Chemical Engineering Research and Design 203 (2024) 1–10

Contents lists available at ScienceDirect

Chemical Engineering Research and Design

journal homepage: www.elsevier.com/locate/cherd

Integrating run-to-run control with feedback control for a spatial atomic layer etching reactor

Henrik Wang [a], Matthew Tom [a], Feiyang Ou [a], Gerassimos Orkoulas [c], Panagiotis D. Christofides [a,b,*]

[a] Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, 90095-1592, USA
[b] Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095-1592, USA
[c] Department of Chemical Engineering, Widener University, Chester, PA 19013, USA

ARTICLE INFO

Keywords:
Run-to-run control
Semiconductor manufacturing
Atomic layer etching
Multiscale modeling

ABSTRACT

Semiconductor manufacturing employs an intricate framework of processes that require accurate design specifications at the nanoscale level. Thermal atomic layer etching fulfills these stringent criteria in an exemplary manner by conducting top-down removal of transistor films to further downsize thicknesses and facilitate transistor densification on wafers. However, it has low productivity. Thus, a spatial sheet-to-sheet reactor is appropriate for achieving high throughput while maintaining substrate quality. In order to continuously regulate the process, a run-to-run (R2R) controller coupled with a continuous feedback proportional–integral (PI) controller is proposed to mitigate a kinetic shift disturbance and a continuous pressure ramp disturbance through a multivariate input correction procedure. A tuning methodology is employed to determine the optimal tuning parameters to enhance the performance of the R2R and PI controller response to their respective disturbances. Results indicate that the combined R2R and PI controller outperforms the sole R2R controller by minimizing the amount of input correction needed to minimize the etching per cycle offset from the setpoint.

1. Introduction

Following the rapid commercialization of electronics, a rising global consumption for semiconductors, which are integral to electronic performance, is generating recurring shortages that can lead to a volatile market that depends on these semiconductors (Voas et al., 2021; Mohammad et al., 2022). In the last two decades, the world has observed semiconductor usage in a variety of applications from smart devices (Lauwers, 2013), biotechnology (Kolahdouz et al., 2022), and computing (Huang et al., 2023), but this overconsumption is problematic for a society that depends on electronics. In addition to their high demand, wafer production comprises over 500 processing steps (Richard, 2023), which reduces productivity due to the precise dimensions and design criteria required to achieve desirable properties: minimal current leakage (Jegadheesan et al., 2020), reduced short-channel effects (Cao et al., 2023), efficient power conversion (Shenai, 2019), and self-alignment behavior to facilitate transistor stacking (Radamson et al., 2020). Particularly, the semiconductor industry has concentrated efforts into optimizing the design of wafer logic components, i.e., transistors, such as metal–oxide–semiconductor field-effect transistors (MOSFETs) with gate-all-around (GAA) designs (Bhol et al., 2022); however, the fabrication for these transistors requires procedures that demand high accuracy in the nanoscale. One fundamental approach uses a top-down method known as atomic layer etching to remove high-κ dielectric oxide films such as Al_2O_3 to allow the size reduction for transistors (Fang et al., 2018).

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 (Huard et al., 2018). 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 (Chiappim et al., 2022) 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

* Corresponding author at: Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, 90095-1592, USA.
E-mail address: pdc@seas.ucla.edu (P.D. Christofides).

https://doi.org/10.1016/j.cherd.2024.01.030
Received 11 January 2024; Received in revised form 16 January 2024; Accepted 17 January 2024
Available online 18 January 2024

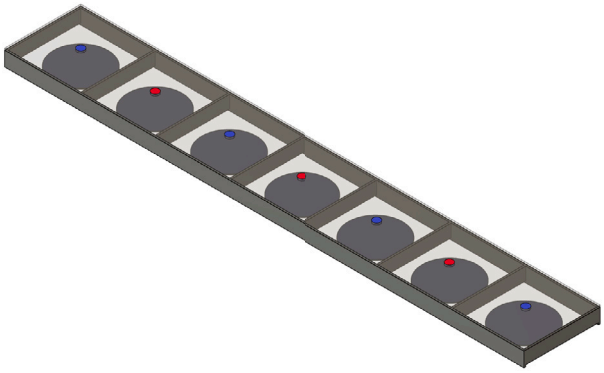


Fig. 1. A three-dimensional depiction of a spatial sheet-to-sheet reactor comprising various zones for HF/TMA exposure and purging.

data is an expensive task that requires numerous cost and materials constraints. Thus, *in silico* multiscale modeling (Li et al., 2013) is an alternative route towards 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 (see Fig. 1) for Al_2O_3 films, that was developed to increase process productivity (Yun et al., 2022b). While this spatial reactor model is a first step towards 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 (Moyne et al., 2018). 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, Yun et al. (2022c) 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 (Derbyshire, 2023). 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 (Yun et al., 2021). 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.

