vega [1992], A review of production planning and scheduling models in the semiconductor industry part I: system characteristics, performance evaluation and production planning, *IIE* **24**, 47–60.

Uzsoy, R., C. Lee, and L.A. Martin-Vega [1994], A review of production planning and scheduling models in the semiconductor industry part II: shop-floor control, *IIE* **26**, 44–55.

Wein, L.M. [1988], Scheduling semiconductor wafer fabrication, *IEEE Transactions on semiconductor manufacturing* **1**, 115–130.

Wein, L.M., and P.B. Chevalier [1992], A broader view of the job-shop scheduling problem, *Management Science* **38**, 1018–1033.

Wiendahl, H.P. [1995], *Load-oriented manufacturing control*, Springer-Verlag, Berlin.

Witte, J.D. [1996], Using static capacity modelling techniques in semiconductor manufacturing, *IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, Boston, 31–35.

Wolf, S., and R.N. Tauber [1986], *Silicon processing for the VLSI era: Volume 1 - process technology*, Lattice Press, Sunset Beach.

Womack, J.P., D.T. Jones, and D. Roos [1990], *The machine that changed the world*, Harper Perennial, New York.

Wood, S.C., S. Tripathi, and F. Moghadam [1994], A generic model for cluster tool throughput time and capacity, *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, Boston, 194–199.

Wu, W.F., J.L. Lang, and J.T. Liao [1998], Static capacity checking system with cycle time considered, *International Symposium on Semiconductor Manufacturing*, Tokyo, 307–310.

You, R-C, C-R Weng, Y-C Chou, H. Wu, and L-C Lu [1999], A tool portfolio planning methodology for semiconductor wafer fabs, *International Symposium on Semiconductor Manufacturing*, 11–14.

Appendix A

Wafer fab model

In this appendix the wafer fab model that was used in Chapter 6 is described. First, a global description is given by dividing the wafer fab into separate components. Then, the specification of each component is described. Finally, the model of the control system is presented. The full source code is not included here. The process routes are not included due to confidentiallity.

The specification language χ [Rooda, 2000] was used to model the wafer fab. A χ specification is object-oriented, which implies that the model is divided into parallel components. Each process captures the behavior of one specific component. For the model, the following components were distinguished:

- Environment,
- Manufacturing system, and
- Control system.

A graphical representation of the model is depicted in Figure A.1. The environment is modelled by process G. This process allows starting new lots into the system and consuming processed lots. The manufacturing system consists of process T and multiple processes E. Process T distributes the lots over the different machine groups. Transport times are considered to be negligible compared to the waiting and process times. Each process E represents a machine group and consists of a buffer process B and several identical machine processes Eqt, this is depicted in Figure A.2. The processes are connected through channels, which represent the material flow through the wafer fab. Upon a request from an idle machine, the buffer sends lots to this machine.

The control system influences the dynamic behavior of the wafer fab by:

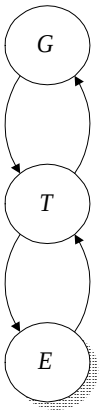

Figure A.1: Architecture of the wafer fab model.

- Releasing,
- Routing, and
- Scheduling.

The control system is not modelled as a process, it is modelled by dedicated functions. The release rules are modelled in a function that is used by process G. The scheduling rules are modelled in functions used by process B. The routing is executed in process T as it distributes the lots over the machine groups according to the process sequence.

A.1 Environment

The behavior of the environment is modelled in process G. Lots are created and released according to a release rule. After being released, the lots are sent to the manufacturing system. Fully processed lots are sent back to process G. Below, an abbreviated specification of process G is presented. Finished lots are received through channel a. Release function nxt_strt_time is used to determine the next start time of a lot. Then a new lot is created using function create_nw_lot and sent to the transporter through channel b.

```
proc G(a : ?lot, b : !lot) =
[[ t : time, x : lot
| t := 0
; *[ true; a?x      --> skip
   [] true; Δt − τ --> b!create_nw_lot(τ); t := τ + nxt_strt_time
   ]
]]
```

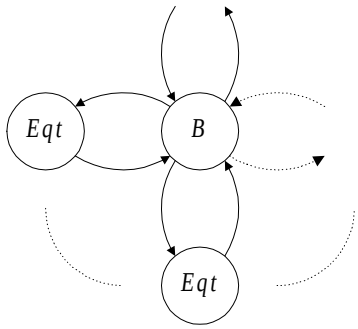

Figure A.2: Equipment family model architecture.

