POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Mechanical Engineering - Ingegneria Meccanica

Author: Riccardo Pomi

Advisor: Prof. Tullio Tolio

Co-advisors: Dr. Maria Chiara Magnanini, Prof. Dragan Djurdjanovic

Academic year: 2022-2023

1. Introduction

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. It becomes imperative to consider structural reconfigurations, improvement initiatives, and operational adjustments while thoroughly evaluating all 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 photolithography. To produce an entire semiconductor wafer, many 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.

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. 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[1]. 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,to develop effective strategies that maximize overall efficiency and performance reducing possible unbalances.

2. Scientific and industrial context

Semiconductor front-end fabs are extremely complex environment. Wafers flow is in constant movement along hundreds of processing machines, where inspection stations redirect defective wafers into rework loops, different threads or discard them in order to achieve the best yield as possible. Numerous processing steps are involved in front-end fabrication, some of them are represented in Fig.1. Photolithography is one of the most critical among the entire process in the fab in a manner that experts in the sector say the fab is built around the process of photolithography. Many criticalities are present during this step mainly because sub-nanometric precision in the alignment between circuit patterns is required for a correct functionality of the final product. 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. To ensure the correct alignment between the layers an overlay metrology station is required. A typical advanced scheme includes an overlay feedback loop that allows the parameters of the stepper to be adjusted and the overlay to be minimized during the process. The overlay error measurement is the bottleneck of the line, it takes a way longer time than photolithography machine. Moreover production flow's logistics could have influence on the behaviour of the manufacturing line, could cause blocking or starvation.

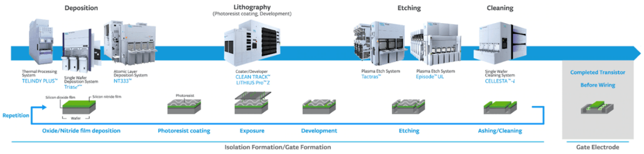

Figure 1: Front-end fabrication steps

At product level, photolithography 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 2. In order to model this phenomena it is used theory based on stream of variation (SoV)[3].

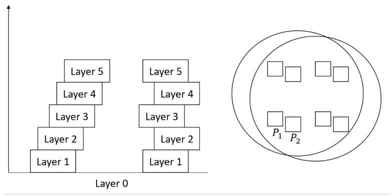

Figure 2: Overlay and stack-up Overlay deviations.

Thus, for the scope of controlling this phenomena in Zang, Djurdjanovic [5] used robust R2R control considering not only overlay, but stack-up overlay error,that is described by the summation of the overlay of non-adjacent layers, by Zernike polynomial based models. The model relate the tool parameters to overlay errors, produce machine settable inputs to minimize those errors[2][5]. Having defined 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.

2.1. Optimal Number of Measurement Markers: Modeling background

This approach 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[6]. By doing so, wafer's inspection would not be perfect so inspection could not detect Bad wafers and let them continue along the manufacturing line. Through [5] is possible to have an estimation of detecting for a certain selected optimal percentage of measurement markers: the probability that a wafer produced is bad and and not detected $\mathbf{P(BND)}$ and the probability that after photolithography a bad wafer is correctly detected and scrapped $\mathbf{P(BD)}$.

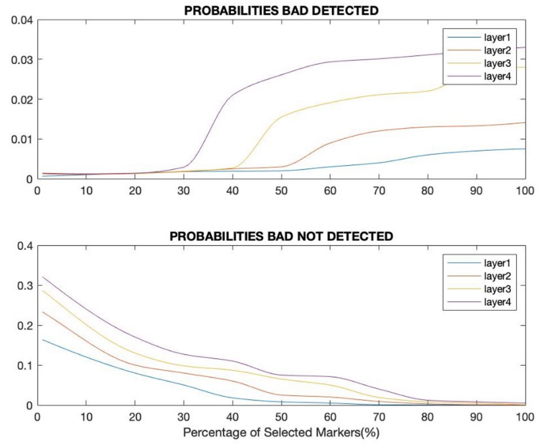

Figure 3: Quality probabilities.

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 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. It is necessary a model that 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 because 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. If it is set an external observer in this case at last inspection machine as shown in Fig. 4, it is possible to notice a part of the flow being discarded or reworked. It is possible with the system model to have an estimation of the probability that some wafer that were defective and was not detected in each previous inspection stage and keep track of it.