This work is organized as follows: Section 2 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 2.1 and the macroscopic computational fluid dynamics model in Section 2.2 and Section 3 examines the development of the R2R (Section 3.1) and feedback (Section 3.2) control system for the S2S reactor.

2. 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 (Cheimarios et al., 2021) through the use of stochastic kinetic Monte Carlo simulations and computational fluid dynamics. This section will summarize the kinetic Monte Carlo algorithm (Yun et al., 2022a) and the multiscale model (Yun et al., 2022b) drawn from prior works.

2.1. Microscopic modeling

Rudimentary thermal ALE processes are characterized by a general two-step process (Engelmann et al., 2015) comprising an initial modification step followed by an etching step. This procedure is accomplished through the use of a gaseous precursor that adsorbs to the surface in a self-limiting manner and a bulky 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 by George (2020). The thermal ALE of Al_2O_3 uses hydrogen fluoride (HF) as the initial precursor reagent and trimethylaluminum (TMA) 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 2.2. 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 (Jansen, 2012).

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 (Christofides and Armaou, 2006). Yun et al. (2022a)

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 (Bortz et al., 1975). 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:

$$k_{ads}(P_a, T) = \frac{\sigma_a P_a A_{site}}{Z_a \sqrt{2\pi m_a k_B T}} \tag{1}$$

$$k_{nonads}(T) = \frac{k_B T}{h} \exp\left(-\frac{E_A}{RT}\right) \tag{2}$$

where σ_a is the sticking coefficient for the adsorbate (e.g., HF and TMA), 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.

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

(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:

$$\sum_{i=1}^{p-1} k_i \leq \Gamma_1 k_{tot} \leq \sum_{i=1}^{p} k_i \tag{4}$$

(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.

$$\Delta t = -\frac{\ln \Gamma_2}{k_{tot}} \tag{5}$$

This kMC process is illustrated in Fig. 2, which shows the evolution of the sites in the $N \times N$ grid.

Following the development of the kMC model, processing times for achieving similar EPC and surface coverage were validated with experimental findings from George (2020). 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 (Yun et al., 2022a). The results of these investigations were used to develop the standard operating conditions for the S2S reactor.

2.2. Macroscopic modeling

A two-dimensional (2D) spatial, sheet-to-sheet (S2S), reactor was then developed from Poodt et al. (2010) and Roozeboom et al. (2012) through the computer-aided design (CAD) modeling software, Ansys SpaceClaim, comprising HF and TMA injection regions that are spatially separated by adjacent N_2 purging zones in Fig. 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 HF and TMA 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 HF, TMA, and N_2 flowrates and the substrate velocity were tested to determine appropriate conditions that maximized the etching per cycle (EPC) (Yun et al., 2022b).

Table 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

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) (ANSYS, 2022a). The mass, momentum, and energy equations are described as follows:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v}) = S_m \tag{6}$$

$$\frac{\partial}{\partial t}(\rho \vec{v}) + \rho(\vec{v} \cdot \nabla)\vec{v} = -\nabla P + \nabla \cdot \left(\bar{\bar{\tau}}\right) + \rho \vec{g} + \vec{F} \tag{7}$$

$$\frac{\partial}{\partial t}(\rho E) + \nabla\left(\vec{v}(\rho E + P)\right) = -\nabla\left(\Sigma h_j \vec{J}_j\right) + S_h \tag{8}$$

where ρ expresses the density of the gaseous species, $\vec{v}$ denotes the velocity of the gases, P is the system pressure, $\bar{\bar{\tau}}$ represents the rank-two stress tensor, $\vec{g}$ is the gravitational acceleration due to Earth, $\vec{F}$ defines the body force subjected onto the gases, h_j and $\vec{J}_j$ 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 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 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 (ANSYS, 2022b).