A.2 Manufacturing system

The manufacturing system consists of process T and multiple processes E. The transport between the machine groups is modelled by process T. A machine group is modelled by process E. The structure of a machine group is depicted in Figure A.2. It consists of a buffer process B and several identical machine processes Eqt.

Process Eqt

Machines can be categorized into several types, according to their behavior. In the model two types of behavior are distinguished: batch machines and cascade machines. Using these two types, the behavior of all machines can be captured.

Batch machines

Batch machines can process one batch of lots at a time. Most of the machine types can be captured in this category. Process *Eqt* sends a signal to the buffer that it is idle and ready to receive and process a batch of lots. This is done using channel s. Then the machine is ready to receive a batch through channel a. After the buffer has sent a batch it will be processed. The processing time is determined using function proc$_{\text{t}}$ime. Then, the batch is sent back to the buffer through channel b. The abbreviated specification is given below.

$$
\begin{array}{l}
\mathsf{proc}\ \mathit{Eqt\_batch}(a : ?\,\text{batch},\ b : !\,\text{batch},\ s : \tilde{}\text{void}) = \\
[\![\ \mathit{free} : \text{bool},\ x : \text{batch},\ t : \text{time} \\
\mid\ \mathit{free} := \text{true} \\
;\ *[\ \mathit{free}\ \ ;\ s\tilde{}\qquad \longrightarrow a\,?\,x;\ \mathit{free} := \text{false};\ t := \tau + \text{proc\_time}(x) \\
\quad [\!]\ \neg \mathit{free};\ \Delta\, t - \tau \longrightarrow b\,!\,x;\ \mathit{free} := \text{true} \\
\quad]\!] \\
]\!]
\end{array}
$$

Cascade machines

An extension to the batch machines is formed by cascading machines. These type of machines can process multiple batches in cascading sequence. The machine receives a batch and after some period of time (but before the batch has been completely processed) the next batch can be received. The abbreviated specification is given below.

$$
\begin{array}{l}
\mathsf{proc}\ \mathit{Eqt\_cascade}(a : ?\,\text{batch},\ b : !\,\text{batch},\ s : \tilde{}\text{void},\ n : \text{nat},\ \mathit{ti} : \text{time}) = \\
[\![\ x : \text{batch},\ \mathit{xs} : (\text{batch} \times \text{time})^*,\ t : \text{time},\ \mathit{free} : \text{bool} \\
\mid\ \mathit{xs} := [\,];\ \mathit{free} := \text{true} \\
;\ *[\ \text{len}(\mathit{xs}) < n \wedge \mathit{free};\ s\tilde{} \\
\qquad \longrightarrow a\,?\,x;\ \mathit{xs} := \mathit{xs} +\!\!+ [x, \tau + \text{proc\_time}(x)];\ \mathit{free} := \text{false};\ t := \tau + \mathit{ti} \\
\quad [\!]\ \neg \mathit{free};\ \Delta\, t - \tau \longrightarrow \mathit{free} := \text{true} \\
\quad [\!]\ \mathit{len}(\mathit{xs}) > 0;\ \Delta\, \mathit{hd}(\mathit{xs}).1 - \tau \longrightarrow b\,!\,\mathit{hd}(\mathit{xs}).0;\ \mathit{xs} := \mathit{tl}(\mathit{xs}) \\
\quad]\!] \\
]\!]
\end{array}
$$

Down time

Machine maintenance or failures cause non available time for production. In the model, it is assumed that no lots enter or leave the machine while it is unavailable. Furthermore, the lot quality is not affected. To model maintenance and failure, two parameters are used: the time between failure (TBF) and the time to repair (TTR). A Weibull distribution is used to model the time between maintenance or failure of the machines. This distribution is widely used because it has large variety of shapes that enable to fit many kinds of data. The probability density function is given by [Montgomery & Runger, 1994]:

$$f_X(x;\alpha,\beta) = \frac{\beta}{\alpha}(\frac{x}{\alpha})^{\beta-1}e^{-(x/\alpha)^{\beta}}.$$

The Weibull distribution depends upon two parameters: the shape factor β and the scale factor α. These factors were determined from actual fab data, assuming that all machines within a group show exactly the same behavior.

