INTERACTIVE CONTROL ENGINEERING
Principles and Applications

Volume 14 Control in Semiconductor Manufacturing

CHAPTER 1

Introduction: Late Measurement and Run-to-Run Correction

Pages 1–22

Edited by John D. Hedengren

How to cite this chapter. "Introduction: Late Measurement and Run-to-Run Correction," in *Control in Semiconductor Manufacturing*, John D. Hedengren, Ed., vol. 14 of *Interactive Control Engineering: Principles and Applications*. Advanced Process Solutions, LLC, 2026, ch. 1, pp. 1–22.

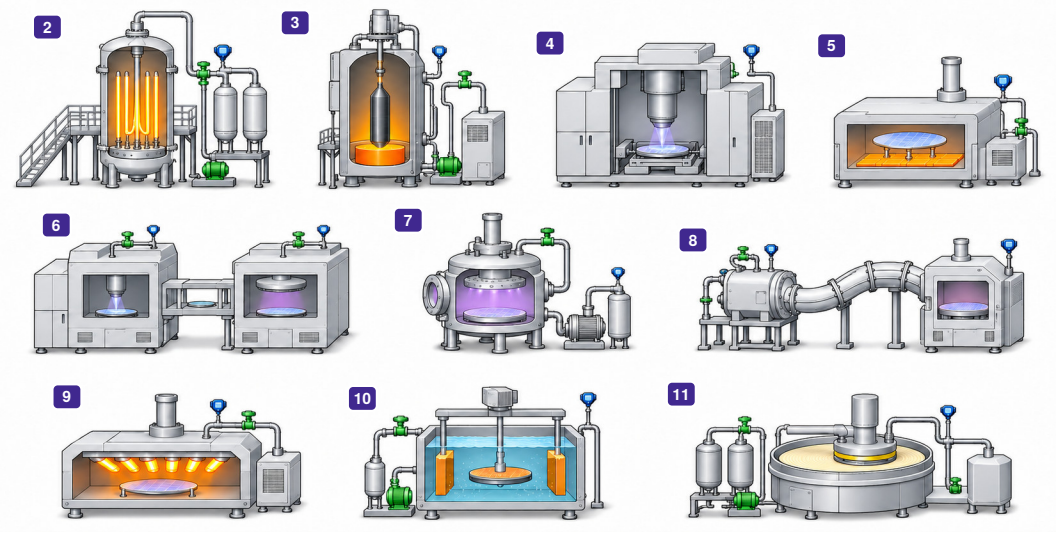
The page numbers in this file are those of the printed volume.

The volume online: contents, other chapters and citation records
https://apmonitor.com/apps/books/v14-semiconductor/

Copyright © 2026 Advanced Process Solutions, LLC. All rights reserved. Produced by Advanced Process Solutions, LLC.

CHAPTER 1

Introduction: Late Measurement and Run-to-Run Correction



The chapters 2 polysilicon reactor 3 crystal puller 4 wafer scanner 5 bake plate 6 lithography and etch 7 plasma etcher 8 ion implanter 9 spike anneal 10 copper plating bath 11 polisher

Notice that the top row makes the silicon and prints the pattern, the middle row transfers it and dopes the wafer, and the bottom row anneals, wires and flattens it. Illustration, not to scale.

IN PLAIN TERMS **Correcting the Next Wafer From the Last One**

The main idea. A chip is built on a thin disk of silicon, called a wafer, in several hundred steps. Most steps leave nothing that can be measured while they run, so the wafer is checked afterward on another machine.

The challenge. That check often comes back after more wafers have been made, and only some wafers are checked at all. Meanwhile the machines change, as their walls get coated, their pads wear and their baths age.

Why it matters. Good control uses each late check to correct the wafers that follow, so that features far smaller than a hair come out the right size.

THREE WAYS TO READ THIS CHAPTER

Explore	8 min	This page, Section 1.2 with its timeline and two History boxes, Figure 1.2, the Did you know? boxes in Sections 1.4.1, 1.4.3 and 1.4.4, and Key points in Section 1.8. No equations.
Study	45 min	The whole chapter in order, working each number of the example M-1, then Exercises 1 to 10.
Practice	5 min	This page, Section 1.4.8 with Table 1.1, Section 1.5 with Figures 1.7 and 1.10, and Section 1.8.

