Title

First-Principles and Machine Learning Modeling for Design and Operation of Area-Selective Atomic Layer Deposition

Author

Tom, Matthew Cheuk-Woh

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

First-Principles and Machine Learning Modeling for Design and Operation of Area-Selective Atomic Layer Deposition

A dissertation submitted in partial satisfaction of the
requirement for the degree Doctor of Philosophy
in Chemical Engineering

by

Matthew Cheuk-Woh Tom

2024

ABSTRACT OF THE DISSERTATION

First-Principles and Machine Learning Modeling for
Design and Operation of Area-Selective Atomic Layer Deposition

by

Matthew Cheuk-Woh Tom

Doctor of Philosophy in Chemical Engineering
University of California, Los Angeles, 2023
Professor Panagiotis D. Christofides, Chair

Semiconductor manufacturing comprises nearly 500 processing steps, where products rely on stringent design criteria to have high-performance characteristics. One of these processing steps includes the fabrication of high-κ oxide films on the surfaces of transistors to minimize current and heat losses, and short-channel effects, which are detrimental to semiconductor longevity. These films demand thicknesses in the nanoscale that are constructed using sequential cycles of atomic layer deposition (ALD) and atomic layer etching (ALE), where precise monolayers of substrate film are deposited and exhibit self-limiting behavior. However, notable challenges in industrial practice include maintaining the accuracy of the deposition and etching processes and the uniformity of the films that are produced, identifying the operating conditions that contribute to optimal product conformation, and developing reactors that maximize the productivity of these atomic layer processes. Additionally, there is insufficient and available data for these processes in industry, which makes their characterization and optimization an obstacle for researchers. Thus, *in silico* modeling has paved the way for producing data that is reflective of data observed in industrial practice. This simulated data is produced through a multiscale computational fluid dynamics

framework that combined microscopic, mesoscopic, and macroscopic phases throughout various time and length scales. This work encompasses several disciplines from reaction characterization through *ab initio* molecular dynamics simulations, rudimentary chemical kinetics laws, and kinetic Monte Carlo methods, reactor optimization and design through computational fluid dynamics, and feedback-based run-to-run and online process control with an application to machine learning for a plethora of atomic layer processes.

The dissertation of Matthew Cheuk-Woh Tom is approved.

Carlos Morales-Guio

Dante Simonetti

Philippe Sautet

Panagiotis D. Christofides, Committee Chair

University of California, Los Angeles

2024

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my gratitude toward my Principal Investigator and advisor, Professor Panagiotis Christofides, who supported me throughout my four years in his research group. Professor Christofides served as a huge mentor, researcher, and role model. Professor Christofides' continuing encouragement, especially when encountering obstacles in my research, also helped me to grow professionally by finding ways to solve critical problems. I am appreciative of his tremendous involvement in my academic career, and my ambition to continue future work in computational fluid dynamics modeling and high-performance computing is owed in part to our work in multiscale modeling.

I am grateful for my family, who have encouraged and supported me throughout my academic career and life. I would like to thank my parents, Harold Tom and Helen Law, and my sisters, Michelle Tom and Karen Tom, for their unconditional love. In particular, I would like to thank my grandmother, Jane Tom, who provided me with housing during my time in research.

I would also like to express my gratitude toward my colleagues and friends, especially Dr. Sungil Yun who served as a mentor and guide when we began collaborating for 3 years. Additionally, Professor Gerassimos Orkoulas had a profound role in my research and I thank him for his support throughout my various research projects. I would also like to extend my appreciation toward my collaborators, Henrik Wang, Feiyang Ou, and Dr. Derek Richard, for their endless guidance and support. Lastly, I would like to acknowledge the members of the Christofides' Research Group: Dr. Fahim Abdullah, Dr. Junwei Luo, Aisha Analjdi, Dr. Yi Ming Ren, Dr. Scarlett Chen, Dr. Mohammed Alhajeri, Berkay Çitmaci, Vito Canuso, and Atharva Suryavanshi.

I would also like to thank Professor Carlos Morales-Guio, Professor Dante Simonetti, Professor Philippe Sautet, and Professor Mathieu Bauchy for serving on my doctoral committee.

I would like to extend my appreciation for the support of the UCLA Hoffman2 Cluster Support Team including Raffaella D'Auria, Shao-Ching Huang, John Pedersen, Brian Pape, and Charles Peterson, for their technical support. Without these staff members, I would have never been able

to resolve some of the most technical problems in high-performance computing.

Additionally, I gratefully acknowledge the financial support for my research from the National Science Foundation (NSF). I would not have been able to elevate my work without their generous contributions.

This dissertation includes the following manuscripts that serve as the foundation of this work:

Chapter 2 contains a version of: Tom, M., Wang, H., Ou, F., Yun, S., Orkoulas, G., & Christofides, P. D., 2023. "Computational fluid dynamics modeling of a discrete feed atomic layer deposition reactor: Application to reactor design and operation." *Computers & Chemical Engineering*, 178, 108400.

Chapter 3 contains a version of: Wang, H., Tom, M., Ou, F., Orkoulas, G., & Christofides, P. D., "Multiscale Computational Fluid Dynamics Modeling of an Area-Selective Atomic Layer Deposition Process Using a Discrete Feed Method," *Digital Chemical Engineering*, 10, 100140, 2024.

Chapter 4 contains a version of: Tom, M., Wang, H., Ou, F., Orkoulas, G., & Christofides, P. D., "Machine Learning Modeling and Run-to-Run Control of an Area-Selective Atomic Layer Deposition Spatial Reactor," *Coatings*, 14, 38, 2024.

Chapter 5 contains a version of: Wang, H., Tom, M., Ou, F., Orkoulas, G., & Christofides, P. D., "Integrating Run-to-Run Control with Feedback Control for a Spatial Atomic Layer Etching Reactor," *Chemical Engineering Research & Design*, 203, 1-10, 2024.

Curriculum Vitae

Education

University of California, Los Angeles — *Sept. 2020 – Mar. 2022*
M.S., Chemical Engineering — Los Angeles, CA

California State Polytechnic University, Pomona — *Sept. 2015 – May 2019*
B.S., Chemical Engineering — Pomona, CA
Minor, Chemistry

Experience

HRL Laboratories, LLC — *May 2024 – Present*
Scientist IV – Computational Engineer — Malibu, CA

Freudenberg Medical, LLC — *Mar. 2020 – Jul. 2020*
Quality Specialist — Baldwin Park, CA

Publications

1. H. Wang, **M. Tom**, F. Ou, G. Orkoulas, & P. D. Christofides, "Integrating Run-to-Run Control with Feedback Control for a Spatial Atomic Layer Etching Reactor," *Chem. Eng. Res. & Des.*, 203, 1-10, 2024.

2. H. Wang, **M. Tom**, F. Ou, G. Orkoulas, & P. D. Christofides, "Multiscale Computational Fluid Dynamics Modeling of an Area-Selective Atomic Layer Deposition Process Using a Discrete Feed Method," *Dig. Chem. Eng.*, 10, 100140, 2024.

3. **M. Tom**, H. Wang, F. Ou, G. Orkoulas, & P. D. Christofides, "Machine Learning Modeling and Run-to-Run Control of an Area-Selective Atomic Layer Deposition Spatial Reactor," *Coatings*, 14, 38, 2024.

4. F. Ou, F. Abdullah, H. Wang, **M. Tom**, G. Orkoulas, & P. D. Christofides, "Sparse Identification Modeling and Predictive Control of Spatially-Distributed Processes," *Chem. Eng. Res. & Des.*, 202, 1-11, 2024.

5. **M. Tom**, H. Wang, F. Ou, S. Yun, G. Orkoulas, & P. D. Christofides, "Computational Fluid Dynamics Modeling of a Discrete Feed Atomic Layer Deposition Reactor: Application to Reactor Design and Operation," *Comp. & Chem. Eng.*, 178, 108400, 2023.

6. Yun, S., H. Wang, **M. Tom**, F. Ou, G. Orkoulas and P. D. Christofides, "Multiscale CFD Modeling of Spatial Area-Selective Thermal Atomic Layer Deposition: Application to Reactor Design and Operating Condition Calculation," *Coatings*, 13, 558, 2023.

7. **M. Tom**, S. Yun, H. Wang, F. Ou, G. Orkoulas and P. D. Christofides, "Multiscale Modeling of Spatial Area-Selective Thermal Atomic Layer Deposition," *Computer Aided Chemical Engineering*, 52, 71-76, 2023.

8. D. Richard, **M. Tom**, J. B. Jang, S. Yun, P. D. Christofides, & C. Morales-Guio, "Quantifying transport and electrocatalytic reaction processes in a gastight rotating cylinder electrode reactor via integration of computational fluid dynamics modeling and experiments," *Electrochimica Acta*, 440, 141698, 2023.

9. **M. Tom**, S. Yun, H. Wang, F. Ou, G. Orkoulas, & P. D. Christofides, "Machine learning-based run-to-run control of a spatial thermal atomic layer etching reactor," *Comp. & Chem. Eng.*, 168, 108044, 2022.

10. S. Yun, F. Ou, H. Wang, **M. Tom**, G. Orkoulas, & P. D. Christofides, "Atomistic-Mesoscopic Modeling of Area-Selective Thermal Atomic Layer Deposition," *Chem. Eng. Res. & Des.*, 188, 271-286, 2022.

11. S. Yun, **M. Tom**, G. Orkoulas, & P. D. Christofides, "Multiscale Computational Fluid Dynamics Modeling of Spatial Thermal Atomic Layer Etching," *Comp. & Chem. Eng.*, 163, 107861, 2022.

12. S. Yun, **M. Tom**, F. Ou, G. Orkoulas & P .D. Christofides, "Multivariable Run-to-Run Control of Thermal Atomic Layer Etching of Aluminum Oxide Thin Films," *Chem. Eng. Res. & Des.*, 182, 1-12, 2022.

13. S. Yun, **M. Tom**, F. Ou, G. Orkoulas & P. D. Christofides, "Multiscale Computational Fluid Dynamics Modeling of Thermal Atomic Layer Etching: Application to Chamber Configuration Design," *Comp. & Chem. Eng.*, 161, 107757, 2022.

14. S. Yun, **M. Tom**, J. Luo, G. Orkoulas & P. D. Christofides, "Microscopic and Data-Driven Modeling of Thermal Atomic Layer Etching of Aluminum Oxide Thin Films," *Chem. Eng. Res. & Des.*, 177, 96-107, 2022.

Awards

Chemical & Biomolecular Engineering: *May 2023*
Teaching Assistant of the Year
University of California, Los Angeles

Chapter 1

Introduction

1.1 Background

Recent advancements in high-performance electronics has been rising in the last decade, leading to major improvements in various applications including computing [43], smart technology [52, 58], biotechnology [57, 128], telecommunications [89], and household appliances [2], as depicted in Fig. 1.1. These devices have improved in efficiency and accessibility as a consequence of semiconductors, which play a critical role in both the performance of these electronic devices and the availability of electronics in the global market. Some of the most high-performance semiconductors have high inventory turnover [60], where reduced supply and increased demand have led to a domino effect, with global shortages observed for electronics that depend on these materials [111]. This nonproductive manufacturing of semiconductors can be attributed to the 500 processing steps required to develop the finished product [95]. Transistors, nanoscale components of semiconducting wafers, have a critical role in semiconductor performance, which is characterized by the size, the number of transistors occupying the wafer [80], the power and current efficiency [103], and computing capability. Semiconductor performance can be enhanced by reducing the length scales, changing the composition (e.g., metal oxide) and architectures like the

Fin Field-Effect (FinFET) [50] and Gate-All-Around (GAA) [69] illustrated in Fig. 1.2, and stacking transistors. However, the fabrication for these transistors is an arduous task requiring numerous steps to achieve desired nanopatterning that is conducive to transistor stacking, and ultimately densification. One step in transistor development is the depositing and etching of thin, high-κ oxide films onto the surfaces of nanowires, which allow electron transport between the source and drain of the transistor. The oxide films play a crucial role in preserving current and heat while simultaneously reducing short-channel effects that degrade semiconductor performance [6, 44]; however, various factors influence the uniformity and thickness of these films. Thus, there is growing interest in pursuing research to design and optimize the processes that contribute to (1) improvements in the conformity of the thin films, which maximize overall semiconductor efficiency and performance and (2) an increase in the productivity of the transistors to meet the rising demand for high-performance wafers.

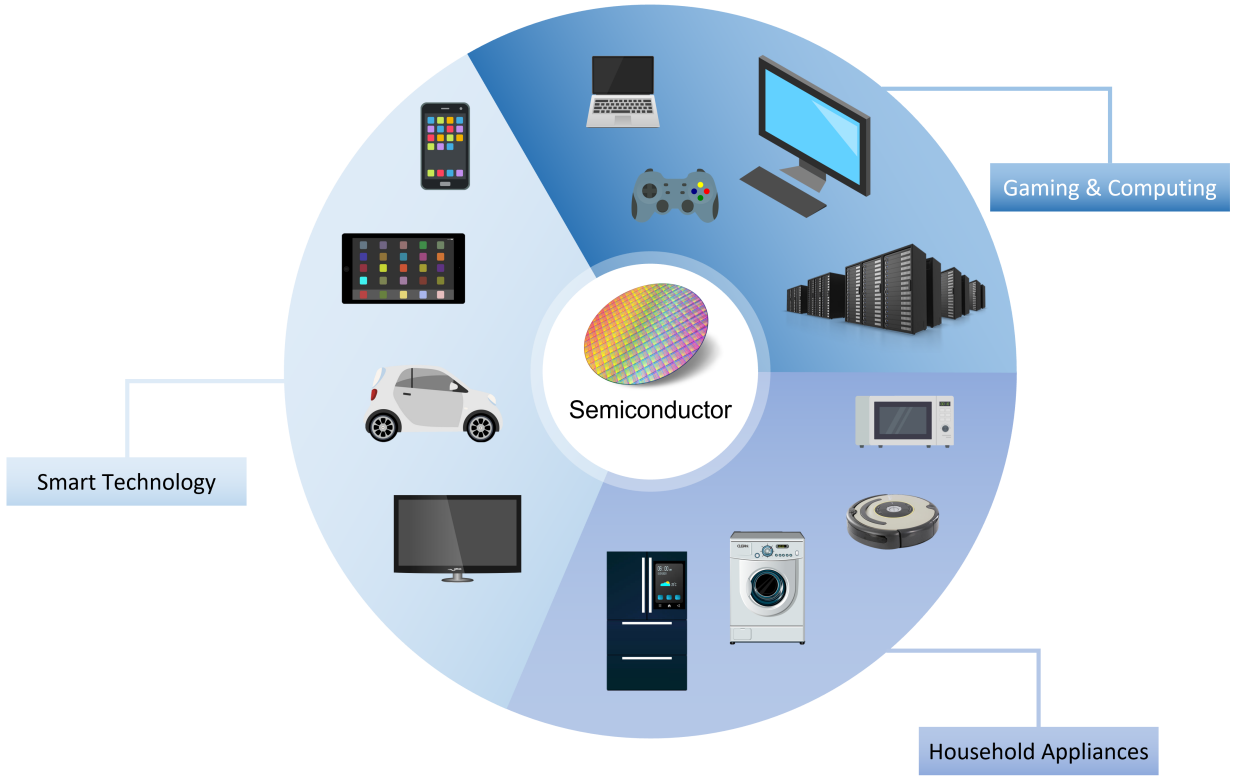
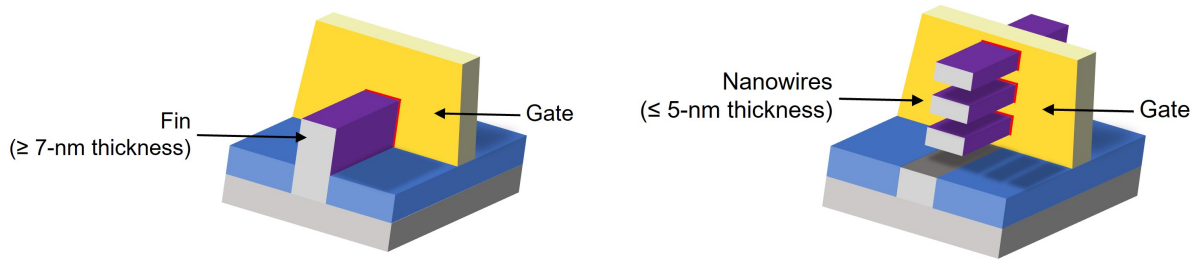


Figure 1.1: A chart depicting the consumption of semiconductors in the electronics industries.



(a) (b)

Figure 1.2: Schematic diagrams of a Fin Field-Effect Transistor (FinFET) and a Gate-All-Around (GAA) transistor in (a) and (b), respectively. FinFETs only permit thickness sizes above 7-nm due to the design of the transistor. Using multiple nanowires allow more current transfer for GAA transistors, which allow sub-5-nm thicknesses. The exterior surface of the fin and nanowires represents the metal oxide insulator.

1.2 Atomic Layer Processing

In the fabrication of these thin films that are characterized by nanoscale dimensions, atomic layer deposition (ALD) and atomic layer etching (ALE) are employed due to their precise controlling of thin-film thicknesses. ALD is a bottom-up procedure where oxide films (e.g., aluminum oxide and silicon oxide) are deposited in monolayers, where self-limiting behavior is observed and prevents permeation of reagent beyond the surface layer of the film [36]. Through this manner, the thin film retains a uniform surface layer and promotes self-aligned structures for transistor stacking. General ALD processes comprise a two-step cycle using gaseous reagents with an initial precursor adsorption step that modifies the surface layer of the substrate and an oxidation step that yields a metal oxide surface. To ensure that reagent and byproduct intermixing is minimized and that self-limiting behavior is observed, cut-in purging is employed between each step, which can be time-consuming depending on the reactor design. Additionally, byproduct removal can be ensured by operating the process under high temperatures that volatilizes these species. This ALD process is depicted in Fig. 1.3, which exemplifies this cyclical procedure.

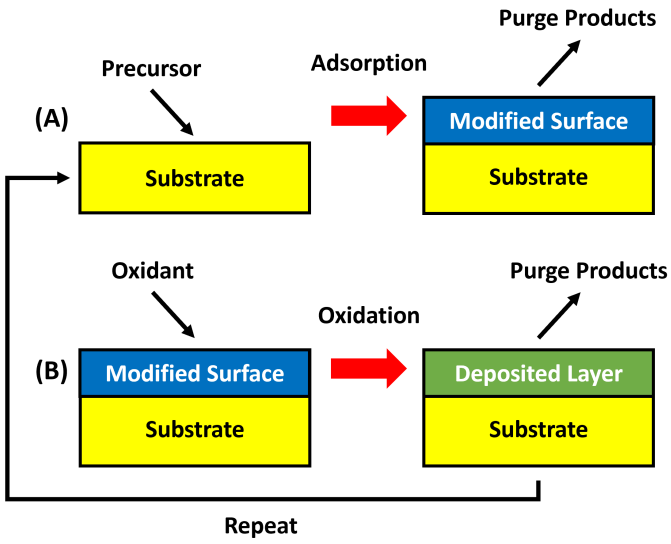


Figure 1.3: A two-step atomic layer deposition process with an initial precursor adsorption and oxidation step. Diagram adapted from George, S. M. [37].

However, ALD may lack accuracy, which would require numerous pre-processing (e.g., lithography) and post-processing (e.g., etching) steps to yield a conforming product. Particularly, industry utilizes a top-down, atomic layer etching (ALE) procedure to allow continued downscaling of film thicknesses and improved film uniformity. In contrast with ALD, generalized ALE schemes adopt a two-step mechanism with an initial precursor modification step followed by an adsorption step with a secondary reagent, which is illustrated in Fig. 1.4. A recurring issue in industrial practice is the time-consuming and laborious process for employing ALD and ALE, where hundreds of cycles conducted sequentially are performed. Such procedure risks a potential for defect formation (e.g., nonuniform deposition and edge growth) as depicted in Fig. 1.5 and reduces the productivity of developing the finished thin-film coating.

One approach to reducing the number of lithographic and ALE steps is to study more *selective* deposition processes that provide greater surface area control for adsorption reactions. Area-selective atomic layer deposition (AS-ALD) is appealing to industry due to the process being dependent on the chemistry of the reactions rather than the operating conditions, by employing

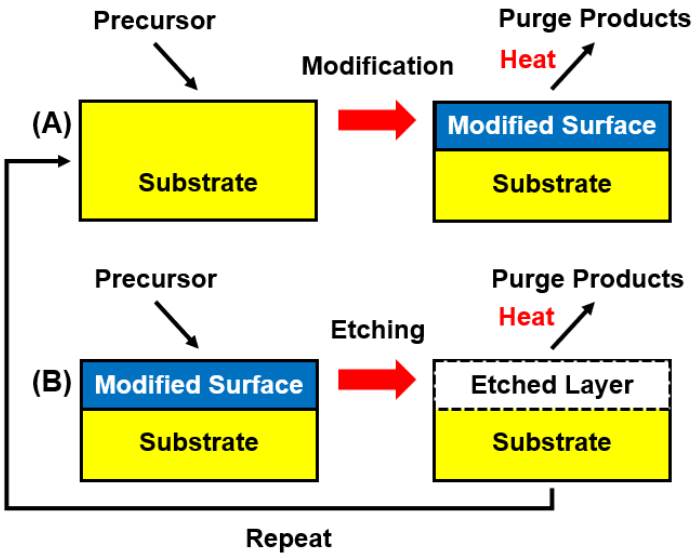
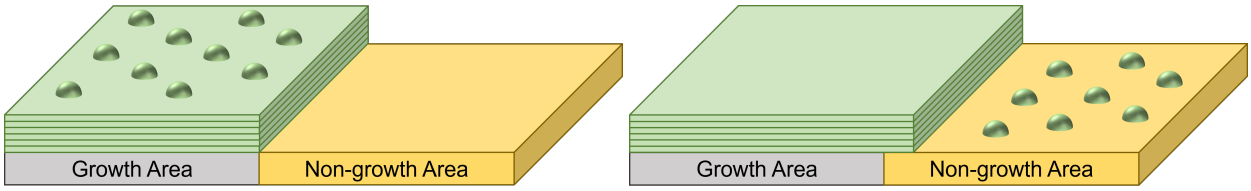


Figure 1.4: A two-step atomic layer etching process with an initial precursor modification and etching step. Diagram adapted from George, S. M. [37].



(a) (b)

Figure 1.5: Depictions of non-ideal atomic layer deposition that leads to (a) nonuniform surfaces and (b) excess growth on neighboring non-growth areas of the substrate.

chemoselective reactions that interact with the surface (i.e., the growth area) and neglect surface growth on undesired locations of the substrate (i.e., the non-growth area). AS-ALD integrates a surface deactivation step to the ALD mechanism, which inhibits deposition on non-growth areas of the substrate through a chemoselective inhibitor [73]. The adsorption of the inhibitor on the non-growth area serves as a protective group to prevent subsequent precursor and reagent interaction with the non-growth area. This three-step AS-ALD process is illustrated in Fig. 1.6 and depicts the substrate comprising non-growth and growth areas that are synonymous of substrate compositions utilized in industrial practice. While there has been growing interest in pursuing AS-ALD,

there is insufficient experimental data to enable the expansion of these processes into industrial settings. Thus, there is motivation to pursue research aimed at data generation and optimization for the scale-up of AS-ALD processes.

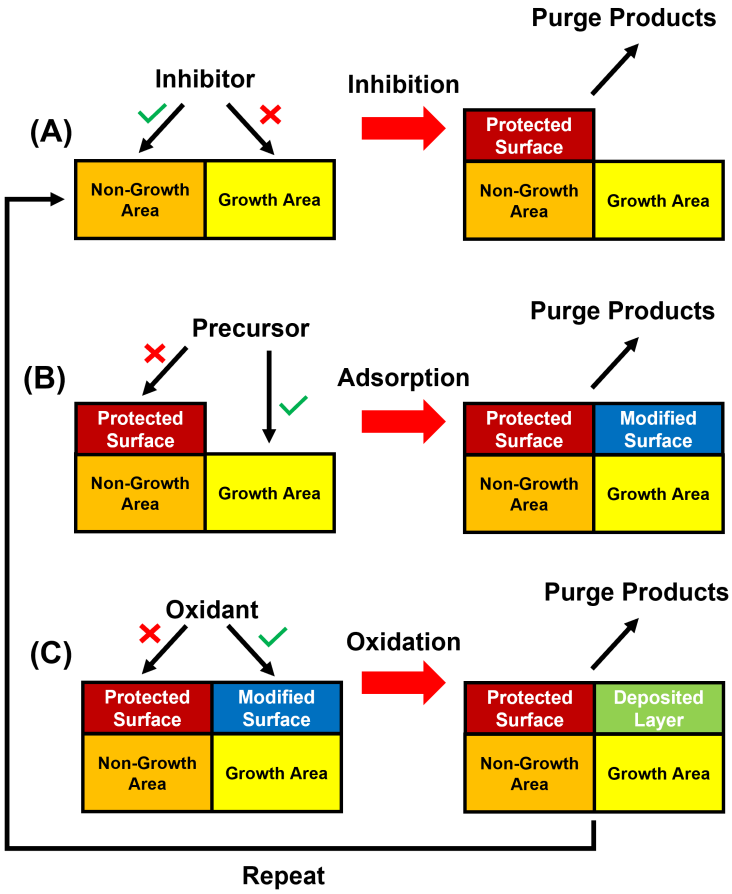


Figure 1.6: A three-step area-selective atomic layer deposition process with an initial precursor modification and etching step. Diagram adapted from George, S. M. [37].

1.3 Multiscale Computational Fluid Dynamics Modeling

A growing issue is the need to supplement incomplete data sets to construct empirical models for AS-ALD processes. Due to the cost and time constraints of scaling experiments to industrial scales, there has been a desire to approach this issue with a more economical solution [42]. A solution to this obstacle is to generate synthetic or simulated data through computational tools and

programs that are cross-validated with industrial or experimental data, as described in Fig. 1.7. The replication of AS-ALD, however, requires a priori knowledge of the various phases influencing the conformity of the substrate due to the occurrence of defects in several time and length scales [75]. Therefore, a multiscale model is applicable for the replication of AS-ALD processes, which comprises microscopic, mesoscopic, and macroscopic domains with each attributed to a time and length scale, and is illustrated in Fig. 1.8.

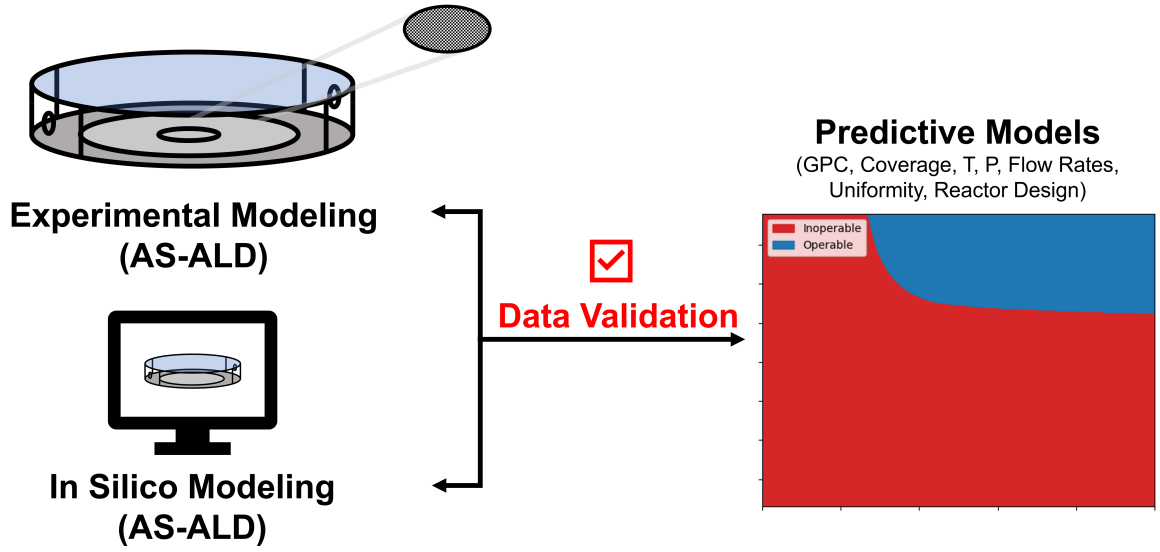


Figure 1.7: A depiction of the use of *in silico* modeling to develop models that establish multivariate input-output relationships that are cross-validated with experimental data and theoretical trends.

The first layer of the multiscale model is the microscopic (i.e., atomistic) layer in which the properties (e.g., thermophysical and kinetic) of molecular and crystalline species, and their associated reactions with one another, are determined through *ab initio* molecular dynamics simulations including density functional theory (DFT), nudged elastic band (NEB) methods, and quasi harmonic approximations (QHA). One pitfall of microscopic modeling is the need for a priori knowledge of reaction pathways for the AS-ALD processes for integration to the mesoscopic surface simulation. For modeling the stochastic behavior of molecular interactions, the kinetic Monte Carlo (kMC) approach is applicable for simulating surface chemistry along the substrate. Lastly,

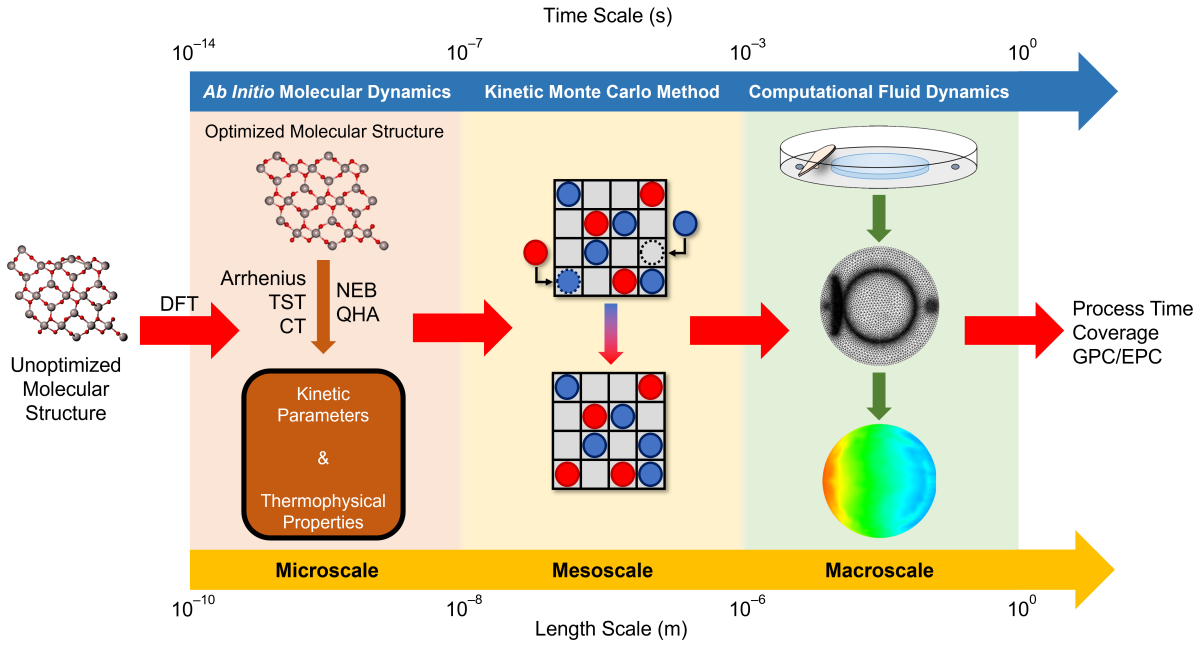


Figure 1.8: A generalized multiscale model relating various domains attributed to time and length scales that allow the scale-up of these processes and the generation of data that is reflective of observed experimental data.

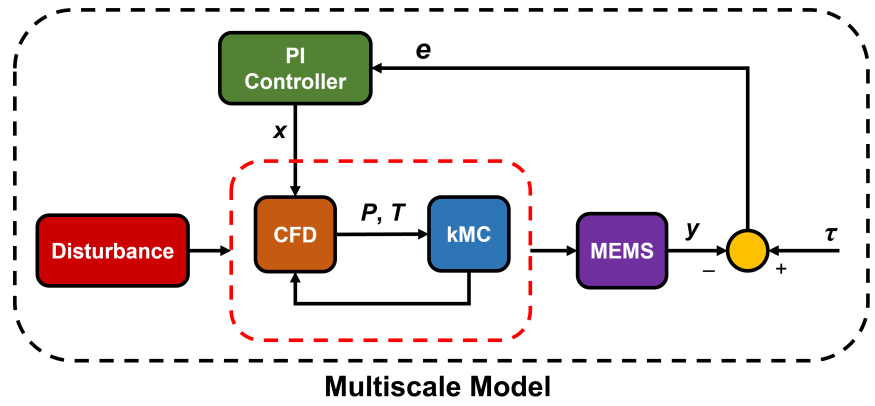
the macroscopic phase employs computational fluid dynamics (CFD) to study the flow behavior of gases in a reactor model, which is necessary to control the uniformity of flow on the substrate surface. The role of the reactor geometry will have a substantial impact on the transport of gases and the productivity of the process.

A frequently visited issue with multiscale simulation is the development of the multiscale programming logic that will enable the conjunction of the various domains. Although the microscopic simulations are treated independently of the mesoscopic and macroscopic simulations, a growing issue is the cyclical framework demanded by connecting the mesoscopic kMC and macroscopic CFD simulations. For example, a previous work [126] used simplified boundary conditions to reduce the complexity of the simulation, at a cost of generating potentially inaccurate data. On the other hand, a prior work [124] utilized a more complex cross-platform programming strategy that conjoins the kMC simulation to the CFD simulation in an external environment that results in

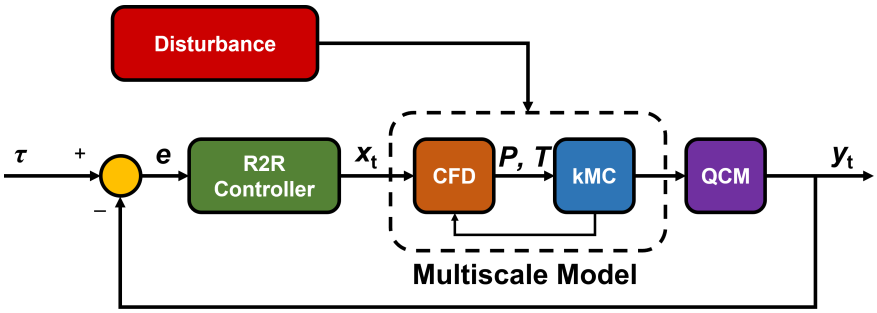
more accurate computation, but at a cost of requiring more simulation time. Thus, there are several options to weigh when considering the organization of the multiscale model.

1.4 Online Feedback and Run-to-Run Control

A recurring issue in semiconductor processing is the observance of defective devices that occur sporadically, making the causation of disturbances difficult to determine and correct. Thus, the incorporation of online feedback and offline run-to-run (R2R) control is necessary to monitor the processes that encounter perturbations in operating conditions. Online feedback controllers such as the proportional-integral (PI) controller can accomplish online monitoring of operating conditions by adjusting input parameters (e.g., reagent flow rates and substrate velocities) supplied to the process, and is able to mitigate unprecedented disturbances such as process shifts and noise. Typically, a Micro Electro Mechanical System (MEMS) comprising numerous sensors are employed to measure reactor operating conditions, as depicted in Fig. 1.9a. In addition to online feedback control, a coupled offline, R2R controller is necessary to monitor parameters that cannot be measured during the process due to the short time scales, such as the growth per cycle (GPC) for deposition processes and the etch per cycle (EPC) for etching processes, which are measured through Quartz Crystal Microbalances (QCM), as described in Fig. 1.9b. One limitation of R2R control is the ability to correct major disturbance shifts, and is also constrained to minor process drifts, which is why a conjoined online feedback and offline R2R control is desired. However, these controllers rely on abundant data sets that enable optimizable adjustment to input parameters through appropriate tuning methodologies. Thus, the integration of PI and R2R control to multiscale CFD simulations can circumvent this issue with insufficient tuning data, while also enabling the development of controllers prior to use in industrial practice.



(a)



(b)

Figure 1.9: Process diagrams of (a) an online feedback (e.g., proportional-integral, PI) and (b) a run-to-run controller that are conjoined to the multiscale models for monitoring the error, e, of the output variables, y, from the process setpoint, τ, by adjusting input variables, x.

1.5 Remarks on Computing Efficiency

A disadvantage of multiscale CFD modeling is the need for robust computing resources to leverage computing efficiency to facilitate the process of collecting data. Depending on the fineness of mesh discretization used in the CFD reactor model, the complexity of the reaction pathways and molecular structures, and the numerical solvers employed, the overall multiscale simulation will be computationally incapable of producing data at a rate that is comparable to data obtained experimentally and industrially. Additionally, increasing central processing unit (CPU) cores, is limited by Amdahl's Law, which forecasts a bounded computing time despite improving parallelization [106]. One approach to improve parallelized computations is to integrate graphics

processing units (GPU) to accelerate simulations and remove computing burden on CPUs, which is effective for a plethora of applications including CFD, machine learning, and artificial intelligence [98]. Additionally, the application of machine learning is capable of reducing the need for repetition of multiscale CFD modeling if sufficient data is previously generated to enable the construction of machine learning models. Lastly, migrating from interpretive programming languages (e.g., Python and MATLAB) and resorting to compiler-based languages (e.g., C/C++ and Fortran) can lead to more efficient parallel simulations. This can be further accelerated through the compilation with open-source solver libraries that are associated with the computer processor.

1.6 Motivation

This dissertation aims to examine the aforementioned challenges encountered in transistor fabrication through *in silico*, multiscale modeling and control of an AS-ALD process while considering the role of computational efficiency. First, the relationship between various delivery geometries for a discrete feed reactor and the uniformity and transport of reagent are explored through CFD simulations. Next, a multiscale CFD model for the AS-ALD of a SiO_2/Al_2O_3 is explored by programming the kMC simulation within the CFD environment through C-language-based user-defined functions (UDFs) to reduce simulation time. The dissertation will also explore the use of machine learning to construct predictive models in the generation of spatiotemporal pressure data for a spatial rotary reactor geometry. Additionally, to further accelerate simulations, the work will study the incorporation of accelerator toolkits through Intel's Math Kernel Library (MKL) via Intel oneAPI. Lastly, various control algorithms are proposed for R2R and online feedback control through exponentially weighted moving averages (EWMAs), machine learning, and PI control.

1.7 Dissertation Structure

This dissertation proposes numerous advancements in the development of AS-ALD processes through the design of various reactor models to optimize the productivity and quality of wafers through multiscale CFD simulations. Additionally, the integration of control systems to these multiscale models that are capable of detecting and mitigating various disturbances are designed through appropriate tuning methodologies. The primary objectives of this dissertation are as follows:

1. To study the impact of reactor configuration design that is applicable to models utilized in industrial practice on reagent transport.

2. To develop multiscale CFD models that (1) produce data that is cross-validated with experimental findings and (2) generate data efficiently.

3. To determine alternative approaches for multiscale modeling through the use of machine learning by generating predictive models for time-series data.

4. To design control systems that employ online feedback and *ex situ* run-to-run control to detect and mitigate various disturbances.

The subsequent sections are organized as follows: **Chapter 2** studies the impact of reactor configuration (i.e., the reagent injection system) on the transport of reagent across the surface of an Al_2O_3/SiO_2 substrate through a macroscopic CFD simulation. Various injection plate geometries are designed with differing characteristic lengths to determine their impact on the speed and length of reagent transport for discrete pulsing times while maintaining constant operating conditions within the reactor. This three-dimensional CFD simulation is conducted through a procedure comprising first of reactor geometry development through computer-aided design (CAD) software, Ansys DesignModeler. Next, a reactor geometry discretization process is conducted through meshing software, Ansys Fluent, and employing various functions that preserve the quality of the mesh.

Lastly, a CFD simulation conducted through Ansys Fluent is constructed by defining pertinent operating conditions including laminar flow behavior, and isothermal and isobaric operation. This work explores the strategy of employing CFD simulation to optimize reactor geometry that results in high-throughput operation by minimizing processing time and in high surface film uniformity on the substrate surface.

Chapter 3 adopts the reactor configurations in Chapter 2 and proposes a three-dimensional multiscale CFD simulation that effectively conjoins the kMC and CFD simulations within the Ansys environment by employing user-defined functions (UDFs). While UDFs offer tremendous flexibility in customizing the CFD simulation, limitations in the number of customizable scalars led to a simplified kMC simulation that is independent of a grid, but rather utilizes a spatial-dependent or "bucket" scheme to study the temporal progression of the film surface coverage. To capture the spatiotemporal progression of the surface coverage, the substrate was partitioned into sections that allowed parallel kMC simulations to be carried out with the CFD simulation. Various multiscale simulations were conducted for each reactor configuration (at constant operating conditions) to study the processing time required to observe complete surface coverage and to maintain surface coverage uniformity.

Chapter 4 integrates a R2R controller to a previously developed three-dimensional multiscale model for a spatial, rotary reactor configuration of an AS-ALD process to detect and mitigate a minor shift disturbance. Due to long simulation times required to generate data that must be obtained sequentially, a machine learning model designed through a multiple-input-single-output (MISO) feedforward neural network (FNN) of time-series surface pressure data. In the design of the R2R controller, an exponentially weighted moving average (EWMA) of a linear regression of an input (substrate rotation velocity) and the output (growth per cycle, GPC) is utilized for adjusting the process input based on the deviation of the measured output from a user-defined setpoint. To resemble the uncertainty in the causes of disturbances, this work applied a minor constant shift pressure and kinetic disturbance to reduce the desired pressure and reaction rate. Through this

strategy, the R2R controller is designed to compensate for the effects of the disturbances by adjusting the reducing/increasing the rotation speed of the spatial reactor while the process is offline, which controls the substrate residence time or exposure time to the reagents.

Chapter 5 proposes the integration of an online feedback and R2R controller to a spatial reactor for an ALE process to compensate for the effects of drift and shift disturbances. Constant pressure drift and kinetic shift disturbances are introduced to the CFD and kMC simulations, respectively, as a means to introduce perturbations to the process. A previously developed two-dimensional, multiscale CFD model for the thermal ALE of Al_2O_3 is utilized to generate sufficient data to construct a linear regression for multiple inputs (substrate velocity and reagent flow rates) and output (etch per cycle, EPC). For the design of the R2R controller, this linear model is used to implement adjustment to the input parameters through an EWMA of the multiple-input-single-output (MISO) model while the process is offline. The EWMA method performs the necessary adjustment by accounting for the error between the measured EPC and the target EPC. To allow for input adjustment while the process is online, a proportional-integral (PI) controller is introduced to the CFD simulation by applying a UDF by measuring the deviation in the operating pressure of the reactor from the setpoint.

Chapter 6 provides a summary of the dissertation.