2.3. 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

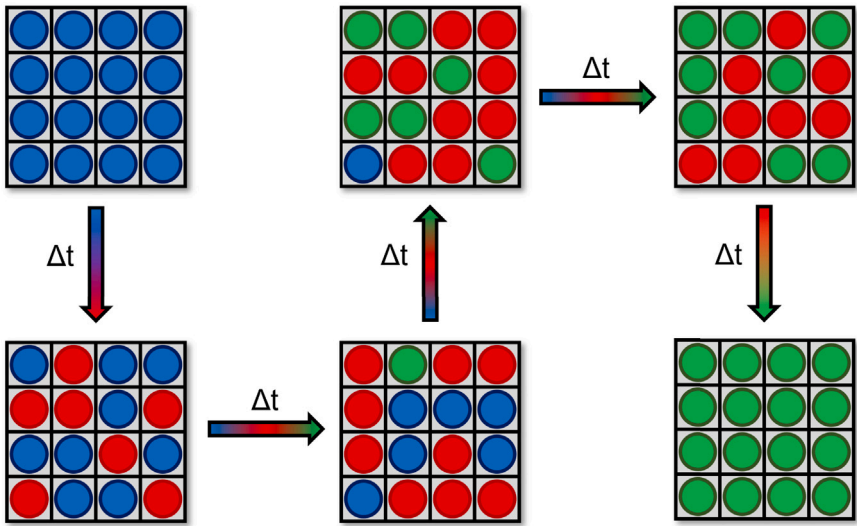


Fig. 2. Graphical representation of the kMC algorithm when conducting the grid approach for the BKL method. Each grid advances to the next grid following the computation of the time update, Δt.

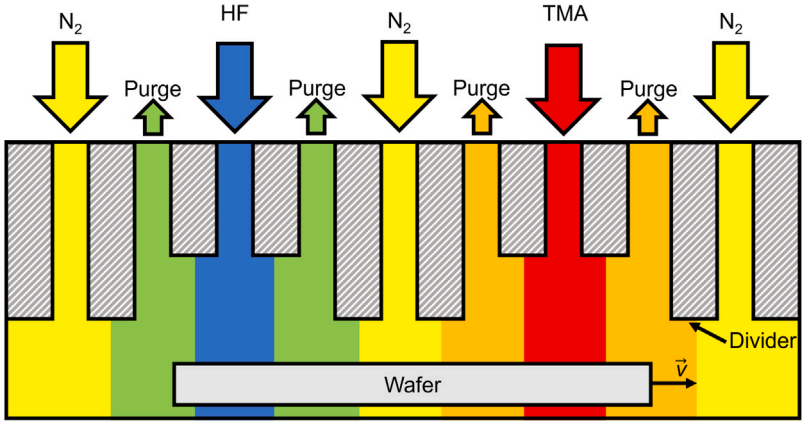


Fig. 3. 2D side projection of the sheet-to-sheet spatial reactor for the thermal ALE of Al_2O_3.
Source: The illustration is adapted from Yun et al. (2022b).

work (Yun et al., 2022b), 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. 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.

Pertinent surface pressure data is extracted from nodal data located on the upper surface of the wafer, which is illustrated in Fig. 5. The pressure field contours illustrate the spatial isolation of the HF and TMA 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.

3. 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 (Lubitz et al., 2014). 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.

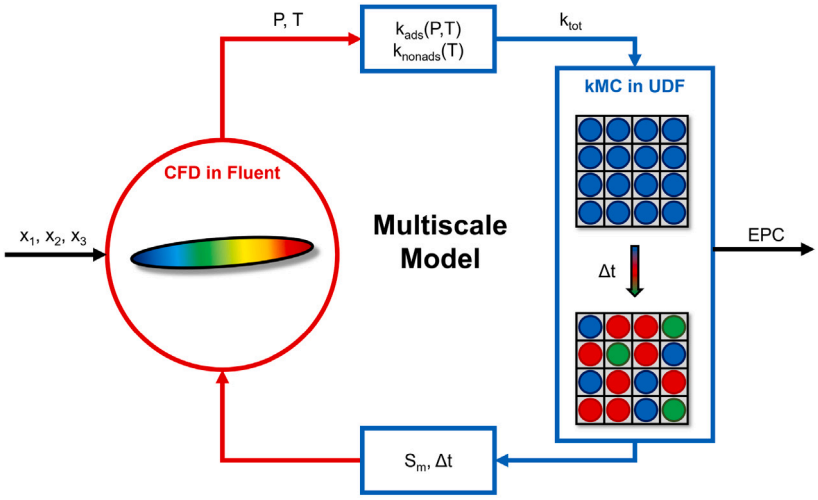


Fig. 4. An illustration of the multiscale simulation that couples the CFD simulation and kMC simulation in Ansys Fluent.