1.1 The Question This Volume Answers

A chip is built on a silicon wafer in several hundred steps, and most of the steps leave nothing that can be measured while they run. The linewidth that an etcher cuts, the dose an implanter delivers and the copper a plating cell lays down are known only after the wafer has left the tool, on a sample of the wafers, and often after other lots have already been processed. The tools themselves drift between those measurements. Chamber walls season, pads wear, baths age and windows warm, and every few days a maintenance event replaces part of the plant. This volume answers one question: how is a step

controlled when the property that matters is measured afterward, late and on a sample, and which techniques recur from one tool to the next?

Seven control problems define the industry, and most chapters meet more than one of them.

- *The product is not measured where it is made.* Chamber sensors read pressure, power and temperature. The wafer's linewidth, thickness or sheet resistance is read on another tool.
- *Measurements arrive late and on a sample.* A reading of one lot returns after later lots have run, and many lots are not measured at all.
- *The tool changes between runs, or while it runs.* Seasoning, wear and aging move the plant from wafer to wafer, and a rod or a crystal that grows during the batch moves it within one run.
- *One wafer, many zones.* A plate, a lamp bank or a carrier head has many actuators that act on one continuous wafer and on each other's zones.
- *A step ends on an event.* An etch is stopped when the film clears, not when a clock runs out.
- *Errors pass from step to step.* A wide line after lithography stays wide after etch unless the etcher is told, so one step can correct the error of the step before.
- *A measurement can be wrong while it looks right.* A pyrometer that assumes the wrong emissivity, a gap voltage that has sparked or a metrology tool with an offset all give readings a loop will act on.

Ten chapters answer the question on ten tools, each with an app that the reader runs in a browser. Chapters 2 and 3 make the material: polysilicon rods that are their own heaters, and a single crystal pulled from the melt. Chapters 4 to 6 pattern it, with a scanner that focuses on a surface it cannot see under the slit, a hundred-zone bake plate and a lithography and etch line run lot by lot. Chapters 7 to 9 etch, implant and anneal, and Chapters 10 and 11 plate the copper wiring and polish it flat.

A reader who has worked through the volume can do the following on a tool that is new to the reader.

- Say what a tool measures, where and when, and which property of the wafer no instrument on it reports.
- Write a run-to-run controller, choose its weight against noise and drift, and find the weight at which a metrology queue makes it unstable.
- Explain why many well-tuned zone loops can overshoot together, and what a model of the zone coupling adds.
- Decide when a step should end on an endpoint signal instead of a timer, and how much over-etch the signal saves.
- Correct one step from a measurement of the step before, and say when the correction lowers the spread.

This chapter develops each of these on one small example and names the chapter that uses it. PID forms and tuning are taken as known from Volume 2, feedforward and cascade from Volume 3, mid-ranging and override from Volume 4, dead time and estimation from Volume 5, interaction and the relative gain array from Volume 6, gain scheduling from Volume 7, run-to-run and batch control in general from Volume 8 and predictive control from Volume 9.

1.2 A Short History

The milestones of Figure 1.1 follow the tools of this volume, from the crystal and the rods that make the silicon to the steps that pattern, etch, implant, anneal, plate and polish it, and to the controllers that correct each step from the one before.

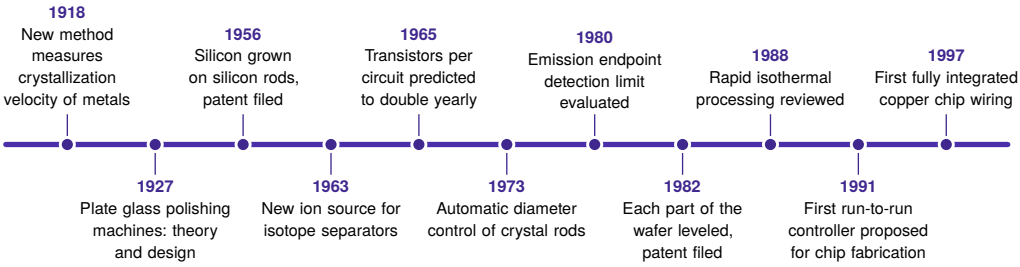


Figure 1.1 Milestones of semiconductor manufacturing and its control. The dates are from [1] (1918), [2] (1927), [3] (1956), [4] (1963), [5] (1965), [6] (1973), [7] (1980), [8] (1982), [9] (1988), [10] (1991) and [11] (1997). A patent is dated by its first filing. Notice that run-to-run control came after most of the tools it now corrects. Events are evenly spaced, not to a time scale.