Chapter 2

Computational Fluid Dynamics Modeling of a Discrete Feed Atomic Layer Deposition Reactor: Application to Reactor Design and Operation

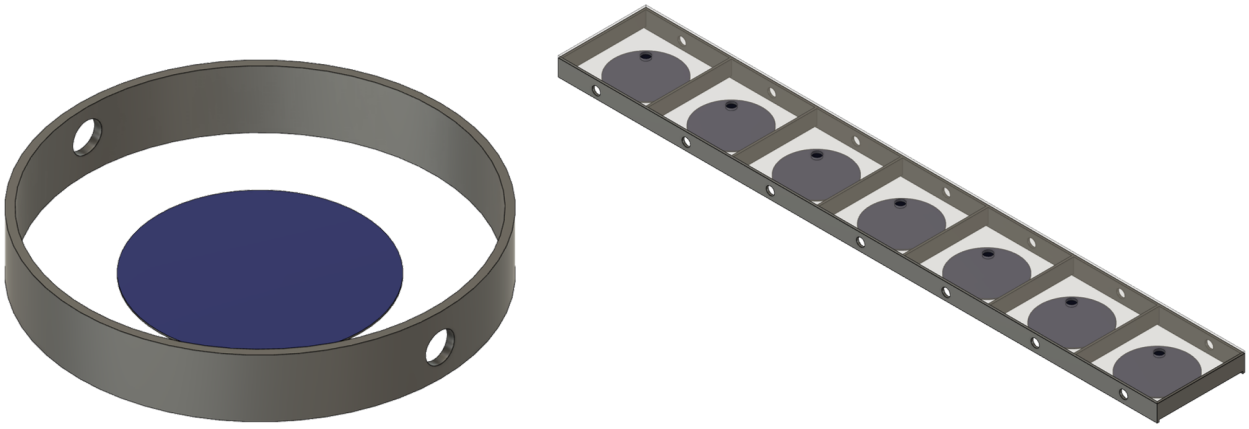
2.1 Introduction

Semiconductor manufacturing industries have faced obstacles attributed to the design of AS-ALD processes due to the lack of data for studying these processes macroscopically. The conformity of the product relies on the delivery system of reagent as well as operating conditions that are intended to purge undesired byproducts and unreacted reagent. Thus, there is motivation for determining the role of reactor geometry on the quality of the film by examining the fluid dynamics on the surface of the substrate. This work, however, neglects the role of the mesoscopic surface kinetics to simplify the simulation model.

The optimization of reactors and operating conditions are necessary for achieving greater prod-

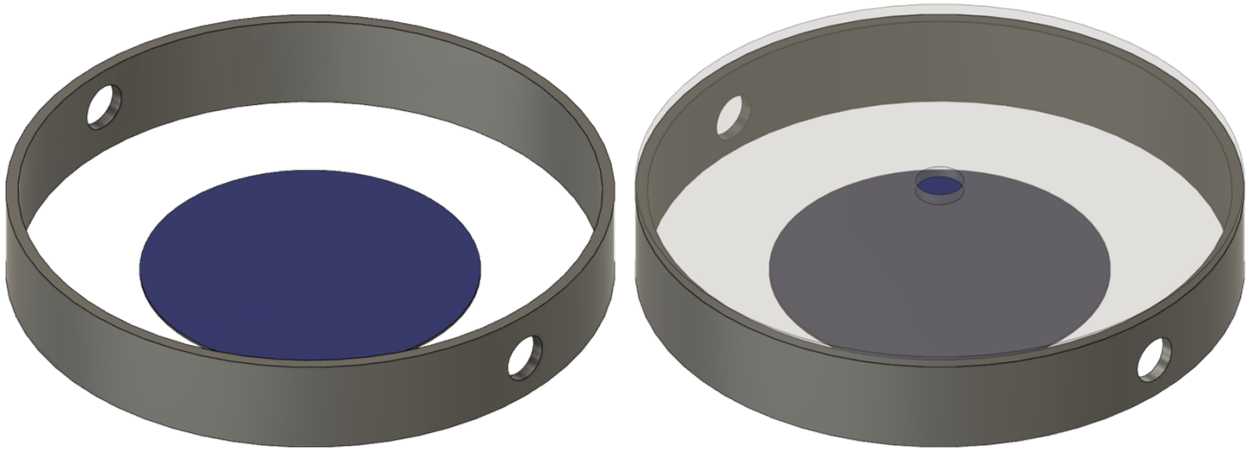
uct yield and oxide film quality. Various reactor modeling proposals and patents have been developed, particularly batch (stationary) [33] and spatial [25, 92] configurations as depicted in Fig. 2.1. Additionally, reactor models have been constructed with different approaches. Variations include the configuration of the reagent delivery system to the substrate through cross or perpendicular (overhead) flow distributions [33, 55] as shown in Fig. 2.2, continuous feed [83, 96] or discrete feed [66] pulses, and fluid partitioning by using dividers such as showerheads [61] and inclined plates [36] as illustrated in Fig. 2.3, or substrate holders [24] to minimize reagent concentration gradients on the surface of the substrate. While the development of these reactors is a first step toward their integration to industrial applications, *in silico* modeling [29] provides an effective approach to studying the behavior of the fluid dynamics for a variety of reactor models, particularly small reactor models, that will improve the process efficiency of the reactor and maximize reagent usage. For instance, spatial reactor models for AS-ALD and ALD perform poorly at attaining complete surface coverage due to the overdosage of reagent to the wafer surface, thereby risking the integrity of self-aligning transistors during this bottom-up fabrication procedure. In particular, this work will examine the fluid dynamics of a stationary-type reactor that employs a perpendicular feed mechanism with a showerhead distributor that will be used for an area-selective atomic layer deposition (AS-ALD) process.

Past works [74, 76, 77] have studied the effectiveness of integrating AS-ALD reaction mechanisms using small molecule inhibitors (SMIs) to reduce post-processing etching and lithography steps to improve the substrate film uniformity [72]. However, the aforementioned works required *in vitro* experiments, which are time-consuming, difficult to replicate in similar operating conditions, and challenging to quantitatively characterize with limited data. Thus, *in silico* modeling facilitates the procedures for gathering data more efficiently while also generating large data sets that align with the findings from experimental works. For instance, prior works [124, 126] have focused on multiscale modeling, an intricate simulation configuration that conjoins microscopic, mesoscopic, and macroscopic modeling [21, 75], of various reactor designs, stationary and spatial,



(a) (b)

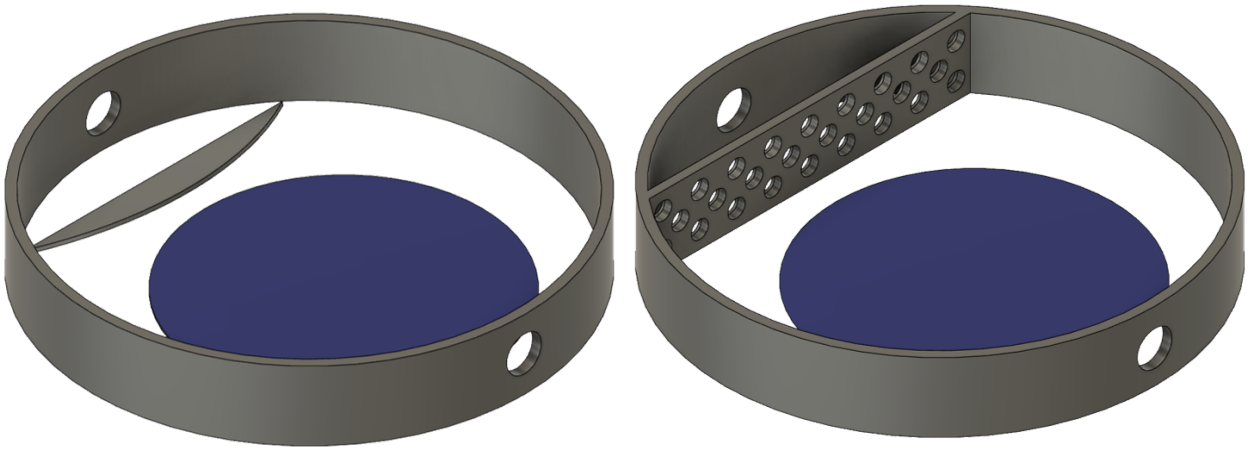
Figure 2.1: (a) Stationary and (b) spatial, sheet-to-sheet, reactor configurations for thin-layer deposition and etching processes.



(a) (b)

Figure 2.2: (a) Cross-flow orientation where feed is introduced parallel to the substrate surface, and (b) perpendicular flow orientation where feed is introduced above the substrate surface for thin-layer deposition and etching processes.

using different reagent delivery systems (showerhead, plate, cross-flow, and perpendicular flow) for ALE processes. Prior work [127] conducted multiscale computational fluid dynamics (CFD) modeling to study the spatiotemporal behavior of reagent distribution in a spatial-type rotary reactor for an AS-ALD process. Several works have also examined the optimization of reactor design for spatial reactors for ALD processes [22, 86, 87] and the optimization of reactor operating con-



(a) (b)

Figure 2.3: (a) Inclined plate and (b) showerhead distributors to control reagent flow uniformity for thin-layer deposition and etching processes.

ditions [28, 29] using CFD. Thus, *in silico* modeling presents an alternative approach for data collection and reactor performance evaluation, which enables the development of advanced technologies that are capable of producing highly conformal thin films by studying the spatiotemporal behaviors of species pressure, temperature, and velocity profiles to further optimize reactor and process design. Particularly, this work will examine the role of the reactor geometry on the fluid dynamics on the substrate surface that can be employed for industrial applications for semiconductor manufacturing.

Motivated by the above considerations, this work will employ computational fluid dynamics (CFD) modeling to study the behavior of the fluid dynamics within the discrete feed reactor model proposed by [66] for an AS-ALD process characterized by [74]. One caveat of the AS-ALD process is that steric hindrance plays an important role in the process. Steric hindrance is caused by bulky molecular species such as bis-diethylaminosilane (BDEAS), and it can introduce surface film deposition nonuniformities and cause incomplete surface coverage due to an excess of molecular interactions when the substrate is exposed to an abundance of reagents [64, 76], which is exemplified in Fig. 2.4. Prior work [122] studied these molecular interactions using a mesoscopic

modeling simulation that is employed with a kinetic Monte Carlo (kMC) algorithm to stochastically simulate the adsorption reactions and the orientation of adsorbates on the substrate surface. The previous works [109, 127] performed multiscale CFD modeling for an AS-ALD process with a silica/alumina substrate for the optimization of an advanced spatial, rotary reactor configuration. However, these works did not account for reactor optimization for mitigating the screening effects induced by steric repulsions, which lead to incomplete surface coverage (i.e., lower observed deposition rates) and ultimately nonuniform thin film surfaces. This work will perform computational optimization of an AS-ALD stationary reactor using a discrete feed method (DFM) to optimize the fluid dynamics of the reagents on the surface of the substrate. Several factors including the evacuation of gases and temporal progression of surface pressure will be discussed to determine saturation pressures that will be beneficial for discrete feed modeling in future multiscale modeling work for the same reactor geometry. For instance, [117] studied the fluid velocity distribution of ALD processes on alumina through CFD modeling and the effect of precursor overdosage on homogeneous flow. Likewise, [17] examined the role of precursor flow rate and residence time on the flow distribution. The ideal reactor has been optimized to have the following characteristics from a physical and computational perspective:

- Wafer surface saturation to initiate chemical adsorption [88].
- Minimal entrainment of gaseous species (i.e., small residence time of gases to reduce steric effects).
- Reactor design that is appropriately meshed to meet simulation standards [3] while yielding realistic simulation time demands.

Additionally, the present work examines the fluid dynamics, particularly the uniformity of the reagent distribution, which is crucial for achieving high film uniformity. Likewise, this work aims to determine the saturation time, which is when the wafer is fully exposed to the reagents, and

purging times, which is when all reagents are completely removed from the reactor, to determine appropriate operating conditions for future multiscale modeling research on the discrete feed mechanism for this reactor design, and to provide an applicable reactor configuration for potential integration into industrial applications. The utilization of the DFM and the optimization of reactor models to counter the effects of steric hindrance will enable the development of large-scale, advanced reactors for ALD and AS-ALD processes by examining the effect of the inlet geometry on the uniform distribution of reagent along the substrate surface, which is a vital condition to enable the production of highly conformal thin film surfaces.

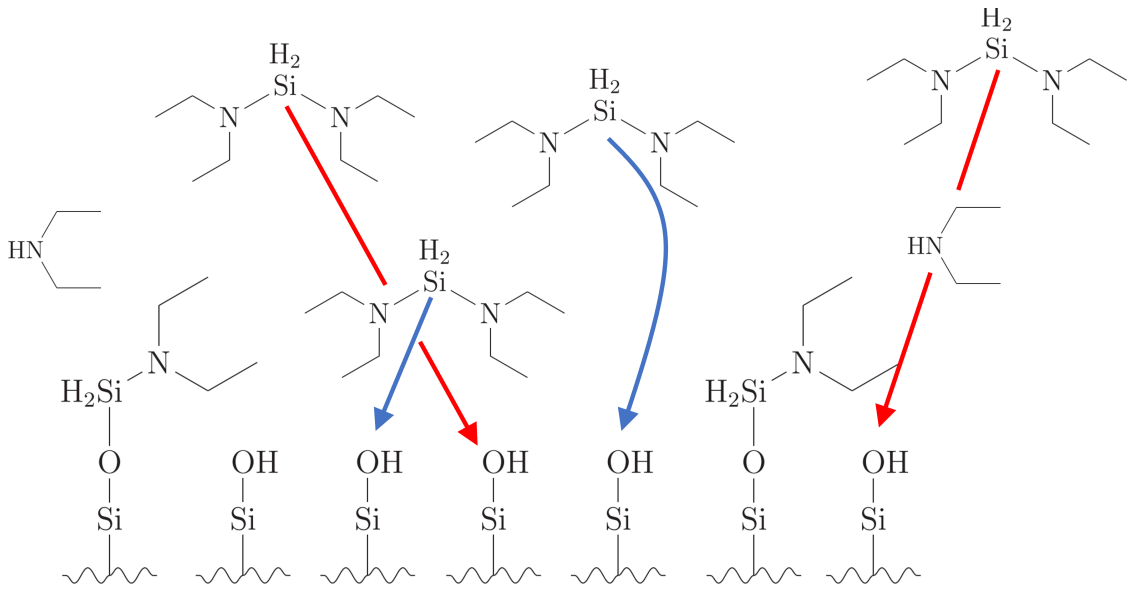


Figure 2.4: Illustration of the steric hindrance screening effect caused by excess reagent or overproduction of byproduct such as diethylamine (DEA), blocking precursors such as bis-diethylaminosilane (BDEAS) from adsorbing to the substrate surface composed of SiO_2. Red arrows indicate screening effects from DEA and excess BDEAS while blue arrows have a probable adsorption path.

This manuscript will be organized in several sections. Section 2.2 will examine the development of the CFD simulation through reactor design and optimization, meshing, and CFD modeling construction through Ansys fluids products, and Section 2.3 will evaluate the fluid dynamics of the reactor model, and discuss procedures intended to maintain film uniformity and maximize the de-

position rate per cycle.

2.2 Computational fluid dynamics modeling framework

This work proposes a computational fluid dynamics (CFD) model for a reactor configuration that is developed through a framework of various software tools (to be discussed below) to replicate the conditions of AS-ALD for a discrete feed method (DFM) approach. The reactor that this paper discusses is modeled loosely using a previously developed design [66] that is constructed through computer aided design (CAD) software. Consequently, the reactor geometry is discretized into a mesh that conforms to software quality criteria [3] while simultaneously reducing the computational strain on the simulation. CFD is later utilized to model the various reactor configurations that are characterized by different reagent delivery systems to study the spatiotemporal behavior of the gases, particularly the development of laminar flow, the evacuation of gases, the formation of vortices, and the distribution of gases. Such conditions will ensure that the reactor is constructed and optimized to ensure that the resulting flow behavior will achieve film uniformity and purge byproducts sufficiently to limit the effects of steric hindrance. This section discusses the procedural steps and assumptions conducted for the construction of the reactor through CAD software, Ansys DesignModeler, and meshing and CFD simulation through Ansys multiphysics software, Fluent.

2.2.1 The Impact of Steric Hindrance

Steric shielding is caused by the bulkiness of molecular species expanding beyond molecular distances between substrate atoms [93, 104]. This repulsion effect is also influenced by the formation of byproducts that hinder the adsorbates and small reactive site distances that facilitate the adsorption of bulky adsorbates. Thus, the self-limiting behavior of AS-ALD, such that monolayers of surface material are deposited with each cycle, is not observed due to the effects of steric hin-

drance [118], and limits the reaction pathway. Following the rate-limited adsorptions, the surface of the wafer will experience an oversaturation of reagent, which will be exhausted and wasteful for the process. To leverage the screening effects induced by the bulky adsorbates, small molecule inhibitors (SMIs) have been integrated into AS-ALD processes [119, 120], where the steric behavior of an SMI was simulated in prior work through kinetic Monte Carlo (kMC) simulation, which uses a stochastic procedure to replicate the conversion of active surface sites in the atomic scale and atomistic methods that employ *ab initio* quantum mechanics simulations to evaluate kinetic parameters [122]. Several works have proposed discrete reagent delivery methods that occur in short pulses with sequential purging pulses to reduce the generation of byproduct species, thereby minimizing the intermolecular collisions with adsorbates [59, 113]. For instance, [84] proposed a pulsed feed method using a numerical study by introducing a steric factor to replicate the rate-limiting behavior of adsorption reactions. [88] conducted a discrete feed method for atomic layer deposition of HfO thin films to mitigate the screening effects by suppressing reagent overdosage, and observed higher growth film rate and improved electrical properties of the film. Motivated by the prior works, this work aims to consider the role of steric hindrance in an AS-ALD process and to adopt reactor configurations that are appropriate for minimizing the role of screening effects for adsorption reactions.

2.2.2 Computational Fluid Dynamics Modeling Equations

The spatiotemporal behavior of the fluid transport is captured by numerically solving the mass, momentum, and energy transport equations, which are the fundamental equations that characterize the motion of the fluids in the reactor. The mass and momentum balance equations are defined by the following expressions, respectively:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \overrightarrow{v}) = S_m \tag{2.1}$$

$$\frac{\partial}{\partial t}\left(\rho\overrightarrow{v}\right)+\rho\left(\overrightarrow{v}\cdot\nabla\right)\overrightarrow{v}=-\nabla P+\nabla\cdot\left(\overline{\overline{\tau}}\right)+\rho\overrightarrow{g}+\overrightarrow{F} \tag{2.2}$$

where ρ is defined to be the density of the fluid mixture, $\overrightarrow{v}$ represents the velocity of the fluid mixture, S_m is the mass transfer source rate, P is the static pressure of the system, $\overline{\overline{\tau}}$ denotes the symmetric, second-order rank, stress tensor term, $\rho\overrightarrow{g}$ is the gravitational body force exerted by the fluid mixture, and $\overrightarrow{F}$ reflects the external body force on the fluid mixture.

Additionally, the energy conservation of the system is expressed by the following equation:

$$\frac{\partial}{\partial t}\left(\rho E\right)+\nabla\left(\overrightarrow{v}\left(\rho E+P\right)\right)=\ -\nabla\left(\Sigma h_j\overrightarrow{J}_j\right)+S_h \tag{2.3}$$

where E represents the internal energy of the system, h_j expresses the sensible enthalpy of fluid species j, S_h denotes the heat transfer source rate, and $\overrightarrow{J}_j$ describes the mass diffusion flux rate of fluid species. The transport equations will be simplified by making assumptions to the reactor design that are elucidated in Section 2.2.5.

2.2.3 Reactor Designs

With the laminar viscous model defined, various three-dimensional (3D) reactor configurations are constructed by modifying the geometry of the gas delivery system to the substrate (e.g., the showerhead and shape of the inlet). The inlet to the reactor is positioned above the showerhead divider to allow the reagent to flow perpendicular to the surface of the substrate, and the reactor outlets are oriented along the lateral sides of the reactor to allow the gases to evacuate in a cross-flow manner. The reactor geometry is constructed using Ansys DesignModeler using a reactor model that employs the discrete feed method (DFM) with a perpendicular flow orientation through a showerhead plate [66]. The reactor has a cylindrical body that is 300 *mm* in diameter and 8 *mm* in height, where the lateral region contains an inlet and outlet for the substrate and for the evacuation of gases, which is illustrated in Fig. 2.5. The gap distance between the inlet (4) to the showerhead

(5) is 3 *mm*, where the showerhead is composed of pores that are 10 *mm* in diameter, and the gap distance between the showerhead and the substrate is 5 *mm*. The outflows (3) have a diameter of 4 *mm*, and the wall is constructed with a sector angle of 40°. Additionally, the wafer (1) is modeled as a thin, 250-*mm* diameter surface which rests on a plate (2), which allows the wafer to enter and exit the reaction zone (6) by way of a rotating conveyor belt that is similar to the rotary reactor modeled in a prior work [127]. A summary of the reactor configuration dimensions are provided in Table 2.1.

Table 2.1: Dimensions for the reactor configurations.

Reactor Dimension	Value
Plate Diameter	290 *mm*
Ring Inlet Outer Diameter	170 *mm*
Ring Inlet Inner Diameter	130 *mm*
Round Inlet Diameter	20 *mm*
Round Outlet Diameter	4 *mm*
Showerhead Diameter	250 *mm*
Showerhead Pores Diameter	10 *mm*
Showerhead Thickness	0.5 *mm*
Showerhead-Wafer Gap Distance	5 *mm*
Inlet-Showerhead Gap Distance	3 *mm*
Wall Sector Angle	40°

The reactor configuration design will be analyzed by changing the geometry of the modeled inlet. Various inlet geometries include a single round inlet (Case 1), multiple round inlets (Case 2), a ring-shaped inlet (Case 3), and a combined ring-shaped and circular inlet (Case 4), which are all depicted in Fig. 2.6. The modification of the inlet geometries serves to provide a crucial understanding of their effect on the uniformity of the reagent distribution along the radial direction of the substrate. Such geometries also aim to prevent an excess of reagent delivery by placing

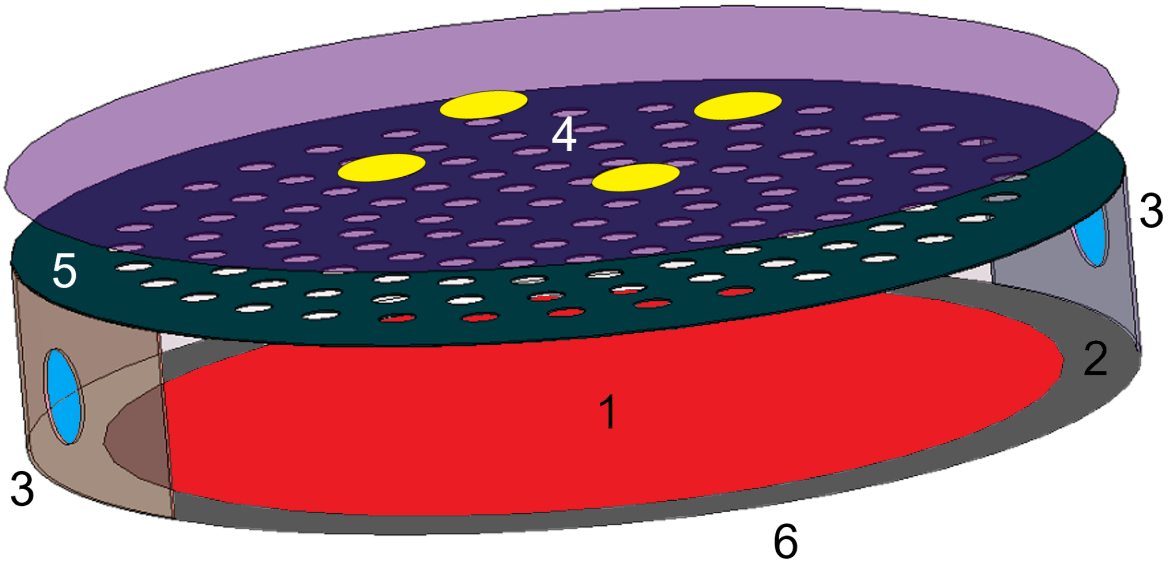
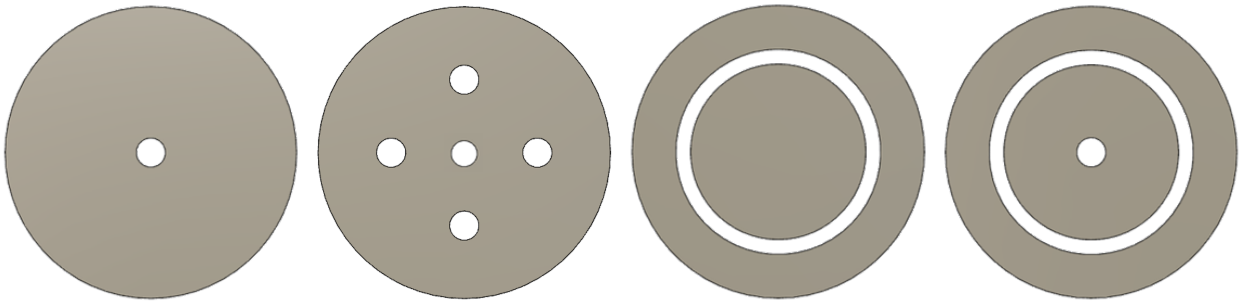


Figure 2.5: Schematic of the proposed reactor constructed using CAD modeling software, Ansys DesignModeler, containing a (1) substrate (in red), (2) substrate holder or bottom plate (in gray), (3) outflows (in blue), (4) inlets (in yellow), (5) showerhead distributor (in teal), and (6) wafer inlet and exit from the reactor chamber, which also serves as an outlet for gases.

the showerhead plate to limit the amount of reagent exposure, thereby reducing potential steric shielding caused by molecular collisions between byproducts and unreacted reagent.



(a) Case 1 (b) Case 2 (c) Case 3 (d) Case 4

Figure 2.6: Inlet geometries composed of the (a) single round inlet, (b) multi-round inlet, (c) ring inlet, and (d) combined round and ring inlet, which are examined for their performance in the reactor model.

2.2.4 Reactor Meshing

To generate spatiotemporal data for the reactor, a finite element approach is integrated into the work by discretizing the reactor geometry into cells through a meshing procedure. The Meshing Mode of Ansys Fluent contains several functions that optimize the efficiency and accuracy of the computations depending on the fineness or size of the mesh while defining the geometry of the cells and the resolution of the mesh. A nonuniform mesh resolution is fundamental to the meshing procedures, which densifies discretized elements in boundary layer zones and disperses cells with increasing distance from the boundary layers. Additionally, several features are available to improve the quality of the mesh to conform to recommended mesh quality criteria by the Ansys guidelines [3] to ensure that simulation accuracy and efficiency are appropriately balanced. Such mesh quality parameters include the minimum orthogonality, the skewness, and the aspect ratio as summarized in Table 2.2, which are important features that are intended to preserve the computational accuracy and efficiency of the simulation.

The minimum orthogonality is a parameter that measures the quality (accuracy and stability) of the mesh. This quality indicator is calculated by finding the minimum value of the normalized dot product of an area vector from a face $\overrightarrow{A}_i$ for a cell i and a vector from the centroid of the cell, i, to the face, $\overrightarrow{f}_i$, and the normalized dot product of area vectors of a face $\overrightarrow{A}_i$ for a cell i and a vector from the centroid of a cell i and the centroid of an adjacent cell that shares the same face $\overrightarrow{c}_i$ for all cells, N, in the mesh. An ideal mesh would have a minimum orthogonality of unity.

$$\text{Orthogonality} = \min\left(\frac{\overrightarrow{A}_i \cdot \overrightarrow{f}_i}{\left|\overrightarrow{A}_i\right|\left|\overrightarrow{f}_i\right|}, \frac{\overrightarrow{A}_i \cdot \overrightarrow{c}_i}{\left|\overrightarrow{A}_i\right|\left|\overrightarrow{c}_i\right|}\right) \quad \forall\, i \in \{1, 2, \ldots, N\} \tag{2.4}$$

The minimum orthogonality is selected for all cells and is expressed in Table 2.2.

The skewness is a measure of the rigidity of a cell, i, from its equilateral counterpart, which is defined as the ratio between the difference of the equilateral cell volume, $V_{eq,i}$ and cell volume $V_{c,i}$

to the equilateral cell volume, as follows:

$$\text{Skewness} = \frac{|V_{eq,i} - V_{c,i}|}{V_{eq,i}} \quad \forall i \in \{1, 2, \ldots, N\} \tag{2.5}$$

The averaged value of the skewness of all cells, N, in the mesh is expressed in Table 2.2, and a skewness close to 0 is desirable.

Lastly, the aspect ratio is a quality parameter that characterizes the stretching of a cell i by comparing the maximum and minimum values of the distance between the centroid of the cell to the centroid of a face in the cell, $d_{f,i}$ or the distance between the centroid of the cell to a node in the cell, $d_{n,i}$.

$$\text{Aspect Ratio} = \frac{\max\left(d_{f,i}, d_{n,i}\right)}{\min\left(d_{f,i}, d_{n,i}\right)} \quad \forall i \in \{1, 2, \ldots, N\} \tag{2.6}$$

The averaged aspect ratio for all cells is evaluated and expressed in Table 2.2.

Several user-defined parameters of the mesh are defined to reduce the number of cells while maintaining a robust quality, which includes the cell growth rate on the surface and inside the reactor volume (1.5) and the minimum and maximum cell lengths (0.36 *mm* and 7.8 *mm*, respectively). The meshing software also employs algorithms intended to improve the organization, structure, and quality of the mesh by removing obsolete cells and restructuring irregular cell geometry in essential boundary layer zones. Additionally, tetrahedral volume cells and triangular surface cells are utilized for the three-dimensional (3D) reactor mesh. The quality of the meshes for each of the reactor configurations are shown in Table 2.2 and indicate that all meshes are within the appropriate tolerances for mesh quality indicators; thus, simulations are conducted with high computational accuracy and efficiency.

Table 2.2: The mesh quality for various reactor configurations in comparison to mesh quality standards provided by Ansys Fluent.

Quality Indicator	Orthogonality	Skewness	Aspect Ratio	Number of Cells
Criteria Range	$0.001 \sim 1$*	0* ~ 0.95	1* ~ 8	N/A
Case 1	0.140	0.443	3.022	1,181,523
Case 2	0.105	0.451	3.040	1,149,495
Case 3	0.097	0.448	3.028	1,171,125
Case 4	0.103	0.449	3.029	1,178,849

*Desired value for ideal mesh quality.

2.2.5 Simulation Development and Parameters

By applying the mass and momentum conservation equations described by Eqs. (2.1) and (2.2), the spatiotemporal behavior of the delivery system of reagent to the substrate surface is studied and then optimized so that the inlet geometry and showerhead will achieve substantial exposure uniformity for small pulse times. Additionally, the design of the reactor outflow must minimize the residence time of the byproducts and excess reagents within the reaction zone, minimize the effect of steric hindrance on surface nonuniformities, and also maximize the amount of deposited material within the deposition cycle. Therefore, the ideal residence time for all species is low as a consequence of the combined pressure force induced by the outflow vacuum pressure and the tendency for gases to migrate toward regions of lower concentration. It is also notable that this work simplifies Eq. (2.1) by neglecting the mass species source term, S_m, due to this work focusing specifically on the fluid dynamics of the system to optimize the species delivery and purging systems for the reactor. Therefore, reactions and chemical kinetics are not considered for this aspect of the work.

The reagent delivery system of the reactor must overcome the mass suction forces generated by the vacuum pressure outflow parameters that cause the gases to migrate radially outward from the

center of the wafer, while simultaneously reducing the residence time of the gases in the reactor. To lessen the potential formation of fluid vortices and eddies, the reagents and carrier gases are delivered in laminar conditions, which also minimizes reagent usage, minimizes reactor size, and simplifies the computational complexity of the model [90]. Thus, the reagent delivery to the wafer surface is influenced by the gravitational force and mass diffusion due to the assumptions that the external body and viscous forces within $\overrightarrow{F}$ in Eq. (2.2) are negligible when the laminar model in Ansys Fluent is defined in the CFD simulation [5]. With the laminar model specified, the diffusion flux rate for gas species j, $\overrightarrow{J}_j$ presented in Eq. (2.3) is defined by the following fundamental expression that is referred to as Fick's Law:

$$\overrightarrow{J}_j = -\rho D_{j,m} \nabla y_j - D_{T,j} \frac{\nabla T}{T} \tag{2.7}$$

where $D_{j,m}$ denotes the mass diffusion coefficient for species, j, y_j represents the mole fraction for species j, and $D_{T,j}$ describes the thermal or Soret diffusion coefficient. The CFD simulation also operates under isothermal conditions, assuming that the reactor has a suitable temperature control system; therefore, in Eq. (2.7), the mass diffusion flux is dependent only on concentration gradients within the fluid mixture.

The fluid flow pattern may be classified by the Reynolds number, Re, which describes flow as being laminar, transient, or turbulent and is defined as follows:

$$Re = \frac{\rho \overrightarrow{v} D}{\mu}$$

By assuming that the dynamic viscosity (μ), density (ρ), and mass flow rate of the fluid for each inlet configuration (hence the velocity, $\overrightarrow{v}$ of the fluid exiting the inlet is constant) in Cases 1 through 4 are constant, the Reynolds number is dependent on the characteristic length, D. The characteristic lengths for the round inlet and ring-shaped inlet depend on the ratio of the surface

area, A_{inlet}, of the inlet and the wetted perimeter, P_{inlet} of the inlet that can be calculated as follows:

$$D_{ring} = \frac{4A_{ring}}{P_{ring}} = \frac{\pi d_o^2 - \pi d_i^2}{\pi d_o + \pi d_i} = d_o - d_i \tag{2.8}$$

$$D_{round} = \frac{4A_{round}}{P_{round}} = \frac{\pi d^2}{\pi d} = d \tag{2.9}$$

The characteristic length of concentric circles described by the ring inlet in Eq. (2.8) depends on the inner and outer diameters (d_o and d_i, respectively), while the characteristic length of a circle described by the round inlet in Eq. (2.9) depends on the diameter, d, of the inlet.

Several user specifications are designated within the simulation to replicate industrial processes, including the mass inflow and outflow rates, temperatures, and pressures, as well as simulation parameters intended to carry out the numerical computations through finite element discretization methods and numerical solver approaches. For instance, the mass inflow and outflow boundary conditions are defined to the reactor model. Mass inflow rates for an arbitrarily chosen gaseous species are defined to have constant flow rates of 2.50×10^{-5} *kg/s* and constant mole fractions of 0.5. For this work, oxygen gas, O_2 is defined to the CFD model with material (e.g., density, viscosity, and thermal conductivity) and thermophysical property data (e.g., standard enthalpy, entropy, and heat capacity) selected from the Ansys Chemkin database. Outflow boundary conditions are specified to ensure that there is no accumulation of gas within the reactor and to prevent backflow, such that the inflow gas flow rate would be equivalent to the outflow gas flow rate. The reactor operating temperature and pressure of the reactor are defined to be 523 *K* and 101.3 *kPa*, respectively. A summary of all parameters defined to the model are provided in Table 2.3. The simulation will be conducted using a pressure-based coupled solver method that optimizes the computation speed by simultaneously solving the transport equations at the expense of requiring more random access memory (RAM) [3]. Additionally, the central processing unit (CPU), which consists of compute cores, has a substantial role in the parallel-computing environment that en-

ables mesh partitioning for simultaneous computing of the CFD simulation. A fixed time step size of 0.001 s is selected for a first-order implicit numerical solver method to numerically solve the transient CFD process model, which satisfies the recommendations by the software for the global Courant number. Simulations are performed through a Linux computer cluster system comprising of two nodes with 36 and 48 computational cores and 384 GB and 512 GB of RAM, respectively, and averaging 6 to 8 hours of simulation time to run 4 seconds of process time.

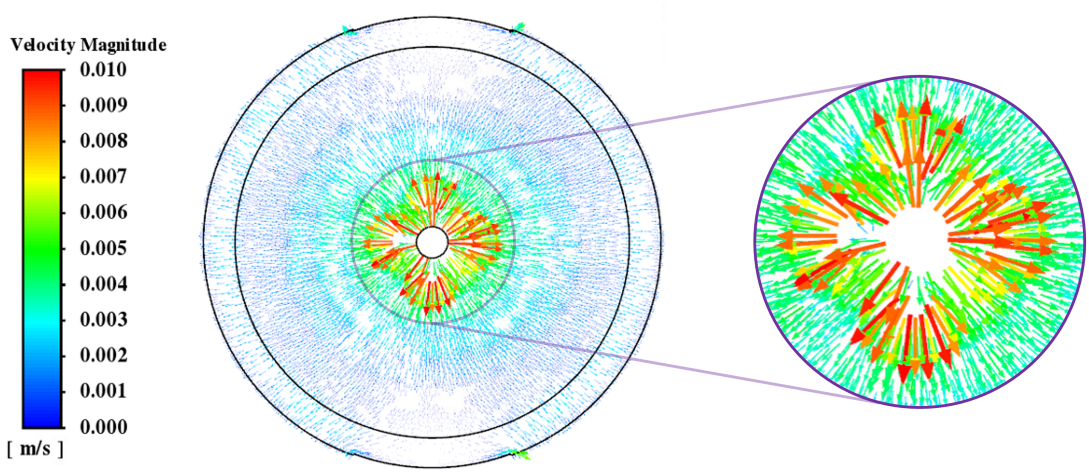
Table 2.3: Parameters specified to the CFD model and solver.

Parameter	Value
Operating Temperature	523 K
Operating Pressure	101.3 kPa
Gas Mass Flow Rate	2.50×10^{-5} kg/s
Gas Mole Fraction	0.50
Time Step Size	0.001 s
Maximum Iterations per Time Step	200

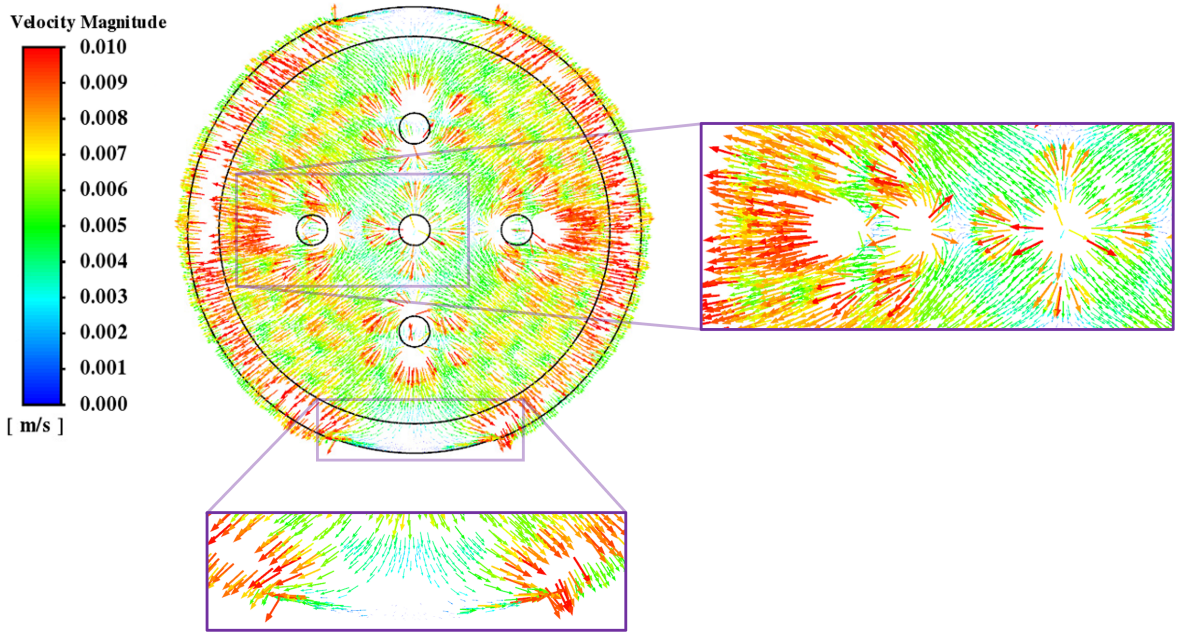
2.3 Simulation Results and Discussion

The computational fluid dynamics (CFD) simulation is performed to determine the role of the inlet configurations on the flow uniformity and understand what conditions will minimize the effects of steric hindrance that are caused by an oversaturation of reagents and byproducts in the vicinity of the substrate surface. To dilute this screening effect, the reagent delivery system limits the amount of exposure that the wafer has at any given time, thereby limiting the rate of reaction, particularly for the initial adsorption steps for the AS-ALD process. Thus, the flow profiles for each of the reactor models, particularly the velocity fields and pathlines illustrated in Figs. 2.7 and 2.8, respectively, and the mass transfer behavior presented in Fig. 2.9, provide a valuable depiction of

the movement of gases on the substrate surface.