The description of process *Eqt* was extended to implement the down time behavior. Two distributions, *dtbf* and *dttr*, were added. Process *Eqt* can experience down time whenever it is idle or processing. During a down event, no lots can enter or leave. Furthermore, the process time of the lots is increased by the duration of the down event. The abbreviated specification is given below. Process *Eqt_cascade* can be extended in the same way.

---

proc *Eqt_batch*(a : ? batch, b : ! batch, s : ˜void, $\alpha_0, \beta_0, \alpha_1, \beta_1$: real) =
⟦ x : batch, tbf, ttr : time, $dtbf, dttr$: → time, $down, free$: bool
| $dtbf := wei(\alpha_0, \beta_0)$; $dttr := wei(\alpha_1, \beta_1)$; $tbf := \tau + \sigma\, dtbf$
; $down :=$ false; $free :=$ true
; ∗[$\neg down \wedge free$; s˜ ⟶ $a\,?\,x$; $free :=$ false; $t := \tau +$ proc_time(x)
⟦ $\neg down \wedge \neg free$; $\Delta\, t - \tau$ ⟶ $b\,!\,x$; $free :=$ true
⟦ $\neg down$; $\Delta\, tbf - \tau$ ⟶ $ttr := \tau + \sigma\, dttr$; $down :=$ true
⟦ $down$; $\Delta\, ttr - \tau$ ⟶ $tbf := \tau + \sigma\, dtbf$; $down :=$ false
]
⟧

Process B

Lots are placed in a queue in front of a machine group. If one of the machines from the group send a signal (or request for a new batch), the scheduling rule is executed to determine which lot (or lots) to send to the machine. Lots that are finished processing are sent back to the transporter.

$$
\begin{array}{l}
\mathsf{proc}\ B(a : ?\,\mathrm{lot},\ b : !\,\mathrm{lot},\ c : (?\,\mathrm{batch})^k \\
\qquad\quad ,\ d : (!\,\mathrm{batch})^k,\ s : (\tilde{}\mathrm{void})^k,\ prmtr : \mathrm{eqt\_char}) = \\
[\![\ x : \mathrm{lot},\ y : \mathrm{batch},\ xs, ys : \mathrm{buffer} \\
|\ xs := [\,];\ ys := [\,] \\
;\ *[\ \mathrm{true};\ a\,?\,x \longrightarrow xs := xs +\!\!+ [x] \\
\quad [\!]\ \mathrm{true};\ c.i\,?\,y \longrightarrow ys := ys +\!\!+ y \\
\quad [\!]\ j : \mathrm{nat} \leftarrow 0..k : \mathrm{len}(xs) > 0;\ s.j\tilde{} \\
\qquad \longrightarrow \langle y, xs\rangle := \mathrm{get\_batch}(xs, prmtr);\ d.j\,!\,y \\
\quad [\!]\ \mathrm{len}(ys) > 0;\ b\,!\,\mathrm{hd}(ys) \longrightarrow ys := \mathrm{tl}(ys) \\
\quad] \\
]\!]
\end{array}
$$

Process T

Process T receives lots from the generator and from the machines. Then it determines where these lots have to go to next and the lots are sent to this destination.

$$
\begin{array}{l}
\mathsf{proc}\ T(a : (?\,\mathrm{lot})^k,\ b : (!\,\mathrm{lot})^k) = \\
[\![\ free : \mathrm{bool},\ xs : \mathrm{buffer}^k,\ x : \mathrm{lot},\ i : \mathrm{nat} \\
|\ free := \mathrm{true};\ j : \mathrm{nat} \leftarrow 0..k : xs.j := [\,] \\
;\ *[\ j : \mathrm{nat} \leftarrow 0..k : \mathrm{true};\ a.i\,?\,x \\
\qquad \longrightarrow \langle x, i\rangle := \mathrm{upd\_rtng}(c);\ xs.i := xs.i +\!\!+ [x] \\
\quad [\!]\ j : \mathrm{nat} \leftarrow 0..k : \mathrm{len}(xs.j) > 0;\ b.j\,!\,\mathrm{hd}(xs.j) \longrightarrow xs.j := \mathrm{tl}(xs.j) \\
\quad] \\
]\!]
\end{array}
$$

A.3 Control system

The control system performs three activities: releasing, routing, and scheduling. The release rate is modelled as a function in process G. Here the amount of lots and the release pattern can be specified. Five different release strategies were implemented.