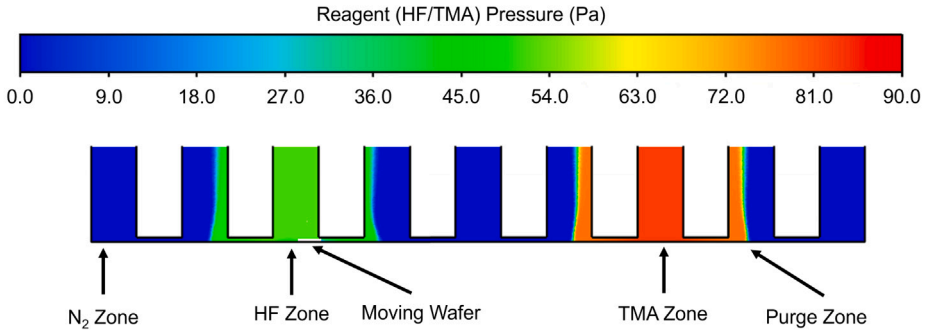


Fig. 5. Surface pressure field data of the 2D S2S reactor at a time of 1.50 s produced from the multiscale simulation.

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 (Coughanowr and LeBlanc, 2009). 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.

3.1. 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 (Butler, 1995). 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 (Andrews, 2022). Sachs et al. (1995) 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. 6.

3.1.1. Linear model

Before the R2R controller is appropriately tuned, a multiple-input-single-output (MISO) linear regression model is generated from 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., Wang and Han, 2013; Yun et al., 2022c). This MISO model is adopted from a prior work (Tom et al., 2022) by assuming negligible Gaussian noise, and relates three manipulated input parameters, the substrate velocity, HF flow rate, and TMA flow rate, to the measured output parameter, etch per cycle (EPC).

$$\hat{y} = \mathbf{B}^T\mathbf{X} + a, \quad \text{where } \mathbf{B} \in \mathbb{R}^3,\ \mathbf{X} \in \mathbb{R}^3,\ a = \alpha + d \tag{9}$$

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

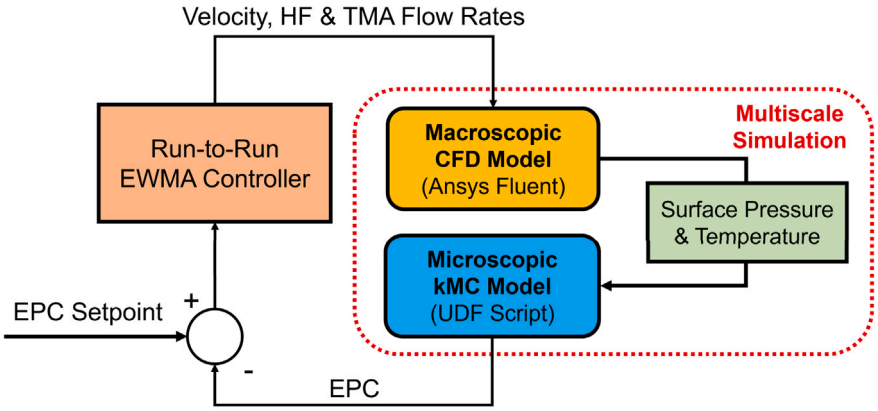


Fig. 6. Process diagram that depicts the conjunction of the R2R controller with the multiscale simulation, where the R2R controller performs input adjustment to the macroscopic CFD model from the calculation of the error between the multiscale model and the target.

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 $\alpha = 0.478$, which is the bias when the system 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 3.1.2. 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. (9) is used in conjunction with an EWMA method to perform manipulated input adjustment, which is elucidated in Section 3.1.2.

3.1.2. 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 (Del Castillo and Hurwitz, 1997). Thus, continuous tuning of the linear regression model described by Eq. (9) is necessary to mitigate these disturbances, which can be accomplished through a translation procedure proposed by Ljung (2010). 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 (Montgomery, 2013). The EWMA method is described by the following expression:

$$a_t = \lambda(y_t - \mathbf{B}^T\mathbf{X}_{t-1}) + (1-\lambda)a_{t-1} \tag{10}$$

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:

$$\min_{\mathbf{B}^T\mathbf{X}_t = c_t} \left\|\mathbf{X}_t - \mathbf{X}_{t-1}\right\|^2 \tag{11}$$

$$\text{s.t. } c_t = \tau - a_t \tag{12}$$

where τ is the target or setpoint of the EPC and $\|\cdot\|$ denotes the l_2 norm. The optimization problem, as derived by Tom et al. (2022) 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:

$$\mathbf{X}_t = \mathbf{X}_{t-1} - \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\left(\mathbf{B}^T\mathbf{X}_{t-1} - c_t\right) \tag{13}$$

3.2. 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 process 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 (Ishikawa et al., 2017). With respect to in-batch feedback control, it is important to note that alternative approaches like data-driven batch control techniques (Chandrasekar et al., 2022) 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.

