POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

An approximate analytical method for the performance evaluation of semiconductor front-end fabrication integrating photolithography inspection strategies

Tesi di Laurea Magistrale in
Mechanical Engineering - Ingegneria Meccanica

Author: **Riccardo Pomi**

Student ID: 10635286
Advisor: Prof. Tullio Antonio Maria Tolio
Co-advisors: Dr. Maria Chiara Magnanini, Prof. Dragan Djurdjanovic
Academic Year: 2022-23

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor, Professor Tullio Antonio Maria Tolio, for teaching me and instilling in me a passion for this research field. Thanks to him, I was able to pursue my dream and embark on what has been a life-changing experience, as well as an enlightening educational journey.

A special thank you also goes to my co-advisor, Dr. Maria Chiara Magnanini, who has supported me throughout the process, inspiring me and always being there with valuable advice and meetings at different times.

Words of gratitude are also due to Professor Dragan Djurdjanovic, who listened to me, inspired me, and helped me even in the most challenging moments of this journey. I hold his teachings in my heart, just like the passion with which he teaches.

At the University of Texas at Austin, I was welcomed as part of a family, where I encountered individuals who enriched me culturally with their diverse backgrounds and cultures.

I consider myself truly fortunate to have a family that supported me in all these decisions, even when it meant changing continents to pursue the choices that will shape my future.

A special thanks to all the people who have crossed my path, shaping its most intricate facets. Thank you for the passion you conveyed, for the hours dedicated to building such an ambitious project, and for giving me the opportunity to play my small part in the evolution of this process. I will carry all of this in my cultural and academic baggage.

This all-encompassing experience has made me a better person, enriched in every sense, and infinitely grateful for the people I have found around me.

Ringraziamenti

Vorrei innanzitutto ringraziare il mio relatore, Professor Tullio Antonio Maria Tolio per avermi insegnato e trasmesso la passione per questo ambito di ricerca; grazie a lui sono riuscito a seguire il mio sogno e partire per quella che è stata un'esperienza di vita, oltre che ad un illuminante percorso formativo.

Un ringraziamento speciale anche alla mia correlatrice, Dottoressa Maria Chiara Magnanini, che mi ha seguito in tutto il percorso, inspirandomi ma soprattutto essendoci sempre con i suoi preziosi consigli e con riunioni a fusorari diversi.

Parole di riconoscenza vanno anche al Professore Dragan Djurjanovic, che mi ha ascoltato, ispirato ed aiutato anche nei momenti più ardui di questo percorso, i suoi insegnamenti me li porto nel cuore come la passione con quale insegna.

Alla University of Texas ad Austin sono stato accolto come in una famiglia, dove ho trovato persone che mi hanno arricchito culturalmente coi loro background e culture differenti.

Mi ritengo davvero fortunato ad avere una famiglia che mi ha supportato in tutte queste decisioni, anche quando si trattava di cambiare continente per inseguire le scelte che detteranno il mio futuro.

Un grazie speciale a tutte le persone che si sono palesate nel mio percorso, dettandone le più articolate sfaccettature, grazie per la passione tramessa, per le ore dedicate a costruire un progetto così ambizioso e a darmi la possibilità di fare la mia piccola parte nell'evoluzione di questo processo.

Porterò tutto ciò nel mio bagaglio culturale e accademico. Questa esperienza a tutto tondo mi ha reso una persona migliore, arricchita in tutti i sensi e infinitamente grata dalle persone che mi son ritrovato attorno.

Abstract

In the present era, inspection strategies employed in multi-stage manufacturing systems have reached a remarkable level of complexity, particularly within the realm of semiconductor manufacturing, where sophisticated technologies are indispensable. However, a prevailing challenge lies in the absence of system-level perspectives that encompass all interconnected stages within a unified framework. This integration aims to bridge the gap and provide a comprehensive understanding of the manufacturing system as a whole.

This thesis has a primary focus on the integration of product-process-system models within a unified framework. To achieve this, a system model is developed using a stochastic modeling approach known as approximate analytical methods. This model takes into account a continuous flow of production, with instances of scrapping in-process and propagation of quality errors.

To validate the output of the model, a comparison is made with a discrete event simulator. This validation process ensures the accuracy and reliability of the model's results. Additionally, a real case study is conducted using a dataset provided by a semiconductor manufacturer based in Austin. This empirical analysis enables the optimization of overlay error measurements at the system level, considering both process control and overall system performance metrics.

The results obtained from the study illustrate the impact of reducing overlay measurement points on the wafer. While this reduction leads to an increase in effective throughput, it also adversely affects the quality of the final product. Moreover, the study reveals that scrapping in the process has downstream effects, causing stages to suffer from starvation. However, appropriate inspection strategies can mitigate these issues.

Keywords: Semiconductor fabrication, Quality strategy, Performance evaluation

Abstract in lingua italiana

Nell'era attuale, le strategie di ispezione impiegate nei sistemi di produzione multi-stadio hanno raggiunto un notevole livello di complessità, in particolare nel settore della produzione di semiconduttori, dove tecnologie sofisticate sono indispensabili. Tuttavia, una sfida prevalente risiede nell'assenza di prospettive a livello di sistema che comprendano tutti gli stadi interconnessi all'interno di un modello unificato.

Questa tesi ha come obiettivo principale l'integrazione di modelli di prodotto-processo-sistema all'interno di un sistema unificato. Per raggiungere questo obiettivo, viene sviluppato un modello di sistema utilizzando un approccio di modellazione stocastica noto come metodi approssimativi analitici. Questo modello tiene conto di un flusso continuo di produzione, con scarto in linea ed tiene conto della propagazione di errori di qualità.

Per convalidare l'output del modello, viene effettuato un confronto con un simulatore ad eventi discreti. Questo processo di convalida garantisce l'accuratezza dei risultati del modello. Inoltre, viene condotto uno studio di un caso reale utilizzando un set di dati fornito da un produttore di semiconduttori con sede ad Austin. Questa analisi empirica consente di ottimizzare le misurazioni degli errori a livello di sistema, considerando sia il controllo del processo che le metriche di prestazioni complessive del sistema.

I risultati ottenuti dallo studio illustrano l'impatto della riduzione dei punti di misurazione degli errori di sovrapposizione sul wafer. Sebbene questa riduzione porti a un aumento della produttività effettiva, influisce negativamente sulla qualità del prodotto finale. Inoltre, lo studio rivela che lo scarto in linea ha effetti a valle, causando problemi di flusso a fine linea. Tuttavia, adeguate strategie di ispezione possono mitigare questi problemi.

Parole chiave: Produzione di semiconduttori, Strategie per la qualità, Valutazione Performance

Contents

Acknowledgements **i**

Ringraziamenti **iii**

Abstract **v**

Abstract in lingua italiana **vii**

Contents **ix**

List of Figures **xiii**

List of Tables **xv**

1 Introduction **1**

1.1 The industrial context . . . 1

1.2 Motivations . . . 2

1.3 Objective . . . 4

1.4 Thesis outline . . . 5

2 Literature Review **7**

2.1 Integrated quality control in multi-stage production systems . . . 7

2.2 Product Level Model . . . 9

2.3 Model-based process control . . . 9

2.4 System-level models . . . 10

2.4.1 Performance evaluation of two-machines lines . . . 11

2.4.2 Performance evaluation of long lines . . . 12

2.4.3 Work in progress scrap . . . 13

3 Recap of Approximate Analytical Model **15**

3.1 Production Line Characterization . . . 15

3.2 Performance evaluation of two machines line . . . 18
3.3 Decomposition method with Continuous Flow . . . 23

4 Overlay Control **27**
4.1 Overlay error measurements . . . 27
4.2 Overlay error model . . . 28
4.3 Robust control and optimal down-selection of markers . . . 31
4.3.1 Robust Control of Overlay Errors . . . 32
4.3.2 Problem Formulation . . . 35
4.3.3 Problem Formulation for the Optimal Number of Measurement Markers . . . 37

5 Reference System **43**
5.1 Product Level . . . 43
5.2 Process Level . . . 44
5.3 System Level . . . 45
5.4 System Scheme . . . 46

6 Problem statement **49**
6.1 Research questions . . . 50
6.2 Outline of the method . . . 51
6.3 Innovative features . . . 51

7 Methodology **53**
7.1 Schematic system . . . 53
7.2 Assumptions . . . 54
7.3 Single-stage model . . . 55
7.3.1 MC: Remark for down states . . . 57
7.4 2M1B Line . . . 58
7.4.1 Quality propagation . . . 60
7.4.2 Starvation . . . 60
7.4.3 Blocking . . . 60
7.4.4 Evaluation . . . 60
7.5 Multi-stage evaluation model . . . 61
7.5.1 Two-level decomposition . . . 61
7.5.2 Integrated machine . . . 63
7.5.3 Building block . . . 64
7.5.4 From buffer-level to machine-level: Lumping . . . 66

7.5.5 State space and state probabilities . . . 66
7.5.6 From machine-level to buffer-level: Partitioning . . . 69
7.5.7 Convergence algorithm . . . 71
7.6 Performance measures . . . 72

8 Numerical Results **75**
8.1 Convergence of proposed algorithm . . . 75
8.2 Comparison with discrete event simulator . . . 76
8.2.1 Simulation model . . . 77

9 Real Case **81**
9.1 Line model . . . 81
9.2 Optimization problem . . . 82
9.3 As-is inspection policy . . . 83
9.4 Solving unbalancing . . . 84
9.5 Sensitivity Analysis . . . 86
9.5.1 Percentage of markers equal in each machine . . . 86
9.5.2 Percentage of markers different in each machine . . . 88
9.5.3 Comparison between two strategies . . . 89
9.6 Importance of optimal set of measurement markers . . . 90

10 Conclusions and future developments **95**
10.1 Future developments . . . 96

Bibliography **97**

A Appendix A **103**
A.1 2M1B Line . . . 103
A.1.1 Parameters of experimental campaign . . . 103
A.1.2 Results . . . 104
A.2 4M3B Line . . . 104
A.2.1 Parameters of the experimental campaign . . . 104
A.2.2 Results . . . 106

List of Figures

3.1 Building Block 15
3.2 Single up-single down Markov chain 16
3.3 System states: internal and boundary. 17
3.4 Graphical representation of machine and buffer level decomposition. 24
3.5 Forward and backward decomposition. 26

4.1 Markers[56] 27
4.2 Dies field errors 29
4.3 Inter/Intra-field error 30
4.4 Stack-up overlay error propagation 30
4.5 Plots of layer-specific yield rate (a) $\Pi(A)$ based on selected markers, (b) $\Pi(B)$ based on unselected markers and (c) $\Pi(A)\Pi(B)$ based on all the markers. 40
4.6 Quality probabilities. 41

5.1 Front-end fabrication processes[1] 43
5.2 Overlay and stuck-up Overlay deviations 44
5.3 External Observer looking at outgoing flow 45
5.4 External Observer looking at outgoing flow at last machine 46
5.5 Physical Reference System 46

6.1 Features to be integrated in analytical models to better represent reality . 50

7.1 Model Reference System 54
7.2 Markov Model 57
7.3 Single-Stage model with down states. 58
7.4 2M1B Line 59
7.5 Building Block 59
7.6 Long Line 61
7.7 Schematic representation of the proposed method. 62
7.8 General Integrated Machine 64

7.9 Relation between the Markov Chains of Integrated Machines (machine-level) and pseudo-machines (buffer-level). 70
7.10 Convergence Algorithm 72

8.1 Convergence paths 75
8.2 Convergence details 75
8.3 Discrete event simulator used 2M1B 78
8.4 Box Plot errors 2M1B line 79
8.5 Discrete event simulator used 4M3B 79
8.6 Box Plot errors 4M3B line 80

9.1 System-Process-Product control 81
9.2 Photolithography line model reduction 82
9.3 Starvation Propagation in current configuration 83
9.4 Starvation Propagation in Optimal configuration 85
9.5 Sensitivity analysis over K_1 and K_2 with optimal % of markers equal for each machine 87
9.6 sensitivity analysis over K_2 cost parameter 88
9.7 sensitivity analysis over K_2 cost parameter 89
9.8 Comparison between two different inspection policies 90
9.9 Throughput endline $TH_{G\&BND}$ 91
9.10 Yield: as $\frac{TH_G}{TH_{INPUT}}$ 92
9.11 Throughput endline of good parts TH_G 92

List of Tables

8.1 Convergence analysis of different configurations 76
8.2 Simulation settings 77
8.3 Parameters of the experimental campaign: Two-Machine Line 78
8.4 Errors on performance measures 79
8.5 Parameters of the experimental campaign: Four-Machine Line 80
8.6 Errors on performance measures 80

9.1 Optimal percentage of markers in each inspection machine given K_1 and K_2 84
9.2 Comparison between 100% markers selected for each machine and optimal configuration found 85

A.1 Full factorial with 2 values for each parameter 103
A.2 Throughput end-line and throughput of scraps results 104
A.3 2^{k-p} fractional factorial plan. $K = 7$,$p = 1$ 106
A.4 Throughput end-line and throughput of scraps results 108

1 | Introduction

1.1. The industrial context

The Manufacturing Sector has always been a key factor to influence a nation's growth. Manufacturing companies are today operating in highly competitive dynamic scenarios with continuously changing conditions. The challenge is to remain competitive in volatile, fast-moving, and customer-driven markets, where even a small fluctuation in the final demand can result in massive decision problem when moving further up the supply chain. Moreover, we are living in a landscape of global change: the increasing global population and the fast development of emerging countries create new markets on one hand but new competitors on the other. The consumer society in which developed countries live, enhance the demand for high-quality products with a high possibility of customization and push companies for a faster time to market.

This scenario will entail higher production capacity to achieve and higher resource and energy consumption for keeping the same living standard for an even larger population. For this reason, a decoupling of the consumption of material and energy from the rising global demand is required(see [2]).Otherwise, the high consumption level will for sure lead to the reduced availability of virgin material and the generation of more and new wastes.

An example of the problems that this might create in Europe, could be the increased scarcity of raw materials used in High-Tech applications and consequently a price increase. This could threaten Europe Energy transition towards renewable sources. The decrease in material availability would increase the dependency of Europe on resource-rich countries like China or other countries worldwide.

The continuous and fast development of new technological solutions has led to shorter product life cycles, challenging companies to plan for facilities whose useful lives are much longer than the life cycle of any individual product it manufactures. Flexibility and reconfigurability are nowadays a must required skill in every manufacturing system. The process in technology has also provided several possibilities to exercise better control over a production plant's performance, both from the point of view of quality and production

logistic. Advanced sensors and machine data collection allow to rapidly inspect several product characteristics with high accuracy and on-line.

The deployment of increasingly complicated designs to enable the administration and control of production processes has improved recently thanks to digitalization. Particularly, the Manufacturing Execution System (MES) has become important to business operations as the primary software module for the implementation of cutting-edge Zero-Defect Manufacturing techniques (ZDM). The implementations of MES have enhanced the control of manufacturing systems for production quality.New sensor technologies in particular have made it possible to collect a wide range of data in real time on manufacturing lines while the process is being carried out at a very fast collection rate. To build a manufacturing system that is adaptable and customer-focused, the issue will be proactive control with appropriate data collection and integrating solutions. These variables result in a transition away from inflexible mass production toward an agile process that can respond with a minimal amount of changeover and production gap cost while always meeting the volume and quality requirements. To strike a balance between efficiency and effectiveness, the production system's complexity must be improved from a global perspective and at various levels. Manufacturing businesses are researching Quality, Production Planning, and Maintenance as essential processes that must be monitored in manufacturing systems to prevent sub-optimal improvement in order to meet these objectives.

1.2. Motivations

The Semiconductor Manufacturing system is recognized as a highly intricate production process comprising four fundamental stages: wafer fabrication, wafer probing, assembly (packaging), and final testing. The initial phase, known as wafer manufacturing or the front-end, incurs significant costs. During this phase, circuits are methodically layered onto the wafer using a series of sequential procedures. Numerous processing steps are involved in this phase.

Consequently, the dynamics, performance, and characteristics of both the process and the end product are determined by an extensive range of factors. Due to the rapidly changing conditions in this market, it becomes imperative to consider structural reconfigurations, improvement initiatives, and operational adjustments while thoroughly evaluating all possible alternative comparisons to devise the most optimal system for a multitude of scenarios.

Photolithography is the crux of IC manufacuring among the entire process in the fab in a manner that experts in the sector say the fab is built around the process of pho-

tolithography. To produce an entire semiconductor wafer, many such steps are performed subsequently and each pattern transfer has a very precise position on the wafer surface. To ensure the correct alignment between the layers an inspection station is required.The inspection, is considered the bottleneck of the line, it takes a way longer time than the other steps. It could be possible to have a faster but less reliable inspection station compromising the knowledge on the product quality but decreasing the cycle time of the bottleneck of the line. But with a single-stage product-process model is not possible to have a general overview of the outcome of a multi-stage production system.

In general, several analytical techniques have been developed to analyze the behavior of manufacturing systems, utilizing equations that assist in making precise decisions during production planning strategies. However, analytical models consist of intricate engineering formulations that are challenging to derive, and they may not always accurately reflect the actual behavior of the system. This is because certain restrictive assumptions may not align with the complexities and dynamics of the system, including improvement initiatives. Another way to address the problem is by simulation tools, as they provide a closer approximation of real-world performance outcomes. Nevertheless, it is important to note that simulating the actual system and extrapolating the results can be time-consuming endeavors.

The good functioning of analytical methods depends on the ability to take into account most of the factors that can affect the behavior of the manufacturing line. Among all of the variables that need to be considered, the quality control system represents a relevant factor for the performance of the system. Currently adopted quality control strategies are mainly single-stage strategies as they do not consider the impact of quality monitoring actions on the economic, logistics, and quality performance of the multi-stage systems in which they are applied.

A deeper understanding of the impact of quality control systems on both the actual quality of the process and product and the performance of the system can be of real help in taking focused decisions when designing the production system.
Moreover aspects such as Quality, process control, production planning, and maintenance that have been treated by scientists and industrialists as separate research areas and have very rarely been considered at a production system-level need to be jointly considered [14]. The growing emphasis on Lean Production has driven researchers to delve deeper into analyzing how system design impacts product quality. In reference [11], it is demonstrated that reducing inventory levels enhances the system's ability to detect quality issues at an earlier stage. However, within the field of manufacturing system engineering, it is widely recognized that the presence of buffers positively affects the production rate of the system.

Buffers effectively decouple machine behaviors, preventing disruptions from propagating upstream and downstream along the production line [28].

Additionally, previous studies have revealed an inverse relationship between operating speed and product quality [50]. Thus, while improving the machine processing rate positively impacts system throughput, it may have a detrimental effect on the system yield, probably the most critical measure in semiconductor manufacturing. The lack of comprehensive understanding often leads to sub-optimal and unbalanced solutions that prioritize one aspect while compromising the overall efficiency of the manufacturing system.

These considerations strongly underline the need for dedicated research activities and highlight their potential impact in terms of enhancing the knowledge of production system behavior and reducing costs for companies. The pursuit of such research is motivated by the desire to achieve a comprehensive understanding of the intricate dynamics within manufacturing systems and to develop effective strategies that maximize overall efficiency and performance.

1.3. Objective

The objective addressed in this thesis is to integrate product model, Stream of Variation Model, process models, Robust control, and a system model that deals with system's performances in order to have an unified framework able to analyse optimal operations. Otherwise considering these models as separate could lead to sub-optimal when considering multi-stage system as a whole. This would lead to obtain a model capable to detect and define propagation of quality errors of production at each different manufacturing stages and to have a reliable performance evaluation of the manufacturing system.
To achieve this goal a new formulation of approximate analytical methods is introduced, using a Continuous flow model developed by Tolio,Matta [57] considering scrapping in process, and quality errors propagation. The work fits in the framework described in 1.2 and tries to make a step further answering to the questions: "Is it possible to combine product-process-system models into a single framework?" and "Could as-is front-end semiconductor manufacturing be improved?"

Subsequently, the effects of these features on the system's productivity are analyzed, and the interrelationships between quality and productivity are investigated. Furthermore, a genuine case study in the realm of semiconductor manufacturing is meticulously examined, aiming to provide substantial support for the development of a robust multi-stage process control model that effectively reduces measurement points in overlay metrology. This endeavor seeks to delve deeper into the intricacies of system-level dynamics, thus offering

profound insights and understanding.

Within the semiconductor industry, the significance of production quality cannot be overstated, particularly in the realm of wafer fabrication, which encompasses a multitude of stages. Considering the inherently low production yield (averaging around 50%) and the stringent requirements imposed on meeting due dates, the timeliness and accuracy of the quality control system become paramount. The protracted flow time further accentuates the need for an agile and responsive quality control system.

Photolithography process, where precise alignment of numerous printed layers is imperative with sub-nanometric precision, presents a critical bottleneck at the inspection station. Consequently, it becomes vitally important to comprehensively assess the impact of reducing measurement points for process control. This evaluation entails examining the influence on both quality and production rate, while adopting a holistic perspective that encompasses the dynamics of the entire system.

1.4. Thesis outline

The thesis is structured in the following chapters:

- In chapter 2. a literature review of the main product-process-system models treated for semiconductor manufacturing will be given.
- In chapter 3. a brief recap of the approximate analytical model proposed in [57] and of the two-level decomposition method presented in [44] is performed.
- In chapter 4. overlay error is explained within its control with a robust approach, then reduction of number of measurement is explained.
- In chapter 5. reference system and explanation of product-process-system models for the specific case are addressed.
- In chapter 6. the problem addressed by the thesis is formalized.
- In chapter 7. the analytical model of single stage and multi-stage is presented. Decomposition is addressed with scrapping and propagation of quality states.
- In chapter 8. numerical results are presented.
- In chapter 9. Real case from semiconductor manufacturing is analyzed.
- In chapter 10. conclusions about the analysis conducted are drawn and hints for future research are given.

2 | Literature Review

2.1. Integrated quality control in multi-stage production systems

A multistage system refers to a system that consists of multiple components, stations, or stages necessary for completing a final product or service. Such systems are prevalent in various industries, and nearly all modern manufacturing processes fall into this category.

The complexity of multistage systems poses significant challenges for effective quality control and improvement. However, recent technological advancements have provided us with tools to understand and overcome these challenges. In discrete manufacturing processes, it is now common to conduct complete inspections at each intermediate operation and employ high sampling rates. The abundance of acquired data creates opportunities for effective quality control systems. Furthermore, advancements in sensors, data acquisition systems, and computer networks have made high-tech methodologies more accessible and affordable for factories. The wealth of Big Data has demonstrated that managing a Multistage Manufacturing System (MMS) requires considering every level involved.

Researchers have studied and developed Cyber-Physical Systems (CPS), which are collaborative computational entities connected to the physical world and its ongoing processes. These systems provide data-accessing and data-processing services via the Internet. In the context of MMS, CPS consists of autonomous and cooperative elements and subsystems connected based on the production context, from processes and machines to production and logistics networks. This technology offers opportunities to achieve the goal of Zero-Defect Manufacturing (ZDM) within the framework of Industry 4.0. (Lee et al., Zhong et al. [39] [65]; Monostori et al. [47])

The abundance of process information in multistage systems has presented significant opportunities for quality improvement, leading to the development of modeling efforts to establish a mathematical description of the interactions between productivity and the quality of the final product.

While there has been considerable interest in analyzing production line performance from a productivity perspective, analytical and simulation-based tools have been developed, but less attention has been given to studying the relationships between quality performance measures, process control, and productivity in production systems. Colledani, Tolio et al. [14] emphasize the need to consider quality, production planning, and maintenance jointly and propose production quality as a new paradigm beyond traditional six-sigma approaches. Sale et al. [55] report several cases from the automotive sector highlighting the impact of different system layouts on product quality and the lack of joint analysis between quality and productivity. The importance of integrating these fields in semiconductor manufacturing is evident in Bassetto et al. [3], where the author proposes a methodology for updating the manufacturing control plan by integrating product, process, and tool data at a system-level perspective. Nonaka et al. [48] demonstrate that the correlation between machine failures and defects in semiconductor manufacturing fabs is the primary cause of quality problems.

Benavent Nacher et al. [54] a dedicated DES based on stream of variation model that has been successfully applied to manage product geometric variation in these systems. In this work, which is focused on the production quality paradigm in a model-based system engineering context,where a digital prototype is proposed to integrate productivity and part quality based on the stream of variation analysis in multistage assembly systems.

Kim et al. [35] investigate how the configuration of production system layouts affects the performance of the quality control system in flow lines. Machines in these systems can experience both operational and quality failures. Under normal operating conditions, machines produce defect-free items, but upon transitioning to the quality failure state, they only produce defective products. The quality control action involves stopping the machine and initiating repair interventions to restore it to a state of perfect quality. The authors consider machines as Markovian models, and the quality control action is represented as a transition that moves the machine from a low-quality state to an inoperative state for the repair process. The study reveals that if processed parts have to pass through multiple stations before inspection, it results in a delay in quality information feedback, reducing the responsiveness and effectiveness of the quality control system. The authors extend their approach to longer production lines in Kim et al. [36].

Colledani and Tolio [12] propose a new analytical method for evaluating the performance of systems where machines are monitored using Statistical Process Control.

The method considers the presence of inspection and integrated stations subject to operational failures and out-of-control states. For the first time, the authors explicitly model

the quality control mechanism within the manufacturing system, establishing a complete link between the two interacting systems. The results highlight counter-intuitive behaviors that can only be accounted for by jointly considering quality and productivity. While the method is initially valid for synchronous lines where machines simultaneously produce and drop parts into the buffer, Colledani and Tolio [11] extend the analysis to production systems where off-line inspections are performed.

The objective of this thesis is to expand the method for asynchronous production lines with scrap and integrate it with quality propagation at product level and model-based process control to have a unique framework that consider dynamics at product, process and system level. This integrated approach aims to make informed decisions on quality policies, considering the comprehensive perspective of quality and productivity. A similar methodology will be adopted to lithography process in a semiconductor manufacturing fab, aiming at a decrease in measurement's reduction for process control in overlay metrology.

2.2. Product Level Model

Researchers have focused on developing models to analyze the flow of product quality errors across multiple stations in a Multistage Manufacturing Process (MMP). A math-based methodology called Stream of Variation (SoV) has been employed to predict potential manufacturing problems downstream. This methodology treats the flow of quality information as analogous to the flow of water in a river, hence the term "stream of variation." These SoV models, initially introduced by Hu [28], establish an analytical connection between quality errors and product/process parameters. They integrate multivariate statistics, control theory, and manufacturing process knowledge into a unified framework, thus eliminating the need for costly trial-and-error fine-tuning in new-product manufacturing processes.

2.3. Model-based process control

Extensive efforts have been dedicated to managing production systems, particularly in the realm of process control.

Initially, SoV models were used for optimal measurement allocation (Sampatraj et al. [46]), identifying root causes of quality errors at the process level (Ding et al. [16]), and designing MMPs (Hu et al. [29]). More recently, research in model-based process control has focused on optimal in-process adjustments of programmable tooling, enabling auto-

matic minimization of errors in product quality based on previous in-line measurements collected along the MMP.

In Djurdjanovic et al. [17], SoV models were employed to derive a model for deterministic feed-forward adjustment of control process parameters. Jiao and Djurdjanovic [32] utilized SoV models for stochastic feed-forward control of quality in multistage manufacturing processes.

Previous research assumed perfect knowledge of process parameters, whether through physics-based approaches (Jiao et al. [33]) or data-driven methodologies (Jin et al. [34]). However, the increasing demand for high-quality products and rapidly evolving technology have led to more complex processes, where uncertainties in model parameter estimates and noise characteristics are inevitable. In Djurdjanovic et al. [18], the authors proposed a method for controlling quality errors in MMPs that is robust to uncertainties in the error flow model. The objective is to minimize product quality variation under worst-case scenarios of potential inaccuracies. However, this approach still assumed perfectly known structural model parameters and modeled noise terms as independent and identically distributed (IID) random vectors.

In Djurdjanovic et al. [61], the researchers relaxed these assumptions and introduced robustness to inevitable inaccuracies in the error model. They overcame the restrictive assumptions of Gaussianity and independence of noise terms. The study emphasized that compared to non-robust control models, the benefits of robustness increase as uncertainty levels and the progression of the manufacturing process rise.

2.4. System-level models

Multi-stage production and transfer lines are made up of a series of machines that are designed to carry out specific operations on raw materials at a predetermined processing rate. The reliability of this type of manufacturing system is heavily dependent on the performance of the machines executing the processes, given the random nature of the phenomena involved. Therefore, a stochastic approach is necessary to obtain appropriate mathematical models to evaluate the system's performance. To help manufacturers in evaluating the performance of their production systems, various models and techniques have been widely researched and developed. Papadopoulos et al. [51] present a classification and review of existing performance evaluation models, which can be broadly categorized into three main tools: Queuing Network models, Markov Chain-based models, and simulation models. While Queuing Networks can provide an exact solution, their use is limited by certain assumptions. Simulation Models, such as Discrete Event Simu-

lation (DES), are frequently used in industry due to their ability to reach a high level of detail. However, they can be time-consuming and require the repetition of experiments to obtain statistically reliable results. Analytical models based on Markov Chains are a middle ground between the above approaches and can provide accurate results with fewer restrictions than Queuing Network models. These models differ based on the assumptions made regarding the flow of parts and the machines' processing times.

The Discrete Deterministic Model consider that processing times are deterministic and equal for all the machines. It was firstly introduced by Buzacott [5] and then studied by Gebennin [22]. The Discrete Exponential Model is characterized by stochastic machines processing times and different among them. Service times are exponentially distributed. This model has been introduced by Gershwin and Berman [25]. The Continuous Deterministic Model is characterized by deterministic machines service times and they can be different among machines. The model treat processed material as a continuous fluid and it is particularly suitable for automatic asynchronous systems. Continuous behavior is also used to approximate discrete production lines. The Continuous approximation of Discrete Deterministic Model assumes the discrete material flow as a continuous fluid with machines operating as valves, and buffers as tanks between them Gershwin et al. [26]. Different authors as Xie et al. [62] and Gershwin et al. [24] adopt this model. In Tolio, Matta & Gershwin [58] the possibility of having multiple failure modes for machines was introduced. Tolio and Ratti [59] evaluate the performance of two-machines lines with generalized thresholds adopting the Continuous Approximation of Discrete Deterministic Model.