The routing of the lots was performed by process T, where it is determined in what queue a lot has to be placed. Three different process flows were taken into account.

The scheduling rule was implemented in a function in process B. Three types of scheduling rules were implemented: FIFO, EDD, and LCT. Below the function get_batch is specified using EDD sorting.

$$
\begin{aligned}
&\textsf{func}\ \text{get\_batch}(xs : \text{buffer},\ prmtr : \text{eqt\_char}) \rightarrow \langle \text{batch} \times \text{buffer} \rangle = \\
&[\![\ x : \text{lot},\ ys : \text{buffer} \\
&\mid\ xs := \text{sort\_edd}(xs) \\
&;\ ys := [x \mid x \leftarrow xs,\ x.recipe = \text{hd}(xs).recipe] \\
&;\ ys := \text{take}(ys, prmtr.btch\_sz) \\
&;\ \uparrow \langle ys, xs - -ys \rangle \\
&]\!]
\end{aligned}
$$

A.4 Data

Table A.1 presents the Weibull parameters for all machine groups in the model. Table A.2 contains a part of the routings that were used.

Table A.1: Weibull parameters for MTBF and MTTR.

Mach grp	**MTBF**		**MTTR**	
	β	α	β	α
01	0.54	4100	0.33	1260
02	1.00	1657	0.54	753
03	0.87	13870	0.51	428
04	0.59	4649	0.54	116
05	0.85	7005	1.00	814
06	0.94	5795	0.53	319
07	0.76	3581	0.54	307
08	0.71	2865	0.55	897
09	1.00	4126	0.54	389
10	0.62	8053	0.54	768
11	0.77	5782	0.54	366
12	0.48	3436	0.49	395
13	1.96	16495	1.48	183
14	1.05	14196	0.45	716
15	0.94	29267	0.46	526
16	0.66	23871	0.64	356
17	1.28	23148	0.52	199
18	0.43	5000	0.49	66
19	1.00	6195	0.49	315
20	1.00	6439	0.47	149
21	1.00	12267	0.47	900
22	1.24	42357	1.00	39
23	0.75	1323	0.57	259
24	1.00	2007	0.64	273
25	0.58	3327	0.52	144
26	2.40	39654	0.85	1110
27	1.00	2867	0.47	247
28	0.40	3115	0.39	176
29	0.66	8887	0.46	1003
30	0.77	2676	0.51	642
31	0.69	1854	0.51	173
32	0.75	3342	0.55	201
33	0.70	2383	0.50	238
34	0.49	1847	0.47	95
35	0.79	3165	0.58	124
36	0.75	2083	0.61	176
37	0.70	2052	0.59	202
38	0.63	2206	0.52	106
39	0.45	1358	0.43	265
40	0.59	2599	0.42	106

Table A.2: Process times.

Process 1		**Process 2**		**Process 3**	
Mach grp	**Proc time**	**Mach grp**	**Proc time**	**Mach grp**	**Proc time**
01	0.01	01	0.01	01	0.01
34	0.02	34	0.02	34	0.02
36	0.01	36	0.01	36	0.01
44	0.01	44	0.01	44	0.01
04	0.06	14	0.03	04	0.06
16	0.03	16	0.03	16	0.03
58	0.04	58	0.04	58	0.04
44	0.01	44	0.01	44	0.01
53	0.04	53	0.04	53	0.04
20	0.17	20	0.17	20	0.17
24	0.15	24	0.15	24	0.15
43	0.01	43	0.01	43	0.01
23	0.23	23	0.23	23	0.23
43	0.01	43	0.01	43	0.01
34	0.02	34	0.02	34	0.02
36	0.01	36	0.01	36	0.01
44	0.01	44	0.01	44	0.01
42	0.00	42	0.00	42	0.00
14	0.06	14	0.06	14	0.06
43	0.01	16	0.03	16	0.03
16	0.03	58	0.04	58	0.04
58	0.03	53	0.03	42	0.00
42	0.00	21	0.23	39	0.03
39	0.03	56	0.07	53	0.03
53	0.03	53	0.04	21	0.23
21	0.23	44	0.01	43	0.01
43	0.01	14	0.03	56	0.07
...	...	...	...	...	...