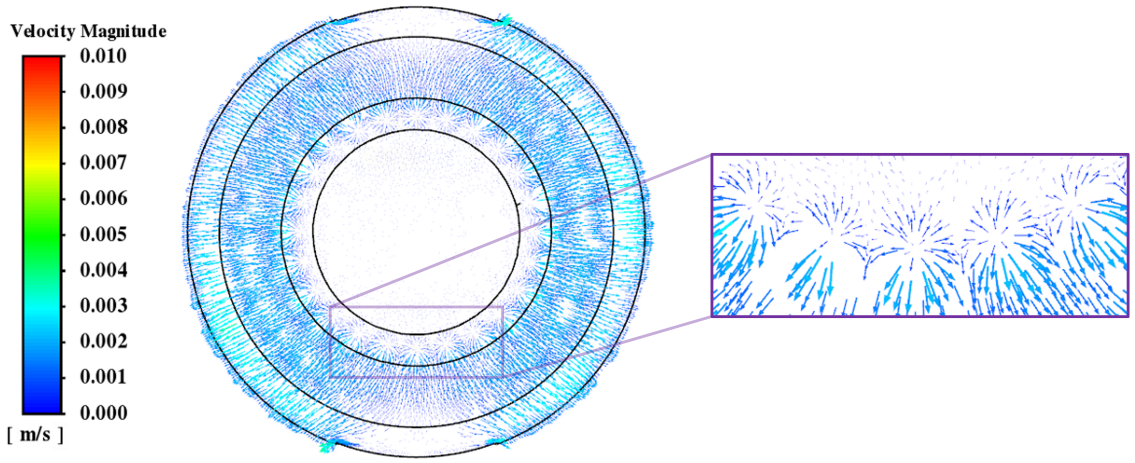


(a) Case 1

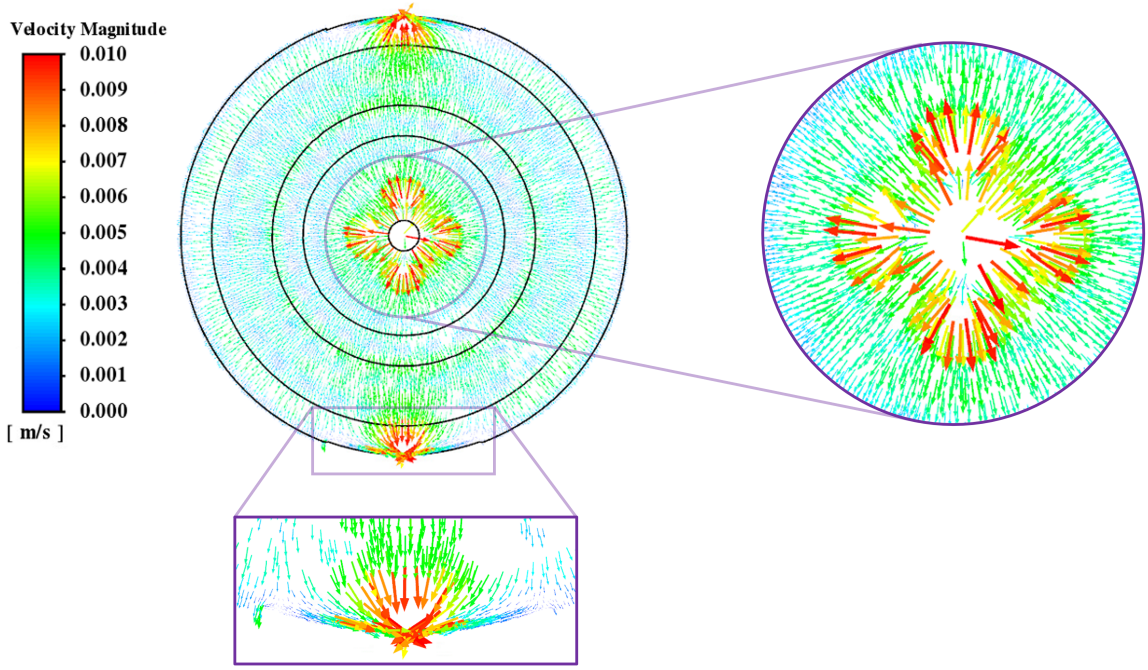


(b) Case 2

The velocity magnitude fields and pathlines provide an understanding of the dispersion of flow in the radial direction, as well as locations for fluid vortex and eddy formation. Fig. 2.7a, for example, presents a homogeneous distribution of flow for the singular and round-shaped inlet reactor



(c) Case 3



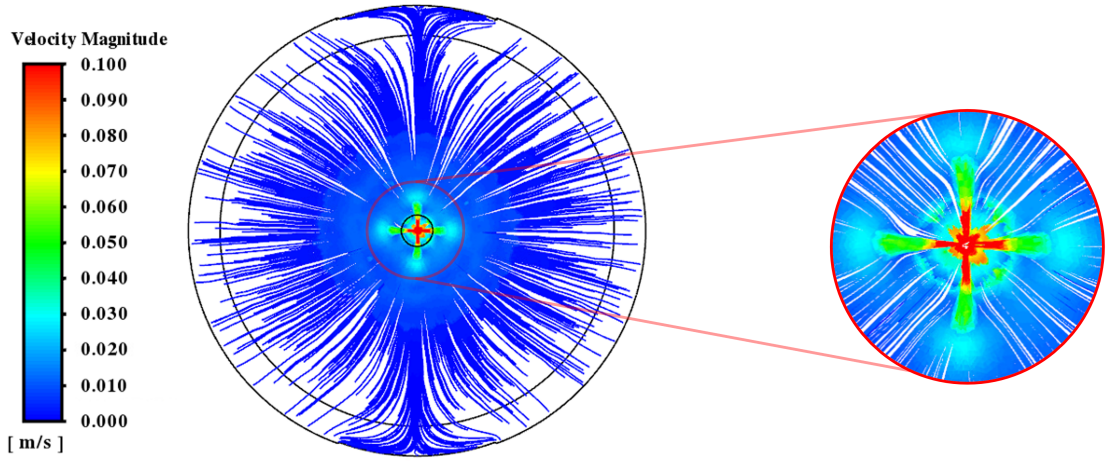
(d) Case 4

Figure 2.7: Velocity magnitude fields for the four reactor configurations at process times of 3 seconds. One-direction radial flow from the center to the outer regions of the reactor wall ensures minimal reagent intermixing that disrupts uniform flow behavior. This disturbed flow is exemplified by Case 2 while Cases 1 and 4 produce a more homogeneous flow.

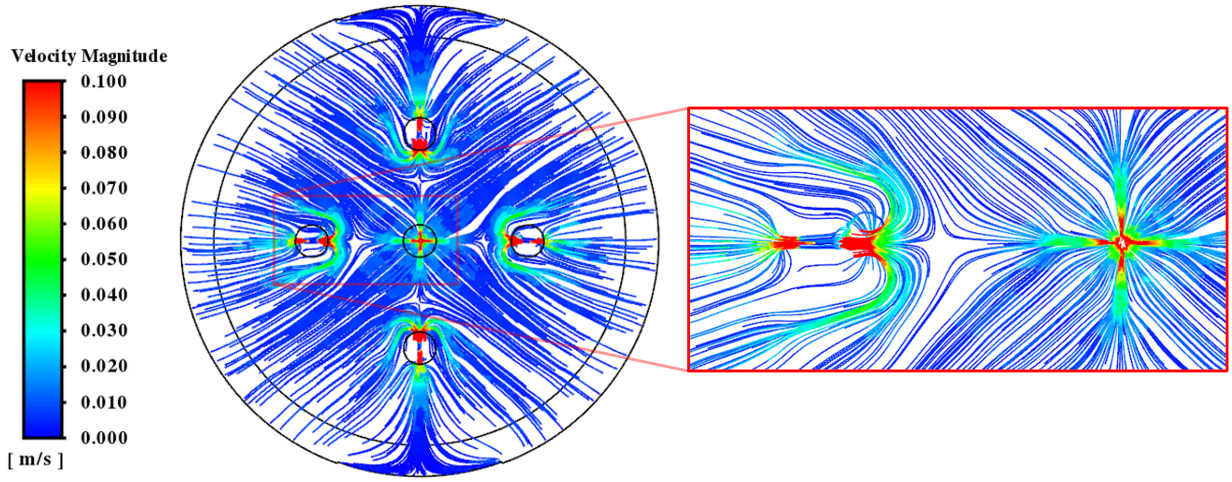
configuration, Case 1, without the formation of vortices or flow perturbations shown in Fig. 2.8a. However, the velocity field for Case 1 indicates that all fluid movement is concentrated on the

central region of the wafer, thereby extending the time to achieve complete saturation of the wafer surface, which is supported by the mole fraction contours for Case 1 in Fig. 2.9a. The oversaturation of reagent to the substrate in a localized region increases repulsions between adsorbates and unexhausted reagent due to the increase in desorption reactions, which dominate over adsorption reactions as the coverage approaches unity [41]. Such results demonstrate that despite the small distance between the showerhead and inlets, the showerhead performs poorly at distributing the reagent due to the gravitational and body forces overcoming the effects of the vacuum forces at the outlet. Consequently, the reagent is unable to diffuse to the outer regions of the substrate in the initial stages of the delivery, which impacts the homogeneity of deposition. Thus, additional geometries, Cases 2 through 4, were created to improve the radial distribution of reagent, thus reducing the potential for large concentration gradients, which contribute to poor surface uniformity and steric repulsions.

The inclusion of multiple, round-shaped inlets for reactor model Case 2 was designed to improve the dispersion of the gases, which is illustrated by the vector fields in Fig. 2.7b; however, there is an increasing likelihood for steric repulsions due to the opposing interactions between the inlet ports and flow field non-smoothness. The over-concentration of reagent resulted in disruptive flow behavior caused by increased convection from the combined gravitational forces and the perpendicular flow configuration. The inhomogeneous fluid flow is representative of the observations made by [117], who concluded that increased reagent pressure acts as a disturbance to the flow field, and leads to nonuniform deposition growth. Thus, the interactions worsened the radial distribution of reagent, which is pictured in Fig. 2.9b and generates highly concentrated regions that appear to increase the possibility of screening effects observed by the mole fraction contours of Case 1 in Fig. 2.9a. Thus, a ring-shaped inlet configuration in reactor Case 3 was developed to improve the uniformity of the flow in the radial direction. Although the flow field presented in Figs. 2.7c and 2.8c indicates greater flow migration in the radial direction, the vacuum pressure of the outlets prevented further gas diffusion into the central regions of the wafer, which is depicted



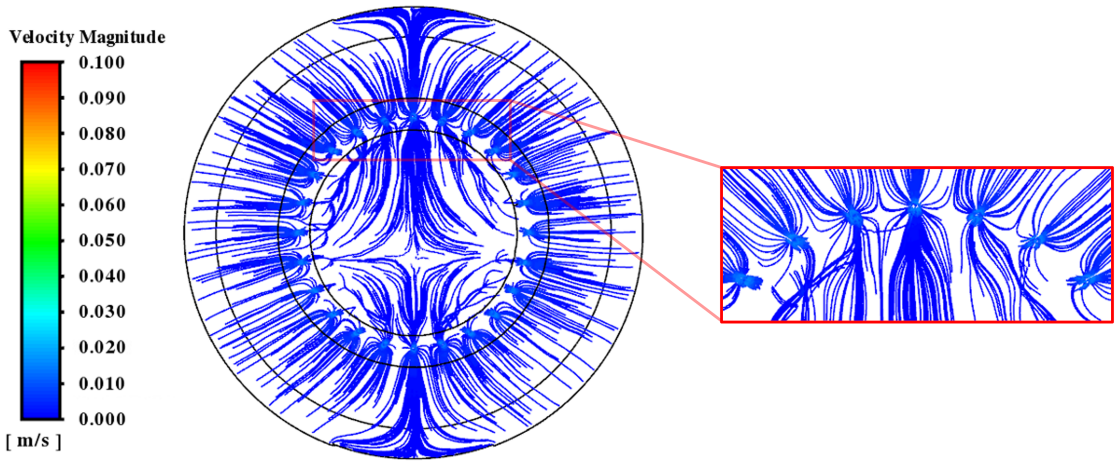
(a) Case 1



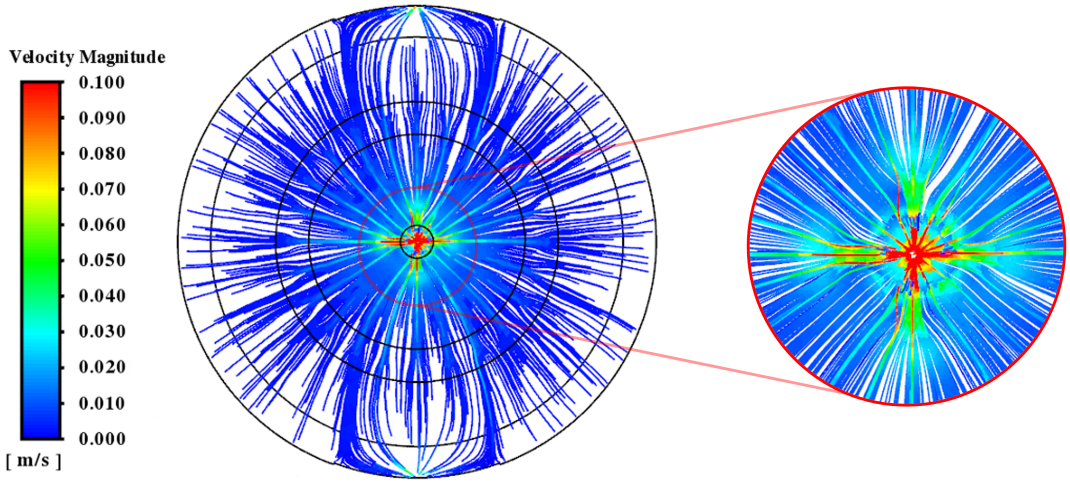
(b) Case 2

in Fig. 2.9c, and limited the potential to achieve complete surface coverage of the wafer. To facilitate the distribution of gases toward the center of the wafer, an additional round-shaped inlet and ring-shaped inlet was positioned in the center of the inlet plate, which resulted in a more uniform movement of reagent in the radial direction by minimizing the gas interactions between both inlet configuration, which is presented in Figs. 2.7d and 2.8d, and improved the distribution of reagent in the radial direction of the wafer in Fig. 2.9d.

Conversely, Reynolds number plots produced in Fig. 2.10 illustrate that the fluid flow in all reactor models are laminar. However, the development of non-smooth flow profiles is influenced by



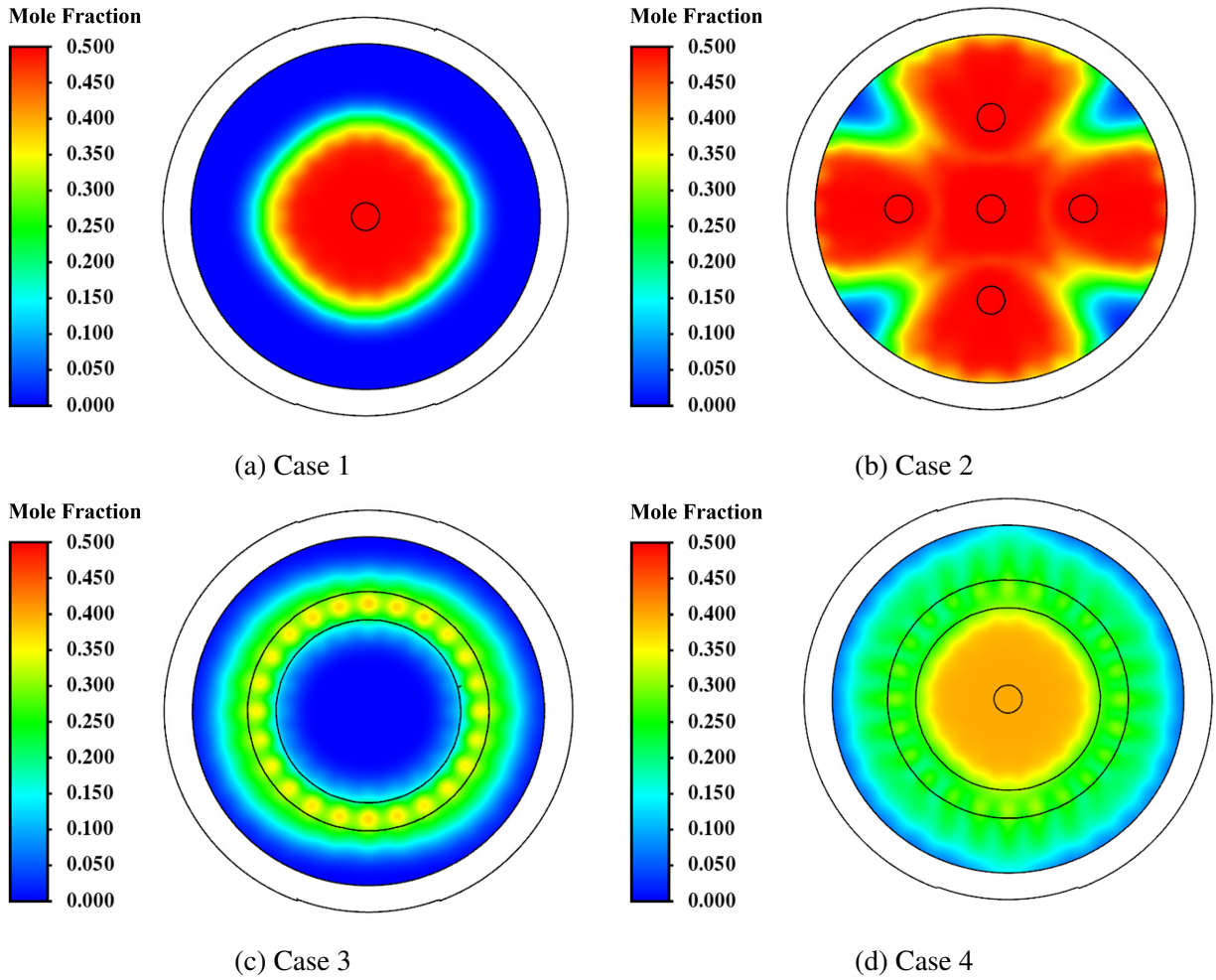
(c) Case 3



(d) Case 4

Figure 2.8: Velocity magnitude pathlines for the four reactor configurations at process times of 3 seconds. The pathlines demonstrate the directionality of flow to ensure effective purging and minimal flow disruption is observed is Cases 1 and 4, while pathlines for Cases 3 and 4 indicate that backflow is possible.

the inlet geometry, which increased the potential for screening effects. The localization of reagent in the central region of the wafer led to the nonuniform flow distribution for reactor configuration Case 1, which is illustrated in Fig. 2.10a. Likewise, Case 2 demonstrates that an overdosage of reagent through a multiple inlet configuration increases the potential for gas entrainment observed by [81] and eventual steric hindrance that restricts maximum deposition of substrate material in



(a) Case 1

(b) Case 2

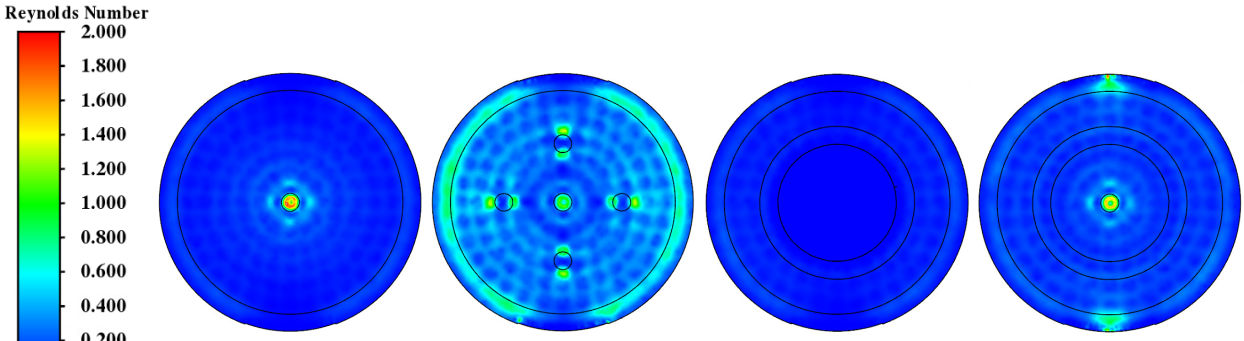
(c) Case 3

(d) Case 4

Figure 2.9: Mole fractions of the gaseous species on the wafer surface for the four reactor configurations at process times of 3 seconds, which illustrate the dispersion of gases in the radial direction as well as complete surface exposure to gases. The contours for Case 4 are indicative of the aforementioned characteristics for radial fluid flow and complete surface exposure.

monolayers shown in Fig. 2.10b. The observed increase in Reynolds number is caused by the increase in characteristic lengths produced by round-shaped inlet geometry in Eq. (2.9), which results in an increase in Reynolds number with added round inlets. As a result, a small number of inlets should be adopted into the inlet geometry to minimize flow non-smoothness. The appearance of turbulence in Case 4, which is illustrated in Fig. 2.10d, is caused by the mixing of the reagents from both inlets, which may produce steric effects due to the pressure difference induced by the vacuum outlets across the wafer surface. The eventual pressure gradient causes variations

in velocity in the radial direction of the wafer, which effectuated Reynolds number fluctuations. In contrast, the ring-shaped inlet in Case 3, visualized in Fig. 2.10c, mitigates turbulent flow due to the symmetry of the inlet and the effective removal of reagent by the vacuum pressure. The observation is supported by the characteristic length for ring-shaped inlet geometry in Eq. (2.8), which reduces the Reynolds number for larger surface area. However, the central region of the wafer is unexposed to reagent as a consequence of this strong pressure difference between the inlet and outlet pressures, which limits reagent migration to the center of the wafer. Hence, further optimization of operating conditions, particularly the inlet flow rates, are needed for future study into the optimization of the AS-ALD process as a whole. From the aforementioned results, the inlet geometry of the AS-ALD reactor has a profound effect on the distribution of the flow, and on the quality of substrate. Furthermore, advanced technologies are able to employ inlet geometries that are structured with characteristic lengths that are capable of minimizing the Reynolds number and controlling the fluid dynamics on the substrate surface. A summary of results is provided in Table 2.4 to illustrate the effectiveness of inlet geometry design on radial flow, gas distribution, purging, and coverage completion within 3 *s* of process time.



(a) Case 1 (b) Case 2 (c) Case 3 (d) Case 4

Figure 2.10: Reynolds number contours of the substrate surface and bottom plate for all reactor configurations at process times of 3 seconds to discuss flow perturbations. Case 3 indicates that laminar flow is uniform throughout the reactor, suggesting that ring-shaped inlet geometry leads to laminar flow, and is supported by the characteristic length computation in Eq. (2.8).

In addition to the distribution of flow, the removal of gases from the reactor have an important

Table 2.4: Comparison of reactor configurations based on criteria.

Reactor	Radial Flow	Uniform Distribution	Effective Purging	Complete Coverage
Case 1	✓	✓		
Case 2				
Case 3	✓	✓	✓	
Case 4	✓	✓	✓	✓

role in minimizing the steric collisions between molecules during the initial adsorption phase. For each reactor model, four outlet ports were generated to prevent backflow of gases into the reaction zone. The velocity pathlines presented in Fig. 2.8 illustrate that all four reactor models effectively purge materials from the reaction zone. Particularly, Case 4 was studied to determine the time to achieve reagent exposure on all surfaces of the wafer and the time to effectively purge all material through a cut-in purging step. Results presented in Fig. 2.11 reveal that 3.8 *s* of reagent feeding is needed to achieve a saturation of the wafer surface, while 7.0 *s* of purging by feeding pure inert species was needed to evacuate all gas species from the reaction chamber after reagent surface saturation was observed. Arguably, the reagent dosage times are lower than that from [17], who determined that longer dosage times are needed to achieve optimal growth rates. However, this work demonstrated the 7.0 *s* of purging time is analogous to the recommendations by [17] of 7.0 *s* to 10.0 *s*. From an economics perspective, the Case 4 reactor configuration demonstrated that minimal reagent loss is observable with the spatially homogeneous distribution of reagent and the reduction in concentration to limit surface adsorption kinetics for preventing steric hindrance generated by screening effects. Thus, the reactor model illustrates that discontinuous feeding of reagent provides sufficient surface exposure within processing times analogous to spatial reactor configurations studied by [124, 127]. A further use of the modeling framework developed in this work would be to generate reactor variable profiles for a variety of operating conditions that can be used to augment experimental data and then use the overall data set to implement data-

driven subspace identification for batch processes [13, 94] to model and improve thin film (product) quality at the end of the batch.

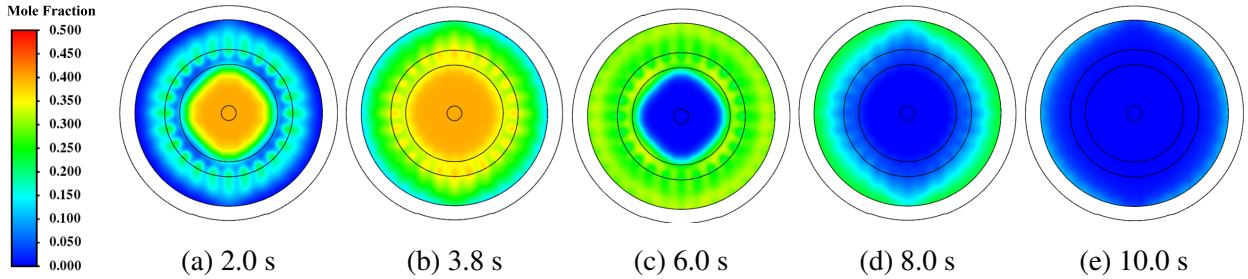


Figure 2.11: Gas mole fraction on the substrate surface at various processing times for the Case 4 reactor geometry. Complete coverage is observed within 3.8 *s* of process time and takes 6.2 *s* or process time to achieve complete purging of reagents on the surface.

2.4 Conclusions

With the rise of high-performance electronics, novel fabrication methods including area-selective atomic layer deposition (AS-ALD) are needed to improve the quality and production rate of semiconducting wafers that require stringent product specifications. The role of reactor configurations, particularly their inlet gas delivery systems, are of increasing interest to maintain wafer surface uniformity and coverage to improve nanopatterning and self-alignment for bottom-up fabrication of transistors. This work examined the impact of a new gas delivery system that employs a showerhead distributor, a perpendicular flow inlet, and cross-flow inlets on the distribution. Additionally, the removal of gaseous species from a reactor chamber to minimize steric hindrance induced by screening effects from excess reagent exposure and overproduction of byproducts in a temporal state was examined. Four inlet configurations were designed, discretized into finite elements, and later simulated through computational fluid dynamics (CFD) software to study the spatiotemporal behavior of the gases and examine the times required to achieve complete wafer exposure to reagents and minimize the residence time of gases through cut-in purging and output geometry

modification. Results indicated that the combined ring-shaped and round-shaped inlet plate as well as four outflow ports were sufficient to achieve all the aforementioned objectives. This geometry can then be integrated into a multiscale CFD model to evaluate the discrete feed method approach on the surface coverage of the wafer.

Chapter 3

Multiscale Computational Fluid Dynamics Modeling of an Area-Selective Atomic Layer Deposition Process Using a Discrete Feed Method

Multiscale models of AS-ALD processes provide a glimpse of various interactions occurring in differing time and length domains, and enables a unique approach to studying the spatiotemporal progression of surface coverage, which allows for the optimization of the processes. The development of multiscale models for AS-ALD are beneficial to the generation of large datasets, but they require a complex cross-platform programming network that couples various simulations into a single framework [75]. Multiscale models apply a combination of atomistic modeling through *ab initio* quantum mechanics computations to evaluate molecular and kinetic property data, mesoscopic modeling to characterize the stochastic surface kinetics through kinetic Monte Carlo methods, and macroscopic modeling to study the spatiotemporal behavior of fluids through computational fluid dynamics. This type of *in silico* modeling framework is beneficial towards studying the behavior

of the AS-ALD process through various time and length scales and towards optimizing reactor configurations with large datasets. This work will study the effects of the reactor geometry by examining multiple discrete feed reactor configurations with the goal of determining the optimal delivery system to produce a high-quality thin film with minimal processing time.

This work is organized as follows: Section 3.1 examines the atomistic modeling of structural, electronic, and thermophysical properties and the mesoscopic modeling of surface scale kinetics, Section 3.2 discusses the development of the macroscopic CFD model of an AS-ALD reactor through Ansys Fluent, Section 3.3 elucidates the multiscale modeling methodology used to conjoin the atomistic-mesoscopic and macroscopic simulations, and Section 3.4 analyzes the multiscale simulation results to determine the optimal reactor geometry that yields minimal process time for achieving full coverage and surface uniformity. A nomenclature of variables is also provided in Table 3.1.

3.1 Atomistic and Mesoscopic modeling

A vital component to understanding the AS-ALD process lies in the kinetics of the surface reactions. In this work, a kinetic Monte Carlo model (kMC) is used to characterize the stochastic nature of surface reactions and determine their dependency on pressure and temperature. This procedure is conducted by first using atomistic modeling techniques via *ab initio* quantum mechanics simulations to derive the reaction rates of possible surface reactions. Then, a kMC algorithm is developed that replicates surface kinetics through a user-defined grid that represents a larger swath of the wafer surface and determines probable reaction pathways for each site on the grid.

The mesoscopic model is one of two integral components in the overall multiscale simulation. Based on the partial pressures and temperature on the surface of the wafer, the extents of reaction in one integration timestep, 0.001 s, are simulated. From this information, the macroscopic model can calculate how much reagent is consumed and how much product is produced, which is accounted

Table 3.1: Table of variables with their respective definitions and units.

Variable	Definition	Units
A_{site}	Surface area of active site	m^2
δt	kMC time advancement	s
Δt	CFD timestep size	s
E	Internal energy of the system	J
E_{act}	Activation energy	$kJ \cdot mol^{-1}$
$\overrightarrow{F}$	External body force	$N \cdot m^{-2} \cdot s^{-1}$
γ_1, γ_2	Uniform random number	–
$\overrightarrow{g}$	Gravitational acceleration of Earth	$m \cdot s^{-2}$
h	Planck's constant	$J \cdot s$
h_j	Sensible enthalpy	$J \cdot kg^{-1}$
$\overrightarrow{J}$	Diffusion flux rate	$kg \cdot m^{-2} \cdot s^{-1}$
k_B	Boltzmann constant	$J \cdot K^{-1}$
$k_{ads,s}$	Adsorption reaction rate constant	s^{-1}
k_{nonad}	Nonadsorption reaction rate constant	s^{-1}
k_{tot}	Sum of L possible reaction rate constants	m^2
m_s	Atomic mass of gaseous reagent	$g \cdot mol^{-1}$
n	Number of active sites	–
P, P_s	Surface pressure for species s	Pa
ρ	Density of gas	$kg \cdot m^{-3}$
R	Ideal gas constant	$J \cdot mol^{-1} \cdot K^{-1}$
σ_s	Sticking coefficient for adsorbate s	–
S_h	Source energy flux generation rate	$J \cdot kg \cdot m^{-3} \cdot s^{-1}$
S_m	Source mass flux generation rate	$kg \cdot m^{-3} \cdot s^{-1}$
T	Absolute temperature of ambient environment	K
$\overline{\overline{\tau}}$	Stress tensor	$N \cdot m^{-2}$
$\overrightarrow{v}$	Fluid velocity	$m \cdot s^{-1}$
Z_s	Coordination number for gas s	–

for in subsequent timesteps.

3.1.1 Reaction rate calculations

The AS-ALD process examined in this work comprises three steps: (A) inhibition, (B) precursor adsorption, and (C) oxidation cycle. This work studies the AS-ALD of an Al_2O_3/SiO_2 substrate using three reagents: acetylacetone (Hacac) as a small molecule and gaseous inhibitor for Step A, bis(diethylamino)silane (BDEAS) as a gaseous precursor for Step B, and ozone (O_3) as a gaseous oxidant for Step C. To characterize all three steps of the AS-ALD process, this work simplifies the complex reaction mechanisms by concentrating on rate-limiting reaction steps determined through *in silico* modeling works [74, 76, 122] and considering each reaction step as an elementary reaction.

All the reactions involved in the AS-ALD process can be classified into two types: adsorption and nonadsorption. Adsorption reactions can be modeled as bimolecular reactions, and as a result, their reaction rate constants, $k_{ads,s}$, for an adsorbate s, can be calculated through Collision Theory of gases. The aforementioned pressure and temperature-dependent formulation is described as follows:

$$k_{ads,s} = \frac{P_s A_{site} \sigma_s}{Z_s \sqrt{2\pi m_s k_B T}} \tag{3.1}$$

where P_s is the partial pressure of the gaseous reagent s, A_{site} is the surface area of a single active site, σ_s is an experimentally determined sticking coefficient unique to the reagent s, Z_s is the coordination number of the gas s, m_s is the atomic mass of the gaseous reagent s, and k_B is the Boltzmann constant.

The reaction rate constants of the nonadsorption reactions, k_{nonad}, are calculated with the temperature-dependent Arrhenius equation, as defined by the following equation:

$$k_{nonad} = \nu \exp\left(-\frac{E_{act}}{RT}\right) \quad \text{where } \nu = \frac{k_B T}{h} \tag{3.2}$$

where h is the Planck constant, E_{act} is the activation energy of the reaction, R is the universal gas constant, and T is the absolute temperature of the reaction. The pre-exponential factor, ν, is calculated using Transition-State Theory (TST) by assuming that the ratio of the partition functions for the transition state and the reactants is unity [47]. This assumption was validated with experimentally determined process times for observing full surface coverage by [76]. The activation energy is found by first using *ab initio* quantum mechanics computations to optimize molecular and crystalline structures via Density Functional Theory (DFT). Then, the activation energies between the reactants and products are determined through Nudged Elastic Band (NEB) calculations where a spring-like force is introduced between the initial and final states of the reaction to generate a user-specified number of intermediate energy stages or "images" that correlate to form a maximum energy that is representative of the activation energy. The aforementioned computations were conducted through the open-source electronic-structure optimization software Quantum ESPRESSO (QE) [38] in a previous work by [122]. Through QE, various hyperparameters including the *k*-points, convergence thresholds, and kinetic energy cutoffs were specified until a molecular structure with minimal observable energy was obtained in their Plane-Wave Self-Consistent Field (PWscf) simulations. It is notable that all of these k values are the reaction rates for a single active site and have units of s^{-1}. In addition to kinetic property data, thermophysical property data of materials involved in the CFD simulation is calculated through density functional perturbation theory and phonon computations to evaluate the vibrational frequencies and utilizing the quasi harmonic approximation (QHA) [7].

3.1.2 Surface kinetics modeling

The kMC algorithm is a stochastic method that uses a set of randomly generated numbers to simulate the random nature of mesoscopic surface reactions in a spatiotemporal manner [14]. The algorithm used in this work is based on the BKL formulation created by Bortz, Kalos, and Lebowitz [8] and has been modified to provide additional insight into the process when the process

has larger reaction times relative to the timestep of the overall multiscale simulation, which is elucidated later on in this section. Prior to the execution of the BKL method in the kMC script, pressure and temperature data extracted from the computational fluid dynamics (CFD) simulation is utilized for the calculation of the reaction rate constants $k_{ads,s}$ and k_{nonad}. Next, the BKL method is conducted by assuming that potential reaction pathways for a given configuration are characterized by a Poisson distribution such that the reaction rate constants, k_i, from a list of possible reactions $i = 1, 2, \ldots, L$ are summed to enable the random selection of reaction pathways. This selection is made possible through the use of a uniformly calculated random number. Subsequently, a time advancement, i.e., the time in which the configuration transitions to the next state, is evaluated using a secondary uniform random number, which is summarized in Algorithms 1 and 2.

The BKL implementation in prior work [127] was conducted in the Python programming language using a grid-dependent procedure where the formulation, which is summarized in Algorithm 1, performs an iterative and ordered BKL procedure for each element in an $X \times Y$ grid. The result after each kMC execution produces grid data where the configuration or state of each element of the grid is discernible. However, this approach consumes memory that is not possible for implementation in user-defined functions (UDFs) in the computational fluid dynamics software, Ansys Fluent, and the integration of Python through cross-platform programming scripts reduces simulation efficiency.

To resolve the aforementioned challenges, a modification to Algorithm 1 is described in Algorithm 2 that contains the critical steps of the kMC methodology employed in this work, which is performed with C programming macros in UDFs that allow the program to communicate with the macroscopic simulation. Specifically, there are functions that use the macroscopic pressure and temperature data to calculate the reaction rates, which are used as inputs for the kMC algorithm. In an effort to preserve memory, a distinguishing feature of Algorithm 2 is the use of a random site selection method where site data is stored in variables that are independent of the location, i.e., the index, in the grid. A more detailed discussion of the memory constraints in UDFs is elucidated

later in this section. Additionally, after the algorithm is completed for a given timestep, the new grid data is converted into the change in coverage, which is then used to calculate the generation and consumption source terms for the reactants and products.

Algorithm 1: Original BKL kMC algorithm in Python with ordered site selection procedure.

Parameters: $P(x, y, t),\ T(x, y, t)$ ▷ *Determined from CFD*

Input: Grid data, $k_{ads,s}(P, T),\ k_{nonad}(T)$ ▷ *Calculated from Parameters*

Output: Grid data, δt

▷ *Let there be X grid rows and Y grid columns*
▷ *Let there be L reactions in the process, and let k_i represent the ith reaction*
$\delta t = 0$
while $\delta t < \Delta t$ **do** ▷ *Running algorithm until kMC timestep is as large as CFD timestep*
    for each species **do**
        if number of species in grid = 0 **then**
            Set the appropriate k_i value(s) to 0 ▷ *Removing impossible reactions*
    $k_{tot} = \sum_{i=1}^{L} k_i$
    for j in $1 : X$ **do**
        for k in $1 : Y$ **do**
            ▷ *Randomly determining if a reaction occurs for each site on the $X \times Y$ grid*
            Randomly select $\gamma_1 \in (0, 1]$
            for r in $1 : L$ **do**
                if $\sum_{i=1}^{r-1} k_i < \gamma_1 k_{tot} \leq \sum_{i=1}^{r} k_i$ **then**
                    if reaction r is possible **then**
                        Execute reaction r
                        ▷ *Steric hindrance for Step B is not shown here.*
    Randomly select $\gamma_2 \in (0, 1]$
    $\delta t = \delta t - \ln(\gamma_2)/k_{tot}$ ▷ *Advancing kMC timestep*

Algorithm 2: Modified kMC algorithm in UDF with random site selection procedure.

Parameters: $P(x, y, t),\ T(x, y, t)$ ▷ *Determined from CFD*

Input: Grid data, $k_{ads,s}(P, T),\ k_{nonad}(T)$ ▷ *Calculated from Parameters*

Output: Grid data, δt

1 $\delta t = 0$

2 **while** $\delta t < \Delta t$ **do** ▷ *Running algorithm until kMC timestep is as large as CFD timestep*

3 Randomly select a site on the grid

4 ▷ *Let there be L possible reactions for the selected site and let k_i represent the ith reaction*

5 $k_{tot} = \sum_{i=1}^{L} k_i$

6 Randomly select $\gamma_1, \gamma_2 \in (0, 1]$

7 **for** r in $1 : L$ **do** ▷ *Going through each possible reaction to randomly select one*

8 **if** $\sum_{i=1}^{r-1} k_i < \gamma_1 k_{tot} \leq \sum_{i=1}^{r} k_i$ **then**

9 Execute reaction r

10 ▷ *Steric hindrance for Step B is not shown here*

11 Determine n, the number of active sites from grid data

12 $\delta t = \delta t - \ln(\gamma_2)/(n k_{tot})$ ▷ *Advancing kMC timestep*

There are two major concerns that the modified kMC algorithm must address. The first appears when one of the reactions has a k value magnitude that is smaller than that of any other reaction; i.e., the rate-limiting step is substantially slower compared to the remaining reactions. A consequence of this order-of-magnitude difference is that reactions that are more *rate-determining* will contribute more to the overall processing time. For instance, in Line 19 of Algorithm 1, which is from the algorithm of the BKL kMC method, the size of the time advancement is directly proportional to $1/k$. Thus, when $1/k$ is larger relative to Δt, the large time advancements will generate incomplete data in the form of a step-wise appearance due to how the algorithm will advance major

portions of the grid concurrently whenever the rate-limiting step is the only available step.

The second area of concern is the memory usage. The BKL method employs a spatiotemporal approach to study the evolution and conversion of active sites that characterize the surface morphology of the substrate. For example, [31] simulated epitaxial growth using a kMC grid to study the surface morphology after each epitaxial cycle but observed computational constraints that limited the grid sizing to an $X \times Y$ grid. In a prior work [109, 127], the kMC algorithm was implemented through an external Python program and conjoined with the macroscopic computational fluid dynamics simulation in Ansys Fluent through a Linux Bash script.

In this work, the kMC algorithm was directly implemented in Fluent through custom User-defined Functions (UDFs) written in the C programming language [5]. UDFs are constructed by integrating Ansys-specific macros that allow extraction and manipulation of nodal data by defining integers attributed to boundaries in the simulation file [4], which is depicted in Fig. 3.1. UDFs also enable the declaration of user-specified variables, referred to by User-defined Memory (UDM) variables, with some restrictions that are detailed as follows. Specifically, following the aforementioned procedures, various macros are employed to enable parallelized computation, which is related to the partitioning procedure employed by the CFD software, iterative computation that is conjoined to the CFD solver, initialization, and executable actions, where C-based programming language is invoked for customized programs for the kMC code. Following the creation of the UDF-script written as a C program, the entire program is compiled through Ansys Fluent to generate the executable C programs, which work in unison with the CFD simulation.

While this approach provides a major increase in computational performance, this method also comes with more restrictions for the new code. The main restriction when implementing UDFs in Fluent is the memory storage, where a maximum of 500 variables can be safely stored for each integration timestep [5]. However, each kMC grid is defined by a 300×300 lattice in this work for a total of 90,000 sites. Thus, the data of each kMC grid must be stored differently so that it can be represented by less than 500 variables.

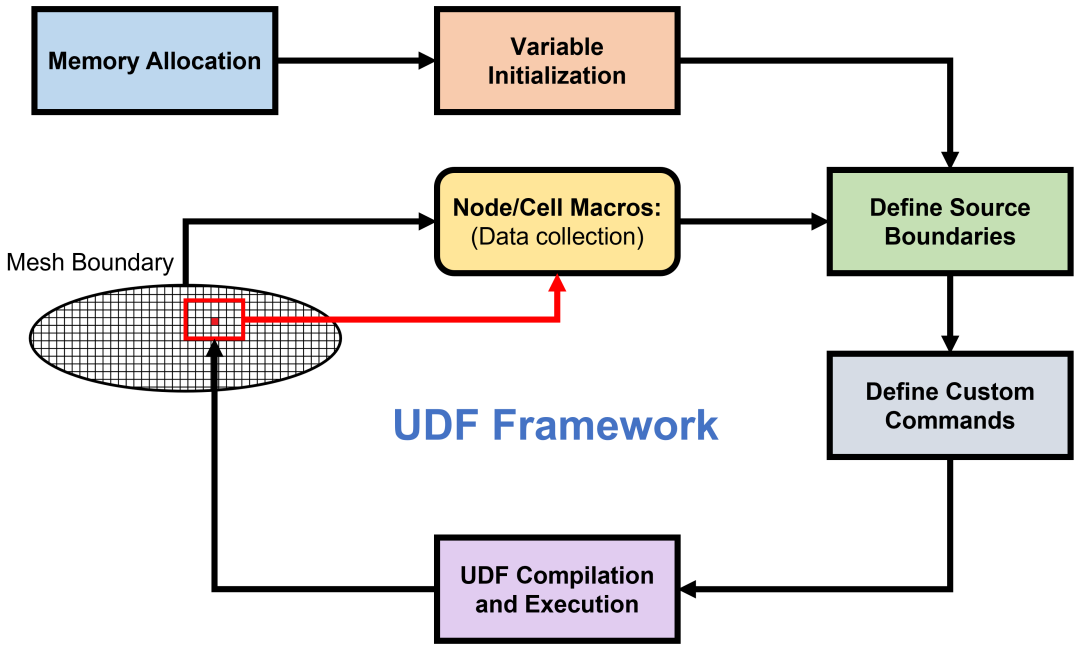


Figure 3.1: A process flow diagram of user-defined function integration in CFD simulations that communicates the nodal data in the mesh.

To rectify the two issues presented, the kMC algorithm used in this work was modified to evaluate the time required for a single active site to react, whereas the BKL algorithm evaluates how the entire grid progresses in a given timeframe [8]. This difference is implemented in the form of three distinctive changes, which are summarized as follows:

(1) The arrangement of occupied sites is neglected by employing a Markov chain in which only one site on the entire grid advances with each step as conducted by [53].

(2) The time advancement computation includes the number of unoccupied sites on the grid, where a reduction in unoccupied sites increases the time progression as employed by [40] and [56].

(3) Grid data is stored as a single integer variable that counts the number of each species, rather than as an array.

The first two modifications resolve the first concern regarding small k values, and the last mod-

ification resolves the second concern regarding memory storage. These adjustments to the BKL formulation are further examined in Section 3.1.2.1 and verified in Section 3.1.2.2.