2.4.1. Performance evaluation of two-machines lines

A production line composed by two machines and one buffer is called two- machine line or Building Block. Various authors have developed performance evaluation models for two-machine lines. Li et al. [42] outline a comparison between the different two-machine line models. Focusing on the deterministic model with continuous approximation of discrete flow, Gershwin [24] presents the procedure to obtain the exact solution considering machines with single failure mode. In this work, the internal states of the buffer are described by probability density function because they are characterized by a continuous change. The boundary states of the buffer are described by probability mass function and, finally, the solution is obtained thanks to a guess. Tolio et al. [58] analyze two-machine lines considering multiple failure modes of the machines and finite buffer capacity. Tolio and Magnanini [44] develops an analytical method for the performance evaluation of deterministic asynchronous two-machine lines proposing a model based on a continuous

representation of the material flow coupled with a threshold based control policy. Tolio and Ratti [59] evaluate the performance of a two-machine line with generalized thresholds. Here the idea is that the machines can behave differently above or below certain buffer levels named thresholds. This means that a machine is described by a particular Markov chain depending on the buffer level. This approach is particularly suitable to control the system by means of thresholds.

2.4.2. Performance evaluation of long lines

The next natural step is to extend the analysis to longer production lines. For the reason of mathematical tractability no exact analytical models are available for such systems. Thus, approximate methods have been developed. In particular two main approaches can be identified in literature: decomposition techniques and aggregation techniques. The basic idea of the aggregation technique (De Koster) [37] is to replace a two-machine one- buffer sub-system by one single equivalent machine. Interesting discussions about aggregation procedures are presented in Chiang et al. [7] and Chiang et al. [8]. Decomposition techniques are the most investigated. The main idea is to decompose the line in a series of two-machine one-buffer sub-systems: provided that analytical models for each sub-problem are given, the performance parameters of the whole line can be computed by means of appropriate iterative procedures. As Levantesi et al. [40] state, the idea of the Decomposition Method is that, in order to obtain an estimate of the performance of the original system, it is necessary to reproduce in each two-machine line a flow of material as close as possible to the flow of material observed in the corresponding buffer of the original line. Blocking and starvation phenomena affect the performance of each machine of the line: this means that the performance of each machine of the line is influenced by the failures of each other machine. Decomposition equations are used to evaluate the influence of remote blocking and starvation phenomena on each Building Block. Since the set of equations is not linear, an iterative algorithm needs to be used to calculate the performance. Gershwin [23] proposed a decomposition method for transfer lines, improved by Dallery et al. [62]. Choong and Gershwin [9] extended this approach to lines with exponentially distributed processing times and Burman [4] improved the method also for asynchronous lines. Tolio and Matta [57] develop the Decomposition Method considering multiple failure modes of the machines resulting in a more realistic representation of the line because each machine can fail in different ways. Colledani et al. [13] discuss the Decomposition Method considering that different types of products are produced. Colledani et al. [10] considers partial or complete blocking and starvation phenomena and develops a more accurate evaluation of the average work in progress. The

Decomposition Method with the continuous Approximation of discrete flow of parts has been applied by Magnanini and Tolio [44] and a comparison with discrete event simulation shows the accuracy of the method.

2.4.3. Work in progress scrap

In a manufacturing system, reworking of the defectives and management of waste or scrap are important issues that call for immediate attention to meet the basic objectives and requirements for lean production system. Perfectly lean systems would be error or defective-free at all stages, but inevitably this is not possible. Hence the production of scrap is inevitable and need to be considered when evaluating the performance of production systems. Indeed when considering a manufacturing line where inspection is performed in-line and the part result to be defective, the cause can be an OOC machine and the parts produced by the upstream machine till the discover of the problem need to be scrapped. Thus, the line have to be unload before the failed machine is repaired. Another case where material need to be scrapped along the line during production is when dealing with goods whose their physical or chemical characteristics fall out of specifications during a stoppage. A valid example is the food industry, where certain processes must be performed in a timely and carefully controlled manner and long failures and disruptions in the production process may cause severe quality deterioration. Also, the extensive exposure of material to certain environments (e.g., heat, humidity, acidity, etc.) is a common fault for in-line scrap. There are countless other examples of manufacturing processes where WIP may be damaged and may have to be scrapped because of stoppages. From a manufacturing systems engineering perspective, when dealing with long linear production lines where inspection is performed within the system, neglecting the impact of the scrap in the performance analysis of the system can lead to massive approximations. In the literature Okamura et al. [49] considers a two-workstation model in which when a workstation fails, the part in it is scrapped. In Jafari et al. [31] the author analyze a transfer line with geometrically distributed uptimes and downtimes in which, when a workstation fails, the part in it is scrapped with a certain fixed probability. Buzacott et al. [6] considers a two-workstation line with no intermediate buffer. Each workstation can accommodate one part. When a workstation fails, the part in it is scrapped as soon as the workstation becomes operative. In all the above works it is assumed that when a failure occurs at a workstation, the one part that is on the work-station is scrapped. Pourbabai [52] describes a model with more than two workstations and nonzero buffers but assumes that if a blockage occurs, the trapped parts are scrapped. More recently Liberopoulos et al. [43] develop a model for bufferless, paced, automatic transfer line where, when a

workstation fails, it stops operating, and so do all the other workstations upstream of it, and the parts trapped in the stopped workstations are scrapped after a maximum amount of time. In Kim et al. [35] the impact of quality on the system productivity is analyzed but no scrap is integrated in the system and the number of defective parts is computed at the end of the line. In Colledani and Tolio [11] an off-line inspection machine controls other machine by means of Statistical Process Control (SPC) and stops the machine that generated the defective part as soon as an OOC is detected. However the material trapped inside the buffers between production and inspection machines is not scrapped and the defective parts need to pass trough all the line where the yield is computed. Moreover the model is valid for synchronous production lines. In this work the scrap for asynchronous machines with inter-operational buffers is modeled and integrated in the performance evaluation in Magnanini, Tolio [45].

3 | Recap of Approximate Analytical Model

This chapter will summarize the performance evaluation models for production systems adopted in this thesis. Firstly, the general characterization of a production line will be shown. Then the performance evaluation of two-machine lines will be discussed, and finally the Decomposition Method will be presented.

3.1. Production Line Characterization

This section will introduce a specific model for machines and buffers on a production line. A production line is a manufacturing system usually organized with a linear development of machines and inter-operational buffers. A production line with K machines and *K-1* buffers is called K-machine line. The simplest system consists of two machines decoupled by a buffer. This system is called Building Block. The upstream machine processes the flow of material and puts it into the buffer, from where the downstream machine takes the material to process it.

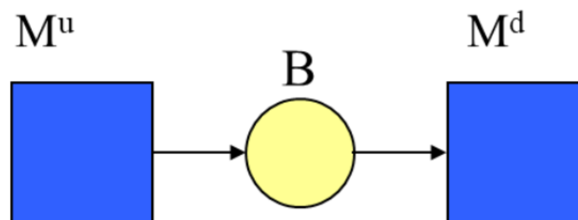

Figure 3.1: Building Block

Machine characterization

Each machine is characterized by a set of operational states called up states and a set of non-operational states called down states. Each state has a different production rate μ_i

defined as the number of parts processed in a time unit and, therefore, it corresponds to the processing speed of the machine. As [15] states, it can be evaluated as: *1/CT* where CT is the average processing time. Up states are those states where $\mu_i \neq 0$. On the contrary down states are those where $\mu_i = 0$. Transitions from up states to down states are known as failure rates. They can occur only if the machine is operational and are called operation dependent. Transitions from down states to up states are known as repair rates. They can occur even if the machine is not working and are called time dependent. Transition rates are exponentially distributed and defined as follows:

$$p_{ij} = 1/MTTF; \quad r_{ij} = 1/MTTF; \tag{3.1}$$

where:
- p_{ij} : failure rate of failure mode j of machine $M\{j\}$
- r_{ij} : repair rate of failure mode j of machine $M\{j\}$
- $MTTF$: mean time to failure of failure mode j of machine $M\{j\}$
- $MTTR$: mean time to repair of failure mode j of machine $M\{j\}$
Note that processing and transitions rates are design parameters provided by the machine manufacturer. For these reasons each machine is modeled as s continuous-time discrete-state Markov Chain.

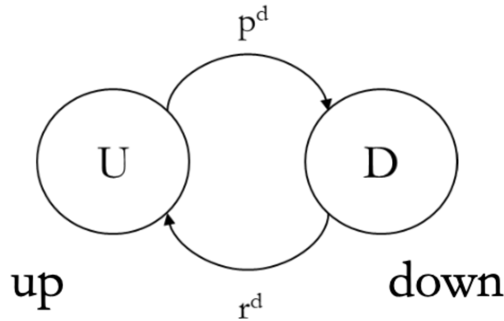

Figure 3.2: Single up-single down Markov chain

Buffer characterization
A buffer is any storage space used to collect parts exiting machine $M\{i\}$and waiting to be processed from machine $M\{i+1\}$ with the aim of decoupling them. Indeed, stops due to breakdown, maintenance, blocking or starvation create imbalances within the system. Using buffers between resources allow machines to be more independent thus keep the

system working for a while in case of a stop on just one machine. The material in the buffer is considered continuous , therefore buffer is characterized by the continuous variable $0 < x < N$ where N is the maximum capacity.
System states
The state of the system is mixed continuous-discrete and is completely defined by the joint states of the machines, which belong to a discrete set, and by the buffer level x which is a continuous variable. Combining both the states of the machines and the buffer level the system state is defined as $S = (S_u; S_d; x)$. Figure 3.3 shows a graphical representation of the system states: every point on the line is a system state and it has the same machine states while the state of the buffer varies between 0 and N. System states are divided in two groups for calculation purpose: internal and boundary states. Internal states are all those states where buffer level is different from the boundaries. Since the material is considered as continuous in the internal states the buffer level changes immediately from state to state

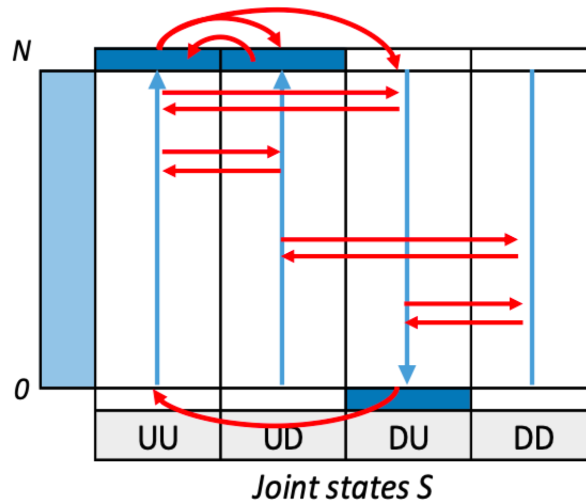

Figure 3.3: System states: internal and boundary.

therefore the time spent in a state is equal to zero. Hence the probability mass of being in a given joint state with a buffer level x is equal to zero:

$$\pi(x; S_u; S_d) = 0 \tag{3.2}$$

Therefore it is necessary to evaluate the probability density function:

$$f(x; S_u; S_d) = 0 \tag{3.3}$$

Once the probability density function is defined, by integrating it the Distribution function is obtained.

$$F(x; S_u; S_d) \tag{3.4}$$

When the Distribution Function is known it is possible to find the probability that the system is in a given joint machine state and in a given interval of buffer levels. For example the probability that the system is in the joint state UU with buffer $0 < x < N$ is:

$$\Pi(UU) = \int f(s; UU) ds \tag{3.5}$$

Boundary States are those particular system states in which buffer level x is equal to 0 or N. In some of those states the system may remain for a finite amount of time therefore a probability mass $(x; Su; Sd)$ exist. In figure 3.3 boundary states are represented as rectangles. The ones filled in blue are those where the system can remain for a finite time.

3.2. Performance evaluation of two machines line

Here, a model for the performance evaluation of two-machine lines with general machines, finite buffer capacity is presented. The model was introduced by [60] and belongs to the family of Continuous Deterministic Models. Continuous means that the flow of parts is described by a continuous material flow. General machines means that each machine can assume different states, each one characterized by its production rate and transition rates to the other states of the machine. Therefore, the model proposed can also be used to evaluate the performance of two-machine lines with finite buffer capacity. The assumptions of the model are:

- The upstream machine is never starved and the downstream machine.
- The quantity $\mu\{S_i^u\}$ is the speed at which the upstream machine processes material while it is in state $\{S_i^u\}$ and is not constrained by the other machine or by the buffer.
- The upstream machine can have multiple up states (in which $\mu\{S_i^u\} \neq 0$) and multiple down states (in which $\mu\{S_i^u\} = 0$). Same consideration applies for downstream machine.
- Buffer capacity is assumed to be finite and the buffer level can change in a continuous fashion
- Machines are unreliable and failures are operation dependent (ODF).

- Processing times include the time to load the part and the time to unload the part.
- Parts are discrete and each machine processes one part at a time.
- The dispatching policy is Fist In First Out (FIFO);
- The blocking discipline is Blocking After Service (BAS).

In order to evaluate the performances of the system, both internal states, and boundary behaviors are analyzed. Internal and boundary equations are developed and solved through equilibrium equations for a generic joint machine state. For the sake of simplicity these equations are not reported here but can be found in [59].

Internal behavior

The behavior of the machines depends on the buffer level, machines are characterized by a vector of states S^u of size I and S^d of size L. Combining the states of the machines, the vector of joint machine states is defined as $S = S^u \otimes S^d$. This vector contains all the possible combinations of the states of the two machines. For the joint machine states in S, the vector $\nu(S)$ of the difference of speed of the two machines in the various joint states is computed as follows:

$$\nu(S) = -I^u \otimes \mu(S^d) + \mu(S^u) \otimes I^d = -\mu(S^d) \oplus \mu(S^u) \tag{3.6}$$

The vector of joint machine states is partitioned as follows:

$$\Upsilon = [S : \nu(S[k]) > 0]$$
$$\Delta = [S : \nu(S[k]) < 0]$$
$$\Phi = [S : \nu(S[k]) = 0]$$

By defining $\Psi = [\Upsilon; \Delta]$, the vector of joint machine states is $S = [\Psi; \Phi]$. The transition rates among the joint states contained in S can be obtained as follows:

$$\hat{Q} = (I^u \otimes \hat{Q}^d) + (\hat{Q}^u \otimes I^d) = \hat{Q}^d \oplus \hat{Q}^u$$
$$\vec{Q} = (I^u \otimes \vec{Q}^d) + (\vec{Q}^u \otimes I^d) = \vec{Q}^d \oplus \vec{Q}^u$$

$$Q = (\hat{Q} + \vec{Q}) - diag[(\hat{Q} + \vec{Q})^T * [1]]$$

and, after rearranging rows and columns of Q:

$$Q = \begin{bmatrix} \Upsilon \to \Upsilon & \Upsilon \to \Delta & \Upsilon \to \Phi \\ \Delta \to \Upsilon & \Delta \to \Delta & \Delta \to \Phi \\ \Phi \to \Upsilon & \Phi \to \Delta & \Phi \to \Phi \end{bmatrix}$$

Boundary behavior

The boundary behavior concerns the behavior of the system when the buffer level corresponds to a upper level x^+ or lower level x^-. It is possible to exit from X when the system state belongs either to Υ^+ or to Δ^- :

$$\Lambda = \begin{bmatrix} \Upsilon^+ \\ \Delta^- \end{bmatrix} = \begin{bmatrix} (x^+, \Upsilon) \\ (x^-, \Delta) \end{bmatrix}; \tag{3.7}$$

It is possible to enter the interior only from one of its boundary (it is not possible to jump directly to the interior). Therefore, it is possible to enter X through one of the system states on the boundaries belonging to Υ^- or Δ^+:

$$\Omega = \begin{bmatrix} \Delta^+ \\ \Upsilon^- \end{bmatrix} \begin{bmatrix} (x^+, \Delta) \\ (x^-, \Upsilon) \end{bmatrix}; \tag{3.8}$$

The rectangular matrix W is defined in the following: the submatrices W define the probabilities that when hitting a boundary, the state of the system changes immediately from a state in Λ to another state in Ω:

$$W = \begin{bmatrix} \Upsilon^+ \to \Delta^+ & \Upsilon^+ \to \Upsilon^- \\ \Delta^- \to \Delta^+ & \Delta^- \to \Upsilon^- \end{bmatrix}; \tag{3.9}$$

All the system states on the upper and lower boundary of X can be grouped in the sets Θ^+ and Θ^- which together form the column vector $\Theta = [\Theta^+; \Theta^-]$.

The rectangular matrix V is defined in the following: the submatrices V define the probabilities that when hitting a boundary the system changes immediately from a state in Λ to a state in Θ:

$$V = \begin{bmatrix} \Upsilon^+ \to \Theta^+ & \Upsilon^+ \to \Theta^- \\ \Delta^- \to \Theta^+ & \Delta^- \to \Theta^- \end{bmatrix}; \tag{3.10}$$

The transition rates matrix T is defined: the submatrices T define the rates that when

the system is in any boundary state Θ it can exit that state and go either to a state in Ω:

$$T = \begin{bmatrix} \Theta^+ \to \Delta^+ & \Theta^+ \to \Upsilon^- \\ \Theta^- \to \Delta^+ & \Theta^- \to \Upsilon^- \end{bmatrix}; \tag{3.11}$$

Finally, the transition rates matrix O is defined: the submatrices O define the rates that when the system is in any boundary state Θ it can exit that state and go to a state in Θ:

$$O = \begin{bmatrix} \Theta^+ \to \Theta^+ & \Theta^+ \to \Theta^- \\ \Theta^- \to \Theta^+ & \Theta^- \to \Theta^- \end{bmatrix}; \tag{3.12}$$

Internal State Equations

As we said equilibrium equations are not reported here but can be found in [59]. Once rearranged, the matrix $\pounds$ can be defined as follows:

$$\pounds = diag[\nu(\Psi)]^{-1}\{Q^T_{\Psi\Psi} - Q^T_{\Phi\Psi}[Q^T_{\Phi\Phi}]^{-1}Q^T_{\Psi\Phi}\} \tag{3.13}$$

By defining with $\Gamma = [\Gamma[1], ..., \Gamma[R]]$ the eigenvalues and with $\Xi = [\Xi[1], ..., \Xi[R]]$ the corresponding independent eigenvectors of $\pounds$, it is possible to obtain:

$$f(x, \Psi) = \Xi \cdot diag(e^{\Gamma x} \cdot C) \tag{3.14}$$

where $C = [C[1], ..., C[R]]^T$ is a vector of appropriate constants to be calculated using boundary equations.

By integrating the probability density functions it is possible to obtain the distribution functions $F(x; S[k])$.

Boundary equations

The node equations of the states $\Theta = [\Theta_1, ..., \theta_R]$ are considered to obtain the first set of boundary equations. This set is expressed as follows:

$$\Pi(\Theta) = B_1 \cdot g(\Lambda) \tag{3.15}$$

where $B_1 = [diag[T \cdot [1]] - ([O]^T - diag[O \cdot [1]])]^{-1}[V]^T$ and $g(\Lambda)$ is the probability flow. This set of equations allows to calculate the probability masses of the boundary states as a function of the probability density functions of the internal states. The second set of boundary equations is obtained by writing the balance equations when getting into the

interior and it is expressed as follows:

$$g(\Omega) = B_2 \cdot g(\Lambda) \tag{3.16}$$

where $B_2 = [W]^T + [T]^T \cdot B_1$

Considering together the 1st and 2nd set of boundary equations, one of these equations is linearly dependent on the others and, therefore, it must be dropped off from the system and substituted with the normalization equation.

Normalization equation

The normalization states that the sum of the probabilities of all the possible states of the system must be equal to one. The normalization equation is:

$$\sum F(x^+) \cdot [1] + \sum \{\Pi(\Theta)^T \cdot [1]\} = 1 \tag{3.17}$$

The resulting system of equations is the following one:

$$\begin{cases} f(x,\Psi) = \Xi \cdot diag(e^{\Gamma x} \cdot C) \\ \Pi(\Theta) = B_1 \cdot g(\Lambda) \\ g(\Omega) = B_2 \cdot g(\Lambda) \\ \sum F(x^+) \cdot [1] + \sum \{\Pi(\Theta)^T \cdot [1]\} = 1 \end{cases} \tag{3.18}$$

Performance measures

Once the probabilities of the internal and boundary states are known, it is possible to evaluate the performance indicators of the two-machine line. Given the conservation of flow, the throughput of the line can be calculated from the upstream or downstream machine as follows:

$$\begin{aligned} TH^u &= \sum [F(x^+, S)^T \cdot \mu_u(S) + \Pi(\Theta)^T \cdot \mu^{\pm}(\Theta)] \\ TH^d &= \sum [F(x^+, S)^T \cdot \mu_d(S) + \Pi(\Theta)^T \cdot \mu^{\pm}(\Theta)] \end{aligned} \tag{3.19}$$

Two common phenomena in production lines can be modeled by means of buffer: blocking and starvation.

Blocking: If only the upstream machine is operational, once the buffer is full the upstream machine is unable to operate and it goes idle. When this situation happens, the upstream machine is said to be blocked by the down-stream machine. The system remains in this state until the downstream machine is repaired. Two different policies of blocking are commonly used: Blocking before service (BBS) and blocking after service (BAS). With BBS the upstream machine cannot process a part from the upstream buffer while it is accepted in case of BAS mechanism. In this last case the machine cannot release the part to the downstream buffer and, therefore, it is blocked. Blocking after service is considered in this thesis.

Starvation: If only the downstream machine is operational, once the buffer is empty the downstream machine is unable to operate and it goes idle. When this situation happens, the downstream machine is said to be starved by the upstream machine. The system remains in this state until the upstream machine is repaired.

When the buffer level drops down to zero a new starvation state S is added to the state space of the downstream machine and the system goes to a joint state where the downstream machine is waiting for the upstream machine to be repaired. On the other hand when the buffer level reaches its maximum a new blocking state B is added to the state space of the upstream machine and the system goes to a joint state where the upstream machine is waiting for the downstream machine to be repaired. Once a new part or a new space is available in the buffer, the starved or blocked machine returns to an operational state.

3.3. Decomposition method with Continuous Flow

This section briefly explain the decomposition method outlined in Magnanini, Tolio [2017] [44] used for this thesis.

The idea behind this method is that of two-level decomposition: performance are evaluated at buffer level in the *Building Blocks* that have an exact analytical solution, and dynamics introduced by downstream and upstream limitations are reported at machine level, with *Integrated Machine*, so that the final structure of a machine includes both local behavior of the physical machine and remote limitations.

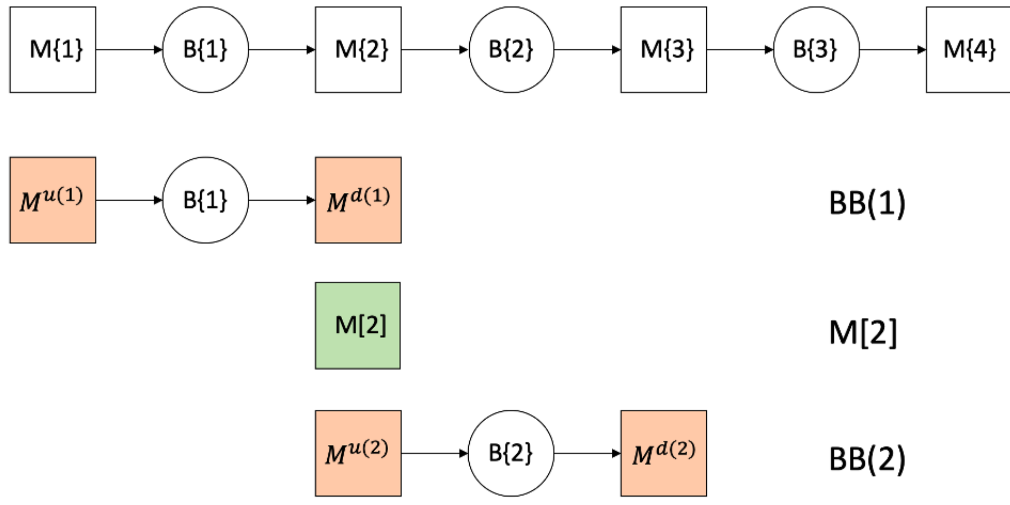

Figure 3.4: Graphical representation of machine and buffer level decomposition.

The assumptions introduced by the model are:

- Machine failures are supposed to happen at the beginning of the operations on a part. Therefore when a failure happens there are no parts partially machined on the machine;
- Under Blocking after service (BAS) policy, the working position of the upstream machine is added to the buffer capacity which therefore becomes N;
- The time in which the part is physically transferred and therefore keeps busy both a position in the buffer and the working position on the machine is considered to be negligible;
- Processing times of the machines are deterministic and may be different between the machines;
- The system is asynchronous i.e. each machine can start or finish a part at any time without synchronization with the other machine;

The application of the decomposition technique consists of three steps:

- $STEP1$: Characterization of two-machine lines (building blocks) with exact analytical solution.
- $STEP2$: Decomposition equations to evaluate the unknown parameters of the pseudo machines of the Building Blocks.
- $STEP3$: Algorithm to solve decomposition equations effciently.

The 1st step of the method is already described in section 3.2.

In the 2nd step of the method Integrated Machines are characterized, one for each machine of the line, by using decomposition equations. Each Integrated Machine integrates, at machine level, the corresponding upstream and downstream limitations descending from the line. Since a failure of one machine affects the behavior of the other machines in the line, the purpose of an Integrated Machine is to take into consideration the blocking and starvation phenomena caused by the interaction of the corresponding machine of the line with the rest of the system. Decomposition equations describe the rate of entering or exiting the limiting states.

The 3rd step of the Decomposition Method applies an algorithm to solve the decomposition equations. The convergence of the throughput is computed by considering that Each Building Block $BB(i)$ represents the entire line centered on the buffer $B\{i\}$ and each Integrated Machine $M[i]$ represents the entire line centered on the machine $M\{i\}$.

In the Building Block the machines are called *pseudo machines* and are updated from the corresponding Integrated Machines. The state space of the upstream pseudo machine $M^u(m)$ takes into consideration the local behavior from the transition rates of the physical machine $M\{m\}$ and the limiting states descending from the upstream part of the line (starvation rates). The state space of the downstream pseudo machine $M^d(m)$ takes into consideration the local behavior from the transition rates of $M\{m+1\}$ and the limiting states descending from the downstream part of the line (blocking rates).

The performance evaluation is performed in two analysis. In the forward analysis, the starvation contributions of each Integrated Machine coming from the upstream part of the line are evaluated. On the contrary, in the backward analysis, the blocking contributions of the Integrated Machines coming from the downstream part of the line are calculated. By imposing the conservation of the flow, the production rate of each Building Block and the related Integrated Machines must coincide. Hence, $BB(i)$ and $M[i]$ are iteratively characterized and solved by means of decomposition equations until the convergence of throughput is satisfied.

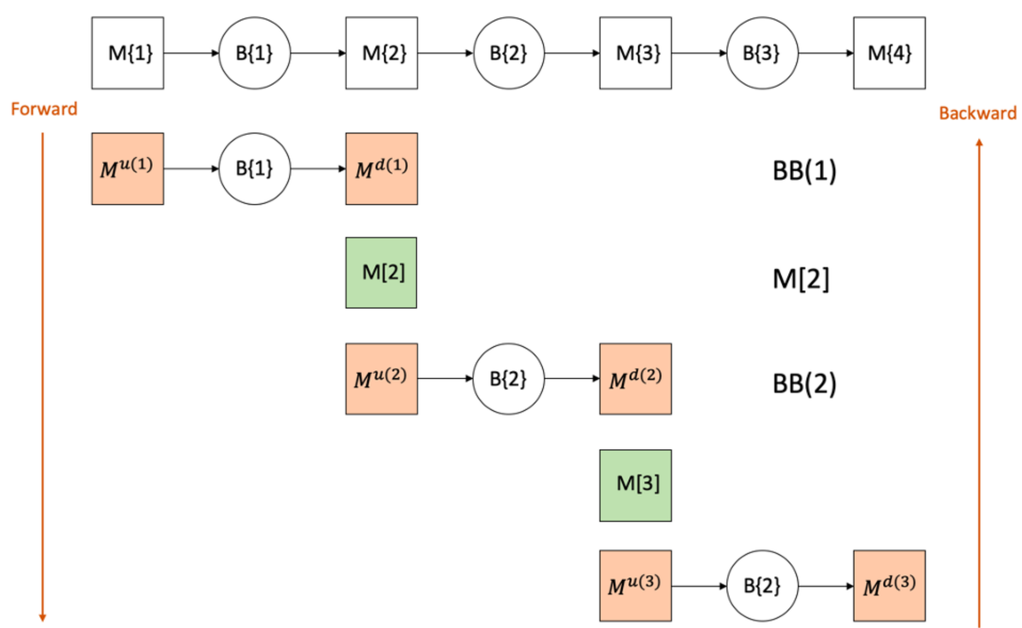

Figure 3.5: Forward and backward decomposition.

4 | Overlay Control

4.1. Overlay error measurements

Overlay measurement involves the creation of specialized patterns known as markers (shown in Figure 4.1) during different lithographic printing stages. These markers are strategically placed on the wafer to enable a metrology tool to accurately measure any overlay errors at specific locations. It's important to note that these markers are printed in areas of the wafer that are not utilized for the integrated circuits (ICs).