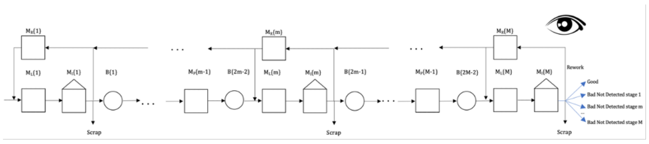

Figure 4: External Observer looking at outgoing flow at last machine.

3. 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. An optimal measurement allocation derived considering only process control can be suboptimal 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.

3.1. Research questions

More specifically the proposed thesis attempts to integrate product/process control developed by Zhang [5] 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.

4. Methodology

4.1. Schematic system

The production system considered is modeled by stations which can be photolithography, inspection stations, inter-operational buffers in a serial layout Figure 6. 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 $M_{L\&I}$, and the system is composed by a total of M stages: $M_{L\&I}(1), ..., M_{L\&I}(m), ..., M_{L\&I}(M)$. Each stage is decoupled by $M - 1$ inter-operational buffers $B(m)$, of finite capacity $N(m)$. Both photolithography and inspection stations are fully reliable, no failures occur in both stages. Whenever 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. The inspection approach adopted is full inspection.

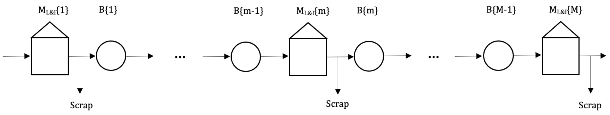

Figure 5: Model Reference System.

4.2. Single-stage model

Each stage is modeled as a Continuous-Time-Markov-Chain and the state-space representation is:

- **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.

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.

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

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

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

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

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

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

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

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

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.

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

Transition rate matrix for Machine m:

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

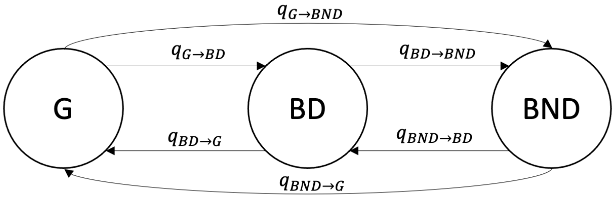

Figure 6: Single-stage Markov model.

4.3. Multi-stage 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, it is based on [4]. 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. 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.

4.3.1. Two-level Decomposition

The multi-stage extends the work presented by Magnanini, Tolio [4]. In Fig.7 is presented the schematic representation of the proposed methodology, and the main steps to build this approach.

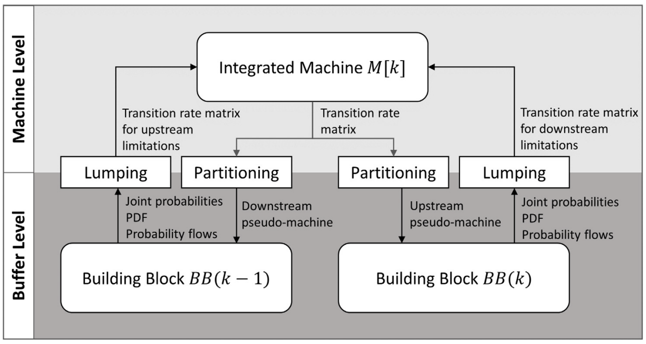
Machine Level
Buffer Level
Integrated Machine $M[k]$
Transition rate matrix for upstream limitations
Transition rate matrix
Transition rate matrix for downstream limitations
Lumping
Partitioning
Partitioning
Lumping
Joint probabilities PDF Probability flows
Downstream pseudo-machine
Upstream pseudo-machine
Joint probabilities PDF Probability flows
Building Block $BB(k-1)$
Building Block $BB(k)$

Figure 7: Schematic representation of the proposed method.

4.3.2. 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 $BB(m)$. 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.
- The remote Blocking states $B[m]$ represent the states in which the Integrated Machine $M[m]$ is downstream limited.
- 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 $BB(m)$.