The Taiwan Semiconductor Manufacturing Company (TSMC) (TSMC, 2024) employs a Micro Electro Mechanical System (MEMS) that measures the deformation of applied pressure in capacitance onto the surface of pressure sensing electrodes (Cheng et al., 2015). These pressure sensors exploit the piezoelectric effect that enables them to change in resistivity when subjected to pressure stresses (Javed et al., 2019). This work considers the role of MEMS for monitoring the surface pressure on the substrate in the HF and TMA 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 used in this work for monitoring and correcting the

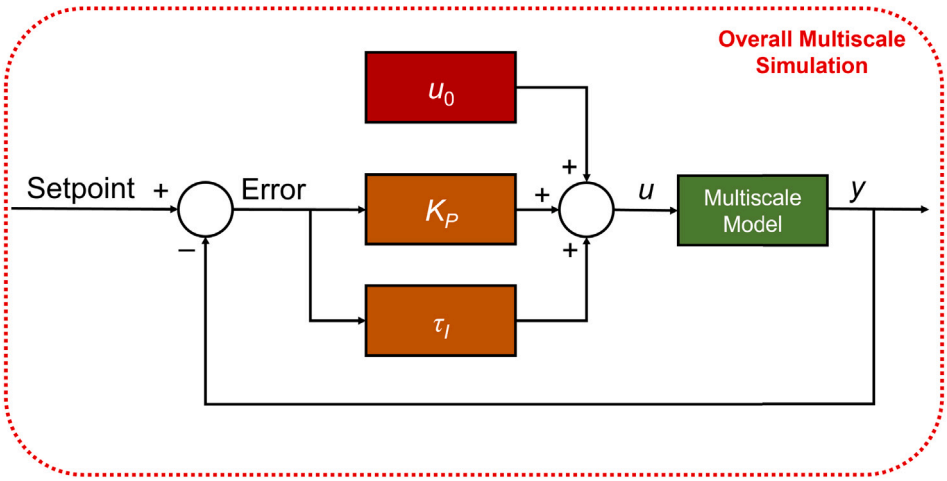


Fig. 7. Process flow diagram depicting the conjoining of the PI controller with the multiscale model to implement correction to the HF and TMA flow rates, u, by accounting for the error between the measured wafer surface pressure y from the target surface pressure. The bias, u_0, is obtained from the R2R controller adjustment of the HF and TMA flow rates.

pressure disturbances and the resulting closed-loop system is displayed in Fig. 7.

PI controllers continuously apply a control action based on the error between the measured pressure and the target pressure, as seen below:

$$u(t) = u_0 + K_0 \cdot K_p \left(e(t) + \frac{1}{\tau_I} \int_0^t e(\tau) d\tau \right) \tag{14}$$

where $u(t)$ is the mass flow rate of HF and TMA 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. (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 (Coughanowr and LeBlanc, 2009). In addition, the PI controller operates in conjunction with the R2R controller, as the latter dictates the starting mass flowrates of HF and TMA 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.

3.3. 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. (2024) 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. (2022c) and Tom et al. (2022) 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 HF and TMA flow rates), or defective and miscalibrated equipment. It is notable that pressure changes influence the rate of adsorption of HF and TMA 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:

$$P_{op} = \begin{cases} P_0 - 50t & \text{for } 0 < t \leq 2 \\ P_0 - 100 & \text{for } t > 2 \end{cases} \tag{15}$$

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.

4. Controller tuning and closed-loop simulation results

4.1. 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 (Oakland, 2003). 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

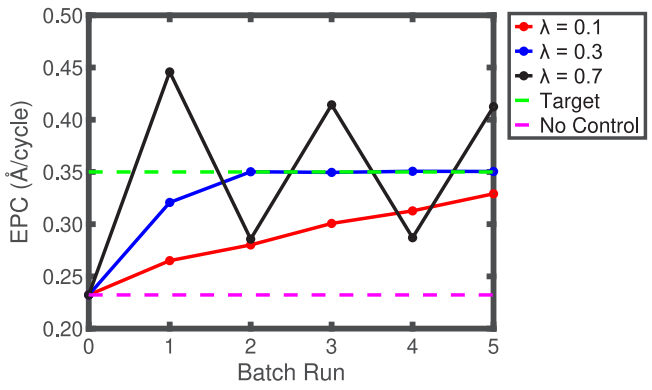


Fig. 8. R2R control plot for various EWMA weights, λ, to determine an optimal weighting parameter that approaches the target EPC in fewer batch runs.

demonstrated in the control plots in Fig. 8. Results illustrate that lower-weight λ requires fewer 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 by Montgomery (2013).

4.2. 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 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. 9.