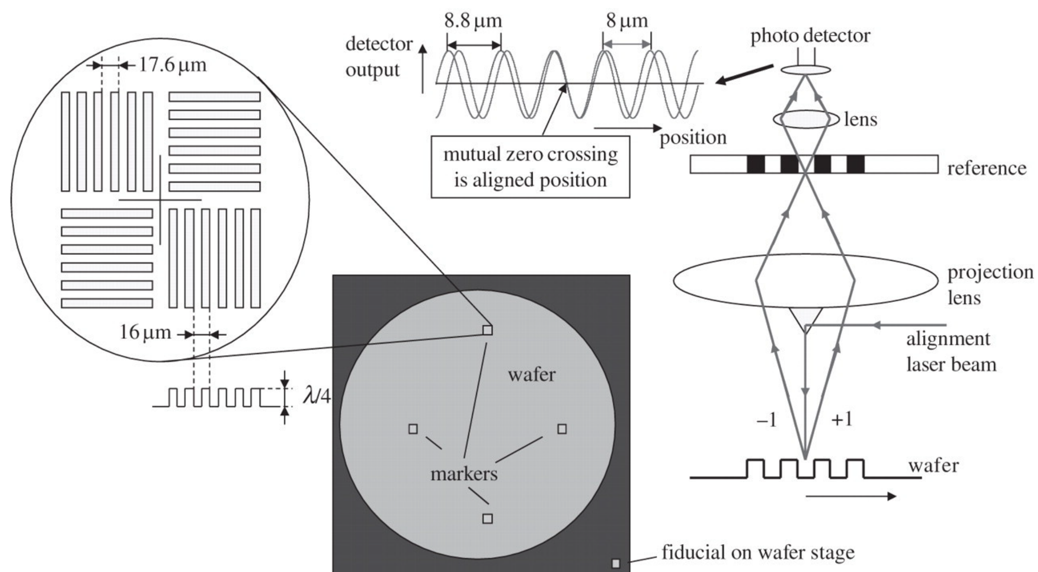

Figure 4.1: Markers[56]

The process of overlay measurement holds significant importance in semiconductor manufacturing due to several key reasons. Firstly, the placement of overlay marks in a grid area reduces the available space for ICs, consequently increasing the cost of each individual IC as the grid size expands.

Furthermore, the time required for inspection during measurements is considerably longer than the actual production time. To measure overlay errors, a camera must focus on each

measurement point. By reducing the number of inspection points, the inspection time can be reduced nearly proportionally. As a result, there exists a trade-off between achieving meticulous inspection and maintaining a high production rate.

It's important to recognize that measuring overlay errors is just one aspect of controlling overlay in a fabrication facility. By carefully planning the number and positioning of measurements, it becomes possible to create a model that fits the resulting data. The coefficients of this model represent physical error terms, which are then fed back into the imaging tool to enhance the overlay accuracy in subsequent printed layers. This reduction in overall overlay errors on the wafers leads to substantial cost savings for the fabrication facility, while simultaneously improving yield.

Given the criticality of overlay errors in ensuring optimal performance of the final product, considerable efforts are dedicated to developing control strategies that ensure alignment between layers within the nanometer scale. These strategies aim to minimize the number of measurement points while maintaining the desired level of accuracy.

4.2. Overlay error model

To enhance resolution and alignment accuracy in the lithography process, it is crucial to model overlay errors and compensate for them within specified tolerances. Typically, overlay errors can be divided into inter-field and intra-field errors [33][41]. Intra-field errors ($E(x, y)$) reflect local effects and vary based on the position within a field (x, y), while inter-field errors ($E(X, Y)$) capture global effects and depend on the overall position of the field's center (X, Y).

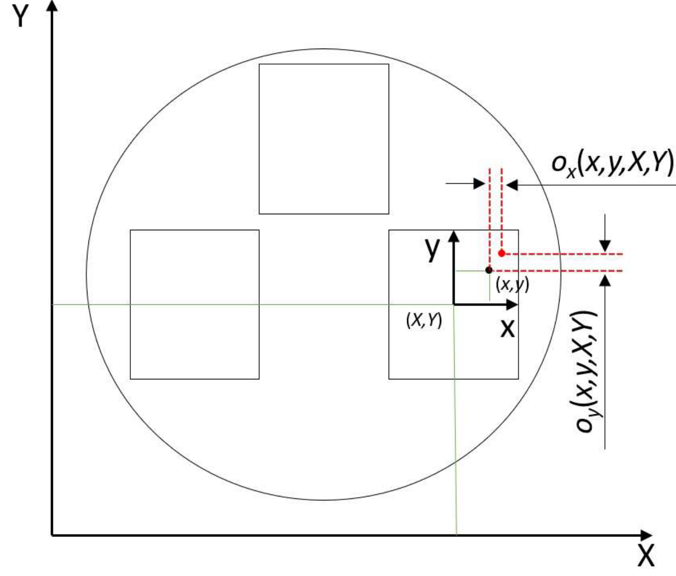

Figure 4.2: Dies field errors

$$
\begin{aligned}
o_x(x,y,X,Y) &= E_x(x,y) + r_x(x,y,X,Y) \\
o_y(x,y,X,Y) &= E_y(x,y) + r_y(x,y,X,Y)
\end{aligned}
\tag{4.1}
$$

Overlay errors can be classified into two categories, systematic and non-systematic. Systematic overlay errors are caused by specific error factors and can be mitigated through appropriate control adjustments. Non-systematic overlay errors ($r(x_i; y_i; X_i; Y_i)$) refer to residuals from the model and cannot be corrected, but efforts can be made to minimize them by reducing manufacturing variances [27]. Equation 4.2 presents a general model that considers error sources related to optics, wafer position, and alignment between them, up to the second order wedge distortion.

$$
\begin{aligned}
o_x(x_i,y_i,X_i,Y_i) = T_x - R_xY_i + M_xX_i + B_xY_i^2 - r_xy_i + m_xx_i - t_xx_i^2 - v_xx_iy_i + w_xy_i^2 + \\
d_{3x}x_i(x_i^2+y_i^2) + d_{5x}x_i(x_i^2+y_i^2)^2 + r_x(x_i,y_i,X_i,Y_i) \\
o_y(x_i,y_i,X_i,Y_i) = T_y - R_yX_i + M_yY_i + B_yX_i^2 - r_yx_i + m_yy_i - t_yy_i^2 - v_yx_iy_i + w_yx_i^2 + \\
d_{3y}y_i(x_i^2+y_i^2) + d_{5y}y_i(x_i^2+y_i^2)^2 + r_y(x_i,y_i,X_i,Y_i)
\end{aligned}
\tag{4.2}
$$

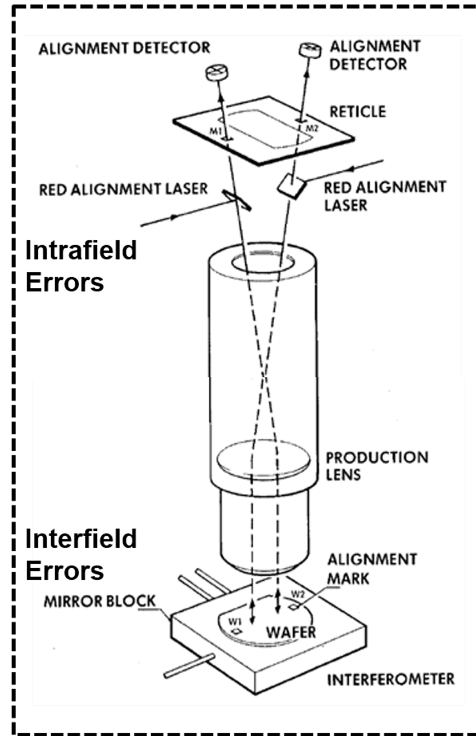

Figure 4.3: Inter/Intra-field error

Apart from layer-to-layer overlay, recent research indicates that cumulative overlay errors across multiple non-neighboring layers also impact device performance. Therefore, achieving a comprehensive multi-layer overlay model and minimizing the overall error necessitates considering stack-up overlay errors Figure 4.4 [21]. The proposed model calculates the stack-up overlay error (Eqn. 4.3) for each layer (k) by summing the overlay error in the current layer with the cumulative stack-up errors from previous layers up to the k_{th} layer [53]. Figure 4.4 illustrates two different scenarios of stack-up errors: a continuous shift towards one side or alternating misalignment as the wafer progresses.

$$
\begin{aligned}
s_x(k) &= s_x(k-1) + o_x(k) \\
s_y(k) &= s_y(k-1) + o_y(k)
\end{aligned}
\tag{4.3}
$$

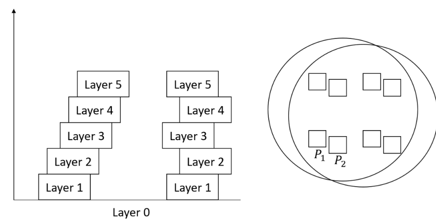

Figure 4.4: Stack-up overlay error propagation

4.3. Robust control and optimal down-selection of markers

Overlay errors and stack-up overlay errors are a critical measure of quality and they are important for the final product performance. Models that relate the tool parameters to overlay errors produce machine settable inputs to minimize those errors, but they are affected by noise and modeling terms not accurately described. For these reasons Gaussianity and independence assumptions of noise terms as well as perfect knowledge of the errors model cannot be satisfied. Ul Haq et al. [61] proposed a methodology for Robust Control of lithography errors that is capable to handle parametric uncertainties in overlay models and not Gaussian noise. In the work the authors define a vector of uncertain parameters delimited by upper and lower bounds and by optimizing the worst-case scenario regarding all the possible uncertainties in the model, a robust multistage control decision can be obtained. In comparison with a non-robust control model, the benefits of robustness increase with higher levels of uncertainty and the progression of the manufacturing process. To feed data into the model overlay error measurements must be performed during the production. Each wafer is characterized by a certain number of markers placed upon the surface where the alignment is assured. Nowadays companies use all of these points to check the right alignment of the wafer dies, but, according to a new study carried by H. Zhang [64], it is possible to use only a few of them and still achieve a good result for the final overlay errors in wafer production. Indeed in [61] the author shows a correlation between residuals of overlay errors model in the lithography process. This means that the errors on the whole layer can be estimated by the measures of just some selected points.

The relationships described above allow to develop a strategical robust measurement point selection model in which the best combination of a given number of measurement points is selected for the robust control of lithography errors, such that the measure of overlay errors at all the candidate measurement points is minimized. The problem is solved by proposing a bi-level robust programming method that minimizes the actual overlay errors for the worst-case scenario regarding main model uncertainties and considers the robustness of control commands. The upper-level optimization task is to select measurement points used for robust control and the lower-level optimization task is to generate the robust control commands based on the data from the selected measurement points. In order to solve this bi-level robust programming model, a nested algorithm is proposed by combining a genetic algorithm with a conservative semidefinite programming approximation and concave quadratic programming, and the robust control command for the worst-case

scenario are obtained. With this method, only a limited number of measurement points are inspected and used to estimate the overlay errors for all the candidate points on the whole layer in order to obtain the control parameters for that layer and to minimize the overall overlay errors for the specific selected pattern of measured points.

4.3.1. Robust Control of Overlay Errors

For layer pattern k of wafer t, let $o^x_{t,k}$ and $o^y_{t,k}$ denote vectors formed by overlay errors in all available measurement markers on the wafer, with $o^x_{t,k}$ denoting errors in the x direction and $o^y_{t,k}$ denoting overlay errors in the y direction on the wafer. In the foundations of control of photolithography overlay errors are Zernike polynomial [41] based models which relate overlay errors to controllable parameters on the photolithography tool via the form

$$\begin{cases} o^x_{t,k} = D^{cx}(u^x_{t,k} + c^x_{t,k}) + e^{ex}_{t,k} + r^x_{t,k} \\ o^y_{t,k} = D^{cy}(u^y_{t,k} + c^y_{t,k}) + e^{ey}_{t,k} + r^y_{t,k} \end{cases} \tag{4.4}$$

where vectors $u^x_{t,k}$ and $u^y_{t,k}$ consist of controllable tool parameters affecting overlay errors in the x and y directions on the wafer, regression matrices D^{cx} and D^{cy} are fully defined by locations of the overlay measurement markers on the wafer, while residual vector terms $r^x_{t,k}$ and $r^y_{t,k}$ account for unmodeled effects and process noise. In general, it is recognized that control of photolithography tools is inherently subject to stochastic actuator uncertainties modeled as

$$\begin{cases} u^x_{t,k} = \bar{u}^x_{t,k} + c^x_{t,k} \\ u^y_{t,k} = \bar{u}^y_{t,k} + c^y_{t,k} \end{cases} \tag{4.5}$$

where $\bar{u}^x_{t,k}$ and $\bar{u}^y_{t,k}$ are vectors of control commands given to the tool, while vectors $c^x_{t,k}$ and $c^y_{t,k}$ model those stochastic actuator uncertainties and are commonly referred to as vectors of process bias terms. Namely, due to exceptionally small scales in which controllable parameters of a photolithography tool reside, unmodeled process dynamics and external noise sources are inevitably significant and cause process bias terms to always be present and continuously change from one wafer to another. A common practice in the industry is to utilize overlay measurements from historical records of previously manufactured wafers to make predictions $\tilde{\hat{c}}^x_{t,k}$ and $\tilde{\hat{c}}^y_{t,k}$ of bias vector term in layer k of wafer t prior to the actual lithography exposure, based on which the relevant control commands $\bar{u}^x_{t,k}$ and $\bar{u}^y_{t,k}$ during the actual exposure can be set to

$$\begin{cases} \bar{u}^x_{t,k} = -\tilde{\hat{c}}^x_{t,k} \\ \bar{u}^y_{t,k} = -\tilde{\hat{c}}^y_{t,k} \end{cases} \tag{4.6}$$

in order to counteract those bias terms. Equation 4.6 describes the so called run-to-run (R2R) paradigm for control of photolithography overlay errors and in its foundations are various methods for dynamic modeling and prediction of bias vector terms, such as Kalman filter-based prediction, various forms of the Exponentially Weighted Moving Average (EWMA) based modeling and prediction methods, or more recently dynamic neural networks [38] and Gaussian Process Regression [30] based approaches. As mentioned earlier, a thorough survey of R2R approaches for control of semiconductor manufacturing processes can be found in [20] and references therein. In this chapter, R2R paradigm will be augmented with considerations of the multistage character of errors in alignment across non-neighboring pattern layers, as well as considerations of robustness of control algorithm performance in the presence of uncertain stochastic terms in the overlay models, as suggested in [61] .Specifically, as we are about to produce pattern layer k on wafer t, let us assume that we have R2R predictions $\tilde{\hat{c}}^x_{t,k}$ and $\tilde{\hat{c}}^y_{t,k}$ of the corresponding bias vector terms, as well as observations of overlay errors of layers 1 through $k-1$ of wafer t measured in a subset of markers defined by a binary vector F_t, which is of the same dimensionality P as the total number of available markers, with one or zero entries in it respectively denoting presence or absence of the corresponding marker in the measurement scheme F_t. Furthermore, following [61] , let us assume that stochastic terms $r^x_{t,k}$,$r^y_{t,k}$,$c^x_{t,k}$ and $c^y_{t,k}$ for layer k of wafer t reside within some upper and lower bounds, which are assumed to be known prior to actual exposure. More precisely, let $r^{ub,x}_{t,k}$ and $r^{ub,y}_{t,k}$ denote upper bounds corresponding on the vector terms $r^x_{t,k}$ and $r^y_{t,k}$ of residuals in the layer-level models and let $r^{lb,x}_{t,k}$ and $r^{lb,y}_{t,k}$ denote the corresponding lower bounds. Later on it will be proposed a statistics-inspired approach for determining these boundaries are based on historical records of overlay errors measured on previously produced pattern layers and wafers, perhaps less formal methods can be used as well, if interested on boundaries please refer to [64] . Furthermore, let $c^{ub,x}_{t,k}$ and $c^{ub,y}_{t,k}$ denote upper bounds corresponding on the bias vector terms $c^x_{t,k}$ and $c^y_{t,k}$ respectively, and let $c^{lb,x}_{t,k}$ and $c^{lb,y}_{t,k}$ respectively denote the corresponding lower bounds. These bounds can be adjusted based on R2R predictions $\tilde{\hat{c}}^x_{t,k}$ and $\tilde{\hat{c}}^x_{t,k}$ of bias vector terms $c^x_{t,k}$ and $c^y_{t,k}$ relevant to pattern k on wafer t, and the corresponding prediction uncertainties.

For measurement scheme F_t, let $o'^x_{t,k}$ and $o'^y_{t,k}$ denote vectors of overlay errors measured in markers F_t on pattern layer k of wafer t. Similarly, let $s'^x_{t,k}$ and $s'^y_{t,k}$ denote vectors of stack-up overlay errors measured in markers F_t on pattern layer k of wafer t. We will

denote this down-selection of overlay and stack-up overlay error measurements as

$$\begin{cases} o'^{x}_{t,k} = F_t \circ o^{x}_{t,k};\ o'^{y}_{t,k} = F_t \circ o^{y}_{t,k} \\ s'^{x}_{t,k} = F_t \circ s^{x}_{t,k};\ s'^{y}_{t,k} = F_t \circ s^{y}_{t,k} \end{cases} \tag{4.7}$$

where $o^{x}_{t,k}$ and $o^{y}_{t,k}$ denote vectors of overlay errors in all markers on pattern layer k of wafer t, while $s^{x}_{t,k}$ and $s^{y}_{t,k}$ denote vectors of overlay errors in all markers on pattern layer k of wafer t. Then, following [61] , for pattern layer k on wafer t, one can use measurements of stack-up overlay errors obtained from markers F_t in the already produced pattern layers 1,2,...,k-1 of wafer t, as well as lower and upper bounds on the uncertain overlay model terms $r^{lb,x}_{t,k}, r^{lb,y}_{t,k}, c^{lb,x}_{t,k}, c^{lb,y}_{t,k}, r^{ub,x}_{t,k}, r^{ub,y}_{t,k}, c^{ub,x}_{t,k}$ and $c^{ub,y}_{t,k}$ to pursue control commands $u^{x*}_{t,k}$ and $u^{y*}_{t,k}$ which robustly minimize a measure of overlay and stack-up overlay errors measured in markers F_t. More precisely, control commands $u^{x*}_{t,k}$ and $u^{y*}_{t,k}$ will be obtained by solving the following optimization problem

$$(u^{x*}_{t,k}, u^{y*}_{t,k}) = \underset{\substack{u^{x}_{t,k}\in\mathbb{R}^{N_x} \\ u^{y}_{t,k}\in\mathbb{R}^{N_y}}}{\arg\min} \max_{\substack{c^{x}_{t,k}, c^{y}_{t,k} \\ r^{x}_{t,k}, r^{y}_{t,k}}} \begin{array}{l} \lambda^x \|o'^{x}_{t,k}\|^2 + \lambda^y \|o'^{y}_{t,k}\|^2 \\ + \alpha^x \|F_t \circ s^{x}_{t,k-1} + o'^{x}_{t,k}\|^2 \\ + \alpha^y \|F_t \circ s^{y}_{t,k-1} + o'^{y}_{t,k}\|^2 \end{array} \tag{4.8}$$

$$subject\ to: \begin{cases} o'^{x}_{t,k} = F_t \circ [D^{cx}(u^{x}_{t,k} + c^{x}_{t,k}) + r^{x}_{t,k}] \\ o'^{y}_{t,k} = F_t \circ [D^{cy}(u^{y}_{t,k} + c^{y}_{t,k}) + r^{y}_{t,k}] \end{cases} \tag{4.9}$$

$$\begin{cases} r^{lb,x}_{t,k} \leq r^{x}_{t,k} \leq r^{ub,x}_{t,k} \\ r^{lb,y}_{t,k} \leq r^{y}_{t,k} \leq r^{ub,y}_{t,k} \end{cases} \tag{4.10}$$

$$\begin{cases} c^{lb,x}_{t,k} \leq c^{x}_{t,k} \leq c^{ub,x}_{t,k} \\ c^{lb,y}_{t,k} \leq c^{y}_{t,k} \leq c^{ub,y}_{t,k} \end{cases} \tag{4.11}$$

One can note that in the objective function 4.8, terms $F_t \circ s^{x}_{t,k-1} + o'^{x}_{t,k}$ and $F_t \circ s^{y}_{t,k-1} + o'^{y}_{t,k}$ describe stack-up overlay errors obtained from markers F_t on pattern layer k of wafer t, while constants $\lambda^x, \lambda^y, \alpha^x$ and α^y are weighting factors which can be used to denote the relative importance of each term in 4.8. Constraint 4.9 expresses the Zernike polynomial-based model 4.4-4.5 for overlay errors $o'^{x}_{t,k}$ and $o'^{y}_{t,k}$ in measurement markers F_t on pattern layer k of wafer t, while constraints 4.10 and 4.11 denote that vector terms $r^{x}_{t,k}, r^{y}_{t,k}, c^{x}_{t,k}$ and $c^{y}_{t,k}$ in the model 4.9 are unknown, but reside within some known lower and upper

bounds. Control algorithm defined by optimization 4.8-4.11 pursues minimization of the worst-case of objective 4.8 which incorporates a measure of overlay and stack-up overlay errors in markers F_t on layer k of wafer t. The worst case of objective 4.8 is assessed regarding uncertain model terms $r^x_{t,k}, r^y_{t,k}, c^x_{t,k}$ and $c^y_{t,k}$ in the overlay errors model 4.9 and an effective method for solving 4.8-4.11 was described in [19] . In the remainder of this section, this ability to solve 4.8-4.11 for any given measurement scheme F_t will be used to obtain measurement schemes which enable best possible performance of the control algorithm 4.8-4.11 under different constraints imposed on the number of measurement markers that are allowed to be retained in the measurement scheme F_t.

4.3.2. Problem Formulation

Given any set of selected overlay measurement markers F_t, we will describe the corresponding control performance $f_{t,k}(F_t)$ at layer k of wafer t using the weighted sum of L2-norms of overlay and stack-up errors in all the candidate markers, as shown below.

$$\begin{aligned} f_{t,k}(F_t) = \lambda^x \|o^x_{t,k}\|^2 + \lambda^y \|o^y_{t,k}\|^2 \\ +\alpha^x \|s^x_{t,k}\|^2 + \alpha^y \|s^y_{t,k}\|^2 \end{aligned} \tag{4.12}$$

Given the robust control commands $u^{x*}_{t,k}, u^{y*}_{t,k}$ obtained for the selected markers F_t using procedure 4.8-4.11, the worst-case performance of the metric 4.12 can be obtained by solving the following optimization problem:

$$J_{t,k}(F_t) = \max_{\substack{c^x_{t,k}, c^y_{t,k} \\ r^x_{t,k}, r^y_{t,k}}} f_{t,k}(F_t) \tag{4.13}$$

$$\begin{cases} o^x_{t,k} = D^{cx}(u^x_{t,k} + c^x_{t,k}) + e^{ex}_{t,k} + r^x_{t,k} \\ o^y_{t,k} = D^{cy}(u^y_{t,k} + c^y_{t,k}) + e^{ey}_{t,k} + r^y_{t,k} \end{cases} \tag{4.14}$$

$$\begin{cases} s^x_{t,k} = s^x_{t,k-1} + o^x_{t,k} \\ s^y_{t,k} = s^y_{t,k-1} + o^y_{t,k} \end{cases} \begin{cases} s^x_{t,0} = 0 \\ s^y_{t,0} = 0 \end{cases} \tag{4.15}$$

$$\begin{cases} r^{lb,x}_{t,k} \leq r^x_{t,k} \leq r^{ub,x}_{t,k} \\ r^{lb,y}_{t,k} \leq r^y_{t,k} \leq r^{ub,y}_{t,k} \end{cases} \tag{4.16}$$

$$\begin{cases} c_{t,k}^{lb,x} \leq c_{t,k}^{x} \leq c_{t,k}^{ub,x} \\ c_{t,k}^{lb,y} \leq c_{t,k}^{y} \leq c_{t,k}^{ub,y} \end{cases} \tag{4.17}$$

where constraints 4.14 and 4.15 respectively calculate overlay and stack-up errors in all the candidate markers, while constraints 4.16 and 4.17 provide lower and upper bounds for the uncertain process bias and residual vector terms in the overlay models. With the worst-case overlay control objective $J_{t,k}(F_t)$ describing performance of any given measurement scheme F_t on layer k of wafer t, and given a constraint that one wishes to use only P_{obj} markers on wafer t, the best-performing set of measurement markers F_t^* on wafer t can be pursued by solving the following optimization problem:

$$F_t^* = argmin_{F_t \in 2^P} \sum_{k=1}^{K} J_{t,k}(F_t) \tag{4.18}$$

$$subject\ to: \sum_{i=1}^{P} F_{t,i} = P_{obj} \tag{4.19}$$

One can observe that objective function 4.18 minimizes the sum of worst-case control performance evaluated on all available candidate markers, across all layers of wafer t, while constraint 4.19 restricts the number of selected markers on wafer t to be P_{obj}. Based on the problem formulation proposed above, let us now express the process of evaluating the objective function in 4.18 for each candidate measurement selection F_t. Given boundaries on the uncertain, stochastic terms for wafer t, the following procedure is conducted, starting with the first layer $k = 1$.

Step 1: For layer k, solve the robust control problem 4.8-4.11 to obtain control commands $u_{t,k}^{x*}$ and $u_{t,k}^{y*}$.

Step 2: For layer k, solve optimization problem 4.13-4.17 and thus characterize worst-case performance of the resulting robust control law, as evaluated on all markers in layer k of wafer t.

Step 3: If $k < K$, obtain the values for stack-up overlays $s_{t,k}^{x}$ and $s_{t,k}^{y}$ corresponding to the worst-case performance $J_{t,k}(F_t)$ and feed them into the next layer $k + 1$. Then, let $k \leftarrow k + 1$ and go the Step 1. If $k = K$, add up layer-specific objectives $J_{t,k}(F_t)$ for all layers and thus obtain the value for the objective function 4.18.