> ⧖ HISTORY **1918 to 1973: Growing the Silicon**
>
> In 1918 Czochralski published a new method for measuring how fast metals crystallize [1], and the puller of Chapter 3 carries his name. A patent first filed in Germany in June 1956 grew pure silicon on self-supporting rods of silicon, joined at their upper ends so that the current passed up one rod and down the other. The rods' temperature was judged by their glowing color, and the current was raised by hand with a rheostat, because the resistance fell as the rods thickened [3]. A paper published in 1973 held the diameter of a pulled silicon crystal automatically. It weighed the crystal and moved the pull rate, and a second controller trimmed the heater power slowly, waiting and seeing, whenever the pull rate left a set band [6]. Chapter 2 and Chapter 3 control both.

The pace of everything after was set by a prediction. In 1965 Gordon Moore predicted, from a few data points, that the number of transistors on an integrated circuit would double about every year. The trend was later adjusted to a doubling every 18 to 24 months and has continued for more than half a century [5]. Polishing came from an older trade. A study of 1927 on plate glass polishing machines related the rate of removal to the pressure and to the relative velocity of pad and glass, and most models of chemical mechanical planarization are still inspired by it [2]. In 1963 a new ion source was described for electromagnetic isotope separators [4]. Some thirty years later it was still the most widely used ion source, and it was used extensively for the production doping of silicon devices [12].

> ⧖ HISTORY **1980 to 1995: Correcting From the Last Run**
>
> In 1980 a study proposed a measure for detecting the end of a plasma etch from the light of the discharge and evaluated its lower limit of detection for films of polycrystalline silicon and silicon nitride [7]. The step could now end on an event, but its result was still measured afterward. A run-to-run controller, which adjusts the recipe lot by lot or wafer by wafer, was first proposed for chip fabrication in 1991 [10]. A 1993 analysis gave a simple condition on the weight and the estimated process gain that keeps such a controller stable [13]. In 1995 a framework combined statistical process control with feedback. Its rapid mode answered sudden shifts such as those caused by maintenance, its gradual mode the slow build-up of deposits in a reactor, and on an epitaxy reactor the rapid mode recovered within three runs [14].

Patterning changed in the 1980s. Chemically amplified resists were proposed in the early 1980s, when exposure tools moved into the deep ultraviolet and their lamps gave too little light there. Rated at first a laboratory curiosity, the resist went quickly into the production of one-megabit memories [15]. A patent application filed in 1982 explained why a wafer could no longer be set square to the lens from three autofocus points. Larger wafers, new materials and each round of exposure and treatment made its surface uneven, so the height of each part of the wafer had to be detected [8]. In 1988 a

review called rapid isothermal processing, which heats one wafer with incoherent light, a superior alternative to the furnace, and it named more accurate temperature measurement as a need [9]. The wiring changed last. In 1997 the first fully integrated copper wiring of a chip technology stacked up to six levels of copper lines [11], the metal that Chapter 10 plates and Chapter 11 polishes flat.

1.3 Notation and Conventions

A *run* is one wafer or one lot, whichever the controller updates on, and k counts runs. A run-to-run controller applies the recipe u_k and later reads the result y_k against a target y^*. The process gain is b, the gain the controller believes is $\hat{b}$, and their ratio is $\xi = b/\hat{b}$. The weight of an exponentially weighted moving average (EWMA) is λ, and m is the number of runs that pass between a run and the arrival of its reading. Linewidths are in nanometers, temperatures in °C with differences in kelvin, and times in seconds or minutes as the tool runs. The framework's h is its integration step and its a the controller action, so neither is used for anything else here. Loop action follows Section 1.6.2: $a = +1$ is reverse acting and $a = -1$ direct acting. A recipe loop whose output rises when the result is above target, such as more trim for a wide line, is direct acting.