3.1.2.1 Derivation of the modified kMC algorithm

To properly make the necessary modifications to the BKL kMC algorithm, it is first important to understand what the kinetic rate represents. Intuitively, the parameter reflects the number of reactions that occurs every second at an active site, which is a region on the substrate surface that is able to undergo a chemical reaction. In other words, the kinetic rate is a measure of the probability that such an event is happening. Additionally, these events can be assumed to follow a Poisson distribution, to react independently of other sites, and to react independently of other possible reactions [20]. As such, probability theory allows the reaction rate to be decomposed into two independent events [39], as shown in Eq. (3.3). First, there is the probability that the site is in a state where the reaction can proceed, $\mathcal{P}(possible)$. Second, there is the probability that the reaction actually proceeds, $\mathcal{P}(proceed)$. This decomposition is expressed as follows:

$$k_{rxn} = \mathcal{P}(reaction)$$

$$k_{rxn} = \mathcal{P}(possible) \cdot \mathcal{P}(proceed) \tag{3.3}$$

where k_{rxn} is the reaction rate of a given reaction for a single active site as calculated in Section 3.1.1 and $\mathcal{P}(reaction)$ is the probability of that reaction taking place in one second.

A similar expression for the reaction rate of the entire grid can be derived by using the number of sites rather than the probability of a site being able to undergo the desired reaction. This expansion is also based on the assumption that these two events are mutually exclusive, yielding:

$$k_{grid} = \mathcal{N}(possible) \cdot \mathcal{P}(proceed) \tag{3.4}$$

where k_{grid} is the average reaction rate for a given kMC grid and $\mathcal{N}(possible)$ is the number of sites that can undergo the desired reaction.

Because of the assumption that each site on the grid is independent of the others, $\mathcal{P}(possible)$ can be related to $\mathcal{N}(possible)$ as follows:

$$\mathcal{N}(possible) = n \cdot \mathcal{P}(possible) \tag{3.5}$$

where n is the number of active sites on the grid. With this relationship, k_{grid} can be related to k_{rxn} as follows:

$$k_{grid} = \mathcal{N}(possible) \cdot \mathcal{P}(proceed)$$

$$k_{grid} = n \cdot \mathcal{P}(possible) \cdot \mathcal{P}(proceed)$$

$$k_{grid} = n \cdot k_{rxn} \tag{3.6}$$

where k_{grid} is the average reaction rate of a given reaction for the entire grid, k_{rxn} is the reaction rate for a single active site, and n is the number of actives sites on the grid.

When there are multiple possible reactions, it is necessary to determine k_{tot}, which is the probability of an unspecified reaction occurring each second. Because it is assumed that the probability of each k is independent of other reactions, the probability that one of two reactions will take place can be found using the inclusion-exclusion principle as follows:

$$\mathcal{P}(k_i \cup k_j) = \mathcal{P}(k_i) + \mathcal{P}(k_j) - \mathcal{P}(k_i \cap k_j)$$

where $\mathcal{P}(k_i \cup k_j)$ is the probability that either reaction i or reaction j will take place, $\mathcal{P}(k_i)$ is the probability that reaction i will occur, $\mathcal{P}(k_j)$ is the probability that reaction j will occur, and $\mathcal{P}(k_i \cap k_j)$ is the probability that both reaction i and reaction j will occur. This equation can be simplified by noting that $\mathcal{P}_i = k_i$ as shown in Eq. (3.3) and that $\mathcal{P}(k_i \cap k_j) = 0$ because the two

reactions are mutually exclusive, which yields:

$$\mathcal{P}(k_i \cup k_j) = k_i + k_j$$

If there are L total possible reactions, these additional reactions can be summed into $\mathcal{P}(k_i \cup k_j)$ to obtain the probability that an unspecified reaction occurs as all of these reactions are independent and mutually exclusive. The expression for this possibility is:

$$\mathcal{P}(k_{tot}) = k_{tot} = \sum_{z=1}^{L} k_z \tag{3.7}$$

where k_{tot} is the possibility of an unspecified reaction occurring and k_z represents the possibility for a specific reaction to occur. Note that, when evaluating an active site in isolation, $k_z = k_{rxn}$; similarly, when evaluating an active site in the context of the entire grid, $k_z = k_{grid}$. Thus, to evaluate the time progression for a single active site at a time, $k_{tot,grid}$ can be represented as follows:

$$\begin{aligned} k_{tot,grid} &= \sum_{z=1}^{L} k_{grid} \\ &= \sum_{z=1}^{L} n \cdot k_{rxn} \\ &= n \cdot \sum_{z=1}^{L} k_{rxn} \\ &= n \cdot k_{tot,rxn} \end{aligned}$$

where $k_{tot,grid}$ is the possibility that an unspecified reaction will occur anywhere on the kMC grid, $k_{tot,rxn}$ is the possibility that an unspecified reaction will occur at a singular active site, and n is the number of active sites on the kMC grid. This formula is employed in Line 12 of Algorithm 2.

The other modification made to the kMC method presented in [122] is the reduction of memory usage as necessitated by the restrictions of Ansys Fluent. The total data stored in between each

timestep must be reduced from 90,000 integers for a 300×300 grid to less than 500 integers. This was done by taking advantage of the fact that the kMC algorithm does not use any positional data; i.e., the simulation is not concerned about the location of the site in the grid, but rather only the state of the grid for each timestep. Thus, instead of using an array with 90,000 entries, the amount of each intermediate species was counted and saved in its own variable. For example, in Step C of the AS-ALD cycle, there are 3 species: V4, V6, and V8. The old method would have 90,000 entries, each one representing a site and tracking whether it is in state V4, V6, or V8. The new method has 3 variables, which are defined as *buckets* that represent the number of V4 sites, the number of V6 sites, and the number of V8 sites. By discarding the unnecessary positional data of the sites, this bucket method is able to store the relevant information of all the kMC grids in 31 variables where 25 variables are reserved for tallying species involved in the Steps A, B, and C of the AS-ALD process and the remaining 6 variables are used for defining the source generation and consumption flux rate terms on the wafer surface boundary conditions.

However, the reaction mechanism for Step B is more complicated than the other two reactions, as steric hindrance plays an essential role in the kinetics of that process. The specific details of the effects of steric hindrance on the BKL formulation can be found in [122], but a summary is as follows. Each site has two adjacent neighbors that apply two conditions on the surface reaction mechanism. The first extra condition is that a site is restricted from certain reactions if a bulky molecule has adsorbed to either neighboring site. Physically, these molecules hinder the primary site from reacting. The second condition is attributed to the final surface reaction, which requires 2 adjacent sites to bond and for both sites to reach the final state. Thus, it is possible for situations to arise where it is impossible for a site to fully react if both of its neighbors have reached completion by bonding with other sites. As a result, these sites must be deactivated so that the kMC algorithm can reach completion.

To implement the first condition described above, the modified kMC algorithm creates buckets that represent the number of adjacent sites that are blocking it. Because each site has two sterically

relevant neighboring sites, there are 3 block status buckets: unblocked, one-block, and two-block. While both the one-block and two-block status represent the target site being unable to undergo certain surface reactions, distinguishing between the two allows the preservation of more positional data about the grid. During the kMC algorithm, whenever the algorithm needs to determine whether the site it randomly selected is blocked, it will do so by randomly selecting a block status.

The second condition is an extra step that takes place after each kMC event. To properly deactivate sites, the following procedure is taken after each iteration of Algorithm 2. By running Algorithm 3, the number of trapped sites that are unable to reach the final V4 state can be accurately represented even after discarding all positional data. After implementing both modifications discussed in this section, the kMC model is able to simulate the surface reactions with

Algorithm 3: Step B steric hindrance locking algorithm.

Variables:

V4: Completely reacted site

IC: Site that will never reach completion

LK: 2 adjacent V4s that cannot trap an unfinished site

▷ *Let there be N total sites*

▷ *Let S_i represent the species of site i*

for j in $1 : N$ **do** ▷ *Going through each site on the grid*

  Randomly select a species S_j

  ▷ *Let S_{adj1}, S_{adj2} be two randomly selected sites that represent the sites adjacent to S_j*

  if $S_j \neq$ V4 AND S_{adj1} = V4 AND S_{adj2} = V4 **then**

    $S_j \longrightarrow$ IC

  else if S_j = V4 AND (S_{adj1} = V4 OR S_{adj2} = V4) **then**

    $S_j \longrightarrow$ LK

    $V4 \longrightarrow$ LK ▷ *This represents the adjacent V4 turning into LK*

greater resolution and obtain high quality results for all the reactions in the AS-ALD process.

3.1.2.2 Verification of the modified kMC algorithm

To verify that the results of the modified kMC algorithm are accurate and valid, comparisons between Algorithms 1 and 2 were examined for both cases with slow reactions (Step B) and cases without slow reactions (Steps A and C).

3.2 Computational fluid dynamics modeling

Computational fluid dynamics (CFD) simulations describe the macroscopic behavior of fluids in larger time and length scales, which enables the scale-up of processes. The integration of CFD is applicable to characterizing the spatiotemporal flow of reagents on the substrate surface, which experiences surface reactions that consume the reagents and generate byproducts. The development of a CFD model requires the construction of a computer-aided design (CAD) model for a three-dimensional (3D) discrete feed reactor system, the performing of a meshing procedure on the CAD model, and the creation of the CFD simulation of the AS-ALD process in the discrete feed reactor model.

3.2.1 Reactor Design

An abundance of reactor models have been investigated to discuss the uniformity of reagent coverage on the substrate surfaces and to improve the productivity of the process. For example, prior work [126] proposed a cross-flow reactor to control the behavior of flow in the azimuthal direction of the substrate for an atomic layer etching process. Additionally, works [32] and [33] suggested using showerhead reactors to improve the uniformity of fluid flow in the radial direction. By considering the challenges attributed to low product throughput, prior work [127] proposed spatial reactor configurations where the reagent is delivered perpendicular to the substrate in a

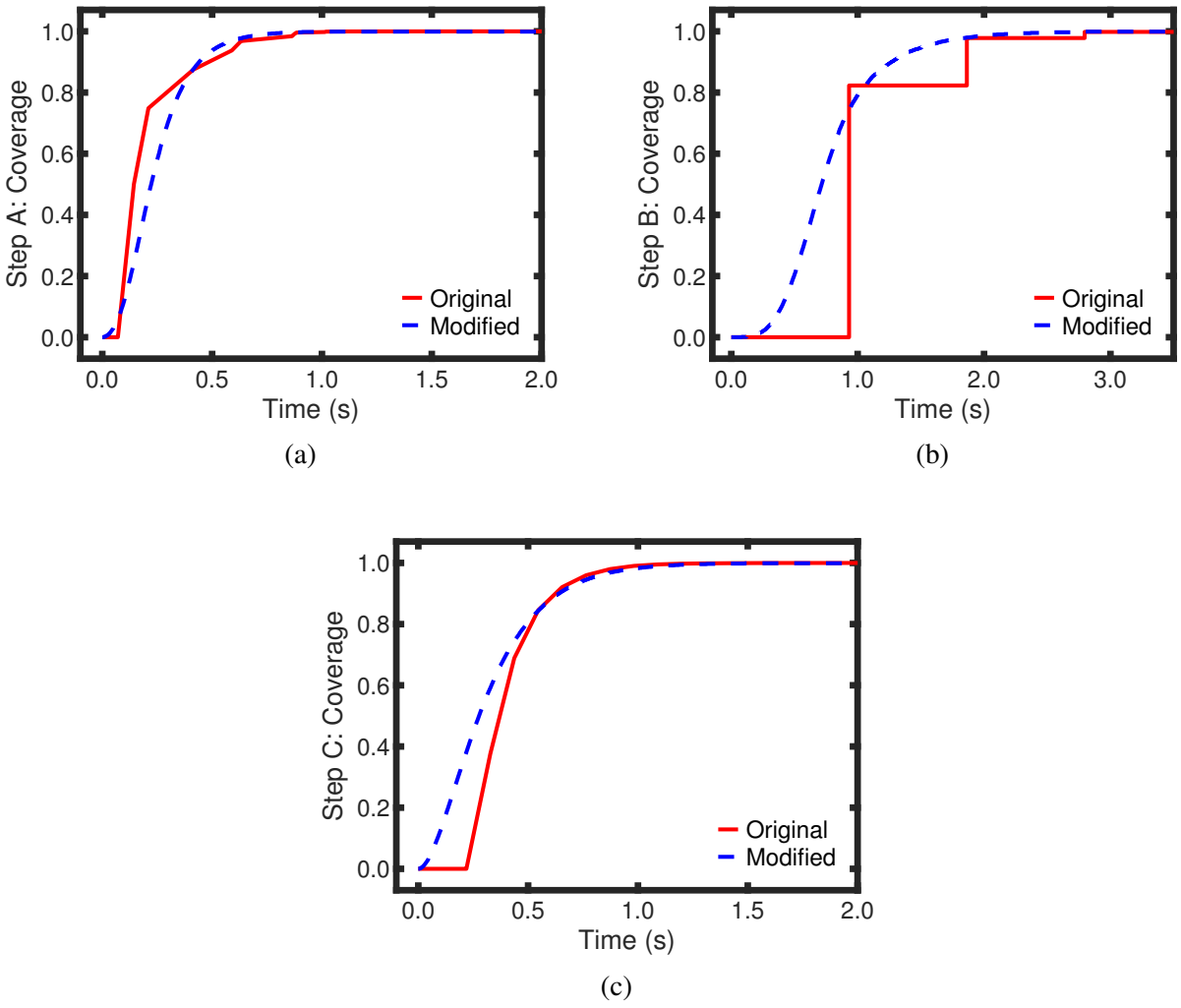


Figure 3.2: Comparison of the original (Algorithm 1) and modified (Algorithm 2) kMC algorithms for Steps (a) A, (b) B, and (c) C. The original kMC processing times to reach full coverage for Steps A, B, and C are 1.018, 2.796, and 1.418 s, respectively. The modified kMC processing times to reach full coverage for Steps A, B, and C are 0.969, 2.793, and 1.487 s, respectively.

continuous feeding mechanism for atomic layer etching and area-selective atomic layer deposition processes, respectively. While the aforementioned reactor models have effectively yielded valuable results in improving product quality and yield, this work considers the impact of steric collisions generated from bulky molecular species including Hacac and BDEAS, which introduces challenges associated with surface uniformity. Thus, there is motivation to develop a reactor that minimizes

steric hindrance induced by screening effects.

This work adopts a previously designed discrete feed reactor [108] inspired by work [66] through Ansys DesignModeler, which delivers reagent perpendicularly to the substrate surface in discrete pulses through an injection plate. The employment of discrete feeding with cut-in purging allows the byproduct species that inhibit adsorption of Hacac and BDEAS on the substrate surface to be regularly removed. The discrete feed reactor, illustrated in Fig. 3.3, situates a showerhead divider that is below and parallel to the injection plate to facilitate the transport of reagents in the radial directions of the substrate, thereby maximizing the exposure of the substrate to the reagent in minimal pulse times. The gap distance between the injection to the showerhead plate is 3 mm, and the gap distance between the showerhead plate to the 200-mm diameter substrate surface is 5 mm. These gap distances are necessary to minimize the volume required to maintain laminar flow behavior [54]. A summary of the reactor dimensions are summarized in Table 3.2. The injection

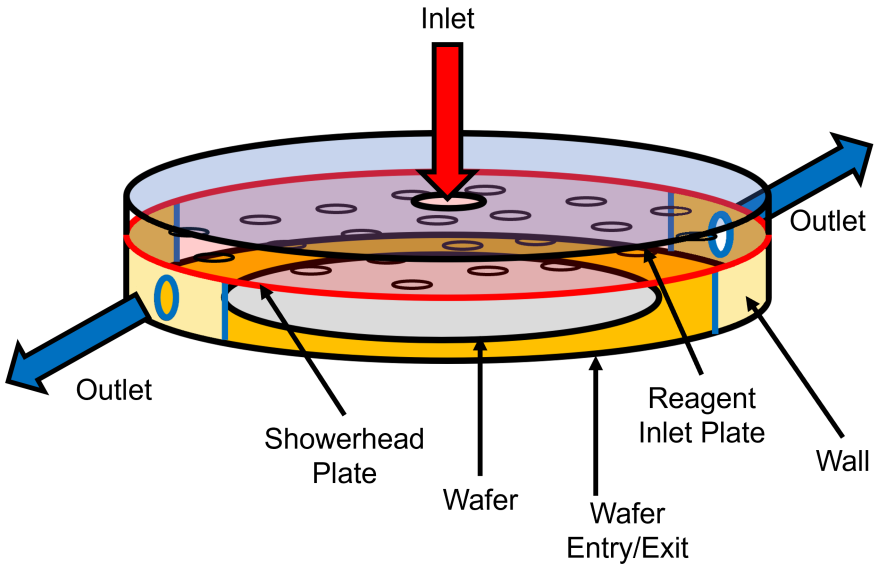


Figure 3.3: Schematic of the discrete feed reactor model for the AS-ALD reaction. The operation of the reactor is conducted in sequential pulses where a mixture of Hacac and N_2, BDEAS and N_2, and O_3 and N_2 are injected in the inlet stream for Steps A, B, and C, respectively. Purged materials including Hacac, H_2, and N_2 for Step A, BDEAS, N_2, and DEA for Step B, and O, O_2, O_3, and N_2 for Step C, are evacuated through the outlets, which include the wafer exit/entry.

plate has a substantial impact on the mass transport of reagents in the radial direction. Thus, various injection plate geometries, which are illustrated in Fig. 3.4, were previously proposed in prior work [108] to observe their impact on the fluid dynamics on the substrate surface. Results from the aforementioned work provided valuable information about the role of characteristic lengths on the rate of mass transfer in the radial direction. This work extends prior macroscopic modeling work for each reactor injection plate geometries by studying their effect on the spatiotemporal coverage and process time required to reach complete surface coverage.

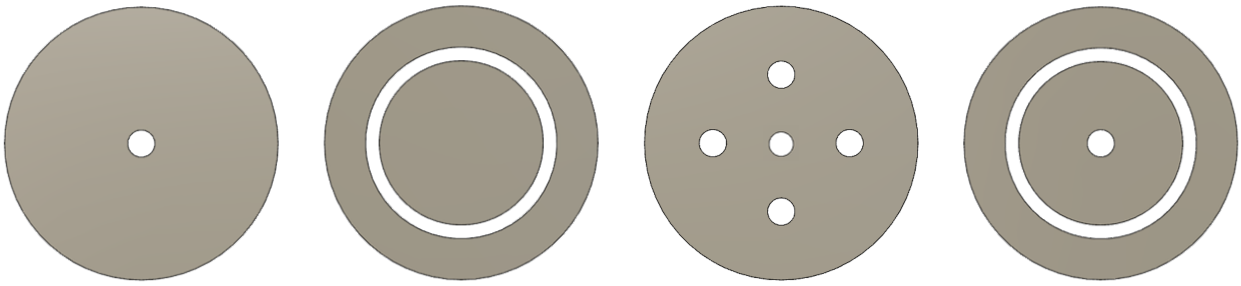
Table 3.2: Dimensions for the reactor configurations.*

Reactor Dimension	Value
Plate Diameter	290 mm
Ring Inlet Outer Diameter	170 mm
Ring Inlet Inner Diameter	130 mm
Round Inlet Diameter	20 mm
Round Outlet Diameter	4 mm
Showerhead Diameter	250 mm
Showerhead Pores Diameter	10 mm
Showerhead Plate Thickness	0.5 mm
Showerhead Plate/Wafer Vertical Gap Distance	5 mm
Reagent Inlet Plate/Showerhead Plate Vertical Gap Distance	3 mm
Wall Sector Angle	40°

*All dimensions are fixed for each reactor configuration.

3.2.2 Meshing

Following the construction of the reactor model, a discretization process is conducted to produce conformal meshes that balance computational efficiency and accuracy when performing the finite element method. Meshes for each reactor model are produced from "Meshing Mode," a



(a) (b) (c) (d)

Figure 3.4: Various feed distributor geometries for (a) Single, (b) Ring, (c) Multi, and (d) Combined reactor configurations.

feature of the multiphysics software, Ansys Fluent, in a prior work [108]. The aforementioned meshes were generated by optimizing mesh quality parameters based on the tetrahedral geometries of the discretized cells, which include the orthogonality, aspect ratio, and skewness [3]. To maximize each reactor configuration mesh, optional remeshing tools were then applied to the irregular surface and volume cells. In additional to maintaining balanced mesh quality, this work aims to minimize the number of cells required to produce the 3-D meshes to reduce the complexity of the computational fluid dynamics simulation, where each reactor configuration comprises 1.1 to 1.2 million cells.

3.2.3 Computational fluid dynamics simulation framework

The macroscopic CFD simulation is constructed by defining boundary, operating, and solver conditions that are specific to the AS-ALD process. This simulation will employ a strategy for

solving the mass, momentum, and energy transport equations, which are described as follows:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \overrightarrow{v}) = S_m \tag{3.8}$$

$$\frac{\partial}{\partial t}(\rho \overrightarrow{v}) + \rho (\overrightarrow{v} \cdot \nabla) \overrightarrow{v} = -\nabla P + \nabla \cdot (\overline{\overline{\tau}}) + \rho \overrightarrow{g} + \overrightarrow{F} \tag{3.9}$$

$$\frac{\partial}{\partial t}(\rho E) + \nabla (\overrightarrow{v} (\rho E + P)) = -\nabla \left(\Sigma h_j \overrightarrow{J}_j \right) + S_h \tag{3.10}$$

where the mass transport equation in Eq. (3.8) is related to the gas-phase species flux, which is represented by the product of the gas-phase species density, ρ, and the velocity of the species, $\overrightarrow{v}$, and is related to the species source generation and consumption flux rate, S_m. The momentum transport equation in Eq. (3.9) relates the rate of momentum per unit volume to the convection, pressure, viscous, and gravitational forces where P is the operating pressure of the reactor, $\overline{\overline{\tau}}$ is the normal two-rank stress tensor, $\overrightarrow{g}$ is the gravitational acceleration constant, and $\overrightarrow{F}$ is the force acting on the system. The energy transport equation defined in Eq. (3.10) describes the relation of the accumulated rate of system energy, E, with the convective, conductive, and energy source generation or consumption, S_h, rates, where h_j and $\overrightarrow{J}_j$ is the sensible enthalpy and mass diffusion flux, respectively, of the gas species j.

Ansys Fluent contains multiple fluid dynamics models that can be used to describe the behavior of the fluid flow. Due to the small reactor sizes and observance of laminar behavior from prior research [108], a laminar fluid model is defined in the simulation. The mass transport is simulated by specifying gas-phase reagent and byproduct species that are present in the Ansys ChemKin database and thermophysical property data generated from experimental works and *ab initio* quantum mechanics calculations discussed in Section 3.1. The source generation and consumption flux rate terms are evaluated through the kMC simulation and defined as boundary conditions on the wafer surface through user-defined functions (UDFs). Additionally, this simulation considered the role of ozone decomposition within the reactor and surface of the wafer. The reactor is also

operated under isothermal and isobaric conditions by assuming that a temperature control system is used to maintain the temperature on the wafer surface and that a vacuum pump is effectively applied to regulate the pressure within the reactor chamber.

A pressure-based coupled solver method is integrated into this work to simultaneously solve the momentum and pressure-based continuity equations in a parallelized algorithm to reduce computation time at a cost of increased memory requirement. To circumvent this issue, CPU-based (central processing unit) nodes were integrated into this work comprising 48 and 36 cores with 512 GB and 384 GB of dynamic random-access memory (DRAM), respectively, and executed through text-user interface (TUI) commands to minimize graphical power. Additionally, a fixed timestep method of step size $\Delta t = 0.001$ s is defined, which is within the Courant number threshold recommended by the default settings for the program. Lastly, under-relaxation factors of 0.5 were assigned to all gas-phase species involved in the mass transport calculations to minimize the potential for divergent or oscillatory residual responses that could potentially be generated from the source flux rate terms evaluated from the kMC simulation.

Reactor	Temperature (K)	Pressure (Pa)	Mole Fraction			Mass Flow Rate (kg/s)
			Hacac	BDEAS	Ozone	
Single	573	300	0.50	0.50	0.20	2.00×10^{-5}
Ring	573	300	0.50	0.50	0.20	2.00×10^{-5}
Multi	573	300	0.50	0.50	0.20	Each Inlet: 4.00×10^{-6}
Combined	573	300	0.50	0.50	0.20	Single: 1.00×10^{-5} Ring: 1.00×10^{-5}

Table 3.3: Operating conditions for each reactor geometry.

3.3 Multiscale modeling

The efficacy and impact that simulations can have naturally depends on their accuracy and precision. Generally speaking, the accuracy of a simulation can always be improved by increasing the computational costs; for example, lowering the integration timestep when numerically solving a differential equation will improve the accuracy of the final answer while increasing the number of calculations that must be made to reach that final answer. Thus, one of the driving motivations for this paper is finding an optimal balance between the accuracy and the computational cost of the simulation.

One commonly used method to improve simulation accuracy without an expensive computational cost is multiscale simulation [116]. This method comprises two simulations that run concurrently and interact with each other: a mesoscopic kMC model that simulates the surface kinetics of the wafer as a function of the pressure and temperature, and a macroscopic CFD model that simulates how the pressure fields within a reactor evolve with time. These two interacting models improve the overall accuracy of the simulation because their domains are intrinsically linked. The surface reactions on the wafer generate and consume products and reagents, which affects the overall pressure fields in the reactor, which affects the reaction rate on the wafer surface. Thus, to improve simulation accuracy, the two models are integrated together in a multiscale framework as shown in Fig. 3.5.

At each integration timestep, $\Delta t = 0.001$ s, the CFD model takes the generation and consumption terms calculated by the kMC model in the previous timestep into account when calculating the pressure fields in the reaction. Then, the kMC model receives information about the species pressure and temperature at the surface of the wafer and uses that to calculate the extent of any surface kinetics, as well as the resulting consumption and generation of species. After repeating this step for multiple timesteps, the multiscale simulation offers a comprehensive understanding of how the wafer surface reactions evolve as time progresses inside the reactor. With the computa-

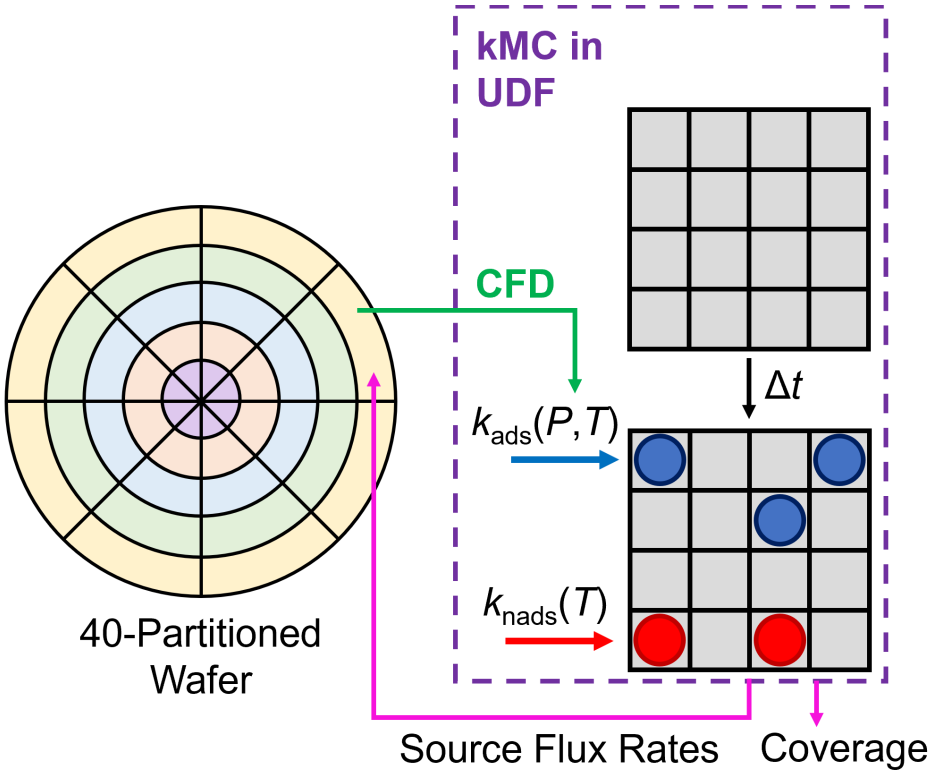


Figure 3.5: Illustration of the multiscale CFD modeling framework. The wafer is partitioned into 40 sections in the CFD simulation to produce a collection of 40 surface pressure and temperature datasets that are used to calculate the reaction rate constants in user-defined functions (UDFs). The kMC simulation, which is performed in the UDF calculates the surface coverage and source flux rate terms that are transmitted to the CFD simulation.

tional resources and numerical simulation specifications stated above, the computing time ranged from 4 to 6 hours for a process time of 3.5 seconds for the 48-core and 36-core nodes, which allows for effective data production.

3.4 Multiscale Simulation Results and Discussion

The aim of the multiscale CFD simulation is to give a quantitative understanding of how various reactor designs affect the process efficiency of the AS-ALD reactions, which is directly tied to the process time. Due to the self-limiting nature of AS-ALD, the process is naturally resistant to over-deposition. Rather, the only drawbacks of overprocessing are unneeded consumption of the

reagents and a decreased product throughput. Thus, there is a large economic incentive to minimize the process time, as this will both reduce the reagent consumption and increase production capabilities.

To evaluate the overall process efficiency, two main criteria are taken into account: the minimum process time and surface coverage uniformity. The minimum process time is the time at which the substrate achieves full coverage for the surface-terminated product, which is directly dependent on the reagent dosage time. On the other hand, the surface coverage uniformity is a measure of how the standard deviation of the coverage on the substrate surface changes with time. Naturally, the lower the overall standard deviation, the less reagent is wasted, and the reverse implies greater reagent wastage. These two criteria are positively correlated, as a low standard deviation distribution implies an effective usage of the reagent, which then implies that the process will be completed quickly.

Fig. 3.6 illustrates the temporal progression of coverage for Steps A, B, and C for each reactor configuration. Generally, the ring-shaped reactor geometry underperforms due to the vacuum pressure forces and steric screening effects from byproducts that result in a lack of fluid transport toward the center of the wafer. Meanwhile, the single and multi-shaped injection plates are conducive towards achieving full surface coverage in minimal processing times by concentrating the reagent toward the center of the wafer. With smaller characteristic lengths, especially for the combined reactor geometry, the injection flow rate is unable to overcome the effects of the vacuum pressure forces. Thus, an initial delay in coverage is observed for the coverage profiles of the combined model when compared to those of the single and multi feed reactors. Table 3.4 also summarizes the processing times required to achieve full surface coverage for each reactor configuration and each step in the AS-ALD ABC cycle, where the single and multi reactor models were observed to have the fastest processing times.

The reactor design with the smallest minimum process time can be quickly determined by examining the average coverage progression across the wafer as a function of time, as shown

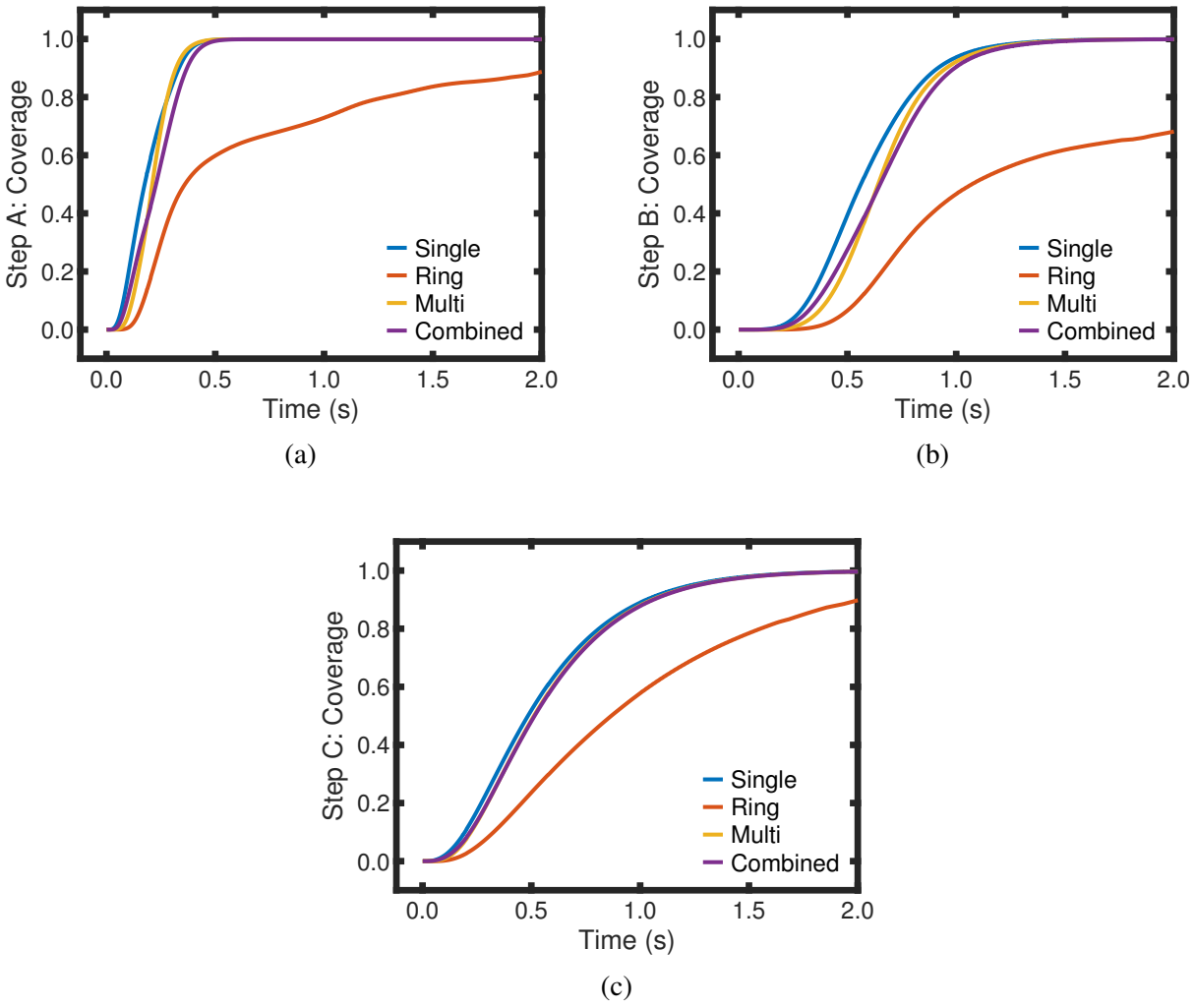


Figure 3.6: Reactor configuration comparison of the temporal progression of the average surface coverage for (a) Step A, (b) Step B, and (c) Step C.

in Fig. 3.6. From Table 3.4, it can be seen that the multi-shaped reactor performed the best overall with a total minimum process time of 3.638 s. However, to understand why this reactor design performed the best, it is necessary to examine more specialized data, such as the standard deviation progression as a function of time and contour plots of the coverage at specific points in time.

The multi-shaped reactor, which has the smallest process times, also consistently has the lowest standard deviation for all points in time. However, the single and combined-shaped reactors have

Reactor	Process Time (s)			Total Process Time (s)
	Step A	Step B	Step C	
Single	0.466	1.431	1.783	3.680
Ring	2.310	4.586	2.596	9.492
Multi	0.453	1.448	1.737	3.638
Combined	0.509	1.508	1.755	3.772

Table 3.4: Computed process times required to obtain full surface coverage for the terminated products from Steps A, B, and C as a function of reactor configuration.

similar standard deviation curves, illustrated in Fig. 3.7, for all three reactions, but the single-shaped reactor consistently outperforms the combined-shaped reactor in terms of process time. The standard deviation plots demonstrate the uniformity of wafer surface coverage for Steps A, B, and C of the AS-ALD process, but lack critical information in determining the spatial dependence of the surface coverage on the injection geometry.

Meanwhile, the contour coverage plots directly reflect the progress of the surface reactions. From Eqs. (3.1) and (3.2), it can be seen that, in an isothermal reactor, the reaction rates are only a function of pressure. Thus, the progress of the surface reaction at a particular point is related to the pressure which that particular point has experienced since the start of the reaction. This means that the contour coverage plots effectively represent how ideal the pressure distribution of a given reactor design is. In Fig. 3.8, all the reactor designs have a high coverage value in the areas directly underneath the precursor dispensers. However, the most valuable information from these contour graphs are the areas with low coverages; these areas represent sections far from the inlet, where the reactor design plays a major role in how high the precursor pressure necessary for carrying out the reactions is. For example, the ring-shaped reactor design has a large area in the center with 0 coverage, which means that almost no precursor is flowing there. This matches the results in [108], which found that the pressure in the center is very low for the ring-shaped reactor. Meanwhile, out of the single, multi, and combined-shaped reactors, the multi-shaped reactor

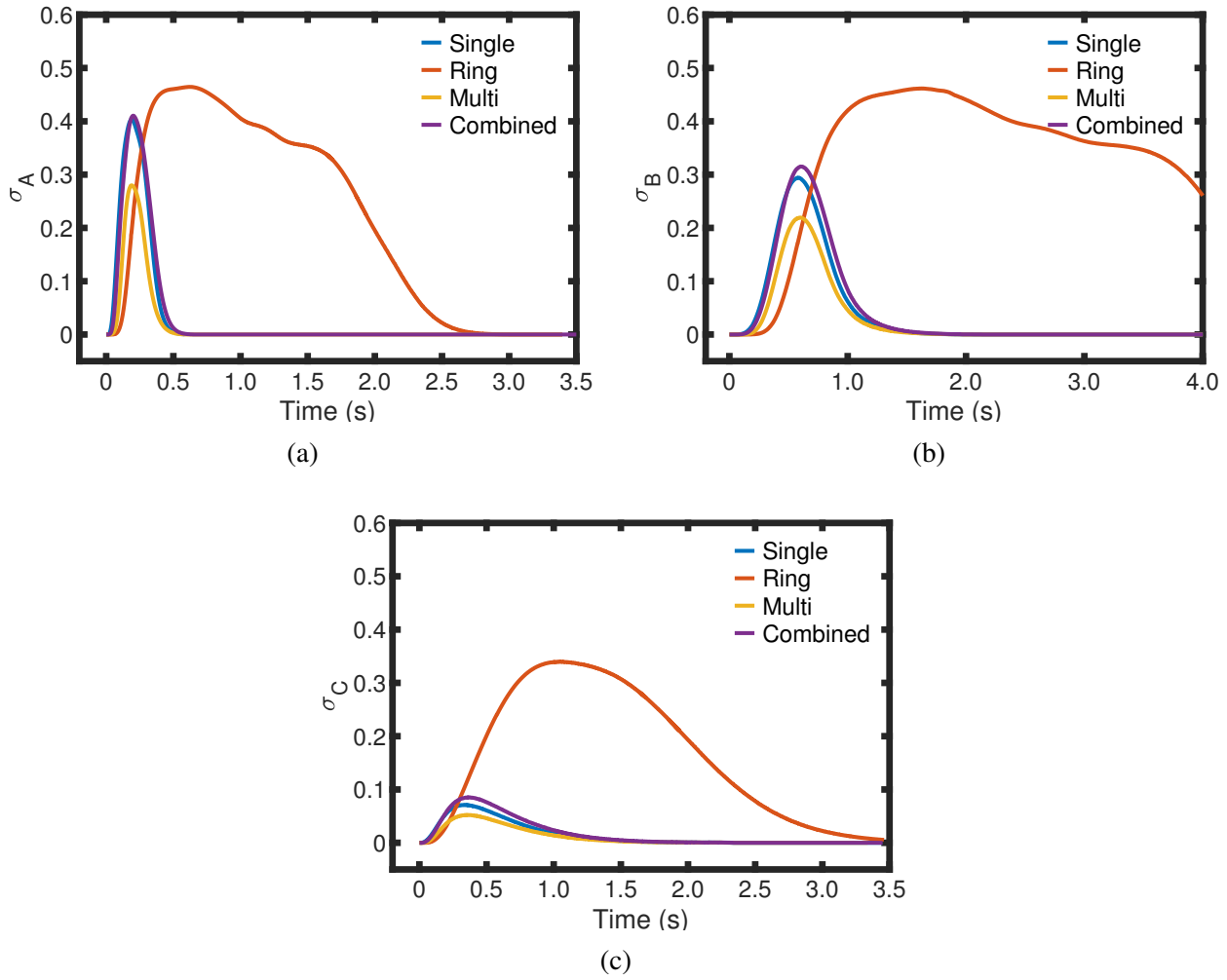


Figure 3.7: Reactor configuration comparison of the temporal progression of the standard deviation, σ, in surface coverage for (a) Step A, (b) Step B, and (c) Step C.

has the best coverage contour plot in that its outer edges have the highest coverage. Even though the coverage at the outer edge is the same for both the single and combined-shaped reactors, the coverage at the middle of the reactor, $r = (12, 16)$ *cm*, is higher for the single-shaped reactor. This means that the single-shaped reactor design has a better pressure distribution in the middle section of the wafer, rendering it superior to the combined-shaped reactor design. The results of the contour plots show that the multi-shaped reactor design is the best, followed by the single, combined, and the ring-shaped in that order. Similarly, the contour plots for steps B and C, as

shown in Figs. 3.9 and 3.10, demonstrate that the multi-shaped reactor performed the best with the other three reactor designs following in the order described above. This correlates with the performances noted in Table 3.4, which shows that contour plots are a good metric for identifying efficient reactor designs. Specifically, they are able to compare and determine what reactor design has a better transient pressure profile. Additionally, Figs. 3.8 to 3.10 illustrate slight asymmetric coverage characteristics that are attributed to numerical error from the first-order implicit method utilized to solve the transient transport equations and randomness that originates from the random number used for reaction selection and time advancement in the kMC simulation. While previous work [108] has demonstrated that the discrete feed reactor models produce uniform flow characteristics, the flow is not a significant indicator of coverage uniformity due to the potential of steric hindrance effects and molecular collisions that impede reagent adsorption.

It is also important to point out that the contour plots provide an important insight; there is a significant difference between the developed pressure profiles at 3.0 s and the initial pressure profiles for each reactor design. The former pressure profiles were examined in a previous work, which concluded that the combined-shaped reactor was the best reactor due to its optimal pressure fields [108]. However, the results presented in Table 3.4 differ substantially, as they show that the multi-shaped reactor is the best reactor with the shortest minimum process time. One explanation that accounts for both of these observances is that the pressure fields at the initial stages of the process play a pivotal role in the overall efficacy of the reactor design. The coverage evolution at the fringes of the substrate, which plays the greatest role in determining the minimum process time, is a function of the entire pressure profile evolution. Thus, it is important for the ideal reactor design to also quickly reach these fringe areas. While the combined-shaped reactor may have more even pressure profiles, the multi-shaped reactor evidently distributes the reagent quicker.

