DEVELOPMENT AND SIMULATION ASSESSMENT OF SEMICONDUCTOR PRODUCTION SYSTEM ENHANCEMENTS FOR FAST CYCLE TIMES

DISSERTATION

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
DOKTORINGENIEUR (DR.-ING.)

VORGELEGT AN DER

TECHNISCHEN UNIVERSITÄT DRESDEN

FAKULTÄT INFORMATIK

DIPL.-ING. KILIAN STUBBE GEB. SCHMIDT
GEBOREN AM 30. Juli 1978 IN Wolfenbüttel

GUTACHTER:

PROF. DR. OLIVER ROSE
TECHNISCHE UNIVERSITÄT DRESDEN
FAKULTÄT INFORMATIK
INSTITUT FÜR ANGEWANDTE INFORMATIK
PROFESSUR FÜR MODELLIERUNG UND SIMULATION

PROF. DR. LARS MÖNCH
FERNUNIVERSITÄT IN HAGEN
FAKULTÄT FÜR MATHEMATIK UND INFORMATIK
LEHRGEBIET UNTERNEHMENSWEITE SOFTWARESYSTEME

TAG DER VERTEIDIGUNG: 29. JANUAR 2010

DRESDEN IM MÄRZ 2010

Contents

1 Introduction

Periods of change are considered disadvantageous times for the creation of optimal results. Fast changing conditions supersede solutions for yesterday's problems. Periods of relative stability, however, enable optimization efforts that have a sustainable effect. In manufacturing we often find highly optimized production solutions in industries that have experienced little change in processing conditions whereas industries experiencing extreme changes often suffer from productivity losses.

Semiconductor manufacturing takes place under constant change of the manufacturing conditions. The leading process technology changes every two to three years which means that the size of the area per feature on the wafer is cut in half. For this change in process technology many processing steps in semiconductor manufacturing have to be changed, new process steps are introduced into the manufacturing route and other steps are removed from the manufacturing route. The speed of this technological change is illustrated in Figure 1.1 (The width of each of the capacity bars corresponds to the production capability in wafer starts.). Every technology reaches peak production in the second or third year after introduction. Afterwards volume is shifted to newer process technology.

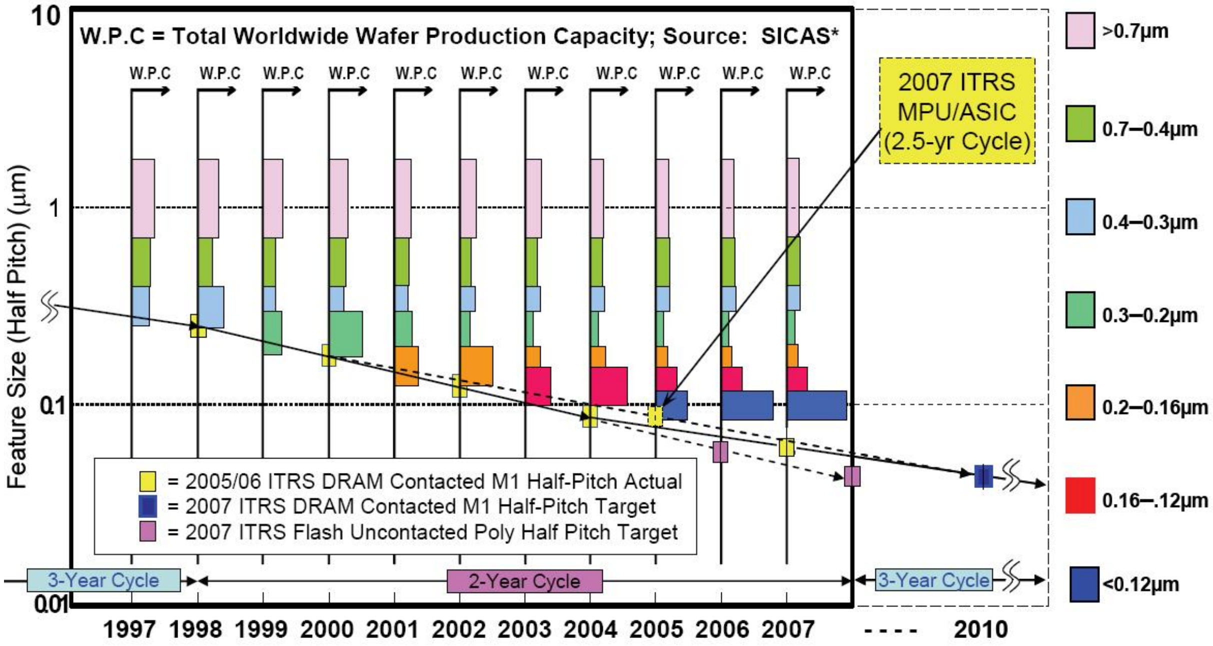


Figure 1.1: Technology cycles illustrated in production wafer capacity by technology and year (source ITRS [SIA07a])

Together with process technology, the product design defines capabilities and performance characteristics of the end product, but also the area necessary for one integrated circuit on the wafer. Miniaturization by new process technologies should lead to smaller areas per integrated circuit but actually this advantage is more than offset by the increasing product complexity (see Subsection 3.1.4 on Moore's Law). In order to maintain reasonable production costs per chip and also to increase productivity larger wafer sizes have been adopted by the industry every 11-13 years during the last decades[Gre07]. Every wafer size step required development of new equipment and manufacturing sites had to be completely refurbished provided they wanted to take part in this wafer size step. This represents another source of recurrent change in semiconductor manufacturing environment[1].

Additionally semiconductor industry has experienced times of extreme growth, which often leads to narrow focus on some aspects only. In the 1990s the fast growing market accepted all products which were more or less up-to-date, therefore product and technology development was given prominence over operational considerations.

For all these different reasons the current production system in semiconductor manufacturing is less than optimal and does not meet the needs of the semiconductor manufacturing companies. There are severals indicators of less than optimal operation. One major indicator is that the cycle time often exceeds 60-80 days.

This represents a major disadvantage, because many operational success factors like lean inventory or fast reaction to customer demand rely on short cycle time (see Section 2.3). Therefore we have chosen a significant reduction in cycle time as objective for production system changes under consideration in this thesis.

Some new approaches for production systems are currently in discussion in the industry, namely the replacement of batch tools with mini-batch or single-wafer tools and the reduction of lot size. We analyze these changes with respect to the effectiveness of the cycle time reduction possible with these changes. Additionally we develop and assess new approaches for production systems in semiconductor manufacturing that are able to deliver substantially shorter cycle times.

The key method in our assessment is discrete-event simulation. It enables us to answer *what-if-questions* i.e. we can assess the cycle time effectiveness of production system changes by creating a simulation model of a fab with the current production system and a changed simulation model with the changed production system. Then we compare the simulation output of both scenarios and evaluate the performance difference.

1.1 Thesis Organization

This thesis is organized as follows. In the first part we discuss the background and introduce methods and simulation model for the production system changes in Chapters 2 through 5. Then we present and discuss the assessment of the different changes in Chapters 6 through 9.

In Chapter 2 we discuss why short cycle time is urgently needed in the semiconductor manufacturing industry and how the production system changes contribute to a company's manufacturing strategy. We outline the semiconductor manufacturing environment in Chapter 3. This includes an introduction to all elements defining or influencing the production system as well as an overview over current cycle time performance and past methods to shorten cycle time. Methods and tools of our analysis are presented in Chapter 4 with a short introduction into discrete-event simulation, queueing theory and gantt-charts. Chapter 5 concludes the first part of the thesis with the presentation of our baseline simulation model and a discussion of the underlying fab as well as verification and validation of the model.

Our assessment of production system changes starts in Chapter 6 with considering the replacement of batch tools. In Chapter 7 we discuss the benefit of reducing the lot size in the baseline model as well as in combination with the changes considered in the preceeding chapter. In the following we assess how fab or tool scaling can leverage the benefit of smaller lot size in Chapter 8. Chapter 9 concludes the second part with the introduction of a new cluster tool type based on the insights gained in previous chapters. We show how such a new cluster equipment could operate and assess the possible benefit.

Finally, we present conclusions of our research in Chapter 10 and give an overview over the perspectives

[1] We do not want to miss mentioning that this changes also represent a possible opportunity in production system design because new equipments have to be developed anyway.

for the industry’s future development and future areas for research.

2 Strategic Motivation for Short Cycle Times

The sharp focus on cycle time within this dissertation originates from their surge in importance. Today, time is one of the critical success factors for competitive manufacturing in general [MST00]. Having very unique characteristics the semiconductor industry has not been captured by this shift to the same extent as did many other industries. However, recent market and product dynamics make the need for advancement more imminent. In the introduction in Chapter 1 we outlined reasons for the suboptimal production system leading to long cycle times. This chapter outlines the general background for the criticality of short cycle time as well as current business environment making substantially shorter cycle time absolutely necessary .

2.1 Company Goals in Current Business Environment

Changes in the economic, technological, sociocultural, and political environment affect business companies [Zäp89]. These changes influence the companies' ability to compete on the market. New risks but also chances arise for the competitive orientation of a company. It is critical that the company management embraces this change and sets the direction for necessary changes in the company's capabilities in order to maintain or increase competitiveness.

We already discussed technology, product design, and wafer size changes in the introduction. Additionally, the business environment of semiconductor manufacturing currently experiences a series of changes.

Doug Grose (Chief Executive Officer at GLOBALFOUNDRIES) characterizes today's market as demanding more diverse product sets with parts ranging wide in volume and both demand and life cycles [Gro07a]. He notes the business challenge to ”[E]effectively manage the mix to deliver the right products at the right time in the right volume" [Gro07b].

Mark Liu (Vice President of Operations at TSMC) similarly observes the market characteristics for semiconductor foundry products for consumer electronics. He notices

- ”continuous price reduction at end market, (...)
- strong price elasticity along the product life time,
- strong inventory effects, (...)
- many products with volume, (...) [and]
- many new applications" [Liu05].

This is consistent with the view of market intelligence service providers. iSuppli Corporation notes that 2006 saw a diversification of memory makers' product portfolio by adding specialty DRAMs designed for non-PC applications. At the same time iSuppli also predicts a decline in DRAM prices of 31% for 2007 [iC07]. In Gartner's DRAM market outlook for 2007 Andrew Norwood (Research Vice President at Gartner) states that ”[C]continued changes in production allocation, fluctuating demand patterns and the unrelenting advance of semiconductor technology make this [DRAM production - KS] a high-risk business for suppliers and investors. (...) DRAM suppliers must acquire the ability to switch capacity between DRAM and NAND flash memory as efficiently as possible to take advantage of prevailing conditions." [Gar06]. This represents a new business challenge for memory makers, which not long ago was considered a one-product commodity business.

The above observations touch different areas of semiconductor manufacturing (memory vs. logic; Integrated Device Maker (IDM) vs. foundry), but have a common theme: A rising product segmentation with higher uncertainties in demand meets continued price decline. This means that companies manufacturing semiconductors have to enhance their capabilities and develop new capabilities to meet this challenge. The core capabilities to develop or improve are

- to react much faster to changes in demand with adjusted product output, and
- to quickly deliver and introduce products in order to be less vulnerable to price declines.

These new capabilities have also been characterized as requiring *agility* from manufacturing [Spl07].

2.2 Manufacturing Strategy

Manufacturing strategy is a functional strategy contributing to the overall company strategy striving to fulfill the company goals [Zah88, BK04]. Within the set of functional strategies, the manufacturing strategy is of central importance as it defines cost, quality and time of the products demanded by the market. Other functional strategies as, e.g., the sales strategy rely on it.

Historically manufacturing strategy has been focused on optimizing cost, however this one-sided optimization often lead to a degradation of competitiveness [Ski86]. Skinner who is widely regarded as one of the manufacturing strategy pioneers therefore demanded that ”... we must set a new, simple but powerful objective for manufacturing: to be competitive." [Ski86]

2.2.1 Definition of Manufacturing Strategy

Building on this objective Skinner defines manufacturing strategy as ”... the competitive leverage required of - and made possible by - the production function ... And it spells out an internally consistent set of structural decisions designed to forge manufacturing into strategic weapon." [Ski84] This definition reflects the significance of the manufacturing strategy for a company’s competitive strategy and ascertains the result of the manufacturing strategy definition process: consistent decisions regarding manufacturing structure.

Henn and Kühnle specify the outcome in a more detailed way. According to them the manufacturing strategy sets the framework for manufacturing operations and the linked capacity allocation and dimensioning of structural elements ([ES99], p. 9-65).

With respect to the timeframe Zahn notes that manufacturing strategies can also be interpreted as decision patterns that are kept up over some period of time ([Zah88], p. 527).

2.2.2 Elements of Manufacturing Strategy

In order to fulfill these definitions a manufacturing strategy has to cover an extensive field. Zahn groups its elements in five categories [Zah94]:

- manufacturing tasks: type and quantity of the products to manufacture.
- manufacturing structure: total capacity, type and capacity of equipments, factory location.
- manufacturing infrastructure: information and communication technology, production planning and control systems.
- manufacturing process: vertical integration, logistics, relation to suppliers.
- human resources: qualification and motivation.

In contrast to this functional classification, Blecker identifies manufacturing concepts as an integrating layer below the manufacturing strategy. These concepts follow a guiding ideal or model to create, guide and develop manufacturing operations. In order to achieve this guiding ideal or model these concepts use specific manufacturing instruments [Ble03]. Popular manufacturing concepts are, e.g., Lean Production, World Class Manufacturing or Computer Integrated Manufacturing (CIM); well-known manufacturing instruments are, e.g., Just-in-time (JIT), Total Quality Management (TQM), Kaizen, or cellular manufacturing [AI05].

It is important to note that these concepts or instruments themselves do not form a strategy. Hayes and Pisano indicate that " ... simply improving manufacturing - by, for example, adopting JIT, TQM, or some other three-letter acronym is not a strategy for using manufacturing to achieve competitive advantage [HP94]", Porter asserts that "Operational Effectiveness is not a strategy" [Por96] and Skinner assesses that competitve advantages in production are founded rather on strategic decisions on capacity and sites and the optimal use of manufacturing capabilities (e.g., modern production technology) than on productivity advances and cost savings ([Ski86], p. 56). The different concepts used need to complement themselves to enable a successful holistic strategy.

2.2.3 Elements for New Semiconductor Manufacturing Strategy

In order to fulfill the company goals identified in Section 2.1 semiconductor manufacturing must deliver substantially shorter cycle times and acquire more flexibility to produce different products. Consequently, a new persuasive strategy has to address these objectives. However, contributing factors to cycle time and inflexibility of todays factories are manifold.

Various approaches have been applied to reduce cycle time in semiconductor manufacturing (see Section 3.6). These traditional improvement approaches have lead to significant cycle time reductions in the past [LH00], however, they are insufficient to reach the desired significant improvement. A different systematic approach is necessary to complement them.

Production system changes have hardly played a role in cycle time reduction efforts, mostly for cost reasons. Anticipating the analysis in Section 3.4, we identify a production system paradigm change including small lot manufacturing, transition to mini-batch and single wafer processing, changes in cluster tool design, and rapid, high volume material handling system as potential enabler. In Figure 2.1, we group these changes under the term *Next Generation Production System*, which we define as a new manufacturing concept implementing this paradigm change in production system design.

Conveniently, this paradigm change is not only an enabler of fast cycle time, it also accomplishes the objective of increased flexibility. Long cycle time is a major source of inflexibility itself preventing rapid changes in the product output. Another major source for inflexibility in semiconductor manufacturing are batch tools which process up to 150 pieces at the same time but often cannot mix wafers of different products in the same run. If batch tools are replaced with mini-batch or single wafer equipment then a higher flexibility is reached. However, flexibility can have a wide range of meanings in manufacturing, therefore we limit the target of increased flexibility in this analysis. We regard only the instances discussed above that come as a by-product when trying to achieve fast cycle time.

Together with renewed efforts in traditional improvement areas the Next Generation Production System concept can form a new promising manufacturing strategy. Yet, the three branches in Figure 2.1 still do not produce the complete manufacturing strategy. Some of the elements defined in Section 2.2.2 as, e.g., the manufacturing tasks are not included. However, those are very different for the different IC makers and the elements defining the success factors of manufacturing in the semiconductor industry are all addressed: speed (Next Generation Production System), cost (Lean Production) and quality (Computer-Integrated Manufacturing).

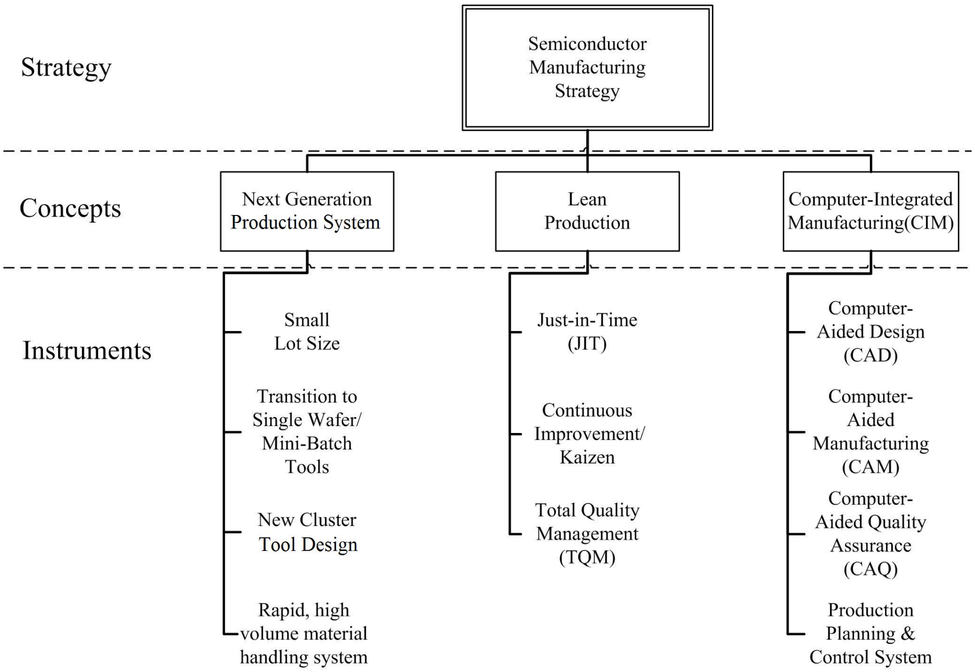

Strategy
Semiconductor Manufacturing Strategy
Concepts
Next Generation Production System
Lean Production
Computer-Integrated Manufacturing(CIM)
Instruments
Small Lot Size
Transition to Single Wafer/ Mini-Batch Tools
New Cluster Tool Design
Rapid, high volume material handling system
Just-in-Time (JIT)
Continuous Improvement/ Kaizen
Total Quality Management (TQM)
Computer-Aided Design (CAD)
Computer-Aided Manufacturing (CAM)
Computer-Aided Quality Assurance (CAQ)
Production Planning & Control System

Figure 2.1: Semiconductor Manufacturing Strategy

2.2.4 Area of Analysis within this Dissertation

This dissertation assesses how effective a Next Generation Production System concept can address the cycle time needs. The reasoning for the need behind the individual instruments is derived and the instruments are tailored to produce the desired effect by means of a theoretical analysis. Furthermore the concept is assessed in a number of simulation scenarios.

This dissertation does not assess in detail how other strategic elements of manufacturing need to adapt in view of this development. E.g., there are implications for lot scheduling and dispatching if the lot size is reduced to very few wafers. As one illustrative example, scheduling and dispatching might then have to take yield analysis needs into account, which is not necessary as long as wafers are grouped together naturally by lots having more wafers than tools have chambers.

This dissertation does not assess either, how the other success factors cost and quality are affected in detail. The measures discussed in this work impact cost and quality, however a thorough assessment would go beyond the scope of this work. Therefore we merely list or roughly discuss the implications, whenever they occur.

2.3 Sustainability of this new Manufacturing Strategy

Manufacturing companies usually have competitors that have sufficient resources and cleverness to realize the success of a new strategy. Therefore a successfull manufacturing strategy will inevitably find its imitator [Gro03]. If the Next Generation Manufacturing concept is successful at one company, then others will copy it. Hence, we have to ask one question to assess the sustainable benefit of this new strategy: which advantages of short cycle time are truly sustainable and survive the competition's catch up?

Short cycle time gives many advantages [HS00]. It enables

- lean inventory,

- faster introduction of new products to market,
- fast yield learning,
- fast excursion finding,
- less reliance on demand forecast, and
- flexibility in product output enabling fast reactions to customer demand.

Apart from the increased flexibility discussed below only one of these benefits is lessened when the competition has caught up: The competitive advantage of introducing new products to market earlier delights the early adopters only for some time. All other benefits are sustainable into the period of *even* competition.

It is more difficult to assess the sustainability of the advantages of higher flexibility. The ability to react quickly to changes in customer demand is clearly sustainable, but the advantage decreases when the competition acquires the same capability. On the other hand the increased flexibility also enables a more diverse product portfolio which can be less susceptible to competition.

The sustainability of most advantages cofirms the whole strategy as promising. However, the extent and the strength of its sustainability depends on the competitive situation of the individual IC maker.

3 Cycle Time in Semiconductor Manufacturing Context

In this chapter, we introduce

- the process of semiconductor manufacturing,
- the entities and the material flow in semiconductor manufacturing,
- the key preformance figures with relevance for cycle time,
- the elements of the production system in semiconductor manufacturing,
- the current cycle time performance, and,
- the traditional methods of cycle time reduction

The use of many manufacturing terms is mixed in both industry and literature, therefore we introduce definitions along with the introductory analysis because we require precise definitions to enable clear analysis.

3.1 Process of Semiconductor Manufacturing

Integrated circuits fabricated in semiconductor manufacturing facilities today consist of up to nearly a billion of transistors, resistors, capacitors, and diodes on a single chip. The size of this chip is only a few square centimeters. In the following we introduce the technological process of manufacturing integrated circuits.

3.1.1 Stages of Semiconductor Manufacturing

The semiconductor manufacturing process can be grouped into these five stages [vZ04]:

1. Material preparation
2. Crystal growth and wafer preparation
3. Wafer fabrication and sort
4. Packaging
5. Final and electrical test

In the *material preparation* stage, the raw material for semiconductor manufacturing is mined and purified. The result of this step is pure silicon with polysilicon structure. In the following second stage, *crystal growth and wafer preparation* the silicon is formed into a crystal with specific structural and electrical parameters. Afterwards the crystal is cut into many thin disks called wafers which are surface treated subsequently. These wafers have a circular shape and their diameter is 300mm in modern semiconductor manufacturing.

In the third step *wafer fabrication and sort* the actual integrated circuits or devices are formed on the wafers' surface. Up too several thousand identical devices can be manufactured on a wafer. The individual devices are called *die* or *chip*. Wafer fabrication can take up to thousand individual operations (including measurement operations), which can be grouped into two major segments representing two major activities. Transistor forming takes place in the front end of line (FEOL) and the wiring of the individual transistors in different connected metal layers takes place in the back end of the line (BEOL).

After wafer fabrication the individual dies are tested electrically to identify those that meet the specification. This test is called *sort* as the dies are sorted into those that match the specified criteria and those that do not.

During the fourth step *packaging* the wafer is cut into individual dies and the ones that met specification in the sorting step are placed into protecting packages. Apart from protecting the die, the package provides a durable connection to, e.g., a printed ciruit board. Afterwards a *final electrical test* is performed to ensure a functioning end product.

Figure 3.1 illustrates the shape of the end products of the individual manufacturing steps.

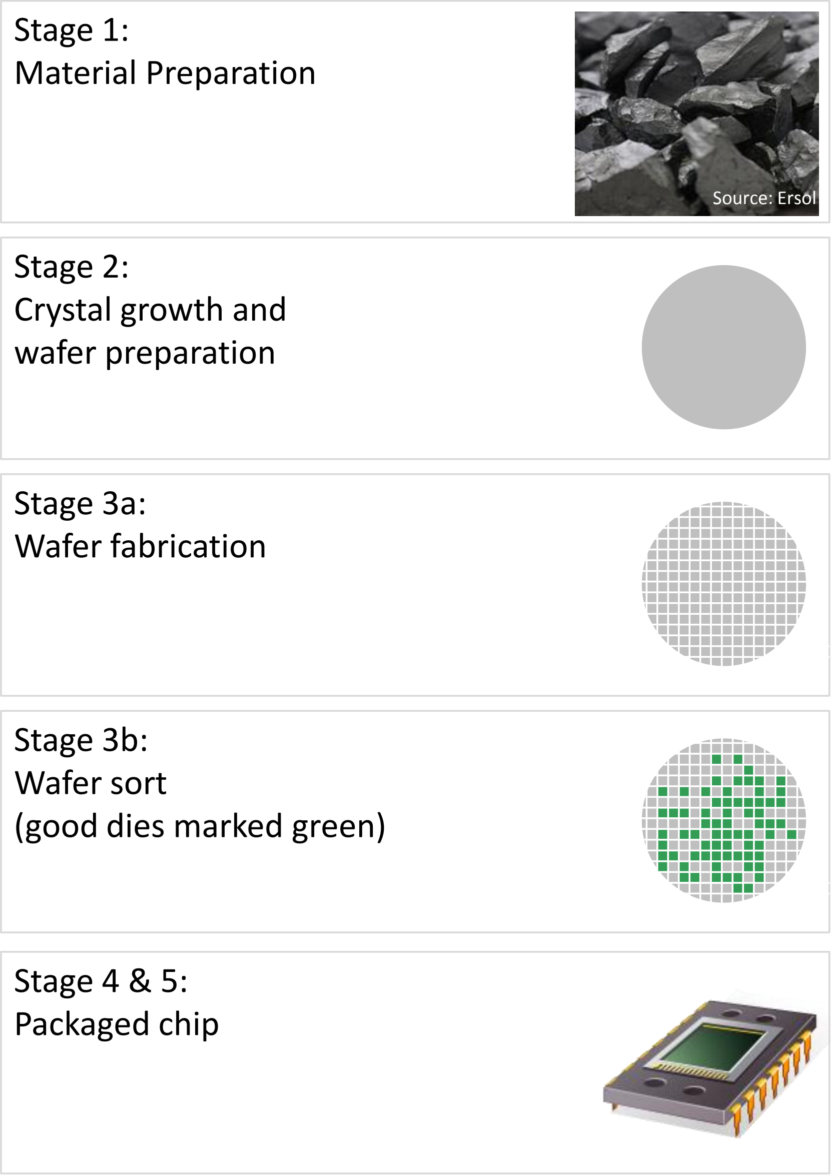


Figure 3.1: Products of the different manufacturing steps

Another differentiation of stages often made is the distinction into the *front-end* part and the *back-end* part of semiconductor manufacturing. In this differentiation the front-end part refers to step 3 only and the back-end part refers to steps 4 and 5. The steps 1 and 2 are seen aas resourcing steps in this view.

All stages are usually performed at different manufacturing plants. By far, most cycle time accrues at the third stage, *wafer fabrication and sort*. Therefore this step is the natural target for cycle time reduction efforts in semiconductor manufacturing and our analysis focuses on this manufacturing step. When refering to semiconductor manufacturing in the following, then we refer to this step only. The manufacturing plants performing this step are called fabs.

3.1.2 Process Steps in Semiconductor Manufacturing

There are numerous different types of integrated circuits for different functions. However, all integrated circuits are made of the same few basic structures and manufacturing processes. The four basic manufacturing operations repeated in various combinations and long sequences are

- layering,
- patterning,
- doping, and,
- heat treatment.

Figure 3.2 illustrates these basic operations, which are explained more detailed in the following.

3.1.2.1 Layering

During layering operations, thin layers are added onto the surface of the wafer. These layers can be of insulating, semi-conducting or conducting material. The processing techniques to generate these layers are either grown oxide or nitride layers, or deposition of various materials. The deposition techniques commonly used are chemical vapor depostion (CVD), physical vapor deposition (PVD), sputtering, and electroplating.

3.1.2.2 Patterning

Patterning describes a series of steps that lead to the removal of selected parts of one or more previously added surface layers. The result of these steps is a pattern on the wafer surface. The removed material can have different shapes. Specifically differentiated are holes in the layer and islands of the remaining material (see Figure 3.2). The patterning steps are grouped into photolithography or masking steps and etching steps.

Photolithography steps consist of two steps[1]. First comes the exposure step in which the pattern for the specific layer is transfered from a photomask onto the top layer of the wafer which consists of photoresist sensitive to light of a specific wavelength. In the second step the photoresist is developed and the unpolimerized parts of the resist are removed.

The etching steps consist of a first step in which the layer below the developed photoresist is removed in the unmasked areas and a second step in which the remaining photoresist is removed.

One layer at a time the various physical parts of the integrated circuit are formed in and on the wafer surface by means of patterning. Patterning is highly critical because it defines the critical dimensions of the resulting product. Correct location of the patterns on the wafer and in relation to the other parts has to be ensured for a functioning end product.

[1] Sometimes the previous layering step of adding photoresist onto the wafer is also accounted for as photolithography step.

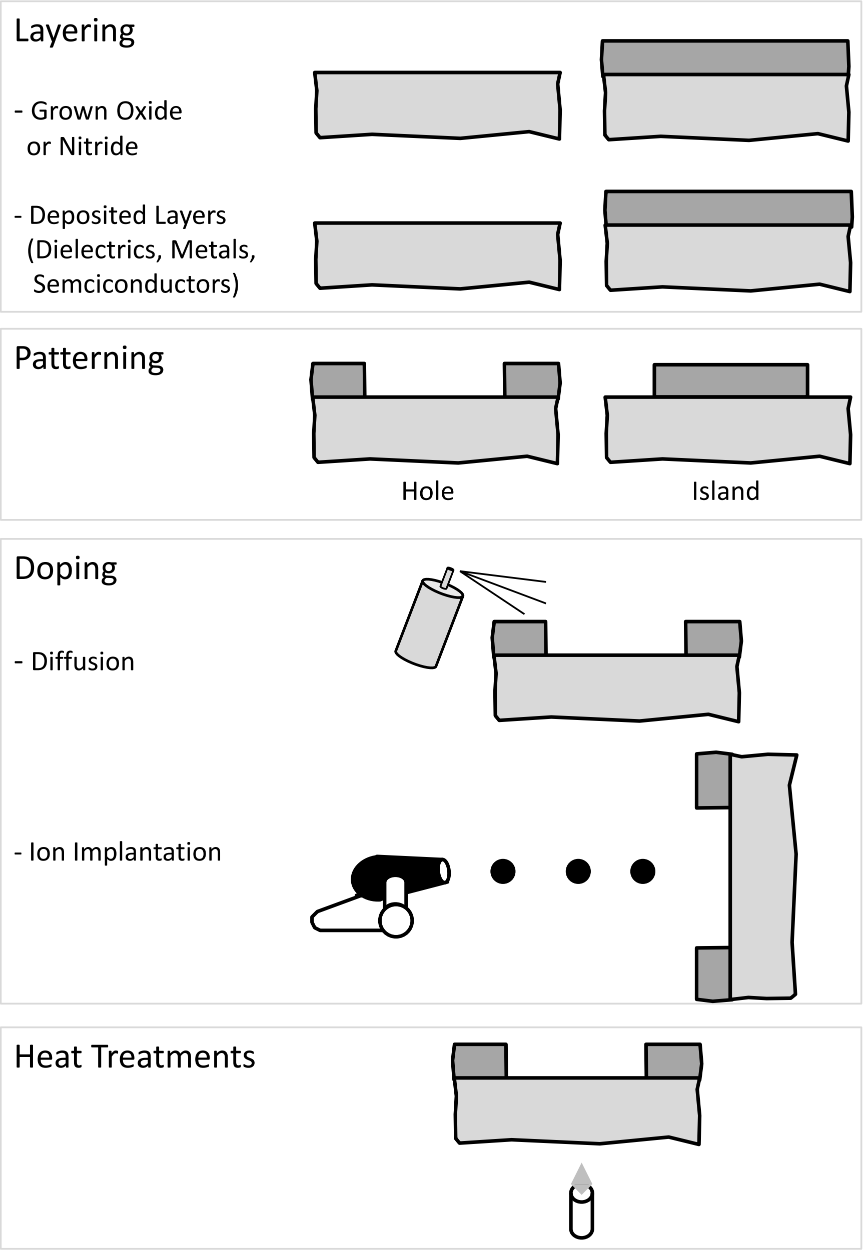


Figure 3.2: Basic operations of semiconductor manufacturing according to [vZ04]

3.1.2.3 Doping

Doping is the process of intentionally inserting impurities into the wafer in order to change its electrical characteristics. Only a small number of dopants is necessary to change the ability of a semiconductor to conduct. Typical doping processes introduce impurities in the order of 1 per 10,000 to 1 per 100,000,000 atoms. Doping can be performed by either diffusion or ion implantation techniques.

Thermal diffusion is a chemical process taking place when the wafer is heated above a specific temperature and is surrounded by vapors of the desired dopant. During the diffusion process dopants move from the vapor into the wafer surface creating thin layers on the wafer surface.

Ion implantation is a physical process. Dopant atoms provided in gas form are ionized, accelerated to a high-speed stream of ions and this stream is swept across the wafer. The momentum of the ions carries them into the wafer surface, with the depth depending on the angle of the stream to the wafers' crystal lattice.

The objective of doping is to create areas on the wafer that are either rich in electrons (n-type) or rich in electrical holes (p-type). These areas form the electrically active regions and N-P junctions which are necessary for operation of the transistors, capacitors, resisitors, and diodes of the integrated circuit.

3.1.2.4 Heat Treatment

Heat treatment operations represent process steps of controlled heating and cooling of the wafer. Typical temperatures reached during this process are 450 to over 1000 degrees Celsius. Heat treatment is, e.g., required to repair damages in the wafers' crystal structure caused by ion implantation or to anneal deposited metals to ensure good electrical conduct.

3.1.3 Example Fabrication Process

Integrated circuit manufacturing starts with a polished blank wafer. Figure 3.3 illustrates the basic operations necessary to form a simple metal oxide silicon-gate transistor structure according to [vZ04]. Although only some intermediate steps are shown, an impression of the number of steps necessary to create this simple transistor can be received. Numerous patterning and layering steps are necessary together with some doping operations (see the p- and n-marked areas) and heat treatment operations (impossible to conceive from the illustration). More complex integrated circuits like current microporcessors or memory devices require much more complex strucures, however.

3.1.4 Product Complexity driving Process Technology Progress - Moore's Law

The famous Moore's Law states that the number of transistors per area doubles approximately every two years. This important trend in semiconductor manufacturing was first observed by Intel co-founder Gordon E. Moore and continues to hold after almost half a century since its publication without showing severe signs of an imminent end of the trend. The engine keeping this trend alive is the miniaturization discussed in the introduction in Section 1. Process technology has to be developed at a rapid pace to enable this trend, because each miniaturization requires altered or new processes at a significant number of process steps. These changes in process technology can refer to slight changes in the process conditions to completely new sequences of additional operations. This also means that every two years routes specifics and equipment throughput change with the introduction of a new technology which also leads to changes in workstation utilization or even the necessity of additional equipments.

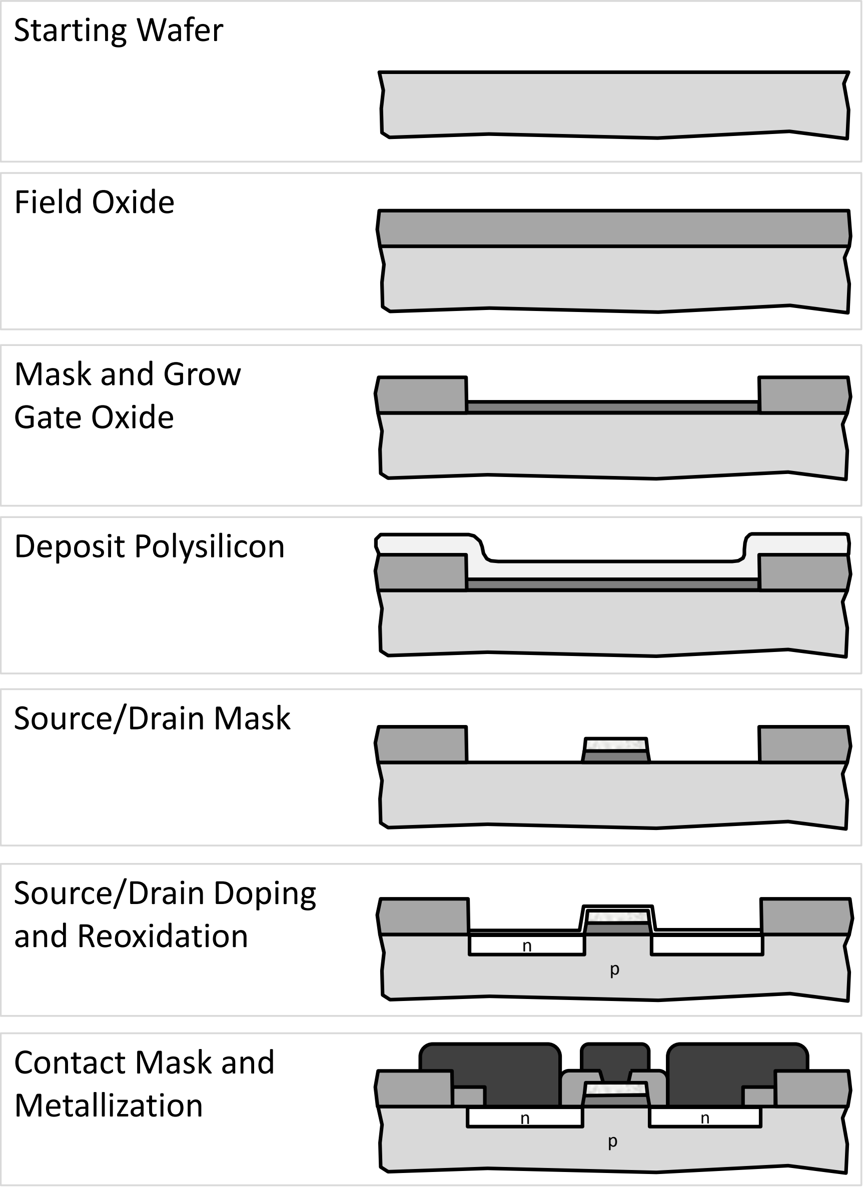


Figure 3.3: Basic operations required to form a simple metal oxide silicon-gate transistor structure according to [vZ04]

3.2 Basic Entities in Semiconductor Manufacturing

We introduce the entities involved in the manufacuring process together with the manufacturing flow defining their relation and interaction.

3.2.1 Definitions

- Wafer: Unprocessed wafers are the raw material of the semiconductor manufacturing process; processed wafers are its end product. Wafers consist of a circular disk of silicon buidling the substrate for the integrated circuits fabricated on it.
- Product: A product specifies the integrated circuit that is manufactured. Usually houndreds to thousands of units of the same product are manufactured simultaneously side to side on a wafer.
- Lot: A lot refers to a set of wafers that traverses a route together. All wafers within a lot have to be of the same product.
- Route: A route describes the process flow that a lot takes. This includes the sequence of operations passed by a lot. Not all operations have to be mandatory; metrology operations usually can be skipped by a specified percentage of lots. Routes begin with lot start and end with lot ship. In semiconductor manufacturing they usually have several hundred operations.
- Operation: An operation is a step within a route. It is associated with a recipe and a workstation.
- Process operation: If the wafer is processed by means of a chemical or physical process and has different characteristics afterwards, then the operation performed is a process operation.
- Metrology operation: In contrast, metrology operations do not change the characteristics of the wafer. Metrology operations are necessary to control the physical and chemical processes and to ensure the quality of the product.
- Skip rate: The skip rate specifies the percentage of lots that can skip an operation. The skip rate r_{skip}'s value range is $0 \leq r_{skip} < 1$ for metrology operations; process operations are not skipped, i.e. their r_{skip} is always zero.
- Sampling rate: In many cases it is sufficient to measure only few wafers and not the full lot at metrology operations. Reasonable sampling is dependent on lot size, therefore we specify it illustrating, e.g., when two wafers of a lot containing 24 wafers are measured, then we specify the sampling rate to be $r_{samp} = 2/24$.
- Tool: Tools refer to the machines performing operations. The term *equipment* is used synonymously.
- Workstation: A workstation is a set of one or more tools that have the same processing capabilities.
- Recipe: A recipe specifies the specific process that is performed on the wafers by a workstation. Typically, workstations can perform different recipes that also vary in duration.
- Utilization: Utilization refers to the percentage of total available production time that a tool or workstation is actually busy.
- Loading: Loading refers to the utilization of the complete manufacturing line. Utilization of the workstation with the lowest capacity equals the loading in the long term.

3.2.2 Relation and Interaction of Basic Entities

We illustrate how these entities relate and interact in Figures 3.4 and 3.5. Figure 3.4 examplifies a route with m operations. Each step has an associated workstation and recipe. The individual lots 1-4 follow the route, but skip some metrology operations. Two important characteristics of semiconductor manufacturing are obvious from the illustration:

- Semiconductor manufacturing has a high degree of reentrancy, i.e. the same workstation is visited several times at different steps.
- Different steps at the same workstation may have the same recipe, but this is rare for different process operatons.

Operation	Workstation	Operation type	Recipe	Lot 1	Lot 2	Lot 3	Lot 4
Lot start				●	●	●	●
Operation 1	Station A	Process	Rec. A1	●	●	●	●
Operation 2	Station B	Process	Rec. B1	●	●	●	●
Operation 3	Station C	Metrology	Rec. C1	●		●	
Operation 4	Station D	Process	Rec. D1	●	●	●	●
Operation 5	Station B	Process	Rec. B2	●	●	●	●
Operation 6	Station C	Metrology	Rec. C1		●		
Operation 7	Station E	Metrology	Rec. E1	●			●
Operation 8	Station B	Process	Rec. B3	●	●	●	●
⋮							
Operation m	Station C	Metrology	Rec. C2	●			●
Lot ship				●	●	●	●

Figure 3.4: Example for a route

Figure 3.5 examplifies operation at a workstation. Provided all tools are busy lots first join the workstation's queue. Whenever one tool becomes available for processing one lot is selected to be processed. After the processing is finished, the lot joins the queue at the next workstation or is processed directly at the next workstation specified by its route.

The differentiation into queue and processing is the basic distinction of a lot's states. Lots are accounted as in queue until the first wafer of the lot enters the tool and then as in process until the last wafer leaves the tool. This represents the distinction into ideally avoidable (queue) and necessary (processing) time which we further explore in the following Section 3.3.

3.3 Fundamental Relations between Throughput, Cycle Time, and Work in Process

Throughput, cycle time and work in process are key performance figures of a fab. They represent

- how many products (throughput) can be manufactured,
- how long the lead time (cycle time) is, and,
- how much capital (work in process) is tied up.

In this section we explore how these fundamental quantities relate.

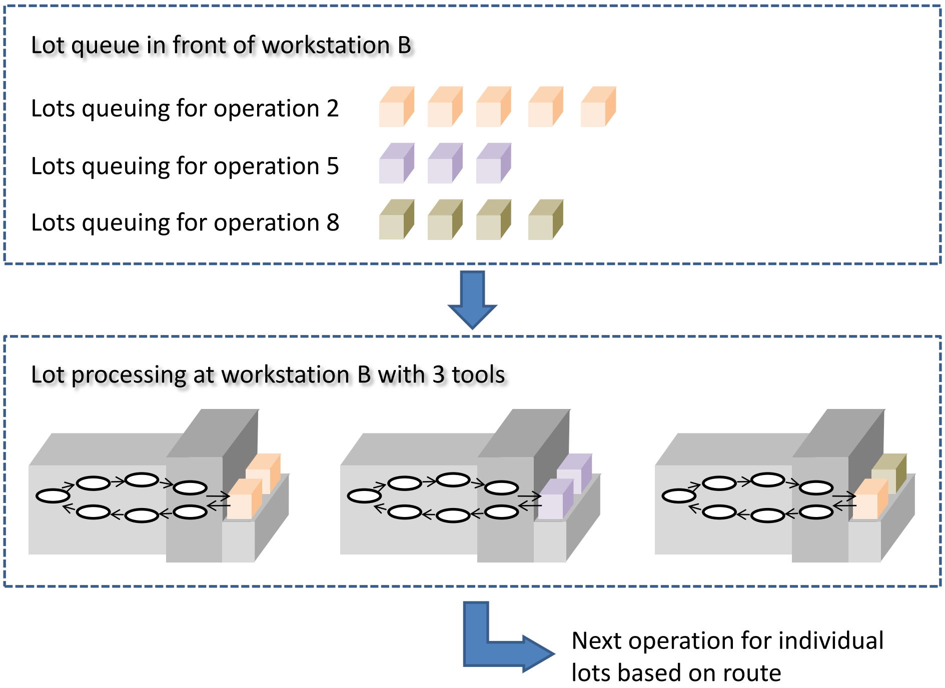

Lot queue in front of workstation B
Lots queuing for operation 2
Lots queuing for operation 5
Lots queuing for operation 8
Lot processing at workstation B with 3 tools
Next operation for individual lots based on route

Figure 3.5: Operating behavior at a workstation

3.3.1 Definitions

- Throughput (THP): Throughput is the quantity of wafers manufactured per unit time. This definition can be applied to the entire fab or individual stations. In order to avoid confusion, we use THP in the station or workstation context and Fab-THP, when denoting the throughput of the entire fab.
- Cycle Time (CT): Cycle time (also called flow time) is the time from releasing a lot into the factory until its processing is completed, i.e., the time that elapses from lot start until the lot reaches the lot ship operation. The time spent at the lot ship operation itself is excluded.
- Work in Process (WIP): Work in Process is the inventory containing all lots that have been started but have not reached the lot ship operation.

3.3.2 Little's Law

Little's law relates the fundamental quantities. It says: The average Work in Process in a stable system is equal to its average Throughput multiplied by its average Cycle Time, or,

$$WIP = THP * CT. \tag{3.1}$$

The only requirement for Little's Law is the stationary property of the system. Real production is rarely stationary, however for an infinite observation time Little's Law holds for all production lines independent of the degree of variability involved. For most less-than-infinite cases, Little's Law is an excellent approximation with the exception of transition periods like loading changes including factory start-ups or changes in product spectrum with different processing requirements.

Little's Law is applicable to a single station, a line segment, a complete line, or a plant consisting of several lines. This makes it widely useful, e.g., we can use it to calculate the missing one of the three fundamental quantities.

3.3.3 Ideal and Actual Performance of Production Lines

We further explore these parameters and study their characteristics for an ideally performing production line.

Bottleneck Workstation Capacity C_0: The bottleneck workstation capacity sets the upper limit to the amount of material that can be processed by the manufacturing line, i.e., it represents the maximum possible FAB-THP. The bottleneck workstation has the highest utilization in the long term and usually a long queue of material waiting to be processed.

Raw Process Time T_0: The Raw Process Time is the sum of the average process times of each operation in the line. Metrology operations are included with their respective skip rates. Alternatively we can define raw process time as the average time it takes for a single lot to traverse an empty line without transport times. The prerequisite of this alternative definition is that always one tool of the workstation is available for processing, i.e., it is not in a downtime[2]. Within the semiconductor industry, transport time is never included in the raw process time. This makes characteristic curves look a bit weird because Cycle Time CT does not approach Raw Process Time T_0 for low loadings as much as seems convincing. Therefore we mark the gap caused by transport times in the characteristic curves illustrating our simulation results later in this work.

Critical WIP W_0: The critical WIP is the WIP level necessary to achieve the line throughput at bottleneck workstation capacity in a manufacturing line without variability. If the WIP is below W_0, then the line is under-utilized even in the no-variability case; if WIP is above W_0, then there is waiting WIP (WIP exists that is not being processed) in any case. Critical WIP is defined by the Bottleneck workstation capacity and raw process time by the following relationship:

$$W_0 = C_0 * T_0. \tag{3.2}$$

These three key figures can be reached simultaneously in an ideal production line with zero variability. However in other cases with variability which includes practicably all real cases, it is impossible to reach this optimal working point. The working points achievable in reality are located on a curve called *characteristic curve*, which is specific for a fab with a defined toolset and a defined product mix. Figure 3.6 illustrates the ideal and actual performance and interdependance of the fundamental quantities discussed in the above paragraphs. There is no indisputable optimal working point on these characteristic curves. Every change on the curve changes at least one parameter to the good and at least one to the bad. The best working point for a specific fab is highly dependent on the company's business model and the market situation.

The characteristic curves reveal one obvious possibility for cycle time reduction: a decrease in FAB-THP, i.e., a lower loading of the fab. However, this also increases the cost of the products manufactured because the capital investment stays the same. Therefore this is not a realistic approach for cycle time reduction, as we expect that fabs already choose the working point that represents the best compromise between cycle time and manufacturing costs.