The chapters keep their own symbols, and several collide. λ is an EWMA weight in Chapter 6 (λ_E, λ_L), Chapter 5 (λ_z), Chapter 9 and Chapter 11, an analysis average in Chapter 10, and a set of regularization weights in Chapter 11. m is the metrology delay in Chapter 6 but an exponent in Chapter 9. θ is the chamber wall condition in Chapter 6, the analysis age in Chapter 10 and a vector of pad parameters in Chapter 11, while Chapter 7 writes its wall coverage as θ_w. ρ is a drift rate, a spectral radius, a ramp factor or a normalized radius, depending on the chapter. Chapter 3 and Chapter 10 write the action as γ, and Chapter 10 uses a for a decay factor $e^{-T/\tau}$, not for the action.

1.4 Measuring Late and Correcting the Next Wafer

1.4.1 The Wafer Flow

Figure 1.2 places each chapter's tool in the flow of a wafer [16]. Trichlorosilane deposits electronic-grade silicon on heated rods, and the rods are melted and pulled into a single crystal that is sliced and polished into wafers. Each layer of a chip is then patterned. A scanner exposes a resist, a bake plate finishes its chemistry and a developer opens the pattern, which an etcher transfers into the film beneath. In the front end an implanter places dopants and a lamp-heated chamber anneals them. In the back end a plating bath fills the etched trenches with copper and a polisher removes the excess. A wafer passes through the patterning steps once for every layer, and between the steps metrology tools measure a linewidth, a thickness or a sheet resistance on a sample of the wafers.

> **+ DID YOU KNOW? From Blue Light to Extreme Ultraviolet**
>
> To shrink the features, the light of lithography moved from blue at 436 nm to ultraviolet at 365, deep ultraviolet at 248 and 193, and extreme ultraviolet at 13.5 nm. Light that short is absorbed by every material, air included, so it is shaped only by mirrors coated with up to 100 layers of molybdenum and silicon [5].

The techniques the chapters share are developed below on one running example, a patterning module called M-1, drawn in Figure 1.3. A lot of 25 wafers enters every 30 min. A bake plate of five zones in a row, each with its own heater and thermocouple, bakes the exposed resist, and the developed linewidth is measured on every lot at once. An etcher trims the line for u seconds at 1.0 nm/s and then etches the film beneath, ending the main etch on a timer or on an optical endpoint. The finished linewidth, 40 nm on target, is measured two lots later. Every number of M-1 is declared here or computed from these declarations by the script kept with this chapter's record. The numbers are chosen for clear arithmetic and are not taken from any one chapter.

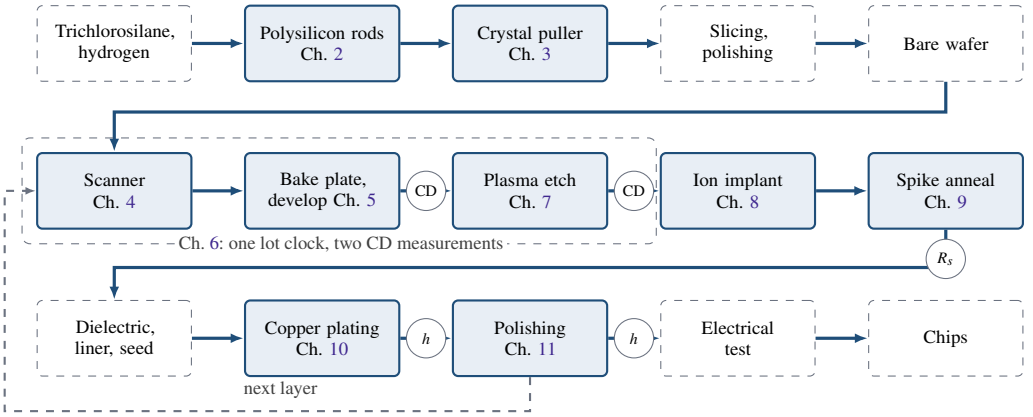


Figure 1.2 The Tools of the Volume in the Flow of a Wafer. Filled Blocks Are the Tools of Chapters 2 to 11, Dashed Blocks Lie Outside the Volume. Circles Mark Measurements on the Wafer After a Step: Linewidth (CD), Sheet Resistance (R_s) and Film Thickness (h). The Dashed Frame Is the Line of Chapter 6, and the Dashed Arrow Repeats the Patterning for the Next Layer.