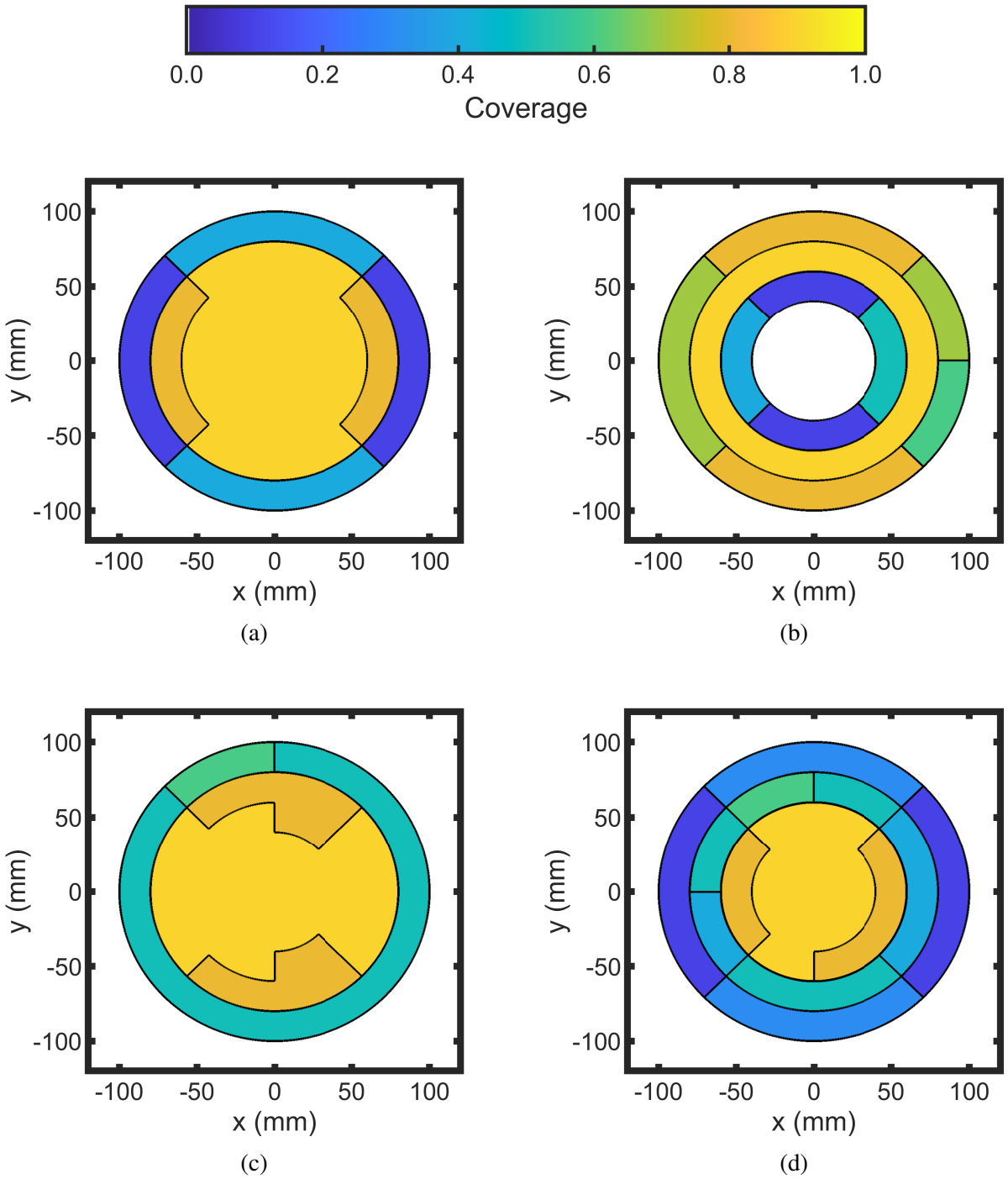


Figure 3.8: Comparison of contour plots of various reactor configurations, (a) Single, (b) Ring, (c) Multi, and (d) Combined, to study the spatial behavior of the surface coverage of the terminated Step A product for a 40-partitioned substrate at a time of 0.3 s.

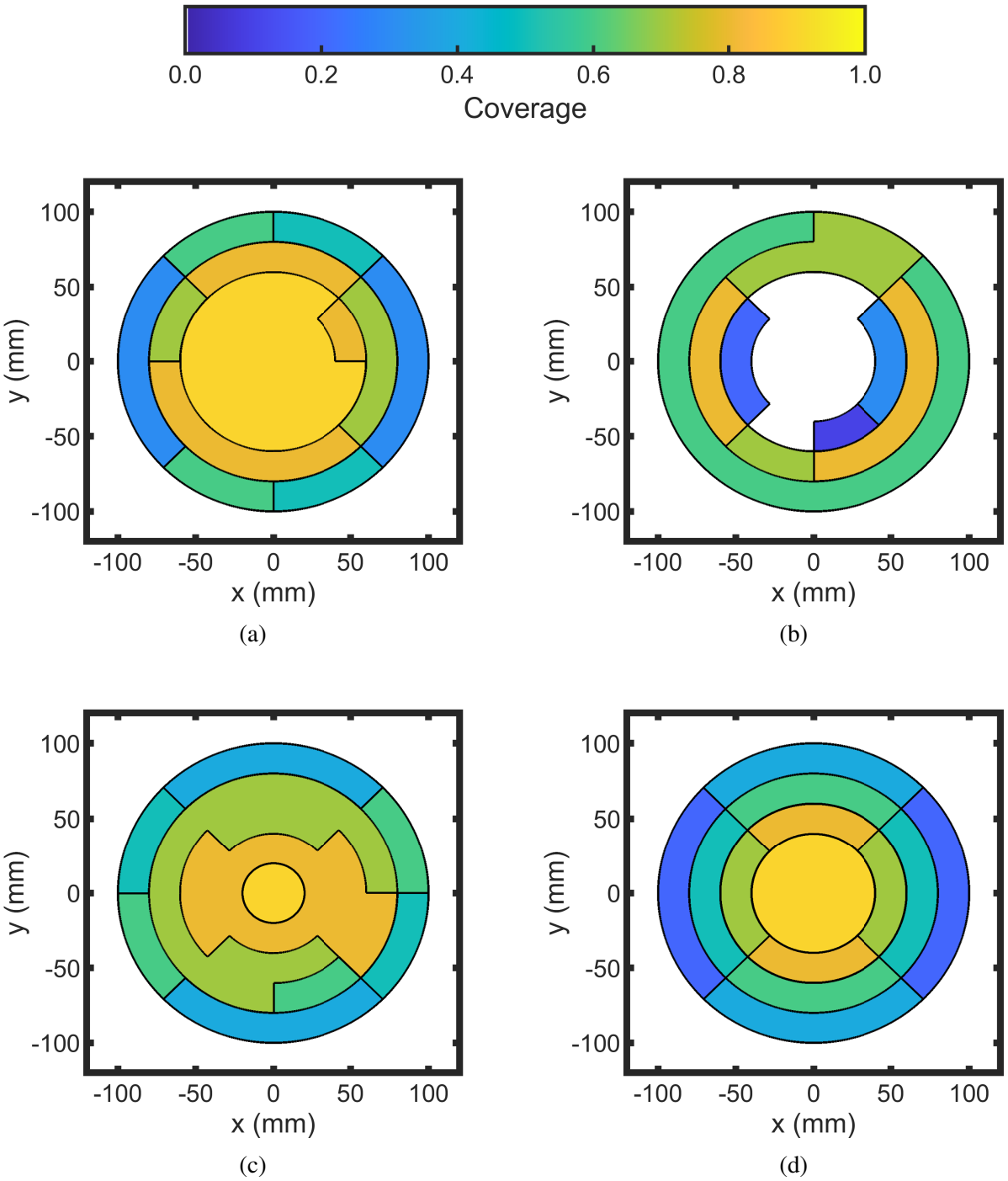


Figure 3.9: Comparison of contour plots of various reactor configurations, (a) Single, (b) Ring, (c) Multi, and (d) Combined, to study the spatial behavior of the surface coverage of the terminated Step B product for a 40-partitioned substrate at a time of 0.8 s.

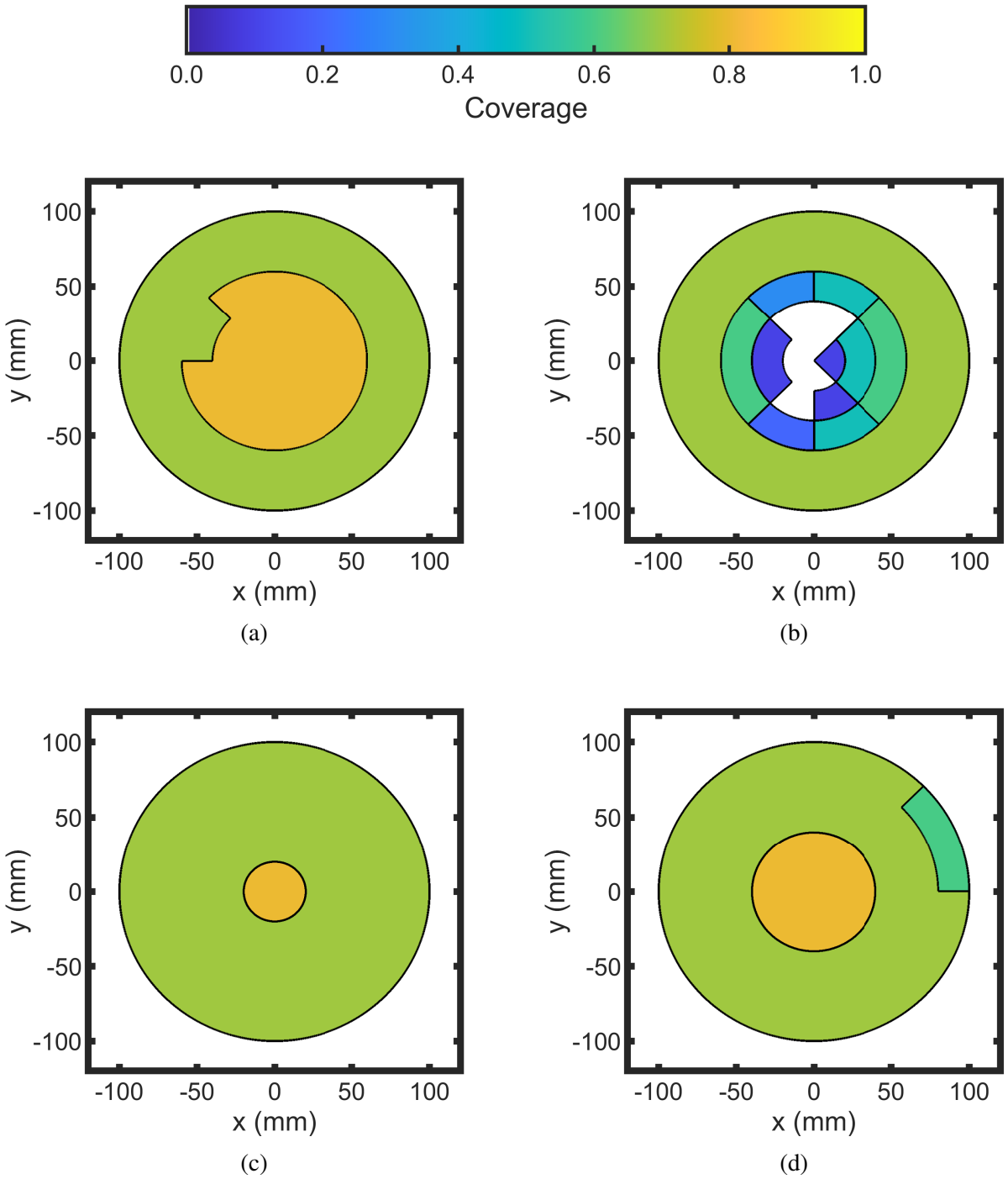


Figure 3.10: Comparison of contour plots of various reactor configurations, (a) Single, (b) Ring, (c) Multi, and (d) Combined, to study the spatial behavior of the surface coverage of the terminated Step C product for a 40-partitioned substrate at a time of 1.0 s.

The minimum process time for each reactor design is defined as the time required to reach 99.9% coverage for each individual reaction. Overall, the multi-shaped reactor has the best results with a total minimum process time of 3.638 s. To understand why this reactor design performs the best, the contour plot is instrumental in revealing the fact that this design spread the precursor to the remote parts of the wafer the quickest out of the examined designs. This detail explains why, even though the combined-shaped reactor has the most even pressure profiles, it ranks third in terms of minimum process time.

3.5 Conclusions

In this work, a multiscale computational fluid dynamics (CFD) model of an area-selective atomic layer deposition (AS-ALD) reactor was developed to study the spatiotemporal progression of surface coverage to allow the scale-up of AS-ALD processes. To improve the accuracy of the multiscale CFD model, this work expanded on the mesoscopic, kinetic Monte Carlo (kMC) algorithm by considering the role of the number of unoccupied atomic sites on the time progression computation. Findings suggested that surface coverage temporal progressions were similar to results conducted in prior work and were able to resolve the ambiguity of surface coverage from prior kMC methodology. Additionally, various reactor injection geometries for the macroscopic CFD simulation were constructed to enhance the mass transfer of reagent on the surface of a semiconductor substrate and study their impact on the process time required to obtain full surface coverage and the uniformity of the surface coverage with respect to time. Results indicated that two reactor models, the multi and the single, required minimal processing time and were characterized by homogeneous, temporal surface coverage.

Chapter 4

Machine Learning Modeling and Run-to-Run Control of an Area-Selective Atomic Layer Deposition Spatial Reactor

The preponderance of high-performance semiconducting materials in advanced electronics is a motivation for increasing their supply and improving inefficient manufacturing practices with processes that yield greater semiconductor quantity and have higher accuracy. For example, the aftermath of semiconductor shortages following the COVID-19 pandemic [35, 78] and the looming political implications [102] regarding the dependence of national security on semiconductors, led to increased government investment toward the strengthening of semiconductor yield and quality through the development of innovative fabrication processes [111]. These investments have been fruitful for increasing semiconductor supply and quality following the rise of atomic layer deposition (ALD) [36, 49], atomic layer etching (ALE) [11, 51], and area-selective atomic layer deposition (AS-ALD) [15, 16, 73] processes, which possess significant roles in the construction of transistors with nanoscale dimensions by precisely depositing or etching monolayers of substrate material in sequential and cyclical procedures. Recently, research has been conducted in

achieving film uniformity for AS-ALD methods without requiring post-processing steps such as ALE to ensure self-alignment during the transistor stacking procedure. AS-ALD has been previously experimented through *in silico*, atomistic-mesoscopic and multiscale computational fluid dynamics (CFD) modeling to study the spatiotemporal behavior of the process in a spatial, rotary reactor configuration [109, 122, 127]. Although AS-ALD is characterized by high accuracy deposition rate, such process requires regulatory monitoring to ensure that quality conformance is maintained, thereby mitigating the risk of semiconductor performance degradation and nonconformance to product criteria.

To ensure process monitoring in short timescale intervals, semiconductor manufacturing industries propose *ex situ* or offline control referred to as run-to-run (R2R) control [82], which performs controller adjustments after each batch run or in this case, after the completion of one cycle of AS-ALD. For example, Critical Manufacturing uses an integrative manufacturing execution system (MES) to regulate semiconductor manufacturing processes for etching and lithography by adopting run-to-run control algorithms that implement suitable control actions [1]. Past works have established run-to-run control for atomic layer etching (ALE) processes conducted in a stationary plate and spatial sheet-to-sheet reactor configuration [107, 125], for plasma-enhanced atomic layer deposition (PEALD) in an inductively coupled plasma (ICP) reactor [121], and for thermal atomic layer deposition (ALD) in a furnace reactor (PEALD) [129]. This work aims to implement R2R controller action on an AS-ALD rotary reactor in an *in silico* experiment that regulates the output, growth per cycle, of a multiscale computational fluid dynamics (CFD) simulation when perturbed by pressure and kinetic shift disturbances by performing feedback after each cycle of the AS-ALD process. To reduce the computational demand for the highly complex multiscale simulation, machine learning and C programming language are substituted in place of CFD and the kinetic Monte Carlo simulations to facilitate the generation of output data for the R2R controller. This work will first examine the development of the multiscale CFD model in Section 4.1, the implementation of machine learning to generate pressure field data in place of the CFD simulation is discussed in Sec-

tion 4.2, and the integration of the R2R controller is elucidated in Section 4.3 with the closed-loop simulations presented in Section 4.4.

4.1 Multiscale computational fluid dynamics model

Area-selective atomic layer deposition (AS-ALD) is defined by a three-step, ABC, process comprising inhibition, adsorption, and oxidation steps that perform selective deposition on growth areas of the substrate [74]. However, conducting experiments in laboratory environments is a time-consuming and laborious process that may yield variable results that depend on idealistic operating conditions. Thus, this section of the manuscript will discuss the development of an *in silico* approach to replicate the experimental laboratory settings by combining atomistic modeling through first principles quantum mechanics simulations to evaluate kinetic parameters inherent to rate-determining reaction pathways, mesoscopic modeling through the stochastic kinetic Monte Carlo (kMC) computations for studying the surface-scale kinetics, and macroscopic computational fluid dynamics (CFD) to evaluate the mass transfer of reagent and byproduct species in the fluid phase. The conjunction of such simulations constitutes a multiscale model that performs simulations through multiple time and length scales, which is advantageous to the development of process scale-up in industrial applications.

4.1.1 Atomistic modeling

The development of an atomistic model is imperative for the calculation of reaction rate constants that for rate-determining reactions involved in each step of the ABC, AS-ALD process. This work utilized an inhibitor, acetylacetone (Hacac), a precursor, bis(diethylamino)silane (BDEAS), and an oxidant, ozone (O_3) for a substrate composed of a non-growth area, aluminum oxide (Al_2O_3), and growth area, silicon oxide (SiO_2) where all reaction rate pathways are elucidated in several reference works [76, 122]. The surface morphologies for Al_2O_3 and SiO_2 were assumed

to be pure in the α [74] and β [97] phases, respectively. Two types of reaction rate constants, the pressure and temperature dependent adsorption, $k_{ads}(P,T)$, and temperature dependent non-adsorption, $k_{nads}(T)$, are computed through Collision Theory and the Arrhenius equation, which depend on multiple variables including the sticking coefficient, σ, for adsorption reactions and the activation energy, E_A, and pre-exponential factor, ν, (evaluated using Transition State Theory) for nonadsorption reactions. Equations for the pressure and temperature dependent Collision Theory and the temperature dependent Arrhenius model are described as follows:

$$k_{ads} = \frac{PA_{site}\sigma}{Z\sqrt{2\pi m k_B T}} \tag{4.1}$$

$$k_{nads} = \nu \exp\left(-\frac{E_A}{RT}\right) \tag{4.2}$$

where P is the surface pressure of the reagent, A_{site} is the surface area of the active site, Z is the coordination number, m is the atomic mass of the adsorbate species, k_B is the Boltzmann constant, T is the ambient operating temperature of the reactor, and R is the ideal gas constant. Additionally, some thermophysical parameters of species such as the specific heat capacity, standard state enthalpy and entropy, density, dynamic viscosity, and thermal conductivity, are needed to reduce the degrees of freedom for the heat transfer dynamics in the macroscopic computational fluid dynamics (CFD) simulations. This work obtains sticking coefficient parameters for various species through an exhaustive literature search where the sticking coefficients for Hacac [36], BDEAS [101], and O_3 [62] were also determined by selecting sticking coefficients from species that are molecularly similar to the species in this work such as that of Hacac, which were validated with experimental processing times [76] to reach full coverage. Thermophysical property data were obtained from the National Institute of Standards and Technology (NIST) database and Material Safety Data Sheets (MSDS) from online references, experimental works, and databases through Ansys Chemkin. Thermophysical property data not found in literature references are calculated through

first principles quantum mechanics computations using density functional theory (DFT), nudged elastic band (NEB) methods, and quasi harmonic approximation (QHA) calculations through the open-source software, Quantum ESPRESSO. The reaction rate constant parameters and thermophysical property data are detailed by Yun et al. [122, 127].

4.1.2 Mesoscopic modeling

The AS-ALD processing occurs when the surface of the substrate is exposed to reagent along the wall-fluid boundary layer. To replicate the stochastic nature of the surface-scale kinetics, this work employs a modified kinetic Monte Carlo (kMC) method based on the algorithm proposed by Bortz, Kalos, and Lebowitz that is sometimes referred to as the n-fold way or BKL algorithm [8]. Whereas the classic BKL algorithm employs a Markov chain that advances the entire grid with each step, the modified kMC used in this project employs a Markov chain that only advances one site of the grid with each step. Specifically, the modified kMC randomly selects a site from the overall grid, followed by a random selection of a reaction pathway j, which then advances the internal timer of the kMC system with the following formulas:

$$k_{tot} = \sum_{i=1}^{N} k_i \tag{3a}$$

$$\sum_{i=1}^{j-1} k_i \leq \gamma_1 k_{tot} \leq \sum_{i=1}^{j} k_i \quad \text{where } j = 1, 2, \ldots, N \tag{3b}$$

$$\delta t = -\frac{\ln \gamma_2}{k_{tot} M} \quad \text{where } t \rightarrow t + \delta t \tag{3c}$$

where k_{tot} is the sum of all possible reaction pathways for a single active site, $\gamma_1, \gamma_2 \in (0, 1]$ are uniformly chosen random numbers, δt is the time evolution that is iteratively summed to the total process time t and represents the time for the active site to convert to the next state, and M is the number of total active sites. One advantageous modification to the conventional BKL approach

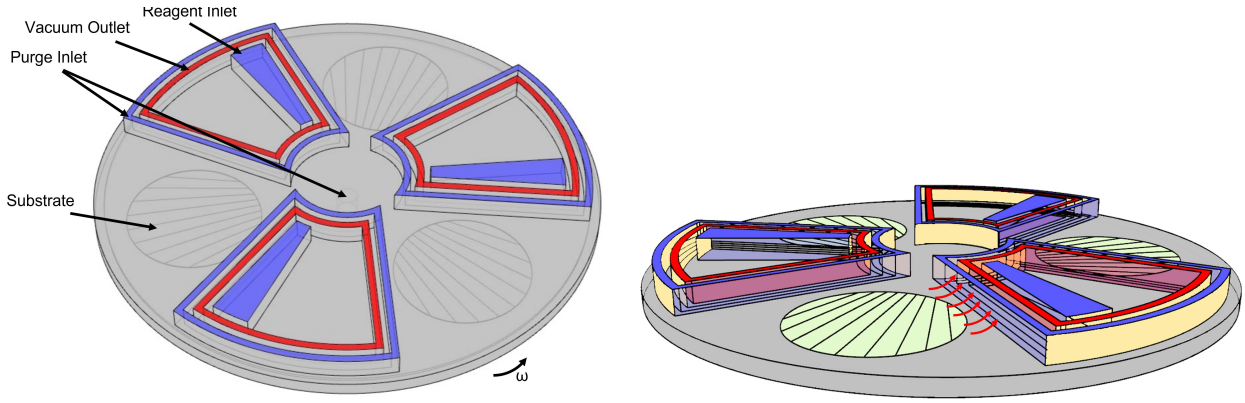
is the inclusion of the number of available active sites (e.g., the number of active sites that have not reached a final state or are unoccupied) to the time evolution described by Eq. (3c). With a decreasing number of available active sites, the probability of adsorption decreases to account for the fact that incoming reagents are less likely to encounter reactive substrate sites, thus causing the time evolution to increase. Additionally, M resembles the changing probability of unoccupied sites due to the nonspontaneous nature of the final reactions for Steps A, B, and C to form the terminated product. Since the number of unoccupied sites decreases with increasing process time, the probability of reagent adsorption will decrease, which builds on the assumptions made by Bortz et al. [8] and Gillespie [39]. For example, Kim et al. accounted for the role of unoccupied sites and steric hindrance effects in their Monte Carlo simulation through a random point searching methodology for an Al_2O_3/SiO_2 substrate [53]. Klement et al. also considered the role of adsorbate collisions and unoccupied active sites on the surface of the substrate for an AS-ALD process on a TiO_2/SiO_2 substrate [56]. This work similarly adopts the random site selection approach while including steric hindrance effects associated with the bulky properties of BDEAS and the free rotation of the adsorbed species, which has been addressed in prior work [122].

To optimize the simulation efficiency and compatibility with multiphysics software, Ansys Fluent, for computational fluid dynamics (CFD) simulations, the kMC script is conducted in the C programming language and utilized with various C compilers including GNU's GCC and Intel oneAPI's ICX, where CPU times are examined in Section 4.2. Additionally, C language has a stronger affinity for running parallel tasks compared to programming languages like Python, which further enhances the speed of the computation.

4.1.3 Macroscopic modeling

The reaction rate constants used in the kMC simulation are dependent on the surface pressure and temperature, as well as whether the reactions are adsorption or nonadsorption reactions. The dynamics of the surface pressure and temperature depend on the design or geometry and the

operating conditions of the reactor used to conduct the AS-ALD process. This work adopts a spatial, rotary reactor configuration that was previous developed [127] and illustrated in Fig. 4.1. The three-dimensional (3D) reactor is optimized geometrically and the operating conditions are methodically chosen to reduce reagent intermixing while establishing uniform substrate exposure to reagent upon entry to the reaction zones. The reactor is operated under isothermal conditions; thus, surface temperature profile is uniform and fixed spatiotemporally, and under isobaric conditions for total reactor pressure. Due to the small volume and sizing of the reactor with height in length scales of 10^{-3} m, a laminar model is specified for the fluid dynamics [90]. Additionally, constant inlet flow rates, reagent concentrations, outlet flow rates, and plate rotation speeds are defined with total inlet and outlet flow rates being equal, which are detailed in Table 4.1.



(a) Top View (b) Side View

Figure 4.1: Schematic of the spatial, rotary reactor used for the AS-ALD process that illustrates the transfer of the wafer through reaction zones by a rotating plate moving at ω *rad/s*.

It is also illustrated that the substrate is partitioned into 10 sections to collect the spatiotemporal behavior of deposition rate and pressure field data. Each partitioned section is assigned an index for identification purposes that is depicted in Fig. 4.3.

The rotary reactor is also discretized into a dynamic mesh comprising tetrahedral and triangular cells that are geometrically constructed thro ugh mesh quality criteria defined by the software, Ansys Fluent. A dynamic mesh is methodically defined to perform remeshing steps that maintain

Table 4.1: Operating conditions of the rotary reactor defined to the multiscale CFD simulation.

Operating Condition	Value
Reactor Pressure	1330 *Pa*
Reactor Temperature	523 *K*
BDEAS Mole Fraction	0.50
O_3 Mole Fraction	0.05
Inlet Mass Flow Rate	2.00×10^{-5} *kg/s*

the mesh quality through each movement of the mesh.

Computational fluid dynamics (CFD) is performed to study the characteristics of the reagent flow along the surface of the substrate by numerically and simultaneously solving the transport phenomena equations for mass, momentum, and heat, which are defined as follows:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \overrightarrow{v}) = S_m \tag{4.4}$$

$$\frac{\partial}{\partial t}(\rho \overrightarrow{v}) + \rho (\overrightarrow{v} \cdot \nabla) \overrightarrow{v} = -\nabla P + \nabla \cdot (\overline{\overline{\tau}}) + \rho \overrightarrow{g} + \overrightarrow{F} \tag{4.5}$$

$$\frac{\partial}{\partial t}(\rho E) + \nabla (\overrightarrow{v} (\rho E + P)) = -\nabla \left(\Sigma h_j \overrightarrow{J}_j \right) + S_h \tag{4.6}$$

where ρ represents the density of reagent, $\overrightarrow{v}$ is the reagent velocity, S_m is the reagent or byproduct consumption and generation source rate, respectively, P is the pressure of the gas-phase species, $\overline{\overline{\tau}}$ is defined as the rank-two stress tensor, $\overrightarrow{g}$ is the gravitational constant, $\overrightarrow{F}$ is the body force subjected onto the gas-phase species, E is the internal energy of the system, h_j is the sensible enthalpy of the species, j, J_j is the mass diffusion flux rate of species, j, and S_h is the heat transfer consumption or generation source rate. The CFD simulation employs a pressure-based coupled solver with a first-order, transient, and implicit numerical solver method with a fixed timestep size of 0.001 *s*.

4.1.4 Multiscale modeling

The conjunction of the atomistic, mesoscopic, and macroscopic simulations produces a multiscale model that enables various time and length scales to be examined. The multiscale model, depicted in Fig. 4.2, is constructed through the application of a user-defined function (UDF) script that is written in the C programming language and integrates C-based macros that are inherent to the Ansys Fluent program. UDFs allow users to customize tasks for specific computations in that are not available in the software client. To discuss the spatiotemporal aspects of the simulation, the program partitions the substrate into 10 sections in the radial direction of the center of mass of the rotary reactor, which is illustrated in Fig. 4.1. The multiscale simulation begins by performing CFD for a timestep size of 0.001 s to calculate the pressure field data on the substrate surface, which is partitioned into 10 sections. This reagent pressure data is calculated through an area-averaged approach for each section on the wafer, and is then sent to UDF to calculate reaction rate constants for adsorption and nonadsorption reactions. Next, the kMC script is executed for each wafer section in a parallel computation procedure until the process time, t, reaches the timestep size of 0.001 s where the output, growth per cycle (GPC), and the source generation and consumption rates for mass and heat are calculated. It is notable that since the reactor is operating under isothermal conditions, heat generation and consumption rates are negligible for this work. Next, the source generation and consumption mass source rate is defined to the surface boundary condition for each section of the wafer, where the next iteration for the multiscale simulation is executed repetitively until the wafer exits the reaction zone. To ensure computational efficiency is sufficient, this work adopted various CPU (Central Processing Unit) nodes comprising 36 and 48 cores with 384 GB and 512 GB of RAM (Random Access Memory), respectively.

A consequence of the partitioning is that earlier sections of the wafer partition will enter the reaction zone before later sections of the wafer where a noticeable delay in complete saturation exposure would be observed. This consideration is depicted from the multiscale simulation results in Fig. 4.3, which illustrates the spatiotemporal behavior of the area-averaged surface pressure for

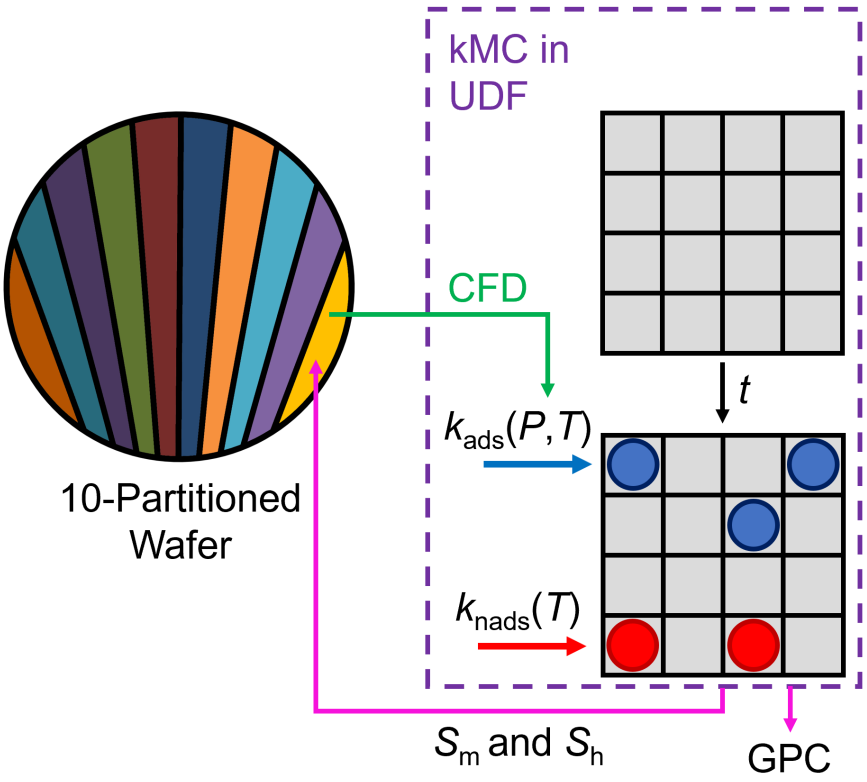
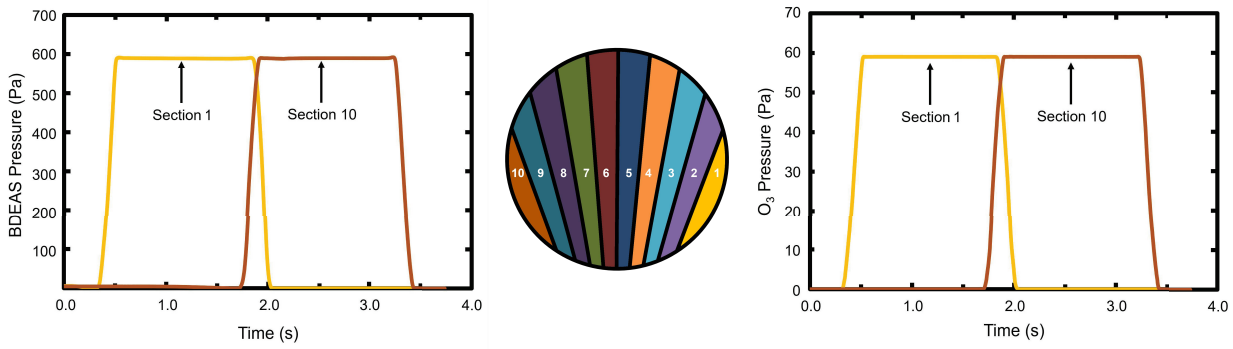


Figure 4.2: The multiscale process diagram that begins with CFD simulation to calculate area-averaged surface pressure and temperature data for evaluating reaction rate constants in the UDF script. Then, the kMC simulation is performed until the simulation reaches the CFD timestep size of 0.001 s to calculate the growth per cycle, GPC, and the source generation terms for mass and heat, S_m and S_h, respectively, which are then defined to the boundary condition on the wafer, which is partitioned into 10 sections.

BDEAS and O_3 for a constant rotation speed of 0.56 *rad/s*.

4.2 Pressure field generation through machine learning

One disadvantage of multiscale simulations are the computational complexity and inefficiency when data must be gathered in short time intervals. The multiscale model defined in Section 4.1 and shown in Fig. 4.2 may require minimum simulation times of 3 days and maximum simulation times of 14 days, depending on the rotation speed of the wafer that is defined to the multiscale model. Lower rotation speeds increase the residence time of the wafer in the reaction zones, which



(a) (b)

Figure 4.3: Multiscale modeling results for a constant rotation speed of 0.56 *rad/s* that depicts the time delay for later sections of the 10-partitioned wafer to observe maximum reagent exposure to (a) BDEAS and (b) O_3. The wafer number is synonymous to the substrate location, which is examined in Section 4.2.

increases the computation time. Thus, this work proposes a machine learning approach to correlate a multi-input-single-output (MISO) data set, to produce pressure field data in place of pressure field data generated through CFD.

4.2.1 Feedforward neural network for MISO system

A feedforward neural network (FNN) model is generated for a multi-input system composing of three variables (rotation speed, process time, and substrate location) to calculate a single predicted output, the growth per cycle (GPC). This FNN model is illustrated in Fig. 4.4, which describes the nodal connections that are conducted to train the FNN. The FNN was constructed using two hidden layers with each including 30 neurons to train and test data sets of 240,000 and 120,000 data points for BDEAS and O_3 pressure, respectively. It is notable that the Hacac pressure is not included in this work, since the GPC is a measurement taken from the growth area of the substrate. A two-hidden layer FNN is described by the following equations that describe the training of the

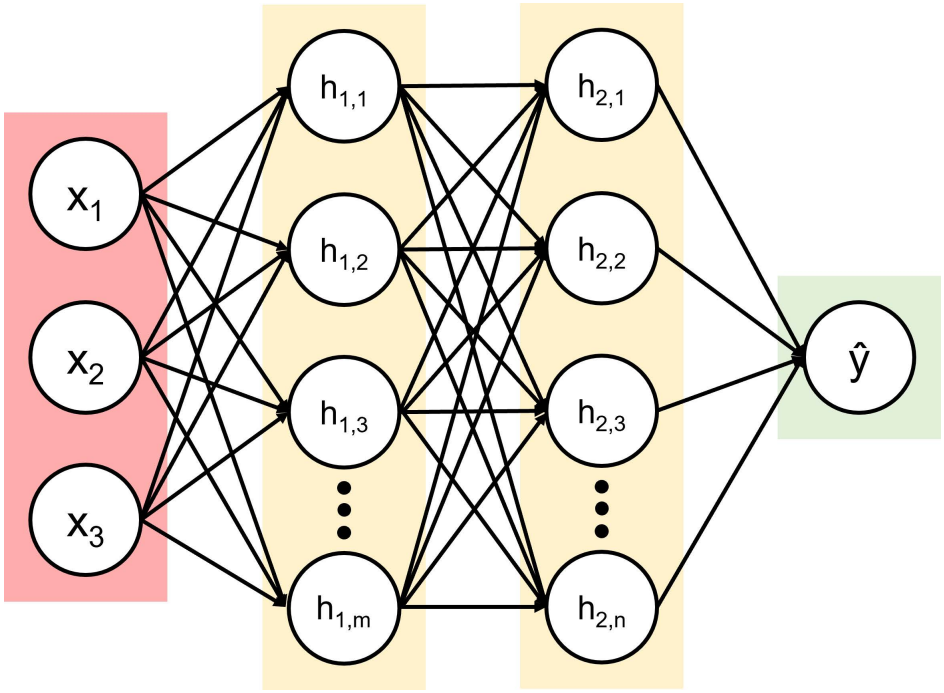


Figure 4.4: A feed-forward neural network with two hidden layers for a three-input-single-output model containing three inputs (rotation speed, process time, and substrate location) and output (growth per cycle).

FNN:

$$\hat{y} = W_2^T H_2 + B_2$$

$$H_2 = f(W_1 H_1 + B_1)$$

$$H_1 = f(W_0 X + B_0)$$

where $\hat{y}$ is the predicted output, GPC, $X \in \mathbb{R}^m$ is the input vector comprising $m = 3$ input parameters (rotation speed, process time, and substrate location), $W_0 \in \mathbb{R}^{n \times m}, W_1 \in \mathbb{R}^{n \times n}, W_2 \in \mathbb{R}^n$ are weights where $n = 30$ neurons for each hidden layer, $B_0 \in \mathbb{R}^n, B_1 \in \mathbb{R}^n, B_2 \in \mathbb{R}$, and $f(\cdot)$ is the rectified linear unit (ReLU) activation function. Additionally, the FNN models for BDEAS and O_3 were generated by using a training size of 90% and testing size of 10% using the Adam optimizer, a learning rate of 0.004, and by minimizing the mean squared error (MSE), which was

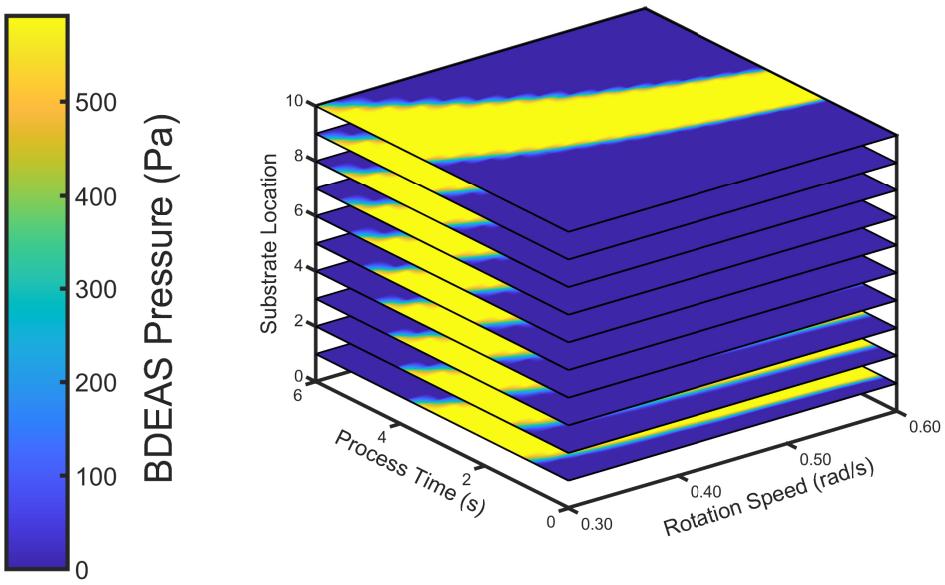
conducted through the open-source Python library, TensorFlow.

Predicted BDEAS and O_3 pressure data from the FNN model are depicted in Fig. 4.5 with the BDEAS and O_3 FNN models having MSEs of 0.0143 and 0.0121, respectively. Fig. 4.3 illustrates that substrate location affects the process time required for the substrate to become completely exposed to reagent due to the time delays for the sections to enter the reaction zone. Furthermore, increasing rotation speed also increases the time needed for the substrate to observe full exposure to reagent, which is reflective of the operation of the rotary reactor while the residence time of the wafer in the reaction zone increases with decreasing rotation speed. The precision and accuracy of the FNN model is presented in Fig. 4.6, which plots a sample of all data points to illustrate the absolute error of the FNN predicted pressures to the multiscale CFD pressures for BDEAS and O_3. Results demonstrate that a majority of points have marginal error, thus enabling the substitution of the FNN model in place of the pressure data generated from the multiscale CFD simulation.

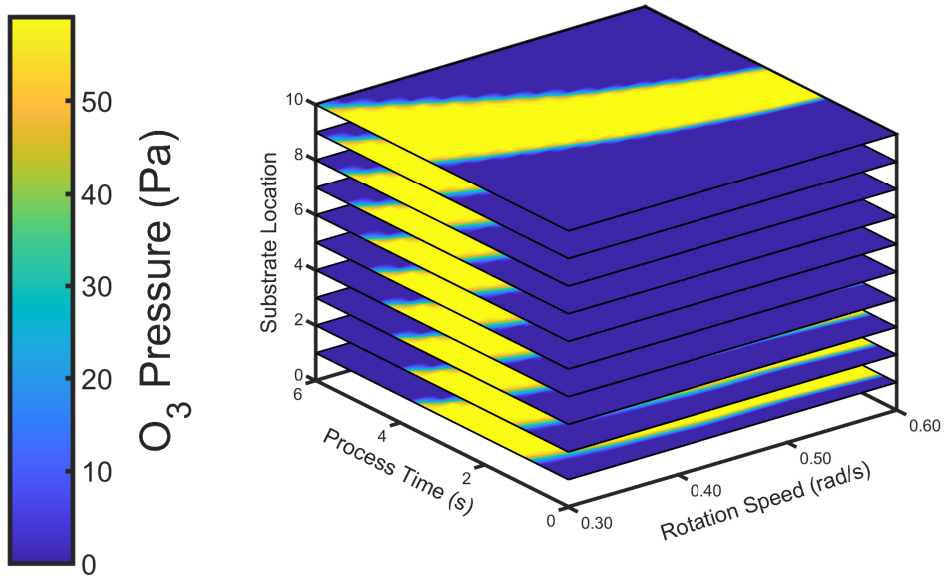
Additionally, with the integration of the FNN model, the multiscale model outlined in Fig. 4.2 substitutes FNN in place of CFD while retaining the conjunction to the kMC script, which is still presented in the C language while removing user-defined function macros from Ansys Fluent's programming language in Fig. 4.7. From the modified multiscale model, simulation time is dramatically reduced from timescales of days to minutes by employing TensorFlow and the Intel oneAPI ICX C compiler toolkit, which is conveyed in Fig. 4.8. The efficiency of the modified multiscale model with FNN integration is sufficient to progress to control work, which requires abundant data generation to produce an accurate input-output model that is described in Section 4.3.