The characteristic curves provide other insights for cycle time reduction, too. Although the ideal performance limits are not achievable in practice, they still provide important insights. First, they mark the improvement potential that is - at least theoretically - realizable by variability reduction. And, secondly,

[2] The raw process time T_0 does not represent the minimum possible cycle time of a production line. The minimum possible cycle time is refered to as *theoretical cycle time* by Sematech and defined as "the summation of cycle times of individual value-added operations at the minimum known process time for a single unit of product for a manufacturing sequence. This definition of theoretical cycle time includes the load, process, and unload times and does not include transportation, set up, queue, downtime, metrology, or production test, which would be considered process inefficiencies." [Sem08] However ideal manufacturing performance cannot mean that metrology operations and actual lot size are disregarded, therefore this defintion would be incorrect to define an ideally performing line.

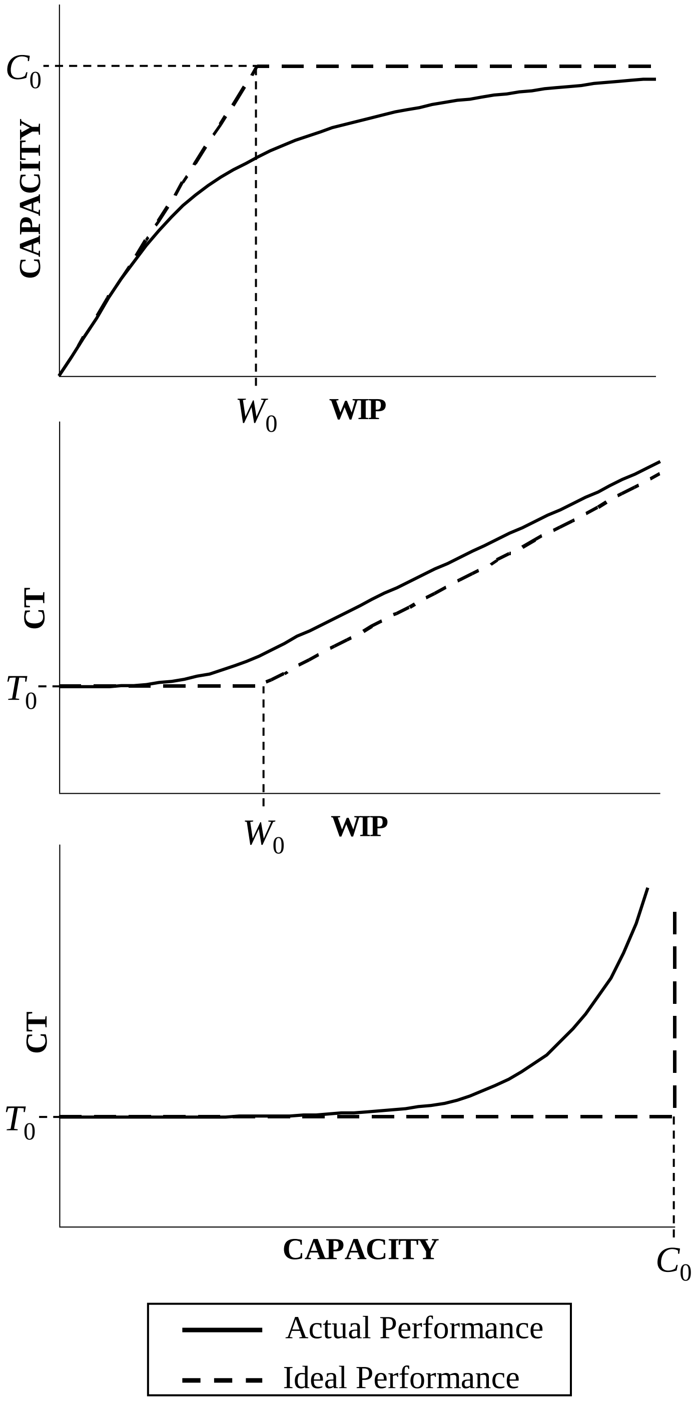


Figure 3.6: Relations between Cycle Time, WIP and Loading

they represent the figures shaped by factors other than variability, that also define the actual performance. Changes in the raw process time T_0, and, consequently changes in critical WIP W_0 lead to new characteristic curves, which can provide better performance while the same or similar sources of varibility remain present.

Traditional cycle time reduction efforts are based on the first insight, targeting a reduction in variability. Our research presented in this dissertation is stimulated on the second insight, i.e., reducing the raw process time, which represents the guiding principle behind the measures discussed in the following chapters.

Additionally, the theoretical limits are basis for another insight: they enable a measurement for CT comparisons (see Subsection 3.3.5).

3.3.4 Role of Variability in Semiconductor Manufacturing

In the previous subsection we stated that variability prevents the manufacturing line's actual performance to match the ideal performance. In this subsection we explore variability sources and their magnitude in more detail.

The major sources of variability are as follows.

- *Planned and unplanned equipment downtimes* lead to far lower equipment availability than in nearly all other industries of mass production. Although the International Technology Roadmap for Semiconductor specifies a process equipment availability of over 94% for 2008 [SIA07b], actual equipment availability is smaller. In our simulation model presented in Chapter 5.1 the least equipment availability is 81% and the average availability is 92%. The impact of these downtimes is severed by its length - in our simulation model the average length of a process equipment downtime is five hours.
- The *reentrant material flow* in semiconductor manufacturing leads to more variability in the lot arrivals at a workstation compared to a flow without reentrancy, because there is no indirect leveling by a specified single previous workstation.
- *Batching equipments* process several lots at the same time resulting in increased variability in lot arrivals at following workstations.
- The *mix of products* in production can also lead to variability, provided the product's routes and/or processes differ significantly.

The first variability cause, equipment downtime, leads to variability in the processing capability of a workstation. Measured in the coefficient of variation of effective process times according to Spearman and Hopp [HS00], the coefficient is significantly bigger than one for many semiconductor equipments, representing high varability according to the classification of variability in the same source.

All four causes together contribute to the variation in lot arrivals. The coefficient of variation in arrivals typically is around one in semiconductor manufacturing, representing moderate variability according to Spearman and Hopp [HS00].

Because of the high variability, the characteristic curve of semiconductor manufacturing fabs is significantly away from the theoretical limits. Therefore, when applying measures to change the theoretical limits, we also have to analyze how this changes the variability impact, i.e., the relative position of the characteristic curve to the theoretical limits.

3.3.5 X-factor as a measurement for CT comparison

The x-factor is defined as the ratio of actual cycle time to raw process time.

$$X - Factor = \frac{CT}{T_0} \tag{3.3}$$

Apart from the definition based on time, as a consequence of Little's Law the x-factor can also be calculated based on WIP, as illustrated in Figure 3.7[3].

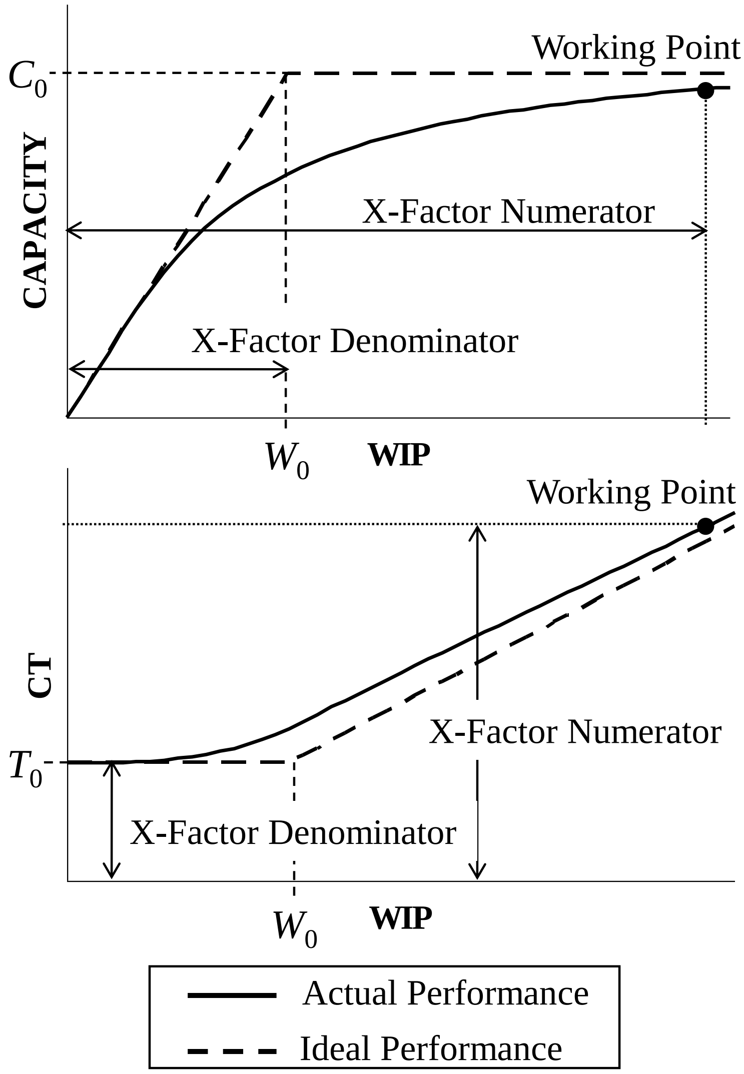


Figure 3.7: Alternate definitions of x-factor

The x-factor is always greater or equal to one and is applicable both on the operation- and on the fab-level. The motivation for the use of the x-factor as a metric for performance comparison is its normalizing property. In this way, routes of different length or fabs having different routes and/or technologies are comparable. To some extend, the x-factor also enables the comparison of operations or a comparison of workstations serving several operations because slow processes naturally have rather long queues and fast processes have rather small queues. However, this is only one influencing factor, and in many cases it is narrow-minded to rely on x-factor comparisons at the operation or workstation level.

[3] Some authors also use the synonym flow factor, e.g., in [Ros99b].

In our analysis the x-factor change associated with the measures discussed in the following chapter is also of interest for a different reason. The change in the x-factor shows whether the improvement in raw process time or total cycle time prevails. Additionally, we have to assess how this figure changes with the measures analyzed because this might change the applicability of x-factor as a benchmarking figure between fabs.

3.4 Elements of the production system

In our research, we examine which changes in the production system can improve cycle time to which extent. In this section we define the elements of the production system, narrow the list of elements down to the ones of interest, and describe advantages and disadvantages of these elements. This description forms the basis of the possible improvement approaches that are discussed in Chapters 6 through 9.

3.4.1 Definition

A production system is defined in a very abstract way as “any of the methods used in industry to create goods and services from various resources” [dB08]. More specifically “a production system may be further characterized by flows (channels of movement) in the process: both the physical flow of materials, work in the intermediate stages of manufacture (work in process), and finished goods; and the flow of information and the inevitable paperwork that carry and accompany the physical flow.” [dB08]

Production systems can be grouped into *project systems* (e.g., ship building), *batch systems* (e.g., food production), and *continuous systems* (e.g., automobile manufacturing).

With respect to the flow of information, automation in the 300mm era of semiconductor manufacturing has brought considerable improvement and set the foundation for enabling a more complex organisation of the physical flow. In our analysis, we regard the basis of a functioning and supportive information flow as being set.

With respect to the physical flow the semicondutor production system can be regarded as a batch system with some subtle adoptions from a continuous system. Apart from the routing which is predetermined (see Section 3.2), the physical flow is defined by WIP control, the equipments and their operation, by the transport between operations, and, by the layout, which we introduce in the following.

3.4.2 WIP Control

Two distinct approaches are applied in semiconductor manufacturing to ensure that lots move through the line in the desired way. One approach defines the lot sequence at equipments, either by dispatching or scheduling, and the other approach, namely lot prority, defines whether significant productivity disadvantages may be acceptable for fast processing of specific lots carrying this priority.

3.4.2.1 Dispatching and Scheduling

Dispatching and scheduling refer to methods that determine the lot sequence at equipments. Whereas dispatching determines the next lot at the time an equipment becomes available for processing, scheduling determines a schedule some time in advance. Because of the complexity inherent to the semiconductor manufacturing production system, the use of dispatching is more common in the industry than scheduling. However, there are some applications of scheduling, e.g., in [LKL02] and [GBL$^+$07].

Dispatching rules can work according to one specific criterion, as, e.g., the FIFO (First-In, First-out)-rule, or the shortest processing time first rule [Ros01b]. More complex rules are possible as well, which can be categorized into

- rules combining several criteria (e.g., the apparent-tardiness-cost-rule),
- conditional rules following different objectives depending on a specific criterion (e.g., fab loading dependent temporary choice of a specific dispatch rule), and,
- multi-level rules sorting lots in several steps after different criteria.

Both dispatching and scheduling have to bring in line different objectives when determining the next lot to dispatch or the lot schedule. These objectives are

- short queueing time at some operations, where quality might be affected by long queueing times,
- high equipment throughput at operations with setups,
- on-time delivery of finished wafers,
- very short total cycle time for high priority lots, and,
- short and predictable total cycle time for normal priority lots.

The first two objectives represent local optimzation targets that have to be considered at specific workstations. They are often worked into the dispatching or scheduling ruleset as limitations, e.g., a maximum number of lots allowed in a workstation queue to limit queueing time or a minimum number of lots between setups.

The last three objectives are fab targets that have to be considered at all workstations. The objectives of on-time delivery of finished wafers on one side and short and predictable total cycle time are not independent of each other, however dispatching and scheduling rules tend to focus on either one. An example of a typical dispatch rule directed towards on-time delivery is the critical ratio rule. This rule favors lots that run behind the necessary progress to achieve the desired delivery date with average cycle time. Additionally the proximity to the end of the line is considered in this rule. In [Ros03] Rose gives a more extensive overview of dispatch rules targeting on-time delivery. Short cycle time is targeted in some dispatching approaches by limiting the consequences of equipment unavailabilities. In this case the momentary capacity situation at workstations of following operations is taken into account on the lot level. If the following operations differ for some lots waiting in queue at a tool, then the momentary capacity situation at following workstations is probably different for these lots and this information can be taken into account for dispatching. An example for the benefit assessment of such an approach is shown in [HF07].

A specific type of dispatch rules are lot start rules. They determine the release of lots into production. Lot start rules try to achieve short cycle and little variation in cycle time by releasing lots in a controlled manner. Some examples are rules that limit WIP like CONWIP [Ros01a] and CONLOAD [Ros99a] rules.

The weight of the objectives and therefore the chosen approach differs by fab profile. A foundry fab is more likely to focus on on-time delivery, whereas a memory fab with little variation in product mix might rather target short cycle time.

The major advantages of dispatching rules are their ease of implementation, their short runtime necessary for instant results and the relatively easy evaluation by simulation. Disadvantageous is the usage of local information on lot and equipment status only. The advantages and disadvantages of scheduling are contrary.

3.4.2.2 Lot Priority Classes

A small share of lots are priority lots. Their task is to provide an exceptionally short cycle time for lots where this is critical. Logically this requires that they are prioritized by dispatching or scheduling. Additionally, the operational policy for these lots might trade some productivity losses for short cycle time. Possibilities are (with generally descending productivity impact)

- holding a tool idle that follows next on the route, enabling instantaneous start of processing once the lot arrives at the workstation,
- holding a load port idle at a tool that follows next on the route, enabling instant loading of the lot onto the tool, and,
- accepting additional setup times although lots of the current setup context are available.

Because of the high productivity impact, the first option of holding a tool idle can be applied for a very limited number of lots only (These lots are often called rocket lots). However, the other two possibilites can be applied more widely. An additional option for saving another bit of cycle time is to hand-carry priority lots between tools which can save a few minutes per operation compared to automated transport. Again, this option is used only for very few lots.

The share of lots that have priority can differ significantly between companies as does their operational policy. ITRS [SIA07b] specifies 6% of total WIP to be priority lots. Out of these 6% one sixth are rocket lots having highest priority and the remainder are hot lots having high, but not top priority.

Use cases for priority lots are widespread; most important are

- development lots for important process improvements,
- qualification and customer sample lots, and,
- lots which have to meet a very aggressive customer due date.

In this dissertation, we focus on normal priority lots, because only normal lot cycle time improvements can address the business needs discussed in Chapter 2. The priority lot cycle time benefit achieved by the replacement of batch tools and by smaller lot size was already subject to our analysis in [Sch07].

3.4.3 Equipments

Two characteristics define the operation of existing semiconductor equipment,

- the batching characteristic, and,
- the tool configuration.

While the tool configuration is of major importance for cluster tool operation only[4], the batching characteristic of a tool is often used as the defining criterion for the operational tool type. A differentiation is made into

- batch tools,
- x-piece tools, and,
- single wafer tools.

In the following we discuss these three tool types and compare both their mode of operation and their specific impact on cycle time.

3.4.3.1 Batch Tools

Batch tools process batches of one or multiple lots at the same time. Figure 3.8 illustrates the common tool design and operation of batch tools. Unlike for other tools it is impossible to keep the carriers containing the lot on the load port while processing takes place, because that would break the space available at the tool front. Therefore carriers are loaded into the internal buffer. Once the full batch, usually spanning several lots in different carriers, is loaded into the buffer area, loading of the batch can begin. The wafers of these lots are then loaded into a batch carrier in the batching area. This batch carrier

[4] There are tool configuration aspects for other tool types, too. E.g., dual boat furnaces are faster than single-boat furnaces and the bath configuration of wet clean sinks determine how flexible a sink can be used. However, these aspects are negligible for our analysis.

is then transfered into the processing area, where the process is performed. Afterwards the unloading process is a mirror image of the loading process.

Typical process applications of batch tools are diffusion or heat treatment processes in furnaces and wet cleans in sinks. These batch tools usually have high throughput capability at comparably low cost, i.e., they have an advantageous cost of ownership. However, batch tools have inherent operational disadvantages, which are

- the long process times inherent to their processing capabilities (see Section 3.4.3.4),
- the long handling times caused by the time-consuming loading and unloading process (see above),
- the batch building time necessary to agglomerate a full batch of several lots, and,
- the batch dissolving time at operations following after a batch operation, caused by the inability of the following workstation to process all lots of a full batch simultaneously. This issue can also be described by the higher lot arrival variability at the following operation caused by a batch of lots arriving at the same time (see Subsection 3.3.4).

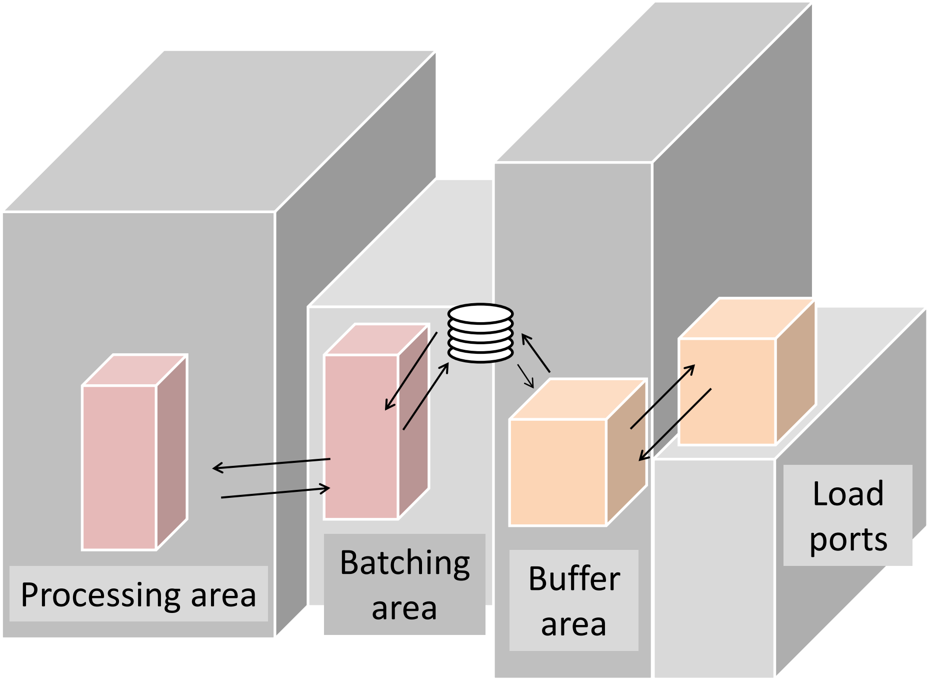


Figure 3.8: Simplified design of batch tools

A recently introduced subgroup of batch tools are mini-batch tools. The prefix *mini* references the smaller batch size compared to normal batch tools that is inherent to mini-batch tools. So far, only mini-batch tools for diffusion and heat treatment applications exist with batch sizes of 25 and 50 wafers.

The operation of mini-batch tools can differ slightly from batch tools. It might not be necessary to have a buffer area, because the load ports can be sufficient to accomodate carriers for two full *mini*-batches enabling continuous processing. Therefore, some mini-batch tools have a simple robot that serves both load ports and the batching area and loads wafers from carriers to the batching area and vice versa. This is similar to the operation of single wafer tools discussed in the next but one subsection.

Apart from the smaller batch size leading to shorter batch buidling and dissolving times, the major advantage of mini-batch tools is their shorter processing time compared to batch tools. However, cost of ownership usually is higher than for batch tools, which has lead to a limited adoption of mini-batch tools in the industry so far. Provided that cycle time advantages can outweigh cost of ownership deficiencies, then this decision flexibility can be used to shape new production systems based on operational considerations.

For most batch tools, also single wafer tool alternatives exist that address the disadvantages of batch tools even better, yet are not widely adopted because of their higher cost of ownership [WEB$^+$03]. Again, provided the better cycle time performance can justify cost of ownership deficiencies, then this decision flexibility for operational considerations can be used to shape new production systems.

3.4.3.2 X-Piece Tools

X-piece tools process batches of x wafers (x < standard lot size) at the same time. Their operation is very similar to batch tool operation. The only significant distinction is the missing internal buffer, as the load ports are sufficient to store all carriers feeding wafers into the tool because of the smaller batch size.

There are relatively few x-piece tools. Implanters used to be x-piece tools with $x = 13$ or $x = 17$, but for process reasons single wafer tools are widely adopted throughout the industry now. A small share of tools are x-piece tools with $x = 2$. Their operational behavior is very similar to single wafer tools.

Because of the decline in use we did not include any x-piece tools in our simulation model.

3.4.3.3 Single Wafer Tools and Cluster Tools

Single wafer tools process batches of single wafers. Figure 3.9 illustrates the tool design and operation of a very simple single wafer tool. The tool consists of the equipment front-end module (EFEM) comprising the load ports, and, the process module. The EFEM is responsible for the link between the material handling system for the inter-equipment lot movement and the equipment internal handling. For this purpose the EFEM contains a simple robot that moves single wafers out of the carrier sitting on the load port into the process module and back after processing. The process module takes and processes only one wafer at a time.

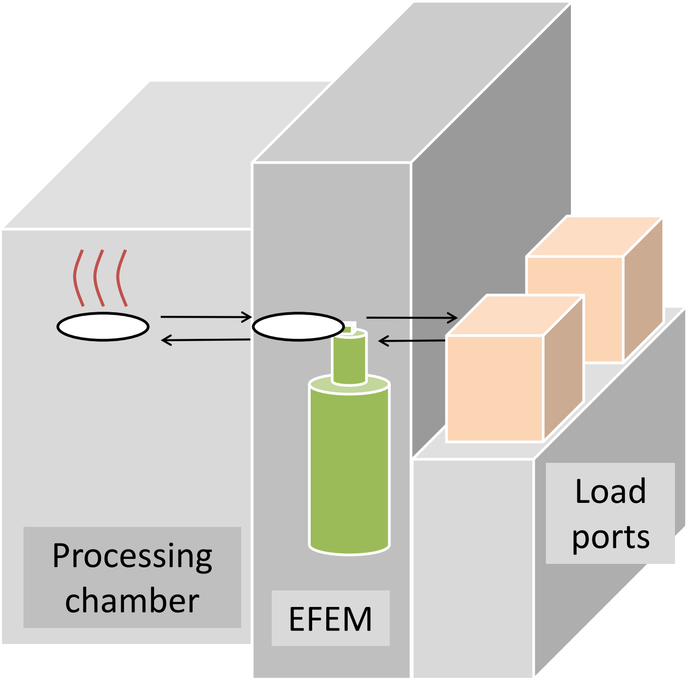


Figure 3.9: Design of a single wafer tool

The predominant share of tools in a fab are single wafer tools. Not all single wafer tools are as simple as in this example. Many single wafer tools contain load locks, which we illustrate in the following cluster tool example. Furthermore, equipments that add at least a second process chamber are called cluster tools, which form a subgroup of single wafer tools which we explore in the following.

Figure 3.10 illustrates the tool design and operation of a cluster tool. The defining characteristic are the more than one process chambers (In this figure we illustrate a cluster tool with three process chambers). By having more than one process chamber, cluster tools integrate process steps following each other or add capacity for the same process step. We define this characteristic of integrating steps or capacity as the *tool configuration characteristic* of cluster tools.

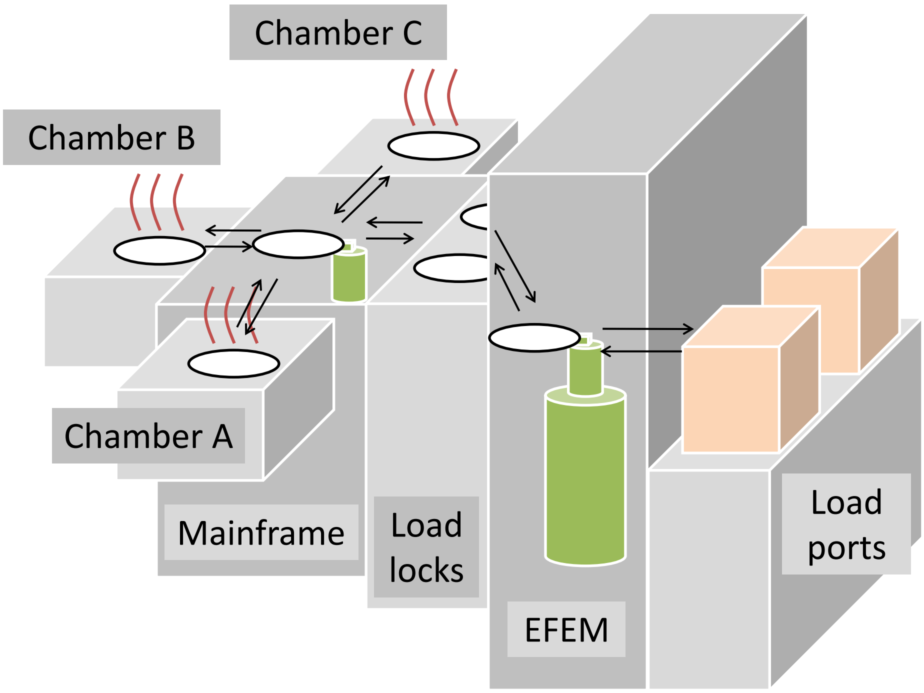


Figure 3.10: Design of a single wafer cluster tool (with load lock)

Specific to cluster tools is a peculiarity of the availability characteristics. Whereas for all other types of equiment downtimes of a subresource lead to the whole equipment being unavailable, this is not a necessary consequence for cluster tools. In cluster tools there can be several chambers for the same process step, hence the unavailability of a single chamber does not block all possible processing paths through the equipment. Therefore processing can still take place, only the processing rate is reduced. One can also say that the equipment is *partially* down in these instances. In some cases the repair of an unavailable chamber can only be performed when the whole cluster is blocked. In these cases the time of partial unavailability enables reduced utilization while, e.g., waiting for a techician to be available. In other cases the broken chamber can be repaired, or a preventive maintenance activity can be performed while the rest of the equipment is productive. This limits the effect of the negetive availability hit by the chamber integration into cluster tools discussed in the next paragraph.

By adding chambers to a tool, cluster tools have brought indisputed advantages to semiconductor manufacturing operations. They save transport time in the case of step integration and cycle time in the case of capacity integration. However, this comes at a cost. Availability characteristics in total and its variability are generally worse for cluster tools (Processing is faster when the same process step can be performed for different wafers of a lot in different chambers at the same time.). In case of the integration of following process steps, downtimes of one step lead to unused capacity of the other steps. And in case of parallel process steps, issues with the mechanical handling take the complete cluster down significantly reducing total capacity available for this process step. Additionally, capacity inequalities of sequential steps lead to unused capacity.

Figure 3.10 also illustrates the integration of load locks into tools. Load locks are independant chambers that perform the transition from atmospheric pressure to vacuum and connect the atmospheric part of a tool with the vacuum part. They are necessary whenever processing shall take place under vacuum conditions and it is unreasonable to perform the pressure change in the process chamber itself. Reasons can be the time and subsequent capacity loss associated with the pressure change or the necessity to keep the wafer under vacuum during the transition between processing steps performed in different chambers. Load locks can be used in both single wafer tools and the subgroup cluster tools.

In Figure 3.11 we show a picture of an actual equipment. At the tool front four orange carriers sit on load ports, the high white tower represents the EFEM and adjacent to its back are loadlock (not visible), mainframe (not visible) and chambers (only one chamber is visible on the right side of the picture.

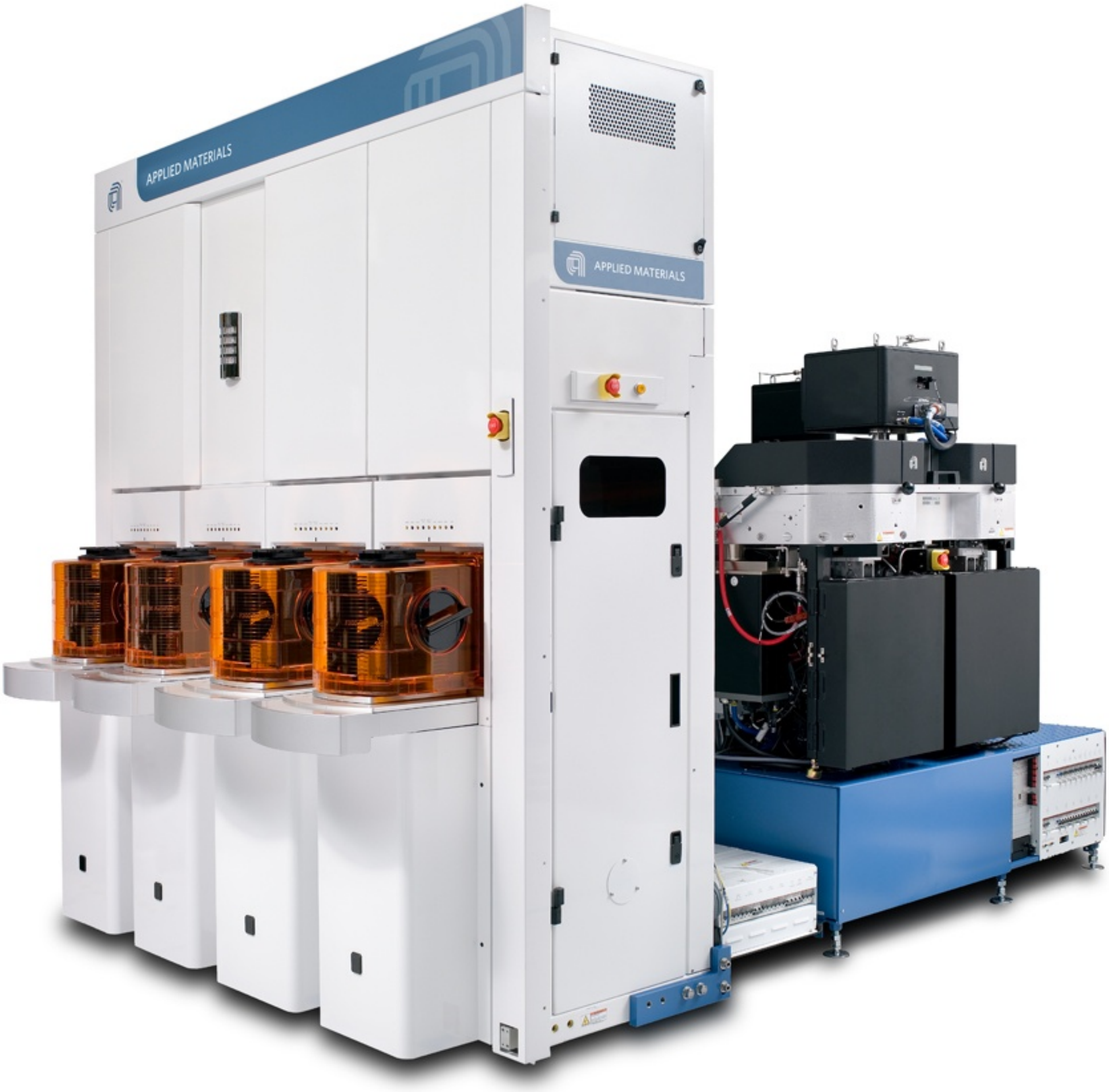


Figure 3.11: Example of an equipment with four load ports: Applied Materials Producer GT (source: BusinessWire)

3.4.3.4 Process Time

The raw process time T_0 represents one of the theoretical performance limits, which we consider for change. In this subsection we define the individual process times at the individual operations and workstations which T_0 consists of and we discuss their qualitative length that is associated with the different tool types presented in the previous subsections.

Definition The process time of a lot at an operation begins when the first wafer is taken out of the carrier and ends when the last wafer is placed back into the carrier. This definition means that the time for the physical or chemical process performed on the wafer is only a part of the process time in our more general manufacturing context. Additionally the time for handling wafers to and from process resources contributes to the process time.

Components of process time Structuring process time we divide it into

- processing PR during which wafers of the lot occupy the limiting subresource(s)[5] of the equipment, and,
- delay DL[6] during which wafers of the lot are in the equipment, but the limiting subresource is occupied by wafers of a different lot, or the limiting subresource is empty[7].

At batch tools, the whole batch occupies the limiting subresource during PR, but at single wafer tools, this is not the case and PR can be further divided into n Processing Intervals PI with n denoting the lot size. PI marks the time interval between consecutive wafer outs of the limiting subresource, independant of the number of entities at the subresource. Figure 3.12 illustrates these process time components for batch and single wafer tools in a simple gantt-chart (For an explanation on Gantt charts see Section 4.3).

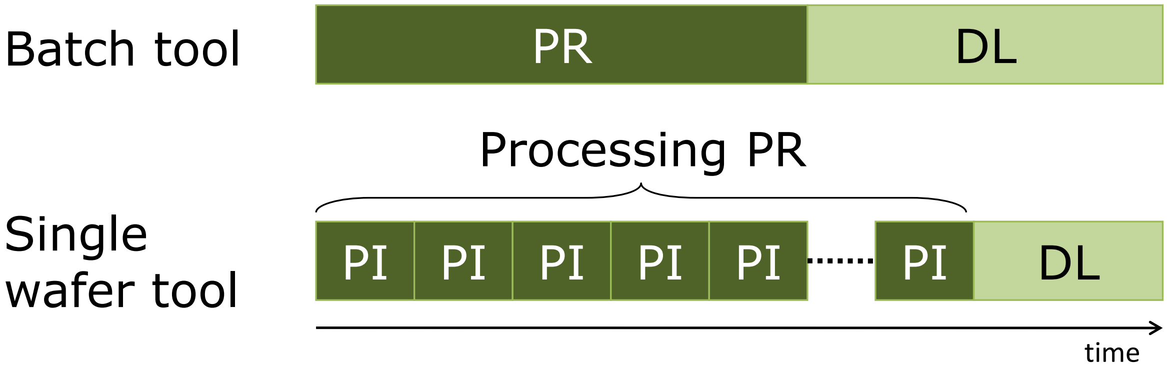


Figure 3.12: Process time components shown for batch and single wafer tools

This classification enables a better assessment of the effects because the smaller components PR and DL have specific characteristics in how they depend on lot size and toolset.

Process time dependencies The length of the physical or chemical process itself has a distinctive effect on process time, but this is in the domain of the process engineer. Longer physical or chemical

[5] The limiting subresource can be one or more chambers performing a process step, but it can also be a robot moving wafers.

[6] In some sources the delay DL is also refered to as first wafer delay.

[7] Perhaps more illustrative, the delay DL marks the time of the first wafer traveling from the carrier to the limiting resource and the last wafer traveling back from the limiting resource to the carrier. At batch tools this includes the formation and division of the batch.

process times are commonly accepted if they increase product performance or yield. Operational considerations are usually of less importance. Therefore we take the physical or chemical process as given and consider the production system design possibilities that define process time.

Considering tool types, the process time of batch tools generally is significantly longer than for mini-batch or single wafer tools. In our fab model batch tools account for 38.3% of process operations, but 55.1% of total PR time and 68.1% of total DL time (summarized over all but the metrology operations). This is due to the long physical or chemical processing time inherent to batch tools and the long loading and unloading process caused by both the wafer reloading to the batch carrier and the carrier handling. The physical or chemical process in alternative single wafer and mini-batch tools is significantly shorter. Additionally they do not have the handling issues, or at least not to the same extent, and therefore enable shorter PR and DL.

Considering lot size, the effect on process time is well known and has been studied in detail by [SRW06] or [Woo96]. Figure 3.13 gives a qualitative overview of process time reduction for smaller lot sizes (In the figure, the relative levels at 25 wafer lot size are general indications for the tool type and not a specific process application.). Process time of batch tools is independant of lot size because the batch contains all wafers of the lot. Process time of x-piece tools improves at multiples of x (in the figure x = 13). Single wafer tool process time improves linearly with lot size reduction[8]. Expressed in subcomponents, the reducution of PR is proportional to lot size reduction whereas DL remains unchanged as traveling time to and from the limiting subresource is not affected by the change in lot size.

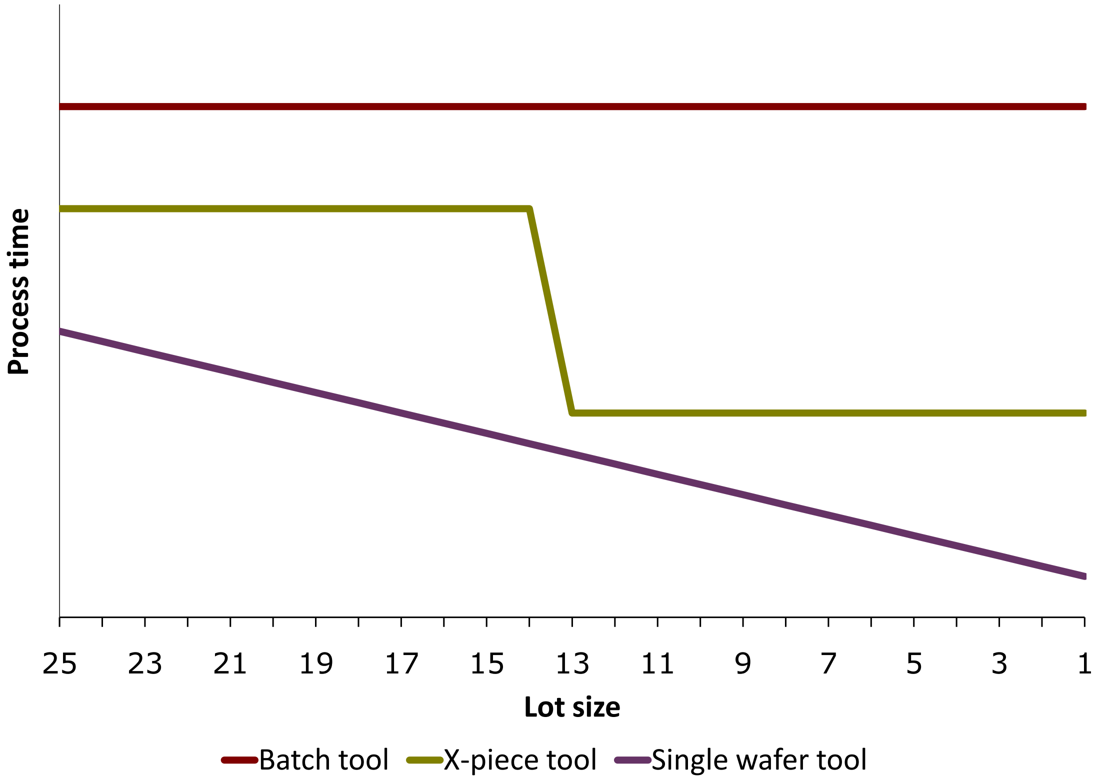


Figure 3.13: Qualitative overview of process time of different tool types at different lot size

Considering tool configuration of cluster tools, there is a high degree of freedom to integrate or disintegrate steps or capacity in production system design. In practice the flexibility is limited because cluster tools usually cannot contain chambers from different suppliers, but this limitation might disappear in the future. Protocol zones, however, form a fixed limitation. It is not possible to integrate steps of different

[8] for some tools the reduction is not completely linear [SRW06]

protocol zones, e.g., of the copper and non-copper areas.

The step integration leads to a reduction in DL, because the wafer can be transported directly from one process chamber performing the first step to another process chamber performing the second step. This saves most of the transport time from the carrier to the limiting subresource and back and, additionally, the time for the inter-equipment transport is saved (see Section 3.4.4.1).

The capacity integration leads to a reduction in PR, because PI is reduced when more entities serve the limiting step. The absolute extent of the reduction in PR caused by capacity integration is reduced with lot size reduction because of the proportional dependence.

3.4.3.5 Cascading

As indicated in the previous subsection, lots or batches are processed by semiconductor equipment in an overlapping fashion, provided $DL > 0$, which is nearly always the case. This operational behavior is called cascading or pipelining.

Figure 3.14 illustrates cascading of lots at single wafer tools in a simple Gantt chart. For illustrative reasons we split DL into two parts. Part 1 (denoted DL p.1) indicates that the first wafer is transported to the limiting subresource and part 2 (denoted DL p.2) indicates that the last wafer is transported back to the carrier. It can be seen in the gantt-chart that the limiting resource is always occupied by wafers of one lot only, as PR does not overlap for different lots. But considering DL as well, processing of lots significantly overlaps, leading to the cascading behavior.

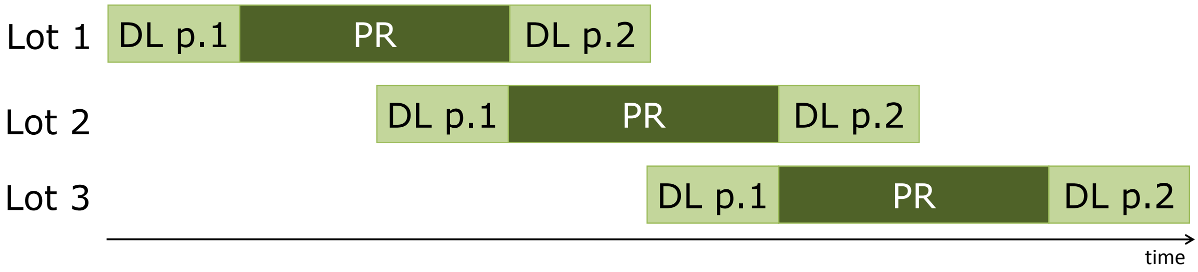


Figure 3.14: Overlapping processing of lots by semiconductor equipment called cascading

The existence of cascading leads to particular operational characteristics. Of specific importance in our analysis is the extent of the cascading as a contributing factor to the carrier exchange time CET, which defines the size of a possible productivity issue at smaller lot size (see Section 7.3.3.1).

3.4.4 Material Handling System

In this section we introduce current material handling system characteristics in semiconductor manufacturing. The material handling system includes the carrier transport between equipments and storage places as well as the storage places for the intermediate buffering of carriers between process operations.

3.4.4.1 Transport

Apart from manual delivery, current generation semiconductor equipment must be capable to receive carriers from floor-based and overhead-transport (OHT) systems [FP00]. While manual handling is intended for exceptional cases only, automated material handling systems perform the bulk load of the deliveries. Some semiconductor manufacturers have chosen floor-based rail-guided vehicles for material handling [WWK+04], but overhead transport systems have seen a much wider adoption in current generation semiconductor manufacturing.

Figure 3.15 shows a picture of an overhead transport vehicle with a carrier and Figure 3.16 illustrates the relative position to the equipments in a sectional view. The track of the vehicle is mounted to the fab ceiling, hence the vehicle moves above the people working in the fab and the transport system does not block manual access to the equipments. Access to the equipments is enabled through the track which crosses the equipments above the load ports and the vehicles ability to lower the carrier on ropes onto the load ports. The vehicle shown in the Figure belongs to the next generation of OHT vehicles and has the additional capability to move and lower carriers from a side position as illustrated in the picture.

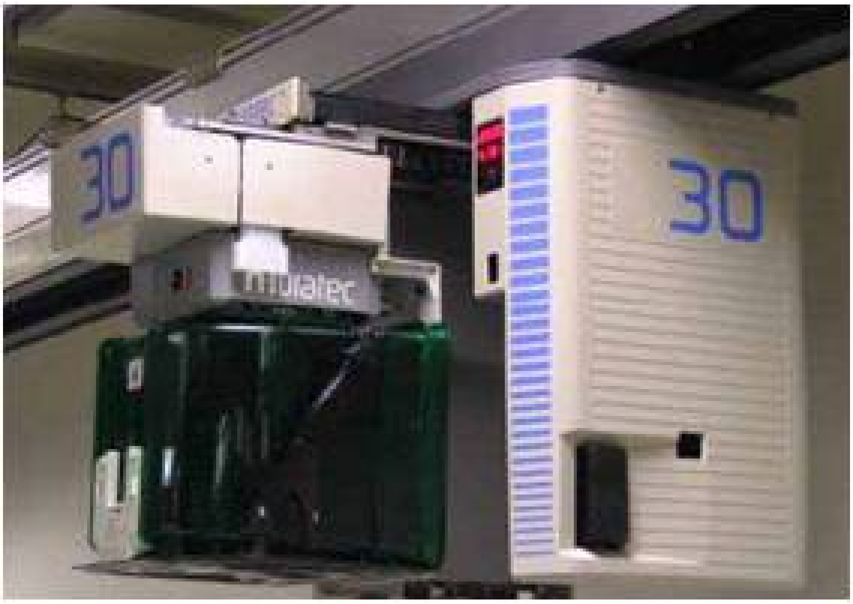


Figure 3.15: OHT vehicle with carrier (green)

Within the OHT material handling systems two subtypes exist.

- In *intrabay/interbay systems* each bay (see Subsection 3.4.5) has its own intrabay OHT system. All intrabay systems are attached to the single interbay transport system through a connecting storage buffer.
- In *unified systems* there is no system separation and the OHT vehicles can travel directly to every point.

Separate intrabay/interbay systems were adopted with the advent of 300mm wafer manufacturing around the year 2000. Due to capacity and speed issues with the storage buffer connection, unified systems are now considered superior and regarded as state of the art.

3.4.4.2 Storage

Carriers have to be stored, while their lot(s) wait for equipments to become available for processing. There are two types of storage places.

- *Stockers* are bulk storage places, which have a capacity of up to 200 carriers. Like equipments, stockers have load ports that connect them to the transport system and an internal robot that moves carriers from the load ports to the storage shelfs. Stockers are used by intrabay/interbay systems as connecting storage. In this case one or two load ports are dedicated for access by each system.
- *Under-track or side-track storages* provide individual storage places for single carriers. They are mounted directly under or to the side of the track (see Figure 3.16). In contrast to stockers these

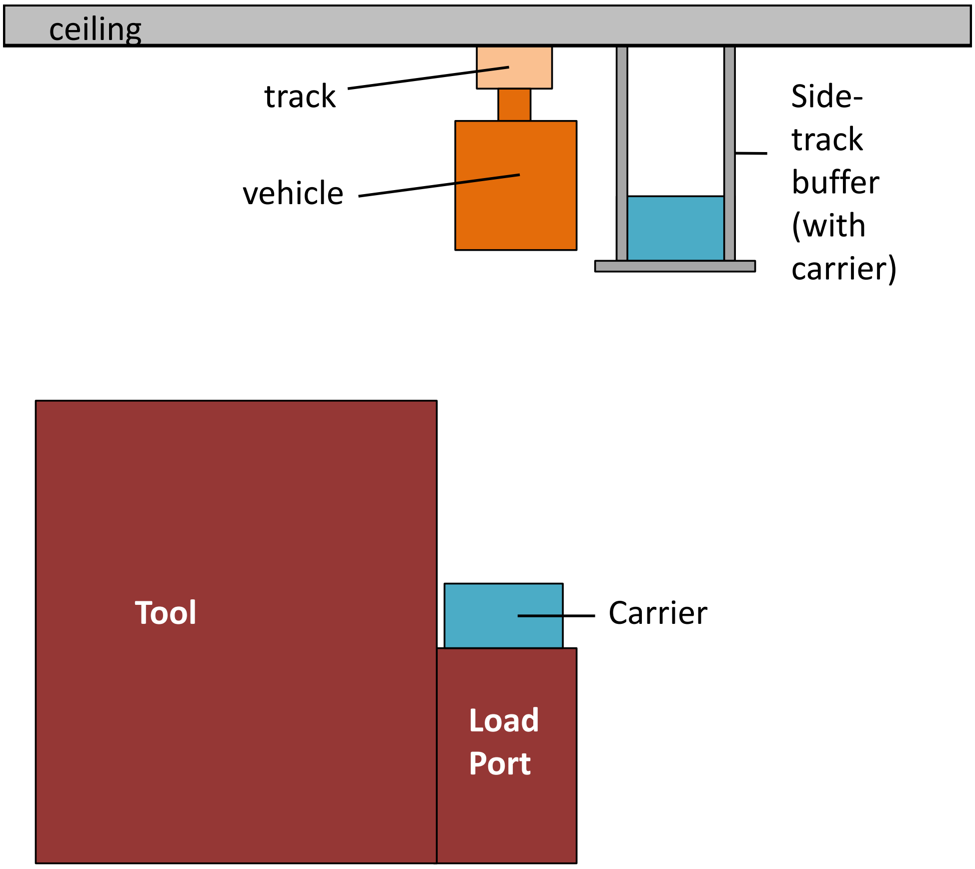


Figure 3.16: Sectional view of OHT system and equipment

storage places do not require floor space and can be located more flexibly. However, total usage is limited by the suitable available track, e.g., junctions, curves etc. are not suitable.