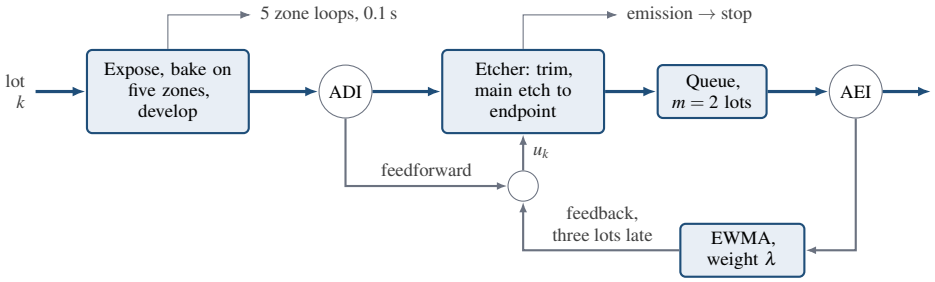


Figure 1.3 The Running Example M-1. The Developed Linewidth (ADI) Is Measured on Every Lot at Once and Can Correct the Etcher's Trim on the Same Lot. The Finished Linewidth (AEI) Is Measured Two Lots Later, so the Feedback Reaches the Trim on the Third Lot After the One It Describes. The Bake Plate Runs Five Zone Loops on a 0.1 s Scan, and the Etcher Ends Its Main Etch on an Emission Signal or a Timer.

1.4.2 Why Little Is Measured in the Chamber

The instruments of a process tool measure the tool. M-1's bake plate reads five thermocouples every 0.1 s and its etcher reads pressure, power and the light of the plasma. None of them reads the linewidth, because the line is tens of nanometers wide and forms under the resist, inside a plasma or under a slit of light that an instrument would block. The linewidth is read on a scanning electron microscope that serves many tools, on a sample of the wafers, after the lot has waited its turn [17]. Measurements therefore come in three kinds: tool sensors in the chamber that are fast and say little about the product, metrology after the step that is slow and says what the product is, and electrical test at the end of the line, weeks later, which no chapter uses.

The clocks of M-1 differ by large factors. One lot of 30 min is 18000 scans of a zone loop, and the after-etch reading of a lot changes the recipe of the third lot after it, 90 min after that lot was etched. A loop on a zone thermocouple corrects within seconds, and a loop on the finished linewidth corrects once a lot on old news. The chapters keep the same split. Chapter 7 closes four chamber loops that settle within 45 s, yet scores each wafer on seven criteria that only the finished wafer shows. Chapter 10 analyzes its bath every 2 h, and each result describes the bath of at least 42 min before. Chapter 11 receives each removal map one or two lots after the wafer was polished, and Chapter 9

measures sheet resistance only after the wafer has left the chamber.

Between metrology results, a model can estimate the product from the tool's own sensors, which the industry calls virtual metrology [18]. Chapter 2 infers the unmeasured center temperature of a growing rod from the supply's volts and amperes, Chapter 10 carries its breakdown products on a model corrected every 8 h by a laboratory, and Chapter 11 keeps a Bayesian estimate of a fresh pad. Estimation in general is the subject of Volume 5.