Samenvatting

Het fabriceren van halfgeleiders is een buitengewoon gecompliceerd productieproces. De complexiteit wordt veroorzaakt doordat er enkele honderden bewerkingen nodig zijn om halfgeleiders te fabriceren. Verder wordt er gebruik gemaakt van zeer geavanceerde bewerkingstechnieken, die telkens verder ontwikkeld worden. Dit alles maakt dat het bouwen van een nieuwe halfgeleiderfabriek vandaag de dag een investering van meer dan 1 miljard dollar vergt, terwijl de kosten van het in bedrijf houden enkele honderden miljoenen dollars per jaar bedragen. Het ontwerp van een fabriek is bepalend voor de te maken investeringen en de te verwachten kosten van het in bedrijf houden. Daarom is een goed ontwerp van cruciaal belang voor een fabrikant om concurrentie voordeel te behalen.

In een halfgeleiderfabriek worden plakken gemaakt, die honderden geïntegreerde schakelingen (ICs) bevatten. De halfgeleiderindustrie kan worden verdeeld in fabrikanten van hoog-volume producten, zoals micro-processoren en geheugen ICs, en fabrikanten van laag-volume producten, zoals consumenten ICs. Hoog-volume producten worden vooral op voorraad gemaakt, terwijl laag-volume producten in kleine hoeveelheden op klant-order worden geproduceerd. Hoog-volume fabrikanten zijn vooral geïnteresseerd in een zo hoog mogelijke doorzet, terwijl de laag-volume fabrikanten ook geïnteresseerd zijn in het realiseren van betrouwbare en korte doorlooptijden. Het realiseren van een hoge kosteneffectiviteit gebeurt door een zo hoog mogelijke bezettingsgraad na te streven. Echter voor het realiseren van betrouwbare en korte doorlooptijden is een verlaging van de bezettingsgraad gewenst. Deze tegenstelling vormt de uitdaging voor het ontwerpen van een halfgeleiderfabriek.

Hoewel er veel onderzoek is gedaan naar het ontwerpen van specifieke delen van een halfgeleiderfabriek, zoals layoutoptimalisering en scheduling, is er in de literatuur geen methode bekend die het volledige ontwerpproces met de bijbehorende ontwerpgereedschappen beschrijft. Het doel van dit proefontwerp is tweeledig: ten eerste het beschrijven van een integrale ontwerpmethode met de bijbehorende ontwerpgereedschappen en ten tweede het toepassen van de voorgestelde aanpak op het ontwerp van Philips' MOS4YOU waferfabriek in Nijmegen.

De in dit proefontwerp voorgestelde ontwerpmethode kan in vier fases worden onderverdeeld: het bepalen van de doelen en randvoorwaarden, het ontwerpen van de architectuur, het ontwerpen van de capaciteit en layout, en het ontwerpen van de operatie. Elke nieuwe fase is een uitbreiding van de voorgaande: er wordt telkens een verdere verfijning aangebracht in het ontwerp. Nadat de doelen en randvoorwaarden zijn beschreven, wordt de architectuur van zowel de materiaalstroom als het besturingsysteem bepaald. Vervolgens worden de benodigde capaciteiten en hun layout bepaald. Tot slot wordt in het ontwerp van de operatie beschreven wat de werkwijzen en besturingsregels zijn.