In principle, the access of OHT vehicles to storage places works as for equipment load ports. In case of the under-track storage or stocker load port, the only difference is the shorter vertical distance, and, in the case of side-track storage, the additional movement of the carrier to the side.

3.4.5 Layout

As being typical for batch production systems, semiconductor facilities use a job shop layout (also called functional layout), i.e., like equipments are grouped together and the different tools of a workstation are situated next to each other. Additional motivating factors in semiconductor manufacturing are [Woo00]

1. the low availability making it desirable to group all tools of the same kind together, so that broken tools can be backed up,
2. the impossibility to have one-to-one assignments of tools to steps[9], and,
3. the sharing of a set of facilities by all tools of a workstation.

Another specific of semiconductor equipment limits the flexibility to alter this general layout paradigm for semiconductor manufacturing. The equipments are grouped into at least three protocol zones (lithography area, non-copper area, copper area), which must be separated.

Apart from this general classification, layout details are dependent on the chosen material handling system. Figure 3.17 illustrates the aisle and chase layout typical for intrabay/interbay systems. The equipments (red) are located on both sides of several bays that stem from a middle aisle. In the figure we only show the equipment of three workstations, indicated by different tones of the red color, to avoid overloading. Each bay is served by its own monorail OHT loop (orange) and stockers (blue) connect these loops via the interbay AMHS system (orange; in this example two lanes).

Figure 3.18 illustrates a typical layout for a unified system. The stockers (blue) are located at the walls of the fab, as they are no longer needed for the carrier exchange between the system components. The outer OHT (orange) loop and the bay loop are connected directly by diverts. The equipments are again located on each side of the bay and on the inner side of the outer loop. Whenever short delivery times to an equipment are necessary, then local under-track or side-track storage must be used in this layout.

3.4.6 Degrees of freedom

Most production system design decisions can only be made in the fab design phase. We outline in the following the degree of freedom regarding the changes discussed in this work depending on the point in time when this decision is made.

- A *running fab* offers only limited possibilities for production system changes. The replacement of all equipments of one type (see Chapter 6) is unthinkable, because of the high cost involved and layout issues. The same applies for replacing most cluster tool equipments with a radically new cluster tool design (see Chapter 9). The reduction in lot size (see Chapter 7) seems possible at first glance. However, many side-conditions discussed as challenges in Section 7.3.3 have to be fullfilled to enable reduced lot size without a loss in productivity and it might not be possible to address them.
- A *fab refurbishment* refers to a reuse of an existing clean-room for a new fab with new equipment and a new material handling system. This offers nearly as much freedom as a greenfield fab, just the layout possibilities might be limited

[9] E.g., there may be three tools in a workstation that serve five steps

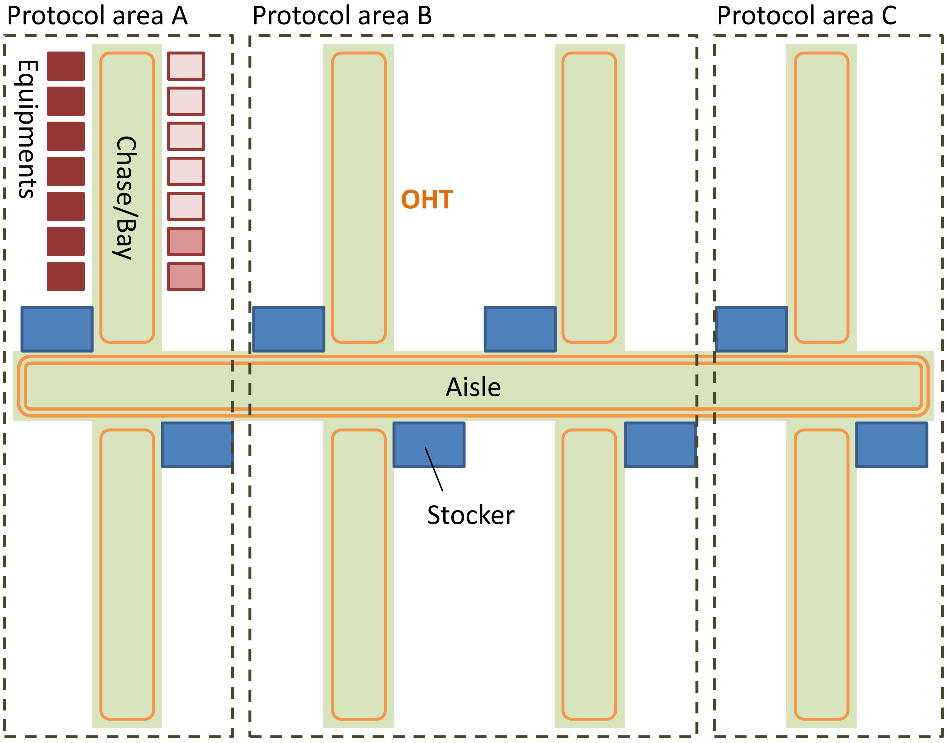


Figure 3.17: Aisle and chase layout for separate intrabay/interbay AMHS systems

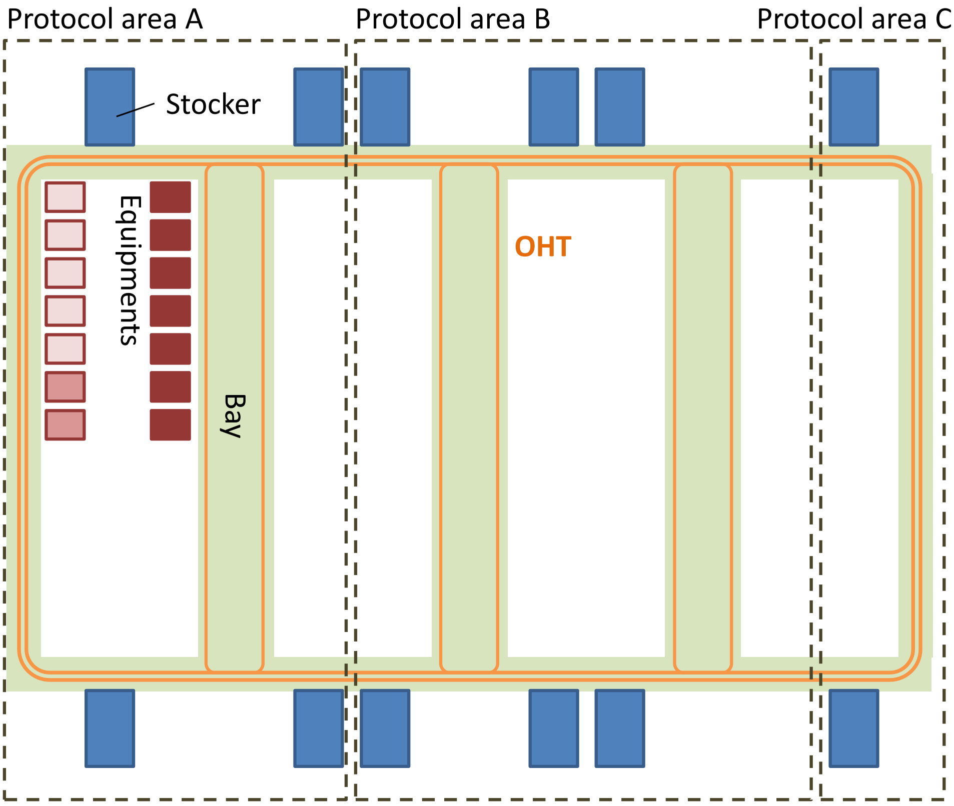


Figure 3.18: Fab layout for unified AMHS systems

- A *greenfield fab* refers to a completely new fab that is constructed on a green field. This offers the full freedom in equipment selection, material handling system design, and, layout definition.

3.5 Cycle Time Components and Performance

In the previous sections, we discussed the origin and dependance of several cycle time components. In this section, we want to recapitulate them and discuss both total cycle time performance of current fabs and the share of its individual components.

3.5.1 Cycle Time Components

On the highest level we divided total cycle time into raw process time T_0 and queue time representing the difference between ideal and actual cycle time performance. On the operational level we divided process time into

- *processing* (PR), denoting the time spent at the actual processing resource, and,
- *delay* (DL) caused by the overlapping processing of consecutive lots often referred to as first wafer delay.

Two queue time components were also identified as

- *transport time* (TT), denoting the time that lots travel between processing and storage locations, and,
- *batch buidling and dissolving time* (BT), denoting both the time spent waiting for other lots to form a process batch, and, the time spent waiting after the batch operation caused by the inability of following workstation to process all lots of the dissolved batch at once.

We classify the remaining part of queue time as

- *remaining queueing time* (QT), denoting the time spent waiting for a processing resource to become available.

Figure 3.19 illustrates the cycle time component attribution.

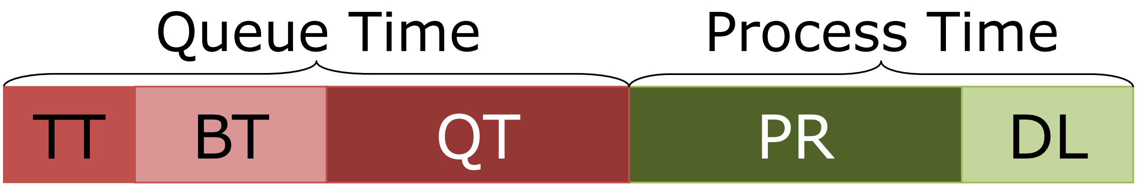


Figure 3.19: Cycle time components

3.5.2 Current Cycle Time Performance

Actual cycle time data is one of the figures that companies rarely share with the public. Therefore the available actual cycle time performance data is limited. The available data has the format days per mask layer [d/ML], which uses the normalizing property of the number of mask operations. A Mask operation represents the lithography step of patterning the wafer, which is then used by the following processes to selectively change the structure of the material below. Because mask operations occur frequently and more or less regulary, they are a possible criterion for the route length. An average route consists of 25-50 mask layers.

In the end of the 1990s Leachman et al performed a fab benchmarking study that included cycle time performance. Cycle time of the benchmarked fabs generally improved over the five-year timeframe, with the best fab slightly below 1.5 d/ML and the average around 2 d/ML [LH00]. In [LKL02] Leachman et al report the resulting cycle time of a cycle time improvement project at Samsung to be 1.35 d/ML in 1999. More recently AMD reported that slightly less than 1.0 d/ML have been achieved as top performance by Fab30 according to [Gro07a].

Another possibility to approximate current cycle time performance is to look at cycle time targets that are defined for the semiconductor industry by inter-company working groups in the International Technology Roadmap for Semiconductors (ITRS). In its Factory Integration chapter of 2007 [SIA07b] they define these cycle time targets for 2008:

- normal lot cycle time: 1.50 d/ML
- x-factor: 3.1
- average AMHS delivery time: 5 min

From these figures we can derive additional details. Based on x-factor and normal lot cycle time we can derive the raw process time to be 0.48 d/ML. And from the average delivery time we can derive a realistic transport time per operation. Conservatively we assume that in 80% of the cases lots are stored intermediately between operations and two deliveries are necessary for one operations. In the remaining 20% lots are delivered directly to the next operation's equipment and only one delivery is necessary. This results in 9 minutes of transport time per operation.

We will compare these insights with the performance of our baseline simulation model in Section 5.4.

3.6 Traditional Approaches to Short Cycle Time

Within traditional cycle time reduction efforts only few target fab-wide improvements. We first discuss these oportunities, specifically *scheduling and dispatching* and *fab scaling*, and then other efforts that lead to spot-improvements.

3.6.1 Dispatching and Scheduling

As discussed in Section 3.4.2.1 dispatching and scheduling tries to optimize lot sequences. One common objective of dispatching and scheduling is to minimize cycle time. Dispatching rules and variants have been assessed and improved extensively and we listed a representative subset of rules and approaches in Section 3.4.2.1.

3.6.2 Fab Scaling

As discussed in Section 3.3.4 variability in availability significantly contributes to queue time in semiconductor manufacturing. One obvious way to overcome this is by fab scaling. In the resulting Mega-fabs there are more equipments in a workstation and therefore the unavailability of one equipment has a smaller effect on the currently available workstation capacity. In [Ros06] Rose analyzes this effect with discrete-event simulation. In a fab of double size the queue time was reduced by up to 70%.

3.6.3 Workstation Capacity Improvement

We distinguish between three ways to improve workstation capacity. These are

- availability improvement,

- throughput improvement, and,
- capacity investment.

The first two ways are continuous improvement efforts that are the daily work of equipment, process, and industrial engineers. These efforts subsume, e.g., reliability improvements, standardized maintenance procedures, optimized process recipes, or more effective wafer sequencing rules for the equipment's internal wafer movement. An examplary application of a subset of these methods is illustrated in [HVG06] and [Sch06].

In some cases it is reasonable to solve a local cycle time problem by the easiest, but most expensive way, which is the investment in another equipment or an equipment upgrade.

3.6.4 Variability Reduction in Equipment Availability

Variability reduction in equipment availability without changing total availability is achieved, e.g., by splitting long maintenance downtimes. An example would be that maintenance tasks that have to be done once a year are not performed all at the same time, but are distributed over the year together with more frequent monthly maintenance tasks. This effort is also a continuous improvement work in the domain of equipment and industrial engineers.

3.6.5 Reduction in the Number of Operations

Operations serve the processing of the product, therefore it is a rare opportunity to reduce the number of operations. The possible opportunities are

- the replacement of a product wafer measurement operation intended to control the quality of a process operation with a measurement on non-product wafers that is run through the same process,
- the removal of particle clean operations, when a sufficiently low particle contamination can be achieved without cleaning,
- the evelopment of in-situ processes, where a second process or measurement is integrated into the previous operation, and,
- the most extreme case of the removal of a complete mask layer, e.g., when the process performed within the mask layer improves performance only and the performance improvement seems unnecessary for a low cost product.

3.6.6 Assessment

With the exception of some spot-improvement efforts which are rather rare, all traditional cycle time reduction efforts target a reduction in queueing time. They led to significant improvements in the past, but seem to reach their limits. The previously cited cycle time of 1.35 d/ML reached by Samsung was achieved with an intrinsic cycle time of 0.9 d/ML [LKL02]. Although the term *intrinsic cycle time* is not explicitly specified, queueing time cannot be part of it and therefore is a smaller contributor than the *intrinsic cycle time*, presumably consisting of RPT, TT, and BT. Therefore the intended systematic production system changes targeting these cycle time contributors are necessary to further reduce cycle time of already well-performing semiconductor fabs.

4 Methods and Tools

In this chapter we present the methods and tools used in the analysis of our production system change measures.

4.1 Discrete-Event Simulation

Discrete-event simulation is our method of choice to evaluate the production system changes of interest. It enables to answer *what-if questions* by creating scenarios with changed conditions and comparing the simulation output of the changed scenarios to the simulation output of the orignal scenario.

4.1.1 Classification and Definitions

In order to assess the potential cycle time benefit of possible production system changes, we have to study how the production system behaves before and after these changes. Figure 4.1 maps different ways to study a system. Because of the high cost, effort and risk involved, both experiments with the actual system and a physical model are unrealistic in the analysis of manufacturing systems, especially if a complete factory shall be analyzed. As this is the case in our analysis of a semiconductor production system, we have to abstract manufacturing operations in a mathematical model. Analytical solutions using mathematical methods to obtain the system's performance figures of interest are only possible if the model is simple enough. One possibility of an analytical solution is queueing theory which we introduce in Section 4.2. However, apart from isolated application possibilites, semiconductor manufacturing is too complex to allow for simple and realistic models that can be evaluated analytically. Therefore, we rely on simulation to evaluate the system's model numerically and use the data gathered to calculate the performance figures of interest.

Banks et al define *simulation* as "the imitation of the operation of a real-world process or system over time. (...) [S]simulation involves the generation of an artificial history of a system, and the observation of that artificial history to draw inferences concerning the operating characteristics of the real system." [BINN00]

The behavior of the system studied with simulation is defined in the *simulation model*. This model takes the form of mathematical and logical relations based on assumptions reagarding system operation [Law07]. Simulation models may be classified as being static or dynamic, deterministic or stochastic, and discrete or continuous [Ros04].

- Static models represent a system only at one particular point in time, whereas dynamic models represent systems as it evolves over time.
- Deterministic models do not contain any probabilistic components, therefore they produce a unique result. In contrast, stochastic models have random inputs that in turn lead to random outputs. Therefore the output must be treated as an estimate of the true results. The confidence, that can be put into the output of stochastic simulation models, increases with the length of the observation, because the influence of randomness decreases.
- In discrete models, the system's state variables change instantaneously at separate points in time, whereas in continuous systems the state variables change continuously over time.

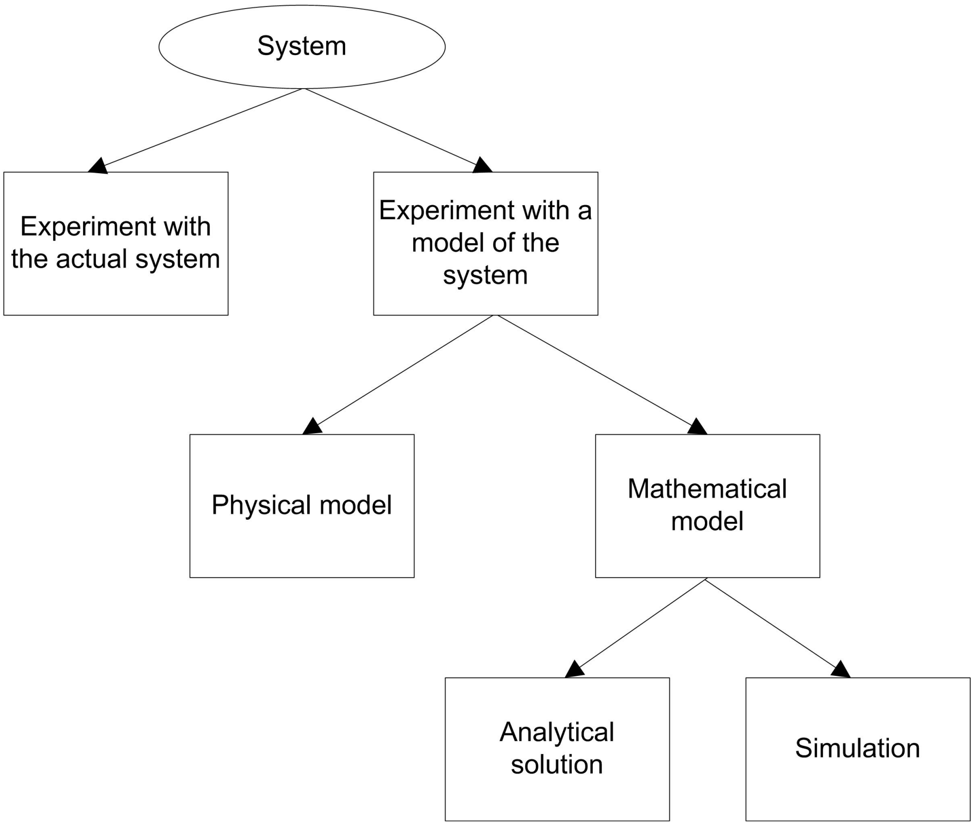

System
Experiment with the actual system
Experiment with a model of the system
Physical model
Mathematical model
Analytical solution
Simulation

Figure 4.1: Ways to study a system according to [Law07]

The simulation model built for our analysis is dynamic, stochastic and discrete. Hence, we use discrete-event simulation for our simulation analysis.

4.1.2 Steps in a Simulation Study

According to [Ros04] and [BINN00] the individual steps of a simulation study are

1. Problem formulation and planning of study,
2. Model conceptualization and data collection,
3. Model translation,
4. Verification,
5. Validation,
6. Experimental design,
7. Simulation runs,
8. Analysis of results, and,
9. Documentation, presentation and possibly implementation.

Some items in this list of items are self-explaining. We describe the simulation-specific steps two to six in more detail in the following.

The step *model conceptualization* refers to the abstraction of the features of the system under study that are essential from an operational point of view. The description of material flow and operation of the different equipments in Chapter 3 falls under this step. The data collected adequately for the abstract model is then entered into the data format required by the simulation software in the next step *model translation*.

Verification of the simulation model follows. Pilot runs of the simulation model are performed and it is checked whether the simulation performs properly. This can be an iterative step requiring quite some debugging. The successive phase is the model *validation*. This means that the model is validated, i.e., performance parameters are compared to the real system to check whether the model accurately represents a real system.

In the *experimental design* phase the alternatives that are to be simulated are determined. For the evaluation of production system changes as we intend, changes are made to the model to represent different system designs and these different models are then run individually for a later comparison of the results. The experimental design also refers to decisions that determine the credibility of the results, i.e., the length of the simulation run, the length of the initialization period, and the number of replications must be determined.

4.1.3 Advantages and Disadvantages of Analysis with Simulation

System analysis with simulation has many advantages, but also some disadvantages. The advantages are (see [Law07] and [BINN00]):

- Many real-world systems elude accurate description by mathematical models that can be evaluated analytically. Simulation is more widely applicable and not subject to this limitation. Therefore it often is the only possibility for an investigation.
- Insights into the operation of the real system can be gained. This includes, e.g., the interaction of variables, the performance sensitivity of variables, or, the reasons for the occurence of certain phenomena.
- Different operating conditions as, e.g., new dispatching rules, operating procedures, loading changes etc. can be tested without disrupting ongoing operation in the real system.
- Alternative system designs can be tested to quantify a possible performance change without actually creating these alternative designs.
- The experimental conditions in simulation can often be controlled better than in an experiment performed with the real system.
- Simulation time can be compressed to enable studying the system over long time or expanded to enable studying the detailed working of a system.

Disadvantages of simulation are:

- Simulation runs of a stochastic model produce output which is based on random variables. Therefore simulation runs provide only estimates of the true performance characteristics of a simulation model for a specific set of input parameters. In order to limit the impact of randomness on the results either several independent runs are performed or the simulation model is run for a very long time.
- The development of simulation models can be time-consuming.
- Simulation requires run-time to generate results and does not generate results at the press of a button.

In our analysis, we use simulation extensively because the semiconductor manufacturing system cannot be described analytically. For some changes under consideration analytical models are applicable under some assumptions, however, and whenever possible we use this alternate method which is unsusceptible to the drawbacks discussed.

4.1.4 Utilization of Simulation in Semiconductor Manufacturing

Discrete-event simulation is used for a number of applications in semiconductor manufacturing. The following list might not be all-embracing, but gives an overview. Simulation is used to

- study the impact of dispatch rules on fab performance, e.g., in [Ros03],
- to generate characteristic curves, e.g., in [BCFR97],
- to assess the benefit of capacity additions,
- to apply simulation-based scheduling in some instances, e.g., in [Rin07]
- to study and optimize the AMHS layout, e.g., in [RPQ05], and,
- to optimize throughput of cluster tools, e.g., in [UR07], [LD05], and [Ben08].

4.1.5 Simulation Software

For our simulation experiments we use the software Factory Explorer 2.8 from WWK [WWK03]. Factory Explorer contains a discrete-event simulator, an MS Excel interface to implement the simulation model and it provides standard ouput reports in MS Excel that provide key figures of interest. It is also possible to implement or change the simulation model in text-files, which provides flexibility regarding automation and to write user-specific output reports that are based on simulation output stored in the same format.

Factory Explorer was developed specifically for semiconductor manufacturing although it is not limited to this application. Therefore it contains specifics that are necessary in semiconductor manufatcuring. The modelling capabilities include, e.g., rework, scrap, splitting, binning, and assembly, which enable simulation models encompassing both front-end and back-end manufacturing. In our simulation model, we do not use all of these features, though.

A substantial advantage of Factory Explorer compared to competing products is its speed. This is essentially important as we want to simulate at small lot sizes, which leads to a significant increase in the number of objects and consequentually presents a significant blow to run time.

4.2 Queueing Theory

Queueing theory is a powerful tool for analyzing queueing behavior as its closed formulas permit exact insights into the relevant factors. Of course, this statement is limited to queueing systems for which these closed formulas exist but it is a very valuable starting point. For those cases where queueing theory cannot provide answers, simulation experiments are the method of choice. These cases include systems where arrival or processing process cannot be approximated with probability distributions or systems where the sequence of operation does follow specific rules as, e.g., for setup avoidance.

The advantages of queueing theory compared to simulation are the provision of exact results, the direct insight in the relevant factors, and the direct calculation without long run times. The main disadvantage is the limitation to systems that perform in a way that can be assessed with queueing theory. This disadvantage will become more apparent in the following discussion.

Queueing systems can be characterized by Kendall's notation as A/B/m with A denoting the distribution of arrival times, B describing the distribution of process times and m denoting the number of parallel machines [GH98]. Semiconductor manufacturing is independent of queuing discipline, therefore we do not specify the queuing discipline in notation or discussion.

In the semiconductor industry, interarrival times to a given workstation are usually highly variable and some research suggests abstracting them as exponential. Therefore, we will use an exponential (Markovian - M) distribution for the arrival process. For the rare exceptions (e.g. workstations with preceding batch operations) that necessitate more general arrival distributions, Whitt [Whi93] gives helpful queueing time approximations.

Process times are usually thought of as being fairly constant. This is a reasonable assumption for the pure process time that we assume as being deterministic. However, if we look at process times from a logistical point of view we have to account for additional delays. These delays can be setups or preventive and corrective maintenance and inflate the process time to a higher value called effective process time t_e. Because these delays only occur for some lots, the effective process time includes variability quantified in its coefficient of variation c_e. The resulting effective process time is a random variable following a general distribution (G) shaped by t_e and c_e.

Refering to the above notation, there are many M/G/m queueing systems in a semiconductor fab. In the special case of a single tool, the M/G/1 queueing system, the queueing time QT is given by

$$QT(M/G/1) = \left(\frac{1+c_e^2}{2}\right)\left(\frac{u}{1-u}\right)t_e \tag{4.1}$$

where u denotes the utilization of the tool. For queueing systems featuring parallel tools, there is, in general, no closed formula that provides exact solutions. However, the approximation

$$QT(M/G/m) = \left(\frac{1+c_e^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)t_e \tag{4.2}$$

gives excellent results although some improvements can be made for low utilizations [BCH79]. In our case the exactness of the above approximation is sufficient though.

Different types of downtimes have different impacts on variability. The most important distinction is between preemptive and non-preemptive downtimes. Preemptive outages occur right in the middle of a process. Typically, these are outages for which there is no control as to when they happen (e.g. failures). In contrast, non-preemptive outages require the tool to be idle before they can happen. This means that we have some control as to exactly when they occur. This is usually the case for planned maintenance activities or setup times.

Setup times elude easy analysis. Intelligent setup and dispatching policies seek to avoid setup occurrence and it is therefore difficult to generally estimate how setup frequency would change with changes in the production system. Because of this lack of clarity we do not analyze the impact of setups with queueing theory.

4.2.1 Preemptive Downtimes

For preemptive downtimes, [HS00] provides formulas for the parameters of interest,

$$t_e = \frac{PR}{A} \tag{4.3}$$

for the effective process time t_e and

$$c_e^2 = c_0^2 + (1+c_r^2)A(1-A)\frac{m_r}{PR} \tag{4.4}$$

for its coefficient of variation c_e. In these formulas A denotes the equipment availability, c_0 the coefficient of variation of the processing time PR (which equals zero for constant PR), c_r the coefficient of variation of repair times and m_r the mean repair time. It is important to note that the availability A refers only to preemptive outages. It is defined as

$$A = \frac{m_f}{m_f + m_r} \tag{4.5}$$

by the parameters mean time to failure m_f and mean time to repair m_r.

4.2.2 Non-preemptive Downtimes

There is no simple queueing formula for non-preemptive outages that is applicable for our analysis. It is required that the downtimes can be attributed completely to a lot, which does not match our definition of preventive maintenance.

Some helpful assumptions on non-preemptive outages can be gained from intuition though:

- At lower utilizations the variability impact of non-preemptive outages is lower than for preemptive outages because longer downtime intervals often happen with no lots in queue.
- At higher utilizations the difference between preemptive and non-preemptive downtimes will disappear because the one lot that can finish processing represents only a small part of the queue.
- The part of effective process time t_e in addition to t_0 is utilization dependent. For low utilizations it is smaller for non-preemptive than for preemptive outages.

4.3 Gantt Charts

A Gantt chart named after the industrial engineering pioneer Henry Gantt (1861-1919) is a bar chart that illustrates a sequence. In Gantt charts the x-axis represents the advancing time. The y-axis lists the resources of interest and their utilization by time is shown horizontally to the resource name on the y-axis. The product utilizing the resource can be marked by either inscription or color. Additionally, it can be illustrated whether the product blocking the resource actually is in process or is waiting for either processing or transport. This is often indicated by a different bar size or a different shade of color. In our Gantt charts, we use both inscription and color to mark the product utilizing the resource and we use a lighter shade of the same color to distinguish waiting from processing.

Figure 4.2 shows an examplary Gantt chart. The chart shows the utilization of three resources A to C with three products P1 to P3. Resource A a first processes P2, is then blocked by P2 until P2 is further processed by Resource B. Directly afterwards Resource A processes P3, which also blocks the resource after processing for some waiting time. Afterwards Resource A stays idle. The utilization of the other resources is read likewise.

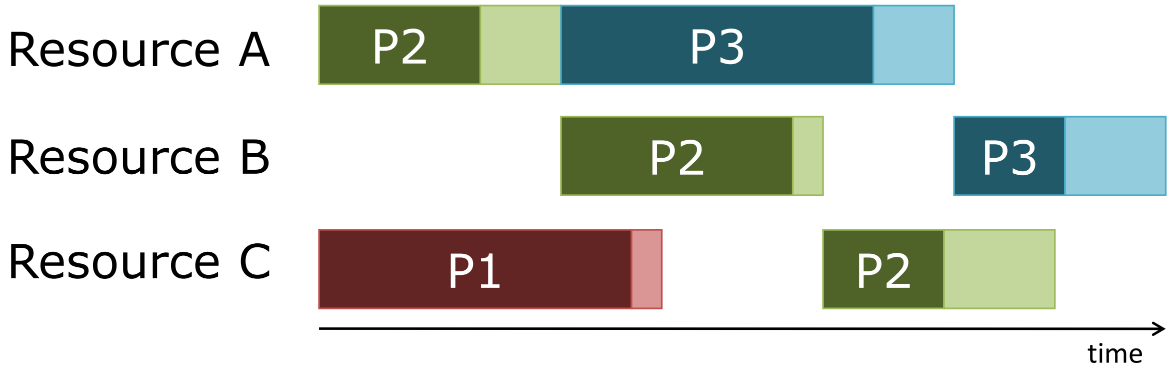


Figure 4.2: Utilization of three resources illustrated in Gantt chart

5 Baseline Simulation Model

Our baseline simulation model has to fulfill several criteria. First, it has to be representative for a typical 300mm semiconductor front-end manufacturing environment. Secondly, the model has to be sufficiently detailed and complex to represent all factors that influence the outcome of the production system changes under consideration. And thirdly, the model has to be simple enough to keep simulation run time at an acceptable level especially at reduced lot sizes.

Information on actual semiconductor routes, workstations, equipment throughput and equipment availability is not shared by semiconductor manufacturers. Therefore we constructed our model based on a model compiled by ISMI in the 300mm planning phase [CA00], which is publicly available. Because of the decade passed since the creation of this model, we adjusted the equipment performance parameters throughput and availability to represent current performance.

5.1 Fab profile

Conforming to our short simulation run time target we have kept the size of our simulation model fab at a small to medium level with respect to flow complexity and wafer starts. The following list gives an overview of additional fab profile characteristics of our simulation model's baseline version.

- Product profile:
  - Products: one product.
  - Flow complexity: 23 mask layers; 7 metal layers.
- Route:
  - 209 process steps.
  - 165 metrology and 4 test[1] steps.
- Tools:
  - 38 process workstations, 175 tools.
  - 9 metrology workstations, 106 tools.
  - 2 test workstations, 40 tools.
- Max. static capacity: 960 waferstarts/day (40 lotstarts per day at 24 wafer lot size).

Two specifications might sound a little odd at first. First, in its baseline configuration the model includes only one product in order to exclude disturbances caused by product variety. We include a scenario with multiple products in Section 7.4.3 though.

Secondly, although current standard lot size is 25 wafers, we have chosen a lot size of 24 wafers as our baseline. However, this change enables a constant batch loading for the lot sizes under consideration of 24, 12, 6, and, 2 wafers. Otherwise we could not distinguish the effects of lot size reduction and batch loading changes. Only the baseline of 24 wafers produces pure results for the lot size reduction.

[1] Test operations are a specific form of metrology steps where the electrical paramters of the individual chips on the wafer are tested instead of general measurements on the wafer.

5.2 Model Conceptualization

Within Chapter 3 we already abstracted equipment operation. We divided process time at an equipent into processing PR summarizing a per-wafer or per-batch process time on the lot level and a delay DL that occurs per lot. In the simulation model, this data is implemented exactly in this way, as a per-wafer or per-batch process time and a per-lot delay, which all are specific to one operation served by one workstation.

Regarding equipment availability, we distinguish two downtime reasons. Failures are downtimes for which we have no control as to when they happen. Therefore we model them with exponentially distributed times between failures and exponentially distributed times to repair. The two figures are specified by a mean time between failures (MTBF) and a mean time to repair (MTTR). In case of downtimes due to preventive maintenance activities, however, we have more control of the timing. These are scheduled activities happening after predefined time or specified processed unit intervals and their length has less variance. Therefore we define them at fixed intervals and use a triangular distribution with a +/- range of 20% for their length. We specify up to six different types of preventive maintenance avtivities per equipment that happen at different intervals (daily to bi-yearly). In order to limit the variability in workstation capacity, we distribute the maintenance activities of the same type over the interval period, i.e., for a workstation with six equipments, we schedule the daily maintenance activities in 4 hour intervals on the different tools[2]. As noted earlier, long and frequent downtimes are typical for semiconductor equipments. In our case, the smallest availability in our model is 81%. All downtimes are modelled as non-preemptive, as Factory Explorer does not allow for preemptive downtimes. This is not a troublesome limitation because at reasonable loadings this does not make a significant difference (see Section 4.2.2).

As an example we list the downtimes of the workstation “E_Dry_Etch_Oxide”.

- Failure: MTBF: 150 hrs; MTTR: 5 hrs
- Daily preventive maintenance: 0.5 hrs
- Weekly preventive maintenance: 2 hrs
- Monthly preventive maintenance: 4 hrs
- Half-yearly preventive maintenance: 10 hrs
- Yearly preventive maintenance: 10 hrs
- Clean after 8000 units: 10 hrs

Due to limitations of the availability data and a not fully developed modeling capability in Factory Explorer, we were unable to model the partial availability at cluster tools. Hence, downtime modeling is uniform for all equipments with complete equipment downs only.

As transport time currently is not a substantial cycle time contributor, we chose not to simulate the material handling system in detail, but to rather include it as a delay per operation. We model transport times with a static delay of 0.15 hrs per actual operation, i.e., there are no transport times associated with skipped metrology operations. Considering conservative 20% direct equipment-to-equipment transports and 80% transports with intermediate storage, the length of 0.15 hrs equaling 9 minutes is consistent with the ITRS [SIA07b], which specifies an average transport time of 5 minutes.

Regarding dispatching we use FIFO-dispatching at all workstations to avoid side effects due to dispatch rule parameters such as due dates. Additionally, we use setup avoidance at applicable workstations and at batch equipment workstations we always wait for the full batch. Regarding only 18 batching contexts at 9 batch workstations with 80 operations the waiting time for additional lots is short compared to the batch processing time, therefore this waiting policy leads to a shorter cycle time.

We have included setup times in the baseline model only where it is definitely unavoidable: At implant operations. For other potential setups, we created separate scenarios which we discuss in Section 7.4.7.

[2] We use the GroupClock feature of Factory Explorer to model this maintenance schedule.

It is unclear how sampling at measurement operation needs to change with lot size reduction. For some applications representing simple process monitoring, the sampling which requires a measurement after a fixed number of wafers is unlikely to change. Other applications might at least depend on sophisticated scheduling efforts to avoid to increase the number of necessary measurements. However, metrology cycle time decreases very significantly with lot size as will be apparent in Section 7.4.2. Therefore, the necessity to assess the question of a possible small increase in measurements in great detail does not arise.

In our simulation model we took a general approach with combined skip rate increase and wafer sampling reduction assuming that the total measurement time stays the same. This might, e.g., mean that the number of wafers that are actually measured stays the same, although this is not necessarily a consequence of the sampling setup used. Table 5.1 specifies the sampling characteristics used in the simulation for all metrology operations independently of their applications.

Table 5.1: Sampling at different lot sizes under consideration.

Lot size	Lot skip rate r_{skip}	Relative lot processing time
24	70%	1
12	77.5%	0.67
6	85%	0.5
2	92.5%	0.33

5.3 Verification

The verification of the simulation model was performed in two steps. First, all results that are not subject to variation are checked against pre-determined values, as, e.g., pre-calculated utilizations based on toolcount and availability in Figure 5.1. Then, plausibility checks were performed to assess whether results influenced by randomness (queueing time) are reasonable compared to single operation assessments performed with queueing models and single operation simulation as in [SR07b].

5.4 Validation

Figure 5.2 illustrates the cycle time performance of the baseline simulation model by component at 92.5% fab loading[3]. Our 92.5% baseline loading represent an almost fully loaded fab, because some extra-capacity at the bottleneck is always needed to buffer for variability, otherwise cycle time goes through the roof.

The cycle time of the baseline scenario equals 0.78 days per mask layer at an x-factor of 1.75. We compare these values with the available actual performance data gathered in Section 3.5.2. The cycle time achieved in the simulation is slightly better than the best performance achieved in today's fabs. However, we did not include all sources of variability into our simulation model. E.g., we exclude lot holds from our model which have a significant impact on cycle time, yet are independant from any of the discussed changes. Therefore, although cycle time performance of the baseline model is challenging, we judge this as a convincing baseline model for our research because the far-reaching changes of lot size reduction or tool replacements are only reasonable compared to well-performing standard fabs. If a

[3] The lower and upper limit of the 95% confidence interval for total cycle time is within 1-2% of the average, therefore we do not show confidence intervals in any total cycle time chart.

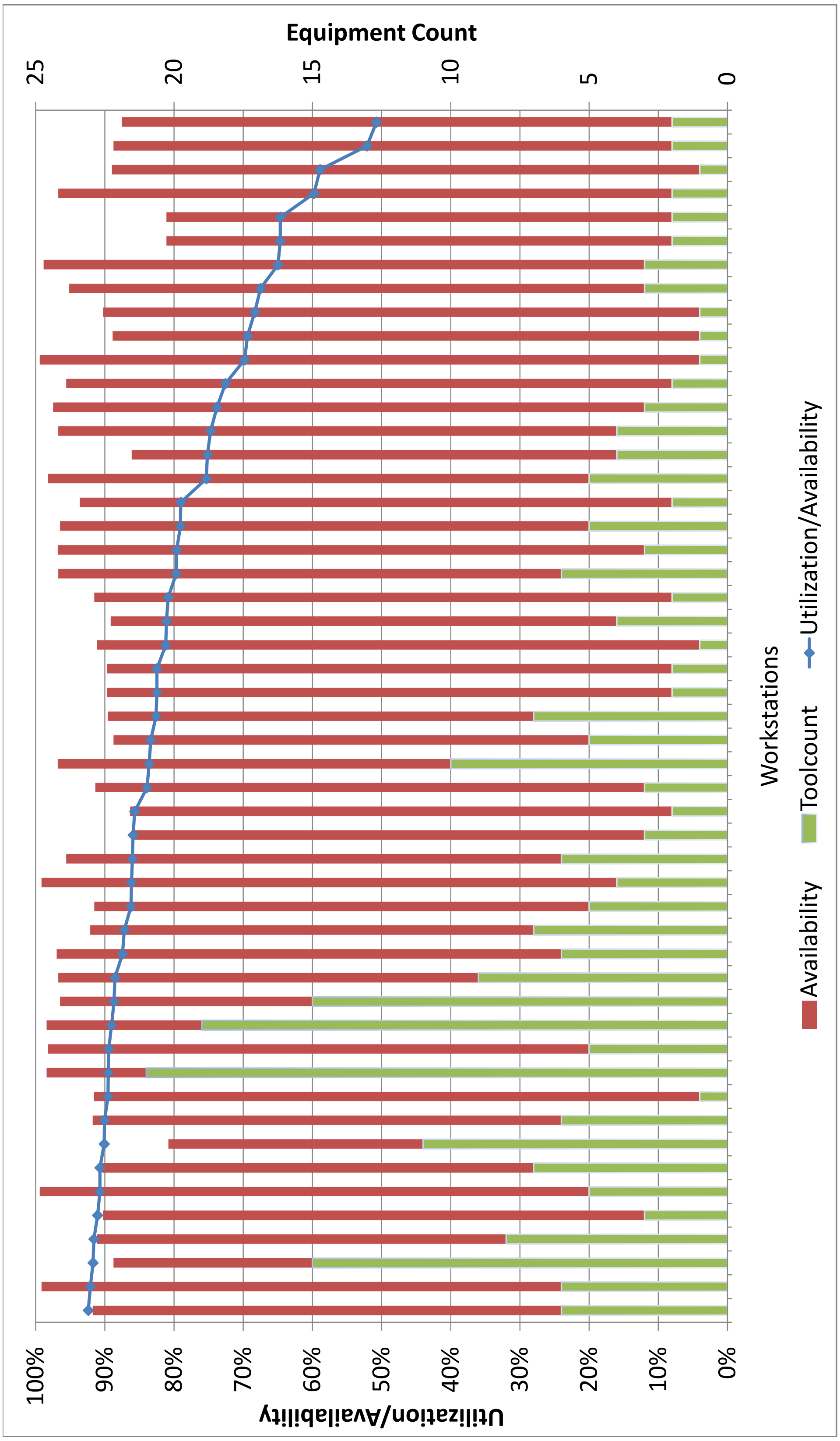


Figure 5.1: Utilization over availability at 92.5% loading

fab's cycle time performance could be significantly improved by conventional methods, then that would be more favorable in comparison.

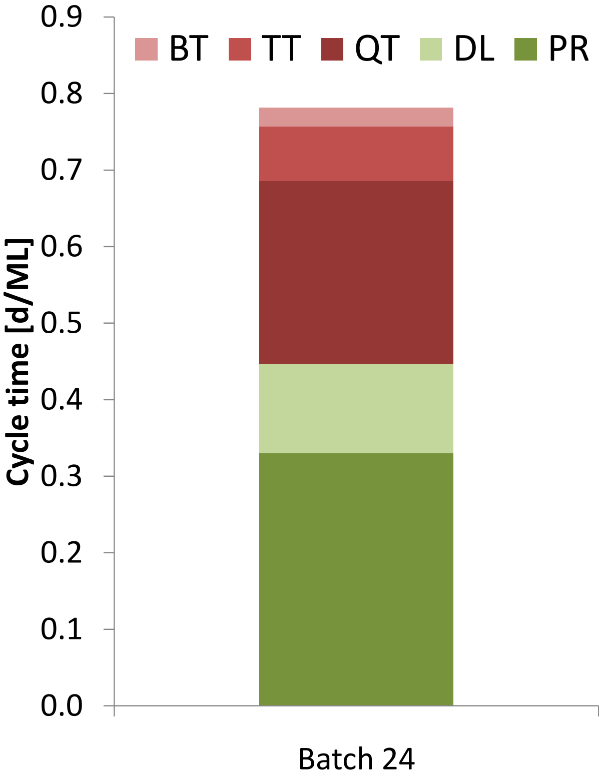


Figure 5.2: Cycle time performance of baseline simulation model at 92.5% loading

5.5 Experimental Design

Each of the following chapters on the assessment of possible changes in the production system contain their own section on the individual experimental design. Some parameters are set universally, though.

As our baseline loading we chose 92.5% and we show most results only for this loading. To allow for the analysis of loading-dependant effects, we we vary the fab loading between 50% and 97.5% in 2.5% increments in all experiments. At each loading we run the simulation for 51 years including a 1 year warm-up period that is discarded for the results (Figure 5.3 representing a cycle time by lot exit chart confirms that the warm-up period is finished before a full year has passed.).

We name all our scenarios with the defining equipment type and the lot size. E.g., in case of our baseline model the scenario name *Batch 24* refers to the usage batch equipment for a defined subset of operations and to the lot size 24.

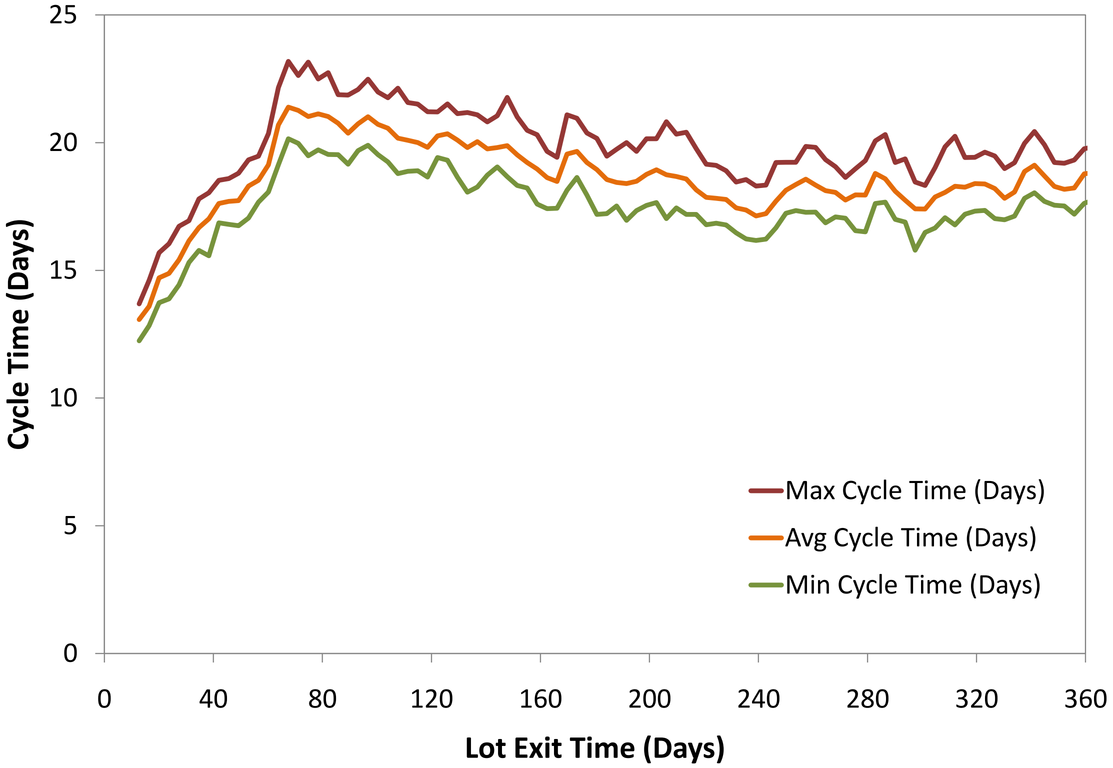


Figure 5.3: Cycle time by lot exit times during warm-up period of baseline simulation model at 92.5% loading

6 Replacement of Batch Tools

In this chapter, we analyze the cycle time gain realized by replacing batch tools with mini-batch or single-wafer tools.

6.1 Literature Review

The cycle time benefit of batch tool replacements with mini-batch and single-wafer tools has been discussed in a number of publications. In the 1990s these discussions were based on 200mm equipments and the emerging cluster tool adoption. Most publications note the benefit in process time, but few actually estimate the total cycle time. As one exception Wood assesses a replacement of most batch tools[1] with single wafer tools in [Woo97] with simulation. His results show that a reduction of 50% in cycle time is possible when single-wafer equipments are used in combination with cluster tools and in-situ metrology without further attributing the benefit to the individual measures. However, prerequisites changed significantly with the 300mm equipment generation, making new analyses necessary.