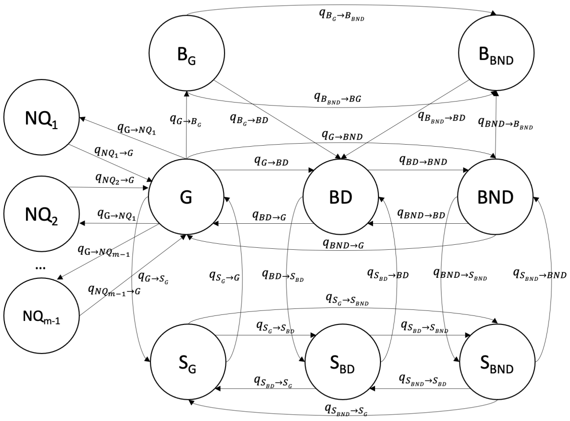

Figure 8: General Integrated Machine.

4.3.3. 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 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 excluding 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: $S^u_{NQ} \otimes S^d$

4.3.4. Lumping

The objective of this step is to characterize the integrated machine based on the output provided by the building block solution, in particular, characterization of the state space and characterization of the transition rate matrix. Let us recall the definition of the transition rate matrix $Q^{[m]}$:

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

The corresponding transition rate matrices can be computed as:

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

$$Q^{[m]}_{LS} = G_{LS}(m-1) \odot [\Pi_L(m-1)]^{-1}$$

$$Q^{[m]}_{LB} = G_{LB}(m) \odot [\Pi_L(m)]^{-1}$$

$$Q^{[m]}_{SL} = Q^{[m]}_{LS} \cdot [\Pi_S(m-1)][\Pi_L(m-1)]^{-1}$$

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

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

$$Q^{[m]}_{LNQ} = \begin{bmatrix} A(1) & ... & A(k) & ... & A(m-1) \\ 0 & ... & 0 & ... & 0 \\ 0 & ... & 0 & ... & 0 \end{bmatrix};$$

$$A(k) = \frac{q_{BND \to G}(k) \cdot [\Pi_{NQ(k)}(k)]}{[\Pi_G(k)]}$$

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

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

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

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

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

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.

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

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:

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

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

4.3.5. 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 $BB(m)$.

In particular, the upstream pseudo-machine $M^u(m)$ is characterized by state space $S^u(m) = [L^{[m]}, S^{[m]}, NQ^{[m]}]$. Similarly, the downstream pseudo-machine $M^d(m)$ is characterized by the state space $S^d(m) = [L^{[m+1]}, B^{[m+1]}]$. A schematic representation of the relation between the pseudomachines at buffer-level and the Integrated Machines at machine-level is provided Fig. 9.

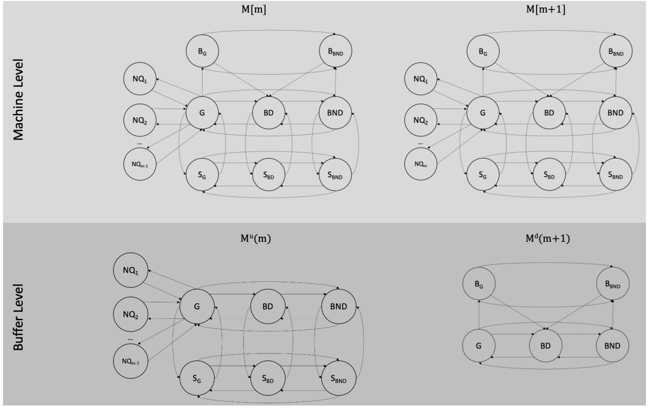

Figure 9: Relation between the Markov Chains of Integrated Machines (machine-level) and Pseudo-machines (buffer-level).

4.3.6. Convergence Algorithm

Forward and backward analysis is performed following methodology proposed by [4].

1. For $m = 1, ..., M$, Integrated Machine $M[m]$ are initialized based on $M\{m\}$.
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 $M^d(m)$.
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 $M^u(m)$.