De volgende relevante ontwerpkeuzes zijn gemaakt en geïmplementeerd in MOS4YOU. De doelen en randvoorwaarden hebben geleid tot een hybride functionele layout. De machines worden verdeeld over hoofdbewerkinggroepen. In elke hoofdbewerkinggroep staat een hoofdtechnologie. Door ondersteunende machines te verdelen over de hoofdbewerkinggroepen, kan een aantal bewerkingen achter elkaar worden uitgevoerd in een groep. Een gedeeltelijk geautomatiseerd transport- en opslag-systeem is gekozen om betrouwbare, beheersbare, en kosten-efficïente transport en opslag te hebben. Tevens wordt flexibiliteit gegarandeerd doordat het transport naar de machines toe handmatig gebeurt. De productie-besturing bestaat uit een vrijgavestrategie en een volgorde-bepalingsstrategie. Orders worden op de vloer vrijgegeven volgens het gewenste activiteiten niveau. Echter, er wordt ingegrepen op basis van vooraf vastgestelde WIP (onderhanden werk) grenzen. Eenmaal op de vloer, wordt de lotvolgorde bepaald zodaning dat er een goede lijnbalans wordt gerealiseerd.

Verbeterstudies zijn uitgevoerd door diverse hoofdbewerkinggroepen afzonderlijk te beschouwen. Hierbij werd rekening gehouden met zeer gedetailleerde informatie. Het resultaat was verbeterde werkplekindeling, verbeterde werkmethoden, en meer verfijnde besturingsregels.

Curriculum Vitae

Edgar van Campen was born on May 30, 1969 in Pijnacker, The Netherlands. After finishing his Atheneum B at the Augustinianum in Eindhoven in 1988, he started his studies at the Eindhoven University of Technology, Department of Mechanical Engineering. He carried out his final project at the Philips Research Laboratories, where he modeled the dynamic behavior of elastomeric materials, to be used as rubber dampers in portable audio systems. He graduated in August 1993. In September 1993 he started following the post-masters program Design of Logistic Control Systems at the Stan Ackermans Institute in Eindhoven. This program was concluded with a final project on designing the layout and transport system of a multi-process multi-product wafer fab. The project was carried out at Philips Semiconductors and concerned the new wafer fab MOS4YOU in Nijmegen. In September 1995 Edgar joined MOS4YOU as manufacture design engineer. His work involved analyzing and modelling of the production capacity and the wafer flow. Since 1998, he has been involved in defining the production control scheduling system and in implementing the rules in production. Besides, since 1995 he carried out a Ph.D. study on the design of multi-process multi-product wafer fabs.

Stellingen

behorende bij het proefontwerp

Design of a multi-process multi-product wafer fab

1. Een gestructureerde aanpak van het ontwerpproces is noodzakelijk voor een goed ontwerp.
    - Dit proefontwerp.
2. Een goede bewerkingsvolgorde vergroot de doorzet van een productie systeem.
    - Dit proefontwerp.
3. Balancering van een *re-entrant flow-line* verkleint de variatie in doorlooptijd en vergroot de doorzet.
    - Dit proefontwerp.
4. Gebruik van de SEMI-E10 standaard is een voorwaarde om *best in class* te produceren.
    - SEMI STANDARDS [1997], *Specification for definition and measurement of equipment reliability, availability, and maintainability*, http://www.semi.org/, San Jose, CA.
    - Dit proefontwerp.
5. Een 300 mm wafer fab met 13 wafers per lot is minder rendabel dan een 200 mm wafer fab met 25 wafers per lot.
6. De verhouding tussen de kwaliteit van een planning en het vermogen van de planner om op tijd te komen is constant.
7. *Prototyping* is een populaire vorm van specificeren.

8. In het belang van de student, de universiteit, en de industrie dient een bedrijfsstage een verplicht onderdeel te zijn in het curriculum van een ingenieursopleiding.

    - Fey, J.J.H [2000], *Stelling 7*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

9. Het bepalen van de karakteristieke parameters ten behoeve van simulatie is moeilijk voor zowel visco-elastisch materiaal als voor een wafer fab.

10. Koken en ontwerpen hebben veel gemeen, planning en gereedschappen bepalen het eindresultaat.

11. Zeilen illustreert het verschil tussen theorie en praktijk, de snelste weg tussen twee punten wordt zelden gerealiseerd door een rechte lijn.

12. Net zoals bij een verbrandingsmotor, bestaat de dynamiek van een wafer fab uit een combinatie van *push* en *pull*.

13. Hoe meer auto's er stil staan, hoe harder het geld rolt.

Edgar van Campen
20 februari 2001