More recent analyses of the cycle time benefit of batch tool replacements are available, but show indifferent results. In [WWK$^+$04] about 15% of the steps were originally performed on batch tools. In a simulation scenario, all batch tools were replaced with single wafer tools resulting in a cycle time benefit of 40%. Wright and Bass show smaller benefits in [BW08] for the replacement of batch tools with single wafer tools. These accumulate to a cycle time reduction of 16.9% for a fab with few different products and 19.1% for a fab with a larger product portfolio. Even smaller estimated benefits are reported in [BML$^+$03]. The cycle time benefit is shown to be 2% (replacement of diffusion furnaces with mini-batch equipments) or 7% (complete replacement of batch equipments with single-wafer equipments) with the authors' annotation that this improvement seems underestimated. The different results can be partially due to specifics of the underlying fab business model, as, e.g., the product portfolio diversity, and to specifics in the underlying simulation model, as, e.g., different baseline utilization of the equipments to be replaced or a different utilization of the replacement equipments if their capacity differs. But the need for a more detailed assessment by different cycle time components is obvious.

6.2 Theoretical Discussion

With the introduction of the semiconductor manufacturing environment in Chapter 3, we already outlined the motivation for the batch tool replacement. We explore the opportunities for improvement and the challenges in more detail in the following.

6.2.1 Cycle Time Reduction Coherences

Based on the discussion in Section 3.4.3.4, we expect a moderate reduction of process time in case of the replacement with mini-batch tools and a massive reduction in case of the replacement with single-wafer tools. The loading process is significantly shortened and the batching process drops. Therefore, within process time the reduction is bound to have a more significant effect on DL than on PR.

[1] Some wet cleans remained on batch tools based on the equipment development status at that time.

By definition the batch building and dissolving time does not occur for single-wafer tools, therefore it is reduced by 100% for the replacement with single-wafer tools. Mini-batch tools also reduce BT to some extent as they require smaller batches. We do not expect changes in QT as a direct effect of the replacements[2]. However, as a consequence of different availability characteristics or different variability in availability due to a different equipment count in the workstation, QT can change as an indirect effect of the replacements. The transport time TT does not change because the same number of operations is necessary to complete the route. Table 6.1 summarizes the expected cycle time reduction by component.

Table 6.1: Expected cycle time reduction by component per tooltype change

Component	mini-batch scenario	single scenario
PR	decrease	big decrease
DL	decrease	big decrease
BT	decrease	elimination
TT	no change	no change
QT	unclear	unclear

With respect to total cycle time the expected reduction depends on the previous contribution of batch equipments to process time and on the extent of BT, which can vary significantly depending on the company profile.

6.2.2 Challenges

There are no challenges involved in the operation of mini-batch or single-wafer tools. In fact, the replacements rather simplify operations because the batch building process is not necessary or at least simplified. Therefore, there is, e.g., no unavoidable locally concentrated transport demand peak.

The major challenge involved in the replacement of batch equipments is the higher cost of ownership of mini-batch or single-wafer equipment. To some extent this can improve with a wider deployment and for single wafer cluster tools integrating more steps as discussed in Chapter 9 because of the more cost-efficient design.

Additionally some process development work still has to be done for a limited number of applications to enable a transfer of the process to mini-batch or single-wafer equipment.

The discussion of these challenges is not in the scope of this work, but we provide estimates of the possible cycle time benefit which in turn might justify some additional cost involved in purchasing and operating mini-batch and single-wafer tools.

6.2.3 Dispatching Considerations

Dispatching rules can consider tooltype specific operational characteristics, especially for batch tools. There are two common rule criteria used to improve cycle time at batch tools.

1. Incomplete batches waiting for additional lots at batch tools can be taken into account at preceeding operations. E.g., lots which are needed to complete a batch are pulled, i.e., have higher priority at preceeding operations. In this way, cycle time of the full batch is optimized.
2. At lot start at the beginning of the route, lots can be started in batches of lots which share the same batching context. This lot start policy has been shown to optimize cycle time.

[2] Because the change in QT is indirect and of less importance compared to the change in other cycle time components, we do not use queueing theory to analyze the cycle time reduction effect of the replacement of batch tools.

For reasons discussed in Section 5.2, the dispatching rules in our simulation analysis all work according to the FIFO rule, i.e., we do not incorporate tooltype specific dispatching. This might represent a small cycle time disadvantage for the scenarios using batch, and also mini-batch tools. However, the only cycle time component that could be improved by incorporating such dispatching is BT.

Apart from these specific rule variants discussed above, there are no commonly applied disatching rules that are tooltype specific.

6.3 Simulation Analysis

In our simulation analysis we assess the cycle time effectiveness of the batch tool replacements in several scenarios. We constructed a mini-batch tool scenario, a single-wafer tool scenario, a multi-product scenario, a hybrid scenario, and a scenario to correctly determine BT.

6.3.1 Experimental Design

In our mini-batch scenario we replace all batch furnaces with corresponding mini-batch tools. The wet clean batch tools remain in the model, because there are no mini-batch tools for this application and batch sizes are already relatively moderate with 50 wafers[3]. The mini-batch furnaces used in the simulation model all have a batch size of 24 wafers. Their throughput is oriented at existing mini-batch tools and is about half the throughput of the batch tools used in the model. As the original batch size is four times the mini-batch size, this means that PR is cut in half by this replacement at furnace operations.

In our single-wafer tool scenario we replace all batch tools with corresponding single wafer tools. For most process applications of batch tools, corresponding single-wafer tools exist, but are not deployed because of worse cost of ownership. In these cases we use throughput and availability characteristics of existing tools for the replacements. If no corresponding single wafer tools exist to date, then we use the throughput of similar tools and applications. Whenever capacity of the single wafer tools requires tool additions or enables tool count reductions, then we adjust the tool count accordingly.

We additionally create an equipment scenario designed for lot size reductions at only a part of the route. In these scenarios, the equipments performing front-end of line operations, i.e., operations before completion of first contact, remain as in the baseline model. In contrast, the batch tools performing back-end of line operations are replaced by single wafer tools. In the lot size reduction scenarios derived from this equipment scenario at standard lot size, lots of the baseline lot size of 24 wafers are split into smaller lots after the last batch tool of the route has been visited. From there on the pure single wafer toolset enables an effective cycle time reduction by reduced lot size. Because of the partial replacement we call this scenario the *hybrid* scenario.

Several factors motivate these scenarios.

1. The share of batch tools is higher in the front-end of line. Therefore replacement with single-wafer tools is less expensive and more effective in the back-end of line.
2. The limitation of lot size reduction to the share of the route that contains only single-wafer tools makes the reduction in cycle time both effective and less susceptible to productivity issues.
3. Foundry semiconductor manufacturers often offer generic transistor design, but customer specific metal layers designed for a specific application. Therefore short cycle time is more important in the back-end of line where the customer specific metal layers are processed. This enables quick reaction to customer demand, whereas in the front-end some bulk demand can be assumed.

[3] In our model the batch size is 48 wafers because of the different baseline lot size.

In our hybrid simulation model the last batch operation is the 107th operation. Until then a total of 28 batch operations are performed. Afterwards the remaining 270 operations are performed at single wafer tools including 52 operations that would be performed on batch tools in the baseline scenario.

In another modification of the baseline scenario, we assess the impact of a broader product spectrum on BT. In this scenario ten different products with an equal volume are processed. The ten products are started in full lots and round robin sequence and do not share the batching context, i.e., batches formed at batch tools do not contain lots of different products. The remaining processing specification is not changed, i.e., the products do not belong to a different process technology.

Additionally, we developed a scenario that enables experimental determination of BT. For this purpose we replaced all batch tools with single-wafer tools, but left all performance parameters like availability, throughput and delay unchanged. Of course, real single wafer tools would have different characteristics, therefore application of this scenario is limited to the experimental estimation of BT.

6.3.2 Batch Building and Dissolving Time BT

The commonly used way to determine a batching time is by measuring the queue time of lots which have not formed a complete batch yet or by calculating it with a batching time formula as in [HS00]. However, there is no such simple way to assess or calculate the batch dissolving time, especially when the dissolving process can stretch over several operations. Therefore we use an experimental approach utilizing the scenario with the single-wafer tools having the performance characterisitcs of the corresponding batch tools. The queueing time disadvantage of the baseline scenario over this single wafer tool scenario is our batch building and dissolving time BT'.

Figure 6.1 illustrates BT' at different loadings. BT' decreases with increasing fab load. Batching time formulas as in [HS00] suggest a linear decrease, therefore we approximate BT with linear regression. We determine BT from BT' first with interpolation and then, because of the high randomness involved in the simulation runs at high loadings, we omit BT' of the four highest fab loadings and extrapolate BT based on BT' of the fab loadings from 50% to 87.5%. The visual impression does not favor one of the two approaches, based on later analysis in Section 7.4.2 we chose the extrapolated values as BT.

Additionally, we display the share of BT' that actually occurs at batch equipments in Figure 6.1. This represents the part of BT' that actually occurs during batch building while the remaining part of BT' represents the batch dissolving time. This is not a clear-cut distinction because a batch dissolving process can also take place at a batch tool with a lower batch size than the previous batch tool, but this is improbable within our route, therefore we suppose that this is a reliable distinction. At lower loadings around 80% of BT' are batch building time and this share decreases to aroud 70% for higher loadings disregarding disturbances caused by the higher influence of randomness at very high loadings. This decrease can be explained by both the smaller time necessary for batch building and less capacity available for batch dissolving because of the higher load.

Our key takeaways of this section are that

1. it is necessary to include the batch dissolving time because significant 30% of BT in the realistic loading range are considered to be batch dissolving time, and,
2. our experimental approach produces reasonable results for BT.

6.3.3 Replacement with Mini-Batch Equipments

Figure 6.2 shows the cycle time results for several scenarios. First, we compare the *Mini-Batch 24 scenario* with the baseline *Batch 24 scenario*. Without any further change the furnace batch tool replacements reduce the total cycle time by 14%.

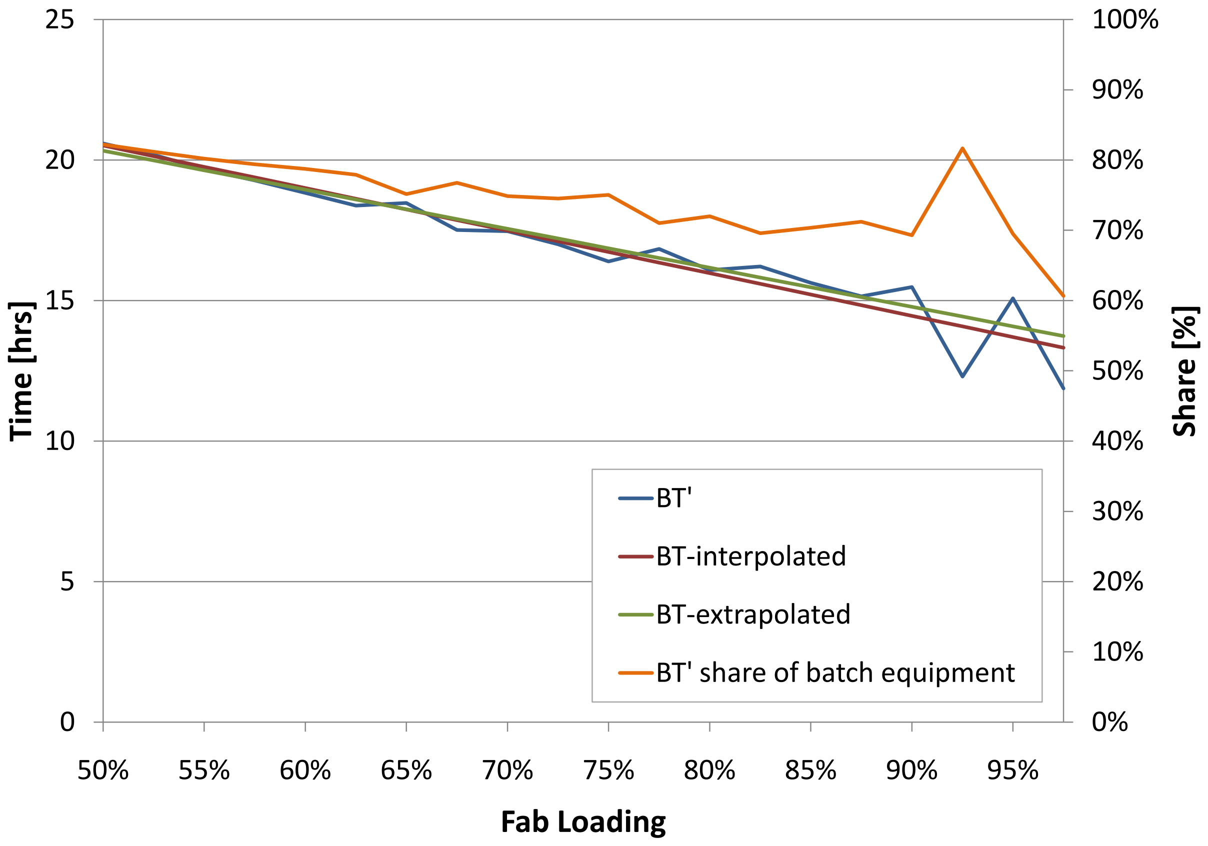


Figure 6.1: Batch building and dissolving time BT' (left axis) for different loadings and share of BT' (right axis) that actually occurs at batch equipments

The relative reduction by component is illustrated in Figure 6.3. BT is reduced very effectively by 60% due to the higher baseline furnace batch size and fewer batching contexts shared between operations at furnace workstations compared to wet clean workstations. QT decreases slightly because of the increased tool count at the mini-batch workstations leading to less variance in the total available capacity. Process time also decreases as derived in Section 6.2.1 with a reduction of 14% in PR and a more significant reduction of 21%in DL caused by shorter loading times at mini-batch operations.

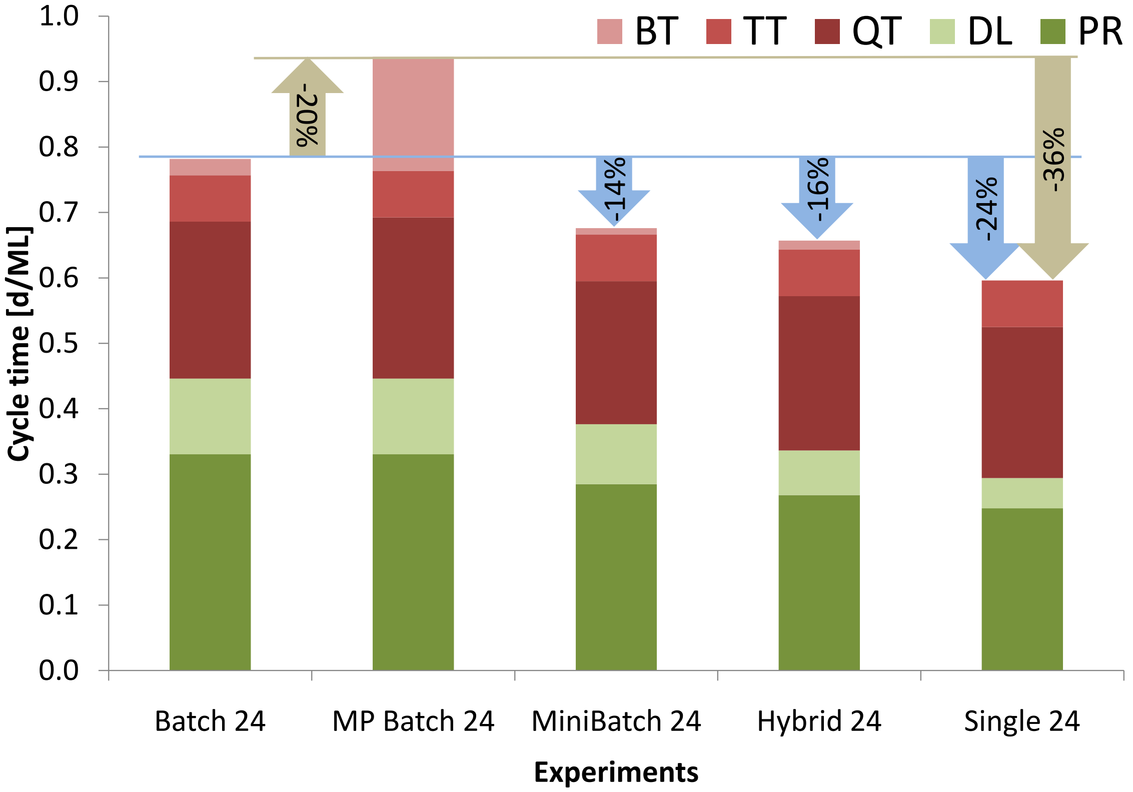


Figure 6.2: Cycle time performance of fab with toolset converted to mini-batch tools or single-wafer tools compared to baseline and multiple products model at 92.5% loading.

6.3.4 Replacement with Single-wafer Equipments

The results of the single-wafer tools scenario are displayed again in Figure 6.2. We compare the *Single 24 scenario* to the baseline *Batch 24 scenario*. The replacement of all batch tools with single wafer tools reduces the total cycle time by 24%.

The relative reduction by component illustrated in Figure 6.3 shows similar relations as in the case of mini-batch tools, but the reductions are bigger in most cases. Batch buidling and dissolving does not occur in single-wafer tool scenarios, hence a reduction in BT by 100% occurs. QT decreases slightly because additional tools are necessary due to lower single-wafer tool capacity and this decreases the variability in availability at these workstations with additional tools. The significant decrease in process time splits into a reduction of 25% in PR and a more significant reduction of 60% in DL because the long loading and unloading times of batch tools do no longer apply.

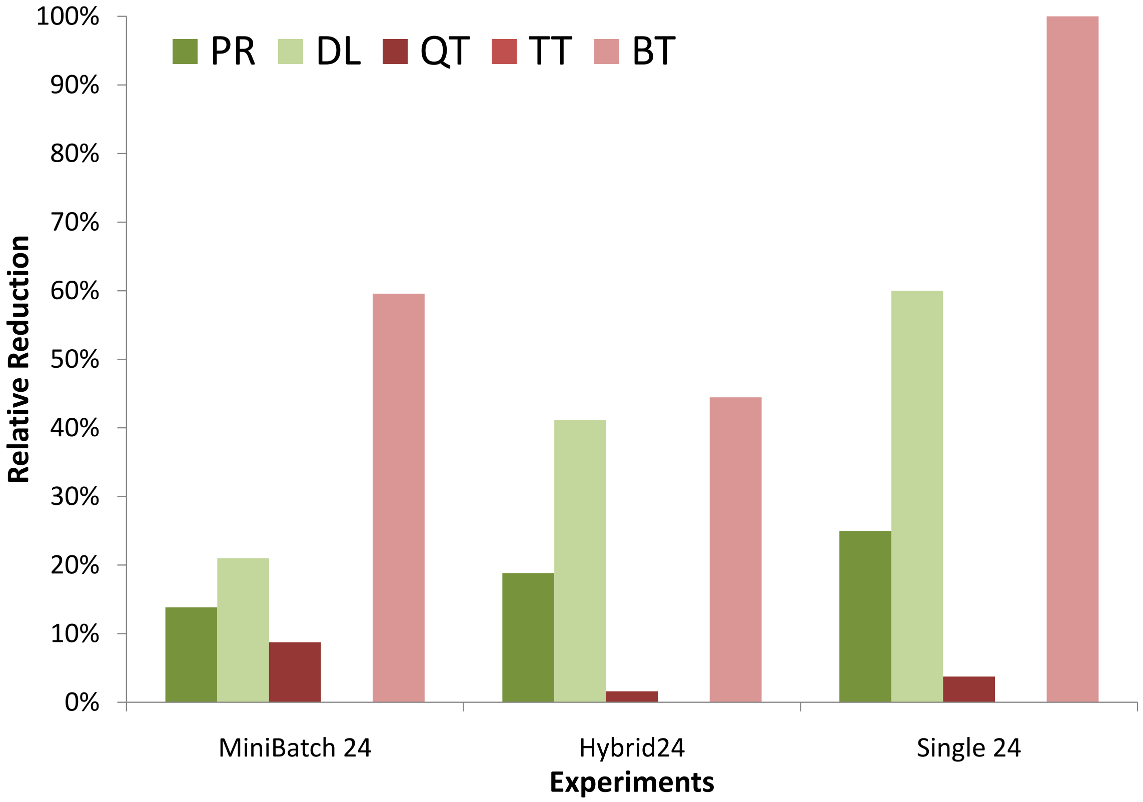


Figure 6.3: Relative reduction per cycle time component for toolset converted to mini-batch or single wafer tools relative to baseline scenario at 92.5% loading.

6.3.5 Hybrid Scenario: Partial Replacement with Single-wafer Equipments

The results of the hybrid tools scenario are displayed in Figure 6.2 as well. We compare the *Hybrid 24 scenario* with the baseline *Batch 24 scenario*. The replacement of all BeoL batch tools with single wafer tools reduces the total cycle time by 16%.

The relative reduction by component is illustrated again in Figure 6.3. Although more than half of the batch operations cease to exist, batch building and dissolving time is reduced by 44% only due to less flexible batching context in the FeoL. This is also less than in the *Mini-Batch 24* scenario. Processing PR and delay DL are reduced by 19% and 41% respectively representing values between the results for the *Mini-Batch 24* and the *Single 24* scenarios. Queueing time QT is reduced only insignificantly and transport time TT does not change.

6.3.6 Product Diversity Considerations

Figure 6.2 also shows the cycle time results of the multiple products scenario. We first compare the multiple products scenario *MP Batch 24* with the baseline scenario *Batch 24*. The results only differ in BT which is significantly higher as in the baseline scenario. It does not increase ten times as does the number of batching contexts in this scenario, however, because batches often do not fall completely apart between batch operations. They are just streamlined.

Next, we compare the multiple products scenario *MP Batch 24* with the single-wafer tools scenario *Single 24*. Because batching does not occur in manufacturing with pure single-wafer toolset, there is no need to create a multiple-products single-wafer tools scenario, the *Single 24* scenario already represents the matching scenario for comparison. The relative reductions by component are the same as in the comparison of the *Single 24* scenario with the baseline scenario *Batch 24*[4], but because of the higher original BT value, the total cycle time reduction reaches 36%.

Of course, with the higher number of batching contexts we consider only one operational aspect of product diversity in semiconductor manufacturing. There are other aspects as, e.g., different routes for different products, however, the number of batching contexts is the defining aspect for possible cycle time reductions achieved by batch tool replacements. Therefore, the number of products with different batching contexts is one major factor to be regarded when such replacements are assessed and an important motivating factor for such a change.

We did not create a separate multiple products scenario for mini-batch tools. The cycle time performance of such a scenario can be easily estimated by a reduction in BT compared to the *MP Batch 24* scenario that is similar to the reduction achieved in the *Mini-Batch 24* scenario compared to the baseline scenario *Batch 24*. All other cycle time components would not differ from the single-product scenario *Mini-Batch 24*.

6.3.7 X-Factor Considerations

Table 6.3.7 lists the x-factors of the scenarios under consideration within this section at a fab loading of 92.5%. Apart from the average x-factor based on the average cycle time, we show the x-factor based on the 95-percentile cycle time as well in order to give an impression of the cycle time variance.

Scenario	X-factor	
	average	95 percentile
Batch 24	1.75	1.92
MP Batch 24	2.10	2.27
Mini-Batch 24	1.80	2.00
Hybrid 24	1.95	2.15
Single 24	2.03	2.27

Table 6.2: X-factor for scenarios at 92.5% fab loading

With respect to the average x-factor, there are measurable differences but no significant change in x-factor between the different scenarios. Therefore x-factor remains a reasonable measure for comparison of fabs with and without batch tools.

With respect to the 95-percentile x-factor, the difference between average x-factor and 95-percentile x-factor is slightly higher in the scenarios with replaced batch equipments. However, this increase is simply due to their smaller raw process time T_0 representing the denominator. The absolute variance in cycle time does not change.

6.4 Conclusions

We summarize the key results of this section with the characterisitic curves for the scenarios *Batch 24*, *MP Batch 24*, *Mini-Batch 24*, *Hybrid 24*, and, *Single 24* displayed in Figure 6.4.

1. Both the mini-batch and the single-wafer tools scenario benefit from the shorter process time intrinsic to their operational characteristics. This can be seen by the different levels of the theoretical limits represented by the sum of raw process time and transport time in the chart.

[4] This is the reason, why we do not include the *MP Batch 24* scenario in the relative component comparison of Figure 6.3

2. BT also causes worse performance of manufacturing with batches. The negative impact is most visible at low loadings where it determines most of the difference of actual performance and the theoretical limits. The size and impact of BT decreases with an increase in loading, but remains significant, especially for the multiple products scenario.
3. The other cycle time components TT and QT play no or no defining role for the difference in cycle time that occurs with different tool type scenarios.

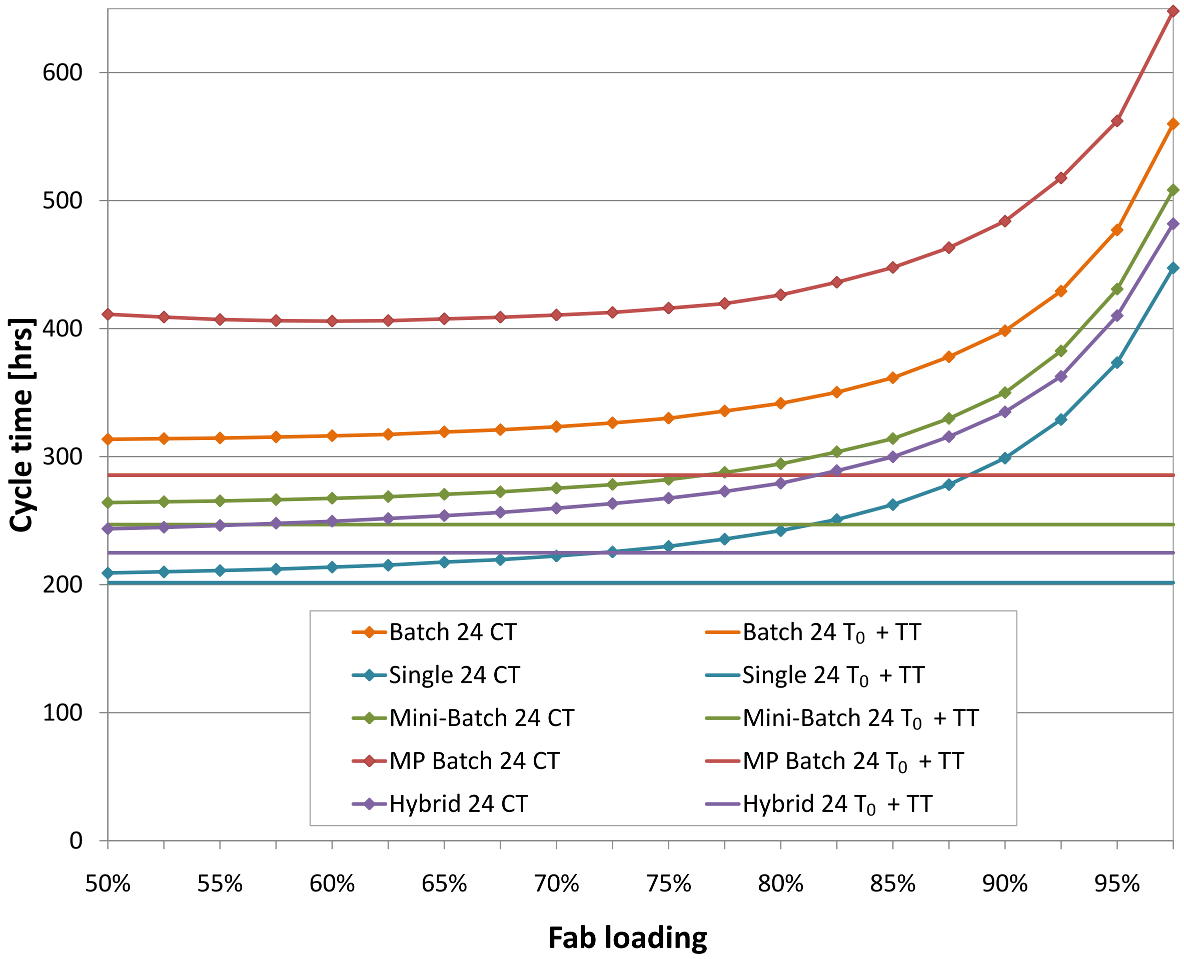


Figure 6.4: Characteristic curve of equipment scenarios at 24 wafer lot size compared to *Batch 24* baseline

The cycle time benefit achieved by the replacement of batch tools is persuasive. Provided that the cost of ownership can be decreased to an acceptable level not too much above that of batch tools, a wide adoption of this approach seems likely in the foreseeable future of semiconductor manufacturing. For very cycle time sensitive manufacturing environments an earlier adoption makes sense. The partial replacement in the BeoL is an interesting intermediate option especially for companies with a business model that is more sensitive to BeoL cycle time.

7 Lot Sizing

In this chapter we assess the cycle time gain that can be achieved by lot size reduction. We published parts of this assessment already in [SR08b] and [SR08a], but extend the analysis significantly in the following.

7.1 Classification and Definition

The computation of optimal lot sizes is a standard topic in operations research. Normally, the lot-sizing problem deals with the basic tradeoff between having many small jobs, which tend to increase setup costs (material, tracking costs, labor, etc) versus having a few large jobs, which tends to increase inventory. However, lot sizing in the semiconductor industry takes place in a different context than in many other industries. Here, lot sizing does not deal with order-specific lot sizes, it deals with the definition of one static standard lot size that is used for the whole factory. Normally this lot size is set to the maximum allowed by the transport carrier and is seldomly studied in detail.

We have already defined the term *lot* in Section 3.2, however, we require a more precise, distinctive definition in this chapter to avoid misunderstandings. The term *lot size* is used ambiguously in the semiconductor industry. It is used as generic term for lot size, transportation size and carrier capacity because in virtually all environments they have the same size. With smaller lot sizes and the possibility to have several several lot sizes but only one carrier capacity simultaneously in a fab this is no longer the case and the distinction between carrier capacity and lot size definitely has to be made. The distinction between lot size and transport size is only necessary for foundry fabs. This dissertation assumes lot size and transport size to be the same und uses the generic term *lot size* for both as this is common throughout the industry. With respect to carrier capacity we do not make a specific assumption, but assume that the chosen carrier capacity is at least equal to the lot size. Considering space requirements it is advisable that carrier capacity does not provide space for more wafers than the lot size, but the availability of carrier capacities might be limited. Therefore some wafer slots in a carrier might be left unused intentionally.

7.2 Literature Review

The cycle time benefit of lot size reductions in semiconductor front-end manufacturing has not been discussed extensively in the literature. In the 1990s there is one notable publication by Wood [Woo97], however still based on 200mm equipments. Depending on the fab loading, the simulation results presented in this publication show a 15-20% reduction in cycle time for a reduction in lot size from 24 to 12 wafers and a toolset including many batch tools. However, prerequisites changed significantly with the 300mm equipment generation, making new analyses necessary.

Some more recent analyses of the cycle time benefit of lot size reductions are available, but do not show comprehensive results. In [WWK$^+$04] the authors report a possible cycle time reduction of 33% for a reduction in lot size from 25 to 13 wafers as a result of their simulation studies. In the underlying simulation model a pure single wafer toolset is assumed, but further details regarding either model or results are not presented.

Another study shows indifferent results [ZWBP08]. The cycle time actually increases from 1.28 D/ML to 1.31 D/ML when the lot size is decreased from 25 to 12 wafers in the baseline model. As additional measures the authors introduce in the following an increase in equipment availability by 5%, a decrease in the length of setup times and first wafer delay (here: DL), a replacement of batch tools with single wafer tools and an optimized cascading strategy. In total these measures lead to 43% reduction in cycle time with 25 wafer lot size and 57% at 12 wafer lot size. The optimized cascading strategy refering to a reduction in the minimum number of wafers necessary to trigger a setup leads to the biggest reduction in cycle time and it remains unclear why this is not applied in the baseline model. However, the size of this reduction is independant of the lot size. Nearly all of the differences in cycle times between the different lot sizes in the results are due to the replacement of batch tools with single wafer tools. At smaller lot sizes the tool replacements lead to a significantly higher reduction in cycle time. We attribute this to the high number of individual batching contexts of 100 products. The increase in equipment availability by 5% and the decrease in the length of setup times and first wafer delay also leads to reductions in cycle time, but they are smaller in size and give only a small advantage compared to the smaller lot size.

All authors fall short of providing details that make the results comprehensible. Consequently the need for a more detailed assessment with respect to different cycle time components and individual influencing factors is obvious.

7.3 Theoretical Discussion

Together with the introduction of the semiconductor manufacturing environment in Chapter 3 we already outlined the motivation for a reduction in lot size. We explore the opportunities for improvement and the challenges more specifically in the following.

7.3.1 Cycle Time Reduction Coherences

As discussed in Section 3.4.3.4, PR decreases proportionally to a reduction in lot size at single wafer tools whereas DL does not change. At batch tools process times do not change at all, but BT increases with lot size reduction provided the same number of wafers are batched. This increase is due to the higher difference between the batch size and the lot size. TT is not directly affected by a reduction in lot size at both single wafer and batch tools, as the number of transports per lot does not change. All of these changes are straightforward, however, the change in QT does not follow such simple principles. Therefore we explore the expected change in QT with queueing theory in the following section. As anticipated result, we expect a slight decrease in QT for smaller lot sizes.

Single wafer tools performing metrology operations form an exception to the previous analysis of the expected cycle time reduction. Refering to our lot size dependent sampling model presented in Section 5.1, which assumes constant total measurement times for all scenarios, lot skip rate at metrology operations increases significantly. Therefore much less lots actually visit the metrology operation and all cycle time components become smaller at single wafer tools performing metrology operations.

We summarize the expected changes per cycle time component, tool type, and operation type in Table 7.1.

Table 7.1: Expected cycle time reduction by component, tool type and operation type

Scenario	Batch tools	Single wafer tools (process)	Single wafer tools (metrology)
PR	no change	proportional decrease	decrease
DL	no change	no change	decrease
BT	increase	n.a.	n.a.
TT	no change	no change	decrease
QT	slight decrease	slight decrease	decrease

7.3.2 Analysis with Queueing Theory

In order to establish a reference for the reduction in QT and the influencing mechanisms, we analyze lot size reductions with queueing theory. This analysis is based on our previous discussion in [SR07b].

The queueing time at a workstation with m tools can be calculated according to Equation 4.2. As first step we identify the parameters that depend on lot size. We generally think of utilization as a function of the number of wafers to be processed that is independent of the lot size[1]. This is obviously true for the number of parallel tools m as well, leaving two parameters defining the queue time changes for lot size reductions, the effective process time t_e and its coefficient of variation c_e.

Considering Equations 4.3 and 4.4, we further decompose the dependence. Of the factors in these equations, we assumed PR to be deterministic. Therefore c_0 equals zero. We further minimize the number of variables by assuming repair times to be exponentially distributed which means that c_r equals one. This leaves the three variables PR, m_r and A defining the queueing time change for smaller lot sizes.

We illustrate this change in an example. In Figure 7.1, we show the relative queuing time for lots of half lot size compared to original full lot size for varying mean time to repair m_r and availability A. The process time PR is a constant 0.5 hrs at full lot size and half that value for the smaller lot size.

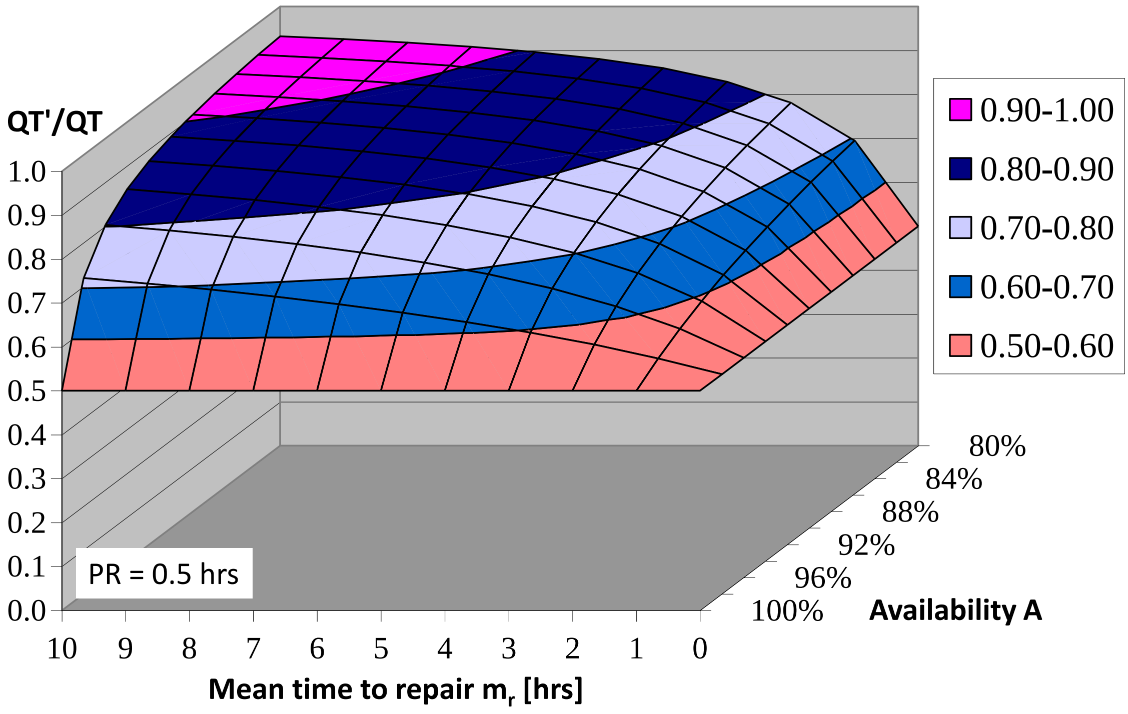


Figure 7.1: Relative queueing time for halving the lot size depending on mean time to repair m_r and availability A

Figure 7.1 illustrates several findings:

- For half the lot size the queueing time is between half the queueing time of full lots and the full lot size value.
- At 100% availability the highest relative queueing time reduction is achieved.

[1] Although we do not analyze the impact of setups with queueing theory, we state for completeness that the assumption of utilization being independent of lot size does not necessarily hold for tools with significant setup frequency.

- Variability degrades the queueing time reduction. Moderate availability reductions already lead to significantly smaller queueing time reductions than maximally possible.
- There is always some reduction in queueing time although the reduction approaches zero for very high variability.

In Figure 7.2, we vary the full lot processing time PR and the availability A for constant mean times to repair m_r of 4 hrs.

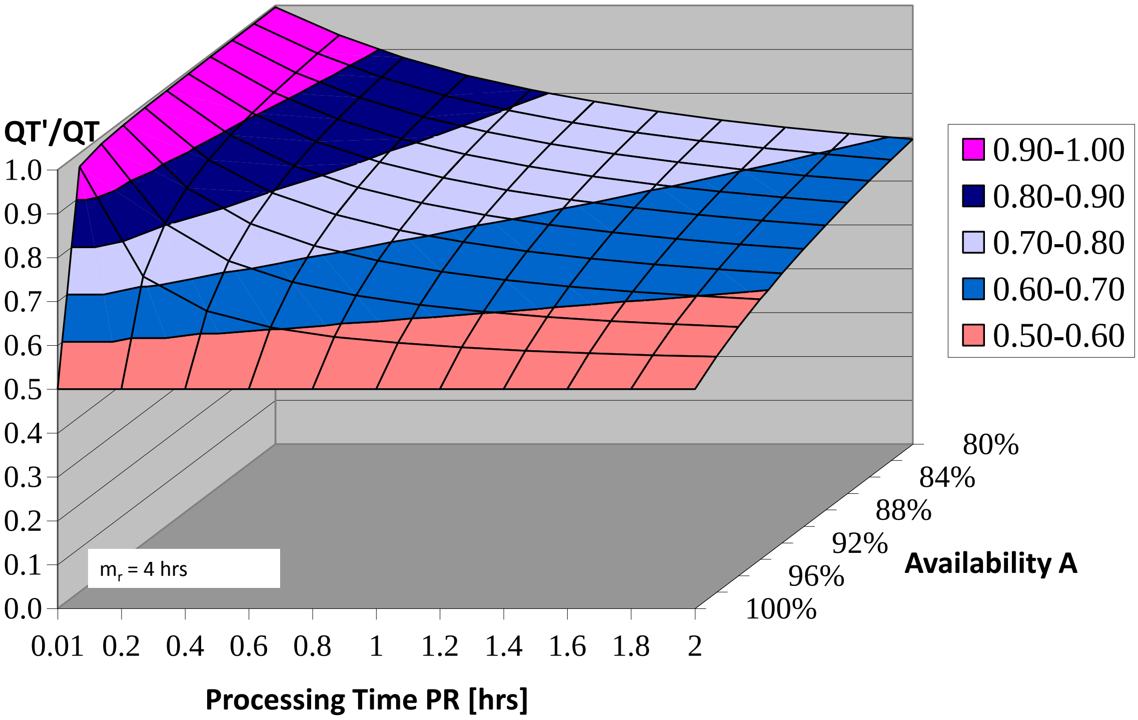


Figure 7.2: Relative queueing time for halving the lot size depending on processing time PR and availability A

The diagram further illustrates the influence of PR on the queueing time reduction. Short process times degrade the queueing time reduction. With respect to the denominator of Equation 4.4, the shape of the PR-curve is concave enabling a wider range of significant queueing time reductions. Yet, the key insight from Figure 7.2 remains that the queueing time reduction is higher at tools running at smaller rates.

With this section we have identified dependent factors and the shape for queueing time reduction achieved by reduced lot size. The result is only valid for the assumptions made for the applicability of queueing theory, but can serve as a benchmark to check the usefullness of simulation results and by quantifying the impact of the variables shaping the queueing time improvement it helps to identify starting points for increasing the cycle time effectiveness of lot size reduction.

7.3.3 Assessment of Challenges

Challenges of lot size reduction can be grouped into equipment productivity and material handling system challenges. However, the distinction is not always clear-cut, as long material delivery times degrade productivity and short delivery times can be part of the solution. We base our assessment on our previous discussion of challenges associated with lot size reduction in [SR07a].

7.3.3.1 Equipment productivity challenges

Scenarios with batch tools represent the first equipment productivity challenge. Across the board batch tools are configured to process batches with a number of wafers that is a multiple of standard lot size. There are three possible ways to load batch tools in a smaller lot size environment:

1. Batch the same number of lots
2. Batch as many multiples of lots as possible without splitting lots
3. Batch the same number of wafers (includes splitting lots into separate batches)

Most toolsets are only capable to perform option (1) because of various design limitations. This in turn corrupts cost of ownership, turning the cost of ownership disadvantage of single wafer tools into an advantage for lot sizes significantly smaller than current standard lot size. Therefore this option is a very unlike choice for operating at smaller lot sizes.

Option (2) is a viable option. Depending on the specific lot size some capacity is lost, but on an acceptable level. However this requires a significant tool redesign of most batch equipments. There is also an alternative solution using a backdoor: Performing lot combine and lot separate operations before and after the batch process sources out a part of the physical batching to a simple tool dedicated to this task. This is an acceptable solution to integrate a limited number of batch tools. However, using it on a large scale creates other issues. The load on the material handling system by this operation increases by large and the additional lot separate and combine operations consumes cycle time and cost capital and engineering resources.

Option (3) is mentioned for completeness, but logistically undesired.

Therefore the most likely approaches to the issue of batch tools are either the redesign of batch tools, so that they can perform option (2) or their replacement with mini-batch or single-wafer tools which do not have these issues. The outsourcing of the batching process might also be used for a limited number of operations.

The second productivity challenge relates to the number of load ports. For continuous processing at full rate several wafers have to be in the equipment simultaneously[2]. Therefore the *in access* status of carriers following each other has to overlap for a specific period of time. With PR and the number of load ports, the overlapping time which is equal to DL defines the window of opportunity for the carrier exchange at a load port, defined as required carrier exchange time. Figure 7.3 illustrates the interaction of the three defining contributors in a Gantt chart.

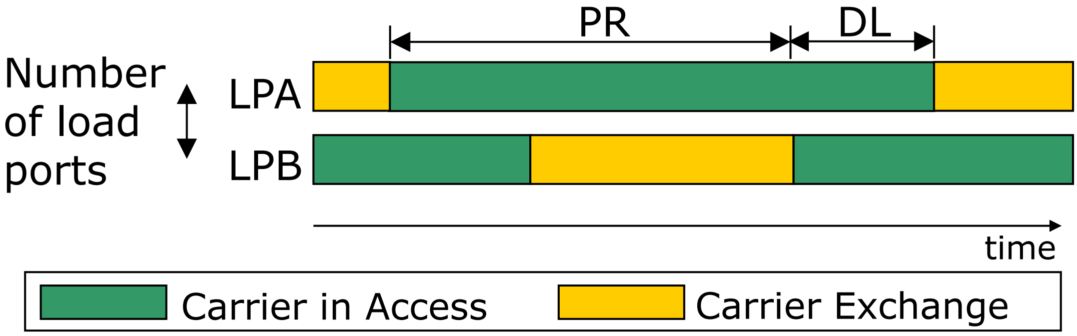


Figure 7.3: Illustration of the factors influencing the required carrier exchange time

[2] In complex lithography equipments more than 50 wafers can be required

Regrettably, only PR decreases with the lot size, while DL does not. Therefore the window of opportunity becomes narrower and eventually negative for smaller lot size. For conventional operations an unavoidable productivity loss would be the consequence. More load ports can be a solution for some toolsets but restrictions with respect to the width of the tool front-end limit the application of this solution approach.

The obvious starting point for the solution of this problem is to shorten the actual time for the carrier exchange. Many attempts have been made to shorten AMHS delivery times, e.g., by the application of under-track-storage [Glü06], [RPQ05]. Another possibility is to eliminate the AMHS dependence and to use an on-equipment buffer directly above the load ports with a dedicated carrier robot [KJ06]. Yet, these approaches cannot solve the instances of requiring negative carrier exchange times. Most likely the removal of empty carriers, whose wafers are already all inside the tool, may solve this problem. This requires new operational scenarios and automation protocols [KEPS06], [RGHT07]. Providing further details on both the issues and solution approaches, we discussed the issue of carrier exchange times extensively in [ZRS$^+$07].

The third productivity challenge relates to recipe dependent setup times. This means that with each change of the process recipe at the tool some setup time is due. This is Most notably at implant tools, where every recipe change requires a tuning of the ion beam that lasts several minutes. Each process step and most technologies require different recipes, therefore this happens very frequently as tools cannot be dedicated to specific recipes. The resulting tool efficiency loss is around 8-25% depending on the product variety of the fab according to [Liu05]. With smaller lot sizes, it can be expected that the chain of wafers of the same recipe is smaller than with the current standard lot size because there is less WIP available. Therefore the efficiency loss increases. Fortunately, the problem does only occur in this magnitude at implant tools otherwise it would be a show-stopper. Obviously, the issue requires action on the tool supplier side, but increased application of scheduling approaches to increase wafer chains of the same recipe might also be part of the solution. We analyze the issue of setup times extensively with simulation in Section 7.4.7.

7.3.3.2 Material Handling System challenges

Smaller lotsizes require more AMHS moves because the same number of wafer starts is distributed among more lots. Figure 7.4 illustrates how AMHS move rates increase while the number of operations per lot and all other conditions do not change. This is not completely accurate as the number of operations performed decreases with lot size because less lot sampling at metrology operations is necessary to adequately monitor the process. However, this reduces the increase only by a small amount.

Todays state-of-the-art unified AMHS systems are limited to around 5000 moves/hour as illustrated by the horizontal red line in Figure 7.4[3] [PP05]. Depending on the size of the factory this system reaches its limitations very soon when smaller lot sizes are deployed. The answer for moderate lot size reductions to 12 or 6 wafers might be to create separate systems again, as they were used in the beginning of the 300 mm era. This solution should involve a more sophisticated connection philosophy based on the experiences with the current system. For extremely small lot sizes it seems that only conveyor based systems are able to deliver the required performance.

7.3.4 Dispatching Discussion

There is no reason why dispatching should have a direct impact on the cycle time effectiveness of lot size reduction. However there are some possibilities of an indirect relation which we discuss in the following.

[3] Non-product lot moves consume a share of AMHS capacity, too - this is neglected in this comparison.