Results demonstrate that increasing τ_I generally increases the time required to eliminate the offset in Fig. 9(a). However, lower τ_I increases oscillatory behavior due to increased influence by the accumulated error term in Eq. (14) on the PI control action. For the process gain, it is illustrated in Fig. 9(b) that lower K_p requires more process time to reach the setpoint. Thus, for the purposes of this work, $\tau_I = 0.66$ and $K_p = 0.60$ were specified to the PI controller.

Table 2
Comparison of averaged errors over 5 batch runs between the target and measured pressures.

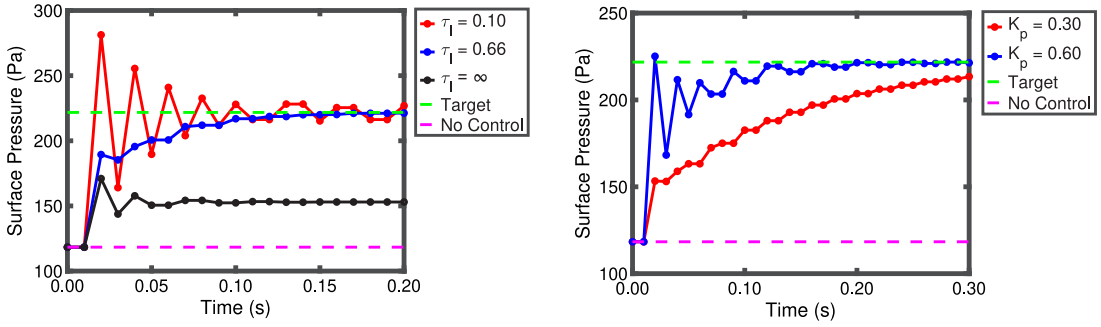
Controller model	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.7$
R2R	0.063	0.025	0.078
PI	0.042	0.042	0.042
R2R & PI	0.019	0.010	0.017

4.3. 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. 10. The measured output, EPC, response illustrated in Fig. 10(a) 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. 10(b)–10(d).

An observable advantage of the conjoined R2R and PI control system is 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 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.

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. 9 show that control actions take less than 0.1 s to be felt on



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

Fig. 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.