4.3 R2R modeling of the SISO process

To develop a paradigm for the control of AS-ALD, the general procedures conducted in *in vitro* environments is first examined to provide insight of the controller mechanism. Area-selective atomic layer deposition (AS-ALD) is a cyclical process composed of a three-step, ABC, proce-

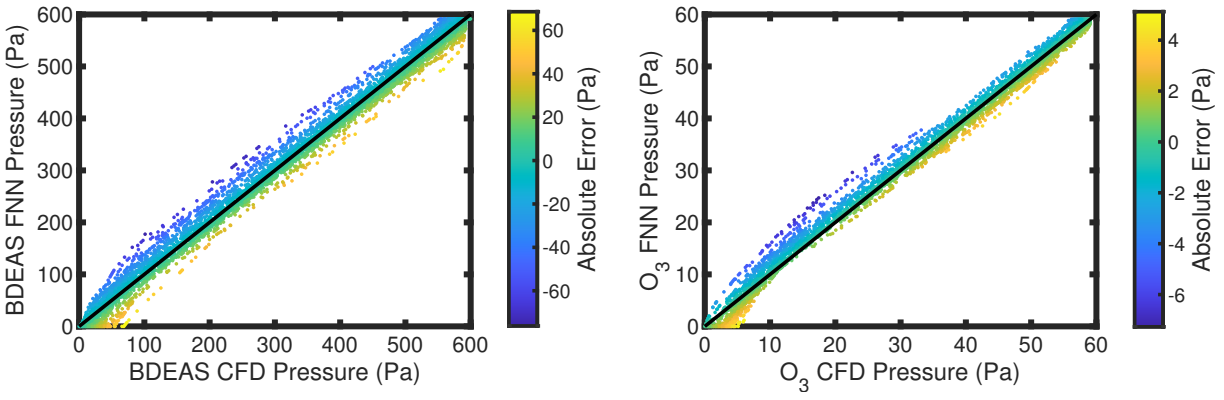


(a)



(b)

Figure 4.5: Contours of predicted pressure field data for (a) BDEAS and (b) O_3 generated from the FNN model for inputs of rotations speed, process time, and substrate location (identified by the partitioned wafer in Fig. 4.3). The FNN models for BDEAS and O_3 have MSE values of 0.0143 and 0.0121, respectively.



(a) (b)

Figure 4.6: Comparison of surface pressure data produced from the FNN predicted model and the multiscale CFD model for (a) BDEAS and (b) O_3 to illustrate the absolute errors from a sample of the total points used to generate the FNN model.

dure composed of inhibition, chemical adsorption, and oxidation sub-cycles that occur in short intervals. Due to these short intervals, which occur in the range of 1 to 5 seconds [63, 74], *in situ* monitoring cannot be performed due to time delays associated with the controller. Thus, *ex situ* statistical process control (SPC) methods such as run-to-run (R2R) control have been employed in the semiconductor manufacturing industry to regulate operating conditions by employing offline adjustments to the input parameter(s) [82] after completion of a batch run or in this work, a cycle of AS-ALD, to return the process to the setpoint. This work will study the impact of these disturbances on the growth per cycle (GPC) on the growth area of the substrate. Thus, Step A of the ABC AS-ALD process is not considered for this component of the R2R control framework.

4.3.1 Run-to-run controller framework

The R2R controller action is conducted by receiving measured output data that is produced from the multiscale model and portrayed in Fig. 4.9. The multiscale simulation is provided the substrate rotation speed as an input variable to evaluate surface pressure data for BDEAS and O_3. Additionally, constant shift disturbances are applied to the FNN or kMC components of the multiscale model in the form of a multiplicative factor to deviate the output, growth per cycle

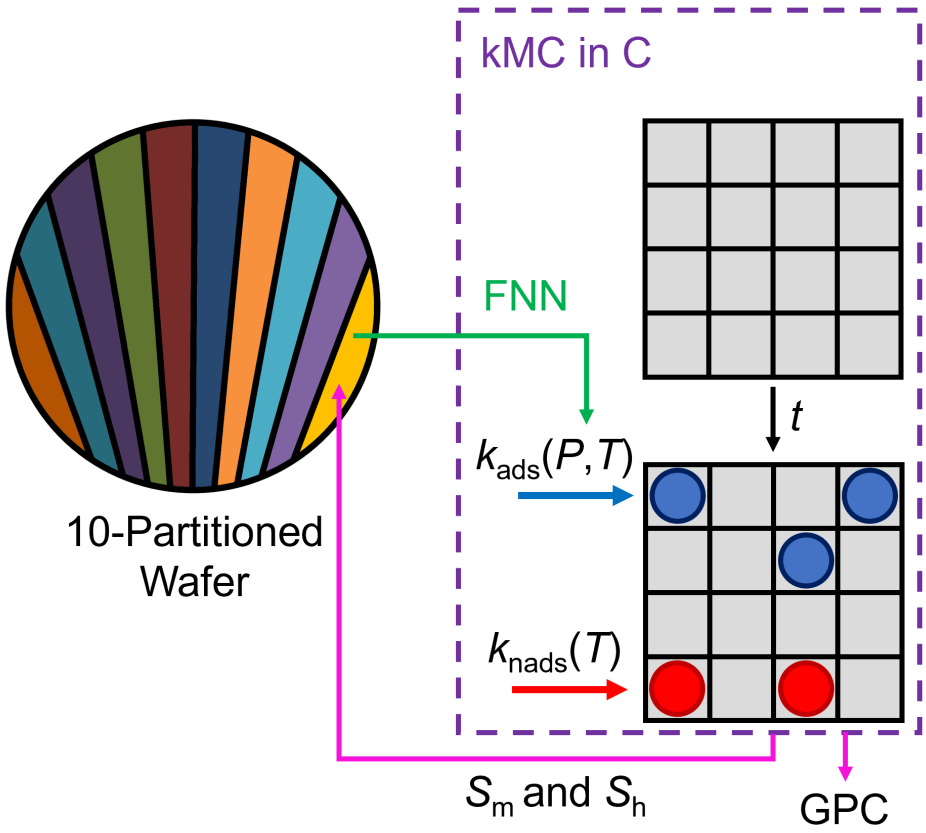


Figure 4.7: The modified multiscale process diagram that substitutes the FNN model for pressure data generation in place of CFD. Additionally, the kMC code is executed in C programming language without the aid of UDFs in Ansys Fluent.

(GPC), from the setpoint. In industrial applications, a Quartz Crystal Microbalance (QCM) [105] is used to measure the GPC by taking the reactor system offline. The resulting GPC is then sent to the R2R controller where an exponentially weighted moving average method is performed to calculate an updated rotation speed for the subsequent batch run, where the formulation is discussed in greater detail in Section 4.3.2.

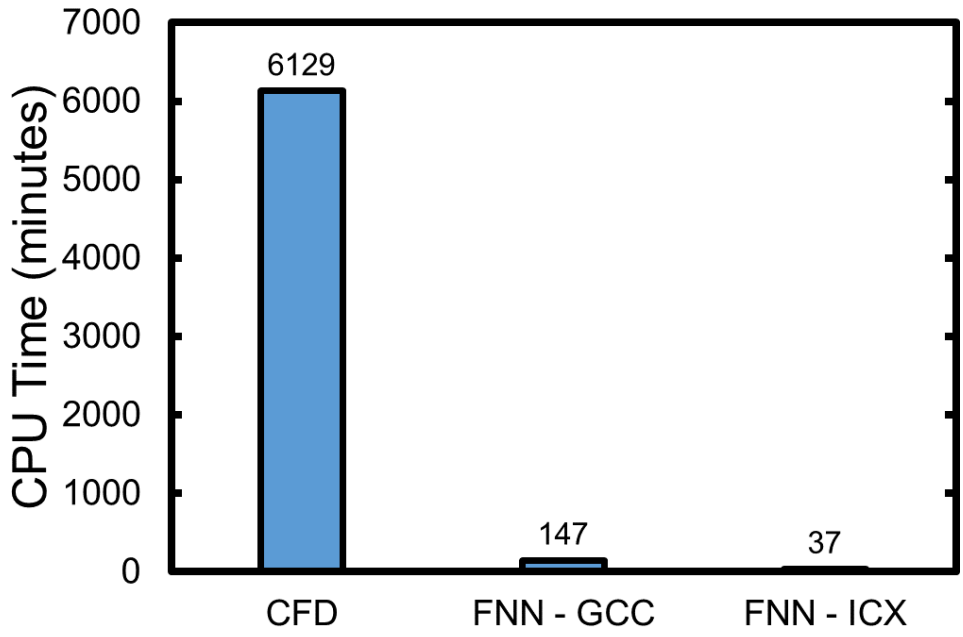


Figure 4.8: Comparison of simulation times with the conventional multiscale CFD simulation to the modified multiscale model with FNN integration using various C compilers, including GNU's GCC and Intel oneAPI's ICX. The CPU time is reduced from timescales of days to minutes through the integration of the FNN model and optimizer tools in the Intel oneAPI toolkits.

4.3.2 Exponentially weighted moving average approach to run-to-run control

Each R2R controller uses a statistical algorithm such as the exponentially weighted moving average (EWMA), the double exponentially weighted moving average (dEWMA), and the autoregressive moving averages (ARMA), to perform controller correction based on the error of the measured output parameter from the target or setpoint. This work will employ the EWMA method to implement controller action due to its capability of detecting and mitigating small-magnitude shift disturbances [10].

The EWMA method relies on a linear regression model of multiscale data of the output variable, growth per cycle (GPC) on the growth area of the substrate, for a range of the input variable, substrate rotation speed, which are obtained offline [27]. The linear regression model is defined as

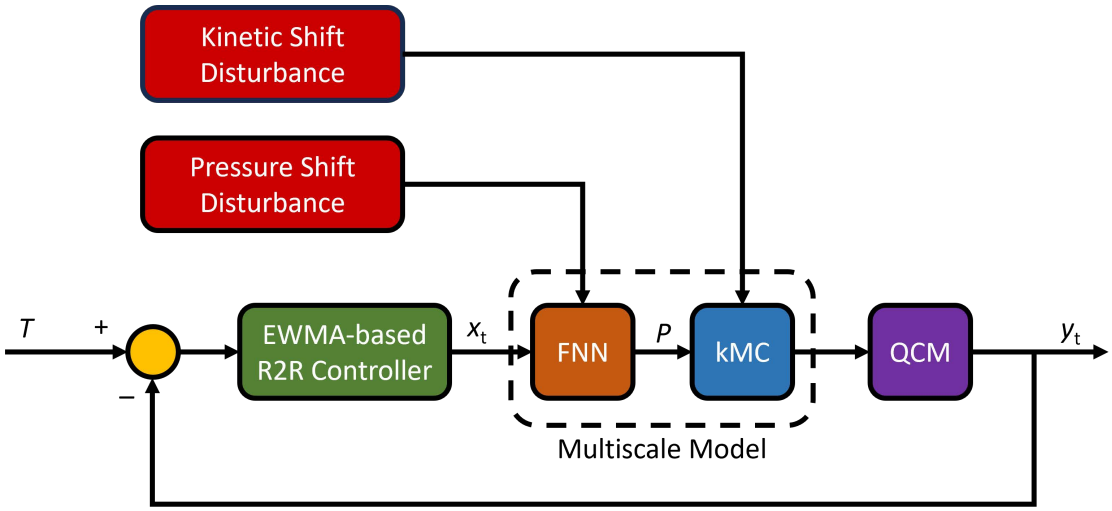


Figure 4.9: Process flow diagram depicting the conjunction of the R2R controller to the multiscale model. The addition of shift disturbances are introduced to the multiscale model in the form of pressure or kinetic perturbations. From industrial perspectives, the deposition process is stopped to measure the output (y_t) growth per cycle offline via a Quartz Crystal Microbalance (QCM). The error generated from the deviation from the setpoint, T is applied to the R2R controller that uses the EWMA method to evaluate an input (x_t), rotation speed, for the subsequent run.

follows:

$$\hat{y} = \alpha + \beta u \tag{4.7}$$

where $\hat{y}$ is the estimated deposition rate evaluated by the linear regression model, α is the bias or intercept, β is the process gain or slope, and u is the input variable. The linear model produced from open-loop data produced from the multiscale simulation of the output, GPC, and input, rotation speed, is presented in Fig. 4.10. The controller-adjusted input variable, μ_t will then be calculated based on the deviation of the output variable from the setpoint, T.

$$\mu_t = \frac{T - a_t}{b} \tag{4.8}$$

where $b = \beta$ and a_t is the adjusted bias at batch run t where $\alpha = a_0$, which is evaluated through an exponentially weighted moving average. A predicted intercept is evaluated by recursively sum-

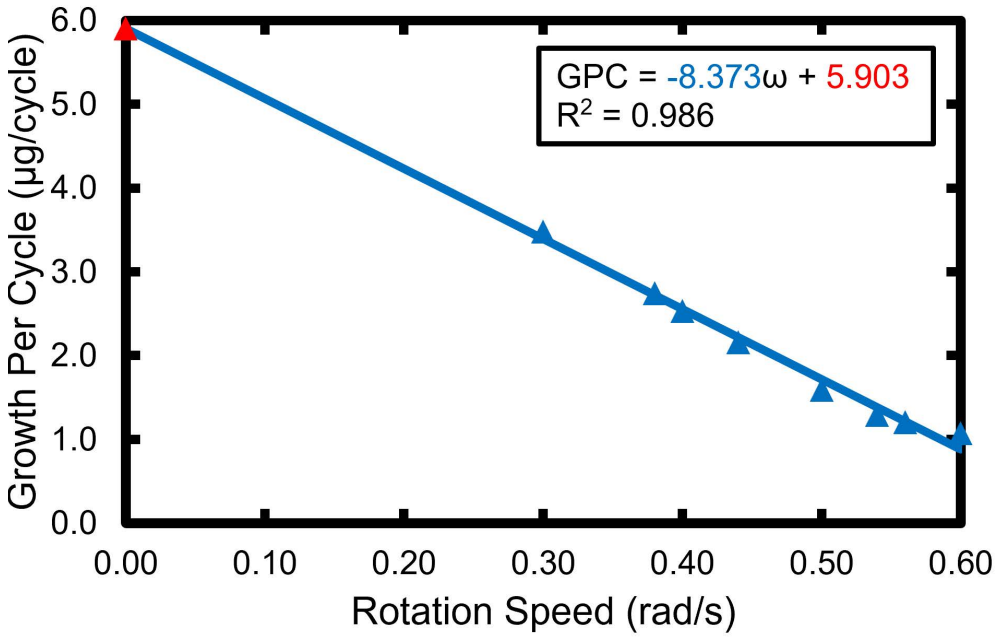


Figure 4.10: The linear model of the output, GPC, and input, rotation speed, from offline multiscale data for evaluating the process gain, β, and bias, α.

ming the errors from each batch run [82], and is defined by the following expression:

$$a_t = \lambda(y_t - bu_{t-1}) + (1 - \lambda)a_{t-1} \tag{4.9}$$

where y_t is the measured GPC evaluated from the multiscale modeling simulation and $\lambda \in (0, 1]$ is a self-determined weighting parameter that is used to preferentially balance the impact of older data on newer data [67]. λ also has a significant role in controlling oscillatory behavior, which is analogous to the role that integral action conducts in proportional-integral (PI) controllers. Thus, lower λ is suggested for the detection of small-magnitude shift disturbances, as recommended by Montgomery [79] and Lucas and Saccucci [71]. An important characteristic of performing an EWMA weight of the intercept of the linear model is the ability to tune modeling data to implement controller action by assuming a constant slope [27]. Such methodology is established by the assumption that disturbances will generally affect all elements of the data-fitted model,

thereby translating all points in the model in the form of a tuning strategy. The adoption of this methodology enables all model-fitted open-loop data to impact the controller update to the input parameters without requiring the generation of a new model for each observation of a disturbance, which may require an abundance of data generation.

4.3.3 Limitations of the EWMA-based R2R controller

Arguably, the EWMA approach has some limitations that are notable for this work. For instance, the EWMA method will utilize a constant λ, which must be determined through extensive experimental studies and is dependent on the magnitude of the shift disturbances [79]. For example, if a large-magnitude error is produced from a shift disturbance with a low λ, additional batch runs would be required for the controller to implement sufficient controller action to minimize the offset. In realistic settings, the magnitude of the shift disturbance is unpredictable, which is why this work focuses on the integration of a constant and minor shift disturbance only. This particular EWMA method also neglects the potential for noise and drift disturbances, which are frequent issues often encountered in industrial applications for semiconductor processing and are attributed to sudden changes in the operation of the reactor [26]. Additionally, this EWMA approach assumes that complete feedback control is implemented for each batch run, thus the reactor will be required to be taken offline after each batch run to measure the output parameter using a Quartz Crystal Microbalance to measure the GPC. This procedure may be impractical for realistic control and is inefficient for achieving high-throughput semiconductor generation, but serves as a starting point towards controller models that implement controller action using an open-loop approach, such as the previously proposed self-tuning model [26]. It is notable that self-tuning strategies using ARMA models are beneficial for mitigating disturbances for drift and noise, but are ineffective for bringing process stability toward the setpoint but rather maintains the process within standard deviations from the setpoint [112].

4.3.4 Compensation of shift disturbances

This work considers the effects of two shift disturbances that are intended to deviate the reactor operation from idealistic conditions. One of these shift disturbances includes a pressure shift disturbance that negatively or positively impacts the GPC of the system by including a constant shift variable intended to decrease or increase the magnitude of the pressure field data in the FNN model. Recall that the GPC is a parameter that is evaluated from the surface coverage computation from the kMC simulation. The reaction rate constants for nonadsorption and adsorption reactions are computed from the temperature-dependent Arrhenius model and the temperature- and pressure-dependent Collision Theory model, respectively. The integration of a pressure and kinetic shift disturbance is intended to increase or decrease the magnitude of the reaction rate constants for the adsorption and nonadsorption reactions, thereby increasing or decreasing the computed surface coverage. For the pressure disturbance, a coefficient is introduced to the machine learning model to increase or decrease the pressure field data while for the kinetic disturbance, a coefficient is introduced to the computation of the total reaction rate constant, k_{tot}, which affects the time advanced computation in Eq. (3c). It is notable that the kinetic shift disturbance is intended to perturb all reaction rate constants (adsorption and nonadsorption), by introducing a constant shift variable intended to decrease or increase their magnitude. The introduction of the pressure shift disturbance reflects the changes in the operating conditions such as that in the input, and kinetic disturbances resemble potential shift disturbances that may be unaccounted for during the AS-ALD process such as equipment failure.

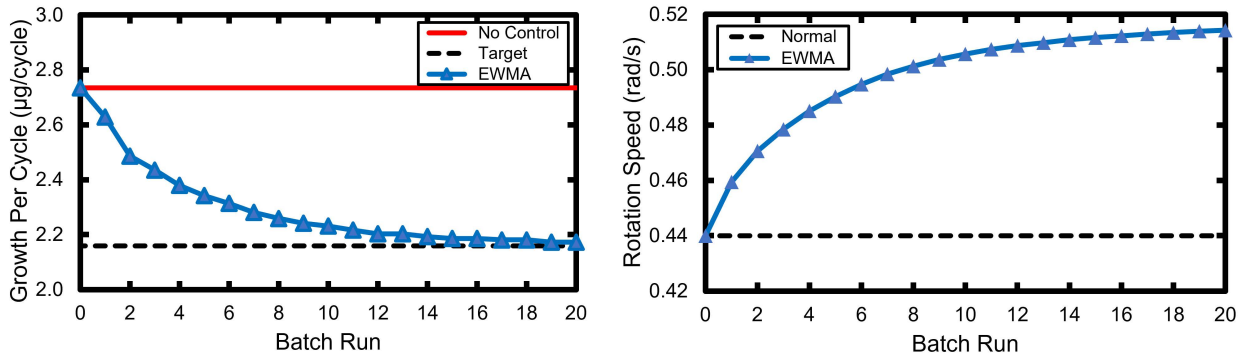
4.4 Closed-loop simulation results

The objective of the R2R controller is to manipulate the rotation speed of the wafer, which congruently affects the residence time of the substrate within the reaction zone. The residence time is a direct indicator of how much reagent is deposited onto the surface of the substrate, where GPC

increases with increasing residence time. By subjecting the AS-ALD process with pressure and kinetic shift disturbances in the multiscale model, the R2R controller must minimize the error from the setpoint by adjusting the rotation speed. This section will examine the response of the R2R controller to various disturbances that positively and negatively affect the GPC on the substrate.

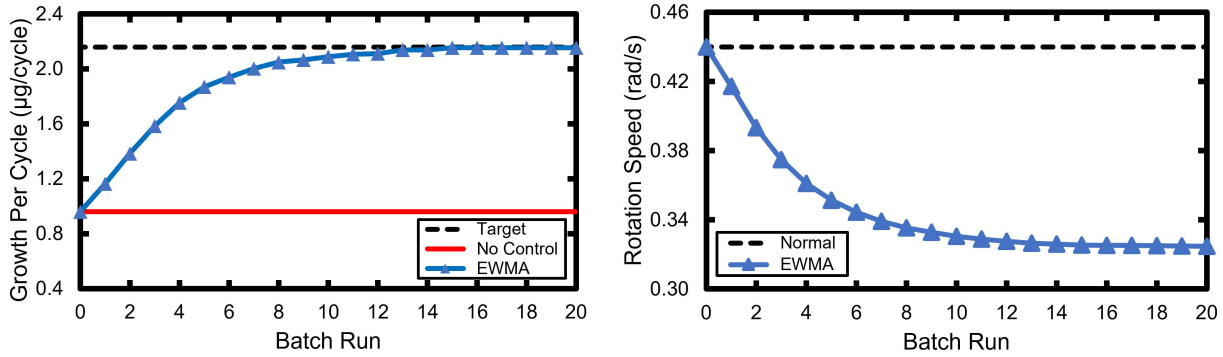
4.4.1 Closed-loop response to pressure disturbances

Pressure disturbances are oftentimes encountered in semiconductor fabrication due to potential changes in the fluid dynamics that are attributed to byproduct generation or changes to the inlet conditions. The pressure has a profound impact on the ability for adsorbates like BDEAS and O_3 to interact with the substrate surface via Collision Theory. Positive shift disturbances on the BDEAS and O_3 pressures were introduced to the FNN model for factors of 0.2 and 0.1, respectively, and their response is illustrated in Fig. 4.11 using an EWMA weight of $\lambda = 0.2$. Findings illustrate that the EWMA approach successfully overcomes the effects of the disturbance within 19 batch runs in Fig. 4.11a by increasing the rotation speed of the wafer depicted in Fig. 4.11b in order to reduce the residence of the wafer in the reaction zone. The results decrease the GPC to the desired setpoint. However, the result suggests that a substantial initial response is needed for the controller to mitigate the effects of the disturbance by requiring lesser batch runs to reach the setpoint. In addition to positive shift disturbances on the BDEAS and O_3 pressures, negative shift disturbances were introduced to the FNN model for pressure generation using factors of 0.2 and 0.1, respectively. The results of the R2R controller action are presented in Fig. 4.12 using an EWMA weight of $\lambda = 0.2$. It is observed that the R2R controller detects the minor shift disturbances induced to the GPC by decreasing the rotation speed in Fig. 4.12b thereby increasing the residence time of the wafer in the reaction zone. The result is depicted by a steady increase in the GPC in Fig. 4.13a where the setpoint is reached at 14 batch runs.



(a) (b)

Figure 4.11: R2R controller action using $\lambda = 0.2$ when perturbing the pressure fields for BDEAS and O_3 with a positive pressure disturbance by factors of 0.2 and 0.1, respectively. The GPC reaches the setpoint at 19 batch runs shown in (a) and the controller action increases the rotation speed illustrated in (b) to reduce the substrate residence time in the reactor. The rate of the rotation speed per batch run also decreases to minimize potential oscillations in the GPC output.



(a) (b)

Figure 4.12: R2R controller action using $\lambda = 0.2$ when perturbing the pressure fields for BDEAS and O_3 with a negative pressure disturbance by factors of 0.2 and 0.1, respectively. The GPC reaches the setpoint at 14 batch runs shown in (a) and the controller action reduces the rotation speed illustrated in (b) to reduce the substrate residence time in the reactor. The rate of the rotation speed per batch run also increases to minimize potential oscillations in the GPC output.

4.4.2 Closed-loop response to kinetic disturbances

Kinetic disturbance factors were applied to the kMC simulations for Steps B and C of the AS-ALD process to represent the perturbation of unknown disturbances that are applied to all reactions

(adsorption and nonadsorption). Fig. 4.13 illustrates the R2R controller response when subjecting the kMC simulations for Steps B and C to a constant kinetic disturbance factor that increases the GPC from the desired setpoint. The R2R controller uses an EWMA weight of $\lambda = 0.2$, which indicates that the setpoint is reached at 14 batch runs in Fig. 4.13a by increasing the rotation speed presented in Fig. 4.13b to compensate for the excess deposition onto the growth area of the substrate. In addition to subjecting the reaction rate constants to positive shift disturbances,

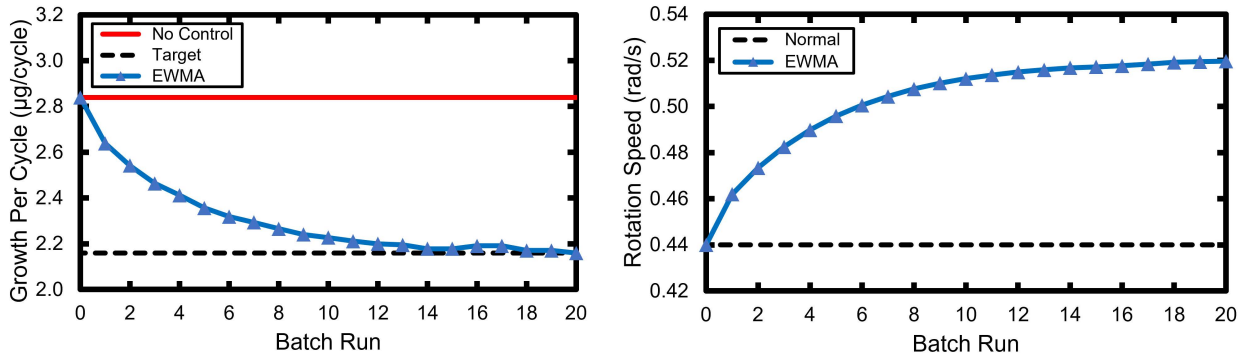
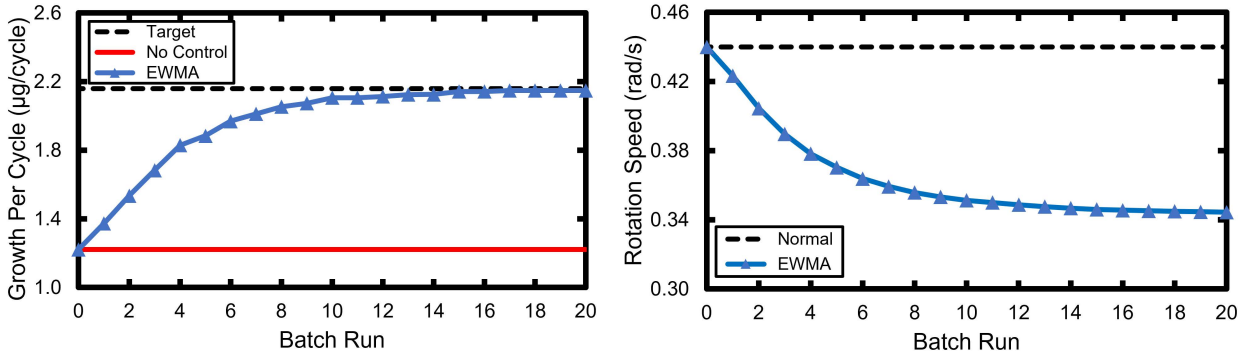


Figure 4.13: R2R controller action using $\lambda = 0.2$ when perturbing the reaction rate constants for Steps B and C with a positive kinetic disturbance by factors of 0.1. The GPC reaches the setpoint at 15 batch runs shown in (a) and the controller action increases the rotation speed illustrated in (b) to reduce the substrate residence time in the reactor. The rate of the rotation speed per batch run also decreases to minimize potential oscillations in the GPC output.

negative shift disturbances were also applied separately to Steps B and C kMC simulations, with results pictured in Fig. 4.14. An R2R controller with an EWMA weight of $\lambda = 0.2$ increased the residence time of the substrate in the reaction zone by decreasing the rotation speed of the wafer depicted in Fig. 4.14b. The result of the adjustment increases the GPC to the setpoint after 16 batch runs, which is shown in Fig. 4.14a.



(a) (b)

Figure 4.14: R2R controller action using $\lambda = 0.2$ when perturbing the reaction rate constants for Steps B and C with a negative kinetic disturbance by factors of 0.1. The GPC reaches the setpoint at 16 batch runs shown in (a) and the controller action decreases the rotation speed illustrated in (b) to increase the substrate residence time in the reactor. The rate of the rotation speed per batch run also increases to minimize potential oscillations in the GPC output.

4.5 Additional Remarks

In the case of offline control, the R2R controller is capable of detecting and correcting minor shifts assuming that the impact of noise and additional shifts are negligible. This assumption is oftentimes not practical for industrial application, due to the stochastic nature of product quality. Thus, it is necessary to consider the role of online feedback control for future work. A combination of online feedback and R2R control, as illustrated in Fig. 4.15 provides an alternative solution for the detection and correction of additional classifications of perturbations to the process. This type of control system [114] is examined in the subsequent section.

4.6 Conclusion

This work developed a run-to-run controller with an exponentially weighted moving average approach to overcome the effects of various pressure and kinetic shift disturbances for an area-selective atomic layer deposition rotary reactor, which are often observed in industrial applications

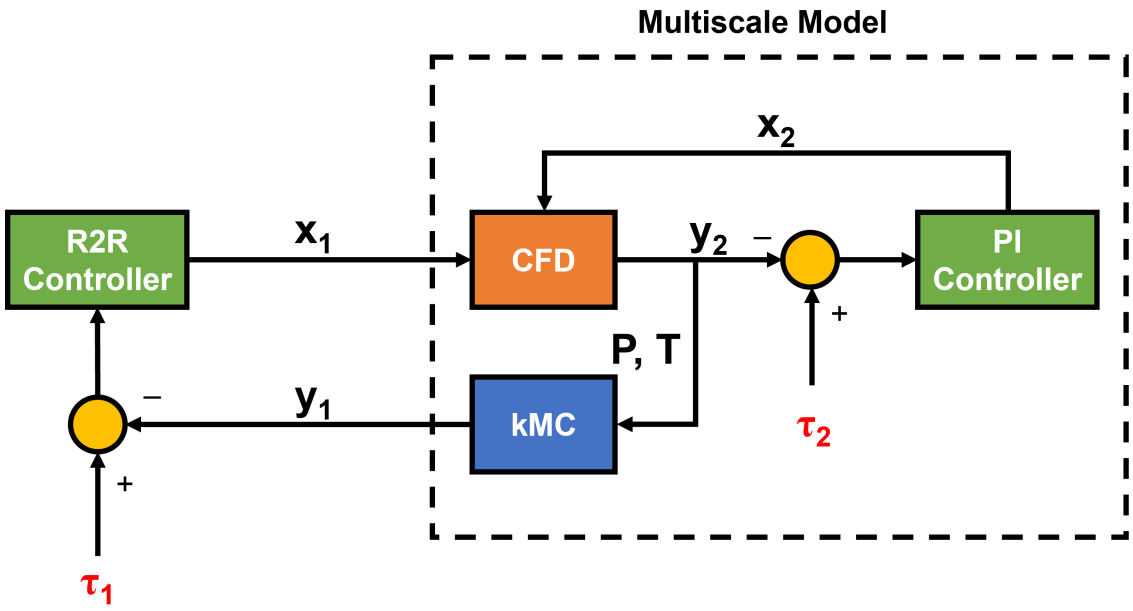


Figure 4.15: A combined online feedback and run-to-run controller that is conjoined to the multiscale model.

for semiconductor manufacturing. To facilitate the process to generate data, this work developed an *in silico* multiscale model that integrates machine learning through a feedforward neural network to generate surface pressure data that is conjoined to a kinetic Monte Carlo simulation written in the C programming language that evaluates the growth per cycle on the growth areas of a semiconductor substrate. An advantageous product of machine learning and C programming language integration for multiscale modeling is the substantial reduction in computation time from timescales of days to minutes. The run-to-run controller implemented input adjustment to the rotation speed of the rotary reactor to regulate the residence time of the substrate within the reaction zones, thereby allowing the controller to bring the process back to the setpoint within a minimum of 14 batch runs.

Chapter 5

Integrating Run-to-Run Control with Feedback Control for a Spatial Atomic Layer Etching Reactor

Unlike bottom-up fabrication approaches such as atomic layer deposition (ALD), atomic layer etching (ALE) is a reversal of ALD, which enables downscaling of the thicknesses of transistors below 10 nm and, under ideal operating conditions, improves surface uniformity, an essential characteristic for transistor alignment [45]. While there are various types of ALE (e.g., plasma and thermal), this work focuses on thermal ALE, which comprises a two-step cyclical process that results in the removal of a monolayer of surface material. However, to study this process in a laboratory setting presents a challenge in developing quantitative, first-principles models that allow the scale-up of thermal ALE processes that are applicable for industrial applications. For instance, ALE requires numerous cycles of etching to produce the finished product, which is a time-consuming task [19] that generates limited data to produce quantitative relationships between the etching rate and operating parameters such as temperature, reagent concentrations, and injection times. Additionally, it is arguable that characterizing the processes with laboratory data is an

expensive task that requires numerous cost and materials constraints. Thus, *in silico* multiscale modeling [65] is a beneficial route toward the scale-up of thermal ALE processes by conjoining microscopic surface kinetics and macroscopic fluid dynamics simulations in a thermal ALE reactor that resemble laboratory results. Through this multiscale simulation, an input-output model between the operating conditions (e.g., reagent flow rates and substrate velocity) and the output (e.g, etching per cycle, EPC) can be established without requiring laboratory experiments that are expensive, wasteful, and time-consuming.

This work adopts a prior multiscale model for a two-dimensional (2D) and spatial, sheet-to-sheet (S2S) thermal ALE reactor [124] for Al_2O_3 films, that was developed to increase process productivity. While this spatial reactor model is a first step toward integration to industrial applications, an essential control system is necessary to maintain process operation and product conformation attributed to equipment aging and changes in operation [82]. To maintain control for discrete processing cycles, a batch-to-batch or run-to-run (R2R) control system is beneficial for implementing control action by accounting for the measured EPC after the completion of each thermal ALE processing cycle. For instance, prior work [125] proposed a multivariable R2R control scheme to mitigate shift disturbances attributed to reagent pressure losses and reductions in adsorbate surface coverage for an inclined plate ALE reactor by manipulating the reagent injection time and flow rate. This work proposes a R2R controller that mitigates marginal kinetic shift disturbances that account for unexplainable perturbations that are oftentimes encountered during the reactor operation. Such control action is performed after each subsequent batch run for manipulated variables that exhibit unstable behavior. However, industry has encountered numerous challenges with real-time monitoring of the process due to time constraints and difficulties of predicting the cycle time of the process [30]. For time-variant manipulated variables such as the flow rate, research has centered on integrating feedback control for on-line measurement, particularly for reagent delivery flow rates that substantially influence the reactor operating pressure [121]. Thus, this work incorporates an on-line feedback controller to mitigate observable ramp disturbances that continuously

change with time (e.g., surface pressures) by implementing proportional-integral (PI) control by adjusting the inlet mass flow rates.

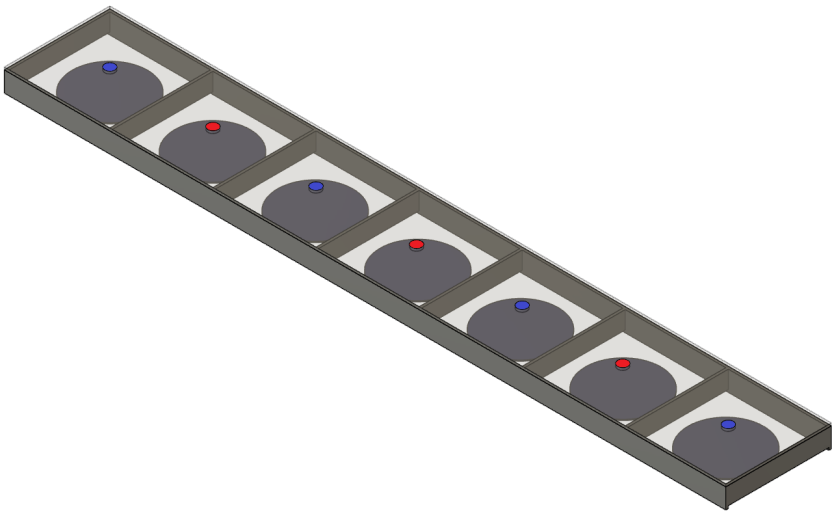


Figure 5.1: A three-dimensional depiction of a spatial sheet-to-sheet reactor comprising various zones for TMA/HF exposure and purging.

This work is organized as follows: Section 5.1 summarizes the multiscale modeling framework for the spatial S2S reactor for the thermal ALE of Al_2O_3 films comprising the microscopic kinetic Monte Carlo simulation in Section 5.1.1 and the macroscopic computational fluid dynamics model in Section 5.1.2 and Section 5.2 examines the development of the R2R (Section 5.2.1) and feedback (Section 5.2.2) control system for the S2S reactor.

5.1 Multiscale Modeling

The thermal atomic layer etching of Al_2O_3 comprises a two-stage, AB process with cut-in purging steps that are spatially separated in a sheet-to-sheet reactor model. To replicate the dynamic behavior of the reactor, an *in silico* multiscale computational fluid dynamics model is proposed that simulates the microscopic surface kinetics occurring on the wafer surface and the macroscopic fluid dynamics in the ambient gas-phase regions of the reactor. This multiscale modeling approach

allows the thermal ALE process to be described in various time and length scales [14] through the use of stochastic kinetic Monte Carlo simulations and computational fluid dynamics. This section will summarize the kinetic Monte Carlo algorithm [123] and the multiscale model [124] drawn from prior works.

5.1.1 Microscopic Modeling

Rudimentary thermal ALE processes are characterized by a general two-step process [34] comprising an initial modification step through the use of a bulky and gaseous precursor that adsorbs to the surface in a self-limiting manner that is succeeded by an etch step with a secondary reagent that removes the modified and volatile surface in a self-limiting manner at high operating temperatures. Situated between these two steps are purging stages to ensure that self-limiting behavior is maintained. For the thermal ALE of Al_2O_3, the proposed reaction mechanism, which assumes an elementary rate law where all reactions are bimolecular, and processing times are obtained from experimental work in [37]. The thermal ALE of Al_2O_3 uses trimethylaluminum (TMA) as the initial precursor reagent and hydrogen fluoride (HF) as the secondary etching reagent, while the spatial S2S reactor model utilizes nitrogen gas (N_2) as the purging material.

The computation of pressure and temperature dependent reaction rate constants are necessary to exemplify the surface kinetics. Thus, the integration of Collision Theory and the Arrhenius model are practical for evaluating reaction rate constants for adsorption, k_{ads}, and nonadsorption reactions, k_{nonads}. However, the Arrhenius model depends on two constants, the activation energy and the pre-exponential factor, that are typically unavailable in literature data, whereas Collision Theory relies on a sticking coefficient factor that is attainable in literature findings. Thus, the use of *ab initio* molecular dynamics simulations is practical for optimizing molecular structures and determining ground-state energy configurations by minimizing electronic energies through Density Functional Theory and applying the Nudged Elastic Band method, where the activation energies are calculated through the open-source software, Quantum ESPRESSO. It is notable that

the software was compiled locally using Intel-based Fortran and C/C++ compilers as part of the Intel oneAPI toolkits to enhance the computation speed and parallelization of multiprocess simulations. Additionally, phonon computations through Density Functional Theory and the Quasi-harmonic Approximation are employed in Quantum ESPRESSO to evaluate thermophysical data to define intensive variables (e.g., specific heat, standard entropy, standard enthalpy) for the macroscopic computational fluid dynamics simulation discussed in Section 5.1.2. To calculate the pre-exponential factor, Transition State Theory is applied, which assumes a negligible dependence on the partition functions for the transition state and reactant [47].

Due to the difficulties of determining the exact configuration, the number of reactions occurring, and the types of reactions manifesting at any given instance of time and location, a microscopic model is necessary to reflect the random nature of realistic ALE reactions. Particularly, a kinetic Monte Carlo (kMC) method is appropriate for this work as it considers the probabilities of the aforementioned mutually exclusive events that describe the configuration of the substrate surface at any location and time [20]. Past work [123] employed a kMC model in the Python programming language, which originated from Bortz, Kalos, and Lebowitz. The BKL method assumes that all potential reactions lie within a Poisson distribution, and it selects a particular reaction through a randomly generated number, and then calculates a time advancement with a secondary random number [8]. The procedure for the BKL approach can be simplified as follows:

(1) An $N \times N$ grid comprising N^2 active sites is declared to the kMC simulation to reflect the initialized wafer prior to thermal ALE processing.

(2) The adsorption and nonadsorption reaction rate constants (k_{ads} and k_{nonads}, respectively) are

evaluated using Collision Theory and the Arrhenius equation, respectively:

$$k_{ads}(P_a, T) = \frac{\sigma_a P_a A_{site}}{Z_a \sqrt{2\pi m_a k_B T}} \tag{5.1}$$

$$k_{nonads}(T) = \frac{k_B T}{h} \exp\left(-\frac{E_A}{RT}\right) \tag{5.2}$$

where σ_a is the sticking coefficient for the adsorbate (e.g., TMA and HF), a, on the Al_2O_3 surface, P_a is the adsorbate surface pressure on the wafer, A_{site} represents the surface area of an Al_2O_3 binding site on the wafer, Z is the adsorbate coordination number, m_a is the atomic mass of the adsorbate, k_B is the Boltzmann constant, T is the surface temperature of the wafer, h is the Planck constant, E_A is the activation energy for the nonadsorbtion reaction, and R is the ideal gas constant.

(3) Next, a set of r possible reactions for the entire $N \times N$ grid is listed and denoted by index i. k_{tot} is then found by summing all of the possible reaction rate constants, k_i. This assumes that each reaction is mutually exclusive, i.e., they are independent events, to employ a Poisson distribution.

$$k_{tot} = \sum_{i=1}^{r} k_i \tag{5.3}$$

(4) A random number, $\Gamma_1 \in (0, 1]$ is selected to determine the reaction pathway; the reaction p that satisfies the following inequality is chosen:

$$\sum_{i=1}^{p-1} k_i \leq \Gamma_1 k_{tot} \leq \sum_{i=1}^{p} k_i \tag{5.4}$$

(5) Lastly, a time advancement, Δt, computation is performed using a secondary random number, $\Gamma_2 \in (0, 1]$ that reflects the time in which the reaction converts the initial state to the

final state.

$$\Delta t = -\frac{\ln \Gamma_2}{k_{tot}} \tag{5.5}$$

This kMC process is illustrated in Fig. 5.2, which shows the evolution of the sites in the $N \times N$ grid.