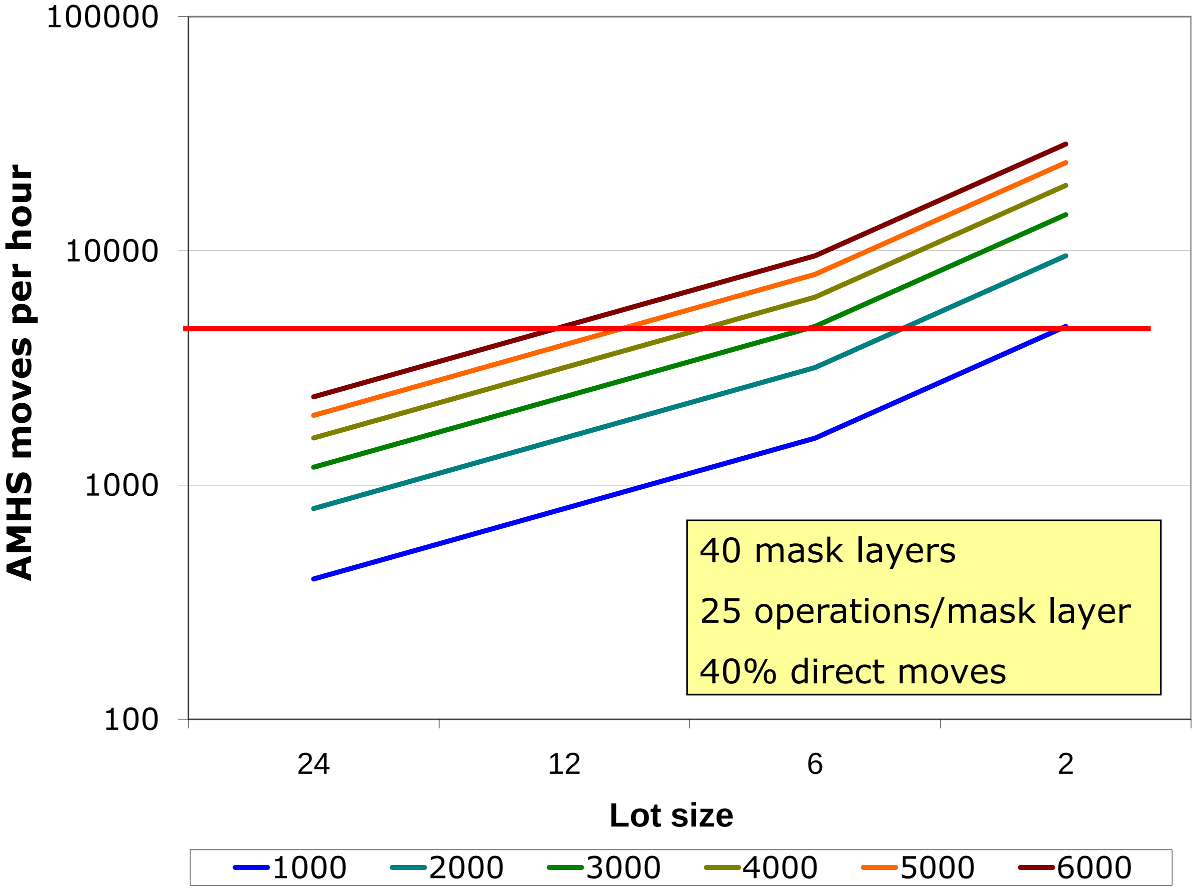


Figure 7.4: AMHS move rate dependent on lot size and wafer starts

1. In Section 7.4.7 we assess the impact of setup times on the cycle time effectiveness of lot size reduction. Dispatching rules at workstations with setups generally should have a setup avoidance criterion [GWS07]. Depending on the specific criterion in the dispatch rules, there can be some relation between the cycle time reduction and the dispatch rule or the specific setup avoidance criterion respectively.
2. In our analysis, we see that variance deteriorates the cycle time effectiveness of reduced lot sizes. Therefore, dispatching rules trying to limit the negative effects of variability, e.g., as discussed in [HF07], might lead to higher cycle time reductions achieved by smaller lot size.
3. Small lot sizes can be a challenge for yield analysis. Currently the *container* lot limits the possible variants for the combination of chambers visited by all wafers of a lot. This is helpful to identify resources responsible for yield issues. With smaller lot size the *container* limits the variants for a smaller number of wafers per lot. Therefore dispatching rules might need adaptations to support yield analysis by providing a specific set of chamber combinations at distinct operations.

For reasons discussed in Section 5.2, the dispatching rules in our simulation analysis all work according to the FIFO rule with a strict setup avoidance rule at workstations that encounter setups.

7.4 Simulation Analysis

In our simulation analysis we assess the cycle time effectiveness of lot size reduction in several scenarios. We use the batch tool scenarios with different product diversity, the mini-batch tool scenario, the hybrid scenario and the single-wafer tool scenario of Chapter 6 as basis and create extra scenarios at reduced lot sizes of 12, 6, and, 2 wafers for each of these equipment scenarios. Additionally, we created scenarios to study the influence of setup times on the cycle time performance.

7.4.1 Experimental Design

The experimental design of the different equipment scenarios was already described. In our lot size reduction scenarios, lot size is the only model input that we change. Refering to our equipment productivity discussion above, batch tools are able to batch the same number of wafers in all scenarios enabled by the standard lot size of 24 wafers chosen in Section 5.1. Therefore no change in loading occurs for the batch tools in our scenario at smaller lot sizes. In our hybid simulation model the lot size remains at the standard 24 wafers until the last batch operation which is the 107th operation. Afterwards lots are split into lots of the reduced lot size 12, 6, or, 2 respectively.

At first glance, it looks weird that we do not additionally change our static transport delay. It is unrealistic to expect that current material handling systems are able to support reduced lot size with the same performance because lot size reduction goes hand in hand with transport volume increase. However, looking at this issue from another side, the need for fast on time delivery increases with lot size reduction and how should that be accomplished if transport times increase? Therefore we have chosen to assume that transport time stays the same for all lot size scenarios keeping in mind that the material handling system might need a significant redesign to achieve this. Additionally, we assume that material handling challenges like reduced carrier exchange times are solved, e.g., as discussed in Section 7.3.3.1.

7.4.2 Toolset including Batch Tools - Single Product Case

Figure 7.5 shows the cycle time results per component for the baseline scenario *Batch 24* and the derived lot size reduction scenarios *Batch 12, Batch 6, and, Batch 2* with *Batch* characterizing the toolset including batch tools and the numbers refering to the lot size of the scenario. Smaller lot sizes lead to a decrease in total cycle time by up to 26%, but it is difficult to see how the individual cycle time components contribute to this result. Therefore we show the relative change of all components in Figure 7.6.

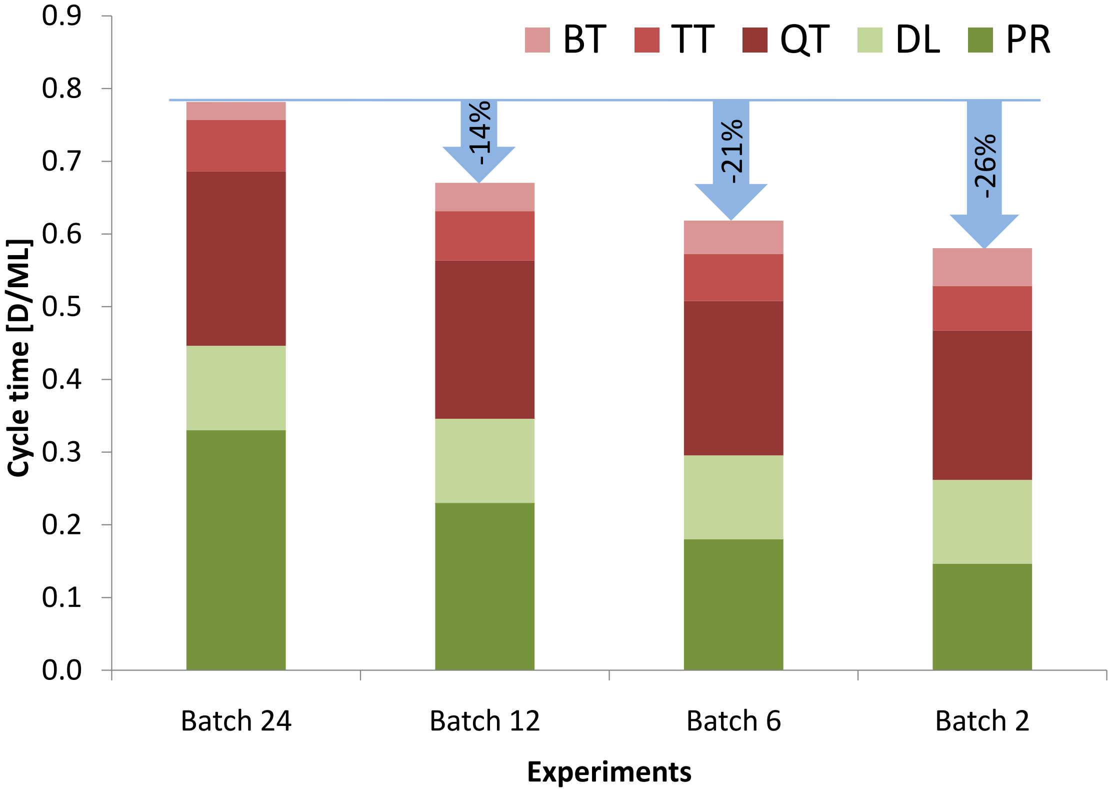


Figure 7.5: Lot size reduction scenarios for toolset including batch tools at 92.5% loading.

- Processing time (PR) is reduced most effectively because of the proportional relation to lot size for single wafer tools.
- Delay (DL) does only change because of the increased lot skip rate at metrology operations. Due to the small DL that lots encounter at metrology tools this results in hardly any change at all.
- Remaining queueing time (QT) decreases only slightly.
- Transport time (TT) decreases only because of the increased lot skip rate at metrology operations resulting in a slight total decrease.
- Batch buidling and dissolving time (BT) increases considerably with lot size reduction. However the total amount is not very significant, partially due to the one product setup.

In the following, we want to highlight three interesting characteristics of the above experiments: The metrology cycle time, the batch building and dissolving time, and the cycle time share accrued to batch tools.

In Section 7.3.1, we discussed that the higher lot skip rate at metrology operations leads to a significant metrology cycle time decrease at smaller lot sizes. Figure 7.7 illustrates this reduction by showing the contribution of metrology tools to total process time and total queue time. The significantly reduced contribution for smaller lot size shows that cycle time of metrology operations is reduced very effectively outperforming the cycle time reduction at process operations.

In Section 6.3.2 we presented our approach to derive the batch building and dissolving time BT. We use this approach to derive values of BT for the reduced lot sizes as well. Figure 7.8 shows BT' and BT for the different lot size scenarios. Whereas for 24 wafer lot size neither the interpolated nor the extrapolated calculation of BT seemed clearly favorable, for the other lot sizes the extrapolated calculation is clearly

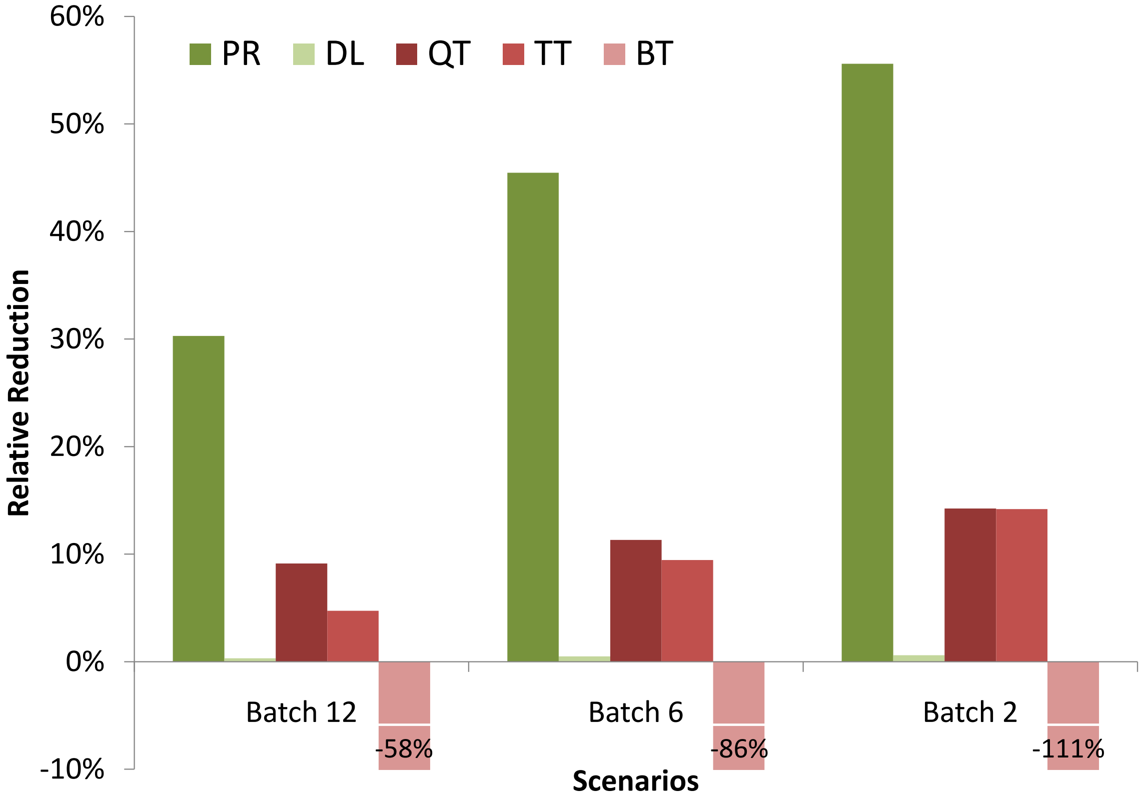


Figure 7.6: Reduction per cycle time component relative to baseline scenario at 92.5% loading.

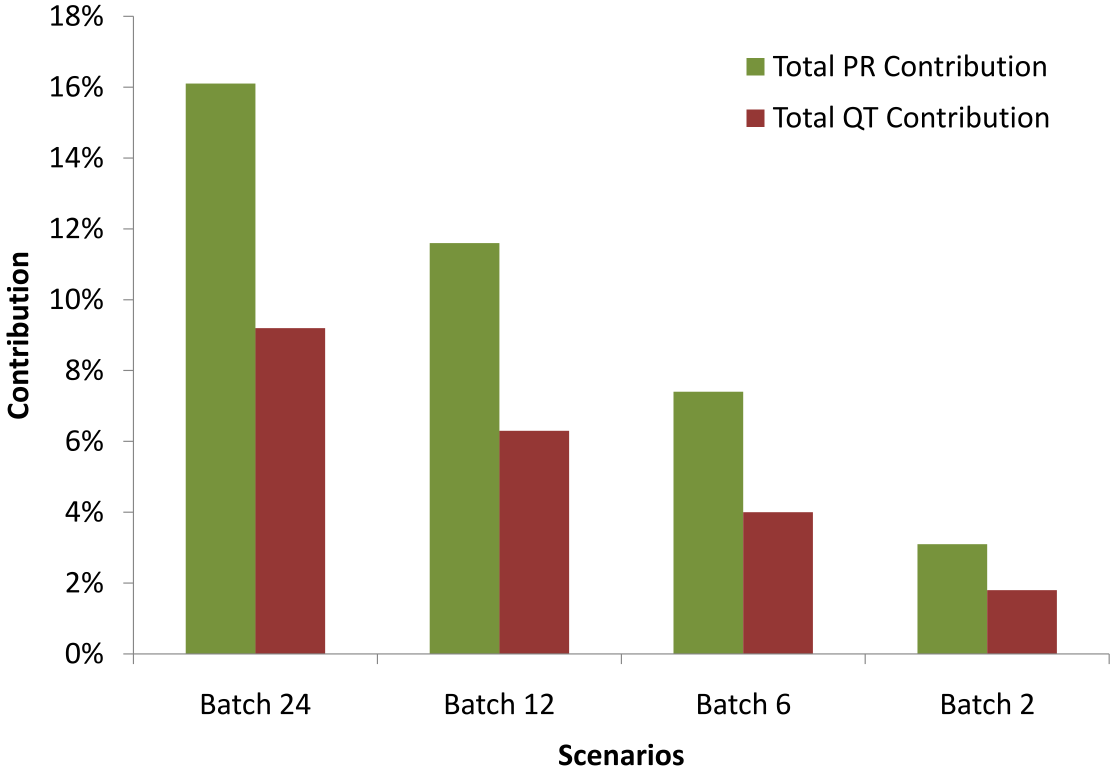


Figure 7.7: Contribution of metrology operations to total PR and QT at different lot size and 92.5% loading.

preferable omitting BT' of the four highest fab loadings. This is obvious from the chart where the BT'-values at high loadings significantly diverge from the direction of the extrapolated BT. BT decreases with increased fab loading and, as discussed, increases with lot size reduction. The slightly indifferent relative direction of the lines referencing BT at different lot sizes shows that the approach does not produce optimal results. However, it is still sufficient for our purpose of illustrating the mechanisms of cycle time reduction by component.

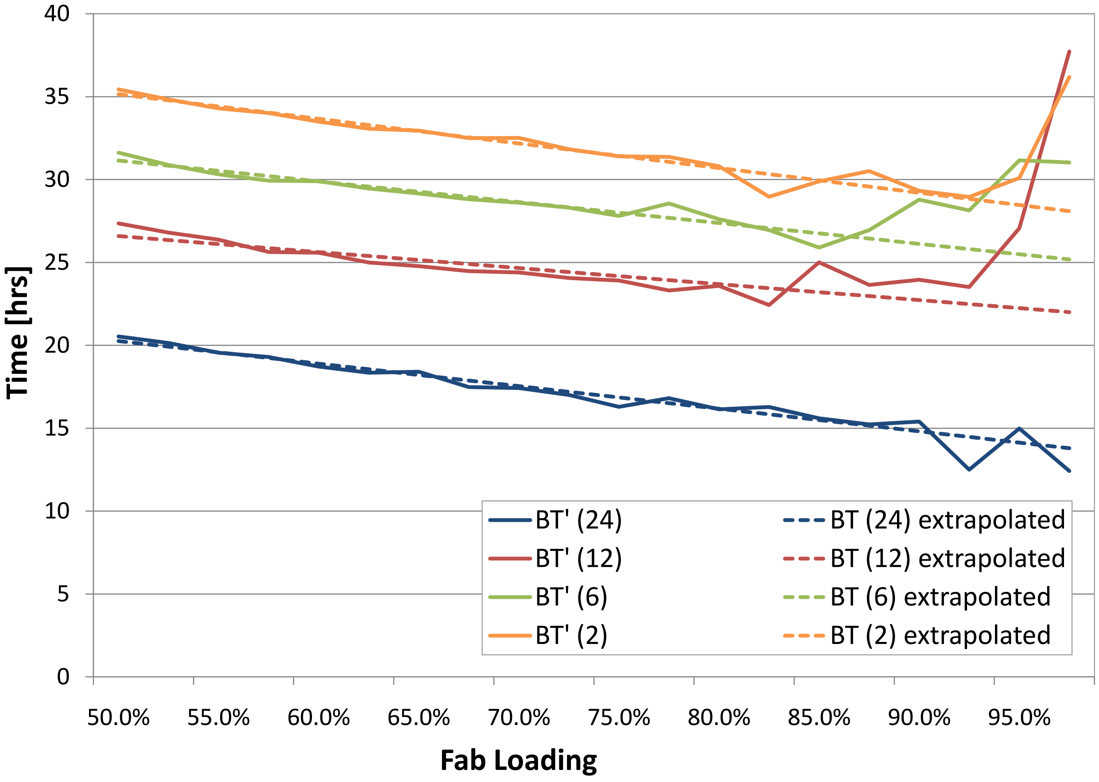


Figure 7.8: BT for different lot size at different loadings.

Figure 7.9 illustrates the share of cycle time that occurs at operations performed by batch tools for the different lot size scenarios. This share of total cycle time increases with lot size reductions confirming batch tools as show-stopper for a more effective cycle time reduction by smaller lot sizes.

7.4.3 Toolset including Batch Tools - Multiple Product Case

Figure 7.10 shows the cycle time results of the multiple products scenario at different lot sizes compared to the baseline scenario at a lot size of 24 wafers. The only difference to the single product case discussed in the previous Section 7.4.2 lies in the size of BT. Because of the higher base value of BT at 24 wafer lot size and the increase of BT at reduced lot size, lot size reduction is about half as effective in terms of cycle time reduction for the multiple products scenario as in the single-product scenario.

7.4.4 Batch Tools replaced with Mini-Batch Tools

Figure 7.11 shows the cycle time results of the mini-batch scenario at reduced lot sizes. In the mini-batch scenario the change to smaller lot sizes is slightly more efficient than in the batch scenario because batch tools account for less cycle time at 24 wafer lot size. Additionally, the mini-batch tools are less susceptible to negative cycle time impacts by different batch contexts because of the lower batch size.

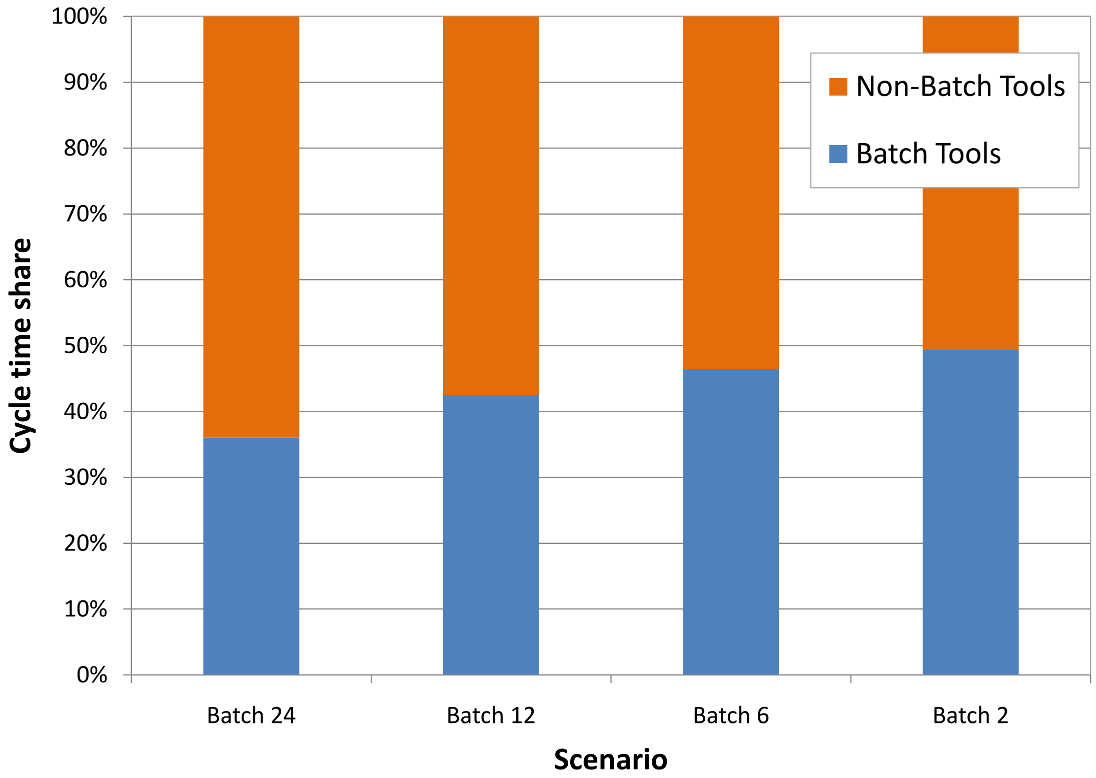


Figure 7.9: Share of cycle time occuring at batch tools for *Batch*-scenarios with different lot sizes at 92.5% fab loading.

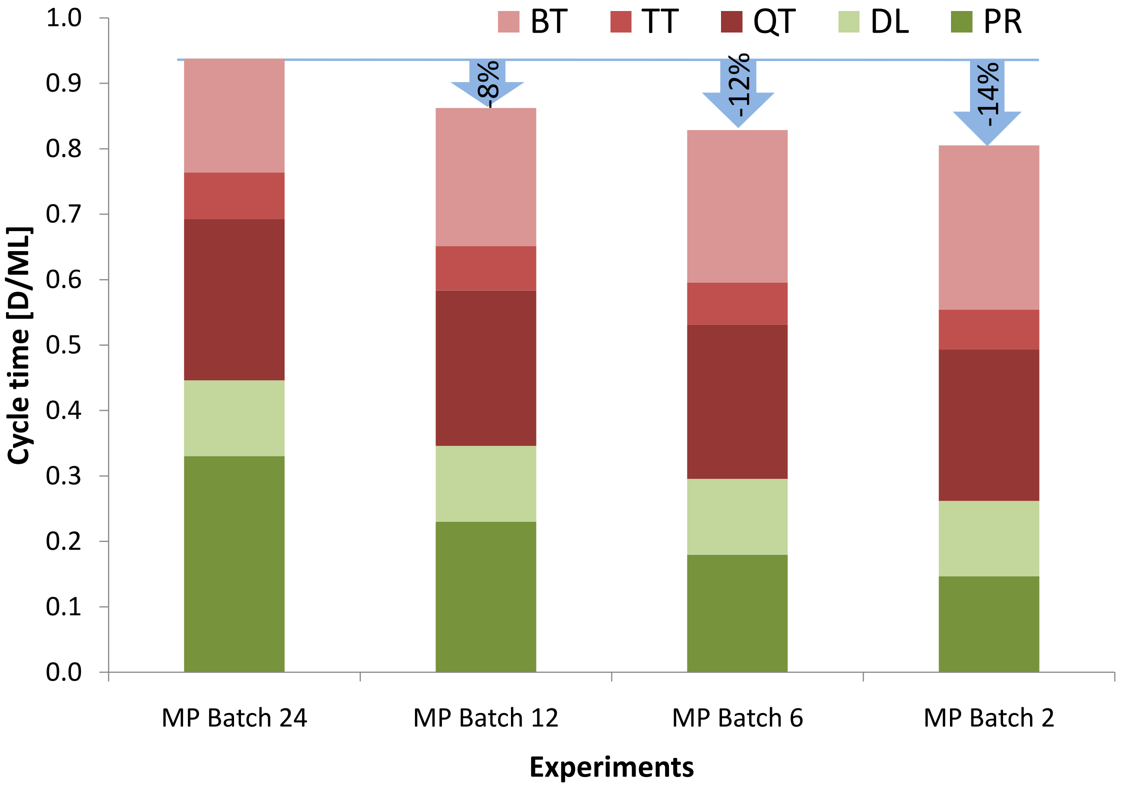


Figure 7.10: Lot size reduction scenarios for toolset including batch tools and multiple products compared to baseline model at 92.5% loading.

However, the cycle time reduction is far off the values realized with a pure single-wafer toolset in the following.

As in the assessment of tooltype replacements in Chapter 6, we do not perform lot size reduction scenarios with multiple products, as the improvement is again easy to estimate. The cycle time performance of such scenarios can be estimated by a reduction in BT compared to the MP Batch scenario that is similar to the reduction achieved in the Mini-Batch scenario compared to the single-product Batch scenario all at the same lot size. All other cycle time components will not differ from the single-product mini-batch scenario.

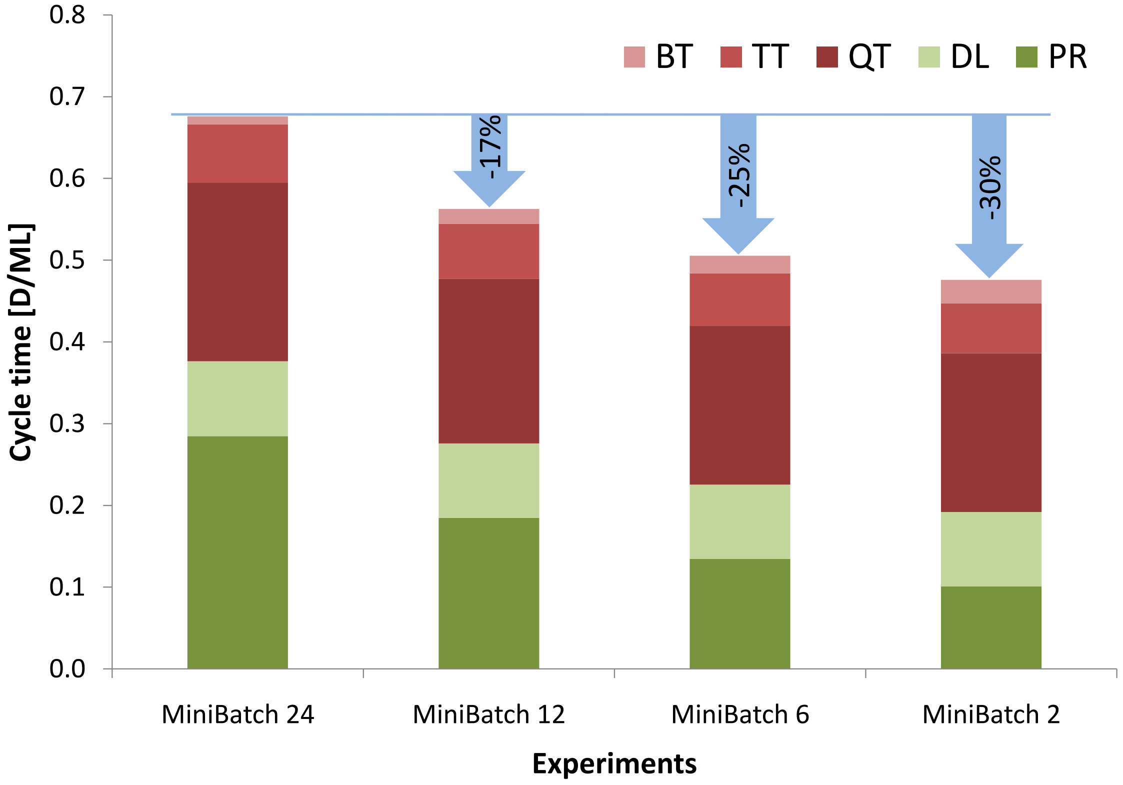


Figure 7.11: Lot size reduction scenarios for toolset converted to mini-batch tools at 92.5% loading.

7.4.5 Batch Tools replaced with Single-Wafer Tools

Figure 7.12 shows the summarized cycle time results for the single-wafer toolset scenarios. In these scenarios lot size reduction is far more effective compared to the batch scenarios because improvements are effective across all workstations and there is no opposite effect from BT.

Figure 7.13 illustrates additional details by showing the improvement per cycle time component.

- Processing time (PR) is reduced very effectively because of the proportional relation to lot size for single wafer tools.
- Delay (DL) hardly changes at all.
- Remaining queueing time (QT) decreases with lot size reduction, however, the decrease is not very significant for reductions below 12 wafers.
- Transport time (TT) changes exactly as for the batch scenario.

As the reduction in QT is not as straightforward as for the other components and is, in addition, loading dependent, we take a closer look at the relative queueing time reduction in Figure 7.14. In this figure, we

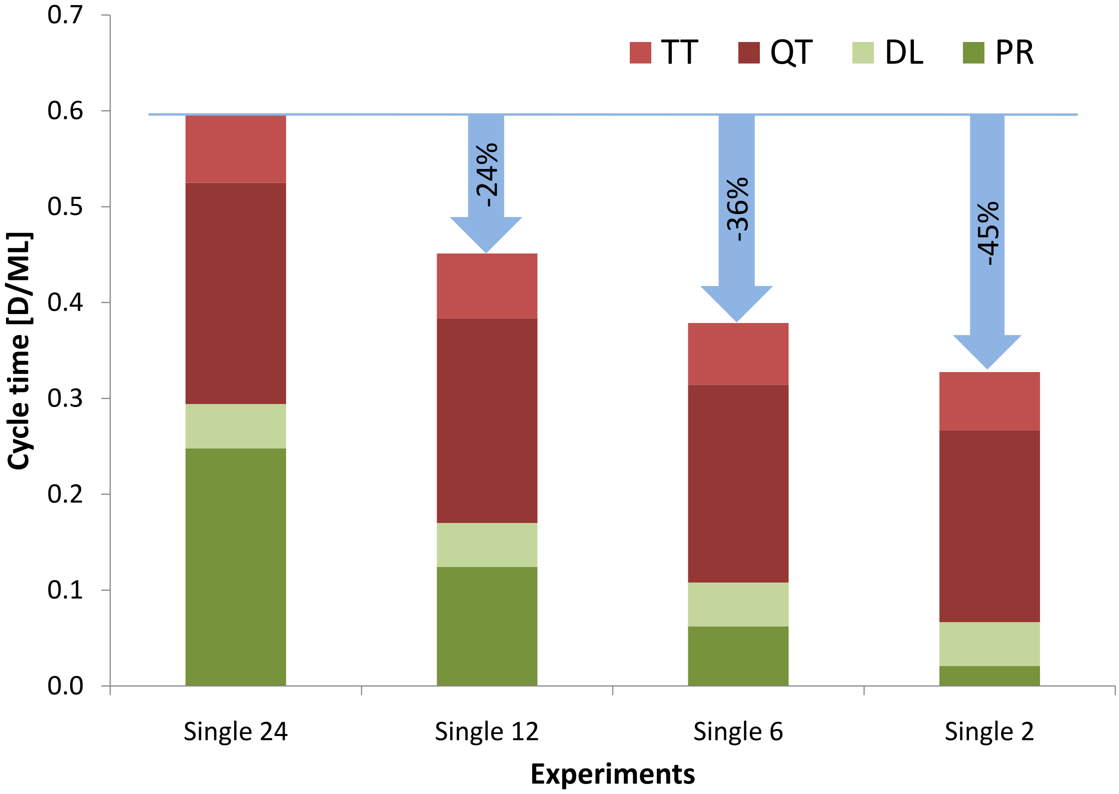


Figure 7.12: Lot size reduction scenarios for toolset converted to single wafer tools at 92.5% loading.

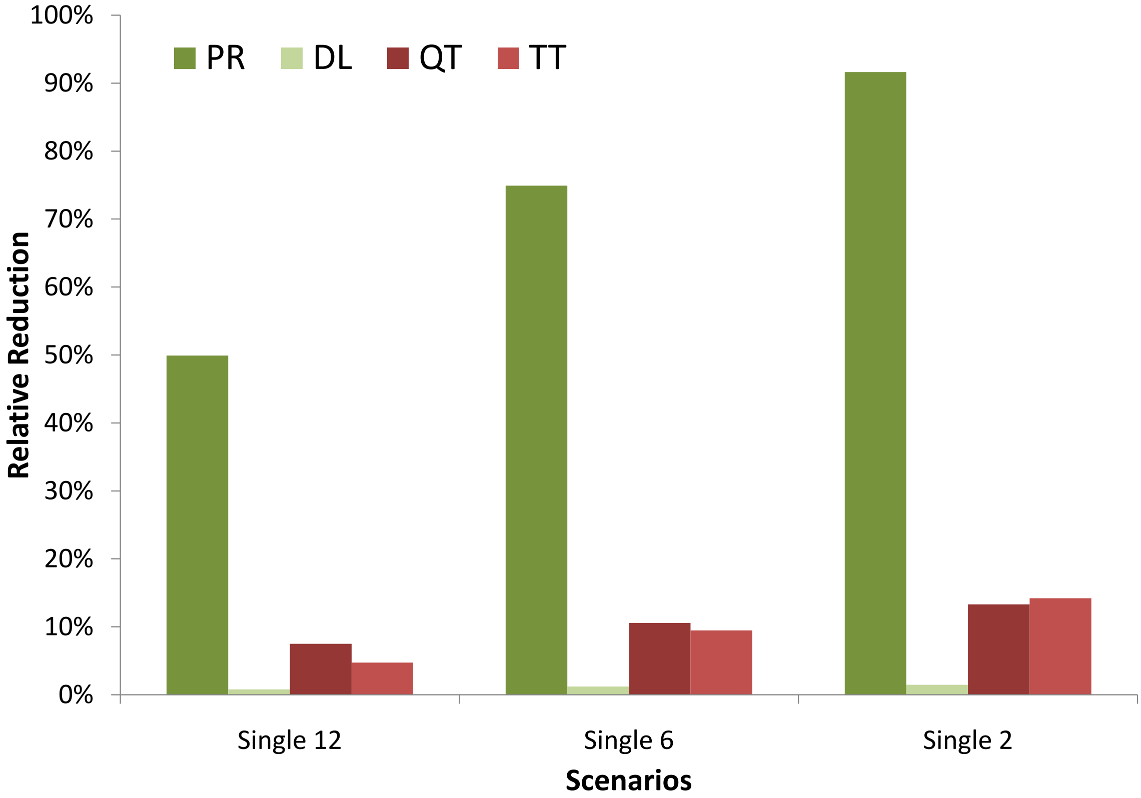


Figure 7.13: Reduction per cycle time component for toolset converted to single wafer tools relative to baseline scenario at 92.5% loading.

show the effect of lot size reduction by a comparison of *Single 12* and *Single 6* scenarios to the *Single 24* scenario.

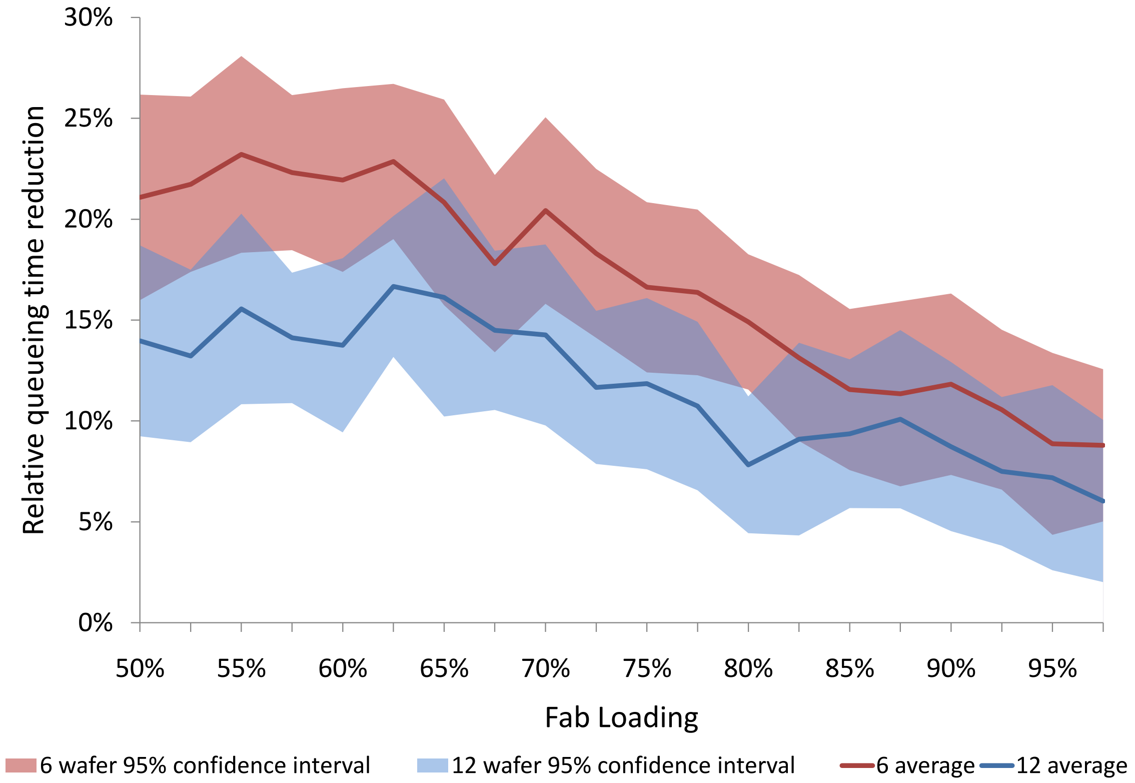


Figure 7.14: Relative queueing time reduction of 12 and 6 wafer lot size compared to 24 wafer lot size for pure single wafer toolset at different fab loadings.

We see that the relative queueing time reduction is much higher at lower utilizations. Under these conditions the negative impact of variability is less corrupting. It is also obvious that the reduction from 12 to 6 is less effective than from 24 to 12. This makes further reductions look unpromising. The 95% confidence interval reaches significant levels here, because variability impacts only this cycle time component and the graph shows a subtraction.

Figure 7.15 illustrates another interesting effect. So far, we showed the cycle time improvement by lot size reduction separately for different toolset scenarios and the effect of tool type replacement for 24 wafer lot size. In this case we show the cycle time improvement of tool type replacements starting at different lot sizes. The more the lot size is already reduced the more effective is a replacement of batch tools. This also means that fabs running already at reduced lot size - for short cycle time or other reasons - replacements with mini-batch or single-wafer tools are even more promising.

7.4.6 Hybrid Scenario

Figure 7.16 shows the summarized cycle time results for the hybrid scenarios with batch tools replaced by single-wafer tools in the BeoL only. In these scenarios lot size reduction is about as effective as in the mini-batch scenarios. Cycle time reduction is achieved only in the BeoL according to the same principles as in the single-wafer tools scenario. Because the cycle time reduction is achieved with only partial a reduction of the lot size, this can represent an interesting option, because batch tool productivity challenges can be avoided and transport demand is increased at smaller rates.

Figure 7.17 illustrates the cycle time improvement in the BeoL by showing the cycle time share accrued

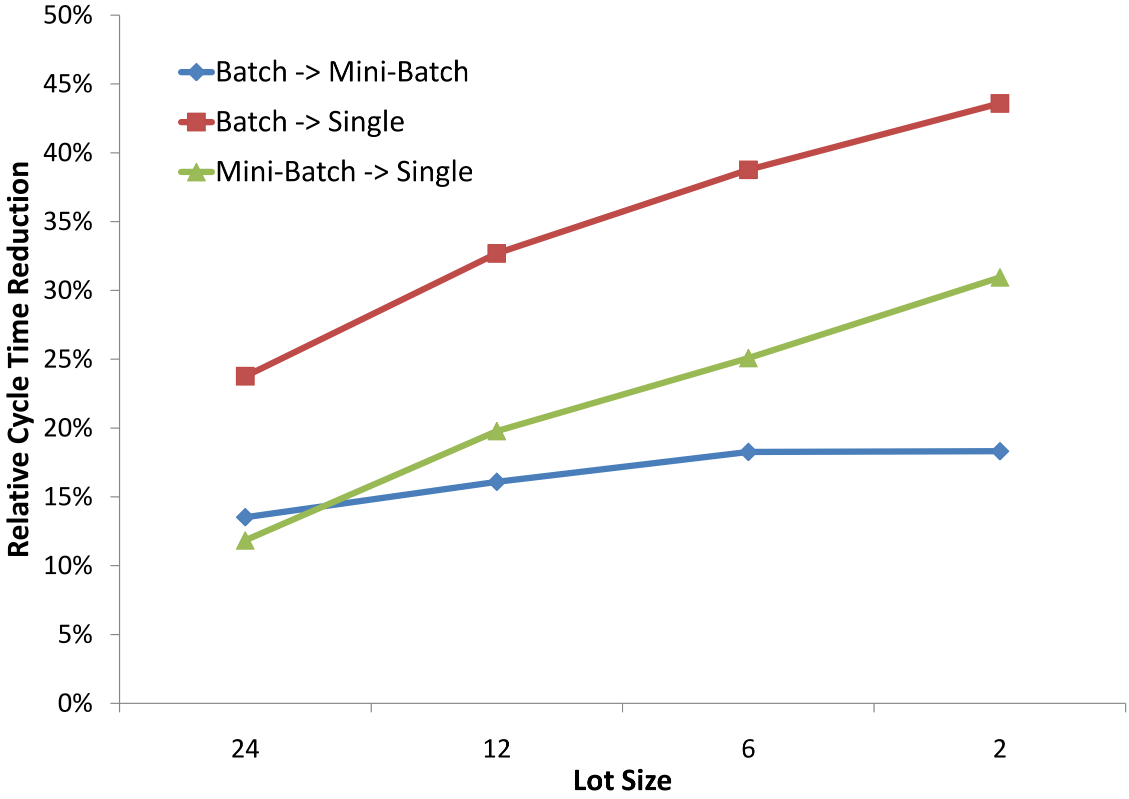


Figure 7.15: Cycle time improvement for toolset conversion at different lot sizes and 92.5% loading.

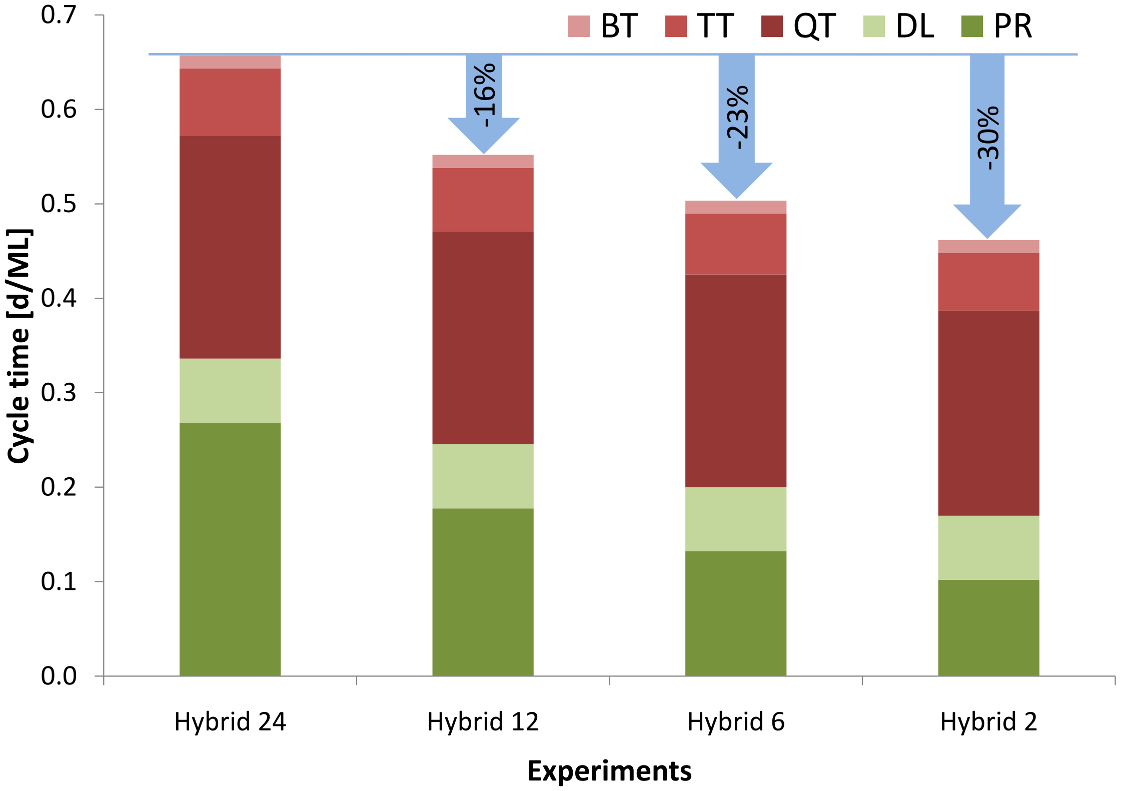


Figure 7.16: Lot size reduction scenarios for toolset converted to single wafer tools in the BeoL with 24 wafer lot size in the FeoL and altered lot size in the BeoL compared to baseline model at 92.5% loading.

to the FeoL and the BeoL. It can be seen that already at 24 wafer lot size cycle time in the FeoL is comparably high for the hybrid scenario. Due to the cycle time reduction at reduced lot size occuring in the BeoL only, the cycle time share accrued to the FeoL further increases with reduced lot size.

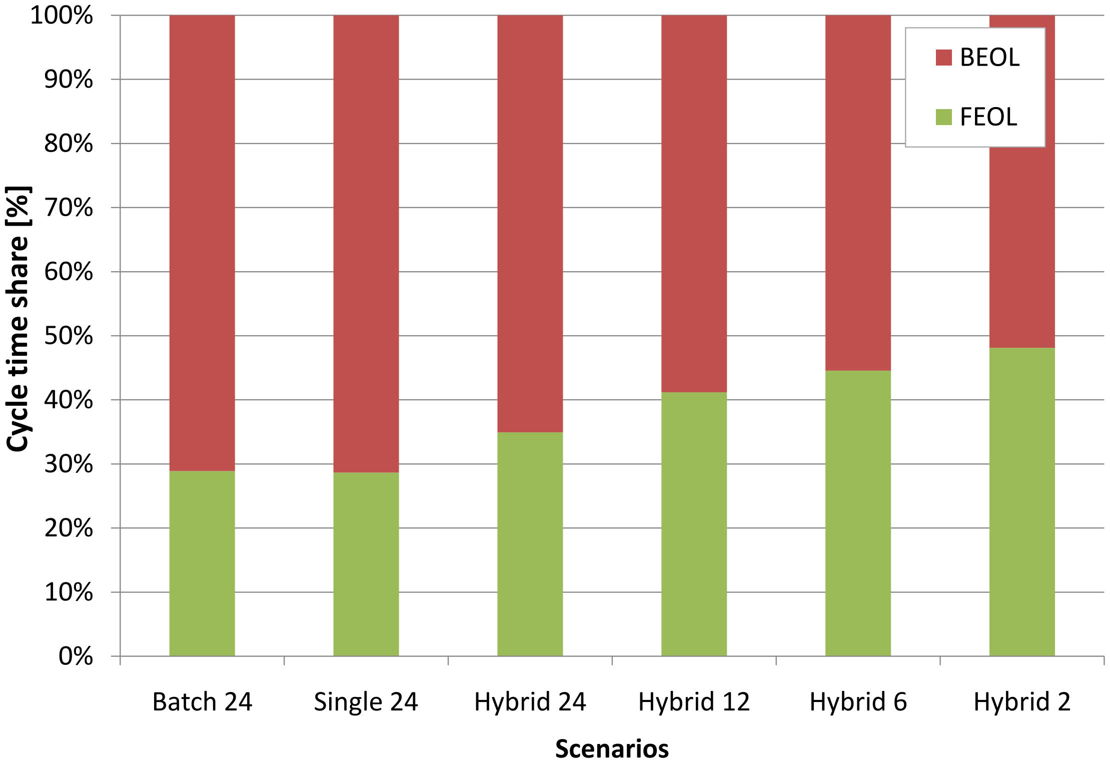


Figure 7.17: Cycle time share accrued to FeoL and BeoL of Hybrid scenarios compared to *Batch 24* and *Single 24* scenarios at 92.5% loading.