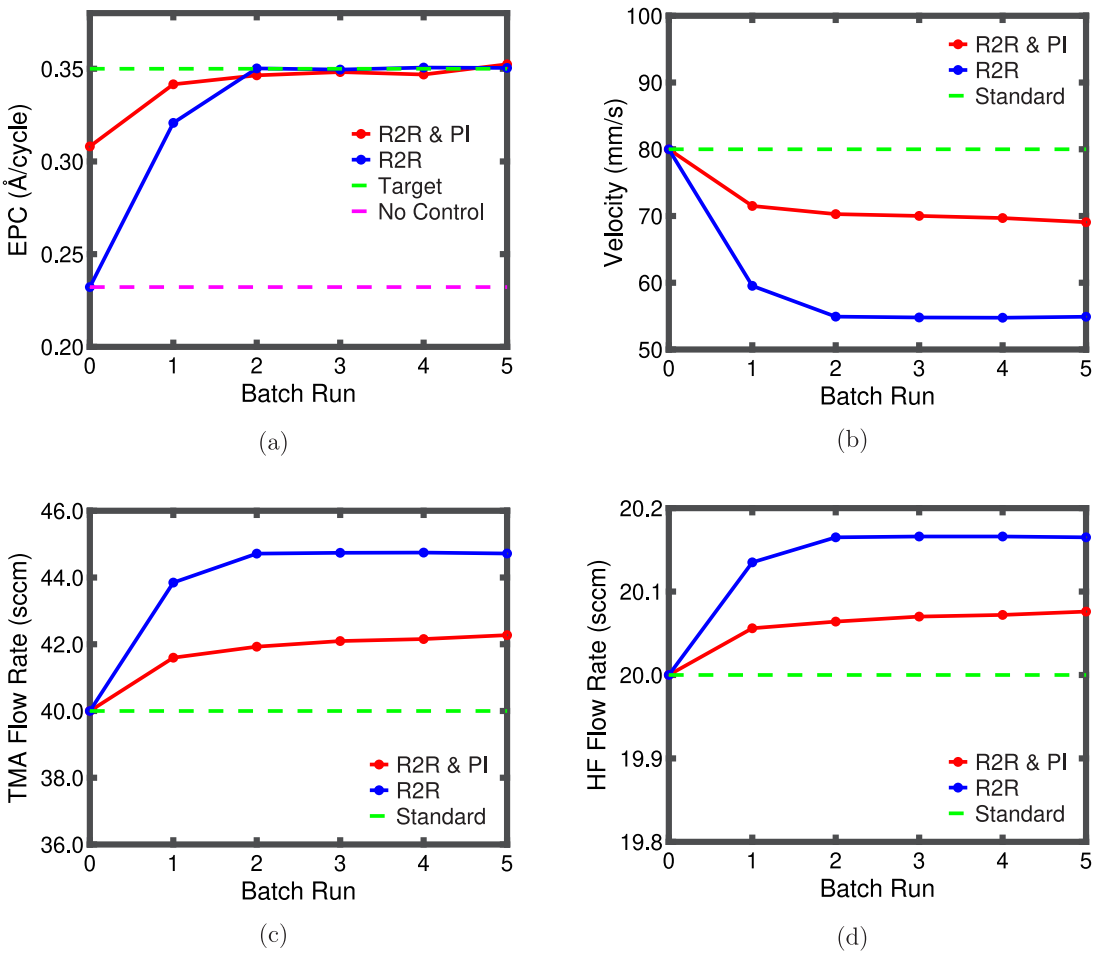


Fig. 10. Comparison of (a) EPC control, (b) substrate velocity, (c) TMA flow rate, and (d) HF flow rate plots for the single R2R controller for a $\lambda = 0.3$ with the combined R2R and PI control system for a $\lambda = 0.3$ and $\tau_I = 0.66$.

the 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.

4.4. 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. 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.

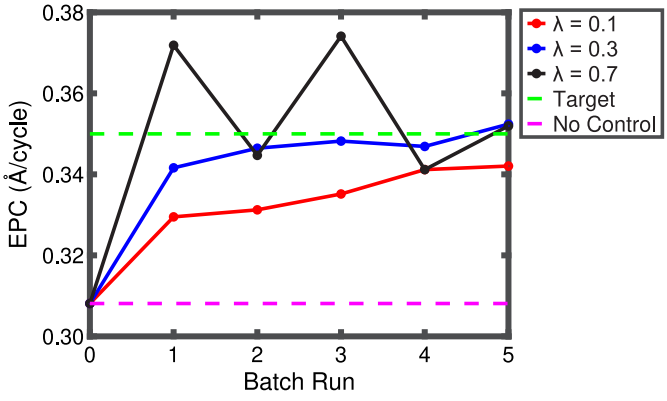


Fig. 11. Comparison of control plots for the combined R2R and PI control system for various EWMA weights, λ to determine an optimal λ that reaches the target in a minimal number of batch runs.

5. 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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Financial support from the National Science Foundation, United States is gratefully acknowledged. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