The condition for termination:

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

4.4. Performance Measures

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

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

Instead the total rejected wafers:

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

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:

$$TH_{BND}^{[M]} = \sum_i \mu(S^{[M]}[i]) \cdot \Pi(S^{[M]}[i])$$

$$with\ i \in S = \{BND, NQ_1, ..., NQ_{M-1}\}$$

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

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

The average buffer level is computed as follow:

$$\overline{x}[m] = x[m] \cdot \int_0^N f(x, S)dx$$

5. Numerical results

Analysis of convergence of decomposition algorithm is assessed and a comparison with *Simulink* Discrete event simulator (DES) is performed. For the sake of compactness numerical results and tables are omitted in this summary. Convergence of the method is achieved always and with a maximum of 11 iteration for 9M8B case, considering a precision of 10^{-15}.This is because the variability in the system is really low. Instead error over performance parameters between model and DES are approximately 2.5%.

6. 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 are scaled but realistic. The objective is to analyze the impact of the different inspection policies to the system performance and dynamics.
Line under study has same structure as in Fig.6 but is a sequence of 4 layers, i.e. 4 photolithography&inspection stages and 3 buffers (4M3B Line).For following analysis has been set up an

optimization problem to evaluate different system configurations, and yet solved by a genetic algorithm:

$$
\begin{aligned}
maxZ = & C_G \cdot TH_G(\underline{\%}, \underline{N}) - \underline{C'_{BD}} \cdot \underline{TH_{BD}}(\underline{\%}, \underline{N}) \\
& - C_{BND} \cdot TH_{BND}(\underline{\%}, \underline{N})
\end{aligned}
$$

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.

configuration parameters:

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

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

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

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

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.

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

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

6.1. As-is issues

Current as-is inspection policy is to measure a certain number of markers at full capacity (100% markers). As it is shown in Figure 10 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.

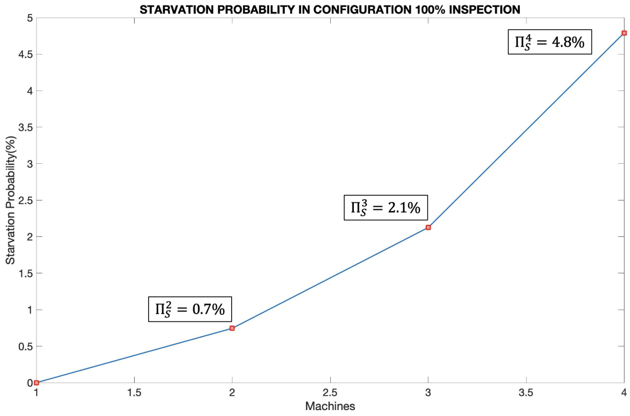

Figure 10: Starvation Propagation.

A trade-off between productivity and quality could be achieved, as an example for fixed $K_1 = 1$,$K_2 = 7.5$ $N = 300$, optimal configuration achieved:

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

Table 1: Optimal configuration

That improves system's performances with a slight degradation on output quality.

	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 2: Performance comparison

6.2. Sensitivity analysis

Sensitivity analysis is performed over cost parameters K_1 and K_2. It is well clear that K_1 is not influencing the system's response so that the only meaningful parameter to be assessed is K_2 (BND cost). It is set an analysis to look for the overall best configuration given K_2, in which each single machine of the configuration can have its own percentage of markers,in Fig. 11 results are displayed.

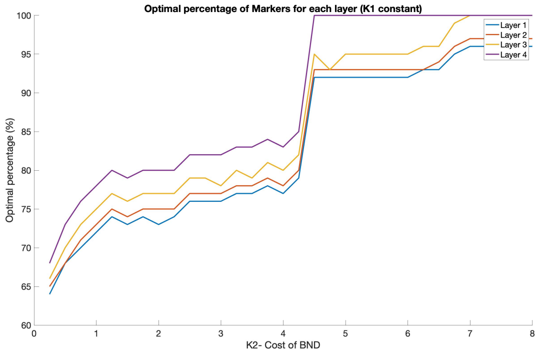

Figure 11: sensitivity analysis.

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.

6.3. Importance of optimal set of measurement markers

Given a percentage $\%^{obj}$ of the total amount of markers available it is possible to select any set/combination of markers $F(\%^{obj})$.
To know details on how optimal selection is performed please refer to [5].
Now, given a $\%^{obj}$, it is presented a comparison of performance results between the traditional operations, using the best set of markers $F^*(\%^{obj})$ and using a generic set of markers $F(\%^{obj})$.
It is assumed that using a generic set of markers $F(\%^{obj})$, probability of not detecting a bad layer $P(BND)$ Fig. 3 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 $F(\%^{obj})$.