7.4.7 Setup Considerations

Setup times occur at some tools after a defined number of wafers have been processed or when process changeovers are necessary. Setups of the first type are usually clean processes that ensure a defined processing environment. Because of the defined interval they can be included in the processing times, and no major lot size dependent effects are expected. However, setups of the second type can have an impact on the queue time decrease effectiveness of lot size reduction. Provided the number of different process operations exceeds the number of tools at the workstation, then each queue time decrease has to go hand in hand with a reduction in the wafer cascade length, i.e., the number of wafers of the same recipe run back to back. This in turn means that setups happen more frequent, which decreases workstation capacity and leads to an increase in queue time. An example for these setups is beamtuning at implant operations that occurs with each process changeover. In our experience setups of this type other than beamtuning can be avoided by dedication provided a sufficient number of tools is available and this is the approach we have taken in our baseline model, i.e., process changeover setups only occur at implant tools.

We embed our setup discussion into the pure single wafer toolset scenarios. In this way, we have no difficulty in distinguishing between setup and batching effects, as both setup avoidance rules and batching per se lead to an agglomeration of lots at the same processing state. At all workstations with setups we use dispatching rules that avoid setups. In the following subsection we assess the setup impact at implant operations and in the next subsection we analyze the effect of additional setups when dedication is not used to minimize setups.

7.4.7.1 Setups in baseline model

Tables 7.2 and 7.3 summarize key figures detailing the effects of lot size reduction at both implant workstations. The low energy implant workstation of Table 7.2 has four tools serving seven operations and the high energy implant workstation of Table 7.3 has two tools serving three operations.

Table 7.2: Setup effects on queue time at low energy implant workstation at 92.5% fab loading

Lot size	Loading	Total QT [hrs]	Setup state	Cascading
24	81%	4.94	6.6%	81.8
12	81%	4.87	9.6%	56.2
6	81%	4.80	12.4%	43.5
2	81%	5.07	14.4%	37.5

Table 7.3: Setup effects on queue time at high energy implant workstation at 92.5% fab loading

Lot size	Loading	Total QT [hrs]	Setup state	Cascading
24	50%	0.79	6.9%	67.0
12	50%	0.72	11.1%	41.7
6	50%	0.75	16.4%	28.2
2	50%	0.80	23.3%	19.8

We want to highlight the following observations:

- The share of setup state increases and the wafer cascading length decreases significantly.
- Lower loading of high energy workstation leads to a higher increase in the share of setup state and to a sharper decrease in the wafer cascading length.
- The reduction in QT is far less than on average (see Figure 7.13) provided QT is reduced at all
- For very low lot sizes QT increases.

7.4.7.2 Scenario with additional setups

In another scenario, we introduce additional setups by dissolving dedications. We combine three back-end etch workstations and two back-end CVD workstations to one workstation each. Because of the different process types performed within the workstation a 24 minute setup time is now necessary whenever the process type group is changed. Again, the dispatch rules at these workstations use an avoid setups policy and additionally a minimum number of tools per setup context is specified.

This setup scenario is quite different to the above example, because the number of setup contexts is significantly lower than the number of tools (three setup contexts at 22 etch tools and two setup contexts at seven CVD tools). Therefore there is no definite need that the setup frequency has to increase to enable queue time reduction.

Tables 7.4 and 7.5 display the total queueing time, the share of setup state per total time, and the number of wafers that are processed between setups. In general the results confirm the observations made above at the implant workstations. However, the negative impact of setups is lower, e.g., the share of the setup state increases less with lot size reduction.

Table 7.4: Setup effects on queue time at etch workstation at 92.5% fab loading

Lot size	Loading	Total QT [hrs]	Setup state	Wafers between setups
24	86%	3.30	1.7%	554
12	86%	3.03	2.0%	471
6	86%	2.99	3.3%	282
2	86%	3.04	4.6%	205

Table 7.5: Setup effects on queue time at CVD workstation at 92.5% fab loading

Lot size	Loading	Total QT [hrs]	Setup state	Wafers between setups
24	79%	3.97	4.3%	639
12	79%	3.77	6.0%	485
6	79%	3.68	8.6%	320
2	79%	3.75	10.2%	269

All examples show that setup times have a negative effect on the queueing time reduction aspired with lot size reduction. Limited occurence of setup times does not jeopardize the concept, however, as queueing time at affected workstations does not become considerably worse.

7.4.8 X-Factor Considerations and Cycle Time Variance Discussion

Table 7.6 lists the x-factors of the major scenarios under consideration within this section at a fab loading of 92.5%. Apart from the average x-factor based on the average cycle time, we show the x-factor based on the 95-percentile cycle time as well in order to give an impression of the cycle time variance.

Table 7.6: X-factor for scenarios at 92.5% fab loading

	X-factor							
	average				95-percentile			
Scenario / Lot size	24	12	6	2	24	12	6	2
Batch	1.75	1.94	2.09	2.22	1.92	2.14	2.31	2.43
MP Batch	2.10	2.49	2.80	3.07	2.27	2.72	3.04	3.38
Mini-Batch	1.80	2.04	2.24	2.48	2.00	2.31	2.51	2.83
Hybrid	1.95	2.25	2.52	2.72	2.15	2.54	2.86	3.12
Single	2.03	2.65	3.51	4.93	2.27	3.07	4.09	5.89

With respect to the average x-factor, there are significant differences in x-factor between the different scenarios, especially for the lot size reduction in scenarios with batch tool replacements. Therefore cycle time performance comparison of fabs with different lot sizes based on x-factor is not advisable.

With respect to the 95-percentile x-factor, the difference between average x-factor and 95-percentile x-factor increases slightly with the reduction in lot size. However, this increase is simply due to the smaller raw process time representing the denominator. The absolute variance in cycle time does not change.

7.5 Conclusions

In the batch scenarios, most of the cycle time gained is due to a decrease in PR at single wafer tools which is partially offset by an increase in BT. Especially with multiple batching contexts this leads to an unconvincing cycle time performance benefit considering the challenges involved with handling reduced lot sizes. Therefore lot size reductions with such a high share of batch tools seems unrealistic, perhaps with the exception of moderate lot size reductions and flexible batching contexts.

The cycle time reduction effectiveness of lot size reductions is improved by replacing batch tools. While it remains unclear whether the improved cycle time performance of twelve wafer lot size with mini-batch tools is already persuasive, this reduction in lot size definitely looks promising for a pure single-wafer toolset. Six wafer lot size might be an interesting option in a fab with reduced variability characteristics, but further lot size reduction still lacks convincing effectiveness regarding the challenges involved.

Figure 7.18 highlights the issue of high variability leading to a less effective reduction in QT. It shows the characterisitic curves for the scenarios *Single 24*, *Single 12*, *Single 6*, and, *Single 2*. While the theoretical limit is reduced significantly by each reduction in lot size, the relative shape of the CT-curves changes only slightly. The negative impact of variability prevents an effective reduction of QT. Therefore we will assess possibilities to reduce the negative impact of variability in the following chapters.

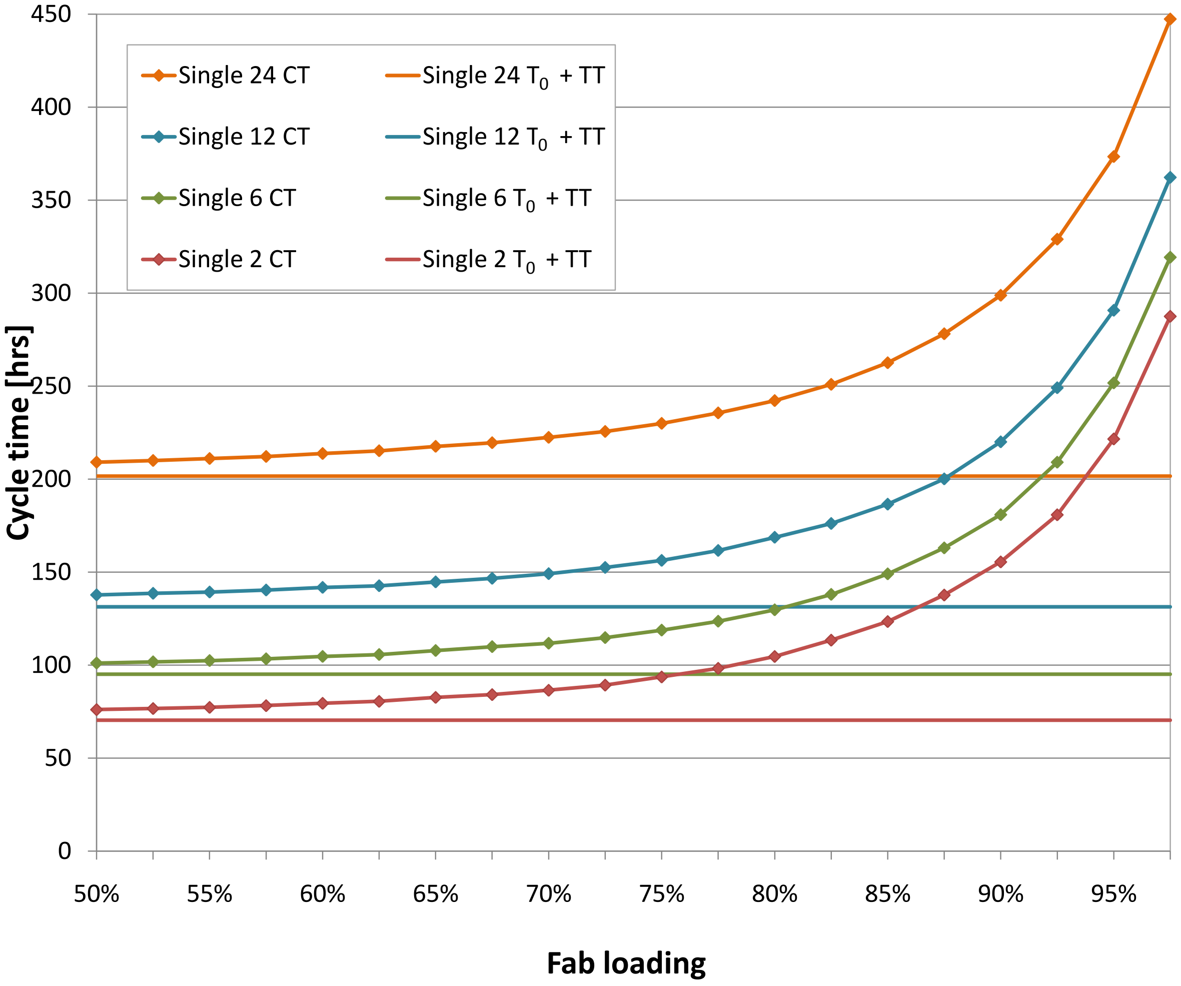


Figure 7.18: Characteristic curves of single wafer tool scenarios at 24, 12, 6, and, 2 wafer lot size

The hybrid scenarios can represent an interesting option under specific circumstances. Advantageous is the limitation of lot size reduction to the route part where a pure single wafer toolset enables a more effective cycle time reduction and is less susceptible to productivity challenges. Circumstances favoring this approach are

- a business model to which short cycle time in the BeoL is specifically important,
- a basis toolset consisting of few batch tools in the BeoL, which makes batch tool replacement in the BeoL less expensive, or,
- BeoL batch tool processes that can be easier transfered to single wafer tools than in the FeoL.

Other configurations of hybrid scenarios are thinkable as well. E.g., it could be interesting to run 12 wafer lot size in the FeoL without replacement of batch tools and utilize a further reduced lot size in the BeoL with a pure single wafer toolset. Alternatively some very limited number of batch tools could remain, e.g., for process or cost reasons, forming another hybrid scenario. There are numerous combination possibilities which might make sense for specific production conditions in a specific company. For those different combinations the results presented in this section can represent starting and reference points for individual analysis.

8 Scaling

In the previous chapter, we ascertained that the negative impact of variability prevents a more effecive reduction of the remaining queueing time QT. The negative impact of variability can often be targeted with scaling. In this case scaling means a higher number of equipments which can provide better buffering for random downtime of equipments. Two possibilities exist for achieving a higher number of equipments. These are

1. the division of current equipments into smaller units forming individual equipments, and,
2. the construction of bigger fabs with higher capacity which require a higher number of unchanged equipments.

8.1 Equipment Scaling

We first discuss the division of current equipments into smaller tools. This is a first analysis of the approach of smaller tools. Therefore we do not examine every possible detail of such an approach. We specifically do not analyze how these tools would look like exactly. We are merely interested in assessing the possible cycle time benefit of such tools which could motivate further work in developing cost-efficient small tool designs.

8.1.1 Theoretical Discussion

The motivation for this analysis is clear. We target a significant reduction in QT, at the baseline lot size of 24 wafers, but even more, we target an effective reduction of QT that goes hand in hand with lot size reductions. In this subsection, we assess the factors leading to a change in the other cycle time contributors and the challenges involved in the approach under analysis.

8.1.1.1 Cycle Time Reduction Coherences

By dissolving cluster tools we loose some of their advantages. In case of dissolving cluster tools with capacity integration through parallel chambers we reduce the processing rate which leads to an increase in PR. And in case of dissolving cluster tools with sequential step integration, we introduce additional tool-internal handling and transports between tools because chambers serving subsequent steps are not situated next to each other anymore, but in different equipments. This leads to an increase in both TT and DL. Hence, QT is the only cycle time component that benefits in this scenario. Some of the other components increase in length, depending on the type of the cluster tool dissolving approach.

Queueing theory confirms the expectation of a more effective reduction in QT. At the baseline lot size, the higher number of entities leads to shorter queueing time according to Equation 4.2. And because of the longer processing times PR, QT is reduced more effectively with lot size reductions as discussed in Section 7.3.2.

8.1.1.2 Challenges

The operational challenges of a fab with these smaller tools with lot size reduction are mostly the same as for the normal single wafer toolset. However, there are two exceptions. First, because of the lower processing rate the smaller equipments have less demanding CET-requirements. As we discussed solutions for this possible issue in Section 7.3.3.1, this does not make much of a difference, though. And secondly, provided cluster tools with sequential step integration are dissolved, then the additional transports place a higher burden on the material handling system. Considering the already challenging increase in transport tasks associated with lot size reduction, this is a major issue.

The biggest challenge of this approach, however, probably lies in a cost-effective realization of smaller tools. In cluster tools several chambers share the handling system of the tool comprising load ports, one or more robots, and, possibly, mainframe and load locks. The cost of these material handling parts of the cluster tools occur only once for several chambers. Smaller tools would lead to an increase in the total number of necessary units of these material handling parts. It is unclear, how this can be achieved in a cost-efficient way.

8.1.2 Simulation Analysis

The basis for the simulation assessment of the cycle time performance of smaller tools is the single wafer tools scenario. A pure single wafer tool environment is necessary for an extensive application of the approach.

8.1.2.1 Experimental Design

Because of the challenges associated with additional transports, we limit the dissolving of cluster tools with sequential step to one application. We dissolve all litho clusters into three tools, one *coating* tool, the *stepper*, and one *developing* tool performing the three major parts of the masking process (These three steps include a number of steps themselves, but we restrain from dissolving those). In case of dissolving cluster tools with capacity integration, we were more daring and dissolved most cluster tools having this characteristic into smaller tools. This process lead to an increase in the process tool count from 212 to 544 tools.

Apart from this change in the toolset with new higher processing time and the additional steps, the experimental design of the single wafer tools scenarios remain unchanged. This includes the four lot size scenarios with 24, 12, 6, and, 2 wafer lot size.

8.1.2.2 Cycle Time Results

Figure 8.1 shows the summarized cycle time results for the *small single wafer tools* scenario. The cycle time performance of the scenarios is compared to the *single 24* scenario and the *single 24 (small tools)* scenario. Already at this level of granularity we can see that the tool dissolving process leads to a significant increase in PR and a significant decrease in QT. Lot size reduction then leads to an effective reduction in total cycle time, but the *single 12 (small tools)* scenario still falls short of the cycle time performance of *single 12* scenario. A reduction in lot size below 12 is necessary for the cluster tool dissolving process to be cycle time effective as is visible by comparison of Figures 8.1 and 7.12.

Figure 8.2 shows the relative cycle time performance by component compared to the *single 24* scenario at 92.5% loading.

- QT is the component targeted with the scenario design and the only component that shows a reduction without exception. The reduction in QT caused by smaller lot sizes is further discussed

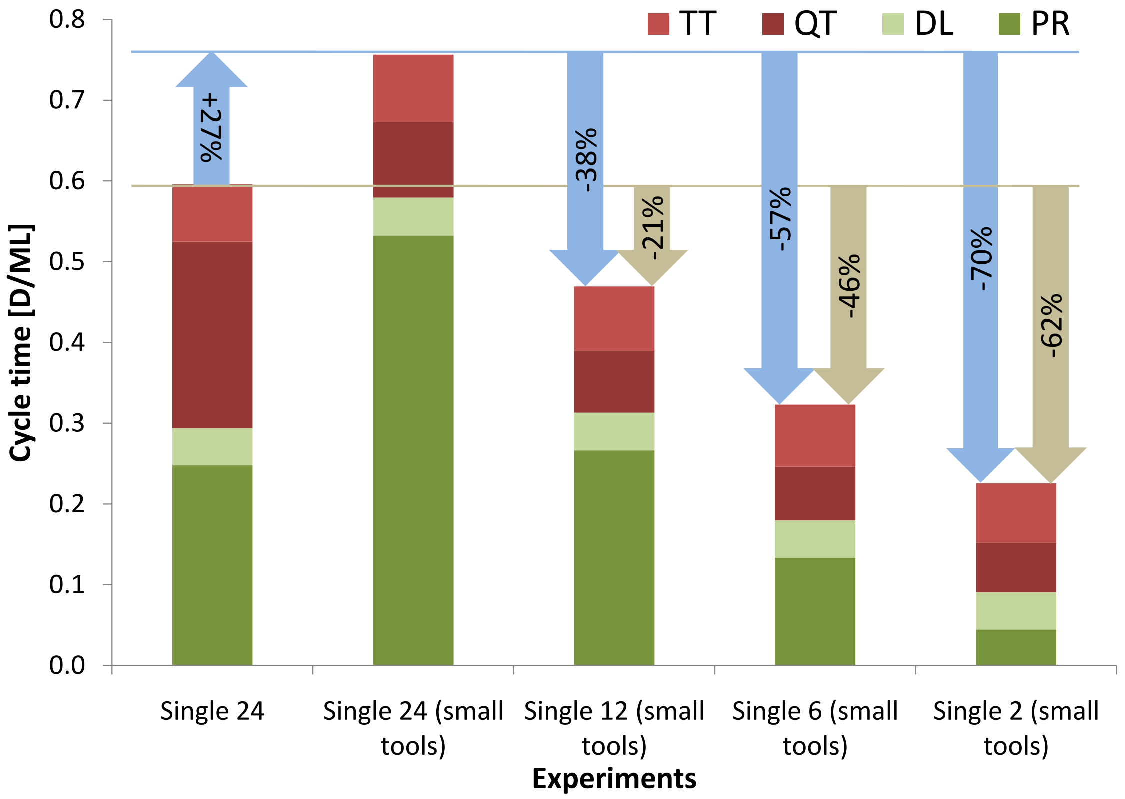


Figure 8.1: Lot size reduction scenarios for pure single wafer toolset (small tools) compared to *Single 24* scenario at 92.5% loading.

in the following subsection.

- TT is negatively affected because of the additional operations leading to additional transport tasks. The reduction caused by smaller lot sizes follows the same principle of fewer lot visits at metrology operations as previously discussed.
- PR is first increased very significantly by 115% through the tool dissolving process. With the reduction in lot size, this disadvantage is reduced, but the performance without tool dissolving is not reached (compare Figure 7.13).
- DL is only slightly affected. Because the dissolving of cluster tools with step integration took only place at a small number of tools without the necessity of vacuum conditions, the negative effect on DL discussed previously has little consequences.

8.1.2.3 Queueing Time Reduction Details

Our target with these scenarios was to achieve a persuasive queueing time reduction. The lot size-independent reduction is significant and clear in its origin in reduced variability effects, but we also want to highlight the reduction possible with smaller lot size with the dissolved cluster tools. Therefore we take a closer look at the relative queueing time reduction achieved by smaller lot size at different utilizations in Figure 8.3. In this figure, we show the effect of lot size reductions by a comparison of Single 12 (small tools) and Single 6 (small tools) scenarios to the Single 24 (small tools) scenario.

This queuing time reduction with smaller lot size is about twice as effective and subject to less variability compared with Figure 7.14 showing the relative queueing time reduction possible with the standard single wafer toolset and lot size reduction. This underlines one of the conclusions of our analysis with queueing theory in Section 7.3.2, that variability degrades the cycle time effectiveness of lot size reduction.

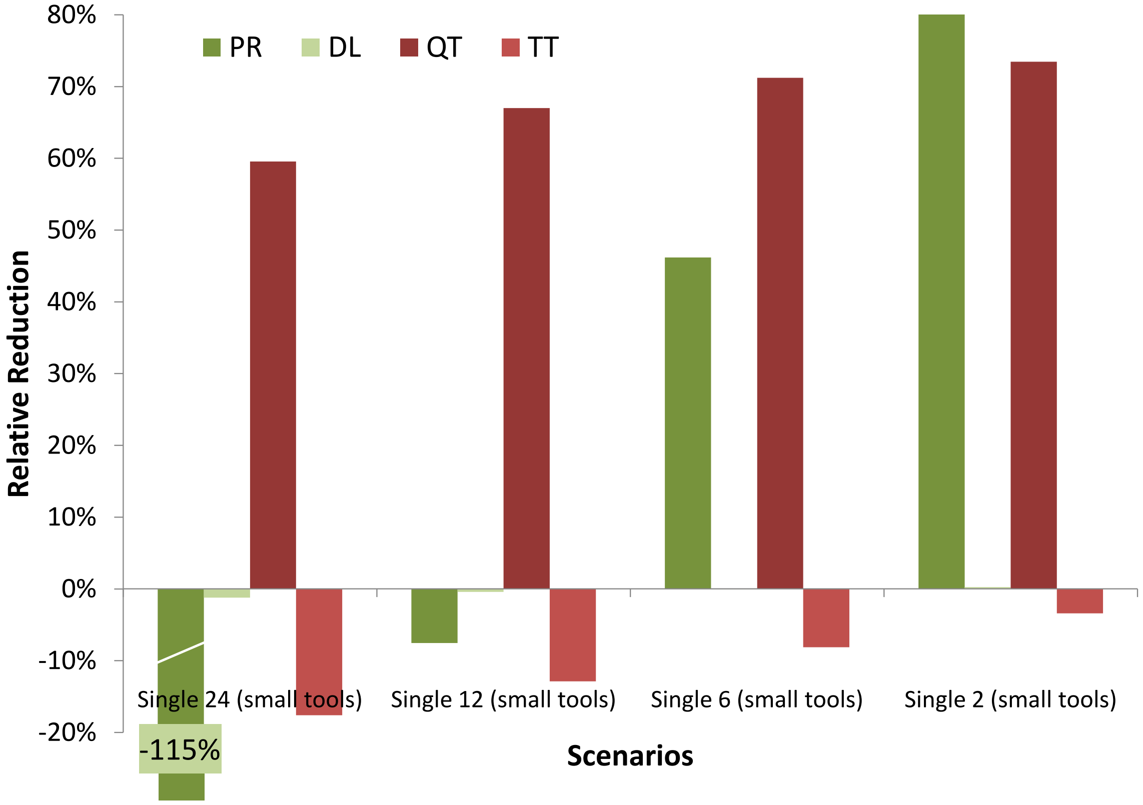


Figure 8.2: Reduction per cycle time component for for pure single wafer toolset (small tools) relative to *Single 24* scenario at 92.5% loading.

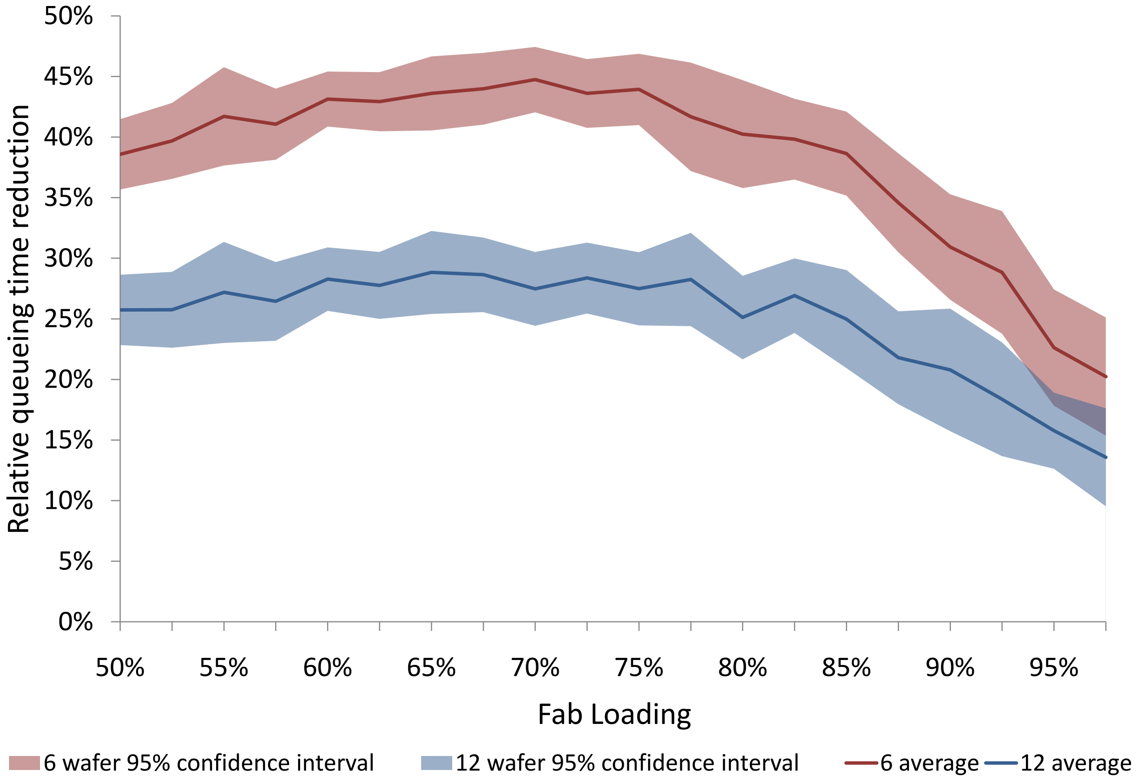


Figure 8.3: Relative queueing time reduction of 12 and 6 wafer lot size compared to 24 wafer lot size for pure single wafer toolset (small tools) at different fab loadings.

8.1.2.4 X-Factor Considerations

Table 8.1 lists the x-factors of the small tools scenarios under consideration within this section at a fab loading of 92.5%. Because of the higher processing time and lower queueing time, the x-factor is very small. Again cycle time performance comparison of fabs with these differences based on x-factor is not advisable.

Table 8.1: X-factor for scenarios at 92.5% fab loading

	X-factor							
	average				95-percentile			
Scenario / Lot size	24	12	6	2	24	12	6	2
Single	2.03	2.65	3.51	4.93	2.27	3.07	4.09	5.89
Single (small tools)	1.31	1.50	1.80	2.48	1.40	1.64	2.02	2.91

Apart from the average x-factor based on average cycle time, we also show the x-factor based on he 95-percentile cycle time enabling an impression of the cycle time variance. Cycle time variance is lower in the *small tools*-scenarios than in the normal single-scenarios. This applies both at the baseline lot size level of 24 wafers and at the reduced lot sizes confirming the effectiveness of this approach with respect to the reduction of variability.

8.2 Fab Scaling

There is a general trend towards bigger fabs in semiconductor manufacturing [O'H07], [Osb07]. There are two major reasons for this trend. First, bigger fabs need more tools per workstation, therefore workstation availability is subject to less variability, which leads to shorter cycle times. And secondly, bigger fabs have higher capacity utilization, because at the capacity demand limit defining whether another equipment is necessary the difference in utilization is highest for small tool counts. E.g., if a second tool is just necessary at a workstation, then capacity utilization is slightly above 50%. But if a fourth tool is just necessary at a workstation, then capacity utilization is slightly above 75%. This effect lowers the relative equipment cost for bigger fabs.

8.2.1 Theoretical Discussion

The benefits of bigger fab size have been studied several times, e.g., in [Ros06], and the prevailing trend seems to confirm them. We extent the area of analysis of previous studies by including the effect of smaller lot sizes.

8.2.1.1 Cycle Time Reduction Coherences

Because of the lower variability inherent to bigger fabs we expect smaller base values of QT at 24 wafer lot size compared to the single wafer tools scenarios at baseline fab size. Additionally we expect that the reduction in QT with lot size is more effective for the same reason. All other cycle time components are not directly affected by bigger fab sizes and should remain unchanged[1].

Unfortunately queueing theory with the approach of effective process times is of no help for assessing the queueing time reduction. The higher equipment count leads to a shorter queueing time at the baseline lot size according to Equation 4.2. But, queueing time reduction caused by smaller lot size, cannot be

[1] There is the possibility that transport times increase slightly for bigger fabs, but we disregard this in our scenarios.

assessed with queueing theory. Considering Equations 4.2 through 4.4, the only difference to the analysis in the previous chapter would lie in a different equipment count m. This, however, would lead to no different result when calculating the resulting relative queueing time reduction. The reason for this result is our exclusive utilization of preemptive downtimes in this approach, which prevents a more effective queueing time reduction for this scenario. Our simulation model does use non-preemptive downtimes, though, which is the case in reality, too.

8.2.1.2 Challenges

The operational challenges of a mega-fab with lot size reduction are mostly the same as for the normal-size fab, however, the total transport demand is increased. Therefore the realization of smaller lot sizes might be even more difficult in a big fab. There are different concepts of transport system separation and interconnection in these big fabs, therefore the challenge depends on the transport system approach and is difficult to assess in general.

8.2.2 Simulation Analysis

The basis for the simulation assessment of the cycle time performance of smaller tools is the single wafer tools scenario. We are interested in studying the size of the reduction in QT with lot size, therefore this represents the best basis.

8.2.2.1 Experimental Design

In the *big fab* scenarios we simply double the toolcount of all workstation of the *single 24* scenario and leave everything else unchanged. As before, we design scenarios for the lot sizes 24, 12, 6, and, 2.

8.2.2.2 Cycle Time Results

Figure 8.4 shows the summarized cycle time results for the *big fab* scenarios. The cycle time performance of the scenarios is compared to the *single 24* scenario and - for reduced lot sizes - to the *single 24 (big fab)* scenario. At the base lot size of 24 wafers, the total cycle time is 28% shorter in the bigger fab. The difference in total cycle time is fully attributed to a smaller QT. Lot size reductions then lead to a reduction in CT of 32% (12 wafer), 48% (6 wafer), and, 59% (2 wafer). This is a higher decrease compared with the reduction in the *single* scenarios of 24% (12 wafer), 36% (6 wafer), and, 45% (2 wafer). The additional reduction in the *big fab* scenarios is again attributable to a more effective reduction in QT, but also to the smaller original QT value putting more weight on the more effective reduction in PR inherent to lot size reductions. PR and all other cycle time components do not change differently compared to the *single* scenarios. Therefore there is no benefit in detailing the relative reduction by cycle time component, we rather analyze the queueing time reduction in more detail in the following.

8.2.2.3 Queueing Time Reduction Details

In Figure 8.3 we illustrate the relative queueing time reduction achieved by smaller lot size at different loadings. We show the effect of lot size reductions by a comparison of Single 12 (big fab) and Single 6 (big fab) scenarios to the Single 24 (big fab) scenario. With the exception of high loadings above 85% this queuing time reduction with smaller lot size is about 2.5 times[2] as effective as compared with Figure 7.14 which shows the relative queueing time reduction possible with the standard single wafer

[2] This ratio is valid for both the 12-wafer and the 6-wafer ratio to 24-wafer lot size.

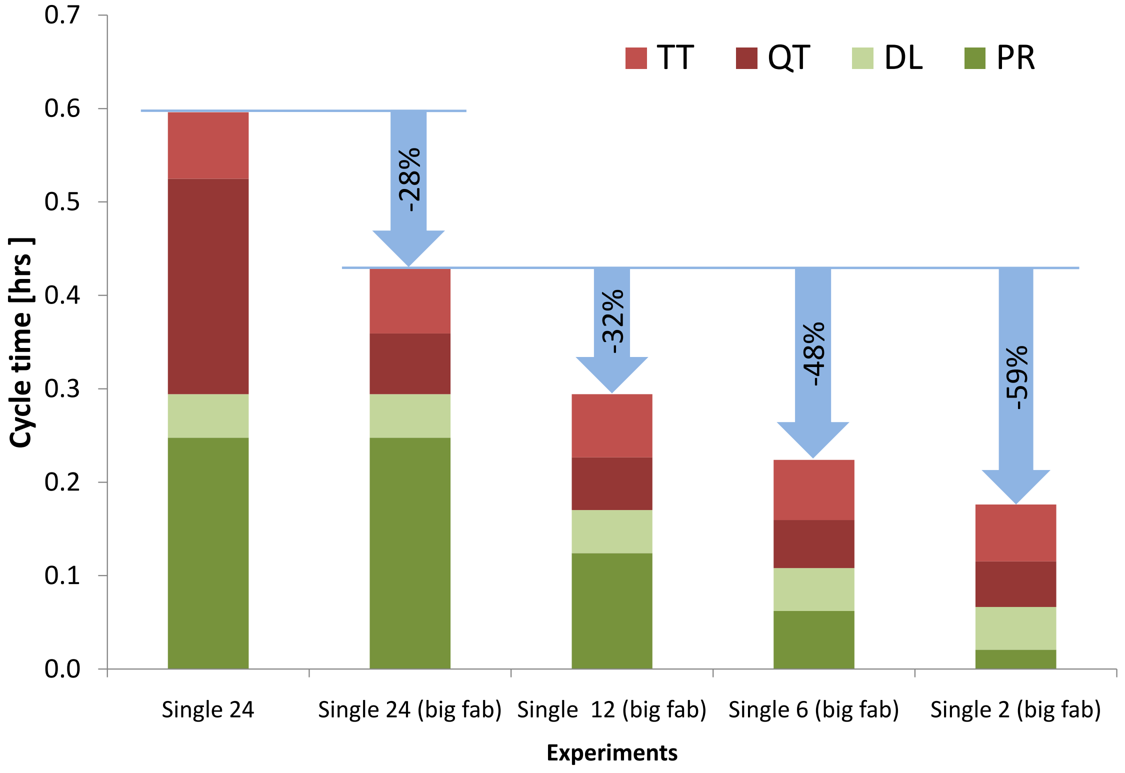


Figure 8.4: Lot size reduction scenarios for pure single wafer toolset (big fab) compared to *Single 24* scenario at 92.5% loading.

toolset and lot size reduction. Above 85% loading this ratio declines and ends at around 1.5 at 97.5% loading. In general the relative queueing time is subject to much less variability indicated by the narrower confidence interval. This again underlines the previous conclusion, that variability degrades the cycle time effectiveness of lot size reduction.

8.2.2.4 X-Factor Considerations

Table 8.2 lists the x-factors of the *big fab* scenarios under consideration within this section at a fab loading of 92.5%. Because of the shorter queueing time, the x-factor is smaller than in the *normal* single wafer tool scenario. Again cycle time performance comparison of fabs with these differences based on x-factor is not advisable.

Table 8.2: X-factor for scenarios at 92.5% fab loading

	X-factor							
	average				95-percentile			
Scenario / Lot size	24	12	6	2	24	12	6	2
Single	2.03	2.65	3.51	4.93	2.27	3.07	4.09	5.89
Single (big fab)	1.46	1.73	2.07	2.65	1.60	1.92	2.35	3.11

Apart from the average x-factor based on the average cycle time, we show the x-factor based on the 95-percentile cycle time as well in order to give an impression of the variance in cycle time. Cycle time variance is lower in the *big fab*-scenarios. This applies both at the baseline lot size level of 24 wafers and at the reduced lot sizes confirming the effectiveness of this approach with respect to variability reduction.

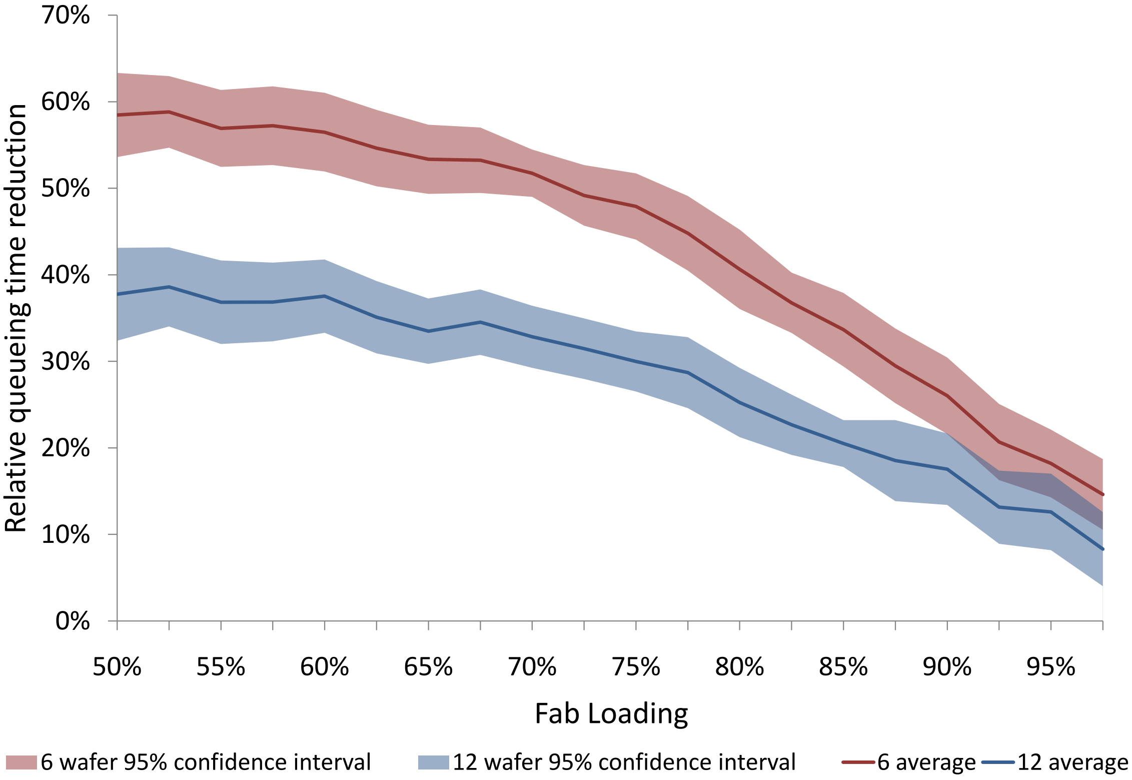


Figure 8.5: Relative queueing time reduction of 12 and 6 wafer lot size compared to 24 wafer lot size for pure single wafer toolset (big fab) at different fab loadings.

8.3 Conclusion

We summarize the key results of this section with the characterisitic curves for the scenarios *Single 24 (small tools)*, *Single 24 (big fab)*, and, *Single 24* displayed in Figure 8.6 (for the benefit of clarity, we do without characteristic curves of reduced lot size here.).

1. Both the *small tools* and the *big fab* scenarios show beneficial reductions in QT. Additionally to the moderate increase of CT with loading visible in the chart, the beneficial QT reductions are also expressed by more effective reductions in QT at smaller lot sizes.
2. Because of the increase in PR, smaller tools are unreasonable at standard lot size and also at moderately reduced lot size. This can be seen by the higher level of the theoretical limits in the chart. Because of the large challenges involved with very small lot sizes of six, and, expecially two wafers, we regard the *small tools* scenarios not as scenarios that are likely to see realization, but as scenarios that can guide further production system design ideas.
3. Additionally to their already superior cycle time performance, big fabs can reduce cycle time more effectively with lot size reductions. However, the issue of providing sufficient transport capacity for reduced lot sizes might be more challenging.

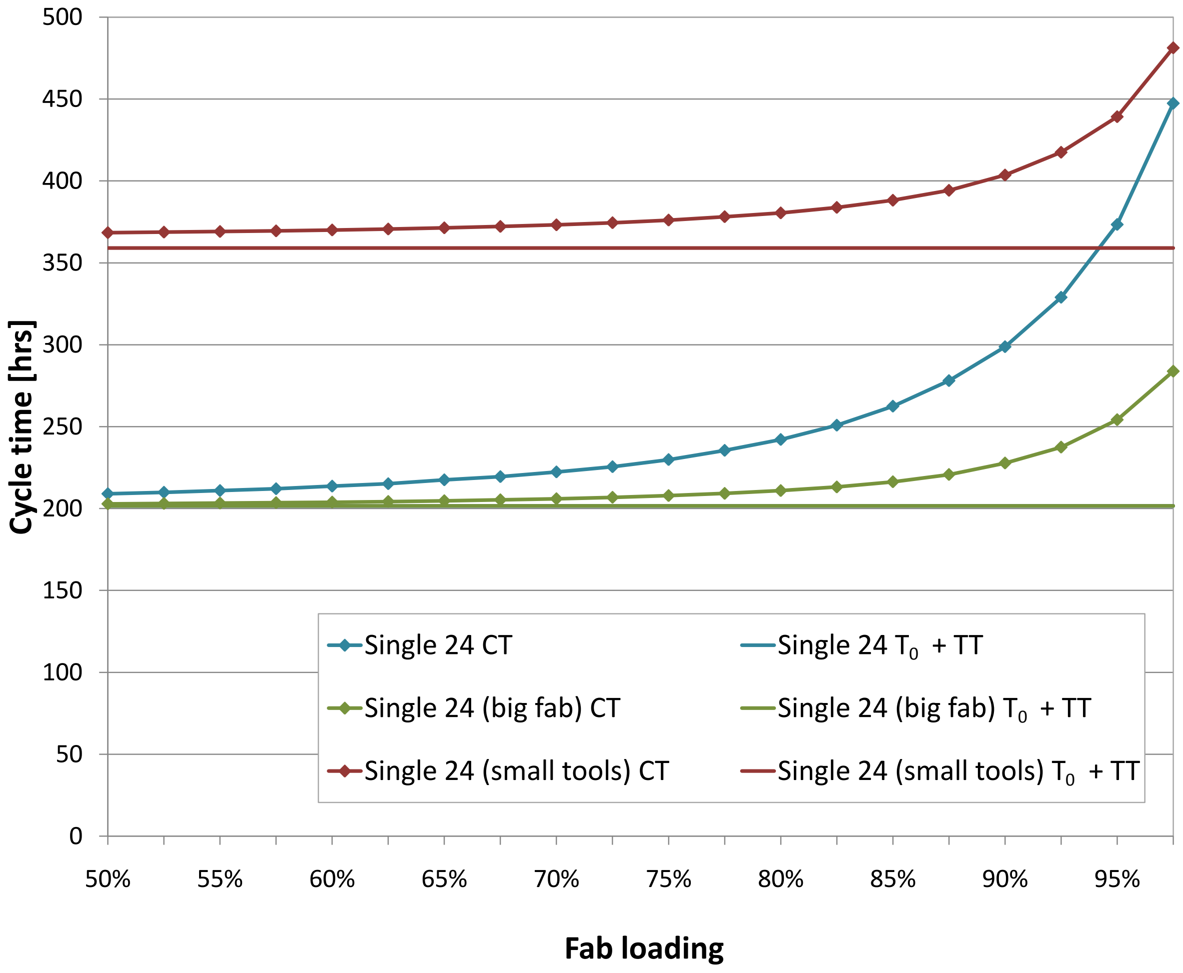


Figure 8.6: Characteristic Curve of scaling scenarios at 24 wafer lot size compared to *Single 24* scenario

9 Lot Streaming

The term *lot streaming* denotes a process of splitting a lot into sublots and then processing these sublots in an overlapping fashion at consecutive process operations in order to accelerate processing [TB93], [Bak93]. Unlike the *lot sizing* decision which leads to a fab-wide transfer size, lot streaming leads to a temporarily reduced transfer size between consecutive operations. In other literature the same behavior is described by the introduction of transfer batches which are smaller than the production lots [GF86]. We use the term *lot streaming* though because it describes the fundamental idea very well in contrast to the *lot sizing* decision.

The lot streaming concept has not been used due to operational considerations in front-end semiconductor manufacturing. However, the integration of stepper and track to combined litho cells effectively enables lot streaming for the three basic operations that are performed: coating (in track), exposure (in stepper), and, development (in track). Because the two tools are physically linked by transfer places for individual wafers, the transfer size between these tools or operations respectively is one wafer. Therefore lots are streamed and processing overlaps, i.e. the processing of wafers at the coating operation is still in progress, while wafers of the same lot already undergo the exposure operation. The decision for this integration is motivated by process quality reasons (The direct transfer enabled a short and reliable time period between the three operations.) and not operational considerations, though.

In back-end semiconductor manufacturing several tools serving following operations are connected by a conveyor-band. This can also be regarded as an application of the lot streaming concept, but the context is not comparable. The material flow in back-end semiconductor manufacturing shows no reentrance and the tools have a very high availability. Therefore this part of semiconductor manufacturing is suitably for the rigid coupling of a conveyor-band.

9.1 Motivation

Several factors motivate an application of the lot streaming concept. In the previous Chapter 7 we came to the conclusion that lot size reduction should not be taken to the ultimum, because the benefit seems diminishable regarding the challenges involved, especially the transport issue. Lot streaming might be a concept that uncovers the improvement potential of one wafer lot size to a significant extent without the challenges.

Additionally the lot streaming concept is suitable to integrate other changes that seem appropriate based on the previous results. First, the increased number of transports necessary at reduced lot sizes have to be addressed. The obvious solution to this issue is to reduce the transport demand, or to hand on transport tasks to a different system. Lot streaming does not directly lead to a reduced transport demand, in fact total transport demand increases, because of the smaller transfer size at the *streamed* operation changes. However, this calls for a different system, and in this way transport demand of the fab-wide AMHS is reduced throught the back-door.

And, secondly the size of the delay DL of the total process time virtually stays the same with lot size reductions and consequently its share of the total process time increases. This makes it a worthy target for an improvement. In the previous paragraph, we identified the need for a different system for the transport for *streaming* operation changes. Provided that this transfer is situated in a tool connection that makes some steps associated with the loading and unloading of wafers obsolete, then DL is reduced. We analyze approaches for this tool connection in the following Section 9.2.

Another consideration worthwile is the definition of the ultimate target. In [Ign08], Ignizio describes an ideal fab. In the ideal semiconductor fab, the "utopian fab,(...) machines would be placed (...) according to the sequence of process steps. There would be one machine per process step and the resulting configuration would form a non-reeentrant, serial, single-unit process flow." The ideal processing unit in this *pipeline* fab design would be a single chip. Of course, this is wishfull thinking. To name just two obstacle, the huge capacity inequality between different tools and process steps at very high equipment costs as well as the high variability in availability forbids this approach for the foreseeable future. But, each transition in this direction that avoids negative side-effects, is a positive step in the direction of the optimum.

9.2 Existing Solutions

How could such a tool connection enabling lot streaming look like? In the literature, we find two proposals that come close to matching the requirements.

9.2.1 EFEM-bond

In [SIA01] different concepts of an EFEM connection were discussed (see Figure 9.1) that could represent first solutions for a limited transition from lot handling to single wafer handling.