4.3.3. Problem Formulation for the Optimal Number of Measurement Markers

The quality of each layer pattern is evaluated by checking whether overlay errors measured at selected markers are within their specification limits. Given lower Specification Limits (LSL) $o_{t,k}^{x,LSL}, o_{t,k}^{y,LSL}, s_{t,k}^{x,LSL}, s_{t,k}^{y,LSL}$ and Upper Specification Limits (USL) $o_{t,k}^{x,USL}, o_{t,k}^{y,USL}, s_{t,k}^{x,USL}, s_{t,k}^{y,USL}$, the probability that overlay and stack-up overlay errors, observed at each selected marker $i \in F_{t,k}^*(P_{t,k}^{obj})$, are within their specification limits can be calculated as:

$$\Pi(\mathcal{A}i) = R_i^{o^x} \times R_i^{o^y} \times R_i^{s^x} \times R_i^{s^y}, \forall i \in F_{t,k}^*(P_{t,k}^{obj}) \tag{4.20}$$

where

$$\begin{cases} R_i^{o^x} = \Pi(o_{t,k,i}^{x,LSL} \leq o_{t,k,i}^x \leq o_{t,k,i}^{x,USL}) \\ R_i^{o^y} = \Pi(o_{t,k,i}^{y,LSL} \leq o_{t,k,i}^y \leq o_{t,k,i}^{y,USL}) \end{cases} \begin{cases} R_i^{s^x} = \Pi(s_{t,k,i}^{x,LSL} \leq s_{t,k,i}^x \leq s_{t,k,i}^{x,USL}) \\ R_i^{s^y} = \Pi(s_{t,k,i}^{y,LSL} \leq s_{t,k,i}^y \leq s_{t,k,i}^{y,USL}) \end{cases} \tag{4.21}$$

For the sake of simplicity it is assumed that overlay and stack-up overlay errors follow Gaussian distributions

$$\begin{cases} o_{t,k,i}^x \sim \mathcal{N}(\mu_{t,k,i}^{ox}, {\sigma_{t,k,i}^{ox}}^2) \\ o_{t,k,i}^y \sim \mathcal{N}(\mu_{t,k,i}^{oy}, {\sigma_{t,k,i}^{oy}}^2) \end{cases} \begin{cases} s_{t,k,i}^x \sim \mathcal{N}(\mu_{t,k,i}^{sx}, {\sigma_{t,k,i}^{sx}}^2) \\ s_{t,k,i}^y \sim \mathcal{N}(\mu_{t,k,i}^{sy}, {\sigma_{t,k,i}^{sy}}^2) \end{cases} \tag{4.22}$$

where means and variances are calculated as

$$\begin{cases} \mu_{t,k,i}^{ox} = D_i^x(u_{t,k}^{*x} + \mu_{t,k}^{cx}) \\ \mu_{t,k,i}^{oy} = D_i^y(u_{t,k}^{*y} + \mu_{t,k}^{cy}) \\ {\sigma_{t,k,i}^{ox}}^2 = D_i^{x\circ 2}{\sigma_{t,k}^{cx}}^2 + {\sigma_{t,k,i}^{rx}}^2 \\ {\sigma_{t,k,i}^{oy}}^2 = D_i^{y\circ 2}{\sigma_{t,k}^{cy}}^2 + {\sigma_{t,k,i}^{ry}}^2 \end{cases} \begin{cases} \mu_{t,k,i}^{sx} = \mu_{t,k-1,i}^{sx} + \mu_{t,k,i}^{ox} \\ \mu_{t,k,i}^{sy} = \mu_{t,k-1,i}^{sy} + \mu_{t,k,i}^{oy} \\ {\sigma_{t,k,i}^{sx}}^2 = {\sigma_{t,k-1,i}^{sx}}^2 + {\sigma_{t,k,i}^{ox}}^2 \\ {\sigma_{t,k,i}^{sy}}^2 = {\sigma_{t,k-1,i}^{sy}}^2 + {\sigma_{t,k,i}^{oy}}^2 \end{cases} \tag{4.23}$$

The probability that this layer pattern is observed to be a good product can be calculated as

$$\Pi(\mathcal{A}) = \prod_{i \in F_{t,k}^*(P_{t,k}^{obj})} \Pi(\mathcal{A}i) \tag{4.24}$$

which can be called as the **yield rate based on unselected makers**. However, this

evaluation of product quality only focuses on the measured overlay errors. There is still a certain possibility that overlay and stack-up overlay errors at markers that are not measured exceed the specification limits but are not observed. To evaluate this probability, we use the estimation $\hat{c}^x_{t,k}(F^*_{t,k})$,$\hat{c}^y_{t,k}(F^*_{t,k})$ of actual realized process bias terms, obtained with overlay errors measured at selected markers $F^*_{t,k}(P^{obj}_{t,k})$, to establish the estimation $\hat{o}^x_{t,k,j}$,$\hat{o}^y_{t,k,j}$,$\hat{s}^x_{t,k,j}$,$\hat{s}^y_{t,k,j}$of overlay and stack-up overlay errors at each marker $j \in \bar{F}^*_{t,k}(P^{obj}_{t,k})$, where $\bar{F}^*_{t,k}(P^{obj}_{t,k}) = 1 - F^*_{t,k}(P^{obj}_{t,k})$ is the set of markers that are not selected. The probability that overlay and stack-up overlay errors estimated at each unselected marker $j \in \bar{F}^*_{t,k}(P^{obj}_{t,k})$, are within their specification limits can be calculated as

$$\Pi(\mathcal{B}i) = R^{\hat{o}^x}_i \times R^{\hat{o}^y}_i \times R^{\hat{s}^x}_i \times R^{\hat{s}^y}_i, \forall i \in \bar{F}^*_{t,k}(P^{obj}_{t,k}) \tag{4.25}$$

where

$$\begin{cases} R^{\hat{o}^x}_j = \Pi(o^{x,LSL}_{t,k,j} \leq \hat{o}^x_{t,k,j} \leq o^{x,USL}_{t,k,j}) \\ R^{\hat{o}^y}_j = \Pi(o^{y,LSL}_{t,k,j} \leq \hat{o}^y_{t,k,j} \leq o^{y,USL}_{t,k,j}) \end{cases} \begin{cases} R^{\hat{s}^x}_j = \Pi(s^{x,LSL}_{t,k,j} \leq \hat{s}^x_{t,k,j} \leq s^{x,USL}_{t,k,j}) \\ R^{\hat{s}^y}_j = \Pi(s^{y,LSL}_{t,k,j} \leq \hat{s}^y_{t,k,j} \leq s^{y,USL}_{t,k,j}) \end{cases} \tag{4.26}$$

For the sake of simplicity it is assumed that overlay and stack-up overlay errors follow Gaussian distributions

$$\begin{cases} \hat{o}^x_{t,k,j} \sim \mathcal{N}(\mu^{\hat{o}x}_{t,k,j}, {\sigma^{\hat{o}x}_{t,k,j}}^2) \\ \hat{o}^y_{t,k,j} \sim \mathcal{N}(\mu^{\hat{o}y}_{t,k,j}, {\sigma^{\hat{o}y}_{t,k,j}}^2) \end{cases} \begin{cases} \hat{s}^x_{t,k,j} \sim \mathcal{N}(\mu^{\hat{s}x}_{t,k,j}, {\sigma^{\hat{s}x}_{t,k,j}}^2) \\ \hat{s}^y_{t,k,j} \sim \mathcal{N}(\mu^{\hat{s}y}_{t,k,j}, {\sigma^{\hat{s}y}_{t,k,j}}^2) \end{cases} \tag{4.27}$$

where means and variances are calculated as

$$\begin{cases} \mu^{\hat{o}x}_{t,k,j} = D^x_j(u^{*x}_{t,k} + \mu^{cx}_{t,k}) \\ \mu^{\hat{o}y}_{t,k,j} = D^y_j(u^{*y}_{t,k} + \mu^{cy}_{t,k}) \\ {\sigma^{\hat{o}x}_{t,k,j}}^2 = D^x_j \mathbb{E}[\Sigma \hat{c}^x_{t,k}(F_{t,k})] D^{xT}_j + {\sigma^{rx}_{t,k,j}}^2 \\ {\sigma^{\hat{o}y}_{t,k,j}}^2 = D^y_j \mathbb{E}[\Sigma \hat{c}^y_{t,k}(F_{t,k})] D^{yT}_j + {\sigma^{ry}_{t,k,j}}^2 \end{cases} \begin{cases} \mu^{\hat{s}x}_{t,k,j} = \mu^{sx}_{t,k-1,j} + \mu^{\hat{o}x}_{t,k,j} \\ \mu^{\hat{s}y}_{t,k,j} = \mu^{sy}_{t,k-1,j} + \mu^{\hat{o}y}_{t,k,j} \\ {\sigma^{\hat{s}x}_{t,k,j}}^2 = {\sigma^{sx}_{t,k-1,j}}^2 + {\sigma^{\hat{o}x}_{t,k,j}}^2 \\ {\sigma^{\hat{s}y}_{t,k,j}}^2 = {\sigma^{sy}_{t,k-1,j}}^2 + {\sigma^{\hat{o}y}_{t,k,j}}^2 \end{cases} \tag{4.28}$$

Then, we calculate the probability that this layer pattern is a perfect product at all the unselected markers as

$$\Pi(\mathcal{B}) = \prod_{j \in \bar{F}^*_{t,k}(P^{obj}_{t,k})} \Pi(\mathcal{B}i) \tag{4.29}$$

which can be called as the **yield rate based on unselected makers**.

The measurement selection method suggested in this Chapter pursues optimality from a purely quality control aspect. However, any measurement down-selection procedure directly affects cycle-times of the resulting process and the knowledge of yield rate behaviors.

However, this evaluation of product quality only focuses on the measured overlay errors. There is still a certain possibility that overlay and stack-up overlay errors at markers that are not measured exceed the specification limits but are not observed.To evaluate this probability, we use the estimation of actual realized process bias terms, obtained with overlay errors measured at selected markers to establish the estimation of overlay and stack-up overlay errors at each marker.

It can be seen that, with a careful selection of measurement markers, decreasing the percentage of selected markers from 100% to 60% has little influence on the estimation of process bias terms, but the distance metric starts to rise rapidly when the percentage of selected markers continue to decrease. This deviation of estimation leads to a decrease in the accuracy of our understanding of yield rate behavior at unselected markers. Figure 4.5 shows the layer-specific yield rate (a)$\Pi(A)$ based on selected markers, (b) $\Pi(B)$ based on unselected markers and (c) $\Pi(A)\Pi(B)$ based on all the markers. We observe that when the percentage of selected markers decreases, in (a), the yield rate $\Pi(A)$ keeps increasing to one, which indicates that we are less likely to identify bad layer patterns through observations. In (b), the yield rate $\Pi(B)$ continues to decrease, and the amount of change is greater than $\Pi(A)$, because $\Pi(B)$ is not only affected by the percentage of unselected markers, but also affected by the reduced accuracy of the understanding of overlay errors at unselected markers. It means that we are becoming less and less confident about our estimation of overlay errors. Therefore we have to admit that based on our current understanding, the probability of them being bad is increasing. Dominated by $\Pi(B)$, in (c), the yield rate $\Pi(A)\Pi(B)$ at all the markers keeps decreasing as well.

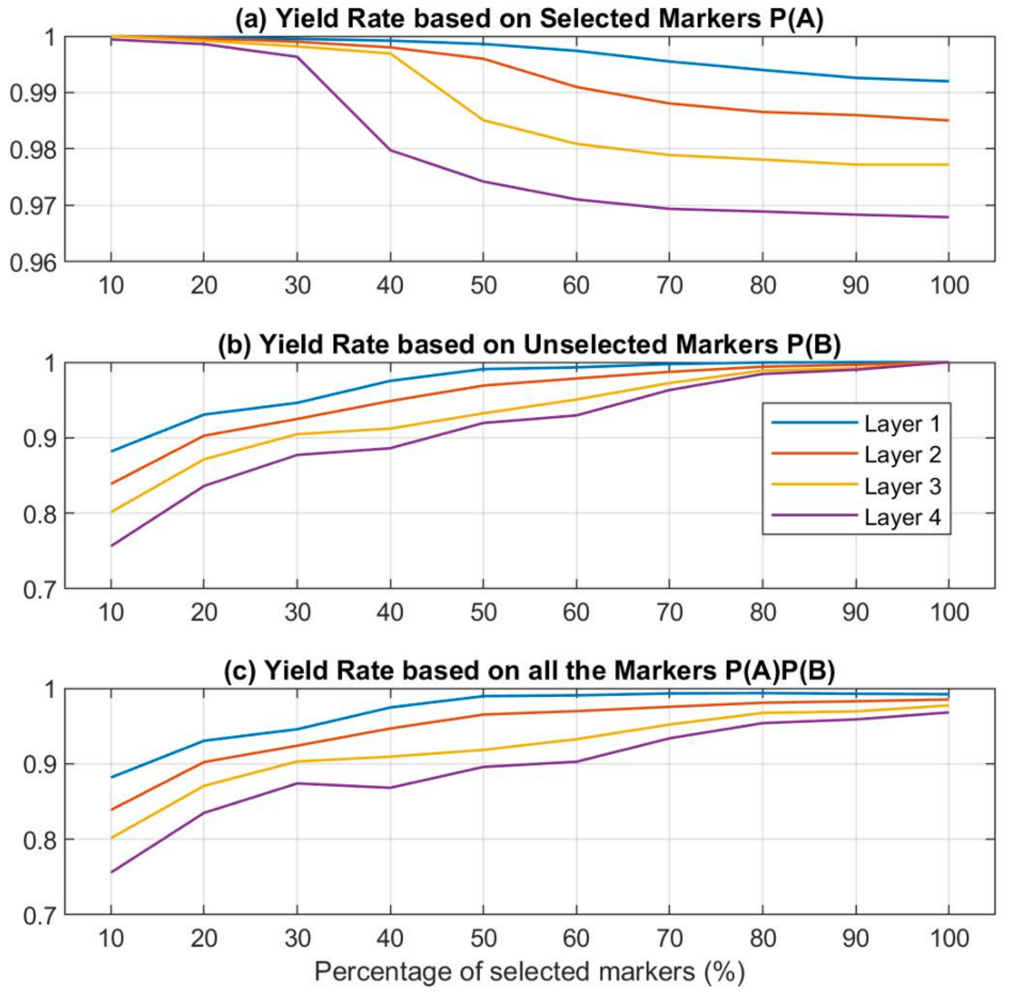

Figure 4.5: Plots of layer-specific yield rate (a) $\Pi(A)$ based on selected markers, (b) $\Pi(B)$ based on unselected markers and (c) $\Pi(A)\Pi(B)$ based on all the markers.

Using yield rate on Selected and Unselected markers in Figure 4.5 it is possible to estimate the quality of the production:

- **P(BD)**: Probability of a Bad wafer is correctly detected and discarded.

$$P(BD) = 1 - P(A) \tag{4.30}$$

- **P(BND)**: Probability that the selected markers are good but the unselected markers are outside boundaries. This will lead to a Misidentified bad layer.

$$P(BND) = P(A) - P(A)P(B) \tag{4.31}$$

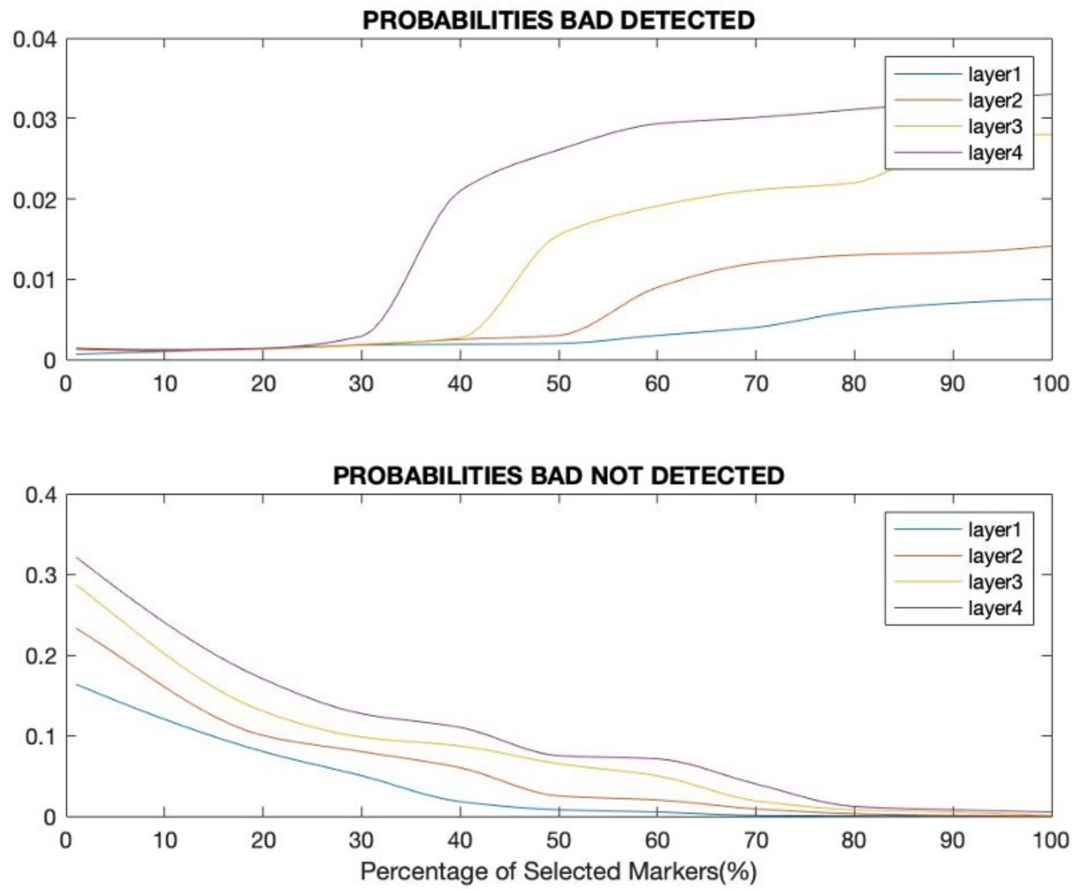

Figure 4.6: Quality probabilities.

5 | Reference System

Due to the precision needed, photolithography is the crux of the IC manufacturing, indeed 33% of wafer manufacturing costs are consumed by it. To produce an entire semiconductor wafer, many such steps are performed subsequently and each pattern transfer has a very precise position on the wafer surface. The alignment of each layer to the previously laid layer is known as overlay and a proper alignment is critical to the quality of the produced devices in order to allow a correct electric current passage in the IC.

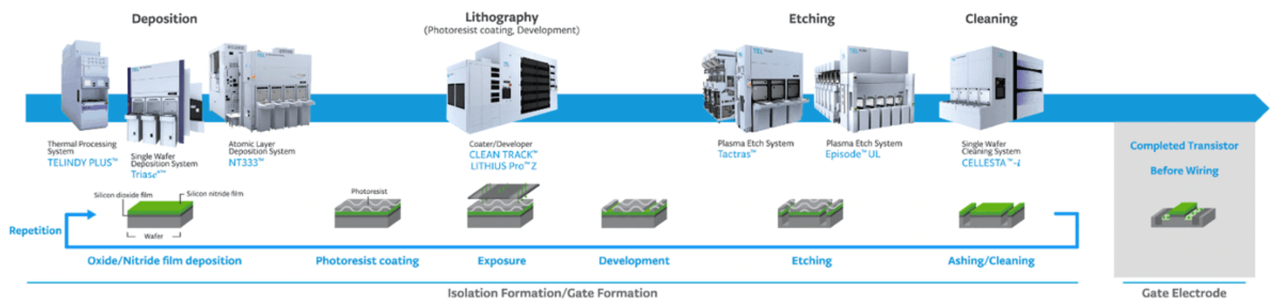

Figure 5.1: Front-end fabrication processes[1]

Photolithography stage is extremely sensible to changes on the production environment, even a single Kelvin could potentially change the photolithography output, or an infinitesimal change of wafer's planarity could lead to a defective electric behaviour of the tiles. To achieve nano-meter scale and be consistent over time it is essential to control not only the behaviour of production machines but to control as well the parameters for each single wafer that has gone through different process conditions. The goal of this thesis is to integrate **product**, **process** and **system-level** models to evaluate the effect at system level, to jointly optimize operations of a semiconductor front-end fabrication line. This three levels are widely studied but the relations among the models still are not explored.

5.1. Product Level

At product level, each stage operates a transformation on the product and may add product deviations, in this case, in form of overlays As we can see from Figure 5.2 Overlay

errors difference between real and design position of layers could lead to a compromised functionality of the Integrated Circuit. In order to model this phenomena it is used theory based on stream of variation, SoV models, firstly introduced in Hu [28] can help in eliminating costly trial-and-error fine-tuning of new product manufacturing processes.

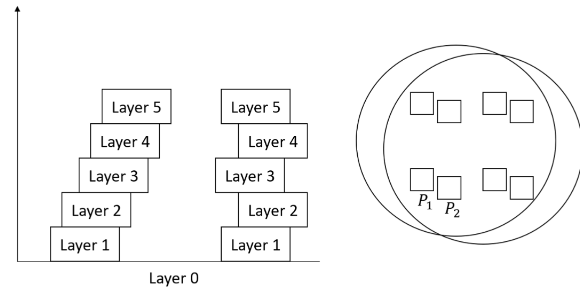

Figure 5.2: Overlay and stuck-up Overlay deviations

5.2. Process Level

It is necessary to define a mathematical model establish an analytical connection between quality errors and process parameters integrating multivariate statistics, control theory as well as manufacturing process knowledge into a unified framework. Thus, for this scope in Zang, Djurdjanovic [64] and previously by Djurdjanovic, Asad [19] used the robust optimization described in Chapter 4 in order to have a R2R control considering not only overlay, but stack-up overlay error, that is described by the summation of the overlay of non-adjacent layers. An optimization framework for the decision-making on the number and selection of measurement markers in photolithography processes, to be able to achieve an acceptable overlay error and at the same time increase the velocity of inspection operation that is time consuming and its cycle time is linearly dependent on the number of markers selected. This lead to a faster production but at the same time decreases the capability of detecting defective layers, that could continue to be processed in the manufacturing line but still with this model it is impossible to address the quality and total output of the manufacturing line.

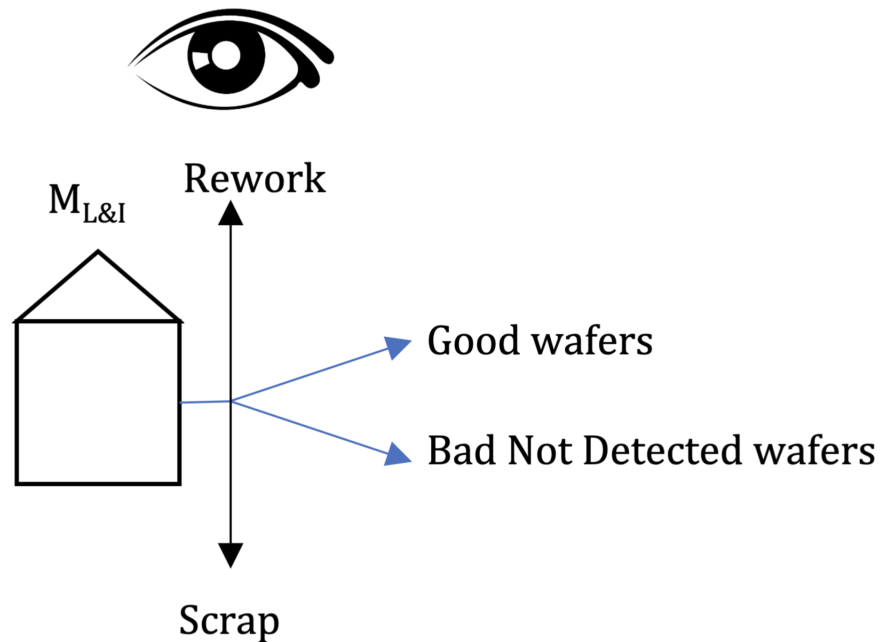

Figure 5.3: External Observer looking at outgoing flow

Process model is addressed in a static way so that it is possible have quality behaviour of the single lithography&inspection stage. This quality behaviour thus is actually decoupled from the real system, it describe the single-stage without considering what happens in upstream machines or how the single-stage influences the downstream machines. So using this approach is not possible to describe the true dynamics of the system. If we set an observer after each inspection station, it can be seen that a part of the inspected wafer would be detected as defective, thus could be scrapped if it is seriously degraded and in previous layers the overlay was nearby boundaries or the wafer could be reworked. The remaining wafers remaining on the manufacturing line are a mixture of Good wafer and defective wafers that are not detected since there is a probability that overlays on markers not inspected could be higher than specifications. From process model and its control it is possible to measure this quantity and estimate the probability that a percentage of inspected wafer is not detected.

5.3. System Level

At this stage Process model that is addressed in a static way for each Lithography&Inspection stages needs to be inserted into a model able to describe system dynamics (starvation&blocking) and quality propagation along the system. Whenever it is selected a percentage of markers that is not at full capacity inspection station could leak defective

wafers as good wafers even if some markers would have been outside boundaries, but they were not measured. Quality propagation is a dynamic behaviour that describe the advancement of defective wafers along the manufacturing thread since this defective wafers were not detected in previous inspection stations. For this purpose could be used different models to better analyse each aspect of dynamic, in this case a continuous Markovian model,system dynamics with a decomposition method based on DDX approach firstly implemented by Dallery Y, David R, Xie X.[63]. Thus with these approaches combined it is possible to have an evaluation of line capacity and its dynamics in terms of influence of different inspection policies combined with the quality measure that allows to characterize quality propagation along production line. As previously, if we set an external observer in this case at last inspection machine, it is possible to notice a part of the flow being discarded or reworked, so the remaining flow is composed by good wafers and a series of defective wafer with bad layers coming from previous inspection stages. So it is possible with the system model to have an estimation of the Probability that some wafer that were defective and not detected in each previous stage and keep track of it, this is helpful in evaluation in scrapping and rework decisions as well as to control propagation of error.

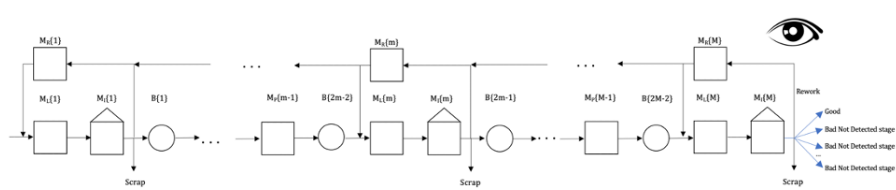

Figure 5.4: External Observer looking at outgoing flow at last machine

5.4. System Scheme

The production system considered is formed by stations which can be photolithography, inspection stations, intermediate process stations and inter-operational buffers in a serial layout with outgoing flows due to detection of bad Layer and exiting flows for operations of rework.

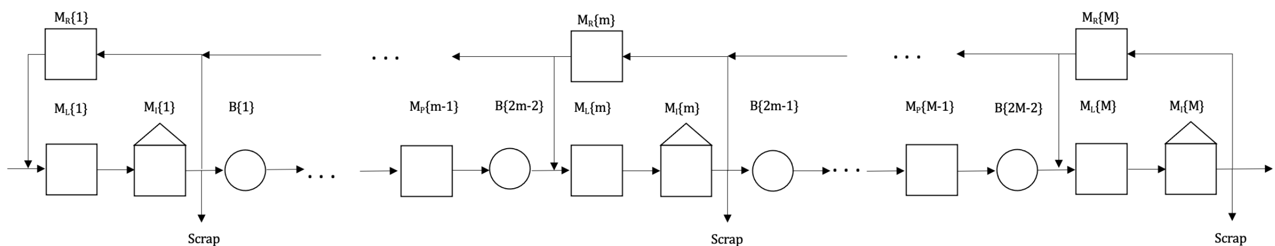

Figure 5.5: Physical Reference System

Each photolithography stage is followed by an inspection station without an inter-operational buffer, therefore these two stages are considered in series and will be considered as an unique stage.
Photolithography stations are the most important machines inside a semiconductors fab, and the goal is that they are working at their best capacity in terms of throughput and quality, these machines are performing a minutely patterned thin film of suitable material over a silicon wafer, flowing in the system.
Inspection stations are those measuring quality of the patterned layers in the upstream manufacturing station through measurement of overlay-errors.
Intermediate process stations are a conglomerate of processes such as photoresist development, etching, cleaning, ion implantation, deposition that are performed in between two photolithography stages.

Both photolithography and inspection stations are fully reliable, no failures occur in both stages, instead intermediate processes have variable production rate, maintenance is needed and local failures can occur.
Those intermediate processes however have lower cycle time than photolithography and inspection stage, moreover total capacity of the machines installed usually is much higher than demanded by photolithography and inspection stages in order to influence as little as possible photolithography stages.
Whenever the inspection station perform the conformity check on the wafer and fine that the patterned layer is defective a parameter tuning of the photolithography station is performed without any delay, in the meanwhile the defective wafer is suddenly unloaded from the inspection machine and rejected from the line or it is sent outside of the line to other rework stations, thus preserve from wasting the capacity of downstream stations in processing wafers that are already defective.
Although, it is important to specify that the tuning could not be perfect thus the next wafer could be defective.
The defective parts could be reworked or entirely scrapped depending on the stage of manufacturing that the error is found, as well of other measures and such as Virtual Metrology that helps to have an estimation of quality measures of photolithography accuracy.
Rework processes are a series of machines external to the manufacturing line that remove the previous layer- through etching process and polishing- after that a post etching inspection is performed to check conformity of previous layer. If it is within specifications the wafer is sent back into the line to photolithography stage, otherwise decision-makers have the option of scrapping the part or rework another layer.

The traditional inspection approach adopted is full inspection(every part that comes out

from the previous manufacturing station will be assessed),moreover the accuracy of the inspection depends on the number of features analyzed thus the inspection time required will increase accordingly.

6 | Problem statement

The goal of this work is to integrate product, process and system-level models to evaluate the effect at system level, to jointly optimize operations.
Indeed, the propagation of multi-stage dynamics has a clear impact on the responsiveness of the quality strategy.
At product level, each stage operates a transformation on the product and may add product deviations, in this case, in form of overlays.
Current process-level control strategies are based on product-level models, as stream of variation, to provide feed-forward control or adaptive quality strategies. Therefore, real importance is given to the investigation of errors propagation in multi-stage manufacturing systems. To ensure the required quality, process control based on models such as SoV was introduced and deeply studied.
Errors propagation is analytically described and incorporated in process control models robust to inaccuracies between process parameters and error generation.
The problem of SoV model-based process control has been assessed in terms of utilizing the measurements obtained up to any given operation, the history of past control actions, as well as model to strategically set controllable process parameters, in such a way that the out-going quality errors are minimized in some sense, without considering their impact on the performance of the system in terms of productivity.
An optimal measurement allocation derived considering only process control can be sub-optimal when considering the multi-stage manufacturing system as a whole. Having more inspections for better modeling of robust control can create bottlenecks and imbalances in the production line flow.
Moreover, the quality addressed by the process control area that looks for the actual magnitude of errors in the features of the product does not consider the quality addressed by the system engineering point of view as the yield and the number of defective parts produced by the system.
Therefore, it is important to evaluate the quality problem both from the process and system point of view.
The need is to integrate model-based process controls in performance evaluation methods

that can highlight their impact on the productivity performances of the line both in terms of productivity indicators (throughput, lead time, WIP) and quality KPIs (yield, effective throughput, defective throughput, scrap rate).
In figure 6.1 on the left the features of a real production system are reported while on the right, the models through which they are integrated in analytical methods are listed.

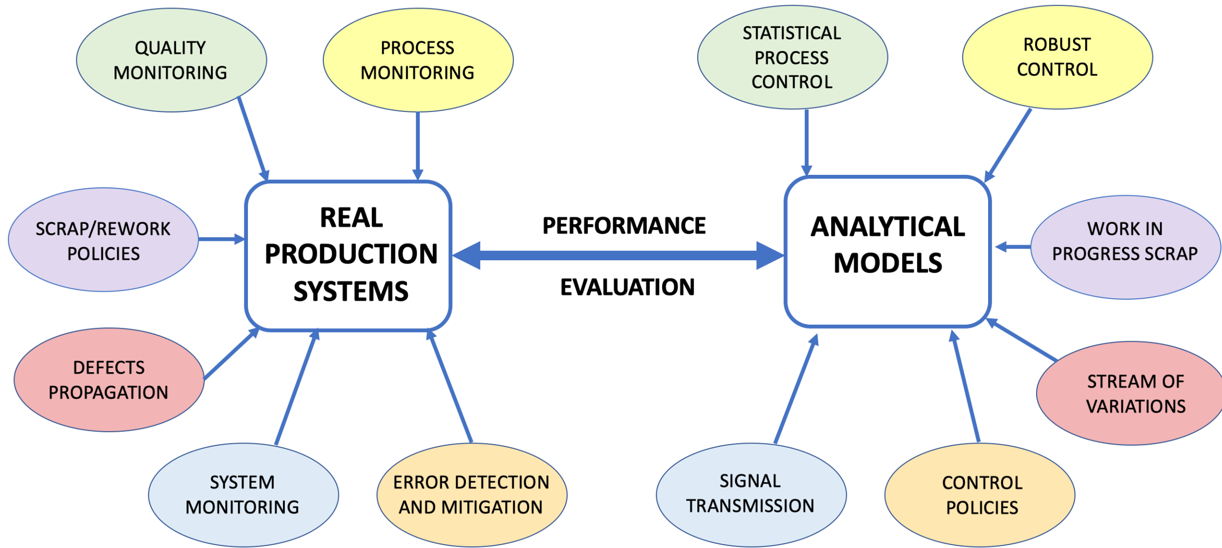

Figure 6.1: Features to be integrated in analytical models to better represent reality

6.1. Research questions

More specifically the proposed thesis attempts to integrate product/process control developed by Zhang [64] with a system model.
The relation between these three control model at different levels is really close but still unexplored.
The ultimate goal is to seek jointly optimized performances of the overall scheme, to see if proposed methodology could improve the traditional semiconductor inspection, that is considered the bottleneck of the manufacturing system.
Indeed this approach could improve not only production KPIs, but could be developed as well to study and optimize system configuration with the right knowledge of system behaviour such as maintenance schedule and its characterization and variation on cycle time of intermediate processes. The outcome of the analysis is to find an optimal trade-off between system productivity and quality of wafers flow.