References

Andrews, M., 2022. Critical manufacturing redefines semiconductor MES. Silicon Semicond. 43, 38–41.
ANSYS, 2022a. Ansys Fluent Theory Guide. ANSYS Inc., Canonsburg, PA.
ANSYS, 2022b. Ansys Fluent User's Guide. ANSYS Inc., Canonsburg, PA.
Bhol, K., Jena, B., Nanda, U., 2022. Silicon nanowire GAA-MOSFET: A workhorse in nanotechnology for future semiconductor devices. Silicon 14, 3163–3171.
Bortz, A.B., Kalos, M.H., Lebowitz, J.L., 1975. A new algorithm for Monte Carlo simulation of Ising spin systems. J. Comput. Phys. 17, 10–18.
Butler, S.W., 1995. Process control in semiconductor manufacturing. J. Vac. Sci. Technol. B 13, 1917–1923.
Cao, W., Bu, H., Vinet, M., Cao, M., Takagi, S., Hwang, S., Ghani, T., Banerjee, K., 2023. The future transistors. Nature 620, 501–515.
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. Comput. Chem. Eng. 161, 107713.
Cheimarios, N., Kokkoris, G., Boudouvis, A., 2021. Multiscale modeling in chemical vapor deposition processes: Models and methodologies. Arch. Comput. Methods Eng. 28, 637–672.
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. J. Micromech. Microeng. 25, 125024.
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.
Christofides, P.D., Armaou, A., 2006. Control and optimization of multiscale process systems. Comput. Chem. Eng. 30, 1670–1686.
Coughanowr, D.R., LeBlanc, S.E., 2009. Process Systems Analysis and Control, third ed. Mcgraw-Hill, Boston.
Del Castillo, E., Hurwitz, A.M., 1997. Run-to-run process control: Literature review and extensions. J. Qual. Technol. 29, 184–196.
Derbyshire, K., 2023. Using ML for improved fab scheduling. https://semiengineering.com/using-ml-for-improved-fab-scheduling/. Accessed 7 January 2024.
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 J. Solid State Sci. Technol. 4, N5054.
Fang, C., Cao, Y., Wu, D., Li, A., 2018. Thermal atomic layer etching: Mechanism, materials and prospects. Prog. Nat. Sci. Mater. Int. 28, 667–675.
George, S.M., 2020. Mechanisms of thermal atomic layer etching. Acc. Chem. Res. 53, 1151–1160.
Huang, A., Meng, S., Huang, T., 2023. A survey on machine and deep learning in semiconductor industry: Methods, opportunities, and challenges. Cluster Comput. 26, 3437–3472.
Huard, C.M., Lanham, S.J., Kushner, M.J., 2018. Consequences of atomic layer etching on wafer scale uniformity in inductively coupled plasmas. J. Phys. D: Appl. Phys. 51, 155201.
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? Japan. J. Appl. Phys. 56, 06HA02.
Jansen, A. (Ed.), 2012. An Introduction to Kinetic Monte Carlo Simulations of Surface Reactions, Vol. 1. Academic Press, pp. 38–119.
Javed, Y., Mansoor, M., Shah, I.A., 2019. A review of principles of MEMS pressure sensing with its aerospace applications. Sensor Rev. 39, 652–664.
Jegadheesan, V., Sivasankaran, K., Konar, A., 2020. Optimized substrate for improved performance of stacked nanosheet field-effect transistor. IEEE Trans. Electron Devices 67, 4079–4084.
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.
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, Springer India, India, pp. 15–24.
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.
Ljung, L., 2010. Perspectives on system identification. Annu. Rev. Control 34, 1–12.
Lubitz, M., Medina, P.A., Antic, A., Rosin, J.T., Fahlman, B.D., 2014. Cost-effective systems for atomic layer deposition. J. Chem. Edu. 91, 1022–1027.
Mohammad, W., Elomri, A., Kerbache, L., 2022. The global semiconductor chip shortage: Causes, implications, and potential remedies. IFAC-PapersOnLine 55 (10), 476–483.
Montgomery, D.C., 2013. Introduction to Statistical Quality Control, seventh ed. John Wiley & Sons, Hoboken.
Moyne, J., Del Castillo, E., Hurwitz, A.M., 2018. Run-to-Run Control in Semiconductor Manufacturing. CRC Press, Boca Raton.
Oakland, J.S., 2003. Statistical Process Control, fifth ed. Butterworth-Heinemann, Oxford.
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. Adv. Mater. 22, 3564–3567.
Radamson, H.H., Zhu, H., Wu, Z., He, X., Lin, H., Liu, J., Xiang, J., Kong, Z., Xiong, W., Li, J., Cui, H., Gao, J., Yang, H., Du, Y., Xu, B., Li, B., Zhao, X., Yu, J., Dong, Y., Wang, G., 2020. State of the art and future perspectives in advanced CMOS technology. Nanomaterials 10, 1555.
Richard, C., 2023. Understanding Semiconductors: A Technical Guide for Non-Technical People. A Press, Berkeley, CA.
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 Conf. Ser. Mater. Sci. Technol. 41, 012001.
Sachs, E., Hu, A., Ingolfsson, A., 1995. Run by run process control: Combining SPC and feedback control. IEEE Trans. Semicond. Manuf. 8, 26–43.
Shenai, K., 2019. High-density power conversion and wide-bandgap semiconductor power electronics switching devices. Proc. IEEE 107, 2308–2326.
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. Comput. Chem. Eng. 168, 108044.
TSMC, 2024. MEMS technology. https://www.tsmc.com/english/dedicatedFoundry/technology/specialty/mems. (Accessed 7 January 2024).
Voas, J., Kshetri, N., DeFranco, J.F., 2021. Scarcity and global insecurity: The semiconductor shortage. IT Prof. 23, 78–82.
Wang, K., Han, K., 2013. A batch-based run-to-run process control scheme for semiconductor manufacturing. IIE Trans. 45, 658–669.
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. Comput. Chem. Eng. 148, 107267.
Yun, S., Tom, M., Luo, J., Orkoulas, G., Christofides, P.D., 2022a. Microscopic and data-driven modeling and operation of thermal atomic layer etching of aluminum oxide thin films. Chem. Eng. Res. Des. 177, 96–107.
Yun, S., Tom, M., Orkoulas, G., Christofides, P.D., 2022b. Multiscale computational fluid dynamics modeling of spatial thermal atomic layer etching. Comput. Chem. Eng. 163, 107861.
Yun, S., Tom, M., Ou, F., Orkoulas, G., Christofides, P.D., 2022c. Multivariable run-to-run control of thermal atomic layer etching of aluminum oxide thin films. Chem. Eng. Res. Des. 182, 1–12.