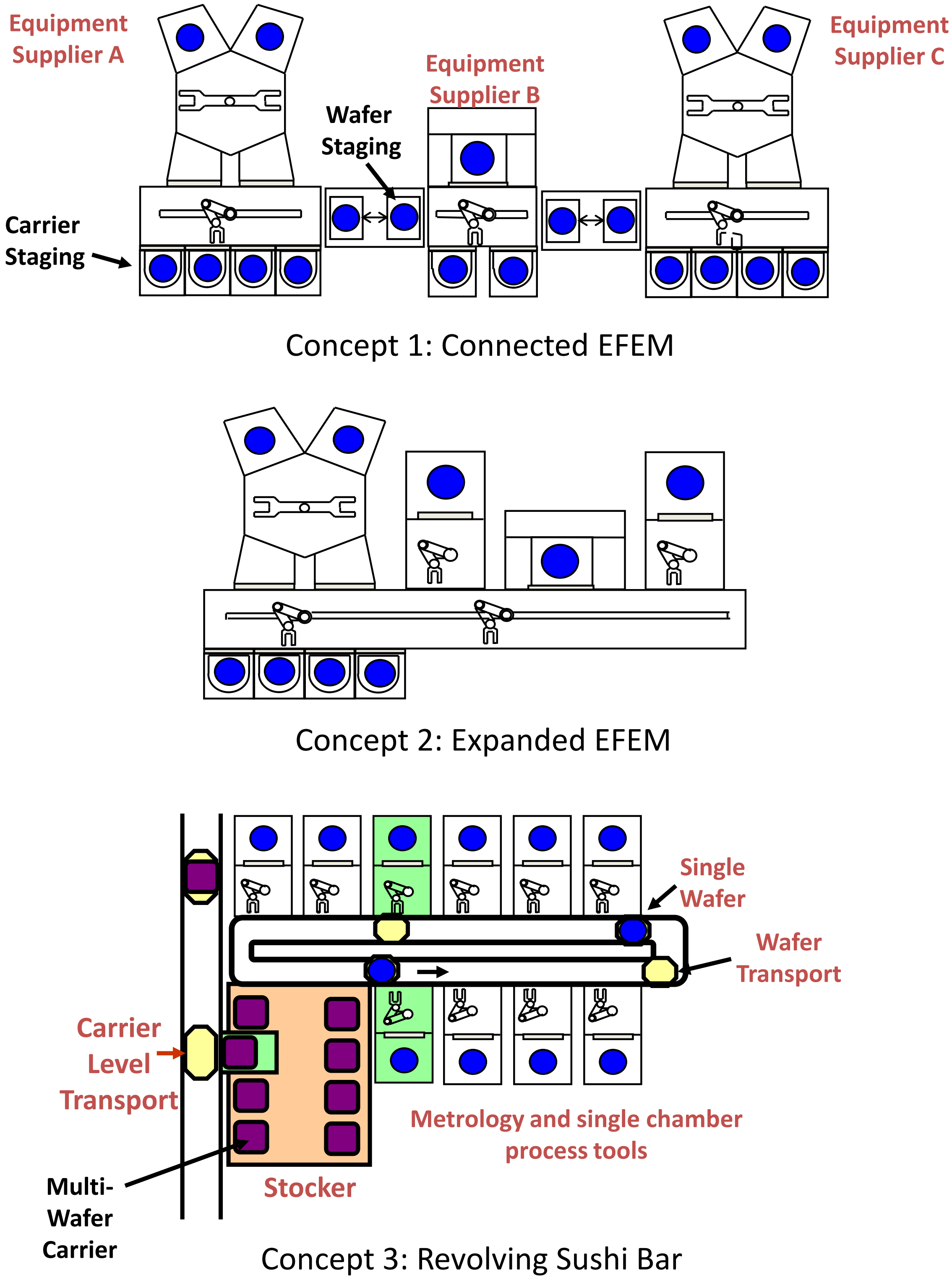
In the first concept, the EFEMs of tools standing next to each other are connected. The wafer stages forming the connection buffer individual wafers and are accessible by the EFEM robots of the adjacent tools. For the connection to make sense, the adjacent tools should frequently perform subsequent process steps. In these cases wafers of a lot would be processed in one tool, then be placed in the wafer staging area, and afterwards be processed on the second tool without visiting the carrier in between. Depending on the empty carrier management, the carrier of the lot would rest on the original load port of the first tool and the wafers would be handled back to it with another visit to the wafer staging area, or, the empty carrier would be transported to a load port of the second tool. Of course, the procedure is extendable to more than two tools.

The concept addresses two of the motivating factors discussed above. It enables lot streaming, as the first wafer of a lot can enter the following tool while other wafers of the same lot are still processed or waiting to be processed at the first tool. And it reduces the transport demand, because the transport between the *streaming* operations is performed by the EFEMs[1]. However, the size of the delay remains unchanged, because the tool-internal wafer handling basically stays the same. The only change is that wafers are loaded to, and unloaded from, the wafer staging area instead of the carrier. There is no reason that this should consume a significantly different amount if time.

The second concept of an expanded EFEM differs only slightly from the first one. Instead of using wafer stages, one expanded EFEM forms the front-end of several tools and its one ore more robots serve all tools. This can speed up the wafer transport between the tools and lead to a slight operational advantage. Other than that the discussion of the previous paragraph on advantages and disadvantages of the connected EFEM applies to the expanded EFEM as well.

The third concept of the revolving sushi-bar features operational enhancements. The bi-directional wafer handling in the expanded EFEM and - depending on the empty carrier management - also in the connected EFEM limits the possible throughput of the EFEM. Therefore the number of tools that can be

[1] To be precise, the possible empty carrier move represents a new transport demand. However, it generates only one move request instead of often two move request between operations because of the intermediate carrier buffering at a storage location and it is a very short move.



Concept 1: Connected EFEM

Concept 2: Expanded EFEM

Concept 3: Revolving Sushi Bar

Figure 9.1: Different EFEM connection approaches discussed in ITRS Factory Integration TWG [SIA01]

integrated into one EFEM-bond is limited. The directed material flow in the revolving sushi-bar EFEM enables more independent transports and avoids the EFEM from becoming a bottleneck. In this way more steps can be integrated into one EFEM-bond. This enlarges the share of operation changes with the possible application of lot streaming, because there are less operation changes requiring an AMHS transport which cannot support lot streaming. Another change is apparent in this concept. Because of the high number of processing resources that can be integrated, the load ports have been replaced by a carrier stocker that can hold more carriers at once. However, the discussed disadvantage of the other concepts remain. There is no change in the material flow that can reduce the size of the delay. It is not possible to operate an area with the physical dimension of the revolving sushi-bar under vacuum, therefore load lock visits are still necessary at every unit connected to it that requires vacuum. Otherwise this could have been a possible delay reduction measure.

A hidden operational advantage of all three concepts is worth further discussion. Downtimes of a tool or chamber would not affect the availability of the other resources in the EFEM-bond. It is easily possible to access the EFEM-bond for a subset of the theoretically possible operations, e.g., because a processing resource is unavailable or capacity of the steps is unequally distributed to save resources (see Section 9.4 for an example). This is in contrast to cluster tool operations, where, e.g., downtimes of processing resources lead to unusable capacity at other steps in the same tool.

Another development is implicitly shown within the three concepts. While the first two concepts plug tools to the EFEM, the third concept plugs individual chambers to the revolving sushi-bar. Requiring a longer development timeline than for the other concepts, this concept assumes that chambers are available individually and conform to standards which enable a flexible allocation of chambers to the material handling part of a tool or system. In the example, two chambers are colored green to illustrate the integration of metrology tools into the bond. However, application of this new flexibility in the revolving sushi-bar is not advantageous. As discussed in the previous paragraph all connected units, in this case chambers, have to provide vacuum themselves if necessary. This means that a load lock has to be attached to each process chamber or that the process of creating vacuum has to be performed in the process chamber. The cost of the extra hardware or the prolonged process time in expensive process chambers is not reasonable. However, the concept of individual chamber allocation is further utilized in the following concept of linear cluster tools discussed in the next paragraph, where it is of more benefit.

9.2.2 Linear Cluster Tools

In [vdM07] and [vdMRPK06] Blueshift Technologies present their linear cluster tool concept. Figure 9.2 shows an illustrative configuration of the linear cluster tool design. Through the addition of a pass thru and a frame with a single-blade robot the configuration is extendable as necessary.

This design breaks with the paradigm of bidirectional material flow in cluster tools. Current cluster tools (see Section 3.4.3.3) handle wafers back to the original carrier at the original load port. In the linear cluster tool design, however, wafers are handled only in one direction leaving the carrier at one side of the tool and entering the original or different carrier at the other side of the tool, hence the name *linear*.

This addresses two disadvantages of conventional cluster tools' material handling. First, the single interface EFEM causes complicated carrier handling and challenging carrier exchange time requirements (see Section 7.3.3.1) and secondly, the bidirectional wafer handling limits throughput and prevents the integration of more chambers into a cluster.

In between the two EFEMs of the linear cluster tools wafers are brought to vacuum in the load lock and then visit one or several chambers. The chambers are accessible by single blade robots that are connected by pass through places. The number of chambers visited by the wafers is dependant on the configuration chosen. It is possible to have only chambers of the same type and then wafers visit only one chamber, leading to high throughput of the cluster tool. If all chambers are of a different type

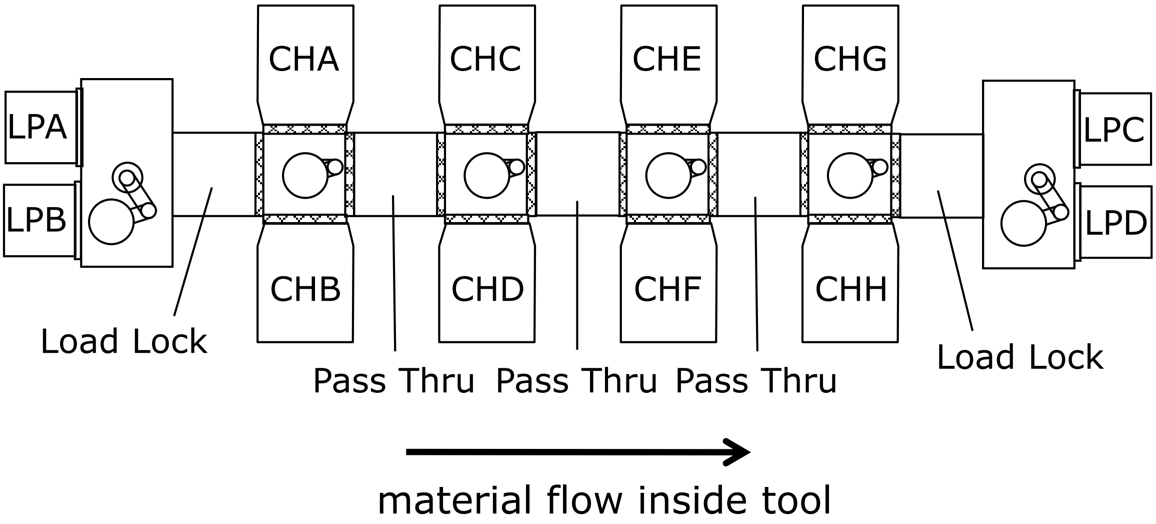


Figure 9.2: Linear cluster tool according to [vdM07]

representing subsequent process steps then wafers might visit all chambers after each other, enabling lot streaming for many subsequent operations.

The linear cluster tool also utilizes individual chambers that are plugged to the *automation part* of the tool. In contrast to the previously discussed example of the revolving sushi bar, the connection of indicidual chambers makes sense, because the vacuum is already provided.

The linear cluster tool concept addresses all three motivating factors. It enables lot streaming, as the wafers of the same lot are processed in different chambers at the same time, which can belong to different operations. And it reduces the transport demand, because the transport between the *streaming* operations is performed within the linear cluster tool[2]. Additionally, the delay is reduced significantly, because the number of load locks as well as the load and unload processes in the EFEM is reduced significantly.

However, there is a negative side-effect. In Section 3.4.3.3, we discussed issues of the integration of sequential steps into cluster tools. These were the capacity inequality between different steps and temporary unavailability of chambers due to failures which both leads to wasted capacity. In the linear cluster tool design these issues become more severe, because more steps can and should be integrated. Consequently effective capacity and cluster availability is both lower than without this integration. The obvious work-around to process several lots in parallel that utilize only selective steps is not persuasive, because the internal material handling does not support the additional number of transports associated with such an operational mode.

9.3 Our Approach

Our approach tries to combine the strengths of the two proposals. We take the linear cluster tool design and add a more flexible accessibility as a feature of the EFEM-bonds. This accessibility has to enable *micro-flows* within the tool that only use a part of the tool's full flow. In this way the available capacity can be utilized flexibly without stressing the tool internal material handling.

Figures 9.3 and 9.4 illustrate our solution. We add access points at the point(s) in the material flow where

[2] To be precise, the possible empty carrier move represents a new transport demand. However, this one short move compares to several times (depending on the number of operations integrated into the cluster tool) two moves between operations (intermediate carrier buffering at a storage location).

wafers shall be fed out of and into the linear cluster tool. Ideally these additional access points do not cost space although they provide the same basic functionality as full EFEMs. Conforming to these space restrictions, we add a loadlock to the stack of the two pass thru places and locate a load port on top of it. The load lock and the load port are served by a robot that is located to the unused side of the pass thru. Because of the lower load that should be placed on these intermediate access points, one load port should be sufficient to serve the needs.

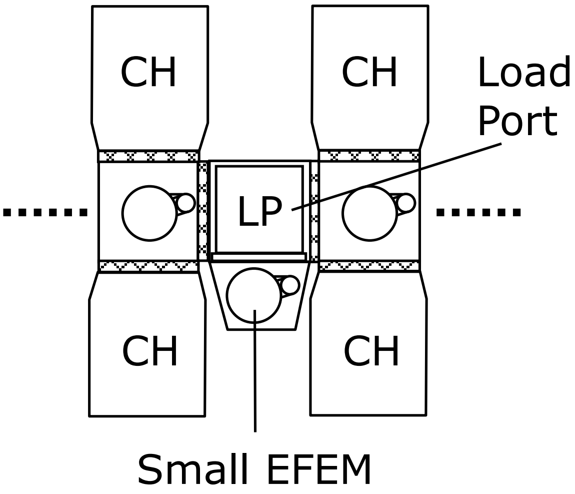


Figure 9.3: Top view: Integration of additional access point with load port into linear cluster tool design (dots indicate that tool continues)

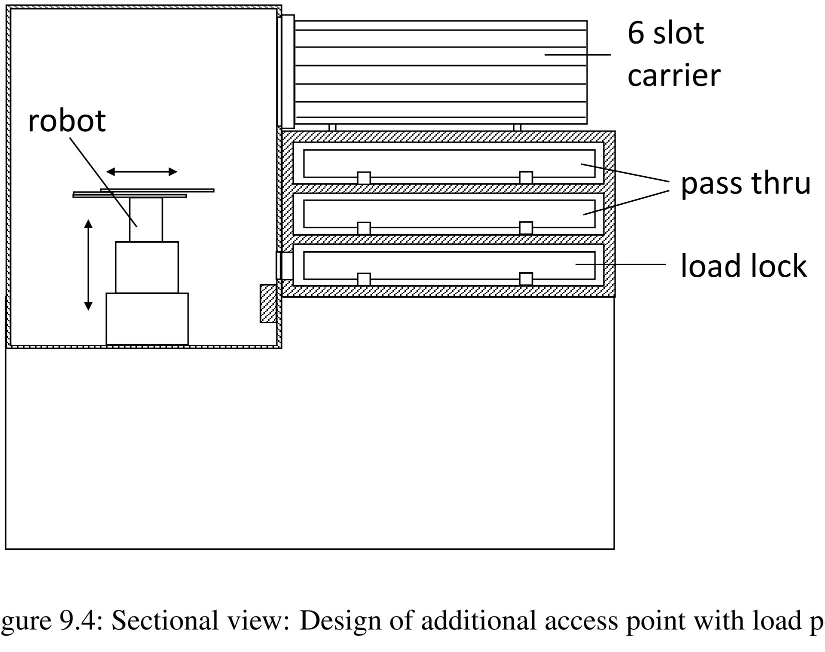


Figure 9.4: Sectional view: Design of additional access point with load port

9.4 Examplary Illustration

We illustrate the operation of the modified linear cluster tools in an example. This example tool group of linear cluster tools serves the following operations and has the following number of chambers available.

- Operation 1: Wet clean process; 9 chambers available
- Operation 2: CVD process; 12 chambers available
- Operation 3: Measurement; 2 chambers available

We distribute the chambers among two linear cluster tools. Because of the odd count of chambers for the wet clean process, tool 1 has 5 chambers for this process and tool 2 has 4 chambers. The other chambers are distributed equally. Figure 9.5 illustrates the resulting tool. The chambers are denoted as

- CHA-CHE (wet clean process; CHE only tool 1),
- CHF-CHK (CVD process), and,
- CHL (measurement).

Both tools have five load ports. Load ports LPA and LPB are responsible for the wafer feeding from the carrier into the tool and load ports LPC and LPD are responsible for unloading the wafers from the tool into the carrier. Load Port LPE represents the intermediate access point, where wafers can be fed in and out between the wet clean and the CVD process. There is no additional access point between the CVD process and the measurement, because of the close load ports LPC and LPD, which can provide access with sufficient performance.

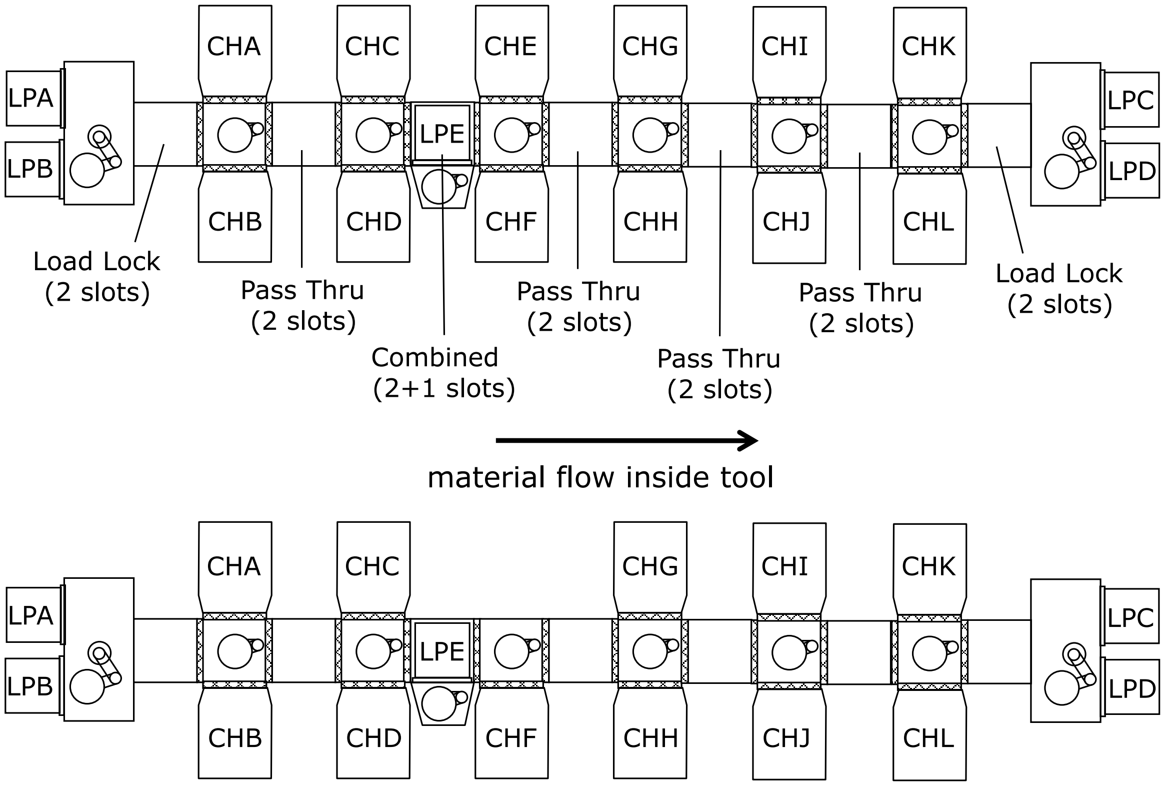


Figure 9.5: Design of linear cluster tool example

The operation of the two linear cluster tools is shown in two simplified Gantt-charts in Figures 9.6 and 9.7. The results are generated with a slightly modified version of the dynamic tool cluster-tool simulator developed in [Bec08]. To avoid confusion, we limit the resources displayed in the Gantt-charts to chambers and load ports. The different colors represent the different carriers and its wafers. Each carrier visits two load ports - once for the delivery of wafers to the tool and once for their pickup (the first three carriers are additionally marked by asterisks to improve orientation.). Most carriers contain six wafers representing the standard lot size in the example. Additionally some carriers contain only one wafer. Those are the wafers intended for a transition between the two tools between the wet clean process and the CVD process. This transition is necessary because of the uneven capacity allocation between the two tools. Similar transitions would be necessary in the case of chamber downtimes that unbalance capacity.

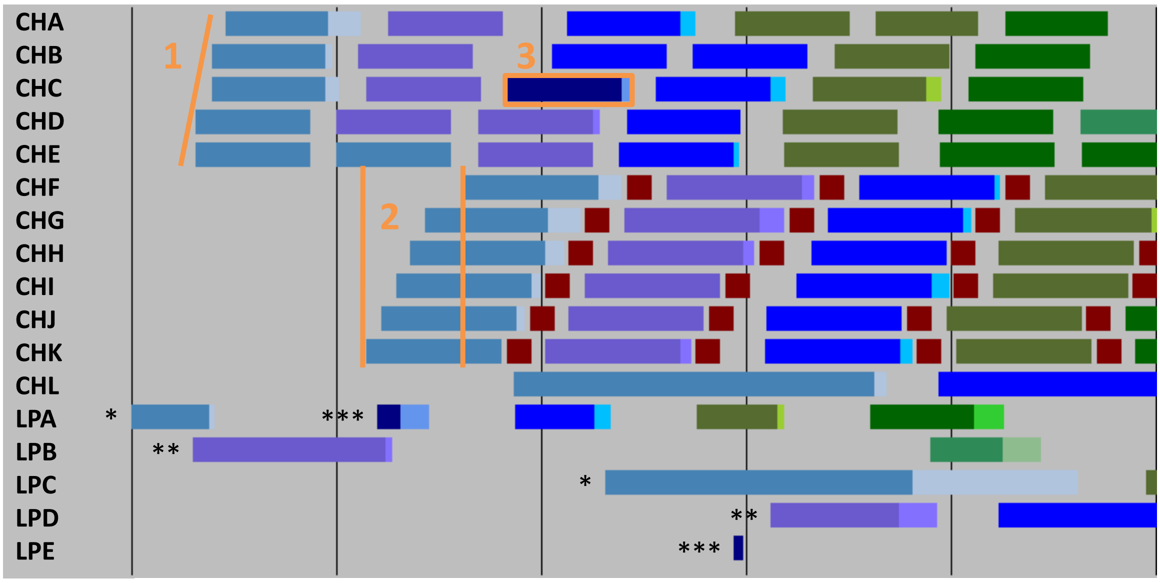


Figure 9.6: Gantt chart of Tool 1 operation

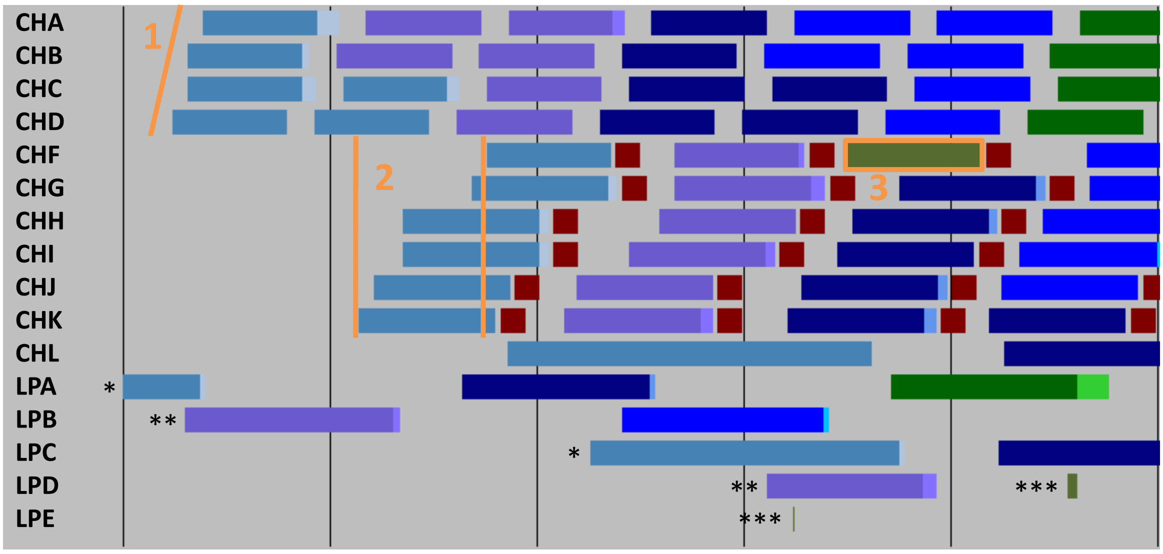


Figure 9.7: Gantt chart of Tool 2 operation

We want to highlight some points of interest in the Gantt chart. Marked with the orange number one an area that shows the importance of parallel chambers that provide the same application. Because of the five or four chambers, respectively, performing the wet clean process many wafers of the same carrier can begin processing very fast. Therefore it remains vital that parallel chambers are available within the same tool as long as lot size is not reduced to one or two wafers.

Marked with the orange number two is the time span that delivers the benefit of lot streaming. Without lot streaming processing at the CVD process chambers could not begin before processing of all wafers of a carrier at the wet clean process is finished. Taking some handling time into account, this is the beginning of the CVD process for the last wafer of the carrier which is marked by the right line. The time span between the left and the right line marks the benefit gained by the earlier processing start enabled by lot streaming. Considering additional handling time if the processes are performed in different conventional cluster tools, the benefit is even bigger than can be illustrated in the example.

Finally, marked with the orange number three is the processing of the first wafer that is transfered between the two tools. The wafer is delivered in the carrier marked by the three asterisks on LPA of tool 1, then processed at CHC and reenters the carrier at LPE. At tool 2 the wafer starts from load port LPE again, is processed in CHF without significantly affecting the processing of the other wafers in the tool, and eventually leaves the tool at LPD.

We want to point out that this presents only one possibility for making use of the capacity split between the tools. It is also possible to steer out wafers of a carrier during processing. Then one additional empty carrier is necessary to make the tool transition and the original carrier then might pick up all its wafers again from two unload load ports of two separate tools. Numerous operational policies are possible to make use of varying operational situations. It will depend on the control and MES capability of a fab which policies are possible in reality.

9.5 Material Handling System Design and Implications

After defining the tool design and operation and illustrating it with an example, we now turn to the material handling system. The operation of the linear cluster tool requires on-time placement and pick-up of carriers at load ports. Especially the empty carriers have to arrive timely to avoid congestion within the tool, which would reduce throughput. With the short transports to intermediate storage locations for empty carriers this calls for a small local transport system serving some tools and a fab-wide system responsible for moving carriers between these local transport systems. In this way response and transport times in the local system are more predictable than in a unified system.

The flexibility that has been incorporated into OHT systems recently makes them suitable for this application. The possibility to move in both directions and the possibility of both under-track and side-track storage are ideal features for the small local transport system. The bi-directional move capability enables using the shortest travel distance. The local storage capability suits the intermediate empty carrier storage and is a flexible option for the carrier transfer between the fab-wide and the local system.

Figure 9.8 illustrates the transfer between the fab-wide and local material handling system in a side-view. Carriers (blue) are delivered by the fab-wide AMHS system (green) to the side-track storage that works as transfer buffer. The local AMHS system takes the carrier out of the side-track storage and delivers it to the equipments' load ports. This transfer operation is similar to the OHT buffering approach for single tools we have outlined in [ZRS$^+$07].

The design of the local AMHS is further illustrated in Figure 9.9 with the two linear cluster tools of the above example. The local AMHS (red) forms a rectangle[3] so that it can serve two linear cluster tools. The EFEMs at the end of the tools are bent in the direction of the AMHS. This enables the AMHS to serve both the load ports at the tool ends and the intermediate load ports. The load ports at the bent EFEMs are not directly under the track but they remain reachable because carriers can be handed off to the side as necessary to reach the side-track storage. The transfer places connecting the systems are side-track storage places as introduced in the above paragraph. In the example shown in the figure we incorporated three transfer places on the left and right side of the tools. Apart from the transfer places there are additional storage places to hold empty carriers. These are implemented as under-track storage places above the automation part of the linear cluster tool.

[3] Practical considerations would most likely lead to a connection between the two OHT systems, that serves to move vehicles in and out for maintenance activities. The connection would not be used in normal operation, therefore it is not shown in

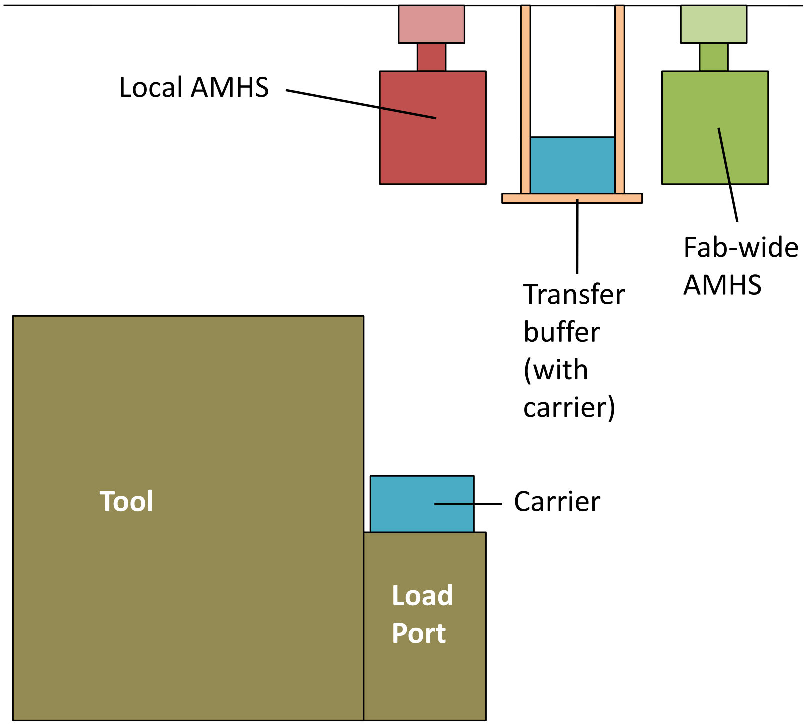


Figure 9.8: Side view of the transfer between fab-wide and local material handling system

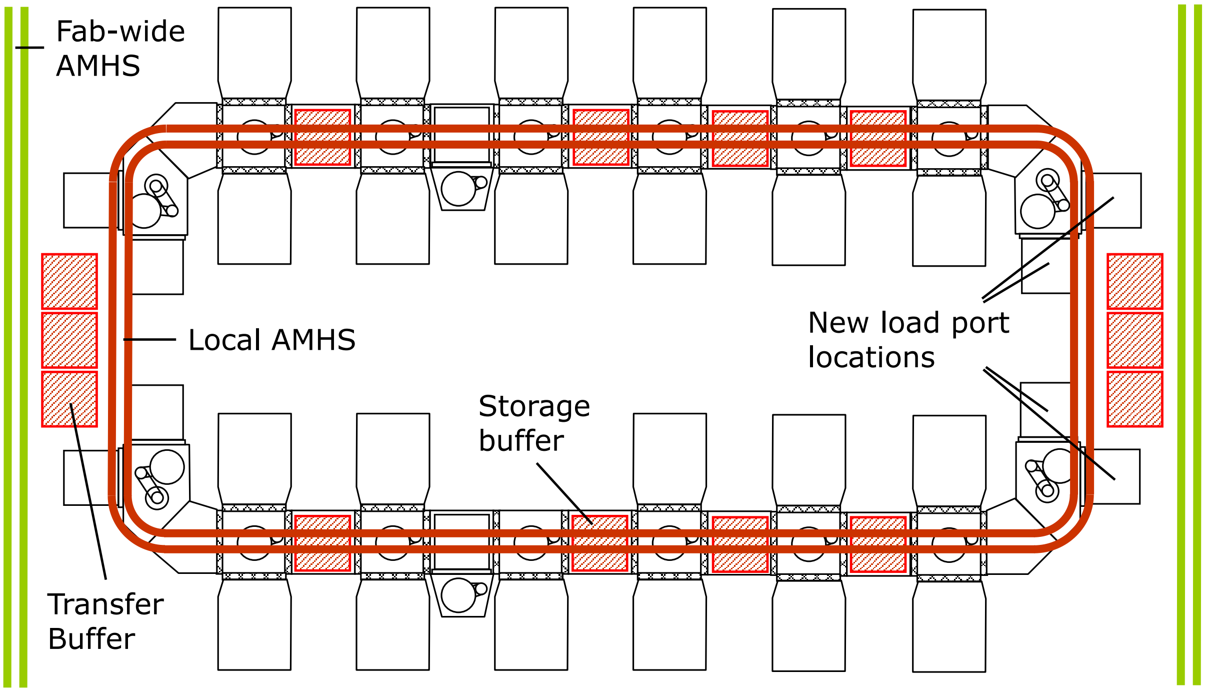


Figure 9.9: Top view of material handling system serving the two linear cluster tools

Figure 9.10[4] illustrates the operation of the local material handling system in a Gantt chart with our two cluster tool example. We display all resources of the local system that can hold carriers. These are

- the storage buffers (implemented as under-track storage places), named SPA-SPE,
- the transfer buffers (implemented as side-track storage places), named TPA-TPE,
- the two vehicles, named Vehicle 1 and Vehicle 2 with Vehicle 1 primarily serving tool 1 and Vehicle 2 primarily serving tool 2, and,
- load ports LPA-LPE of both tools, tool 1 and tool 2.

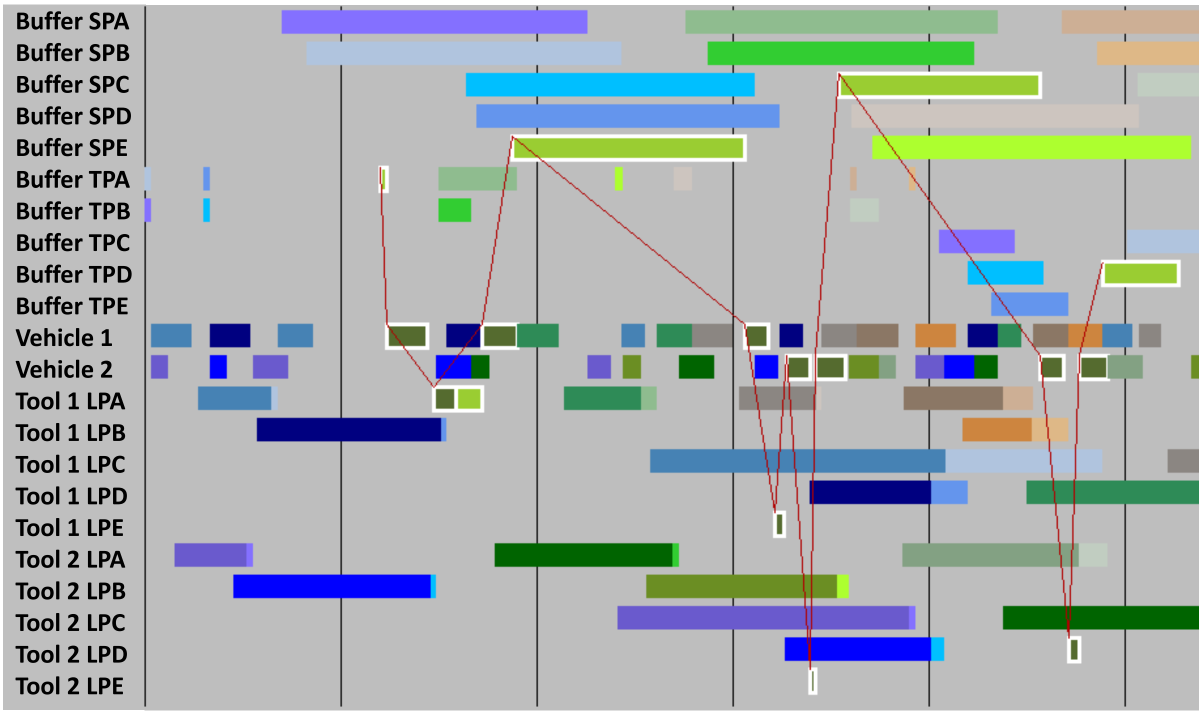


Figure 9.10: Gantt chart of material handling system operation

In the chart we highlight the path of one carrier visiting both tools. The carrier enters the system at the transfer place TPA. Vehicle 1 then transfers it to load port LPA and after its wafer is fed into the system loads it onto the storage place SPE. When the associated wafer waits in the load lock for delivery to load port LPE, then the carrier is picked up by Vehicle 1 again and placed on LPE, where the wafer is fed into it. Vehicle 2 then picks the carrier up and transfers it to LPE of tool 2. Then the wafer is fed into the carrier and Vehicle 2 transfers the wafer to the storage port SPC. After the wafer finished processing in tool 2 and is ready for delivery to the load port, Vehicle 2 picks the carrier up again and transfers it to the load port LPD. After the wafer is fed into the carrier, Vehicle 2 picks it up and places it on the transfer port TPD, where it leaves the local material handling system.

The Gantt chart shows that the material handling system environment we envision for the created linear cluster tools works. The number of vehicles might have to be increased in case of tools with higher throughput because vehicle utilization is already high in our example.

the figure.

[4] Note that the storage locations show only light colors when a carrier is present because there is no actual process

9.6 Simulation Analysis

For the assessment of the new production system created with the modified linear cluster tools we adjust our simulation model to reflect a fab operating with these tools. In the following, we first describe the steps which we performed for this adjustment and then discuss the cycle time results.

9.6.1 Experimental Design

In order to create integrated linear cluster tools, we first divided the process flow into segments of several process step sequences. The specific segments were chosen with respect to protocol zone (lithography area, copper area, non-copper area), process reasonability[5] and segment similarity.

As second step, we grouped like segments and built linear cluster tools of the associated process resources. In total we created 17 different segment groups and consequently 17 different linear cluster tool workstations. As a starting point for building these new cluster tool workstations we used the smaller tool units established for the scenario with small tools. According to the capacity needs of the segment groups we distributed these small tool units to the new cluster tool workstations. Because of the static tool unit count the distribution of the tool units is not exactly the same as the capacity ratio between the segment groups, but represent the closest fit.

In step 3, we determined more closely how the actual cluster tools look like. We limited the number of tool units for one new cluster tool to a maximum of twelve. Most cluster tool workstations are rather small with this limitation. 13 of the 17 workstations have two to four tools (and two of the four bigger workstations are lithography and implant which are discussed in the following). Within one workstation not all cluster tools look the same because the tool unit count does not allow for an even distribution. However the unit count does not differ by more than one.

Two of the cluster tools built form an exception: The lithography clusters of track and scanner already represents an abundant agglomeration of process steps and therefore they are not suitable for further integration. We assume only one modification from the baseline scenario for litho tools: the addition of carrier access points between scanner and track which enable feed-in and feed-out of lots after coating/before exposing and after exposing/before developing. In this way, we obtain litho clusters that operate similarly to the new cluster tools.

The second exception are implant tools which are impossible to integrate into the linear cluster tool concept because of their physical dimension. Therefore we assume no change for implant tools.

With respect to the process flow some adjustments have to be made in the model to reflect the new operational design. We now have to distinguish between transport size and processing lot size and we have transports to and from intermediate access points with associated lot de-streaming that only occur for a fraction of the lots.

In order to integrate this behavior into the model, we introduce transport operations between at all points of possible (but not always required) transports, i.e. between all operations that can be performed by the same cluster tool. This operation uses a dummy tool of a dummy tool workstation with infinite capacity. The operation's processing time is the time necessary for de-streaming the transport-size lot and the transport time is the uniform transport time of 0.15 h. However, in the normal case this extra transport operation is skipped, because the lots continue processing at the current linear cluster tools. The ratio of lots, which perform the transport operation are calculated from the probability of three cases:

- Unplanned maintenance share at the following operation.
- Planned maintenance share at the following that exceeds planned maintenance share of the current

[5] e.g., it makes sense to have an etch operation with the following clean operation and the related metrology operations in one cluster tool, because this enables fast feedback

operation (This assumes that planned maintenance activities are synchronized within a cluster tool.).

- Capacity overloading share caused by the unequal distribution of the small tool units.

Some characteristics of metrology operations have to be highlighted. In some cases, metrology units could not be integrated into all cluster tools of a workstation, because the metrology tool unit count does not allow for it. Therefore there is a higher share of additionally necessary transports of metrology operations. However, because we allow the transport of single wafers to metrology operations, a high skip rate at metrology operations follows. Therefore the additionally necessary transports for metrology operations do not carry much weight with respect too their contribution to total average cycle time.

9.6.2 Cycle Time Results

We start the presentation of our results with a comparison of the number of inter-equipment transports. Figure 9.11 displays the number of transports for the baseline scenario and the scenarios with the new modified linear cluster tools. Due to the integration of subsequent operations into a tool, the total transport demand is lower for the scenarios with the modified linear cluster tools. Additionally, because individual wafers are transported to metrology operations regardless of lot size, the transport demand of metrology operations does not increase with lot size reduction which leads to a smaller total increase.

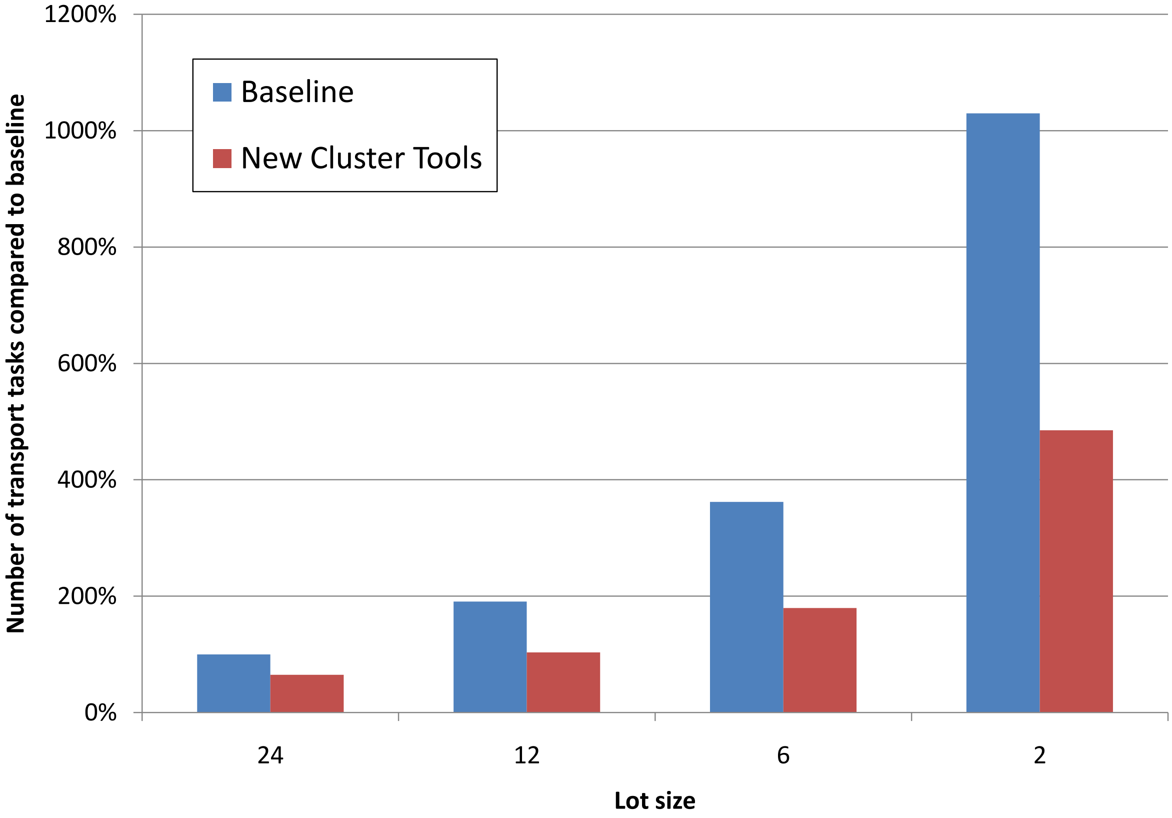


Figure 9.11: Transport demand of scenarios with baseline toolset and toolset consisting out of new linear cluster tools for different lot sizes

At six wafer lot size the transport demand is still less than twice that of our baseline scenario at 24 wafer lot size. This still seems achievable considering that the window of opportunity for transports to a transfer buffer is not as small as for deliveries to a tool load port. Additionally it has to be taken account that some inter-equipment transports can be performed by the local transport system. The transport demand of the two wafer lot size scenario still seems beyond the capablity of a vehicle-based system, though.

Figure 9.12 shows the cycle time results of the simulation runs of the model with the new modified linear cluster tools. Compared to the *Single 24 scenario* the cycle time is reduced very effectively even without lot size changes. Lot size reduction, in these scenarios refering to the standard transport size between the local material handling systems, achieves additional benefit. However, at least two wafer lot size does not deliver the benefits necessary to justify the tremendous transport demand and in many cases twelve waver lot size might already deliver sufficient cycle time performance.

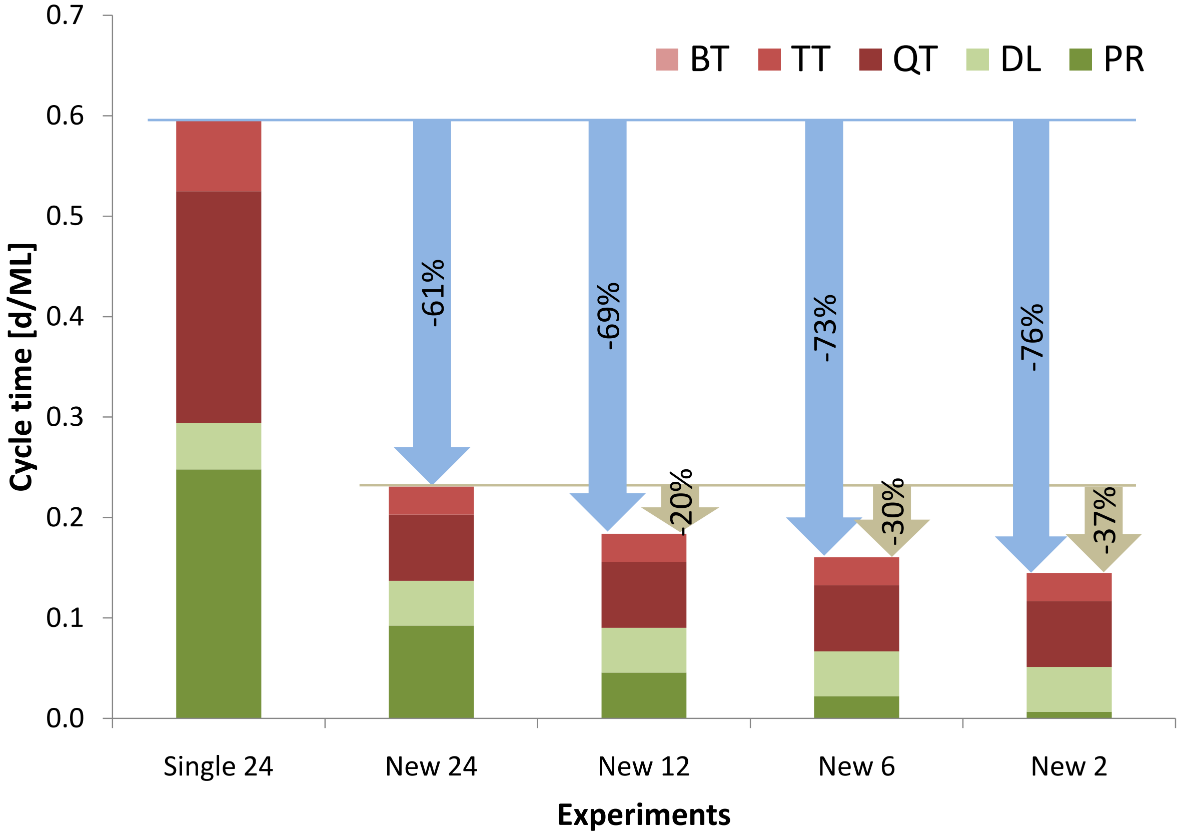


Figure 9.12: Comparison of new linear cluster tool scenario at different lot size and to *Single 24* scenario

The difference in cycle time between the different lot size lies in the process time only. Because of the flexible allocation of wafers to chambers used in the simulation model queueing time does not change. Within the process time most of the difference lies in PR as logical for lot size changes.

Figure 9.13 compares different equipment scenarios at 6 wafer lot size and the baseline scenario. The biggest benefit is provided by the scenario with the new linear cluster tools. In comparison with the small tools scenario we see, that the queueing time QT is very similar as expected, because we use the individual chambers with the same flexibility. In total the cycle time performance of the scenarion with the new modified cluster tools seems well-balanced with regard to the individual components. In other scenarios, e.g., the Single 6 scenario the cycle time reduction is unbalanced between queueing time and process time.