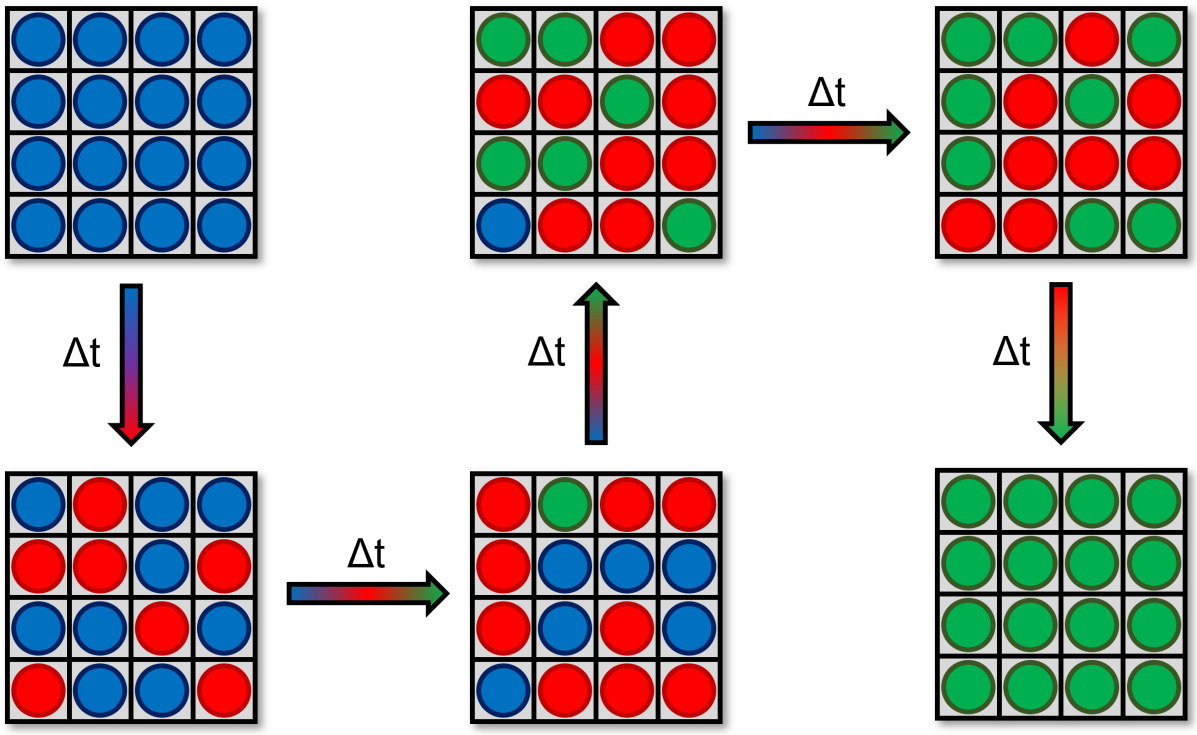


Figure 5.2: Graphical representation of the kMC algorithm when conducting the grid approach for the BKL method. Each grid advances to the next grid following the computation of the time update, Δt.

Following the development of the kMC model, processing times for achieving similar EPC and surface coverage were validated with experimental findings from [37]. Additionally, a predictive model for a multiple-input-single-output dataset was constructed through a feedforward neural network (FNN) to correlate input parameters, pressure and temperature, with the output parameter, processing time [123]. The results of these investigations were used to develop the standard operating conditions for the S2S reactor.

5.1.2 Macroscopic Modeling

A two-dimensional (2D) spatial, sheet-to-sheet (S2S), reactor was then developed from works [91] and [99] through the computer-aided design (CAD) modeling software, Ansys SpaceClaim, comprising TMA and HF injection regions that are spatially separated by adjacent N_2 purging zones Fig. 5.3. Following the construction of the reactor, a 2D dynamic mesh discretization procedure was conducted using finite elements with triangular geometry, and an optimized mesh with balanced mesh quality was obtained by integrating the remeshing and refinement tools in Ansys Workbench. To optimize the reactor model, multiple macroscopic computational fluid dynamics simulations were conducted through Ansys Fluent for various gap distances, i.e., the distance between the wafer and divider walls, to determine the distance that minimizes TMA and HF intermixing. Ultimately, it was found that a gap distance of 5 *mm* was suitable for the reactor design. Additionally, various reactor operating conditions including the TMA, HF, and N_2 flowrates and the substrate velocity were tested to determine appropriate conditions that maximized the etching per cycle (EPC) [124].

The numerical simulation is performed using a pressure-based coupled solver that simultaneously solves the mass and momentum equations to reduce computation clock time at a cost of requiring more random access memory (RAM) [3]. The mass, momentum, and energy equations are described as follows:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \overrightarrow{v}) = S_m \tag{5.6}$$

$$\frac{\partial}{\partial t} (\rho \overrightarrow{v}) + \rho (\overrightarrow{v} \cdot \nabla) \overrightarrow{v} = -\nabla P + \nabla \cdot (\overline{\overline{\tau}}) + \rho \overrightarrow{g} + \overrightarrow{F} \tag{5.7}$$

$$\frac{\partial}{\partial t} (\rho E) + \nabla (\overrightarrow{v} (\rho E + P)) = -\nabla \left(\Sigma h_j \overrightarrow{J}_j \right) + S_h \tag{5.8}$$

where ρ expresses the density of the gaseous species, $\overrightarrow{v}$ denotes the velocity of the gases, P is the system pressure, $\overline{\overline{\tau}}$ represents the rank-two stress tensor, $\overrightarrow{g}$ is the gravitational acceleration due to

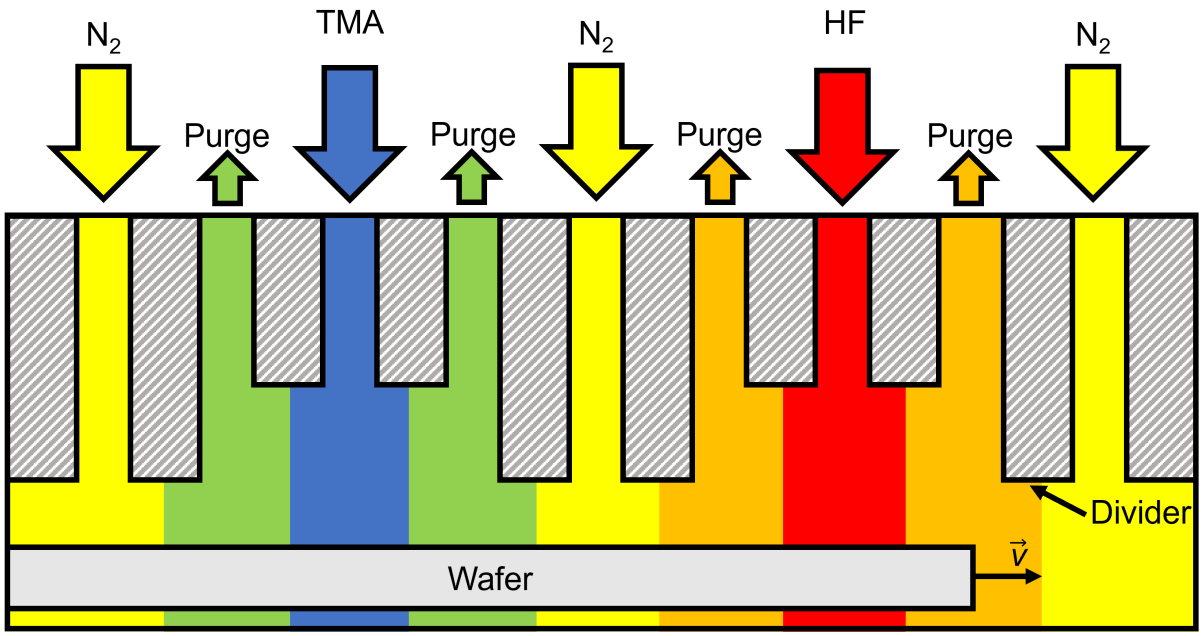


Figure 5.3: 2D side projection of the sheet-to-sheet spatial reactor for the thermal ALE of Al_2O_3. The illustration is adapted from [124].

Earth, $\overrightarrow{F}$ defines the body force subjected onto the gases, h_j and $\overrightarrow{J}_j$ are the sensible enthalpy and mass diffusion flux of species j, respectively, and S_m and S_h are source generation and consumption flux rates for the mass and energy equations, respectively. To reduce the complexity of the simulation, the reactor is assumed to operate under isothermal and isobaric conditions, which are made possible by the inclusion of temperature controllers that maintain temperature uniformity on the wafer surface and a vacuum pump to maintain the pressure within the reactor chamber. A table of standard operating conditions for the thermal ALE, S2S reactor is provided in Table 5.1.

Additionally, a parallelization procedure is employed that partitions the reactor mesh based on the number of compute cores available in the central processing unit (CPU). For this work, multiple compute nodes comprising 36 and 48 cores, and 384 GB and 512 GB of random access memory (RAM) were used. The numerical simulation also adopts a first-order implicit method to solve the transient transport equations using a timestep size of 0.001 s. To simulate the movement of the wafer through each zone of the reactor, a dynamic mesh procedure was integrated into the

Table 5.1: Standard operating conditions for the spatial, thermal ALE, sheet-to-sheet reactor.

Reactor Operating Condition	Value
Operating Pressure	300 *Pa*
Operating Temperature	573 *K*
Substrate Velocity	80 *mm/s*
HF Flow Rate	20 *sccm*
TMA Flow Rate	40 *sccm*

simulation by defining a constant substrate velocity. To ensure the quality of the mesh is maintained through each discrete movement, a smoothing and remeshing procedure with application default settings were specified [5].

5.1.3 Multiscale Modeling

The juncture of the microscopic and macroscopic simulations is a tedious process that requires cross-platform programming to enable the exchanging of output data from each simulation. In the previous work [124], the multiscale model adopted a Linux Bash Shell script to enable the transfer of data between the macroscopic model in Ansys Fluent and the mesoscopic kMC model in the Python programming language. However, for this work, the kMC model is directly integrated in Ansys Fluent through the C programming language in customizable user-defined functions (UDFs). This new multiscale integration scheme results in faster simulation times due to the simpler code architecture and retains all the accuracy of the method used in the previous work. The program executed for each simulation is summarized in the following steps and illustrated in Fig. 5.4:

(1) The CFD simulation in Ansys Fluent is executed through a Linux Bash script and runs for a processing time of Δt.

(2) Once Δt is reached, the CFD simulation records pressure and temperature data that is stored through a custom UDF.

(3) The CFD simulation is paused while the kMC simulation is executed in C-language inside of Ansys Fluent. It calculates the time advancement, EPC, and source generation and consumption flux rate terms. When the time advancement reaches Δt, the source generation and consumption flux terms are used to update the corresponding variables through UDFs.

(4) The CFD simulation is executed for the subsequent time advancement, and the cyclical loop continues until the wafer reaches the end of the reactor.

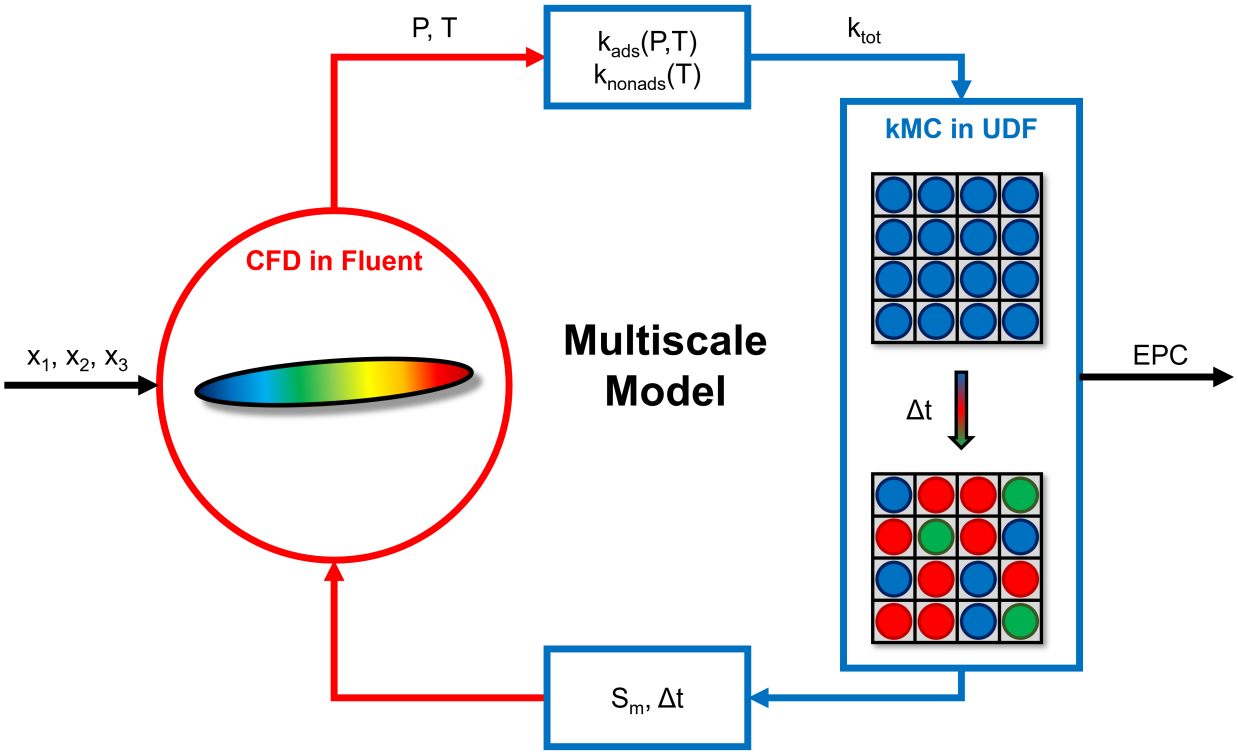


Figure 5.4: An illustration of the multiscale simulation that couples the CFD simulation and kMC simulation in Ansys Fluent.

Pertinent surface pressure data is extracted from nodal data located on the upper surface of the wafer, which is illustrated in Fig. 5.5. The pressure field contours illustrate the spatial isolation of

the TMA and HF reaction zones, which is made possible by the addition of adjacent N_2 injection and purging zones. The wafer, which is simulated with a constant velocity, is represented by a "floating" wall boundary to prevent the formation of irregular cell geometry as a consequence of remeshing procedures defined to the dynamic mesh.

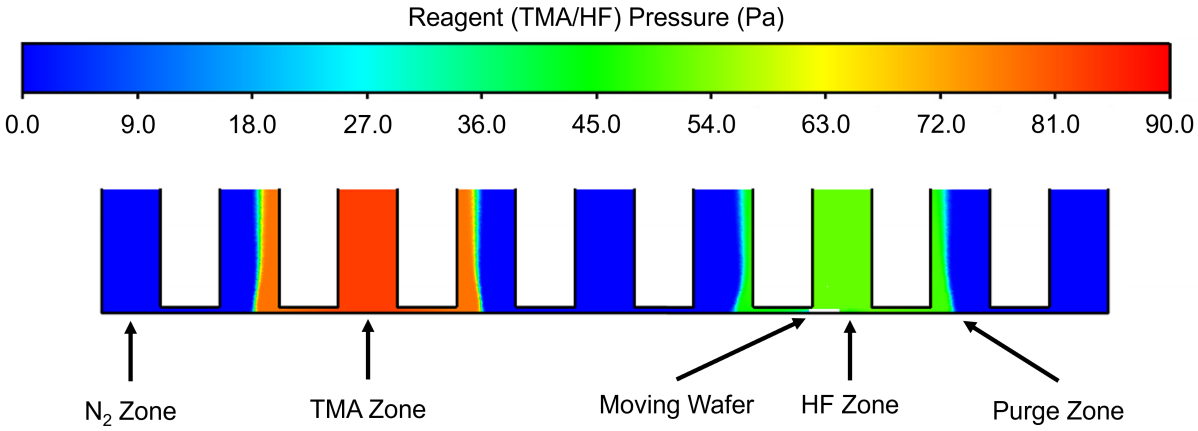


Figure 5.5: Surface pressure field data of the 2D S2S reactor at a time of 1.50 *s* produced from the multiscale simulation.

5.2 Process Control

While a multiscale model is beneficial for studying the optimal operating conditions desired to maximize wafer quality and productivity, these conditions generally encounter disturbances that disrupt the ideal behavior of the thermal ALE process. Disturbances can be classified into two bifurcations, shifts and drifts, that can dramatically change the process operation if undetected. Additionally, thermal ALE requires expensive reagents such as TMA, which are costly and toxic; thus, it is imperative to minimize the amount of unused reagent [70]. Therefore, the integration of a process control system is desired to regulate the thermal ALE operation by exploiting a multivariate input parameter correction procedure that is conducted in an optimal manner.

One major advantage of simulated models is that process control systems, which improve the robustness of the overall system, can be economical in both cost and time [23]. In this work,

two forms of process control systems are examined: an ex situ run-to-run (R2R) controller and an in-line proportional-integral (PI) controller. For multivariate processes with fast dynamics, the R2R controller is advantageous. The R2R controller implements ex situ or off-line control action after the completion of a thermal ALE batch run. This multivariate input correction is evaluated using an algorithm such as the exponentially weighted moving average (EWMA) of a linear model that relates the input parameters to the output. However, the R2R controller performs poorly at detecting process shifts that dramatically change the dynamics of the process during the batch run. Thus, PI control, which is conducted in-line, is practical for implementing continuous control action within the duration of the thermal ALE cycle as it consistently measures process data to adjust an input parameter. Such R2R and PI controllers require experimental and deterministic tuning to ensure that the process offset is minimized in a minimal number of batch runs.

5.2.1 Run-to-Run Controller

A run-to-run (R2R) controller is beneficial for semiconductor processing due to the short time intervals required to complete a single cycle of thermal ALE, and also for its capability to implement control actions for slow dynamic systems that require multivariate process control [9]. This form of control has been integrated in manufacturing execution systems (MES) by Critical Manufacturing to monitor the behavior of deposition, etching, and lithography processes in wafer fabrication [1]. A work [100] also studied various tuning approaches for R2R control with semiconductor processes that were purposefully disturbed in the form of a closed-loop tuning methodology. R2R controllers employ an ex situ form of process control in which control actions are performed after a batch run finishes, which is when a sensitive metrology device (e.g., Quartz Crystal Microbalance) measures the mass loss off-line in industrial practice. Following the aforementioned procedure, an EWMA algorithm can be employed to determine the control actions that modify the input parameters to overcome the effects of disturbances while also reducing the offset in minimal batch runs. This R2R control process is depicted by the process flow diagram in Fig. 5.6.

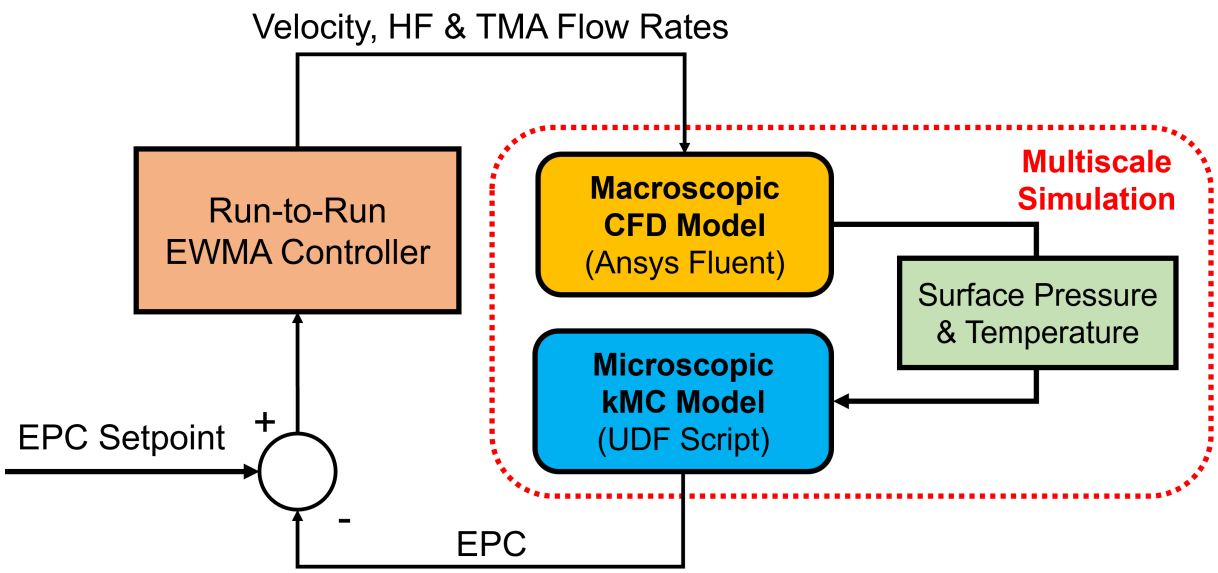


Figure 5.6: Process diagram that depicts the conjunction of the R2R controller with the multiscale simulation, where the R2R controller performs input adjustment to the macroscopic CFD model from the calculation of the error between the multiscale model and the target.

5.2.1.1 Linear model

Before the R2R controller is appropriately tuned, a multiple-input-single-output (MISO) linear regression model of off-line data obtained from the multiscale simulation. However, it is notable that the type of model is chosen based on the deterministic trend of the dataset [e.g., 115, 125]. This MISO model is adopted from a prior work [107] by assuming negligible Gaussian noise, and relates three manipulated input parameters, the substrate velocity, TMA flow rate, and HF flow rate, to the measured output parameter, etch per cycle (EPC).

$$\hat{y} = \mathbf{B}^T\mathbf{X} + a, \quad \text{where } \mathbf{B} \in \mathbb{R}^3,\ \mathbf{X} \in \mathbb{R}^3,\ a = \alpha + d \tag{5.9}$$

where $\hat{y}$ is the predicted EPC output, $\mathbf{B} = 10^{-3}\begin{bmatrix} 0.0121 & 0.346 & -1.84 \end{bmatrix}^T$ is a vector containing process gains or coefficients for each input parameter, $\mathbf{X} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T$ is the manipulated input vector comprising the HF flow rate (x_1), TMA flow rate (x_2), and the substrate velocity (x_3),

and a represents the corrected bias term of the linear model by accounting for the effects of the correction d from the bias term $\alpha = 0.478$ that has no observable disturbance. In this work, the updated bias term a is employed to distinguish from the bias α of the linear model generated from off-line data that is undisturbed. The role of a is vital for the calculation of the updated input variables through an exponentially weighted moving average of the bias term a, which is further elucidated in Section 5.2.1.2. To ensure the accuracy of the predicted model, the mean squared error (MSE) metric, which describes the averaged deviation from the estimated and experimental EPC, was determined to be 4.236×10^{-4}. The linear model described in Eq. (5.9) is used in conjunction with an EWMA method to perform manipulated input adjustment, which is elucidated in Section 5.2.1.2.

5.2.1.2 Exponentially weighted moving average

In order to implement control action, R2R controllers depend on an algorithm that accounts for the error generated from the deviation of the measured output from the setpoint or target. One challenge often encountered in industrial practices is the lack of the data generation to construct empirical models that can effectively gather deterministic trends with perturbed data sets [27]. Thus, continuous tuning of the linear regression model described by Eq. (5.9) is necessary to mitigate these disturbances, which can be accomplished through a translation procedure proposed in [68]. By assuming that the process gain, B, is independent of the disturbances, the process model is translated for a value d. This updated bias, a, is determined through an exponentially weighted moving average (EWMA) of the bias, α, to sum the errors following the completion of each batch cycle [79]. The EWMA method is described by the following expression:

$$a_t = \lambda(y_t - \mathbf{B}^T\mathbf{X}_{t-1}) + (1-\lambda)a_{t-1} \tag{5.10}$$

where a_t is the updated bias for the subsequent batch run, t, that depends on the previous bias, a_{t-1}, y_t represents the observed EPC evaluated from the multiscale CFD simulation, and λ is an exponential weight that is strictly determined from experimental research. An advantageous feature of the EWMA algorithm is that the recursion strategy reduces EPC offset by summing errors generated from historical data in the form of "integral action." However, the EWMA method requires that an optimal λ be chosen to minimize offset while requiring a minimal number of batch runs to meet this criteria. The subsequent adjustment to the input parameters, $\mathbf{X}_{t-1}$ is calculated by minimizing the sum of the least squares of all input parameters, which is described by the following minimization problem:

$$\min_{\mathbf{B}^T\mathbf{X}_t=c_t} \|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2 \tag{5.11}$$

$$\text{s.t. } c_t = \tau - a_t \tag{5.12}$$

where τ is the target or setpoint of the EPC and $\|\cdot\|$ denotes the l_2 norm. The optimization problem, as derived in prior work [107], which utilizes the partial derivatives of the Lagrange function to create the following formula that describes the computation of an optimal input, $\mathbf{X}$, for the subsequent batch run, t:

$$\mathbf{X}_t = \mathbf{X}_{t-1} - \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\left(\mathbf{B}^T\mathbf{X}_{t-1} - c_t\right) \tag{5.13}$$

5.2.2 Feedback Controller

Continuous process control is desired for processes that observe fast dynamics and have sensitive responses to perturbations. Feedback control is beneficial for regulating the dynamical behavior of the thermal ALE process in the S2S spatial reactor due to potential disturbances that may influence the standard operating pressures of the reactor. A limitation of R2R control is that adjustment is employed after the completion of an etching cycle, which results in a lack of pro-

cess monitoring while the etching process is conducted and can introduce nonconformal surface impurities that degrade transistor performance. Additionally, the R2R controller requires sensitive measuring apparatuses such as the Quartz Crystal Microbalance that requires off-line measuring to record the EPC, which limits the detection of disturbances that occur during the operation of the reactor (i.e., in situ monitoring). Therefore, the monitoring of another measurable parameter, the surface pressure, is practical for use in on-line feedback control and continuous pressure supervision is necessary to control the frequency of collisions between molecular species and on reactor walls that can negatively influence the behavior of the initial adsorption reactions for Steps A and B [46]. With respect to in-batch feedback control, it is important to note that alternative approaches like data-driven batch control techniques [12] may be used instead of the proportional-integral control schemes. Such model-based approaches may lead to achievement of additional performance requirements like reducing batch time.

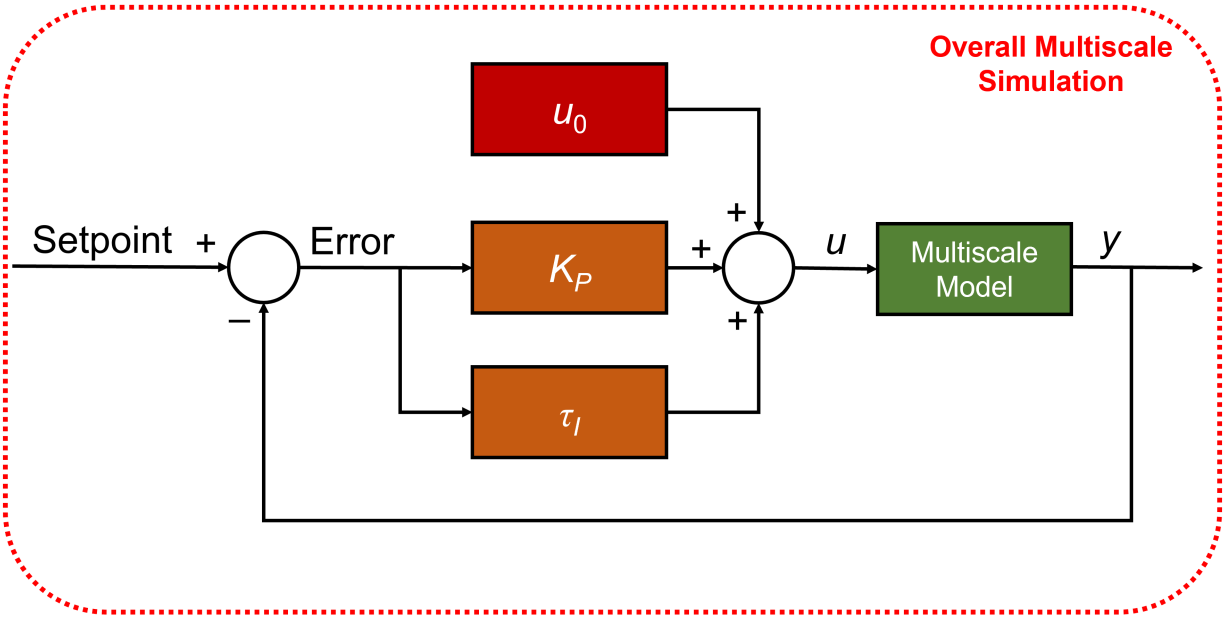


Figure 5.7: Process flow diagram depicting the conjoining of the PI controller with the multiscale model to implement correction to the TMA and HF flow rates, u, by accounting for the error between the measured wafer surface pressure y from the target surface pressure. The bias, u_0, is obtained from the R2R controller adjustment of the TMA and HF flow rates.

The Taiwan Semiconductor Manufacturing Company (TSMC) [110] employs a Micro Electro Mechanical System (MEMS) that measures the deformation of applied pressure in capacitance onto the surface of pressure sensing electrodes [18]. These pressure sensors exploit the piezoelectric effect that enables them to change in resistivity when subjected to pressure stresses [48]. This work considers the role of MEMS for monitoring the surface pressure on the substrate in the TMA and HF reaction zones and implements flow rate adjustments to account for perturbations in the surface pressure. To ensure that the MEMS is continually implemented in real-time, PI control is practical for monitoring and correcting the pressure disturbances. PI controllers continuously apply a control action based on the error between the measured pressure and the target pressure, as seen below:

$$u(t) = u_0 + K_0 \cdot K_p \left(e(t) + \frac{1}{\tau_I} \int_0^t e(\tau) d\tau \right) \tag{5.14}$$

where $u(t)$ is the mass flow rate of TMA and HF taken at time t, u_0 is a bias term that is evaluated from the R2R controller input for x_2 and x_3 in Eq. (5.13), K_0 is a conversion term to correlate mass flowrate and pressure, K_p is the proportional gain, $e(t)$ is the error measured by the system at time t, and τ_I is the integral time constant. K_p represents the proportional adjustment to the current error, and it drives the system towards the setpoint but stabilizes at an offset away from the setpoint. However, the $(K_0 K_p/\tau_I) \int_0^t e(\tau) d\tau$ term represents the integral adjustment of the overall error, and it drives the system from the offset to the setpoint. Thus, a well-tuned PI controller will be able to quickly drive the surface pressure of the wafer to the target pressure without any overshoot [23]. In addition, the PI controller operates in conjunction with the R2R controller, as the latter dictates the starting mass flowrates of TMA and HF while the former adjusts them in real time. In this manner, the PI pressure controller effectively maintains the desired partial pressures of the reagents on the wafer surface even in the presence of pressure disturbances.

5.2.3 Disturbances

In industrial practice, there are a variety of disturbances that may affect the thermal ALE process environment and result in deviations of the EPC from the setpoint. For instance, a disturbance in the wafer surface temperature can affect the EPC and the temperature uniformity of the surface. Additionally, perturbations in the operating pressure can be attributed to a failure in the vacuum pump, which is needed to remove excess reagent and byproducts from the reaction chamber, or changes in the reagent feed composition and flow rate. To reduce the complexity of the simulation, the reactor is assumed to operate isothermally with a temperature control system that maintains surface temperature uniformity and the standard operating temperature of the reactor, which has been developed using a model predictive controller with sparse identification modeling in prior work (Ou et al., 2024). Tom et al. also introduced a pressure disturbance that reduces the probability of reagent adsorption. While there are numerous disturbances encountered in the reactor operation, the generalization of the disturbances through their agglomeration into a "kinetic" disturbance simplifies the control system. For example, Yun et al. and Tom et al. made this simplification to reduce perturbations through general kinetic shift and process drift disturbances, which resemble disturbances such as side-wall deposition and corrosion on reactor surfaces (Butler, 1995), by decreasing the reaction rate constants described in Section 2.1. The decreasing of the reaction rate constants are intended to exemplify the uncertainties surrounding disturbance identification, which is generally difficult to predict in fast dynamics operation in real time. Additionally, the role of ramp disturbances in the wafer surface pressure must be considered as a consequence of competing side reactions, immediate changes in the operating conditions (e.g., the TMA and HF flow rates), or defective and miscalibrated equipment. It is notable that pressure changes influence the rate of adsorption of TMA and HF in the initial reaction mechanism for Steps A and B, respectively. Thus, the Collision Theory equation, which evaluates the temperature- and pressure-dependent adsorption reaction rate constant, is influenced by the pressure disturbance. To introduce these disturbances to the multiscale model, a kinetic shift disturbance is applied to the kMC simulation

by multiplying all reaction rate constants by a multiplicative factor of 0.8 to reduce the rate of kinetics for the overall process. Meanwhile, a ramp pressure disturbance is introduced into the CFD simulation by defining an operating pressure that is reduced linearly for the first two seconds and then maintained constant at the final value, which is expressed by the following equation:

$$P_{op} = \begin{cases} P_0 - 50t & \text{for } 0 < t \leq 2 \\ P_0 - 100 & \text{for } t > 2 \end{cases} \tag{5.15}$$

where P_{op} is the operating pressure in *Pa*, P_0 is the starting operating pressure of 300 *Pa*, and t is time. Essentially, the operating pressure falls from 300 *Pa* to 200 *Pa* over the course of 2 *s*, after which the operating pressure is maintained at 200 *Pa*.

5.3 Controller Tuning and Closed-loop Simulation Results

5.3.1 Tuning of the R2R Controller

The value for the exponential weight, λ, affects the amount of influence that historical data has on the R2R control action [85]. For the purposes of this work, various weighting parameters were studied to determine their impact on the controller performance when subjected to the kinetic shift disturbance. The observed impact of the controller correction to the inputs on the EPC output for several λ is demonstrated in the control plots in Fig. 5.8. Results illustrate that lower-weight λ requires lesser batch runs to sufficiently approach the setpoint. Thus, the R2R controller performance depends more on older batch data, which aligns with the tuning suggestions made in [79].

5.3.2 Tuning of the PI Controller

Appropriate tuning of the PI controller is imperative in ensuring the elimination of offset is obtained with minimal process time. PI control can introduce oscillatory response depending on

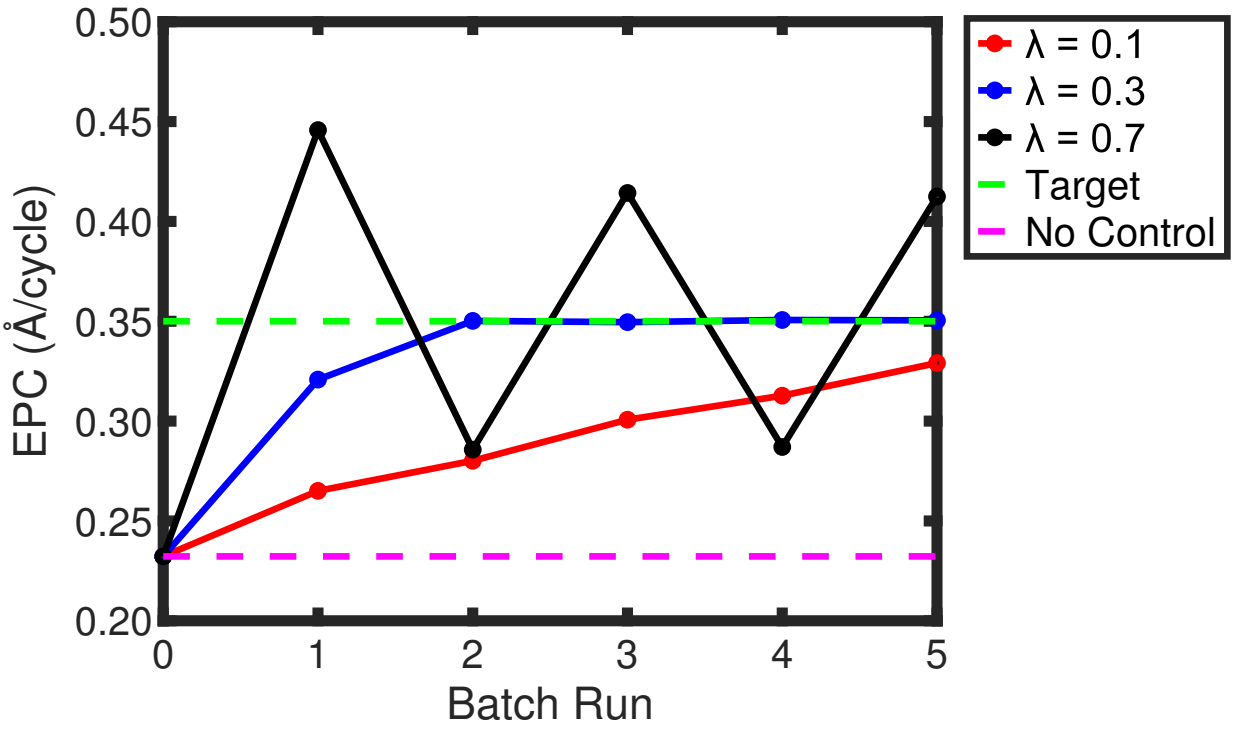
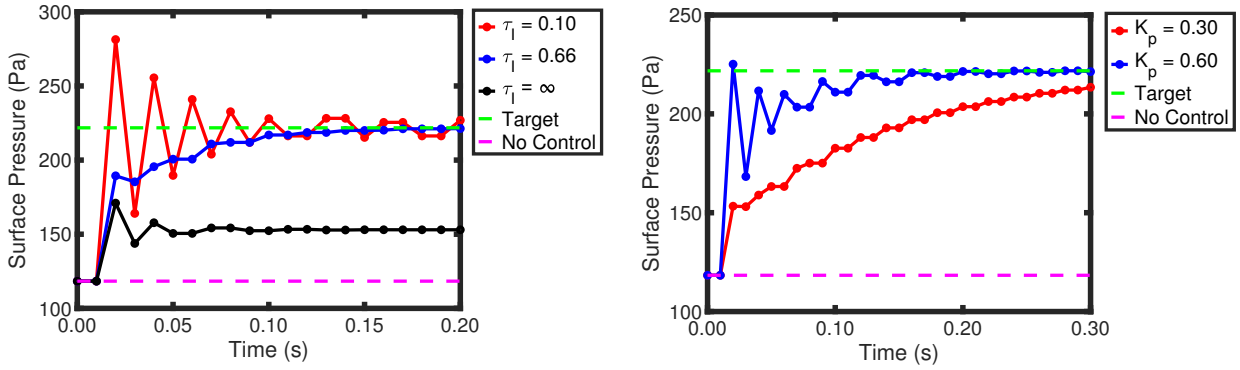


Figure 5.8: R2R control plot for various EWMA weights, λ, to determine an optimal weighting parameter that approaches the target EPC in lesser batch runs.

the value of the integral time constant defined to the controller, which can be mitigated with the integration of derivative control or by appropriately tuning the PI tuning parameters through a closed-loop tuning approach by introducing the ramp disturbance to the multiscale CFD simulation. The latter procedure is studied in this work. While there are numerous tuning methodologies (e.g., Ziegler-Nichols and Cohen-Coon), this work employs a systematic approach to studying the behavior of the process response with various integral times, τ_I, and proportional gains, K_p. The tuning procedure applies a constant K_p value for multiple τ_I, and vice versa, to determine the optimal tuning parameters for the PI controller, and the controller response is presented in Fig. 5.9. Results demonstrate that increasing τ_I generally increases the time required to eliminate the offset in Fig. 5.9a. However, lower τ_I increases oscillatory behavior due to increased influence by the accumulated error term in Eq. (5.14) on the PI control action. For the process gain, it is illustrated in Fig. 5.9b that lower K_p requires more process time to reach the setpoint. Thus, for the purposes



(a) $K_p = 0.60$ (b) $\tau_I = 2.00$

Figure 5.9: Controller responses to various τ_I at constant $K_p = 0.60$ in (a) and various K_p at a constant $\tau_I = 2.00$ to determine optimal parameters that eliminate offset in minimal time.

of this work, $\tau_I = 0.66$ and $K_p = 0.60$ were specified to the PI controller.