A fast measurement can also be wrong while it looks right. A pyrometer converts radiance to temperature with an emissivity it is given. Radiance from a wafer at absolute temperature T grows as εT^4, so a total-radiation pyrometer set to $\hat{\varepsilon}$ reads the temperature of Equation (1.1), and a loop that holds the reading on T_{SP} holds the wafer elsewhere.

$$T_{read} = T\left(\frac{\varepsilon}{\hat{\varepsilon}}\right)^{1/4}, \qquad T = T_{SP}\left(\frac{\hat{\varepsilon}}{\varepsilon}\right)^{1/4} \tag{1.1}$$

At a setpoint of 1050 °C a film stack whose emissivity is 10 % higher than assumed is held at 1018.85 °C, 31.15 K low, while the trend shows the setpoint exactly [19]. Chapter 9 builds its pyrometers on this relation and finds the same 31.2 K, and its sheet resistance comes out 26.3 % high on a nitride cap that the pyrometers take for bare silicon. A real narrow-band pyrometer makes about a third of this error, which the chapter also computes.

A metrology offset does the same to a run-to-run loop. If M-1's after-etch microscope reads 0.5 nm high, a loop with integral action settles the reported linewidth on 40 nm and the true one on 39.5 nm, and Chapter 6 measures 0.8 nm of real material removed for an offset of 0.8 nm. A third kind of wrong measurement is one whose physical basis has gone. In Chapter 8 a spark halves the extraction voltage for milliseconds, the current reading stops describing the source, and a gate that marks it invalid cuts the beam overshoot after the spark from 14.75 to 0.29 %.

1.4.3 Run-to-Run Control

M-1's etcher trims the developed line, and the finished linewidth of lot k is $y_k = d_k + b\,u_k + v_k$, where u_k is the trim time, $b = -1.0\,\mathrm{nm/s}$ is the process gain, d_k is everything else the line does and v_k is noise. At the design point, a developed line of 60 nm, a fixed etch bias of 5 nm and 15 s of trim give 40 nm. Nothing measures d_k. A run-to-run controller estimates it from the readings and sets the next recipe so that the estimate lands on target [14], as in Equation (1.2).

$$\hat{d}_k = \lambda\left(y_k - \hat{b}\,u_k\right) + (1-\lambda)\,\hat{d}_{k-1}, \qquad u_{k+1} = \frac{y^* - \hat{d}_k}{\hat{b}} \tag{1.2}$$

The estimate is an exponentially weighted moving average of the disturbance each reading implies, and $0 < \lambda \le 1$ sets how much of each reading it believes. Substituting the recipe into the average gives the incremental form of Equation (1.3).

$$u_{k+1} = u_k - \frac{\lambda}{\hat{b}}\left(y_k - y^*\right) \tag{1.3}$$

It is a discrete integral controller with no proportional or derivative term, sampled once a run, and $\lambda/\hat{b}$ is its integral gain. With $b < 0$ the trim rises when the line is wide, so the loop is direct acting. Chapter 6 writes the same pair of equations for its etch trim and its exposure dose, Chapter 9 corrects its soak time from each wafer's sheet resistance, and Chapter 10 its plating time from each wafer's thickness. Chapter 5 runs a hundred of these loops, one per zone, on the measured linewidth map, and Chapter 11 one per profile mode.

> ✦ DID YOU KNOW? **A Third of the Cost Is Patterning**
>
> Lithography is the most frequently used process in a chip fab and typically accounts for 30 to 35 percent of the total process cost. A linewidth held within 10 % of its target is customarily required of a high-performance process [10].

The weight trades noise against drift [13]. With the gain known and the reading available before the next run, the error $e_k = y_k - y^*$ obeys $e_{k+1} = (1-\lambda)\, e_k$ plus the change in the disturbance between the two runs. Against white noise of variance σ^2 the loop has nothing to correct and passes a filtered copy of the noise into the recipe, so the output variance rises to $2\sigma^2/(2-\lambda)$. At $\lambda = 0.4$ that is 25 % more variance, 11.8 % more spread, than no control at all. Against a drift of δ per run the loop must move the recipe every run, and it can do so only from an error, so it settles with the offset of Equation (1.4).

$$e_\infty = \frac{\delta}{\lambda} \tag{1.4}$$

A drift of 0.05 nm per lot leaves 0.125 nm at $\lambda = 0.4$ and 0.25 nm at 0.2. Chapter 6 measures this residual against the same formula, 0.2500 nm at a drift of 0.075 nm per lot and $\lambda = 0.30$. Chapter 11 adds a slope term for its wearing pad, the double EWMA that removes a steady ramp's offset in theory; the chapter measures how far it reduces the offset on a wearing pad.