6.2. Outline of the method

In this thesis, the problem discussed above is considered.
So, to obtain a more realistic model of manufacturing lines, able to integrate process control, we added to the Approximate Analytical Method described in Chapter 7 some new features that better model the impact of quality control to the line performances.
The method presented in the next chapters models with continuous-time mixed-state Markov chains the asynchronous manufacturing line dynamic where quality propagation is determined by quality inspection policy. Additional states are added along the line to consider propagation of quality errors along the manufacturing line.
Instead flow split is treated as a variation of the production rate at buffer level.
Moreover, in the real case, robust model-based control is applied to the process of semiconductor photolithography and integrated into the performance evaluation of the system for an optimal quality inspection policy.

6.3. Innovative features

Given the complexity of the manufacturing line explained in the above chapter it is possible to identify some themes that are not completely covered by existing literature of the proposed method.
Most relevant challenges are:

- Relation between quality measure given by process model and product model, in other words the identification, the single-stage model
- Splitting of flow due to detection of defective wafer and subsequent scrapping in process
- System model for propagation of Inspection error in the case that a lower number of markers are inspected.
- Splitting of flow due to detection of defective wafer and subsequent rework in machines disconnected from the main manufacturing line with a stochastic production rate and subsequent splitting if the previous layer is defective as well.

First three challenges are addressed in this thesis as described in Chapter 7, that represents the innovative features with respect to the state of the art.

7 | Methodology

In this chapter is proposed the formalization of the analytical model implemented to solve the described problem.
A Continuous Deterministic model will be described for performance evaluation of production lines with quality policies and in-process scrapping of parts.
Firstly, the characteristic of the reference system will be analyzed within the assumptions.
Then an analysis on single-stage modeling for the quality control is discussed and the actual formalization of the model with scrap is discussed.
Lastly multi-stage modeling is presented with a two-level decomposition method adopted for performance evaluation of long line.

The goal of this approach is to provide a methodology to address with a Markovian Model the production quality of a manufacturing line, with states that represents the quality status of the parts as *BAD DETECTED*: the defective parts that are scrapped after inspection operations and *BAD NOT DETECTED*: that represent the defective parts measured but not detected by the inspection station and will go through other manufacturing stations.

7.1. Schematic system

The production system considered is modeled,simplifying the system described in Chapter 5 by stations which can be photolithography, inspection stations, inter-operational buffers in a serial layout Figure 7.1.
Each photolithography stage is followed by an inspection station without an inter-operational buffer, therefore these two stages are considered in series and will be considered as an unique stage.
Intermediate process stations are a conglomerate of processes, dynamics of the production line is not affected by intermediate processes that could be a reasonable assumption since usually there is always capacity available and cycle time of single stages is lower than photolithography&inspection stage, so they are removed by the schematic model.

Both photolithography and inspection stations are fully reliable, no failures occur in both stages.
Whenever the inspection station perform the conformity check on the wafer and fine that the patterned layer is defective a parameter tuning of the photolithography station is performed without any delay, in the meanwhile the defective wafer is suddenly unloaded from the inspection machine and rejected from the line, thus preserve from wasting the capacity of downstream stations in processing wafers that are already defective.
Although, it is important to specify that the tuning could not be perfect thus the next wafer could be defective.
The inspection approach adopted is full inspection(every part that comes out from the previous manufacturing station will be assessed),moreover the accuracy of the inspection depends on the n° of features analyzed thus the inspection time required will increase accordingly.

In Figure 7.1 is presented the model reference system that is formed by $M_{L\&I}\{m\}$ that are photolithography&inspection stations that are a single stage since there is not a buffer between them and are consiedere in series.After each stage of photolithography&inspection is located a buffer, so there will be in total $(M-1)$ buffers $B\{m\}$ of finite capacity $N(m)$.

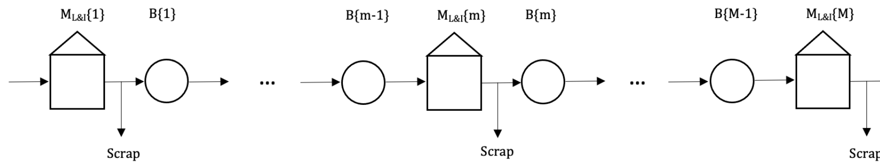

Figure 7.1: Model Reference System

7.2. Assumptions

General Assumptions of the system:

1. The system is asynchronous, machines can start or finish a part at any time instant without synchronization with other stations.
2. The material flow in the system is modeled through a continuous deterministic asynchronous line.
3. The presence of defective parts entering the first machine is not considered.
4. Rejection in-process.
5. The upstream machine is never starved.

6. The downstream machine is never blocked.
7. Each machine has its own deterministic service time.
8. Buffer capacities are finite.
9. The dispatching policy is First In First Out (FIFO).

Assumptions on the machines:

1. Only one part type is produced.
2. Each machines can have different operational mode.
3. Service times include the time to load and unload the part.
4. Machines are reliable.
5. Blocking After Service (BAS) policy.

7.3. Single-stage model

The characterization of the Markovian model is done considering an external observer that is looking at the wafers flowing out from the inspection station and looking at the probability of quality state of the outgoing wafer that could be **good (G)**, **defective & detected (BD)** and the probability that a wafer is **defective & not detected (BND)**, this states represents the **state-space representation**.

Machines are modeled by means of Continuous Time-Discrete State Markov Chains. From probabilities defined in Chapter 4 it is possible to have estimation of wafer's quality thus to create Markovian Model of the conglomerate state Lithography and Inspection, with states-space that represents the quality status of the parts as:

- **GOOD-(G)**: represent correct status of the production.
- **BAD DETECTED-(BD)**: the defective wafers that are scrapped after inspection operations.
- **BAD NOT DETECTED-(BND)**: represent the defective parts measured but not detected by the inspection station and will go through other manufacturing stations.

Production rate

Litho&Inspection stage is considered as two machines in series, thus cycle time is the summation of both Lithography and Inspection times. Cycle time of lithography is considered equal for each stage.

$$CT^{\{m\}} = CT_{litho} + CT^{\{m\}}_{inspection}$$

Inspection time as it is shown, depends almost linearly on the number of Markers observed. So the more markers, control commands decide to select, the more CT will increase and decrease overall productivity. Moreover it is the time relation between inspection and lithography cycle time as:

$$CT^{\{m\}}_{inspection} = 1.5 \cdot CT_{litho} \cdot \%markers^{\{m\}}$$

for completeness we define production rate for each state of the machine $\mu^{\{m\}}$ as:

$$\mu^{\{m\}} = \frac{1}{CT^{\{m\}}} \; \frac{[wafers]}{[t.u.]}$$

Since each Markovian State represent quality state of the part and it machine is considered fully reliable each state have production rate that characterize machine.

$$\mu^{\{m\}}(S[G]) = \mu^{\{m\}}(S[BND]) = \mu^{\{m\}}(S[D]) = \mu^{\{m\}}$$

Transition Rates

Machines are considered reliable but have different states of production quality. According to a new study carried by H. Zhang [64] it is possible to have an estimation of the probabilities of the states described above, so transition are calculated as a function of probabilities considered in Figure 4.5. Moreover each transition- is important to notice- is considered time dependent.

The transition from G and BND to BD are equal and is calculated as in

$$q^{\{m\}}_{G\to BD} = q^{\{m\}}_{BND\to BD} = \frac{P(BD)^{\{m\}}}{CT^{\{m\}}} \tag{7.1}$$

meaning that the mean time to move to BD is $P(BD)^{\{m\}^{-1}} \cdot CT^{\{m\}}$

The transition from G and BD to BND are equal and is calculated as in

$$q^{\{m\}}_{G \to BND} = q^{\{m\}}_{BD \to BND} = \frac{P(BND)^{\{m\}}}{CT^{\{m\}}} \tag{7.2}$$

meaning that the mean time to move to BND is $P(BND)^{\{m\}^{-1}} \cdot CT^{\{m\}}$

The transition from BND to G is calculated as in

$$q^{\{m\}}_{BND \to G} = \frac{1 - P(BND)^{\{m\}} - P(BD)^{\{m\}}}{CT^{\{m\}}} \tag{7.3}$$

Instead the transition from BD and BND to G are equal to production rate, since after each cycle they could move to a good production.

$$q^{\{m\}}_{BD \to G} = \frac{1}{CT^{\{m\}}} = \mu^{\{m\}} \tag{7.4}$$

Transition rate matrix for Machine m:

$$Q\{m\} = \begin{bmatrix} 0 & q^{\{m\}}_{G \to BD} & q^{\{m\}}_{G \to BND} \\ q^{\{m\}}_{BD \to G} & 0 & q^{\{m\}}_{BD \to BND} \\ q^{\{m\}}_{BND \to G} & q^{\{m\}}_{BND \to BD} & 0 \end{bmatrix};$$

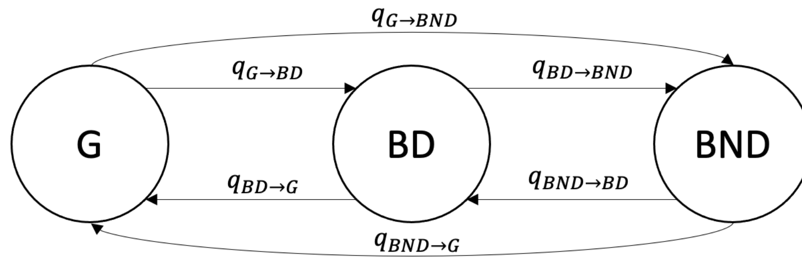

Figure 7.2: Markov Model

7.3.1. MC: Remark for down states

The model represented in this chapter and followings is extended integrating as well down states that could represent starvation characteristics of the intermediate processes, or maintenance and set-up processes, Figure 7.3. In here is presented the Single-stage with down states, Multi-stage evaluation model presented in Section 7.5 is valid for this methodology as well.

It is important to underline that since the knowledge about these intermediate processes

and its characteristics are strictly protected by manufacturer's privacy policies it is considered to delimit the case study and the formalization described in this chapter not considering these down states.

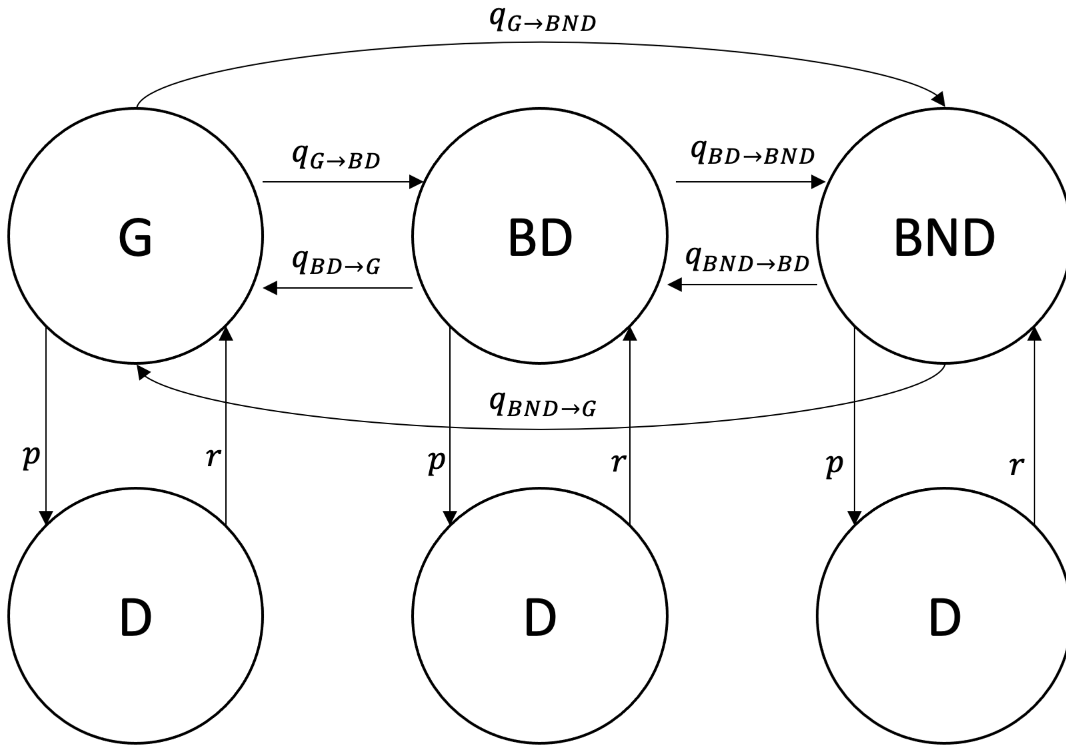

Figure 7.3: Single-Stage model with down states.

7.4. 2M1B Line

In this section a 2M1B line is considered to present the modeling assumptions for in-line rejection and a detailed analysis of quality propagation along production stages as well as buffer dynamics. In Figure 7.4 is presented a 2M1B line extrapolated from the whole model. It is considered here the most relevant case where $\mu^u > \mu^d$, a decreasing velocity profile. The other scenarios where $\mu^u <= \mu^d$ lead to a Building Block with absent buffer dynamics, it would be always empty because downstream machine does not have downstates.

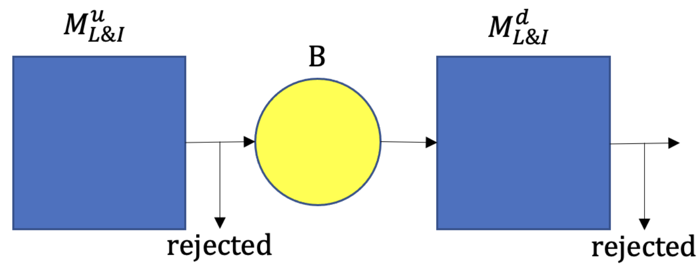

Figure 7.4: 2M1B Line

If it is set an external observer at intermediate buffer it is evident that whenever upstream machine detect a defective wafer this will not be sent to the buffer and it is immediately rejected, so the apparent production rate of state Bad Detected is set at zero even though from perspective of upstream single stage it is processing parts $\mu^u(BD) = 0$.

In Figure 7.5, is presented the continuous-time mixed continuous- and discrete-semi-Markov Chain for the Building Block.The set of states characterizing each pseudo-machine is enlarged as follows.

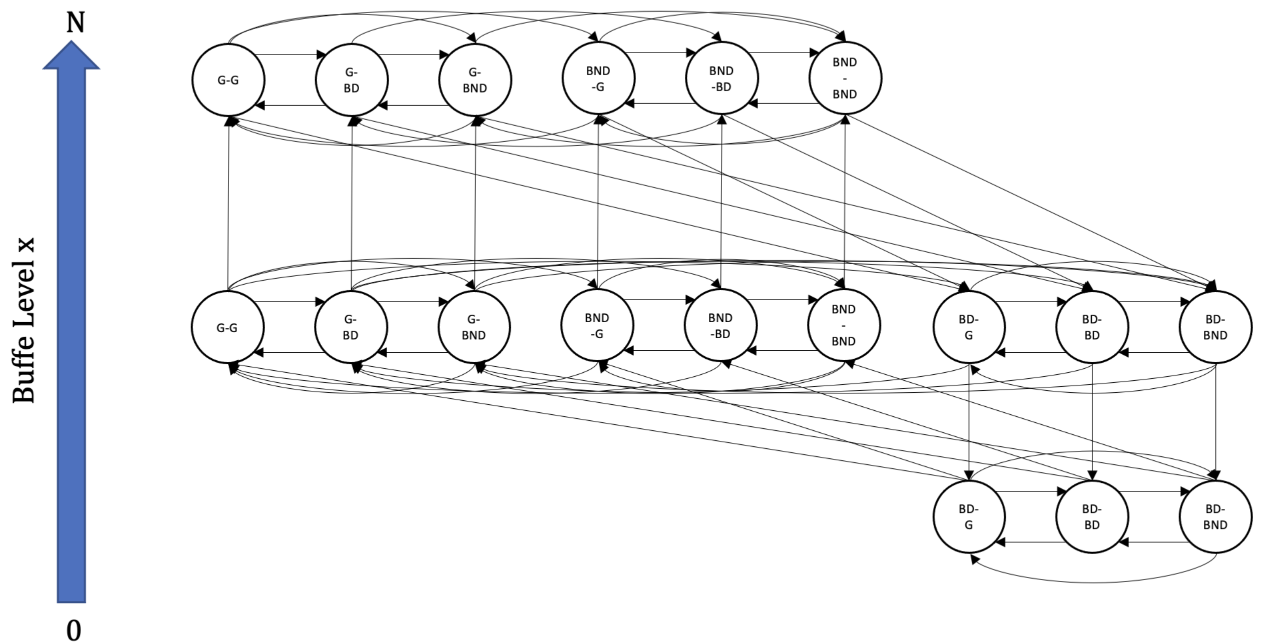

Figure 7.5: Building Block

In Figure 7.5, is presented the continuous-time mixed continuous- and discrete-semi-Markov Chain for the Building Block.The set of states characterizing each pseudo-machine is enlarged as follows.

7.4.1. Quality propagation

It is possible to notice from Figure 7.5 that some states that in upstream machine do not detected the bad production will add a layer on top of the bad not detected one and propagate along the line without noticing them. The states above described are $S_{NQ} = \{(BND, G); (BND, BD); (BND, BND)\}$ considering them at all level, with and without buffer limitation. A state is added in state-space of downstream machine representing the condition of being producing a wafer that still have a defective layer, therefore the abbreviation NQ "Non-Quality" is used.

7.4.2. Starvation

A new state is added to the state-space of the upstream machine, representing the condition of being 'blocked', therefore the abbreviation B is used. This state communicates with the rest of the system by means calculated transitions that will be better explained in next sections.

7.4.3. Blocking

A new state is added to the state-space of the downstream machine, representing the condition of being 'starved', therefore the abbreviation S is used. This state communicates with the rest of the system by transitions.

7.4.4. Evaluation

Steady-state probabilities

The steady-state probabilities $\Pi(S(m))$ of the joint states (S^u, S^d) are found by integrating the probability density function $f(S(m))$ on the modeled buffer variable $x, x = [0; N]$. Indeed, the model provides the complete mapping of the system states in terms of joint steady-state probabilities. The vector of steady-state probabilities is computed as follows:

$$\Pi((S^u, S^d)) = \int_0^N f(S(m))dx \tag{7.5}$$

Probability flow matrix

The probability flow vector $g(S(m))$ represents the probability flow exiting a state belonging to the Markov Chain. It is defined as:

$$g(S(m)) = \nu(S(m)) \cdot f(S(m)) \tag{7.6}$$

It is convenient to associate the probability flow $g(S(m))$ in relation to the destination state. Therefore, let $G(m)$ be the probability flow matrix of Building Block $BB(m)$ defined as:

$$G(m) = [g(S(m)_{IN} \to S(m)_{OUT})] \tag{7.7}$$

7.5. Multi-stage evaluation model

The objective of the approach presented herein is to provide a method to accurately evaluate the steady-state performance of multi-stage asynchronous manufacturing systems. To extend the method to long lines the decomposition approach is used,as presented in [45]. Each machine in the line can be described as Integrated Machine $M\{m\}m = 1, ..., M$ Similarly, for each buffer in the line $B(m), m = 1, ..., M - 1$, a two-machine line BB is built.

Integrated Machines give a representation of the line centered in machines, while $BB(m)$ represent the line centered in the buffer m. Their solution must be the same and also BBs must be coherent with each other.

Therefore, the overall solution of the system evaluation comes from a linearized system of equations where joint influences among Building Blocks are computed and Integrated Machines are characterized. Hence, the characterization of Integrated Machines is used to link one Building Block to another, in order to guarantee the homogenization of the performance evaluation. In Figure 7.6 is presented a scheme of a long line considered in the following discussion.

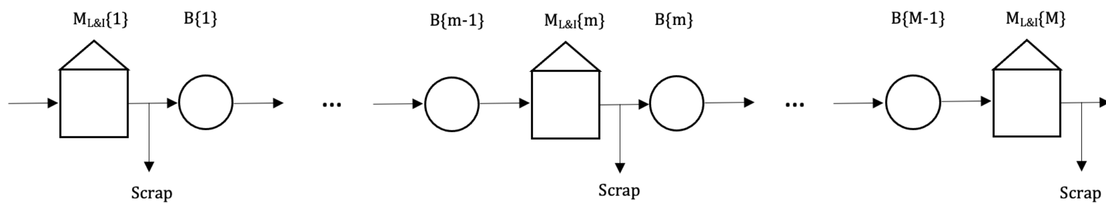

Figure 7.6: Long Line

7.5.1. Two-level decomposition

The multi-stage extends the work presented by Magnanini, Tolio [45]. It has been highlighted that a manufacturing system is a combination of controlled dynamics coming from the joint influence of the resources. Indeed, each machine has its own behavior that may - or may not - influence other machines in the line.

The propagation of effects do overlap, and makes the overall system complex to evaluate.

Buffers have the role of mitigating the propagation effects along the line by acting as filters.
In fact, buffers introduce a delay in the propagation of limiting phenomena, such as starvation or blocking, whereas machines introduce a delay in the propagation of the resumption of flow, once the limiting condition has finished.In order to do so, the model takes two different viewpoints on the manufacturing system. In Fig.7.7 is presented the schematic representation of the proposed methodology.

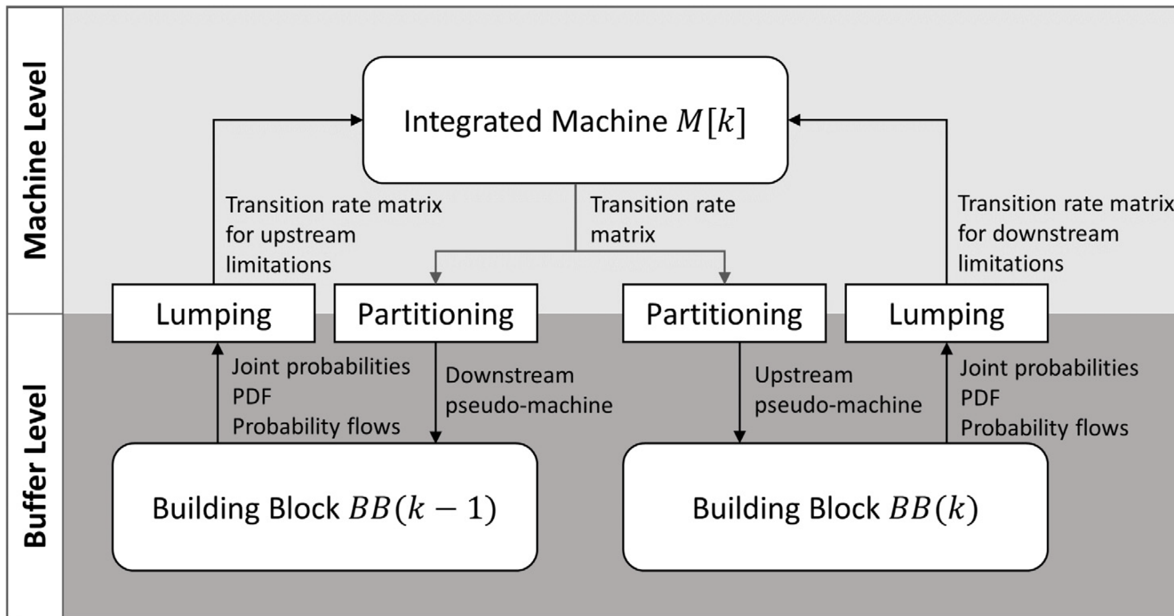

Figure 7.7: Schematic representation of the proposed method.

Machine Level

At machine level, the Integrated Machines are characterized. Each Integrated Machine is modeled by a continuous-time discrete-state Markov Chain that represents the resulting dynamics of the original machine once it has been inserted in the system. Therefore it adds to the behavior of the original machines the states originated from a limited or controlled behavior.

Buffer Level

At buffer level, the Building Blocks model the inflow and outflow of each buffer with respect to the machines in the line. Each Building Block is modeled by a continuous-time mixed continuous- and discrete-state Markov Chains.

7.5.2. Integrated machine

An *Integrated machine* $M[m]$ takes information regarding upstream phenomena limiting it from $BB(m-1)$ and information regarding down-stream phenomena from the $BB(m)$. Therefore, the characterization of the Integrated Machine $M[m]$ grounds on the evaluation of the two adjacent Building Blocks $BB(m-1)$ and $BB(m)$.

Moreover, equations have to be developed in order to derive transition rates at machine-level from the controlled transitions at buffer-level, and fully characterize the Integrated Machine. As stated above, the Integrated Machine $M[m]$ represents the correspondent original machine $M\{m\}$ once it has been inserted in the system. It adds to the behavior of the original machine in isolation, named Local states $L[m]$, three state partitions:

- The remote Starvation states $S[m]$ represent the states in which the Integrated Machine $M[m]$ is upstream limited and they are defined based on the dynamics which are explicit in Building Block $BB(m-1)$.
- The remote Blocking states $B[m]$ represent the states in which the Integrated Machine $M[m]$ is downstream limited and they are defined based on the dynamics which are explicit in Building Block $BB(m)$, they could comprehend both slowdown and blocking phenomena.
- The Non-Quality states $NQ[m]$ represent the states in which the Integrated Machine $M[m]$ is processing defective layers that were still defective in previous machines and sending them to the downstream stage $BB(m)$.
  The remote Bad Not Detected States $NQ[m]$ represents a measure of quality of production from Building Block $BB(m)$, the probability that the upstream machine start producing a bad not detected layer is propagated through buffer $B(m)$ to machine $M[m]$, thus production stage M will have $M-1$ NQ states.
  Each single NQ state is a state lumping of different remote and local states since whenever upstream machine is in state BND.

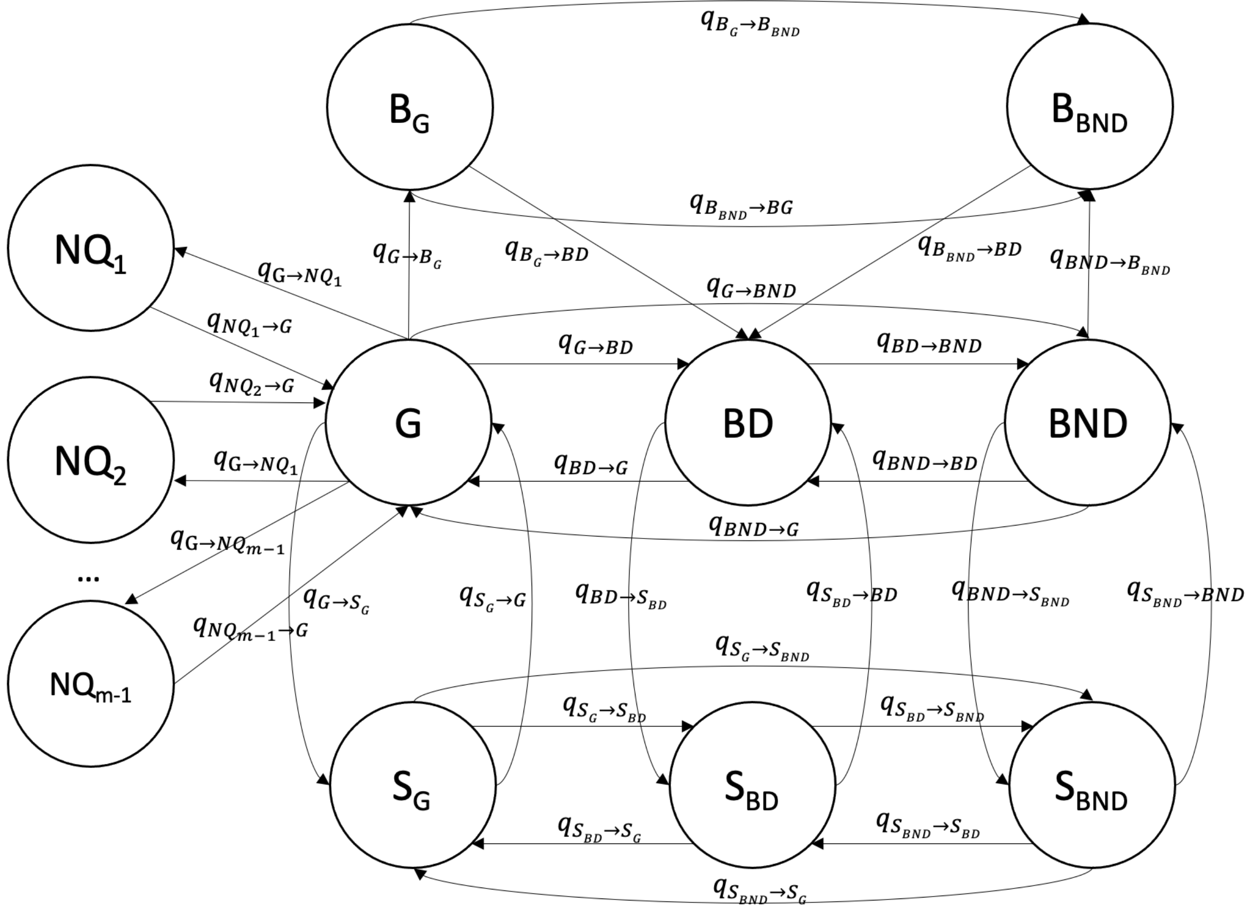

Figure 7.8: General Integrated Machine

7.5.3. Building block

The Building Block $BB(m)$ is a two-machine one-buffer line representing the inflow and outflow of the overall system centered on the considered buffer.
The inflow is modeled using the upstream pseudo-machine $M^u(m)$, and the outflow is modeled using the downstream pseudo-machine $M^d(m)$.
The state $S(m)$ of the identified Building Block $BB(m)$ is represented by the triplet $S(m) = (x, S^u, S^d)$, where x is a continuous variable representing the buffer level, S^u is a discrete variable representing the states of the upstream pseudo-machine and S^d is a discrete variable representing the states of the downstream pseudo-machine.
The duplet (S^u, S^d) denotes the joint machine states.
Hence, the total number of joint machine states includes:

- The joint machine states when no limitation occurs and no quality propagation: $S^u_{(B,BND,NQ)} \otimes S^d_{(S,NQ)}$ where $S^u_{(B,BND,NQ)}$ denotes all possible upstream states ex-

cluding the blocking state B, the upstream Bad Not Detected state BND as well as upstream states of Non-Quality Propagation NQ , $S^d_{(S,NQ)}$ denotes all possible downstream states excluding the starvation state S as well as upstream states of Non-Quality Propagation NQ and $\otimes$ denotes the Kronecker product.