9.6.3 X-Factor Considerations

Table 9.1 lists the x-factors of the new cluster tools scenarios under consideration within this section at a fab loading of 92.5%. Although process time is reduced significantly, the x-factor is smaller than in the *normal* single wafer tools scenario. In addition x-factor does not increase as steep with lot size reduction

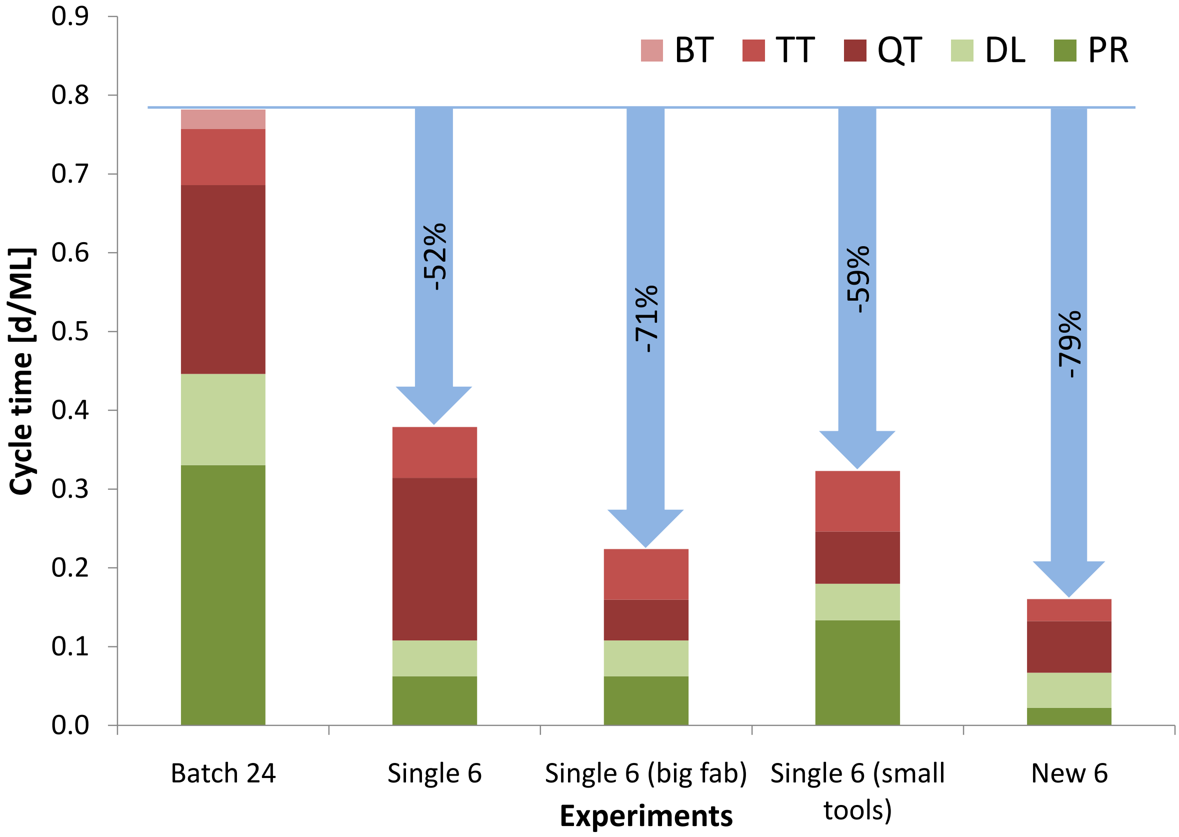


Figure 9.13: Comparison of equipment scenarios

underlining the strong queueing time reduction. Again cycle time performance comparison of fabs with these differences based on x-factor has its pitfalls.

Table 9.1: X-factor for scenarios at 92.5% fab loading

	X-factor							
	average				95-percentile			
Scenario / Lot size	24	12	6	2	24	12	6	2
Single	2.03	2.65	3.51	4.93	2.27	3.07	4.09	5.89
New Cluster	1.68	2.04	2.40	2.83	1.89	2.35	2.82	3.38

Apart from the average x-factor based on the average cycle time, we show the x-factor based on the 95-percentile cycle time as well in order to give an impression of the cycle time variance. Cycle time variance is lower in the *new cluster*-scenarios. This applies both at the baseline lot size level of 24 wafers and at the reduced lot sizes confirming the effectiveness of this new approach with respect to variability reduction.

9.7 Conclusions

We summarize the key results of this section with the characterisitic curves for the new linear cluster tool scenario at 24 wafer lot size and the Single 24 scenario displayed in Figure 9.14 (for the benefit of clarity, we do without characteristic curves of reduced lot size here.). Both the theoretical limits and the relative shape of the CT-curves are reduced very effectively confirming the potential of the lot-streaming approach. It is possible to outperform the cycle time reduction achieved by conventionally reduced lot sizes without running into the same operational problems that make their realization look difficult.

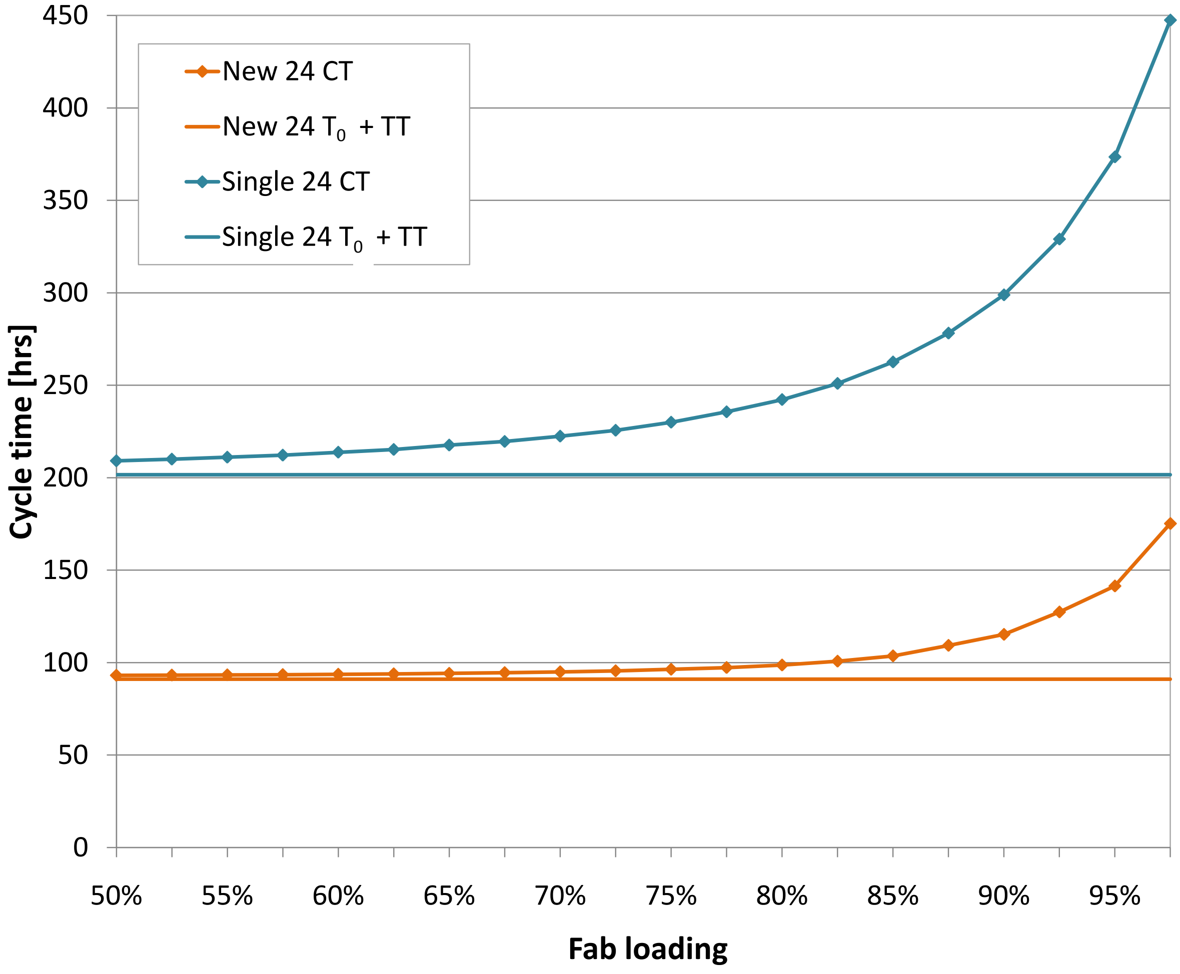


Figure 9.14: Characteristic curve of the new linear cluster tool scenario at 24 wafer lot size compared to *Single 24* scenario

10 Conclusions

In this dissertation, we analyzed different approaches for production system design changes that enable shorter cycle times in semiconductor manufacturing and therefore can be substantial elements for a modern manufacturing strategy. The changes under analysis were

- the replacement of batch tools with mini-batch or single-wafer tools,
- the reduction of standard lot size,
- equipment or fab size scaling, and
- new linear cluster tools enabling lot streaming.

The replacement of batch tools with mini-batch or single wafer tools shows persuasive cycle time advantages especially for fabs with a diverse product portfolio that could be even higher in case of different technologies in the same fab. A decrease in cost of ownership of the respective mini-batch or single wafer tools to a lower level increasing competitiveness compared to batch tools from a cost perspective is a prerequisite for wide adoption, though. Provided this can be achieved it seems likely that in the foreseeable future semiconductor manufacturing is performed without batch tools. For fabs with a very cycle time sensitive business model an earlier adoption makes sense. The replacement of batch tools in the BeoL only is an interesting intermediate option especially for companies with a business model that is more sensitive to BeoL cycle time.

The reduction in standard lot size also leads to a reduction in cycle time. This approach, however, lacks persuasiveness if lot size is reduced to very small values, because variability impacts jeopardizes the cycle time effectiveness of the lot size reduction approach. Considering the toolset, the cycle time reduction achieved with smaller lot size is higher if batch tools are replaced with mini-batch tools, and again higher if the replacement occurs with single wafer tools. The hybrid configuration with batch tools in the FeoL only and smaller lot size in the BeoL represents an interesting intermediate option. We also show that lot size reduction leads to less or no cycle time reduction at workstations with setup times, but this does not compromise the approach if setup occurences do not prevail. However, companies that consider lot size reduction have to evaluate how the extent of setups inherent to their product spectrum, tools and process capabilities influences the possible cycle time gain and seek opportunities to avoid setups efficiently.

In our scaling scenarios we targeted a reduction of the negative impact with scaling. We considered increasing the number of individual tools by creating smaller tools or by increasing the fab size. Both approaches lead to reduced queueing time and a more effective reduction in queueing time achieved by smaller lot sizes. In the experiment with smaller tools the cost for this improvement is a significantly increased process time leading to an unfavorable total cycle time. Therefore this experiment can only serve as guidance for future equipment design trying to achieve the queueing time benefit without the disadvantage of increased processing time. Fab scaling does not lead to such a disadvantage, therefore the queueing time advantage transfers to a cycle time advantage. Hence, fabs of bigger scale can achieve a more significant cycle time reduction by smaller lot sizes.

Based on the insights of previous experiments and cluster tool design suggestions in the literature we created a new cluster tool type in our lot streaming scenarios addressing the disadvantages of current designs. These scenarios show persuasive results. Both theoretical limits and queueing time are reduced very effectively by this approach using a linear cluster tool design with additional access points integrating several operations into one cluster tool.

Apart from these direct assessments of production system changes, we have also produced indirect results that provide beneficial insights and enable simplified adaption of the analysis.

Our cycle time assessment by component enables detailed analysis of mechanisms and cycle time performance of the discussed production system changes. This assessment by component also provides for an easy estimation method for other fabs by assuming the same relative evolvement by component. Provided current cycle time performance by component is available or can be gathered, the cycle time gain achieved by the production system changes can be estimated for a different fab profile with a specific share of batch tools, specific batching context, and specific processing times.

10.1 Development Perspectives for Process and Automation Equipment Industry

Not all of our changes are possible with currently available equipments and material handling systems. Some development work is necessary to strengthen or create the possibility to put the changes under consideration into practice. Development prospects that support or enable the production system changes are listed in the following by the specific production system changes.

- Replacement of batch equipments
  - Development of mini-batch and single-wafer tool alternatives with more competitive cost of ownership.
- Lot size reduction
  - AMHS systems supporting small lot size of 12 or 6 wafers and correspondingly higher move rates (Development of an AMHS system for an even smaller lot size, as, e.g., 2 wafers does not seem promising.).
  - Material handling at the EFEM allowing for the removal of empty carriers and their nearby storage.
  - Equipment availability improvements and reduction in the variability of availability.
  - Elimination or reduction of setup times.
- Lot streaming[1]
  - Disintegration of process and automation part of cluster tool equipment enabling new equipment design incorporating automation part and process chambers from different suppliers (This includes the definition and application of common standards for the physical and the control interface.).
  - Development of automation part for linear cluster tool with additional access points.
  - AMHS systems supporting on-time delivery of wafers to linear cluster tools and the intermittent transport of wafers with flexible transport size.
  - Manufacturing Execution System (MES) and process control system development to support the distinct increase of flexibility in material flow.

10.2 Future Areas of Analysis

The development perspectives discussed form the basis for further research to support the development and the realization of the production system changes under consideration. The following subject areas need further examination.

In our analysis we considered FIFO-dispatching only in order to avoid side effects due to dispatch rule

[1] We have no bullets for the *scaling*-analysis as the *small tools*-scenarios show no substantial benefit and the *big fab*-scenario does not show substantially different development needs than previous scenarios with lot size reduction.

parameters such as due dates (see Section 5.2). Further analysis should show how our results compare to the outcome of simulation scenarios considering different dispatching rules. Specific attention should be given to the development and assessment of new dispatching approaches that can limit the negative impact of variability on the effectiveness of cycle time reduction achieved with smaller lot size.

Scheduling could play an important role in future WIP Control picking up where improvements by dispatching are no longer possible. When determining the lot sequence at a workstation, scheduling takes additional information on the equipment and lot status at other workstations into account, therefore it has the potential to better cope with the effects of variability. Hence, it seems likely that scheduling approaches may lead to an increased cycle time reduction effectiveness of the measures discussed. Scheduling might also enable more efficient material transport, because transport needs are known in advance which enables optimization in the order and assignment of AMHS tasks to vehicles.

We listed development perspectives for new AMHS systems in the previous Section 10.1. These developments have to be analyzed with simulation or other means in order to identify the design changes that are able to address the increased performance needs of the different scenarios.

The new equipment design approach which we developed in this dissertation, needs additional analysis. The most suitable size for the linear cluster tools as well as the corresponding number of access points has to be determined. And a simulation model comprising the complete AMHS has to confirm reasonable operation which we showed at an example comprising a local AMHS with two linear cluster tools only.

In our analysis we have quantified the operational benefits, but only discussed the business case. Determining the financial benefits of the cycle time reductions we showed as well as putting a specific price tag on each improvement scenario forms the basis for a reasonable decision to put the most beneficial approach into practice. This analysis will depend significantly on the company's business model, i.e., whether it is a foundry, a logic maker, or a memory device maker.

Finally, the approach will have to stand against reality. At some point verification in a pilot line will be necessary to show that the benefits are sustainable when transfering the approach from models to the real world.

Bibliography

[AI05] Naciye Akca and Andre Ilas. Produktionsstrategien: Überblick und systematisierung. Arbeitsbericht Nr. 28; Institut für Produktion und Industrielles Informationsmanagement; Universität Duisburg-Essen, 2005. available at `http://www.pim.uni-essen.de/fileadmin/Publikationen/Arbeitsbericht_Nr._28.pdf`, accessed 06/13/2009.

[Bak93] Kenneth R. Baker. A comparative study of lot streaming procedures. *OMEGA International Journal of Management Science*, 21(5):561–566, 1993.

[BCFR97] Steven Brown, Frank Chance, John W. Fowler, and Jennifer Robinson. A centralized approach to factory simulation. *Future Fab International*, 1997.

[BCH79] Onno J. Boxma, J. W. Cohen, and N. Huffels. Approximations of the mean waiting time in an m/g/s queueing system. *Operations Research*, 27:1115–1127, 1979.

[Bec08] Matthias Becker. *Simulation von Clustertools in der Halbleiterfertigung: Entwicklung eines dynamischen Clustertool-Simulators*. Vdm Verlag Dr. Müller, 2008.

[Ben08] Jozsef Daniel Benke. *Development and Assessment of Wafer Transport Sequencing Rules for Clustertools: Development of Sequencing Rules for Clustertools by Means of Dynamic Tool Simulation*. Vdm Verlag Dr. Müller, 2008.

[BINN00] Jerry Banks, John S. Carson II, Barry L. Nelson, and David M. Nicol. *Discrete-Event System Simulation*. Prentice Hall, third edition, 2000.

[BK04] Thorsten Blecker and Bernd Kaluza. *Entwicklungen im Produktionsmanagement*, chapter Produktionsstrategien - ein vernachlässigtes Forschungsgebiet?, pages 4–27. A. Braßler and Hans Corsten, München, 2004.

[Ble03] Thorsten Blecker. *Moderne Produktionskonzepte für Güter- und Dienstleistungsproduktion*, chapter Entwurf eines auf Internet-Technologien basierenden Produktionskonzepts, pages 273–316. H. Wildemann, München, 2003.

[BML$^+$03] Olivier Bonnin, Dominique Mercier, Didier Levy, Martin Henry, Isabelle Pouilloux, and Eric Mastromatteo. Single-wafer/mini-batch approach for fast cycle time in advanced 300-mm fab. *IEEE Transactions on Semiconductor Manufacturing*, 16(2):111–120, May 2003.

[BW08] Eddy Bass and Robert Wright. Modelling semiconductor factories for equipment and cycle time reduction opportunities. *Future Fab International*, 24:50–55, 1 2008.

[CA00] Elizabeth Campbell and Jim Ammenheuser. 300 mm factory layout and material handling modeling: Phase ii report. ISMI Technical Publication, 2000. available at `http://ismi.sematech.org/docubase/document/3848beng.pdf`, accessed 06/13/2009.

[dB08] Encyclopædia Britannica. production system. Encyclopædia Britannica Online, 2008. available at `http://www.britannica.com/EBchecked/topic/478032/production-system`, accessed 06/13/2009.

[ES99] Walter Eversheim and Günther Schuh. *Produktion und Management. Betriebshütte*. Springer- Verlag, Berlin, 1999.

[FP00] Jackie Ferell and Margaret Pratt. I300i factory guidelines: Version 5.0. Technology Transfer Document, International Sematech, 2000. available at `http://www.sematech.org/docubase/document/3311geng.pdf`, accessed 06/13/2009.

[Gar06] Gartner. Gartner says worldwide semiconductor memory revenue grows 22 percent in 2006; market to experience 10 percent growth in 2007. Press Release, 2006. available at `http://www.gartner.com/it/page.jsp?id=499334`, accessed 06/13/2009.

[GBL$^+$07] Nirmal Govind, E. W. Bullock, He Linling, B. Iyer, M. Krishna, and C. S. Lockwood. Operations management in automated semiconductor manufacturing with integrated targeting, near real-time scheduling, and dispatching. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 363–370, 2007.

[GF86] Eliyahu M. Goldratt and Robert E. Fox. *The race*. North River Press, 1986.

[GH98] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc., 1998.

[Glü06] Detlev Glüer. Improving amhs performance with undertrack storage systems. Presentation at Innovationsforum Dresden, 2006.

[Gre07] Arieh Greenberg. Next generation factory challenges and opportunities. Keynote Presentation at Advanced Semiconductor Manufacturing Conference, 2007.

[Gro03] Strategieinstitut Boston Consulting Group. *Clausewitz: Strategie denken*. Deutscher Taschenbuch Verlag, München, 2003.

[Gro07a] Doug Grose. Call to action on the next generation factory. Keynote Presentation at 4th ISMI Symposium on Manufacturing Effectiveness, 2007.

[Gro07b] Doug Grose. Collaborating for efficiency. Presentation at SEMICON West, 2007.

[GWS07] Gero Grau, Jörg Weigang, and Kilian Schmidt. Improving dispatch rules for cascading tools. In *IEEE Conference on Automation Science and Engineering*, pages 261–264, 2007.

[HF07] Myoungsoo Ham and John W. Fowler. Balanced machine workload dispatching scheme for wafer fab. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 390–395, 2007.

[HP94] R.H. Hayes and G.P. Pisano. Beyond world class: The new manufacturing strategy. *Harvard Business Review*, 72:77–86, 1994.

[HS00] Walace J. Hopp and Mark L. Spearman. *Factory Physics: The Foundations of Manufacturing Management*. Irwin McGraw-Hill, Singapore, 2000.

[HVG06] Stefan Hempel, Steffen Volt, and Wolfram Grundke. Lithography cell productivity improvement approaches. *Semiconductor Fabtech*, 32:83–89, 2006.

[iC07] iSuppli Corporation. Price plunge to slow dram revenue growth in 2007. Press Release, 2007.

[Ign08] James Ignizio. What are the alternatives to 450 mm wafers. *Future Fab International*, 25:60–64, 4 2008.

[KEPS06] Sushant Koshti, Eric Englhardt, Amit Puri, and Vinay Shah. E87+ (e87 plus) -an extension to the semi e87 specification for carrier management for improving tool productivity and enabling lot size reduction. In *ISMI International Symposium on Semiconductor Manufacturing*, 2006.

[KJ06] Gregor Kübarth and Gerd Jungmann. Alternative buffer solution for 300mm waferlots.

Presentation at Innovationsforum Dresden, 2006.

[Law07] Averill M. Law. *Simulation Modeling & Analysis*. McGraw Hill, fourth edition, 2007.

[LD05] Todd LeBaron and Joerg Domaschke. Optimizing robot algorithms with simulation. In *Winter Simulation Conference*, pages 2211–2217, 2005.

[LH00] Robert C. Leachman and David A. Hodges. Competitive semiconductor manufacturing. Presentation, 2000. Competitive Semiconductor Manufacturing Program Update.

[Liu05] Mark Liu. The advanced foundry in the consumer electronics era. Keynote Presentation at 2nd ISMI Symposium on Manufacturing Effectiveness, 2005.

[LKL02] Robert C. Leachman, Jeenyoung Kang, and Vincent Lin. Slim: Short cycle time and low inventory in manufacturing at samsung electronics. *interfaces*, 32(1):61–77, 2002.

[MST00] Peter Milling, Uwe Schellenbach, and Jörn-Henrik Thun. The role of speed in manufacturing. *First World Conference on Production and Operations Management POM Sevilla 2000*, 2000.

[O'H07] Michael O'Halloran. Future fab trends. In *ISMI Symposium on Manufacturing Effectiveness*, 2007.

[Osb07] Mark Osbourne. 300mm fab trends. In *ISMI Symposium on Manufacturing Effectiveness*, 2007.

[Por96] Michael Porter. What is strategy. *Harvard Business Review*, 74:61–78, 1996.

[PP05] Jeffrey S. Pettinato and Devadas Pillai. Technology decisions to minimize 450mm wafer size transition risk. *IEEE Transactions on Semiconductor Manufacturing*, 18(4), November 2005.

[RGHT07] Jan Rothe, Ralf Georgi, Alfred Honold, and Eberhard Teichmann. Optimized performance of automated material delivery by improved exception handling. In *International Symposium on Semiconductor Manufacturing*, pages 1–4, October 2007.

[Rin07] Robert Ringel. Simulation based scheduling - the fab30 furnace scheduler. Presentation at 9th workshop „Simulation und Leistungsbewertung von Fertigungssystemen“ at University of Technology Dresden, 2007.

[Ros99a] Oliver Rose. Conload - a new lot release rule for semiconductor wafer fabs. In *Winter Simulation Conference*, pages 850–855, 1999.

[Ros99b] Oliver Rose. Estimation of the cycle time distribution of a wafer fab by a simple simulation model. In *SMOMS*, 1999.

[Ros01a] Oliver Rose. Conwip-like lot release for a wafer fabrication facility with dynamic load changes. In *SMOMS '01 (ASTC '01)*, pages 41–46, 2001.

[Ros01b] Oliver Rose. The shortest processing time first (sptf) dispatch rule and some variants in semiconductor manufacturing. In *Winter Simulation Conference*, pages 1220–1224, 2001.

[Ros03] Oliver Rose. Comparison of due-date oriented dispatch rules in semiconductor manufacturing. In *Industrial Engineering Research Conference*, 2003.

[Ros04] Oliver Rose. Simulationstechnik zur systemanalyse. Vorlesung, 2004.

[Ros06] Oliver Rose. Economy of scale effects for large wafer fabs. In *Winter Simulation Conference*, pages 1817–1820, 2006.

[RPQ05] Kristin Rust, Nipa Patel, and Peng Qu. Modeling practical methods of material storage in a 300 mm fab. In *IEEE International Symposium on Semiconductor Manufacturing*, pages 25–28, 2005.

[Sch06] Kilian Schmidt. Optimizing tool internal wafer sequencing to improve cluster tool throughput. In *ISMI Symposium on Manufacturing Effectiveness*, 2006.

[Sch07] Kilian Schmidt. Improving priority lot cycle time. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 117–121, 2007.

[Sem08] Sematech. Definitions. Website, 2008. available at `http://ismi.sematech.org/publications/dictionary/t_to_th.htm`, accessed 06/13/2009.

[SIA01] Semiconductor Industry Association SIA. The international technology roadmap for semiconductors - factory integration - material handling backup section. Presentation, 2001. available at `http://www.itrs.net/Links/2001ITRS/Links/Factory/Material\Handling\Systems\Backup.ppt`, accessed 06/13/2009.

[SIA07a] Semiconductor Industry Association SIA. The international technology roadmap for semiconductors - executive summary (edition 2007). ITRS release, 2007. available at `http://www.itrs.net/Links/2007ITRS/Home2007.htm`, accessed 06/13/2009.

[SIA07b] Semiconductor Industry Association SIA. The international technology roadmap for semiconductors - factory integration (edition 2007). ITRS release, 2007. available at `http://www.itrs.net/Links/2007ITRS/Home2007.htm`, accessed 06/13/2009.

[Ski84] Wickham Skinner. Operations technology: Blind spot in strategic management. *Interfaces*, 14:116–125, 1984.

[Ski86] Wickham Skinner. The productivity paradox. *Harvard Business Review*, 64:55–59, 1986.

[Spl07] Mike Splinter. The future depends on an agile fab. *Semi Quarterly Report to Members Winter 2007*, 2007.

[SR07a] Kilian Schmidt and Oliver Rose. Development and simulation assessment of semiconductor fab architectures for fast cycle times. In *Doktorandenforum der SimVis*, 2007.

[SR07b] Kilian Schmidt and Oliver Rose. Queue time and x-factor characteristics for semiconductor manufacturing with small lot sizes. In *IEEE Conference on Automation Science and Engineering*, pages 1069–1074, 2007.

[SR08a] Kilian Schmidt and Oliver Rose. Development and simulation assessment of semiconductor fab architecture enhancements for fast cycle times. In *Doktorandenforum der SimVis*, 2008.

[SR08b] Kilian Schmidt and Oliver Rose. Simulation analysis of semiconductor manufacturing with small lot size and batch tool replacements. In *Winter Simulation Conference*, 2008.

[SRW06] Kilian Schmidt, Oliver Rose, and Joerg Weigang. Modeling semiconductor tools for small lotsize fab simulations. In *Winter Simulation Conference*, pages 1811–1816, 2006.

[TB93] Dan Trietsch and Kenneth R. Baker. Basic techniques for lot streaming. *Operations Research*, 41(6):1065–1076, 1993.

[UR07] Robert Unbehaun and Oliver Rose. Predicting cluster tool behavior with slow down factors. In *Winter Simulation Conference*, pages 1755–1760, 2007.

[vdM07] Peter van der Meulen. Linear semiconductor manufacturing logistics and the impact on cycle time. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 111–116, 2007.

[vdMRPK06] Peter van der Meulen, Raymond Ritter, Patrick Pannese, and Christopher Kiley. An advanced equipment automation architecture for ultraclean manufacturing of 450 mm wafers. In *ISMI Symposium on Manufacturing Effectiveness*, 2006.

[vZ04] Peter van Zant. *Microchip Fabrication*. McGraw Hill, 5 edition, 2004.

[WEB+03] Ronald A. Weimer, Denise Marra Eppich, Kevin L. Beaman, D. Carl Powell, and Fernando González. Contrasting single-wafer and batch processing for memory devices. *IEEE Transactions on Semiconductor Manufacturing*, 16(2):138–146, May 2003.

[Whi93] Ward Whitt. Approximating the gi/g/m queue. *Production and Operations Management 2*, 2:114–161, 1993.

[Woo96] Samuel C. Wood. Simple performance models for integrated processing tools. *IEEE Transactions on Semiconductor Manufacturing*, 9(3):320–328, August 1996.

[Woo97] Samuel C. Wood. Cost and cycle time performance of fabs based on integrated single-wafer processing. *IEEE Transactions on Semiconductor Manufacturing*, 10(1), February 1997.

[Woo00] Samuel C. Wood. *Handbook of Semiconductor Manufacturing Technology*, chapter Factory Modeling, pages 1103–1122. 2000.

[WWK03] Inc. Wright Williams & Kelly. *User Manual Factory Explorer 2.8*, 2003.

[WWK+04] Takayuki Wakabayashi, Shinichi Watanabe, Yoshiaka Kobayashi, Tsutomu Okabe, and Atsuyoshi Koike. High-speed amhs and its operation method for 300mm qtat fab. *IEEE Transactions on Semiconductor Manufacturing*, 17(3):25–28, 2004.

[Zah88] Erich Zahn. *Handbuch strategischer Führung*, chapter Produktionsstrategie, pages 515–542. H.A. Henzler, Wiesbaden, 1988.

[Zah94] Erich Zahn. *Handbuch Produktionsmanagement*, chapter Produktion als Wettbewerbsfaktor, pages 241–258. Hans Corsten, Wiesbaden, 1994.

[Zäp89] Günther Zäpfel. *Strategisches Produktionsmanagement*. de Gruyter, Berlin, New York, 1989.

[ZRS+07] Olaf Zimmerhackl, Jan Rothe, Kilian Schmidt, Les Marshall, and Alfred Honold. The effects of small lot manufacturing on amhs operation and equipment front-end design. In *International Symposium on Semiconductor Manufacturing*, pages 185–189, October 2007.

[ZWBP08] Emrah Zarifoglu, Robert Wright, Chad Bubela, and Joey Preece. Modelling semiconductor factories for equipment and cycle time reduction opportunities, part ii. *Future Fab International*, 25:54–59, 4 2008.

A Route in Baseline Model with Performance Characteristics

Operation name	Workstation	PR [hrs]		DL [hrs]	Batch	Skip rate
		per wafer	per batch	per lot	ID	per lot
1_Clean_Pre_OxAn	W_Wet_Bench		0.333	0.500	7_1	
2_Oxation_Sac	D_Furn_FastRmp		2.217	0.667	1_1	
3_Meas_Film	M_Meas_Film	0.167		0.008		70%
4_LPCVD_Nitre	D_RTP_Nitr	0.058		0.075		
5_Meas_Film	M_Meas_Film	0.167		0.008		70%
6_Expose_AA	L_Litho_DUV	0.013		0.488		
7_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
8_Meas_CD	M_Meas_CD	0.167		0.008		70%
9_Etch_AA	E_Dry_Etch_A	0.040		0.120		
10_Plasma_Strip	E_Dry_Strip	0.013		0.021		
11_Clean_O3	W_VP_HF_Bench		1.250	0.250	6	
12_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
13_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
14_Meas_CD	M_Meas_CD	0.167		0.008		70%
15_Clean_Pre_OxAn	W_Wet_Bench		0.333	0.500	7_1	
16_Oxation	D_Furn_FastRmp		2.217	0.667	1_2	
17_Meas_Film	M_Meas_Film	0.167		0.008		70%
18_Oxe_STI	F_CVD_Ins	0.020		0.038		
19_Meas_Film	M_Meas_Film	0.167		0.008		70%
20_Densification	D_Furn_FastRmp		2.217	0.667	1_3	
21_CMP_AA	P_CMP_Ins	0.028		0.138		
22_Wet_Strip	W_Wet_Bench		0.333	0.500	7_3	
23_Dry_Strip	E_Dry_Strip	0.033		0.050		
24_Clean_Pre_OxAn	W_Wet_Bench		0.333	0.500	7_1	
25_Oxation_Sac	D_Furn_FastRmp		2.217	0.667	1_4	
26_Meas_Film	M_Meas_Film	0.167		0.008		70%
27_Expose_Implant	L_Litho_I	0.011		0.406		
28_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
29_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
30_Meas_CD	M_Meas_CD	0.167		0.008		70%
31_Implant	I_Implant_HiE	0.008		0.025		
32_Implant	I_Implant_HiE	0.008		0.025		
33_Implant	I_Implant_LoE	0.011		0.031		
34_Plasma_Strip	E_Dry_Strip	0.013		0.021		
35_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
36_Expose_Implant	L_Litho_I	0.011		0.406		
37_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
38_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
39_Meas_CD	M_Meas_CD	0.167		0.008		70%

40_Implant	I_Implant_HiE	0.008		0.025		
41_Implant	I_Implant_LoE	0.011		0.031		
42_Implant	I_Implant_LoE	0.011		0.031		
43_Plasma_Strip	E_Dry_Strip	0.013		0.021		
44_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
45_Wet_Strip	W_Wet_Bench		0.333	0.500	7_3	
46_Clean_O3	W_VP_HF_Bench		1.250	0.250	6	
47_Oxation_Gate	D_RTP_OxAn	0.033		0.058		
48_Meas_Film	M_Meas_Film	0.167		0.008		70%
49_LPCVD_Poly	D_Furn_Poly		5.000	0.667	4	
50_Meas_Film	M_Meas_Film	0.167		0.008		70%
51_Expose_Gate	L_Litho_DUV	0.013		0.488		
52_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
53_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
54_Meas_CD	M_Meas_CD	0.167		0.008		70%
55_Etch_Gate	E_Dry_Etch_Gate	0.040		0.200		
56_Meas_Film	M_Meas_Film	0.167		0.008		70%
57_Inspect_Visual	M_Insp_Visual	0.250		0.008		70%
58_Plasma_Strip	E_Dry_Strip	0.013		0.021		
59_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
60_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
61_Meas_CD	M_Meas_CD	0.167		0.008		70%
62_Clean_Pre_OxAn	W_Wet_Bench		0.333	0.500	7_1	
63_Oxation	D_RTP_OxAn	0.020		0.045		
64_Meas_Film	M_Meas_Film	0.167		0.008		70%
65_Expose_Implant	L_Litho_I	0.011		0.406		
66_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
67_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
68_Meas_CD	M_Meas_CD	0.167		0.008		70%
69_Implant	I_Implant_LoE	0.011		0.031		
70_Plasma_Strip	E_Dry_Strip	0.013		0.021		
71_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
72_Expose_Implant	L_Litho_I	0.011		0.406		
73_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
74_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
75_Meas_CD	M_Meas_CD	0.167		0.008		70%
76_Implant	I_Implant_LoE	0.011		0.031		
77_Plasma_Strip	E_Dry_Strip	0.013		0.021		
78_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
79_Clean_Pre_OxAn	W_Wet_Bench		0.333	0.500	7_1	
80_LPCVD_TEOS	D_Furn_TEOS		5.000	0.667	5	
81_Meas_Film	M_Meas_Film	0.167		0.008		70%
82_Etch_Spacer	E_Dry_Etch_Spacer	0.033		0.067		
83_Meas_Film	M_Meas_Film	0.167		0.008		70%
84_Clean_Pre_OxAn	W_Wet_Bench		0.333	0.500	7_1	
85_RTP_Anneal/Ox	D_RTP_OxAn	0.033		0.058		
86_Meas_Film	M_Meas_Film	0.167		0.008		70%
87_Expose_Implant	L_Litho_I	0.011		0.406		
88_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%

89_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
90_Meas_CD	M_Meas_CD	0.167		0.008		70%
91_Implant	I_Implant_LoE	0.011		0.031		
92_Plasma_Strip	E_Dry_Strip	0.013		0.021		
93_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
94_Expose_Implant	L_Litho_I	0.011		0.406		
95_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
96_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
97_Meas_CD	M_Meas_CD	0.167		0.008		70%
98_Implant	I_Implant_LoE	0.011		0.031		
99_Plasma_Strip	E_Dry_Strip	0.013		0.021		
100_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
101_Clean_Pre_OxAn	W_Wet_Bench		0.333	0.500	7_1	
102_RTP_Anneal	D_RTP_OxAn	0.020		0.045		
103_Dry_Strip	E_Dry_Etch_Spacer	0.033		0.067		
104_Meas_Film	M_Meas_Film	0.167		0.008		70%
105_Plasma_Strip	E_Dry_Strip_Ins	0.013		0.021		
106_Clean_Post_Strip	W_Wet_Bench		0.333	0.500	7_2	
107_Clean_O3	W_VP_HF_Bench		1.250	0.250	6	
108_PVD_Ti/Co	F_PVD_Met_TiCo	0.017		0.183		
109_Inspect_Visual	M_Insp_Visual	0.250		0.008		70%
110_RTP_Silice	D_RTP_OxAn_Con	0.020		0.045		
111_Wet_Strip Ti/Co	W_Wet_Bench_Ins		0.333	0.500	8_1	
112_RTP_Anneal	D_RTP_OxAn_Con	0.020		0.045		
113_CVD_Nitr/TEOS	F_CVD_Ins_Con	0.020		0.038		
114_Meas_Film	M_Meas_Film	0.167		0.008		70%
115_CVD_BPSG	F_CVD_Ins_Con	0.020		0.038		
116_Meas_Film	M_Meas_Film	0.167		0.008		70%
117_Densification	D_Furn_OxAn_Ins		2.217	0.667	3_1	
118_CMP_BPSG	P_CMP_Ins_Con	0.028		0.138		
119_Meas_Film	M_Meas_Film	0.167		0.008		70%
120_Inspect_PLY	M_Insp_Def	0.250		0.008		70%
121_APCVD_Ox	F_CVD_Ins_Con	0.022		0.045		
122_Expose_Contact	L_Litho_DUV	0.013		0.488		
123_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
124_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
125_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
126_Etch_Contact	E_Dry_Etch_Con	0.040		0.120		
127_Plasma_Strip	E_Dry_Strip_Ins	0.013		0.021		
128_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
129_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
130_CVD_Ti/TiN	F_CVD_Met_Con	0.017		0.050		
131_CVD_W	F_CVD_MetW	0.014		0.036		
132_CMP_W	P_CMP_Met	0.018		0.115		
133_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
134_Coat_LowK	L_CoatOnly	0.008		0.075		
135_Cure	D_Furn_Nitr		2.000	0.667	2_2	
136_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
137_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%

138_CVD_Oxe	F_CVD_Ins_Oxe	0.010		0.040		
139_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
140_Expose_Line	L_Litho_DUV	0.013		0.488		
141_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
142_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
143_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
144_Etch_LowK	E_Dry_Etch_LowK	0.040		0.200		
145_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
146_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
147_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
148_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
149_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
150_PVD_Barrier	F_PVD_Met	0.022		0.245		
151_CVD_Cu	F_CVD_Met_Cu	0.022		0.045		
152_Electro_Cu	P_Electroplate	0.020		0.180		
153_CMP_Cu	P_CMP_Met	0.028		0.138		
154_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
155_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
156_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
157_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
158_Test	T_Test_Wet	0.250		0.100		
159_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
160_CVD_Nitre	F_CVD_Ins_Nitre	0.010		0.040		
161_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
162_Coat_LowK	L_CoatOnly	0.008		0.075		
163_Cure	D_Furn_Nitr		2.000	0.667	2_2	
164_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
165_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
166_CVD_Oxe	F_CVD_Ins_Oxe	0.010		0.040		
167_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
168_Expose_Via	L_Litho_DUV	0.013		0.488		
169_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
170_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
171_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
172_Etch_Oxe	E_Dry_Etch_Oxe	0.017		0.100		
173_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
174_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
175_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
176_Expose_Line	L_Litho_DUV	0.013		0.488		
177_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
178_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
179_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
180_Etch_LowK	E_Dry_Etch_LowK	0.040		0.200		
181_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
182_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
183_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
184_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
185_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
186_PVD_Barrier	F_PVD_Met	0.022		0.245		

187_CVD_Cu	F_CVD_Met	0.022		0.045		
188_Electro_Cu	P_Electroplate	0.020		0.180		
189_CMP_Cu	P_CMP_Met	0.028		0.138		
190_Post CMP Clean	W_Wet_Brush_Bench		0.250	0.500	9	
191_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
192_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
193_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
194_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
195_CVD_Nitre	F_CVD_Ins_Nitre	0.010		0.040		
196_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
197_Coat_LowK	L_CoatOnly	0.008		0.075		
198_Cure	D_Furn_Nitr		2.000	0.667	2_2	
199_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
200_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
201_CVD_Oxe	F_CVD_Ins_Oxe	0.010		0.040		
202_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
203_Expose_Via	L_Litho_DUV	0.013		0.488		
204_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
205_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
206_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
207_Etch_Oxe	E_Dry_Etch_Oxe	0.017		0.100		
208_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
209_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
210_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
211_Expose_Line	L_Litho_DUV	0.013		0.488		
212_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
213_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
214_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
215_Etch_LowK	E_Dry_Etch_LowK	0.040		0.200		
216_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
217_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
218_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
219_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
220_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
221_PVD_Barrier	F_PVD_Met	0.022		0.245		
222_CVD_Cu	F_CVD_Met	0.022		0.045		
223_Electro_Cu	P_Electroplate	0.020		0.180		
224_CMP_Cu	P_CMP_Met	0.028		0.138		
225_Post CMP Clean	W_Wet_Brush_Bench		0.250	0.500	9	
226_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
227_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
228_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
229_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
230_CVD_Nitre	F_CVD_Ins_Nitre	0.010		0.040		
231_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
232_Coat_LowK	L_CoatOnly	0.008		0.075		
233_Cure	D_Furn_Nitr		2.000	0.667	2_2	
234_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
235_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%

236_CVD_Oxe	F_CVD_Ins_Oxe	0.010		0.040		
237_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
238_Expose_Via	L_Litho_DUV	0.013		0.488		
239_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
240_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
241_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
242_Etch_Oxe	E_Dry_Etch_Oxe	0.017		0.100		
243_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
244_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
245_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
246_Expose_Line	L_Litho_DUV	0.013		0.488		
247_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
248_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
249_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
250_Etch_LowK	E_Dry_Etch_LowK	0.040		0.200		
251_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
252_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
253_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
254_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
255_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
256_PVD_Barrier	F_PVD_Met	0.022		0.245		
257_CVD_Cu	F_CVD_Met	0.022		0.045		
258_Electro_Cu	P_Electroplate	0.020		0.180		
259_CMP_Cu	P_CMP_Met	0.028		0.138		
260_Post CMP Clean	W_Wet_Brush_Bench		0.250	0.500	9	
261_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
262_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
263_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
264_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
265_CVD_Nitre	F_CVD_Ins_Nitre	0.010		0.040		
266_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
267_Coat_LowK	L_CoatOnly	0.008		0.075		
268_Cure	D_Furn_Nitr		2.000	0.667	2_2	
269_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
270_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
271_CVD_Oxe	F_CVD_Ins_Oxe	0.010		0.040		
272_Expose_Via	L_Litho_DUV	0.013		0.488		
273_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
274_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
275_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
276_Etch_Oxe	E_Dry_Etch_Oxe	0.017		0.100		
277_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
278_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
279_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
280_Expose_Line	L_Litho_I	0.011		0.406		
281_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
282_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
283_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
284_Etch_LowK	E_Dry_Etch_LowK	0.040		0.200		

285_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
286_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
287_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
288_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
289_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
290_PVD_Barrier	F_PVD_Met	0.022		0.245		
291_CVD_Cu	F_CVD_Met	0.029		0.055		
292_Electro_Cu	P_Electroplate	0.020		0.180		
293_CMP_Cu	P_CMP_Met	0.028		0.138		
294_Post CMP Clean	W_Wet_Brush_Bench		0.250	0.500	9	
295_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
296_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
297_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
298_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
299_CVD_Nitre	F_CVD_Ins_Nitre	0.010		0.040		
300_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
301_Coat_LowK	L_CoatOnly	0.008		0.075		
302_Cure	D_Furn_Nitr		2.000	0.667	2_2	
303_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
304_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
305_CVD_Oxe	F_CVD_Ins_Oxe	0.010		0.040		
306_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
307_Expose_Via	L_Litho_I	0.011		0.406		
308_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
309_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
310_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
311_Etch_Oxe	E_Dry_Etch_Oxe	0.017		0.100		
312_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
313_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
314_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
315_Expose_Line	L_Litho_I	0.011		0.406		
316_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
317_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
318_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
319_Etch_LowK	E_Dry_Etch_LowK	0.065		0.267		
320_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
321_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
322_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
323_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
324_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
325_PVD_Barrier	F_PVD_Met	0.022		0.245		
326_CVD_Cu_80nm	F_CVD_Met	0.029		0.055		
327_Electro_Cu	P_Electroplate	0.020		0.180		
328_CMP_Cu	P_CMP_Met	0.028		0.138		
329_Post CMP Clean	W_Wet_Brush_Bench		0.250	0.500	9	
330_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
331_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
332_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
333_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	

334_CVD_Nitre	F_CVD_Ins_Nitre	0.010		0.040		
335_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
336_Coat_LowK	L_CoatOnly	0.008		0.075		
337_Cure	D_Furn_Nitr		2.000	0.667	2_2	
338_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
339_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
340_CVD_Oxe	F_CVD_Ins_Oxe	0.010		0.040		
341_Meas_Film	M_Meas_Film_Ins	0.167		0.008		70%
342_Expose_Via	L_Litho_I	0.011		0.406		
343_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
344_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
345_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
346_Etch_Oxe	E_Dry_Etch_Oxe	0.017		0.100		
347_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
348_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
349_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
350_Expose_Line	L_Litho_I	0.011		0.406		
351_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
352_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
353_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
354_Etch_LowK	E_Dry_Etch_LowK	0.065		0.267		
355_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
356_Inspect_Visual	M_Insp_Visual_Ins	0.250		0.008		70%
357_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
358_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
359_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
360_PVD_Barrier	F_PVD_Met	0.022		0.245		
361_CVD_Cu	F_CVD_Met	0.029		0.055		
362_Electro_Cu	P_Electroplate	0.020		0.180		
363_CMP_Cu	P_CMP_Met	0.028		0.138		
364_Post CMP Clean	W_Wet_Brush_Bench		0.250	0.500	9	
365_Meas_CD	M_Meas_CD_Ins	0.167		0.008		70%
366_Inspect_PLY	M_Insp_Def_Ins	0.250		0.008		70%
367_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
368_Test	T_Test_Wet	0.200		0.100		
369_Clean_Metal	W_Wet_Bench_Ins		0.333	0.500	8_3	
370_CVD_TEOS/Nitre	F_CVD_Teos	0.040		0.110		
371_Expose_Pad	L_Litho_I	0.011		0.406		
372_Meas_Overlay	M_Meas_Overlay	0.167		0.008		70%
373_Etch_PAD	E_Dry_Etch_Pad	0.025		0.125		
374_Plasma_Strip	E_Dry_Strip_Ins	0.013		0.021		
375_Clean_Post_Strip	W_Wet_Bench_Ins		0.333	0.500	8_2	
376_Anneal_Metal	D_Furn_OxAn_Ins		1.500	0.667	3_2	
377_Test	T_Test_Sort	0.500		0.100		

B Setup Specification in Setup Scenario

Workstation	Setup context	Avg. duration [hrs]	Max. tools per context
I_Implant_HiE	all operations	0.083	1
I_Implant_LoE	all operations	0.083	1
F_CVD_Ins_BackEnd	Nitride/Oxide	6	
E_Dry_Etch_BackEnd	Oxide/LowK/Pad	6	

Acknowledgments

I am indepted to my doctoral advisor Prof. Dr. Oliver Rose. He met my wish to pursue a Ph.D. under his supervision with enthusiasm and encouraged me to dare this challenge. During the course of my Ph.D. studies he supported me with sound guidance, profound discussions, and organizational advice necessary for successful execution. At several conferences I enjoyed the companionship and I am glad to say that I could not have wished for a better doctoral advisor.

I have great collegues at AMD in Dresden. Our collaboration and technical dialog contributed to my understanding and knowledge of the semiconductor production system and many discussions on various aspects of new production system approaches have been helpful in giving ideas a specific shape consistent with the manufacturing environment. Specifically I want to thank my manager Thomas Quarg for his confidence in my abilities and his support regarding my Ph.D. work, Gunnar Flach for introducing myself to the analysis and optimization of tools, Ken Wallers for initiating my participation in ISMI working groups regarding the next generation factory, and Lothar Mergili for his inspirational ideas on lean manufacturing in the semiconductor industry.

My parents have raised me with great care, passionate support, and last but not least, by example. I owe them much and continue to appreciate their advice.

This work would have been impossible without the loving support of my wife Frauke. Her encouragement and understanding helped me over times of stagnancy and technical challenges and she generously accepted the commitment of my leisure time to this work. This holds especially true as our son Gabriel was born during the course of my Ph.D studies and she had to carry the bulk of the first months' time and effort in taking care of him.