5.3.3 Integrated Run-to-Run Control and Feedback Control

The inclusion of simultaneous R2R and PI control is necessary to perform controller adjustment by measuring both the EPC after the completion of one thermal ALE cycle through a Quartz Crystal Microbalance offline and the wafer surface pressure through MEMS sensors. The performance of the combined R2R and PI control system is compared to that of a conventional R2R system in the form of controller response to the kinetic shift and pressure drift disturbances, which is presented in Fig. 5.10. The measured output, EPC, response illustrated in Fig. 5.10a indicates that the individual R2R control system requires one less batch run to reduce the offset from the setpoint compared to that of the combined R2R and PI control system. However, a consequence of the faster response at mitigating the disturbance requires a larger expenditure of reagent and a substantial increase in residence time, which is not ideal for thermal ALE operation. When investigating the controller adjustment to the manipulated input variables, results illustrate that the combined R2R and PI control system performs better than that of the single R2R control system by requiring higher substrate residence times (i.e., lower substrate velocities), and reduced

reagent consumption (i.e., lesser reagent flow rates) after the completion of one thermal ALE cycle in Figs. 5.10b to 5.10d. An observable advantage of the conjoined R2R and PI control system is

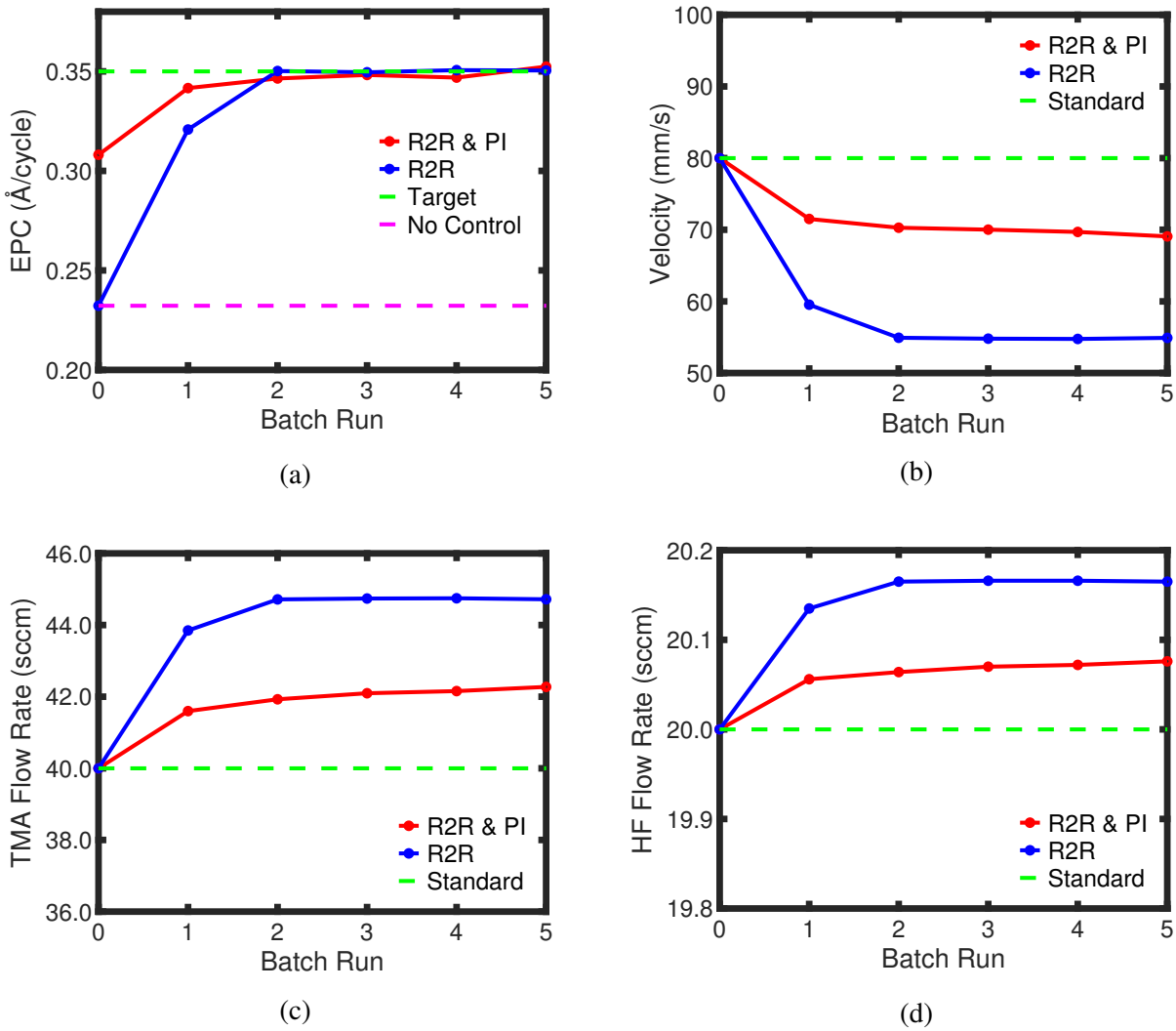


Figure 5.10: Comparison of (a) EPC control, (b) substrate velocity, (c) TMA flow rate, and (d) HF flow rate plots for the single R2R controller for a $\lambda = 0.3$ with the combined R2R and PI control system for a $\lambda = 0.3$ and $\tau_I = 0.66$.

the ability for the controller to implement correction within the batch run to mitigate the effects of the pressure disturbance, while also reducing the effects of the kinetic disturbance. Due to the regulation of the reagent flow rates for the PI controller, a reduction in wafer velocity is pronounced as a consequence of the R2R controller performing velocity correction following the completion

of the batch run. The performance of each control system is further expressed in terms of the averaged EPC error across all batch runs in Table 5.2, which illustrates that the combined R2R and PI control system reduces the averaged error substantially. Additionally, the exponential weight of $\lambda = 0.3$ results in minimal averaged EPC error.

Table 5.2: Comparison of averaged errors over 5 batch runs between the target and measured pressures.

Controller Model	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.7$
R2R	0.063	0.025	0.078
PI	0.042	0.042	0.042
R2R & PI	0.019	0.010	0.017

The primary objective of the combined R2R and PI control system is to have a fast response time, which was achievable with deterministic tuning parameters. A fast response time precludes a reduction of unused reagent and exposure time needed to obtain complete surface coverage of the terminated oxide film. For example, the PI controller results in Fig. 5.9 show that control actions take less than 0.1 *s* to be felt on the wafer surface. Because the reactor knows where the wafer is as well as how quickly the wafer is moving, the reactor can only apply the control actions when the wafer is within the reaction zone. Thus, when the wafer is in the purge zone, the control actions can be disabled, minimizing the usage of expensive reagents.

5.3.4 Robustness

While previous efforts were conducted to study the impact of the weighting parameter for a single R2R control system, this section further investigates the role of the exponential weight, λ, for the combined R2R and PI control system. Previously, it was mentioned that the PI control reduces the offset in the pressure, but due to the HF and TMA flow rate corrections coinciding with the adjustments made to the HF and TMA flow rates by the R2R controller, the combined PI

and R2R control mitigate the EPC offset. Therefore, the impact of the R2R controller, and the λ that is supplied to the R2R controller, on the EPC correction can be limited. Fig. 5.11 depicts the connected R2R and PI control system response to the kinetic and pressure disturbances for multiple λ. Results indicate that despite the aforementioned assertion, λ largely introduces oscillatory behavior with increasing λ, which makes it difficult to discern the impact of λ on the output. Also, $\lambda = 0.1$ requires additional batch runs due to the slower controller response and effect of recent EPC error on the subsequent batch run.

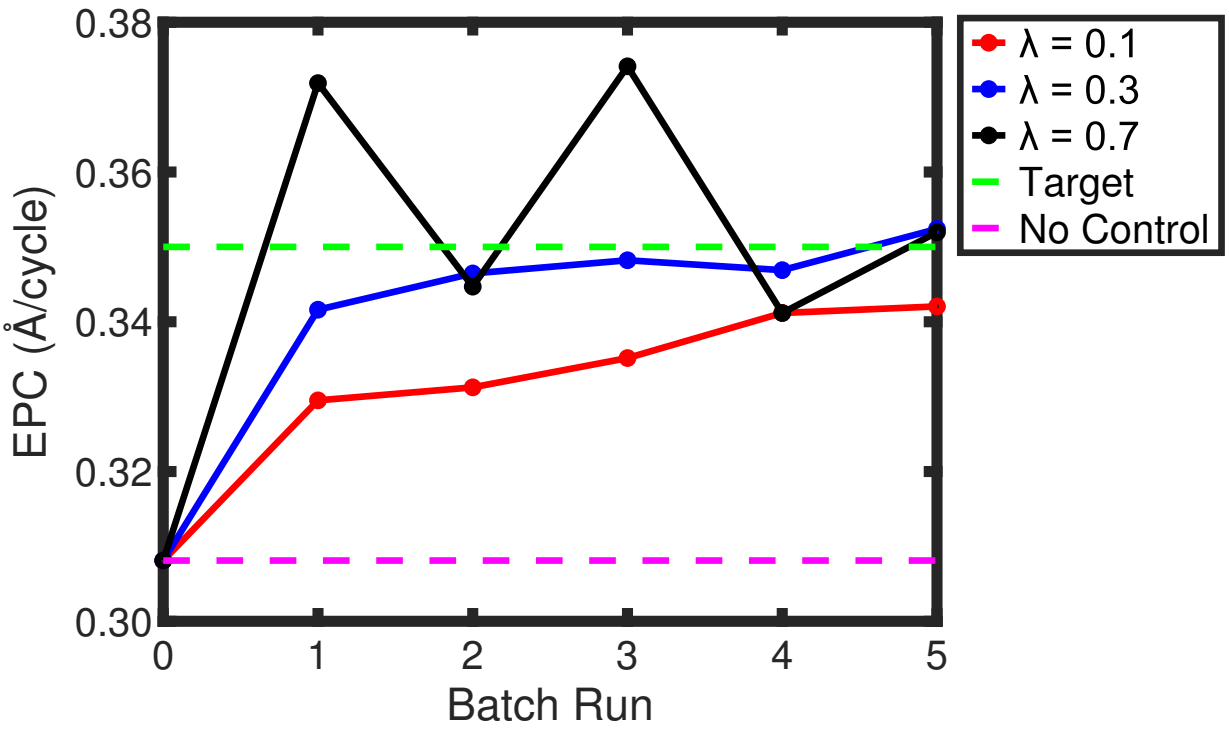


Figure 5.11: Comparison of control plots for the combined R2R and PI control system for various EWMA weights, λ to determine an optimal λ that reaches the target in a minimal number of batch runs.

5.4 Conclusion

Thermal atomic layer etching (ALE) is a crucial procedure to enable the fabrication of downscaled transistors, which occupy semiconducting wafers. However, thermal ALE is characterized by being inaccurate and nonproductive due to the demanding design criteria required to produce high-performance semiconducting chips. Thus, an *in silico* multiscale modeling approach is adopted to determine optimal operating conditions to produce highly conformal transistor films for a spatial, sheet-to-sheet reactor that is recognized for increasing product throughput. Previous works have focused on efforts to integrate run-to-run (R2R) control systems with exponentially weighted moving average (EWMA) algorithms to compensate for the effects of perturbations to the thermal ALE process through a multivariate control procedure; however, continuous feedback control is needed to improve the correction of disturbances within the batch process. This work designed a conjoined R2R and Proportional-Integral (PI) control system that implements control action both continuously and after the completion of a thermal ALE cycle. The combination of both control systems successfully optimized control performance through the tuning of both controllers, which led to observable reduction in input parameter deviation from their standard operating conditions.

Chapter 6

Conclusion

This dissertation discussed some of the current obstacles encountered in the semiconductor processing industries, particularly in the fabrication of thin oxide films on the surfaces of transistor nanowires. To navigate around the issue with imprecise deposition and nonproductive atomic layer processes that require numerous pre- and post-processing steps, area-selective atomic layer deposition (AS-ALD) was proposed as a solution to the matter. However, this process lacks characterization to enable scale-up in industrial practices. In particular, industry has faced issues with process optimization to the lack of experimental and industrial data; thus, this dissertation employed *in silico* multiscale computational fluid dynamics modeling as a way to produce data at a more efficient rate. Additionally, the impact of reactor geometry, particularly the delivery system of reagent to the substrate, was studied to determine optimal reactor models that employ a discrete feed mechanism, which maximizes productivity, minimizes reagent consumption, and film uniformity. Lastly, this dissertation explored various control systems that detect and mitigate pressure and kinetic disturbances applied to the multiscale simulations by integrating online feedback and run-to-run (R2R) control.

Chapter 2 examined the impact of the reactor delivery system on the transport of reagent onto the substrate surface. Stationary reactor models that employ a discrete feeding mechanism were

designed with small volumes and utilized a perpendicular flow orientation through a showerhead distributor to minimize the potential for steric collisions. Differing inlet plate geometries (single-, ring-, multi-, and combined single- and ring-inlet) were created to determine their impact on the delivery of reagent. The spatiotemporal progression of the reagent transfer was captured using macroscopic computational fluid dynamics (CFD) for various reactor configurations under the assumption that reactor operating conditions are constant. Results indicated that smaller characteristic lengths, while effective at reducing potential turbulent flow, led to a lack of reagent transport during the dosing time; thus, the selection of the reactor relies on delivery systems that are conducive to increasing reagent movement by requiring lesser dosage time.

Chapter 3 designed a multiscale CFD simulation of the AS-ALD process for the stationary reactor models to study the impact of the reactor geometry on the film surface coverage. In order to preserve the efficiency of the simulation, the kinetic Monte Carlo (kMC) simulation was integrated into the CFD software environment through user-defined functions (UDFs) written in the C programming code, which facilitate parallel computations. The resulting multiscale CFD model allowed the tracking of the spatiotemporal progression of the surface coverage, which discovered that some reactor models were more favorable than others for some steps of the AS-ALD process. For example, the precursor adsorption step must overcome a statistically improbable collision on the substrate surface by requiring a higher initial concentration of reagent, which is observed by the single-inlet reactor model. The combined-inlet reactor model was more successful at achieving maximum surface coverage with minimal processing time for Steps A and C of the AS-ALD cycle.

Chapter 4 discusses the integration of a R2R controller into a spatial rotary reactor of an AS-ALD process. A constant kinetic shift and pressure shift disturbance were introduced separately to a multiscale CFD simulation, resulting in process deviation from the user-defined setpoint. To overcome this disturbance, a R2R controller utilizing an exponentially weighted moving average (EWMA) of a linear input (substrate rotation speed) and output (growth per cycle) model generated from offline multiscale CFD data that is undisturbed was implemented. The integration of

R2R controller was able to mitigate the effects of the disturbances in less than 20 batch runs, and effective exponential weight tuning was able to overcome the effects of noisy measured output.

Chapter 5 developed a conjoined control system comprising an online feedback and R2R controller of a spatial, sheet-to-sheet reactor of an atomic layer etching (ALE) process. A constant kinetic shift and a constant pressure drift were defined to the multiscale CFD simulation to reflect disturbances observed in industrial practice. To overcome these disturbances, an online feedback control in the form of proportional-integral (PI) control was designed by measuring the substrate surface pressure and adjusting the reagent flow rates. This PI controller was successfully tuned to balance the level of proportional control to allow for faster correction and integral control to minimize process offset from the setpoint. Lastly, an R2R controller using an EWMA of a linear model for a multiple input (substrate velocity and reagent flow rates) and single output (etching per cycle) was utilized to implement adjustment of the inputs offline. The results found that the combined PI and R2R control system was able to reduce reagent consumption and reduce unnecessary process time when compared to an individual R2R controller.

Bibliography

[1] Andrews, M., 2022. Critical manufacturing redefines semiconductor MES. Silicon Semiconductor, 43, 38–41.

[2] Anitha, V. C., Banerjee, A. N., & Joo, S. W., 2015. Recent developments in TiO_2 as n- and p-type transparent semiconductors: Synthesis, modification, properties, and energy-related applications. Journal of Materials Science, 50, 7495–7536.

[3] ANSYS, 2022. Ansys Fluent Theory Guide. ANSYS Inc., Canonsburg, PA, USA.

[4] ANSYS, 2022. Ansys Fluent Customization Manual. ANSYS Inc., Canonsburg, PA, USA.

[5] ANSYS, 2022. Ansys Fluent User's Guide. ANSYS Inc., Canonsburg, PA, USA.

[6] Asenov, A., Wang, Y., Cheng, B., Wang, X., Asenov, P., Al-Ameri, T., & Georgiev, V. P., 2016. Nanowire transistor solutions for 5nm and beyond. In: Proceedings of 17th International Symposium on Quality Electronic Design, 269–274, Santa Clara, CA, USA.

[7] Baroni, S., Giannozzi, P., & Isaev, E., 2010. Density-functional perturbation theory for quasi-harmonic calculations. Reviews in Mineralogy and Geochemistry, 71, 39–57.

[8] Bortz, A. B., Kalos, M. H., & Lebowitz, J. L., 1975. A new algorithm for Monte Carlo simulation of Ising spin systems. Journal of Computational Physics, 17, 10–18.

[9] Butler, S. W., 1995. Process control in semiconductor manufacturing. Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena, 13, 1917–1923.

[10] Carson, P. K. & Yeh, A. B., 2008. Exponentially weighted moving average (EWMA) control charts for monitoring an analytical process. Industrial & Engineering Chemistry Research, 47, 405–411.

[11] Carver, C. T., Plombon, J. J., Romero, P. E., Suri, S., Tronic, T. A., & Turkot Jr., R. B., 2015. Atomic layer etching: An industry perspective. ECS Journal of Solid State Science and Technology, 4, N5005.

[12] Chandrasekar, A., Garg, A., Abdulhussain, H. A., Gritsichine, V., Thompson, M. R., & Mhaskar, P., 2022. Design and application of data driven economic model predictive control for a rotational molding process. Computers & Chemical Engineering, 161, 107713.

[13] Chandrasekar, A., Zhang, S., & Mhaskar, P., 2023. A hybrid hubspace-RNN based approach for modelling of non-linear batch processes. Chemical Engineering Science, 281, 119118.

[14] Cheimarios, N., To, D., Kokkoris, G., Memos, G., & Boudouvis, A. G., 2021. Monte carlo and kinetic monte carlo models for deposition processes: A review of recent works. Frontiers in Physics, 9, 631918.

[15] Chen, R. & Bent, S. F., 2006. Chemistry for positive pattern transfer using area-selective atomic layer deposition. Advanced Materials, 18, 1086–1090.

[16] Chen, R., Kim, H., McIntyre, P. C., Porter, D. W., & Bent, S. F., 2005. Achieving area-selective atomic layer deposition on patterned substrates by selective surface modification. Applied Physics Letters, 86, 191910.

[17] Chen, Y., Li, Z., Dai, Z., Yang, F., Wen, Y., Shan, B., & Chen, R., 2023. Multiscale CFD modelling for conformal atomic layer deposition in high aspect ratio nanostructures. Chemical Engineering Journal, 472, 144944.

[18] Cheng, C. L., Chang, H. C., Chang, C. I., & Fang, W., 2015. Development of a CMOS MEMS pressure sensor with a mechanical force-displacement transduction structure. Journal of Micromechanics and Microengineering, 25, 125024.

[19] Chiappim, W., Neto, B. B., Shiotani, M., Karnopp, J., Gonçalves, L., Chaves, J. P., Sobrinho, A. d. S., Leitão, J. P., Fraga, M., & Pessoa, R., 2022. Plasma-assisted nanofabrication: The potential and challenges in atomic layer deposition and etching. Nanomaterials, 12, 3497.

[20] Christofides, P. D. & Armaou, A., 2006. Control and optimization of multiscale process systems. Computers & Chemical Engineering, 30, 1670–1686.

[21] Christofides, P. D., Armaou, A., Lou, Y., & Varshney, A., 2009. Control and Optimization of Multiscale Process Systems. Birkhäuser, Boston, MA, USA.

[22] Cong, W., Li, Z., Cao, K., Feng, G., & Chen, R., 2020. Transient analysis and process optimization of the spatial atomic layer deposition using the dynamic mesh method. Chemical Engineering Science, 217, 115513.

[23] Coughanowr, D. R. & LeBlanc, S. E., 2009. Process Systems Analysis and Control. Mcgraw-Hill, Boston, MA, USA, 3rd edition.

[24] Dahmen, K. H., 2003. Chemical vapor deposition. In: Meyers, R. A. (Ed.), Encyclopedia of Physical Science and Technology, 787–808. Academic Press, New York, NY, USA, 3rd edition.

[25] De la Huerta, C. M., Nguyen, V. H., Dedulle, J. M., Bellet, D., Jiménez, C., & Muñoz-Rojas,

D., 2018. Influence of the geometric parameters on the deposition mode in spatial atomic layer deposition: A novel approach to area-selective deposition. Coatings, 9, 5.

[26] Del Castillo, E., 1996. A multivariate self-tuning controller for run-to-run process control under shift and trend disturbances. IIE Transactions, 28, 1011–1021.

[27] Del Castillo, E. & Hurwitz, A. M., 1997. Run-to-run process control: Literature review and extensions. Journal of Quality Technology, 29, 184–196.

[28] Deng, Z., He, W., Duan, C., Chen, R., & Shan, B., 2016. Mechanistic modeling study on process optimization and precursor utilization with atmospheric spatial atomic layer deposition. Journal of Vacuum Science & Technology A, 34, 01A108.

[29] Deng, Z., He, W., Duan, C., Shan, B., & Chen, R., 2016. Atomic layer deposition process optimization by computational fluid dynamics. Vacuum, 123, 103–110.

[30] Derbyshire, K., 2023. Using ML for improved fab scheduling. https://semiengineering.com/using-ml-for-improved-fab-scheduling/. Accessed: 2024-01-07.

[31] DeVita, J. P., Sander, L. M., & Smereka, P., 2005. Multiscale kinetic monte carlo algorithm for simulating epitaxial growth. Physical Review B, 72, 205421.

[32] Dobkin, D. M. & Zuraw, M. K. (Eds.), 2003. Principles of Chemical Vapor Deposition, Volume 1. Springer Dordrecht.

[33] Elers, K. E., Blomberg, T., Peussa, M., Aitchison, B., Haukka, S., & Marcus, S., 2006. Film uniformity in atomic layer deposition. Chemical Vapor Deposition, 12, 13–24.

[34] Engelmann, S. U., Bruce, R. L., Nakamura, M., Metzler, D., Walton, S. G., & Joseph, E. A., 2015. Challenges of tailoring surface chemistry and plasma/surface interactions to advance atomic layer etching. ECS Journal of Solid State Science and Technology, 4, N5054.

[35] Frieske, B. & Stieler, S., 2022. The “semiconductor crisis” as a result of the COVID-19 pandemic and impacts on the automotive industry and its supply chains. World Electric Vehicle Journal, 13, 189.

[36] George, S. M., 2010. Atomic layer deposition: An overview. Chemical Reviews, 110, 111–131.

[37] George, S. M., 2020. Mechanisms of thermal atomic layer etching. Accounts of Chemical Research, 53, 1151–1160.

[38] Giannozzi, P., 2009. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. Journal of Physics: Condensed Matter, 21, 395502.

[39] Gillespie, D. T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics, 22, 403–434.

[40] Giménez, M. C., Del Pópolo, M. G., & Leiva, E. P. M., 2002. Kinetic Monte Carlo study of electrochemical growth in a heteroepitaxial system. Langmuir, 18, 9087–9094.

[41] Holmqvist, A., Törndahl, T., & Stenström, S., 2012. A model-based methodology for the analysis and design of atomic layer deposition processes-part i: Mechanistic modelling of continuous flow reactors. Chemical Engineering Science, 81, 260–272.

[42] Horstemeyer, M. F., 2010. Practical Aspects of Computational Chemistry: Methods, Concepts and Applications, Chapter: Multiscale Modeling: A Review, 87–135. Springer Netherlands, Dordrecht, Netherlands.

[43] Huang, A., Meng, S., & Huang, T., 2023. A survey on machine and deep learning in semiconductor industry: Methods, opportunities, and challenges. Cluster Computing, 26, 3437–3472.

[44] Huang, J., 2022. Research progresses on suppressing the short-channel effects of field-effect transistor. Highlights in Science, Engineering and Technology, 27, 361–367.

[45] Huard, C. M., Lanham, S. J., & Kushner, M. J., 2018. Consequences of atomic layer etching on wafer scale uniformity in inductively coupled plasmas. Journal of Physics D: Applied Physics, 51, 155201.

[46] Ishikawa, K., Karahashi, K., Ichiki, T., Chang, J. P., George, S. M., Kessels, W. M. M., Lee, H. J., Tinck, S., Um, J. H., & Kinoshita, K., 2017. Progress and prospects in nanoscale dry processes: How can we control atomic layer reactions? Japanese Journal of Applied Physics, 56, 06HA02.

[47] Jansen, A. P. J. (Ed.), 2012. An Introduction to Kinetic Monte Carlo Simulations of Surface Reactions, Volume 1. Academic Press.

[48] Javed, Y., Mansoor, M., & Shah, I. A., 2019. A review of principles of MEMS pressure sensing with its aerospace applications. Sensor Review, 39, 652–664.

[49] Johnson, R. W., Hultqvist, A., & Bent, S. F., 2014. A brief review of atomic layer deposition: From fundamentals to applications. Materials Today, 17, 236–246.

[50] Jurczak, M., Collaert, N., Veloso, A., Hoffmann, T., & Biesemans, S., 2009. Review of FINFET technology. In: 2009 IEEE International SOI Conference, 1–4.

[51] Kanarik, K. J., Lill, T., Hudson, E. A., Sriraman, S., Tan, S., Marks, J., Vahedi, V., & Gottscho, R. A., 2015. Overview of atomic layer etching in the semiconductor industry. Journal of Vacuum Science & Technology A, 33, 020802.

[52] Khakifirooz, M., Fathi, M., & Wu, K., 2019. Development of smart semiconductor manufacturing: operations research and data science perspectives. IEEE Access, 7, 108419–108430.

[53] Kim, H. G., Kim, M., Gu, B., Khan, M. R., Ko, B. G., Yasmeen, S., Kim, C. S., Kwon, S. H., Kim, J., Kwon, J., Jin, K., Cho, B., Chun, J. S., Shong, B., & Lee, H. B. R., 2020. Effects of Al precursors on deposition selectivity of atomic layer deposition of Al_2O_3 using ethanethiol inhibitor. Chemistry of Materials, 32, 8921–8929.

[54] Kim, J., Chakrabarti, K., Lee, J., Oh, K. Y., & Lee, C., 2003. Effects of ozone as an oxygen source on the properties of the Al_2O_3 thin films prepared by atomic layer deposition. Materials Chemistry and Physics, 78, 733–738.

[55] Kimes, W. A., Moore, E. F., & Maslar, J. E., 2012. Perpendicular-flow, single-wafer atomic layer deposition reactor chamber design for use with in situ diagnostics. Review of Scientific Instruments, 83, 083106.

[56] Klement, P., Anders, D., Gümbel, L., Bastianello, M., Michel, F., Schörmann, J., Elm, M. T., Heiliger, C., & Chatterjee, S., 2021. Surface diffusion control enables tailored-aspect-ratio nanostructures in area-selective atomic layer deposition. ACS Applied Materials & Interfaces, 13, 19398–19405.

[57] Kolahdouz, M., Xu, B., Nasiri, A. F., Fathollahzadeh, M., Manian, M., Aghababa, H., Wu, Y., & Radamson, H. H., 2022. Carbon-related materials: Graphene and carbon nanotubes in semiconductor applications and design. Micromachines, 13, 1257.

[58] Lauwers, L., 2013. Semiconductor technology enabling smart electronics. In: Chakravarthi, V. S., Shirur, Y. J. M., & Prasad, R. (Eds.), Proceedings of International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking (VCASAN-2013), 15–24, Bangalore, India. Springer India.

[59] Lee, A. J., Lee, S., Han, D. H., Kim, Y., & Jeon, W., 2023. Enhancing chemisorption efficiency and thin-film characteristics via a discrete feeding method in high-k dielectric

atomic layer deposition for preventing interfacial layer formation. Journal of Materials Chemistry C, 11, 6894–6901.

[60] Lee, C. W., Cho, T. L., & Kim, M. S., 2017. Is there a better semiconductor firm in taiwan? Management and Economics Review, 2, 37–46.

[61] Lee, F., Marcus, S., Shero, E., Wilk, G., Swerts, J., Maes, J. W., Blomberg, T., Delabie, A., Gros-Jean, M., & Deloffre, E., 2007. Atomic layer deposition: An enabling technology for microelectronic device manufacturing. In: Proceedings of IEEE/SEMI Advanced Semiconductor Manufacturing Conference, 359–365, Stresa, Italy.

[62] Lee, G., Lee, B., Kim, J., & Cho, K., 2009. Ozone adsorption on graphene: Ab initio study and experimental validation. The Journal of Physical Chemistry C, 113, 14225–14229.

[63] Lee, J. M., Lee, J., Oh, H., Kim, J., Shong, B., Park, T. J., & Kim, W. H., 2022. Inhibitor-free area-selective atomic layer deposition of SiO_2 through chemoselective adsorption of an aminodisilane precursor on oxide versus nitride substrates. Applied Surface Science, 589, 152939.

[64] Li, J., Tezsevin, I., Merkx, M. J. M., Maas, J. F. W., Kessels, W. M. M., Sandoval, T. E., & Mackus, A. J. M., 2022. Packing of inhibitor molecules during area-selective atomic layer deposition studied using random sequential adsorption simulations. Journal of Vacuum Science & Technology A, 40, 062409.

[65] Li, J., Ge, W., Wang, W., Yang, N., Liu, X., Wang, L., He, X., Wang, X., Wang, J., & Kwauk, M., 2013. From Multiscale Modeling to Meso-Science: A Chemical Engineering Perspective. Springer, Berlin.

[66] Lin, S. C., Wang, C. C., Tien, C. L., Tung, F. C., Wang, H. F., & Lai, S. H., 2023. Fabrication of aluminum oxide thin-film devices based on atomic layer deposition and pulsed discrete feed method. Micromachines, 14, 279.

[67] Liu, K., Chen, Y., Zhang, T., Tian, S., & Zhang, X., 2018. A survey of run-to-run control for batch processes. ISA Transactions, 83, 107–125.

[68] Ljung, L., 2010. Perspectives on system identification. Annual Reviews in Control, 34, 1–12.

[69] Loubet, N., Hook, T., Montanini, P., Yeung, C. W., Kanakasabapathy, S., Guillom, M., Yamashita, T., Zhang, J., Miao, X., Wang, J., Young, A., Chao, R., Kang, M., Liu, Z., Fan, S., Hamieh, B., Sieg, S., Mignot, Y., Xu, W., Seo, S. C., Yoo, J., Mochizuki, S., Sankarapandian, M., Kwon, O., Carr, A., Greene, A., Park, Y., Frougier, J., Galatage, R., Bao, R., Shearer, J., Conti, R., Song, H., Lee, D., Kong, D., Xu, Y., Arceo, A., Bi, Z., Xu, P., Muthinti, R., Li, J., Wong, R., Brown, D., Oldiges, P., Robison, R., Arnold, J., Felix, N., Skordas, S., Gaudiello, J., Standaert, T., Jagannathan, H., Corliss, D., Na, M. H., Knorr, A., Wu, T., Gupta, D., Lian, S., Divakaruni, R., Gow, T., Labelle, C., Lee, S., Paruchuri, V., Bu, H., & Khare, M., 2017. Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET. In: Proceedings of Symposium on VLSI Technology, T230–T231, Kyoto, Japan.

[70] Lubitz, M., Medina, P. A., Antic, A., Rosin, J. T., & Fahlman, B. D., 2014. Cost-effective systems for atomic layer deposition. Journal of chemical education, 91, 1022–1027.

[71] Lucas, J. M. & Saccucci, M. S., 1990. Exponentially weighted moving average control schemes: Properties and enhancements. Technometrics, 32, 1–12.

[72] Mackus, A. J. M., Bol, A. A., & Kessels, W. M. M., 2014. The use of atomic layer deposition in advanced nanopatterning. IEEE Electron Device Letters, 6, 10941–10960.

[73] Mackus, A. J. M., Merkx, M. J. M., & Kessels, W. M. M., 2019. From the bottom-up: Toward area-selective atomic layer deposition with high selectivity. Chemistry of Materials, 31, 2–12.

[74] Mameli, A., Merkx, M. J. M., Karasulu, B., Roozeboom, F., Kessels, W. E. M. M., & Mackus, A. J. M., 2017. Area-selective atomic layer deposition of SiO_2 using acetylacetone as a chemoselective inhibitor in an ABC-type cycle. ACS Nano, 11, 9303–9311.

[75] Maroudas, D., 2000. Multiscale modeling of hard materials: Challenges and opportunities for chemical engineering. AIChE Journal, 46, 878–882.

[76] Merkx, M. J. M., Sandoval, T. E., Hausmann, D. M., Kessels, W. M. M., & Mackus, A. J. M., 2020. Mechanism of precursor blocking by acetylacetone inhibitor molecules during area-selective atomic layer deposition of SiO_2. Chemistry of Materials, 32, 3335–3345.

[77] Merkx, M. J. M., Angelidis, A., Mameli, A., Li, J., Lemaire, P. C., Sharma, K., Hausmann, D. M., Kessels, W. M. M., Sandoval, T. E., & Mackus, A. J. M., 2022. Relation between reactive surface sites and precursor choice for area-selective atomic layer deposition using small molecule inhibitors. The Journal of Physical Chemistry C, 126, 4845–4853.

[78] Mohammad, W., Elomri, A., & Kerbache, L., 2022. The global semiconductor chip shortage: Causes, implications, and potential remedies. IFAC-PapersOnLine, 55, 476–483.

[79] Montgomery, D. C., 2013. Introduction to statistical quality control. John Wiley & Sons, Hoboken, NJ, USA, 7th edition.

[80] Moore, G. E., 1998. Cramming more components onto integrated circuits. Proceedings of the IEEE, 86, 82–85.

[81] Mousa, M. B. M., Oldham, C. J., & Parsons, G. N., 2015. Precise nanoscale surface modification and coating of macroscale objects: Open-environment in loco atomic layer deposition on an automobile. ACS Applied Materials & Interfaces, 7, 19523–19529.

[82] Moyne, J., Del Castillo, E., & Hurwitz, A. M., 2018. Run-to-Run Control in Semiconductor Manufacturing. CRC Press, Boca Raton, FL, USA.

[83] Muñoz-Rojas, D., Nguyen, V. H., De la Huerta, C. M., Jiménez, C., & Bellet, D., 2019. Spatial atomic layer deposition. In: Mandracci, P. (Ed.), Chemical Vapor Deposition for Nanotechnology, 3–27. IntechOpen, London, UK.

[84] Muneshwar, T. & Cadien, K., 2016. A_xBA_xB... pulsed atomic layer deposition: Numerical growth model and experiments. Journal of Applied Physics, 119, 085306.

[85] Oakland, J. S., 2003. Statistical Process Control. Butterworth-Heinemann, Oxford, UK, 5th edition.

[86] Pan, D., 2021. Density functional theory (DFT)-enhanced computational fluid dynamics modeling of substrate movement and chemical deposition process in spatial atomic layer deposition. Chemical Engineering Science, 234, 116447.

[87] Pan, D., Jen, T. C., & Yuan, C., 2016. Effects of gap size, temperature and pumping pressure on the fluid dynamics and chemical kinetics of in-line spatial atomic layer deposition of Al_2O_3. International Journal of Heat and Mass Transfer, 96, 189–198.

[88] Park, T. J., Kim, J. H., Jang, J. H., Kim, U. K., Lee, S. Y., Lee, J., Jung, H. S., & Hwang, C. S., 2011. Improved growth and electrical properties of atomic-layer-deposited metal-oxide film by discrete feeding method of metal precursor. Chemistry of Materials, 23, 1654–1658.

[89] Petti, L., Münzenrieder, N., Vogt, C., Faber, H., Büthe, L., Cantarella, G., Bottacchi, F., Anthopoulos, T. D., & Tröster, G., 2016. Metal oxide semiconductor thin-film transistors for flexible electronics. Applied Physics Reviews, 3, 021303.

[90] Ponraj, J. S., Attolini, G., & Bosi, M., 2013. Review on atomic layer deposition and applications of oxide thin films. Critical Reviews in Solid State and Materials Sciences, 38, 203–233.

[91] Poodt, P., Lankhorst, A., Roozeboom, F., Spee, K., Maas, D., & Vermeer, A., 2010. High-speed spatial atomic-layer deposition of aluminum oxide layers for solar cell passivation. Advanced Materials, 22, 3564–3567.

[92] Poodt, P., Cameron, D. C., Dickey, E., George, S. M., Kuznetsov, V., Parsons, G. N., Roozeboom, F., Sundaram, G., & Vermeer, A., 2012. Spatial atomic layer deposition: A route towards further industrialization of atomic layer deposition. Journal of Vacuum Science & Technology A, 30, 010802.

[93] Puurunen, R. L., 2005. Surface chemistry of atomic layer deposition: A case study for the trimethylaluminum/water process. Journal of Applied Physics, 97, 121301.

[94] Rashid, M. & Mhaskar, P., 2023. Are neural networks the right tool for process modeling and control of batch and batch-like processes? Processes, 11, 686.

[95] Richard, C., 2023. Understanding Semiconductors: A Technical Guide for Non-Technical People. Apress, Berkeley, CA, USA.

[96] Ritala, M. & Leskelä, M., 2002. Chapter 2 - Atomic layer deposition. In: Singh Nalwa, H. (Ed.), Handbook of Thin Films, 103–159. Academic Press, Burlington, MA, USA.

[97] Roh, H., Kim, H. L., Khumaini, K., Son, H., Shin, D., & Lee, W. J., 2022. Effect of deposition temperature and surface reactions in atomic layer deposition of silicon oxide using bis(diethylamino)silane and ozone. Applied Surface Science, 571, 151231.

[98] Rojek, K., Wyrzykowski, R., & Gepner, P., 2021. AI-accelerated CFD simulation based on OpenFOAM and CPU/GPU computing. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V. V., Dongarra, J. J., & Sloot, P. M. A. (Eds.), Computational Science – ICCS 2021, 373–385, Cham, Switzerland. Springer International Publishing.

[99] Roozeboom, F., Kniknie, B., Lankhorst, A. M., Winands, G., Knaapen, R., Smets, M., Poodt, P., Dingemans, G., Keuning, W., & Kessels, W. M. M., 2012. A new concept for spatially divided deep reactive ion etching with ALD-based passivation. IOP Conference Series: Materials Science and Engineering, 41, 012001.

[100] Sachs, E., Hu, A., & Ingolfsson, A., 1995. Run by run process control: Combining SPC and feedback control. IEEE Transactions on Semiconductor Manufacturing, 8, 26–43.

[101] Schwille, M. C., Schössler, T., Schön, F., Oettel, M., & Bartha, J. W., 2017. Temperature dependence of the sticking coefficients of bis-diethyl aminosilane and trimethylaluminum in atomic layer deposition. Journal of Vacuum Science & Technology A, 35, 01B119.

[102] Shattuck, T. J., 2021. Stuck in the middle: Taiwan's semiconductor industry, the U.S.-China tech fight, and cross-strait stability. Orbis, 65, 101–117.

[103] Shenai, K., 2019. High-density power conversion and wide-bandgap semiconductor power electronics switching devices. Proceedings of the IEEE, 107, 2308–2326.

[104] Siimon, H. & Aarik, J., 1997. Thickness profiles of thin films caused by secondary reactions in flow-type atomic layer deposition reactors. Journal of Physics D: Applied Physics, 30, 1725–1728.

[105] Song, S. K., Saare, H., & Parsons, G. N., 2019. Integrated isothermal atomic layer deposition/atomic layer etching supercycles for area-selective deposition of TiO_2. Chemistry of Materials, 31, 4793–4804.

[106] Sun, X. H. & Chen, Y., 2010. Reevaluating amdahl's law in the multicore era. Journal of Parallel and Distributed Computing, 70, 183–188.

[107] Tom, M., Yun, S., Wang, H., Ou, F., Orkoulas, G., & Christofides, P. D., 2022. Machine

learning-based run-to-run control of a spatial thermal atomic layer etching reactor. Computers & Chemical Engineering, 168, 108044.

[108] Tom, M., Wang, H., Ou, F., Yun, S., Orkoulas, G., & Christofides, P. D., 2023. Computational fluid dynamics modeling of a discrete feed atomic layer deposition reactor: Application to reactor design and operation. Computers & Chemical Engineering, 178, 108400.

[109] Tom, M., Yun, S., Wang, H., Ou, F., Orkoulas, G., & Christofides, P. D., 2023. Multiscale modeling of spatial area-selective thermal atomic layer deposition. In: Kokossis, A. C., Georgiadis, M. C., & Pistikopoulos, E. (Eds.), Proceedings of 33rd European Symposium on Computer Aided Process Engineering, Volume 52 of Computer Aided Chemical Engineering, 71–76. Elsevier, Athens, Greece.

[110] TSMC, 2024. MEMS technology. https://www.tsmc.com/english/dedicatedFoundry/technology/specialty/mems. Accessed: 2024-01-07.

[111] Voas, J., Kshetri, N., & DeFranco, J. F., 2021. Scarcity and global insecurity: The semiconductor shortage. IT Professional, 23, 78–82.

[112] Wang, C., 2013. A study of R2R control improvement using adjustment limit to reduce frequency of control. South African Journal of Industrial Engineering, 24, 102–110.

[113] Wang, H., Wang, Z., Xu, X., Liu, Y., Chen, C., Chen, P., Hu, W., & Duan, Y., 2019. Multiple short pulse process for low-temperature atomic layer deposition and its transient steric hindrance. Applied Physics Letters, 114, 201902.

[114] Wang, H., Tom, M., Ou, F., Orkoulas, G., & Christofides, P. D., 2024. Integrating run-to-run control with feedback control for a spatial atomic layer etching reactor. Chemical Engineering Research and Design, 203, 1–10.

[115] Wang, K. & Han, K., 2013. A batch-based run-to-run process control scheme for semiconductor manufacturing. IIE Transactions, 45, 658–669.

[116] Wehinger, G. D., Ambrosetti, M., Cheula, R., Ding, Z. B., Isoz, M., Kreitz, B., Kuhlmann, K., Kutscherauer, M., Niyogi, K., Poissonnier, J., Réocreux, R., Rudolf, D., Wagner, J., Zimmermann, R., Bracconi, M., Freund, H., Krewer, U., & Maestri, M., 2022. Quo vadis multiscale modeling in reaction engineering? – A perspective. Chemical Engineering Research and Design, 184, 39–58.

[117] Xiong, S., Jia, X., Mi, K., & Wang, Y., 2021. Upgrading polytetrafluoroethylene hollow-fiber membranes by CFD-optimized atomic layer deposition. Journal of Membrane Science, 617, 118610.

[118] Xu, W., Haeve, M. G. N., Lemaire, P. C., Sharma, K., Hausmann, D. M., & Agarwal, S., 2022. Functionalization of the SiO_2 surface with aminosilanes to enable area-selective atomic layer deposition of Al_2O_3. Langmuir, 38, 652–660.

[119] Yarbrough, J., Shearer, A. B., & Bent, S. F., 2021. Next generation nanopatterning using small molecule inhibitors for area-selective atomic layer deposition. Journal of Vacuum Science & Technology A, 39, 021002.

[120] Yarbrough, J., Pieck, F., Grigjanis, D., Oh, I. K., Maue, P., Tonner-Zech, R., & Bent, S. F., 2022. Tuning molecular inhibitors and aluminum precursors for the area-selective atomic layer deposition of Al_2O_3. Chemistry of Materials, 34, 4646–4659.

[121] Yun, S., Ding, Y., Zhang, Y., & Christofides, P. D., 2021. Integration of feedback control and run-to-run control for plasma enhanced atomic layer deposition of hafnium oxide thin films. Computers & Chemical Engineering, 148, 107267.

[122] Yun, S., Ou, F., Wang, H., Tom, M., Orkoulas, G., & Christofides, P. D., 2022. Atomistic-

mesoscopic modeling of area-selective thermal atomic layer deposition. Chemical Engineering Research & Design, 188, 271–286.

[123] Yun, S., Tom, M., Luo, J., Orkoulas, G., & Christofides, P. D., 2022. Microscopic and data-driven modeling and operation of thermal atomic layer etching of aluminum oxide thin films. Chemical Engineering Research & Design, 177, 96–107.

[124] Yun, S., Tom, M., Orkoulas, G., & Christofides, P. D., 2022. Multiscale computational fluid dynamics modeling of spatial thermal atomic layer etching. Computers & Chemical Engineering, 163, 107861.

[125] Yun, S., Tom, M., Ou, F., Orkoulas, G., & Christofides, P. D., 2022. Multivariable run-to-run control of thermal atomic layer etching of aluminum oxide thin films. Chemical Engineering Research & Design, 182, 1–12.

[126] Yun, S., Tom, M., Ou, F., Orkoulas, G., & Christofides, P. D., 2022. Multiscale computational fluid dynamics modeling of thermal atomic layer etching: Application to chamber configuration design. Computers & Chemical Engineering, 161, 107757.

[127] Yun, S., Wang, H., Tom, M., Ou, F., Orkoulas, G., & Christofides, P. D., 2023. Multiscale CFD modeling of area-selective atomic layer deposition: Application to reactor design and operating condition calculation. Coatings, 13, 558.

[128] Zhang, A. & Lieber, C. M., 2016. Nano-bioelectronics. Chemical Reviews, 116, 215–257.

[129] Zhang, Y., Ding, Y., & Christofides, P. D., 2019. Integrating feedback control and run-to-run control in multi-wafer thermal atomic layer deposition of thin films. Processes, 8, 1–18.