- The joint machine states when downstream limitations occur, i.e. the upstream machine is blocked: $B \otimes S^d_{(S,NQ)}$
- The joint machine states when upstream limitations occur, i.e. the downstream machine is starved: $S^u_{(B,BND,NQ)} \otimes S$.
- The joint machine states that represents the states of bad production in *current* stage: $S^u_{BND} \otimes S^d$
- The joint machine states that represents the states of bad production in *previous* stages: $S^u_{NQ} \otimes S^d$

According to the selected Building Block model, the solution method returns the steady state probability density functions $f(x, S^u, S^d)$ (PDF) for each joint machine state (S^u, S^d) as a function of the buffer level x.
The proposed method is valid if other similar two-stage models are considered, which are based on the Markovian representation of machines, and for which the solution method returns the PDF.
Based on the PDF, the following main output can be computed:
the steady-state probabilities of the joint machine states $\pi(S^u, S^d)$ as:

$$\pi(S^u[z], S^d[j]) = \int_0^N f(x, S^u, S^d)dx \quad \forall z \forall j \tag{7.8}$$

and the boundary probability flows between joint machine states as a function of the empty or full buffer levels, $G(0, S^uS^d \rightarrow S^uS^d)$ and $G(N, S^uS^d \rightarrow S^uS^d)$ respectively, as:

$$G(0, S^uS^d \rightarrow S^uS^d) = (\mu(S^u) - \mu(S^d)) \cdot B_2 \cdot f(0, S^u_{(B)}[z], S^d[j])dx \quad \forall z \forall j \tag{7.9}$$

$$G(N, S^uS^d \rightarrow S^uS^d) = (\mu(S^u) - \mu(S^d)) \cdot B_2 \cdot f(N, S^u[z], S^d_{(S)}[j])dx \quad \forall z \forall j \tag{7.10}$$

where B_2 is a Boolean matrix that defines the possible boundary transitions between joint machine states.

7.5.4. From buffer-level to machine-level: Lumping

The objective of this step is to characterize the integrated machine based on the output provided by the building block solution, in particular, (i) characterization of the state space and (ii) characterization of the transition rate matrix.

7.5.5. State space and state probabilities

The state space of each Integrated Machine $M[m]$ is defined from the corresponding machine in isolation $M\{m\}$ and from the Building Blocks $BB(m-1)$ and $BB(m)$.

$$L^{[m]} = S\{m\} \tag{7.11}$$

$$B^{[m]} = B(m) \otimes S^{d}_{(S,NQ)}(m) \tag{7.12}$$

$$S^{[m]} = S^{u}_{(B,BND,NQ)}(m-1) \otimes S(m-1) \tag{7.13}$$

$$\begin{cases} NQ^{[m]} = \cup(S^{u}_{BND}(m-1) \otimes S^{d}_{NQ}(m-1); S^{u}_{NQ}(m-1) \otimes S^{d}(m-1)) \ \forall m > 2; \\ NQ^{[2]} = \cup(S^{u}_{BND}(m-1) \otimes S^{d}_{NQ}(m-1)) \ m = 2 \end{cases} \tag{7.14}$$

Since the Markovian model have many number of operational stages that increases along the line because of states NQ the state space could become extremely large, thus leading to problems of state explosion.
This problem can be addressed by means of lumping. From the output of the building blocks, the steady-state probabilities of the states of the integrated machine $M[m]$ can be computed through partial lumping:

$$\Pi_L[i] = \sum_{j} \pi_{(m)}(S^u[i], S^d[j]) \quad \forall i \tag{7.15}$$

$$\Pi_L[i] = \sum_{z} \pi_{(m-1)}(S^u[z], S^d[i]) \quad \forall i \tag{7.16}$$

$$\Pi_B[j] = \pi_{(m)}(B, S^d[j]) \quad \forall j \tag{7.17}$$

$$\Pi_S[z] = \pi_{(m-1)}(S^u[z], S) \quad \forall z \tag{7.18}$$

$$\Pi_{NQ}[m-1] = \sum_{l} \pi_{(m-1)}(S^u[BND], S[l]) \quad \forall l \tag{7.19}$$

We can observe that the steady-state probabilities of local states in $M[m]$ can be derived equally from both $BB(m-1)$ and $BB(m)$, where the original machine $M\{m\}$ is included either in the downstream pseudo-machine $M^d(m-1)$ or the upstream pseudo-machine $M^u(m)$.
In contrast, the steady-state probabilities of the blocking states in the Integrated Machine $M[m]$ are derived from the Building Block $BB(m)$, where the original machine $M\{m\}$ is downstream limited by the remainder of the system.
Similarly, the steady-state probabilities of the starvation states in the Integrated Machine $M[m]$ are derived from the Building Block $BB(m-1)$, where the original machine $M\{m\}$ is upstream limited by the remainder of the system.

7.5.5.1. Transition rate matrix & Production Rates

At the machine-level, the goal is to define the transition rates in order to have a complete characterization of the transition rate matrix $Q^{[m]}$ belonging to the continuous-time discrete-state Markov Chain of each Integrated Machine $M[m]$. Let us recall the definition of the transition rate matrix $Q^{[m]}$:

$$Q^{[m]} = \begin{bmatrix} Q_{LL} & Q_{LS} & Q_{LB} & Q_{LNQ} \\ Q_{B} & Q_{SS} & Q_{SB} & Q_{SNQ} \\ Q_{BL} & Q_{BS} & Q_{BB} & Q_{BNQ} \\ Q_{NQL} & Q_{NQS} & Q_{NQB} & Q_{NQNQ} \end{bmatrix}; \tag{7.20}$$

In the following, the decomposition equations used to define the sub-matrices are introduced.
There are two sets of equations: the first defines the transition rates to enter and exit the limiting states, and represents the computation of transition rates at the machine level from the controlled transitions at the buffer level, and the second defines the transition rates among remote states.

Entering and exiting the limiting states

This set of equations defines the transition rates for entering and exiting the limiting states. These equations are based on the balance equations for the continuous-time Markov chain (CTMC).
The balance equations are based on probability flow, and it is generally computationally intractable to solve this system of equations for most queuing models [44]. However, in

this case the output from the Building Block evaluation contributes to the solution of these equations.
The corresponding transition rate matrices can be computed as:

$$Q_{LS}^{[m]} = G_{LS}(m-1) \odot [\Pi_L(m-1)]^{-1} \tag{7.21}$$

$$Q_{LB}^{[m]} = G_{LB}(m) \odot [\Pi_L(m)]^{-1} \tag{7.22}$$

$$Q_{B}^{[m]} = Q_{LS}^{[m]} \cdot [\Pi_S(m-1)][\Pi_L(m-1)]^{-1} \tag{7.23}$$

$$Q_{BL}^{[m]} = \begin{bmatrix} 0 & q_{B_G \to BD}(m) & 0 \\ 0 & q_{B_{BND} \to BD}(m) & 0 \end{bmatrix}; \tag{7.24}$$

Entering and exiting the Non-Quality states

This set of equations defines the transition rates for entering and exiting the states that defines the quality state of bad not detected wafers along the line.
In order to simplify the Markovian model of the Integrated machine $IM[m]$ it was performed, lumping of NQ states have only one transition to Good state even though theoretically transition could occur in every other state.
Since the other transitions are much less frequent and the probability of the other states are many orders of magnitude inferior, the flux of probability going to other states would have been negligible.
Transition rate matrix for NQ is described as:

$$Q_{NQL}^{[m]} = \begin{bmatrix} q_{BND \to G}(1) & 0 & 0 \\ ... & ... & ... \\ q_{B_{BND} \to G}(m-1) & 0 & 0 \end{bmatrix}; \tag{7.25}$$

$$Q_{LNQ}^{[m]} = \begin{bmatrix} \frac{q_{BND \to G}(1) \cdot [\Pi_{NQ(1)}(1)]}{[\Pi_G(1)]} & ... & \frac{q_{BND \to G}(m-1) \cdot [\Pi_{NQ(m-1)}(m-1)]}{[\Pi_G(m-1)]} \\ 0 & ... & 0 \\ 0 & ... & 0 \end{bmatrix}; \tag{7.26}$$

Transitions among limiting states

This set of equations define the transition rates among limiting states of the same type

therefore but with different number of states,

$$Q^{[m]}_{SS} = Q^{[m]}_{LL} \tag{7.27}$$

$$Q^{[m]}_{BB} = Q^{[m]}_{LL} \tag{7.28}$$

$$Q^{[m]}_{NQNQ} = Q^{[m]}_{NQS} = Q^{[m]}_{NQBL} = [0] \tag{7.29}$$

$$Q^{[m]}_{SB} = Q^{[m]}_{SNQ} = [0] \tag{7.30}$$

$$Q^{[m]}_{BS} = Q^{[m]}_{BNQ} = [0] \tag{7.31}$$

Production Rates

Due to state lumping, both in NQ and B partition needs to be updated production rate since new state is the sum of many others with different characteristics.

States that are producing and delivering bad detected layers $S_{NQ_{m-1}} = \{(BND - G), (BND - BD), (BND - BND)\}$ are lumped into a single state called NQ_{m-1} and its production rate is scaled considering that the state $(BND - BD)$ is no-productive from point of view of $BB(m + 1)$ because it will discard that wafer.

$$\mu_{NQ_{m-1}} = \frac{\sum \mu(S_{NQ_{m-1}}) \cdot \Pi(S_{NQ_{m-1}})}{\sum \Pi(S_{NQ_{m-1}})} \tag{7.32}$$

The downstream limitations has been lumped since with long lines the number of state would increase exponentially, more specifically B_G is the lumping of boundary states $S_{B_G} = \{(G - G), (G - BD), (G - BND)\}$ and B_{BND} is the lumping of boundary states $S_{B_{BND}} = \{(BND-G), (BND-BD), (BND-BND)\}$. Their production rate are scaled considering that the states have different production rates thus:

$$\mu_{B_G} = \frac{\sum \mu(S_{B_G}) \cdot \Pi(S_{B_G})}{\sum \Pi(S_{B_G})} \tag{7.33}$$

$$\mu_{B_{BND}} = \frac{\sum \mu(S_{B_{BND}}) \cdot \Pi(S_{B_{BND}})}{\sum \Pi(S_{B_{BND}})} \tag{7.34}$$

7.5.6. From machine-level to buffer-level: Partitioning

Based on the characterization of the machine level, the input to the buffer level can be defined in terms of the state space and transition rate matrix of the pseudo-machines for

each building block $BB(m)$.
A schematic representation of the relation between the pseudomachines at buffer-level and the Integrated Machines at machine-level is provided Fig. 7.9.

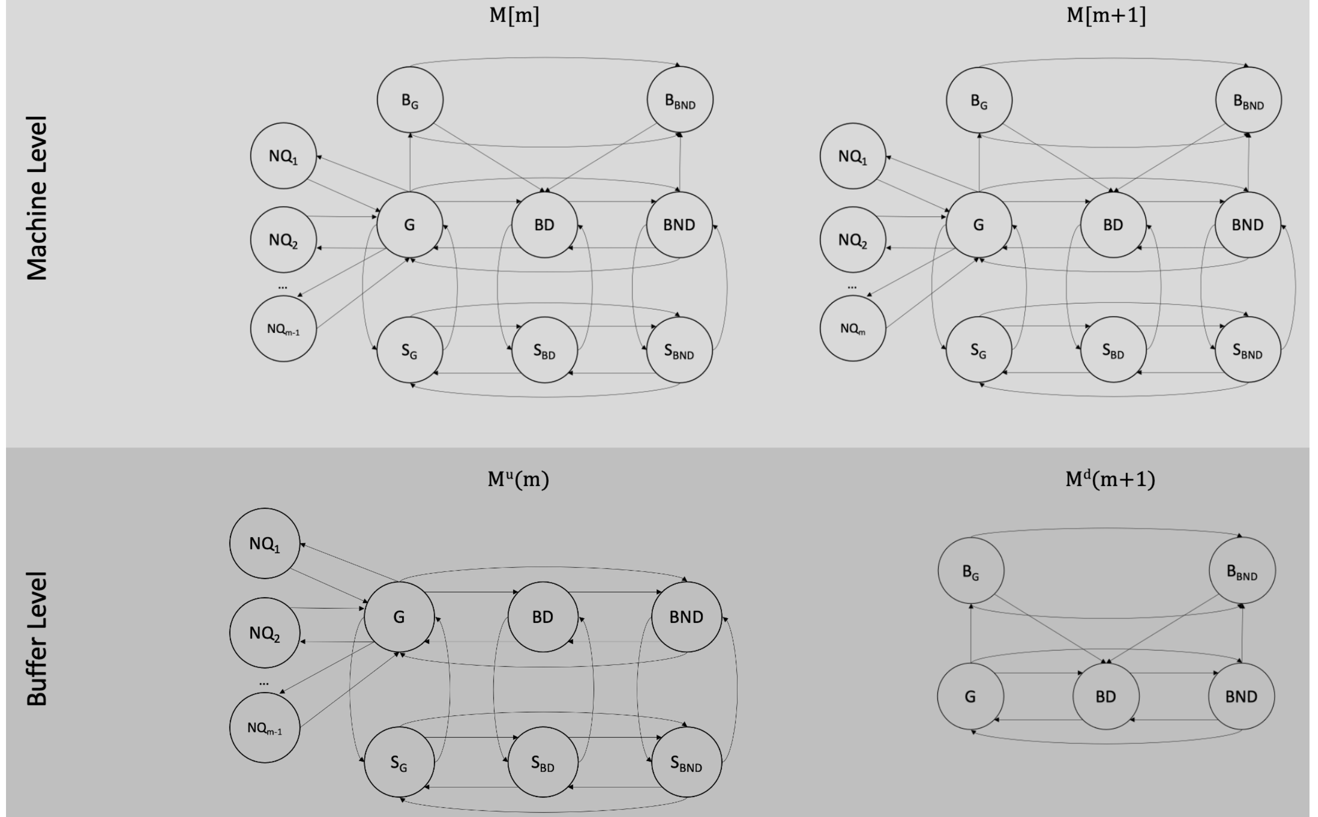

Figure 7.9: Relation between the Markov Chains of Integrated Machines (machine-level) and pseudo-machines (buffer-level).

In particular, the upstream pseudo-machine $M^u(m)$ is characterized by state space $S^u(m) = [L^{[m]}, S^{[m]}, NQ^{[m]}]$. The corresponding transition rate matrix $Q^u(m)$ is:

$$Q^u(m) = \begin{bmatrix} Q_{LL}^{[m]} & Q_{LS}^{[m]} & Q_{LNQ}^{[m]} \\ Q_{B}^{[m]} & Q_{SS}^{[m]} & Q_{SNQ}^{[m]} \\ Q_{NQL}^{[m]} & Q_{NQS}^{[m]} & Q_{NQNQ}^{[m]} \end{bmatrix}; \tag{7.35}$$

Similarly, the downstream pseudo-machine $M^d(m)$ is characterized by the state space $S^d(m) = [L^{[m+1]}, B^{[m+1]}]$. The corresponding transition rate matrix $Q^d(m)$ is

$$Q^d(m) = \begin{bmatrix} Q_{LL}^{[m+1]} & Q_{LB}^{[m+1]} \\ Q_{BL}^{[m+1]} & Q_{BB}^{[m+1]} \end{bmatrix}; \tag{7.36}$$

7.5.7. Convergence algorithm

Forward and backward analysis is performed. During the first forward evaluation, starvation and non-quality states are added to the state space of integrated machines whilst in the first backward analysis blocking limiting states are added.
Once all the states have been added transition rates of integrated machines are iteratively defined from building block.
On the other, building blocks are built with pseudo-machines iteratively defined from integrated machines. An iterative procedure is required.
The condition for termination of the iterations is found in the convergence of the steady-state probabilities of the integrated machines M[m], using norm $L2$ to evaluate difference between probabilities.

$$Diff(m) = \sqrt{\sum_j (\Pi_j(m) - \Pi_j(m-1))^2} \tag{7.37}$$

Step 1: For $m = 1, ..., M$, Integrated Machine $M[m]$ are initialized based on $M\{m\}$.

Step 2: For $m = 1, ..., M-1$

(a) Characterization of upstream and downstream pseudo-machines $M^u(m)$ and $M^d(m)$ from $M[m]$ and $M[m+1]$.

(b) Evaluation of Building Block $BB(m)$, based on $M^u(m)$, $M^d(m)$ and $B(m)$.

(c) Characterization of Integrated Machine $M[m+1]$ based on the downstream pseudo-machine $M^d(m)$.

Step 3: For $m = M-1, ..., 1$

(a) Characterization of upstream and downstream pseudo-machines $M^u(m)$ and $M^d(m)$ from $M[m]$ and $M[m+1]$.

(b) Evaluation of Building Block $BB(m)$, based on $M^u(m)$, $M^d(m)$ and $B(m)$.

(c) Characterization of Integrated Machine $M[m]$ based on the downstream pseudo-machine $M^u(m)$.

A graphical representation of the decomposition method is reported in Figure 7.10

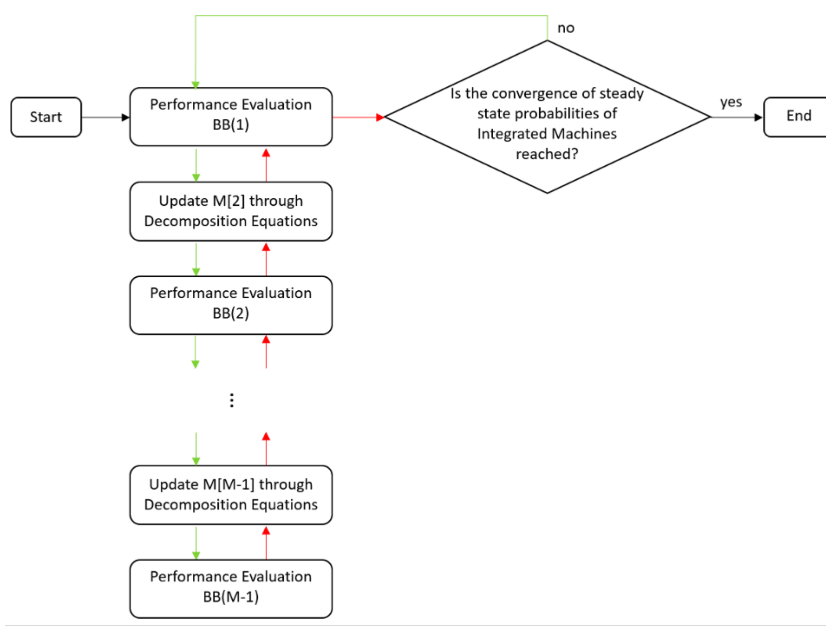

Figure 7.10: Convergence Algorithm

7.6. Performance measures

Throughput

At buffer level, the Integrated Machine $M[m]$ can be seen either as the upstream pseudo-machine $M^u(m)$, in relation to buffer $B(m)$, or as the downstream pseudo-machine $M^d(m-1)$, in relation to buffer $B(m-1)$.

$$TH^{u(m)} = \sum_i \mu(S^u[i]) \int_0^N f(x, S^u[i], S^d[l])dx \tag{7.38}$$

$$TH^{d(m-1)} = \sum_l \mu(S^d[l]) \int_0^N f(x, S^u[i], S^d[l])dx \tag{7.39}$$

Since the defective parts are unloaded from each buffer without being sent to the downstream machine, the conservation of throughput cannot be satisfied. Hence the total throughput of the line must be computed on the machines of the last BB.

$$TH^{u(m)} \neq TH^{d(m-1)}$$

It is possible to calculate the throughput of BD wafers at each lithography&inspection

stage as:

$$TH_{BD}^{[m]} = \mu(S^{[m]}[BD]) \cdot \Pi(S^{[m]}[BD]) \tag{7.40}$$

Instead the total rejected wafers:

$$TH_{BD} = \sum_{m=1}^{M} TH_{BD}^{[m]} \tag{7.41}$$

Moreover, NQ states allow to have at the end of the line an estimate of the throughput of defective wafers not detected through all manufacturing stages. In last Integrated Machine $IM[M]$ states: $S = \{BND, NQ_1, ..., NQ_{M-1}\}$ with production rates $\mu = \{\mu_M, \mu_{NQ_1}, ..., \mu_{NQ_{M-1}}\}$ have a throughput of BND wafers as:

$$TH_{BND}^{[M]} = \sum_{i} \mu(S^{[M]}[i]) \cdot \Pi(S^{[M]}[i]) \; with \; i \in S = \{BND, NQ_1, ..., NQ_{M-1}\} \tag{7.42}$$

Instead throughput of good wafers at the end of the line is:

$$TH_{G}^{[M]} = \mu(S^{[M]}[G]) \cdot \Pi(S^{[M]}[G]) \tag{7.43}$$

Average Buffer Level

The average buffer level is computed as follow:

$$\overline{x}[m] = x[m] \cdot \int_{0}^{N} f(x, S)dx \tag{7.44}$$

8 | Numerical Results

In this chapter, results from analysis conducted on the model are exposed. Firstly analysis of convergence of decomposition algorithm is assessed. Then, performance evaluation of the proposed method has been compared to simulation, and results from this validation are presented: the structure of the experimental campaign, simulation model, and actual validation results are presented.

8.1. Convergence of proposed algorithm

Although the proof of convergence is currently under study, all calculated cases reached convergence in a limited number of iterations. The number of iterations required to reach convergence strongly depends on the length of the selected line.

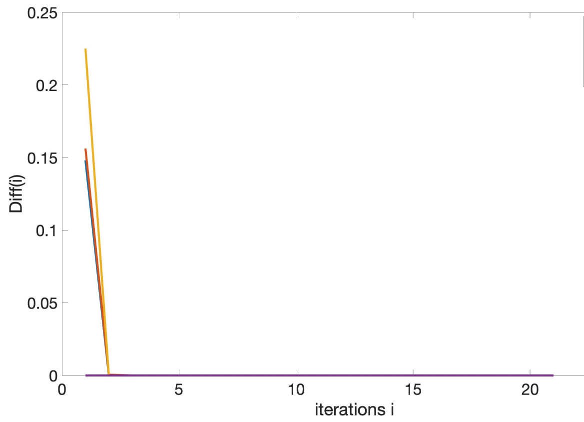

Figure 8.1: Convergence paths

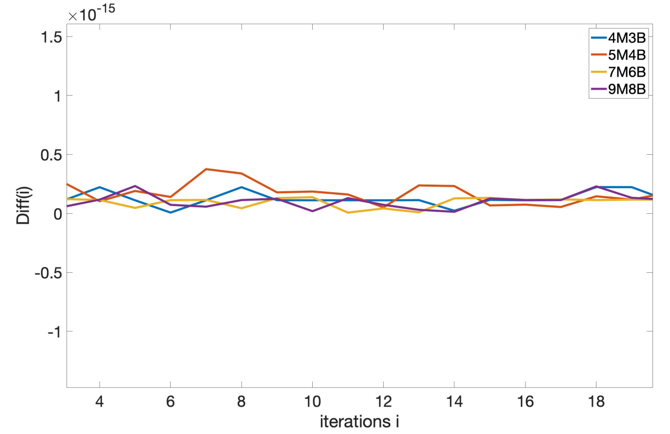

Figure 8.2: Convergence details

For all the calculated cases, the $L2$ norm of probability of a fixed m machine, for each iteration is calculated.

$$Diff^m(i) = \sqrt{\sum_j (\Pi_j^m(i) - \Pi_j^m(i-1))^2} \tag{8.1}$$

Fig. 8.1 shows the average difference in L2 per iteration for machine m . Fig. 8.2 focuses on a smaller scale in order to show the details of the iteration measures also for small values. The algorithm is set to reach a precision of 10^{-15} on Matlab.
We observed that a fair value of precision, e.g. 10^{-8}, was already reached for a low number of iterations in all the tested layouts. This can result in clear advantages when the proposed model is used as the evaluation kernel for optimization algorithm.
It is interesting to notice that $iter_{min} = 1$ since whenever the configuration is with all machines equal, there is no buffer dynamics and solution will converge immediately.

Line	Group	$Iter_{max}$	$Iter_{min}$
4M2B	N=10	4	1
	N=100	5	1
	N=300	6	1
5M3B	N=10	4	1
	N=100	5	1
	N=300	7	1
7M6B	N=10	6	1
	N=100	9	1
	N=300	10	1
9M8B	N=10	6	1
	N=100	9	1
	N=300	11	1

Table 8.1: Convergence analysis of different configurations

8.2. Comparison with discrete event simulator

The proposed analytical model has been validated through the comparison of its performance evaluation results to a simulation model results. The configuration parameter of the manufacturing line proposed in the previous chapter are:

- Capacity of each buffer: N
- Cycle time of photolithography machines equal for each one: C_I
- Percentage of markers selected in each inspection stage i: P_i^*

8.2.1. Simulation model

The simulation model adopted has been designed in *SimEvents*, the Discrete Event Simulation tool of *Simulink*, which is the dynamic system simulation package integrated in MATLAB. The model presents the following characteristics:

- The first machine of the line is never starved and the last machine of the line is never blocked.
- Machines cycle times are deterministic and include the time to load and unload parts.
- First In First Out dispatching policy is adopted.
- The blocking discipline is Blocking After Service.

Simulation settings are reported in Table8.2

Setting Parameter	**Value**
Runs	10
Run length	3000000 t.u.
Warm-up length	100000 t.u

Table 8.2: Simulation settings

For validation of the proposed model, fractional factorial plan with 64 cases, combining 6 different configuration parameters with two values each, has been designed. Results from analytical model and simulation model are compared through the percentage error of the measured performance:

- Steady-state throughput end-line: TH_{END}
- Steady-state scrap throughput: TH_{BD}

The cases composing the experimental campaign have been identified with the Minitab tool Design of Experiment, to select the combinations maximizing the power of the campaign. Three different lines has been analyzed:

- 2M1B Line
- 4M3B Line

Results from analytical model and simulation model are compared through the percentage error of the measured performance:

- steady state throughput: $err\%_{TH^{END}} = \frac{TH^{END}_{mod} - TH^{END}_{sim}}{TH^{END}_{sim}} \cdot 100$
- steady state scrap throughput: $err\%_{TH^{BD}} = \frac{TH^{BD}_{mod} - TH^{BD}_{sim}}{TH^{BD}_{sim}} \cdot 100$

The configuration parameters are:

- Buffer capacity for each buffer: N
- Cycle time of photolithography: $C_I \frac{hours}{wafer}$
- Percentage of markers at machine i: P^*_i

Experimental plan 2M1B Line

In Figure 8.3 is possible to see the *Simulink* model used for the proposed analysis. It is possible to notice splitting of flow after each server that model scrapping in process.

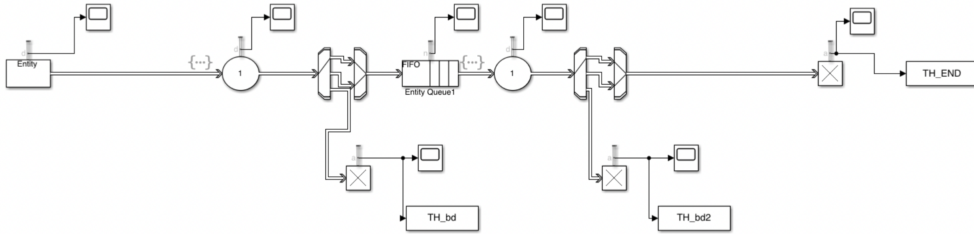

Figure 8.3: Discrete event simulator used 2M1B

The total amount of configuration parameters is 4. Parameters for this case are reported in Table 8.3 Details about parameters for all cases are reported in Appendix A.

Level	N	C_I	P^*_1	P^*_2
Low	10	0.1	50	65
High	300	0.25	70	100

Table 8.3: Parameters of the experimental campaign: Two-Machine Line

Results

It is possible to notice from Figure 8.6 that error $err\%_{TH^{END}}$ has a mean and variability really low compared with the error $err\%_{TH^{BD}}$ that is much higher but still within 2%.
Thus is reasonable since the probability of BD is really low so from each iteration there could be a substantial difference even though the run length and warm-up length are set at high values.
Thus results are appropriate and the error does not to seem abnormal patterns, this result validate the approximate analytical model.