The best weight depends on how the disturbance moves. Suppose M-1's d_k is a random walk with steps of 0.2 nm per lot, plus white noise of 0.5 nm. With the variances q of the steps and r of the noise, the error variance under Equation (1.2) with a known gain is Equation (1.5).

$$\sigma_e^2 = \frac{(1-\lambda)^2 q + \lambda^2 r}{1-(1-\lambda)^2} + (m+1)\, q + r, \qquad \frac{\lambda_*^2}{1-\lambda_*} = \frac{q}{r} \tag{1.5}$$

where m is the metrology delay of Section 1.4.4, zero for now. The weight λ_* that minimizes it is the steady gain of a Kalman filter for this disturbance [20]. For M-1, $q/r = 0.16$ and $\lambda_* = 0.328$, which gives a spread of 0.610 nm. Figure 1.4b shows the trade. A weight of 0.1 lets the walk get ahead, 0.688 nm, a weight of 0.8 chases noise, 0.677 nm, and with no drift any weight is worse than none.

1.4.4 Metrology Delay

M-1's after-etch reading of lot k arrives after lot $k+2$ has been etched, so it first changes the recipe of lot $k+3$. A step in d on lot j therefore produces $m+1 = 3$ lots at full error before the recipe moves at all, whatever the weight. If the controller simply adds a fraction of the newest reading's error to its estimate, the closed loop is Equation (1.6), with $K = \lambda\xi$ the loop gain [21].

$$e_{k+1} = e_k - K\, e_{k-m}, \qquad z^{m+1} - z^m + K = 0 \tag{1.6}$$

Placing a root on the unit circle, $z = \mathrm{e}^{j\omega}$, separates the characteristic equation into a phase condition that fixes $\omega = \pi/(2m+1)$ and a magnitude condition that gives the stability boundary of Equation (1.7).

$$K_{max} = 2 \sin \frac{\pi}{2(2m+1)} \tag{1.7}$$

It returns 2, 1, 0.618, 0.445 and 0.347 for $m = 0$ to 4, so the first lot of queue halves the usable gain. The weight and the gain ratio enter only as a product. A controller that believes the trim rate is half the true one runs at twice the loop gain it was tuned for, so $\lambda = 0.4$ with $\xi = 2$ gives $K = 0.8$ and diverges at $m = 2$.

Figure 1.4a shows the cost on M-1 at $\lambda = 0.4$. With no queue the error falls by 0.6 per lot and stays within 10 % of the step after five lots. With $m = 2$ the dominant root is 0.882, the error rings with a period near ten lots, and it stays within 10 % only after 20 lots. At $\lambda = 0.7$, outside the boundary, it grows.