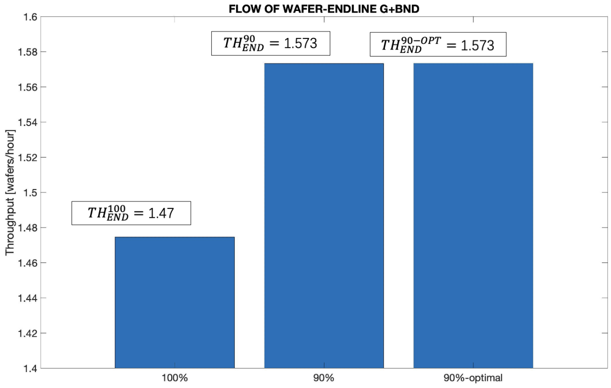

Figure 12: Throughput endline $TH_{G\&BND}$

Fig. 12 shows the throughput of wafers flowing from the last machine of the system in the three different configurations where all machines have the same number of markers selected.
It is evident that inspection of 100% of available markers increases cycle time thus throughput of the line is low, on the other hand decreasing percentage allows the whole line to produce faster.

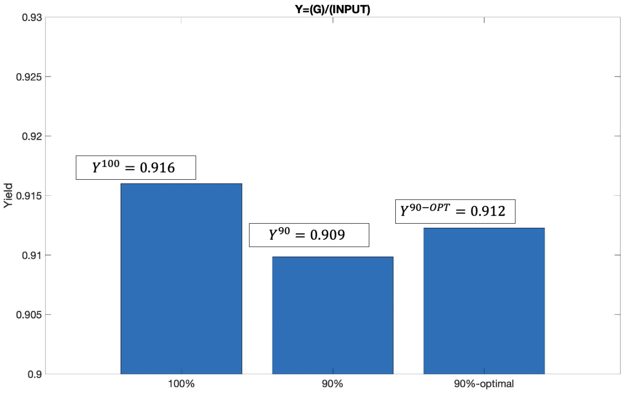

Figure 13: Yield: as $\frac{TH_G}{TH_{INPUT}}$

Fig. 13 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.

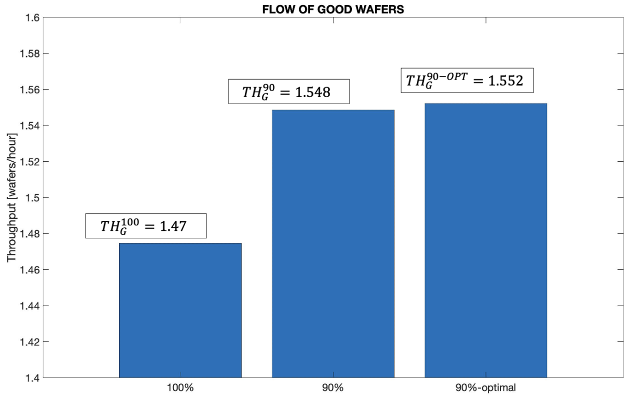

Figure 14: Throughput endline of good parts TH_G

Lastly it is possible to show 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 used is the optimal one Fig. 14.

7. Conclusions

In this thesis, it is developed an approximate analytical method for performance evaluation of asynchronous production lines where machines and lithography machines are controlled through robust control on the overlay error correction, where 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 and system's propagation of quality errors is modeled.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, and could improve the the traditional inspection policy.

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. 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.

References

[1] Marcello Colledani and Tullio Tolio. Performance evaluation of production systems monitored by statistical process control and on-line inspections. *International Journal of Production Economics*, 2009.

[2] Dragan Djurdjanović, Asad Ul Haq, Maria Chiara Magnanini, and Vidosav Majstorović. Robust model-based control of multistage manufacturing processes. *CIRP Annals.*, 2019.

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

[4] Maria Chiara Magnanini and Tullio Tolio. A markovian model of asynchronous multistage manufacturing lines fabricating discrete parts. *Journal of Manufacturing Systems*, 2023.

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

[6] Huidong Zhang, Tianheng Feng, and Dragan Djurdjanovic. Dynamic down-selection of measurement markers for optimized robust control of overlay errors in photolithography processes. *IEEE Transactions on Semiconductor Manufacturing*, 2022.