Performance	err[%]		
	MIN	MEAN	MAX
TH_{END}	0.041	0.192	0.445
TH_{BD}	0.177	0.817	1.586

Table 8.4: Errors on performance measures

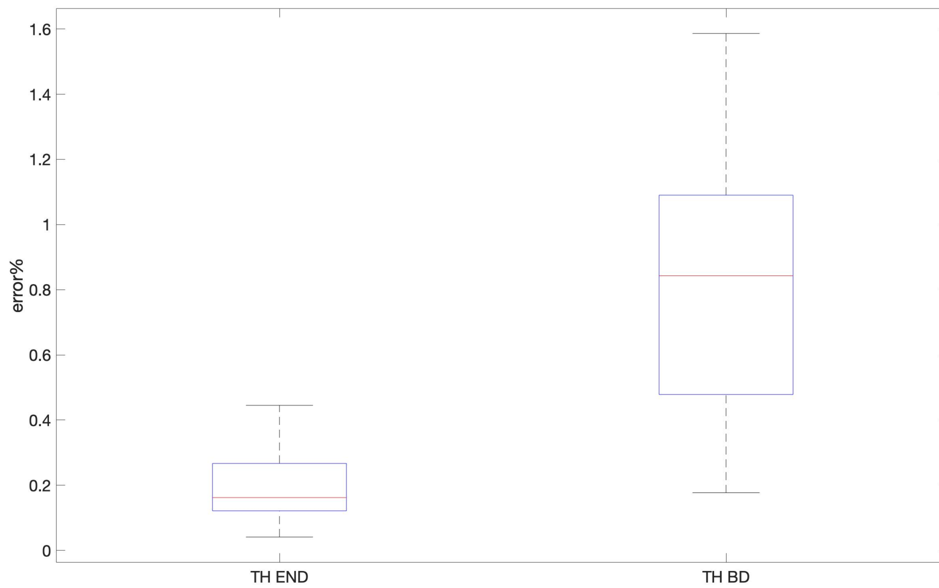

Figure 8.4: Box Plot errors 2M1B line

Experimental plan 4M3B Line

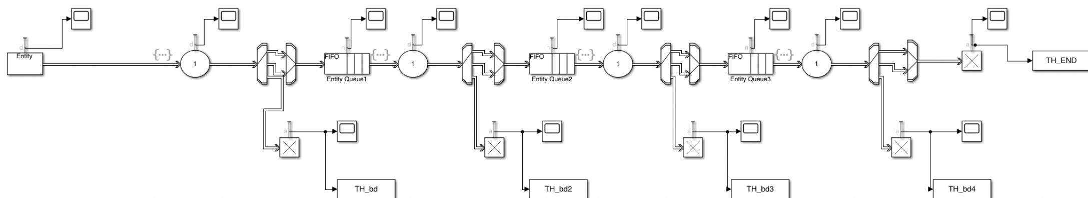

Figure 8.5: Discrete event simulator used 4M3B

The total amount of configuration parameters is 6. Parameters for this case are reported in Table 8.5 Details about parameters for all cases are reported in Appendix A.

Level	N	C_I	P_1^*	P_2^*	P_3^*	P_4^*
Low	10	0.1	50	55	60	70
High	300	0.25	80	85	90	100

Table 8.5: Parameters of the experimental campaign: Four-Machine Line

Results

It is possible to notice from Figure 8.6 that error in both performance errors are in the order of 2% and the variability of the error of BD parts is higher.

Compared with 2M1B results it is possible to see that with increasing of machines increases $err\%_{TH^{END}}$ and it is similar in mean and variance to $err\%_{TH^{BD}}$, but still variance of $err\%_{TH^{BD}}$ is higher.

Thus results are appropriate and the error does not to seem abnormal patterns, this result validate the approximate analytical model.

Performance	err[%]		
	MIN	MEAN	MAX
TH_{END}	0.072	0.992	1.898
TH_{BD}	0.067	0.983	2.172

Table 8.6: Errors on performance measures

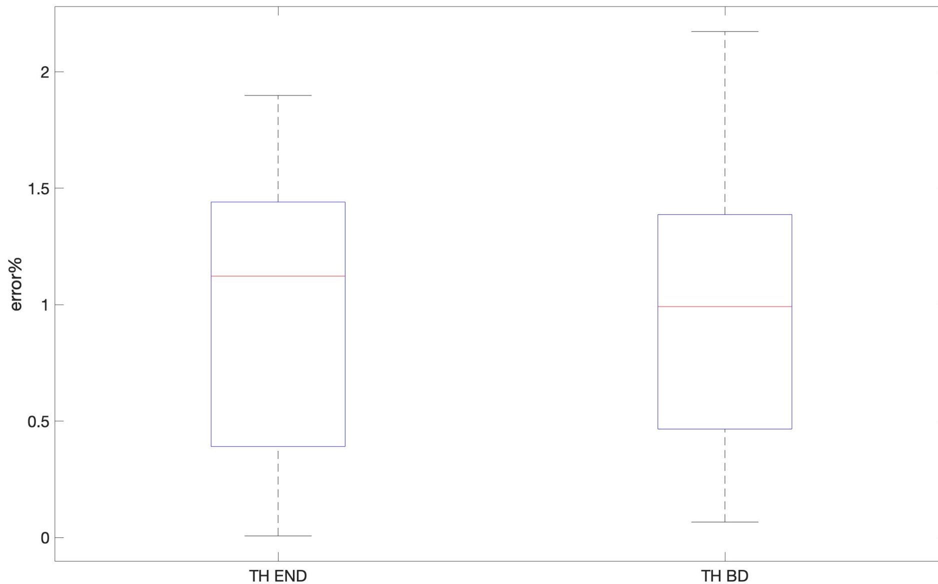

Figure 8.6: Box Plot errors 4M3B line

9 | Real Case

In this chapter, the model application to an industrial case is presented.
The real datased is provided by a semiconductor foundry based in Austin TX, thus for privacy reason data is scaled but realistic. A performance evaluation model that can jointly consider production system and process parameters and quality related effect is of vital importance for taking optimal decisions with a more complete view of the problem. To obtain a joint two-level control, both at system and process level the model must take into account all the parameters related to both these aspects such as the number of inspected points on each wafer, the number of sampled wafer from each lot and the buffer capacity.
The objective is to analyze the impact of the different inspection patterns to the system performance and dynamics.

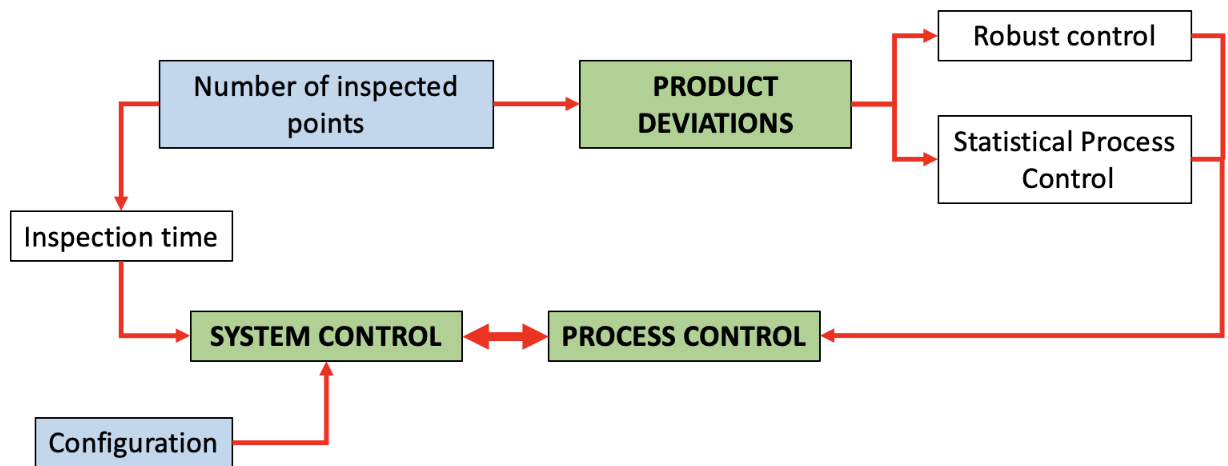

Figure 9.1: System-Process-Product control

9.1. Line model

It is presented the model used in which following results are based. Real-dataset is provided for a 4 machine line, quality measures described in Section 4.3.3 are used.

In Figure 9.2 is presented the system where all fixed hypothesis and assumptions in Section 7.1 are valid.

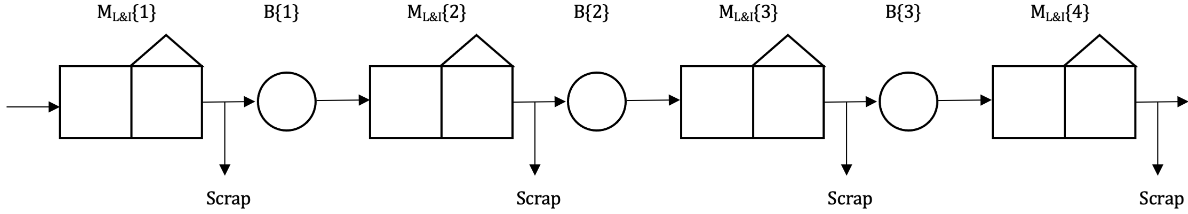

Figure 9.2: Photolithography line model reduction

9.2. Optimization problem

In this section is addressed the optimization problem set to have numerical results.

$$max\ Z = C_G \cdot TH_G(\underline{\%}, \underline{N}) - \underline{C'_{BD}} \cdot \underline{TH_{BD}}(\underline{\%}, \underline{N}) - C_{BND} \cdot TH_{BND}(\underline{\%}, \underline{N}) \tag{9.1}$$

with Performance parameters:

- TH_G: Flow of good wafers at end of line.
- TH_{BND}: Flow of bad not detected wafers at end of line.
- TH_{BD}: Flow of Bad Detected Layers at each inspection machine.

Decision variables that represents the system configuration:

- $\underline{\%}$: Percentage of Markers used in each inspection station.

$$40\% \leq \%^{[m]} \leq 100\%$$

- $\underline{N}$: Buffer capacity in each $B(m)$.

$$10 \leq B(m) \leq 300$$

and cost parameters:

- C_G:Revenue per unit of flow.
- C_{BND}:Cost of BND wafers per unit flow.

$$C_{BND} = K_2 \cdot C_G\ with\ K_1 = 0, 0.25, ..., 8$$

- $\underline{C_{BD}}$: Cost of BD wafers per unit flow in each inspection stage.

$$\underline{C_{BD}}(4) = K_1 \cdot C_G \; with \; K_2 = 0, 0.2, ..., 1$$

$$\underline{C_{BD}}(m) = \frac{m}{4} \cdot \underline{C_{BD}}(4) \; with \; m = 1, 2, 3$$

9.3. As-is inspection policy

Current as-is inspection policy is to measure a certain number of markers at full capacity. For memory circuits less and for logic circuits quite more, but rules to have a certain number of points are based on each single manufacturer experience rules.
Each stage (Lithography+Inspection) has same cycle time since inspections use all markers available to check overlay errors.
For following numerical results is considered a $CT_L = 0.25 \frac{hour}{wafer}$ and $CT_I = 1.5 \cdot 0.25 \frac{hour}{wafer}$. If it is considered a four machine line with scrapping in process the only variability introduced in the system is defined by quality scrapping in process.
As it is shown in Figure 9.3 there is a starvation issue due to the out-flowing of parts, the flow is not conserved and downstream machines are affected by this behaviour.

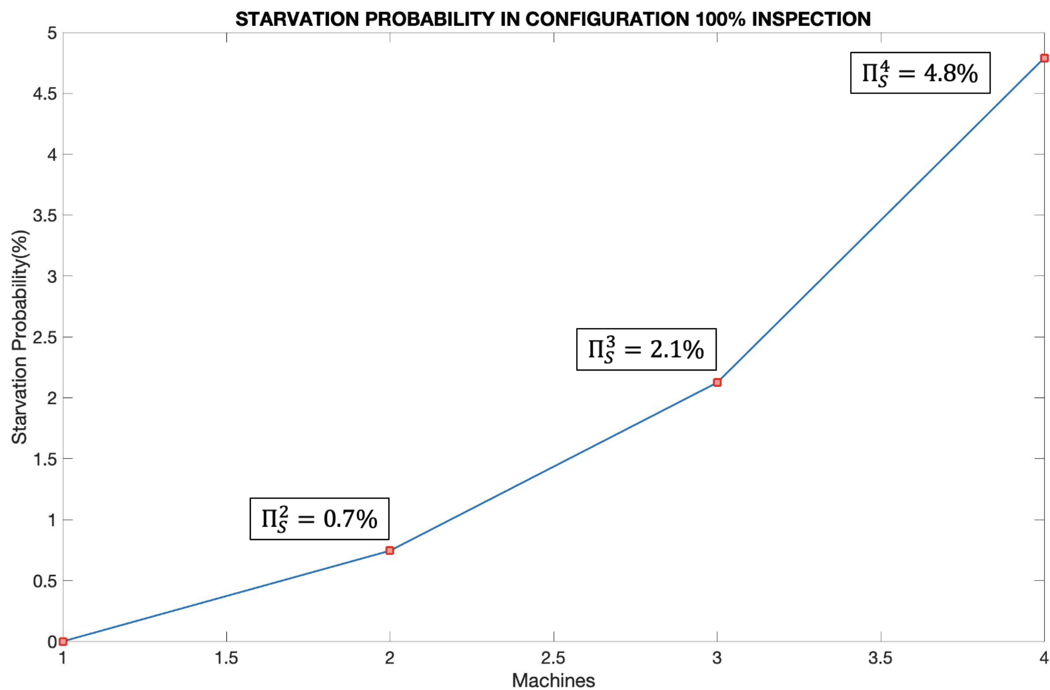

Figure 9.3: Starvation Propagation in current configuration

It is evident that this phenomena affects the whole productivity of the manufacturing line

since downstream machines are highly impacted by starvation effect, on the other hand measuring the whole set of available markers on the wafer's layer enable production to be completely reliable, thus all wafer outgoing from last machine will be perfect at first approximation.
A trade-off between productivity and quality could be achieved and it is discussed in next sections. This trade-off could be achieved because at the end of front-end fabrication electric testing is implemented at wafer level- know as probe testing- where full inspection is performed.

9.4. Solving unbalancing

Even though theoretically line is balanced since each machine has same cycle time, due to scrapping in process it becomes unbalanced since flow is not conserved.
In order to overcome this problem, from an intuitive point of view, could be possible to decrease production rate of upstream machines to have a higher flow in first stages to compensate scrapping along the production line.

It was set a Genetic algorithm using the optimization problem described in 9.2 to solve the problem described above and to find the best solution given certain cost parameters. It was set cost parameters such as:

- $K_2 = 7.5$
- $K_1 = 1$
- $B(1) = B(2) = B(3) = 300$

So it was given an extreme value at K_1 (cost of BND flow), and the highest value for K_2 (cost of BD flow).

Optimal Configuration

	M1	**M2**	**M3**	**M4**
$\%Mrks$	96	97	100	100

Table 9.1: Optimal percentage of markers in each inspection machine given K_1 and K_2

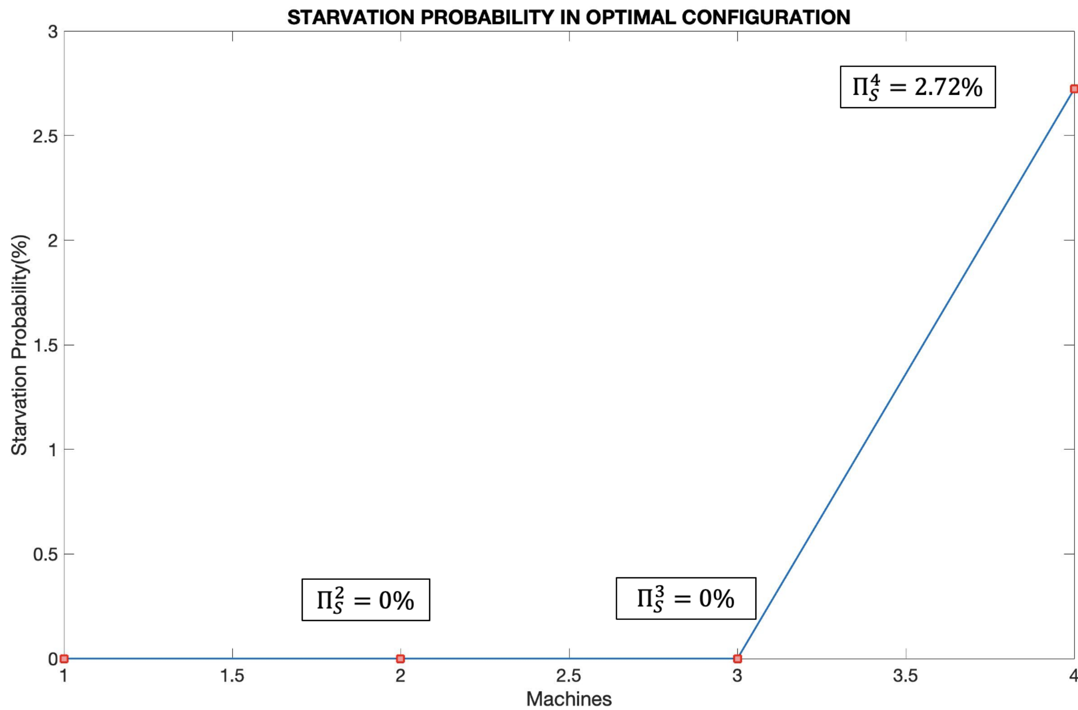

Figure 9.4: Starvation Propagation in Optimal configuration

It is evident from Figure 9.4 that starvation has been limited in the optimal configuration found by the optimization problem.

Performance comparison

	100% conf.	opt. conf.
$\Pi_S^4\%$	4.8	2.72
$TH_G[\frac{w}{h}]$	1.474	1.507
$TH_{BND}[\frac{w}{h}]$	0	0.0004
$TH_{BD}[\frac{w}{h}]$	0.125	0.1274
$Yield[\frac{TH_G}{TH_{IN}}]$	0.9217	0.9218

Table 9.2: Comparison between 100% markers selected for each machine and optimal configuration found

Performance measures of as-is policy 100% line and the optimal configuration are compared in Table 9.2.

It is possible to notice that starvation probability as already seen has been reduced.

TH_G,throughput of Good parts from last stage has been increased even though the cycle time at last stage is the same, since starvation time is now lower.
On the other hand decreasing percentage of selected markers has increased TH_{BND} since inspection in first stages is perfect and previously deposited layers cannot be detected by later inspection stages.
TH_{BD} should have been decreased since decreasing % of markers would lead to a inferior scrapping in process but since production rate in first stages has increased more than the decrease in the $\Pi(BD)$.
$Yield$ on optimal configuration is slightly better than as-is configuration.

9.5. Sensitivity Analysis

Optimal configuration given by optimization problem is higly dependant on cost parameters of BD-BND flow i.e. K_1,K_2, it is essential to perform a sensitivity analysis over this parameters to see how the solution would differ.

In next sections sensitivity analysis is performed in a fist approach where all inspection machines in the line have the same % of markers selected where starvation phenomena still affect the manufacturing line and a second approach, where each machine can assume a different % of markers to overcome starvation effect.

Cost and parameters fixed for following analysis:

- $0 \leq K_1 \leq 1$
- $0 \leq K_2 \leq 8$
- $B(1) = B(2) = B(3) = 300$

9.5.1. Percentage of markers equal in each machine

First sensitivity analysis is performed with the hypothesis of maintaining same cycle time between each inspection state i.e. same percentage of markers.

It is important to say that within this analysis it is hypothesized that intermediate processes can handle a higher production capacity thus they will not become line's bottlenecks.

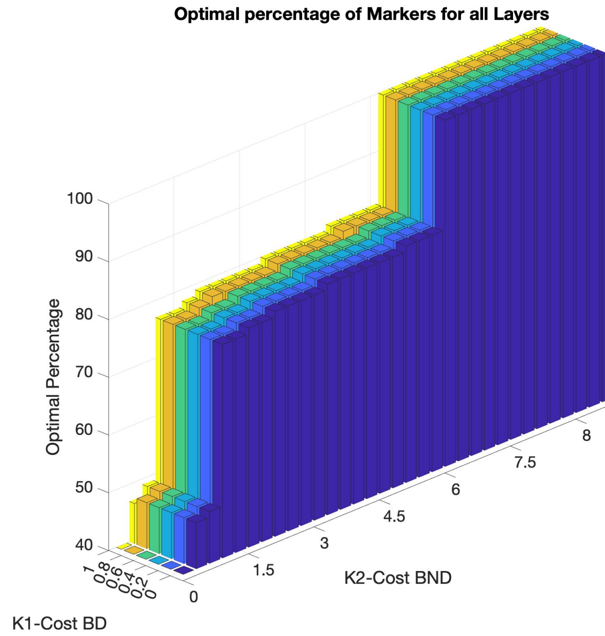

Figure 9.5: Sensitivity analysis over K_1 and K_2 with optimal % of markers equal for each machine

From Figure 9.5 it is evident how cost K_1 does not affect overall solution, there is no variation of optimal percentage of markers over K_1 axis.

An interpretation for this result could be that the influence of increasing K_1 cost is balanced by productivity of the line that is highly influenced by other cost.

On the other hand K_2 is the main parameter that leads trade-off between line's productivity and quality. In Figure 9.6 it is shown the trend, as K_2 increase its weight, optimal percentage increases accordingly balancing the trade-off, until a value of $K_2 \approx 5$ where $K_2 \cdot TH_{BND}$ overweight $C_G \cdot TH_G$ and % of markers jump suddenly to 100%.

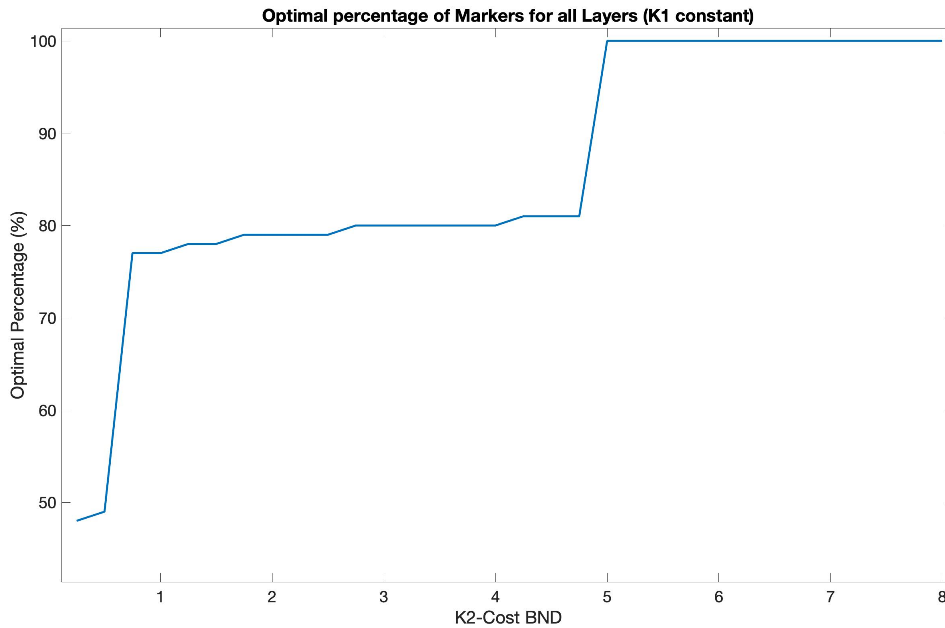

Figure 9.6: sensitivity analysis over K_2 cost parameter

9.5.2. Percentage of markers different in each machine

In order to overcome starvation effects on the line and to optimize the whole production line it is convenient to consider each inspection station with different percentage.

Cost of BD wafers i.e K_1 does not affect considerably optimal consifiguration as in previous sensitivity analysis. Only results over BND cost will be assessed.

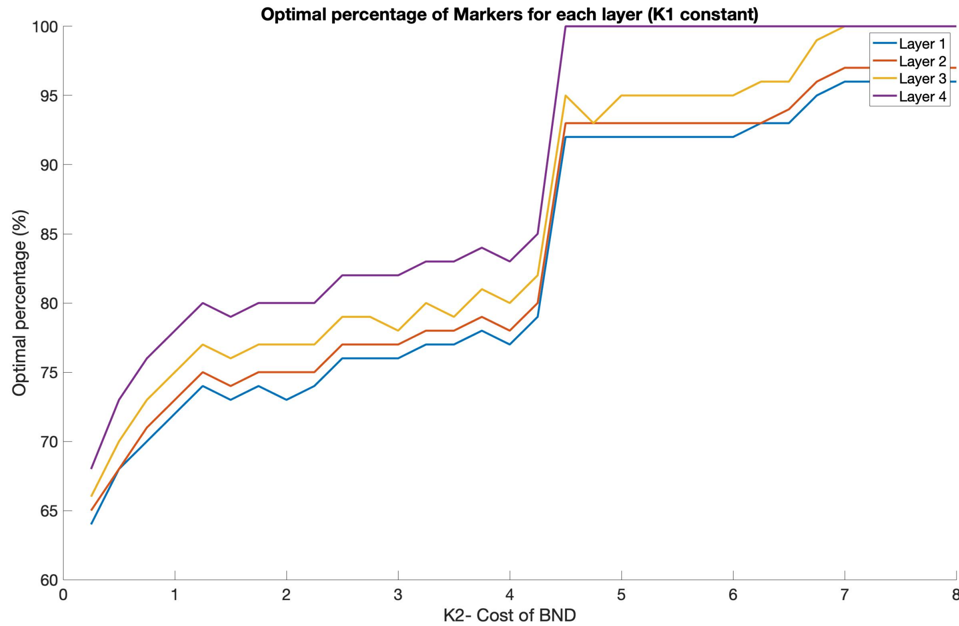

Figure 9.7: sensitivity analysis over K_2 cost parameter

In Figure 9.7 it is shown the trend similar to the previous analysis, as K_2 increase its weight, optimal percentage increases accordingly balancing the trade-off.
Each layer has its own percentage of selected markers and it is increasing with downstream layers, this is explainable mainly for two reasons:

- To overcome unbalancing of the line, so to reduce starvation at last production stage.
- Stack-up overlay error increases along the manufacturing stages since, mathematically is the summation of the whole set of overlays deposited in each stage, so it is appropriate to select more markers in later stages so that the bad layers will be detected.

9.5.3. Comparison between two strategies

The aim of this section is to compare the strategies described in previous paragraphs:

- Optimal configuration with percentage of markers equal in each inspection stage.
- Optimal configuration with percentage of markers that could differ in each single inspection stage.

In Figure 9.8 is presented the comparison of the different approaches described.

It is evident that for every cost value for each cost parameter the strategy where each inspection station can change its percentage of inspected markers is better than the strategy where every machine has same percentage of inspected markers.

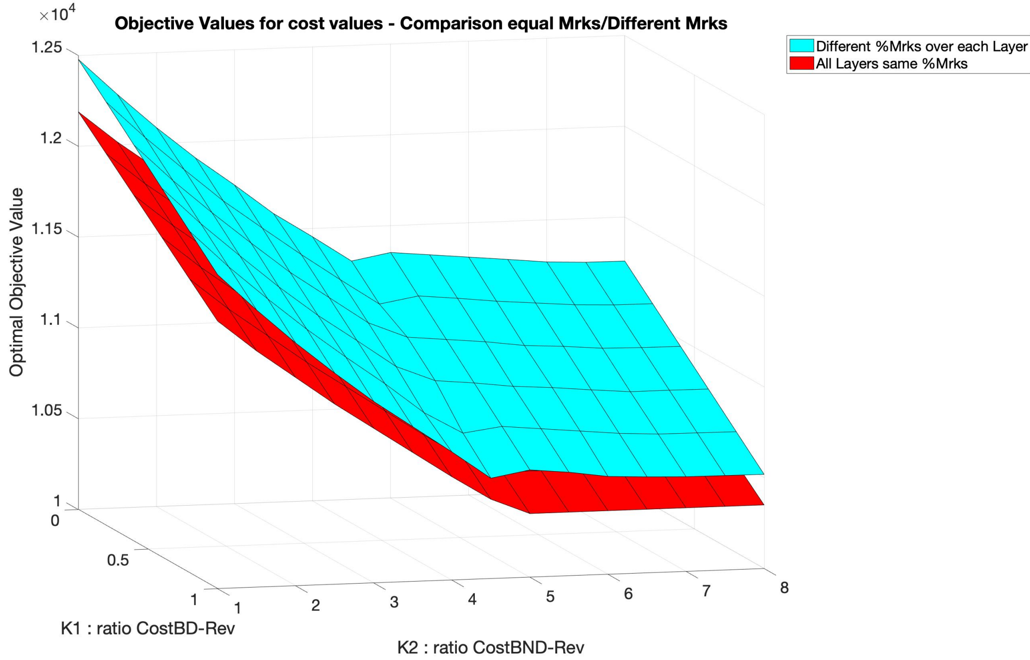

Figure 9.8: Comparison between two different inspection policies

9.6. Importance of optimal set of measurement markers

Given a percentage P^{obj} of the total amount of markers available it is possible to select any set/combination of markers $F(P^{obj})$.

The objective of [64] is to to find the best set of markers $F^{*}(P^{obj})$ that minimize the difference between the distribution of process bias $\hat{c}^{x}(F(P^{obj})), \hat{c}^{y}(F(P^{obj}))$estimated with observations from selected markers and the distribution of $\hat{c}^{x}(F^{Tot}), \hat{c}^{y}(F^{Tot})$ estimated with observations from all the candidate markers.

To know details on how optimal selection is performed please refer to [64].

Now, given a P^{obj}, it is presented a comparison of performance results between be the as-is operations, using the best set of markers $F^{*}(P^{obj})$ and using a generic set of markers $F(P^{obj})$.

It is assumed that using a generic set of markers $F(P^{obj})$, probability of not detecting a bad layer $\Pi(BND)$ 4.6 will increase by a 20% from optimal case.

It could be better or even be worse. The objective is to compare these results to enhance the importance to make the right decision on the selection on the best $F(P^{obj})$.

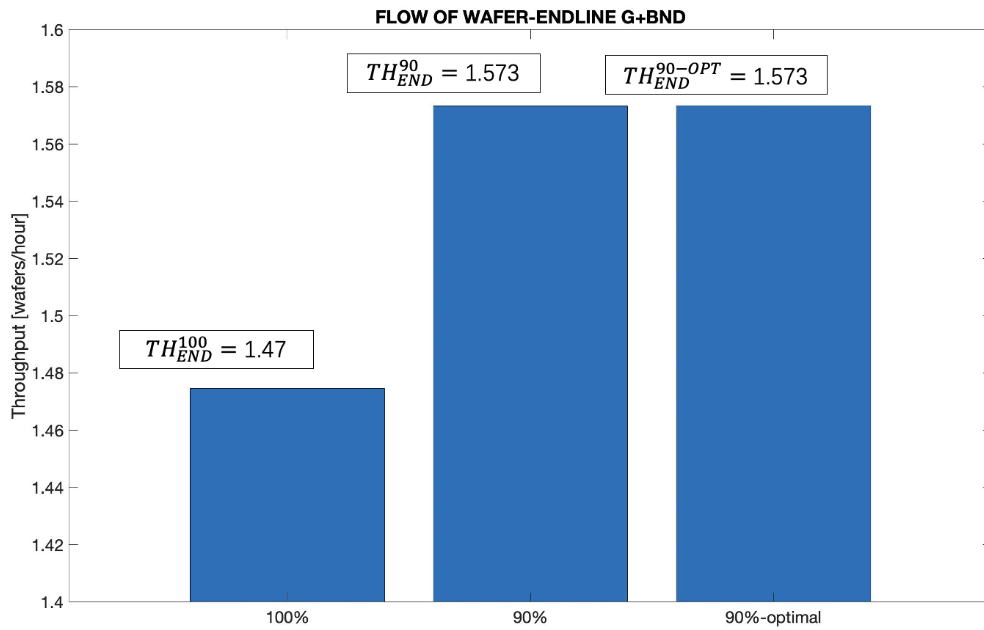

Figure 9.9: Throughput endline $TH_{G\&BND}$

Figure 9.9 shows the throughput of wafers flowing from the last machine of the system in the three different configurations where all machines has the same number of markers selected.

It is evident that inspection of 100% of available markers increases cycle time thus throughput of the line low, on the other hand decreasing percentage (as an example 90%) allows the whole line to produce faster.

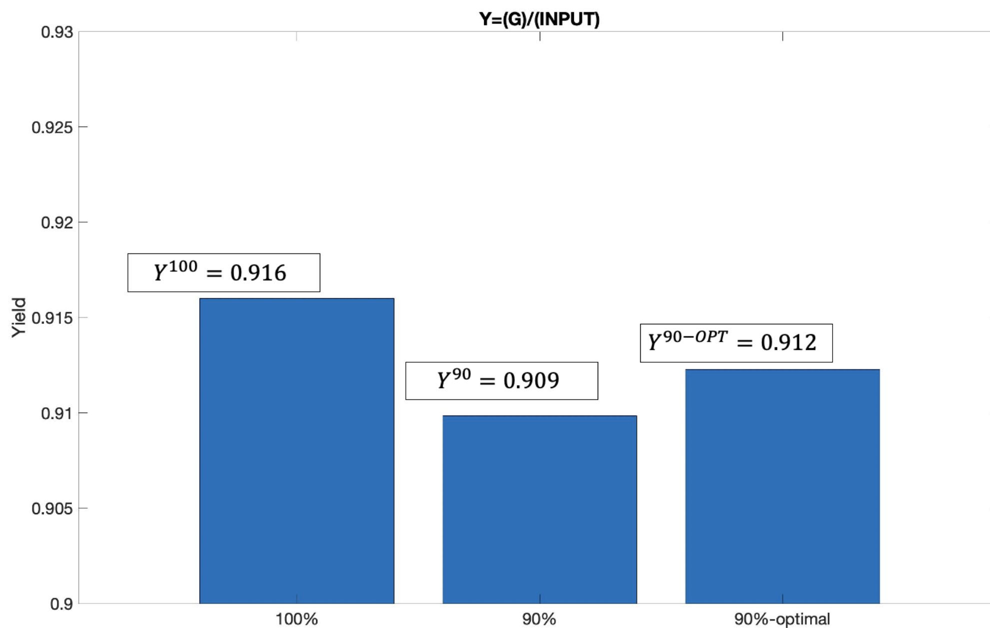

Figure 9.10: Yield: as $\frac{TH_G}{TH_{INPUT}}$

Figure 9.10 highlights how production yield changes among different configurations: in 100% configuration line is the best possible since no BND are produced in the system. Configurations with 90% selected markers are slightly worse but using the optimal set of markers increases the chances of discovering BND, scrapping them and not using production capacity on defective wafers.

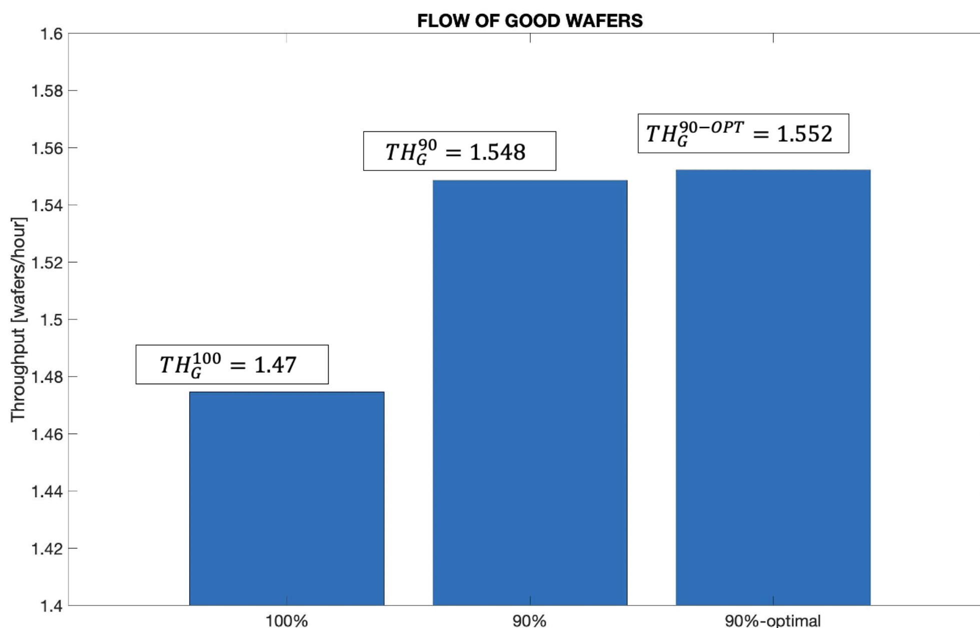

Figure 9.11: Throughput endline of good parts TH_G

Lastly it is possible to show 9.11 that throughput of Good wafers combining two results above described leads to have a higher throughput of good wafers in the case of selecting 90% instead of using the complete set of markers only if the set of markers used is the optimal one.

10 | Conclusions and future developments

In this thesis, it is developed an analytical method for performance evaluation of asynchronous production lines where machines and lithography machines are controlled through robust control on the overlay error correction and inspection station can statistically evaluate production quality.The method is based on a continuous-time mixed-state Markov Chain representation of the line, with a continuous material flow.
Moreover in-process scrap of defective parts is implemented.

Methodology with exact solution is formulated for two-machine line and then it is extended to longer lines with an approximate method, based on a two-level decomposition approach.

The decomposition method consists in decomposing the original line in many two-machine lines, as many as the number of buffers, and transferring information to machines in the form of transition rates to enter and exit blocking and starvation states and including quality information concerning previous stages on production line.

An experimental campaign has been conducted and performance evaluated through the proposed analytical method has been compared with results from discrete event simulation. Comparisons have proven the accuracy of the results provided.

The real case study is the result of a nine months research program carried out at the University of Texas at Austin, where the study of the processes involved in the manufacturing of integrated circuits is of main concern, because of the outpost factory pool that the city represents for semiconductors.

The final results show that when considering in a unique framework process control and system engineering the optimal solution can be different from the one derived considering only one aspect.

More specifically it is shown that decreasing in a optimal way the percentage of selected markers could be beneficial in terms of productivity compromising just a little bit the

quality output, indeed decreasing this percentage layer-to-layer allows the line to balance the starvation brought by the scrapping in process.

10.1. Future developments

The thesis fits in the research area of analytical modeling of manufacturing systems for performance evaluation. Moreover it addresses the need to consider in a unique framework production logistic and process control.
Therefore, many possibilities for future developments are present.
Some are more closely related to the model proposed:

- Integrate intermediate processes in between photolithography&inspection stages; as a conglomerate of machines with stochastic cycle time with no failures.
- Flow splitting to rework stations with stochastic cycle time decoupled by a finite buffer. This flow will merge in previous lithography station.
- Consider flow as batches of N wafers into a cassette/wafer carriers.

Bibliography

[1] URL https://www.tel.com/sustainability/report/hq95qj0000000pdf-att/sr2020_03_e.pdf.

[2] Sustainable consumption and production, 2022. URL https://www.un.org/sustainabledevelopment/sustainable-consumption-production/.

[3] S. Bassetto and A. Siadat. Operational methods for improving manufacturing control plans: Case study in a semiconductor industry. *Journal of Intelligent Manufacturing,*, 2009.

[4] M. H. Burman. *New results in flow line analysis.* PhD thesis, Massachusetts Institute of Technology, 1995.

[5] J. A. Buzacott. Modelling manufacturing systems. *Robotics and Computer Integrated Manufacturing,*, 1985.

[6] J. A. Buzacott. A review of: Stochastics models of manufacturing systems. *European Journal of Engineering Education*, 1993.

[7] S. Y. Chiang, C. T. Kuo, and S. M. Meerkov. Bottlenecks in serial production lines: theory and application. *IEEE Transactions on Robotics and Automation*, 2000.

[8] S. Y. Chiang, C. T. Kuo, and S. M. Meerkov. Bottlenecks in serial production lines: Identification and application. *Mathematical Problems in Engineering*, 2001.

[9] Y. F. Choong and S. B. Gershwin. A decomposition method for the approximate evaluation of capacitated transfer lines with unreliable machines and random processing times. *IIE transactions*, 1987.

[10] M. Colledani and S. B. Gershwin. A decomposition method for approximate evaluation of continuous flow multi-stage lines with general markovian machines. *Annals of Operations Research,*, 2013.

[11] M. Colledani and T. Tolio. Performance evaluation of production systems moni-

tored by statistical process control and on-line inspections. *International Journal of Production Economics*, 2009.

[12] M. Colledani and T. Tolio. Integrated analysis of quality and production logistics performance in manufacturing lines. *International Journal of Production Research*, 2011.

[13] M. Colledani, F. Gandola, A. Matta, and T. Tolio. Performance evaluation of linear and non-linear multi-product multi-stage lines with unreliable machines and finite homogeneous buffers. *IIE Transactions (Institute of Industrial Engineers)*, 2008.

[14] M. Colledani, T. Tolio, A. Fischer, B. Iung, G. Lanza, R. Schmitt, and J. Váncza. Design and management of manufacturing systems for production quality. *CIRP Annals - Manufacturing Technology*, 2014.

[15] Y. Dallery and S. B. Gershwin. Manufacturing flow line systems: a review of models and analytical results. *Queueing systems*, 1992.

[16] Y. Ding, D. Ceglarek, and J. Shi. Fault diagnosis of multi-stage manufacturing processes by using state space approach. *Journal of Manufacturing Science and Engineering,*, 2002.

[17] D. Djurdjanovic and J. Zhu. Stream of variation based error compensation strategy in multi-stage manufacturing processes. In *In ASME 2005 International Mechanical Engineering Congress and Exposition,*, 2005.

[18] D. Djurdjanovic, Y. Jiao, and V. Majstorović.. Multistage manufacturing process control robust to inaccurate knowledge about process noise. *CIRP Annals - Manufacturing Technology,*, 2017.

[19] D. Djurdjanović, A. U. Haq, M. C. Magnanini, and V. Majstorović. Robust model-based control of multistage manufacturing processes. *CIRP Annals.*, 2019.

[20] T. F., P. T., L. Z., and C. S. Survey on run-to-run control algorithms in high-mix semiconductor manufacturing processes. *IEEE Transactions on Industrial Informatics*, 2015.

[21] R. C. Farrow and I. C. Kizilyalli. Alignment mark fabrication process to limit accumulation of errors in level to level overlay,. 2002.

[22] E. Gebennini, A. Grassi, C. Fantuzzi, S. B. Gershwin, and I. C. Schick. Discrete time model for two-machine one-buffer transfer lines with restart policy. *Annals of Operations Research*, 2013.

[23] S. B. Gershwin. Efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Operations Research*, 1987.

[24] S. B. Gershwin. *Manufacturing Systems Engineering*. PTR Prentice Hall, 1994.

[25] S. B. Gershwin and O. Berman. Analysis of transfer lines consisting of two unreliable machines with random processing times and finite storage buffers. *AIIE Transactions*, 1981.

[26] S. B. Gershwin and J. E. Schor. Efficient algorithms for buffer space allocation. *Annals of Operations Research,*, 2000.

[27] S. C. Horng and S. Y. Wu. Compensating the overlay modeling errors in lithography process of wafer stepper. In *In Proceedings of the 2010 5th IEEE Conference on Industrial Electronics and Applications, ICIEA 2010*, 2010.

[28] S. J. Hu. Stream-of-variation theory for automotive body assembly. *CIRP Annals - Manufacturing Technology*, 1997.

[29] S. J. Hu, J. Ko, L. Weyand, H. A. Elmaraghy, T. K. Lien, Y. Koren, H. Bley, G. Chryssolouris, N. Nasr, and M. Shpitalni. Assembly system design and operations for product variety. *CIRP Annals - Manufacturing Technology*, 2011.

[30] W. J. and M. S. Gaussian process regression for virtual metrology-enabled run-to-run control in semiconductor manufacturing. *Transactions on Semiconductor Manufacturing*, 2017.

[31] M. A. Jafari and J. G. Shanthikumar. An approximate model of multistage automatic transfer lines with possible scrapping of workpieces. *IIE Transactions (Institute of Industrial Engineers)*, 1987.

[32] Y. Jiao and D. Djurdjanovic. Joint allocation of measurement points and controllable tooling machines in multistage manufacturing processes. In *IIE Transactions (Institute of Industrial Engineers)*, 2010.

[33] Y. Jiao and D. Djurdjanovic. Stochastic control of multilayer overlay in lithography processes. *IEEE Transactions on Semiconductor Manufacturing,*, 2011.

[34] R. Jin and J. Shi. Reconfigured piecewise linear regression tree for multistage manufacturing process control. *IIE Transactions (Institute of Industrial Engineers)*, 2012.

[35] J. Kim and S. B. Gershwin. Integrated quality and quantity modeling of a production line. *OR Spectrum,*, 2005.

[36] J. Kim and S. B. Gershwin. Analysis of long flow lines with quality and operational failures. *IIE Transactions (Institute of Industrial Engineers)*, 2008.

[37] M. B. D. Koster. Estimation of line effciency by aggregation. *International Journal of Production Research*, 1987.

[38] Kuo, H. F., and A. Faricha. Artificial neural network for diffraction based overlay measurement. *IEEE Access*, 2016.

[39] J. Lee, B. Bagheri, and H. A. Kao. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, 2015.

[40] R. Levantesi, A. Matta, and T. Tolio. Performance evaluation of continuous production lines with machines having different processing times and multiple failure modes. *Performance Evaluation*, 2003.

[41] H. J. Levinson. *Lithography process control*, volume 28. SPIE Press, 1999.

[42] J. Li, D. E. Blumenfeld, and J. M. Alden. Comparisons of two-machine line models in throughput analysis. *International Journal of Production Research*, 2006.

[43] G. Liberopoulos, G. Kozanidis, and P. Tsarouhas. Performance evaluation of an automatic transfer line with wip scrapping during long failures. *Manufacturing and Service Operations Management*, 2007.

[44] M. C. Magnanini and T. Tolio. Decomposition of asynchronous automated long lines with finite buffer capacity. In *In 11th Conference on Stochastic Models of Manufacturing and Service Operations*, 2017.

[45] M. C. Magnanini and T. Tolio. A markovian model of asynchronous multi-stage manufacturing lines fabricating discrete parts. *Journal of Manufacturing Systems*, 2023.

[46] S. S. Mandroli, A. K. Shrivastava, and Y. Ding. A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes. *IIE Transactions (Institute of Industrial Engineers)*, 2006.

[47] L. Monostori, B. K. ad ar, T. Bauernhansl, S. Kondoh, S. Kumara, G. Reinhart, O. Sauer, G. Schuh, W. Sihn, and K. Ueda. Cyber-physical systems in manufacturing. *CIRP Annals*, 2016.

[48] Y. Nonaka, Y. Suginishi, A. Lengyel, M. Ono, and K. Sugimoto. Tsunami effect prediction methodology for critical resource analysis. *In Manufacturing Systems and Technologies for the New Frontier.*, 2008.

[49] K. Okamura and H. Yamashina. Analysis of the effect of buffer storage capacity in transfer line systems. *AIIE Transactions*, 1977.

[50] J. H. Owen and D. E. Blumenfeld. Effects of operating speed on production quality and throughput. *International Journal of Production Research*, 2008.

[51] C. T. Papadopoulos, J. Li, and M. E. O'Kelly. A classification and review of timed markov models of manufacturing systems. *Computers and Industrial Engineering*, 2019.

[52] B. Pourbabai. Optimal utilization of a finite capacity integrated assembly system. *International Journal of Production Research,*, 1990.

[53] L. Qin, J. X. Yu, and B. Ding. *Advances in Databases: Concepts, Systems and Applications*. Number pages 850–862. Springer Berlin Heidelberg., 2007.

[54] B. N. S., R. C. P., R. S. F., and A.-N. J. V. Multidomain simulation model for analysis of geometric variation and productivity in multi-stage assembly systems. *Applied Sciences*, 2020.

[55] M. L. Sale and R. A. Inman. Survey-based comparison of performance and change in performance of rms using traditional manufacturing. *International Journal of Production Research*, 2003.

[56] R.-H. M. Schmidt. Ultra-precision engineering in lithographic exposure equipment for the semiconductor industry. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2012.

[57] T. Tolio, A. Matta, and F. Jovane. A method for performance evaluation of automated flow lines. *CIRP Annals*, 1998.

[58] T. Tolio, A. Matta, and S. B. Gershwin. Analysis of two-machine lines with multiple failure modes. *IIE Transactions (Institute of Industrial Engineers)*, 2002.

[59] T. A. M. Tolio and A. Ratti. Performance evaluation of two-machine lines with generalized thresholds. *International Journal of Production Research*, 2018.

[60] T. Tullio. Performance evaluation of two-machine line with multiple up and down states and finite buffer capacity. In *Eighth conference on stochastic models of manufacturing and service operations.*, 2011.

[61] H. A. U. and D. Djurdjanovic. Robust control of overlay errors in photolithography processes. *IEEE Transactions on Semiconductor Manufacturing*, 2019.

[62] X. L. Xie. Approximate analysis of transfer lines with unreliable machines and finite buffers. *IEEE Transactions on Automatic Control*, 1989.

[63] D. Y, D. R, and X. X-L. An efficient algorithm for analysis of transfer lines with unreliable machines and finite buffers. *IIE Transactions*, 1988.

[64] H. Zhang. *Integrated Operational Decision-Making in Flexible Manufacturing System with Considerations of Quality and Reliability*. PhD thesis, The University of Texas at Austin, 2019.

[65] R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman. Intelligent manufacturing in the context of industry 4.0: A review.

A | Appendix A

In this appendix chapter, the detailed set of experimental parameters for different model validation campaigns and the results obtained are reported.

A.1. 2M1B Line

A.1.1. Parameters of experimental campaign

Case	Parameters			
	N	C_I	P_1^*	P_2^*
1	300	0,25	50	100
2	10	0,1	70	65
3	10	0,25	70	100
4	10	0,1	50	65
5	300	0,25	70	65
6	300	0,1	50	100
7	10	0,1	70	100
8	10	0,25	50	65
9	300	0,1	70	100
10	300	0,25	70	100
11	300	0,25	50	65
12	300	0,1	70	65
13	10	0,1	50	100
14	10	0,25	50	100
15	300	0,1	50	65
16	10	0,25	70	65

Table A.1: Full factorial with 2 values for each parameter

A.1.2. Results

Case	TH_{END}			TH_{BD}		
	Sim	Mod	err%	Sim	Mod	err%
1	1,555497	1,568658	0,147618	0,036122	0,035905	0,603726
2	4,757086	4,808545	0,179017	0,069152	0,069435	0,408079
3	1,583735	1,57298	0,135787	0,035202	0,034794	1,171368
4	4,935215	5,002781	0,445174	0,071265	0,071916	0,906022
5	1,923399	1,923418	0,291065	0,028045	0,027774	0,977172
6	3,948087	3,921644	0,209282	0,088338	0,089762	1,586212
7	3,962035	3,932449	0,263769	0,087749	0,086985	0,877566
8	2,006657	2,001112	0,380604	0,028476	0,028766	1,0089
9	3,888691	3,932449	0,107304	0,088014	0,086985	1,183084
10	1,557668	1,57298	0,049862	0,035075	0,034794	0,807507
11	2,017018	2,001112	0,041275	0,029213	0,028766	1,553769
12	4,751055	4,808545	0,161945	0,069234	0,069435	0,289596
13	3,871708	3,921644	0,270286	0,089137	0,089762	0,69678
14	1,571513	1,568658	0,149782	0,035841	0,035905	0,177054
15	4,96168	5,002781	0,16201	0,072119	0,071916	0,28223
16	1,940197	1,923418	0,083841	0,027622	0,027774	0,548883

Table A.2: Throughput end-line and throughput of scraps results

A.2. 4M3B Line

A.2.1. Parameters of the experimental campaign

Case	Parameters					
	N	C_I	P_1^*	P_2^*	P_3^*	P_4^*
1	10	0,25	80	55	90	70
2	300	0,1	80	55	90	70
3	300	0,1	50	85	60	70
4	10	0,25	50	85	60	100
5	10	0,1	50	85	60	100
6	10	0,25	50	85	60	70
7	300	0,25	50	85	60	100

8	10	0,1	50	85	90	70
9	10	0,1	50	55	60	70
10	300	0,1	50	55	60	100
11	10	0,1	80	55	60	100
12	300	0,1	50	85	90	100
13	300	0,25	80	85	60	100
14	300	0,25	80	55	90	100
15	300	0,25	50	85	90	70
16	10	0,25	80	55	60	100
17	10	0,1	50	55	90	70
18	10	0,1	80	85	60	100
19	300	0,1	50	55	60	70
20	10	0,1	80	55	60	70
21	300	0,1	80	85	90	70
22	300	0,25	50	55	90	100
23	10	0,1	50	85	90	100
24	300	0,25	80	85	90	70
25	10	0,25	50	55	90	70
26	10	0,25	50	55	60	70
27	10	0,1	80	85	90	100
28	300	0,25	80	85	90	100
29	300	0,25	50	55	60	100
30	300	0,25	80	85	60	70
31	300	0,1	80	85	90	100
32	10	0,25	80	55	60	70
33	10	0,25	80	55	90	100
34	300	0,25	50	85	90	100
35	10	0,1	80	85	90	70
36	300	0,25	80	55	60	70
37	10	0,1	80	55	90	100
38	300	0,25	80	55	90	70
39	10	0,1	50	55	90	100
40	300	0,25	50	85	60	70
41	10	0,25	80	85	90	100
42	300	0,1	80	85	60	100
43	10	0,1	50	85	60	70

44	10	0,1	80	55	90	70
45	10	0,25	80	85	90	70
46	300	0,1	80	55	60	70
47	300	0,1	50	85	60	100
48	300	0,1	80	55	90	100
49	300	0,1	80	55	60	100
50	10	0,1	50	55	60	100
51	300	0,1	50	85	90	70
52	300	0,25	80	55	60	100
53	300	0,25	50	55	90	70
54	300	0,1	80	85	60	70
55	10	0,25	80	85	60	70
56	10	0,25	50	55	60	100
57	300	0,25	50	55	60	70
58	10	0,25	50	55	90	100
59	10	0,25	50	85	90	70
60	300	0,1	50	55	90	70
61	10	0,25	50	85	90	100
62	300	0,1	50	55	90	100
63	10	0,25	80	85	60	100
64	10	0,1	80	85	60	70

Table A.3: 2^{k-p} fractional factorial plan. $K = 7, p = 1$

A.2.2. Results

Case	TH_{END}			TH_{BD}		
	Sim	Mod	Err%	Sim	Mod	Err%
1	1,591764	1,608194	1,021646	0,109912	0,110125	0,193098
2	4,071322	4,020485	1,264437	0,272635	0,275313	0,972605
3	4,026455	4,082613	1,375526	0,306891	0,307195	0,098979
4	1,503701	1,532224	1,861514	0,133718	0,136688	2,1723
5	3,82846	3,83056	0,054809	0,345949	0,341719	1,237823
6	1,609198	1,633045	1,460268	0,125009	0,122878	1,734488
7	1,5271	1,532225	0,334499	0,139035	0,136688	1,717331
8	4,014997	3,963005	1,311941	0,377062	0,369958	1,920198

9	4,778713	4,724826	1,140521	0,303754	0,302484	0,419848
10	3,847556	3,834	0,353572	0,316659	0,315941	0,227159
11	3,827269	3,863785	0,945082	0,258532	0,257015	0,590409
12	3,775555	3,822852	1,237218	0,405571	0,410328	1,15942
13	1,527595	1,544615	1,10186	0,117284	0,117948	0,562744
14	1,554168	1,543524	0,689608	0,121524	0,122996	1,196714
15	1,583238	1,58529	0,129444	0,14517	0,147989	1,904556
16	1,569596	1,545514	1,558187	0,100668	0,102806	2,08006
17	4,01895	3,960512	1,475518	0,326798	0,322429	1,355075
18	3,914905	3,861536	1,382088	0,299554	0,294869	1,588954
19	4,692773	4,724849	0,678879	0,299873	0,302485	0,863572
20	4,298921	4,298133	0,018327	0,237478	0,233309	1,787192
21	4,06969	4,022264	1,1791	0,316777	0,317132	0,111779
22	1,530868	1,530578	0,018978	0,145053	0,145246	0,133515
23	3,87347	3,822286	1,339089	0,403784	0,410304	1,589015
24	1,636114	1,608906	1,691098	0,126937	0,126853	0,066645
25	1,563092	1,584205	1,332708	0,129576	0,128972	0,468792
26	1,885581	1,88993	0,230126	0,120081	0,120994	0,754297
27	3,793351	3,856305	1,632514	0,345064	0,348766	1,061535
28	1,531063	1,542591	0,747335	0,139386	0,139509	0,088231
29	1,528775	1,5336	0,314606	0,127223	0,126377	0,670064
30	1,638737	1,658616	1,198509	0,10873	0,107081	1,540012
31	3,800935	3,856478	1,440253	0,347452	0,348773	0,378805
32	1,687312	1,719253	1,857864	0,0943	0,093323	1,046561
33	1,572753	1,543458	1,898015	0,125459	0,122994	2,004402
34	1,507906	1,529141	1,388667	0,16229	0,164131	1,121706
35	4,004985	4,02219	0,427753	0,320717	0,317128	1,131744
36	1,687244	1,719253	1,861811	0,09356	0,093323	0,25343
37	3,854585	3,858644	0,105184	0,304945	0,307486	0,826303
38	1,609445	1,608194	0,077811	0,110807	0,110125	0,619396
39	3,782111	3,825922	1,145121	0,359326	0,363101	1,039896
40	1,649934	1,633045	1,034197	0,123447	0,122878	0,463287
41	1,542634	1,542522	0,007242	0,138381	0,139507	0,80703
42	3,802716	3,861537	1,523239	0,292222	0,294869	0,897809
43	4,07103	4,082613	0,283713	0,307975	0,307195	0,253856
44	3,967749	4,020485	1,311677	0,281133	0,275313	2,113858

45	1,594286	1,608876	0,906833	0,127795	0,126851	0,744225
46	4,284089	4,298133	0,326751	0,234086	0,233309	0,333038
47	3,867429	3,830563	0,962409	0,345174	0,341719	1,01081
48	3,834407	3,858809	0,632382	0,311589	0,30749	1,332963
49	3,92494	3,863785	1,582765	0,253377	0,257015	1,415573
50	3,815648	3,833999	0,478647	0,316916	0,315941	0,308346
51	3,938986	3,963225	0,611589	0,36366	0,369973	1,706192
52	1,532021	1,545514	0,873057	0,103298	0,102806	0,478795
53	1,560682	1,584205	1,484858	0,127444	0,128972	1,184103
54	4,070543	4,146539	1,832752	0,268919	0,267703	0,454117
55	1,641658	1,658614	1,022311	0,105628	0,107081	1,357462
56	1,528853	1,5336	0,309497	0,124958	0,126377	1,122339
57	1,859179	1,88994	1,62759	0,121647	0,120994	0,53946
58	1,525711	1,530369	0,304343	0,147055	0,145241	1,249517
59	1,614298	1,585202	1,835465	0,146092	0,147983	1,277792
60	4,017572	3,960512	1,440734	0,316055	0,322429	1,976937
61	1,507842	1,528915	1,378301	0,165367	0,164122	0,758712
62	3,828728	3,826445	0,059673	0,36038	0,363116	0,753563
63	1,524373	1,544614	1,310452	0,118266	0,117947	0,269962
64	4,192323	4,146535	1,104261	0,263775	0,267703	1,467403

Table A.4: Throughput end-line and throughput of scraps results