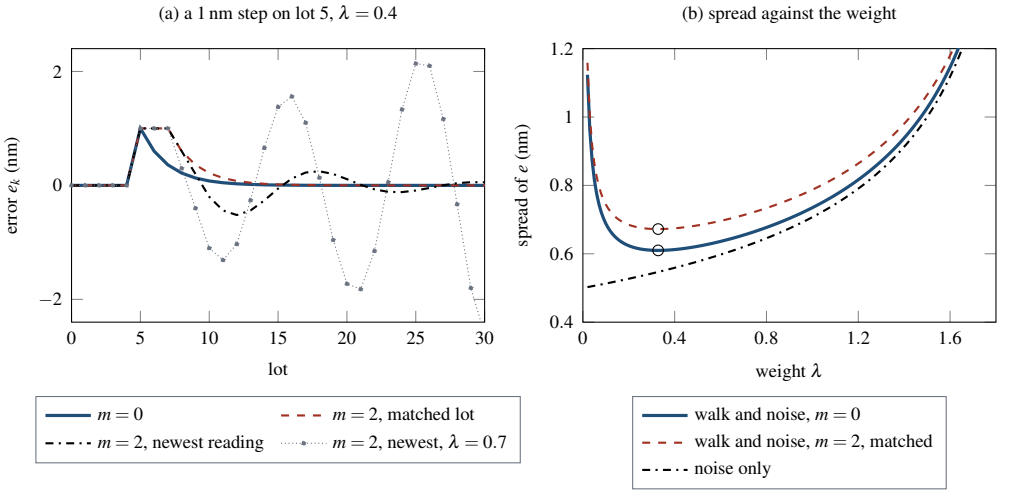


Figure 1.4 Run-to-Run Control of M-1's Trim. (a) The Error After a 1 nm Step With No Queue (Thick Solid), With a Two-Lot Queue and the Matched-Lot Estimator (Dashed), With the Newest-Reading Estimator (Dash-Dotted) and With That Estimator at a Weight Beyond the 0.618 Boundary (Dotted, Dots). (b) The Spread Against the Weight for a Random Walk of 0.2 nm per Lot Plus 0.5 nm of Noise, Without and With the Queue, and for Noise Alone. Open Circles Mark the Best Weight, 0.328.

The boundary belongs to the estimator, not to the plant. A controller that records which recipe made each lot can regress the delayed reading on that recipe, $\hat{d}_k = \lambda(y_{k-m} - \hat{b}\,u_{k-m}) + (1-\lambda)\hat{d}_{k-1}$, and with the gain known the delayed term then cancels. The error still sits at full value for $m+1$ lots, then falls by $1-\lambda$ per lot at any queue length, as the dashed curve shows: inside 10 % after seven lots. Chapter 6 measures both estimators on one line and finds the matched-lot boundary 3.24 times the other at $m = 2$. Two costs remain with the matched lot. The dead band of $m+1$ lots stays, and against a drifting disturbance each lot of queue adds the variance q of one more step, as the $(m+1)\,q$ term of Equation (1.5) states. For M-1 the queue raises the best spread from 0.610 to 0.672 nm without moving the best weight.

Sampling one lot in n_s lengthens the time between corrections by n_s with no change of λ, and Chapter 6 measures a second cost in the feedforward of Section 1.4.7. A reading that describes an old state of the tool must also be compared with the model of that time. Chapter 11 replays each late removal map against the pad state and recipe of the wafer that made it [22], and Chapter 10 compares each analysis with its model at the moment the sample left the bath. Its continuous version of the boundary is a trim gain of 210 for a 42 min analysis age, which would rise to 744 with an age of 2 min. Delay in continuous loops is the subject of Volume 5.

> ✦ DID YOU KNOW? **Most Products Run Only a Few Lots**
>
> A fab runs many products on each tool, and few of them often. In one manufacturing study more than 70 percent of the products had fewer than ten runs [10], so a controller that keeps one estimate per product can wait a long time between its readings.

1.4.5 Multizone Thermal Control

M-1's bake plate has five zones in a row. Each zone holds $C_z = 10\,\mathrm{J/K}$, loses $U = 0.5\,\mathrm{W/K}$ to the surroundings and exchanges $k_x = 1.0\,\mathrm{W/K}$ with each neighbor, so its temperature rise θ_i above ambient obeys Equation (1.8).

$$C_z \frac{\mathrm{d}\theta_i}{\mathrm{d}t} = P_i - U\theta_i - k_x \sum_{j \in \mathcal{N}(i)} (\theta_i - \theta_j), \qquad C_z \frac{\mathrm{d}\theta}{\mathrm{d}t} = \mathbf{P} - \mathbf{A}\theta \tag{1.8}$$

where $\mathcal{N}(i)$ is the set of neighbors and $\mathbf{A}$ is the conductance matrix. Its steady gain is not one number. A power pattern shaped like an eigenvector of $\mathbf{A}$ produces a temperature pattern of the same shape, and for zones in a row with insulated ends the eigenvectors are cosines with j half-waves across the plate. Their gains and time constants are Equation (1.9).

$$G_j = \frac{1}{U + 2k_x[1 - \cos(j\pi/N)]}, \qquad \tau_j = C_z G_j, \qquad j = 0, \ldots, N-1 \tag{1.9}$$

On M-1, with $N = 5$, the uniform pattern has a gain of 2.0 K/W and a time constant of 20 s, and the pattern that alternates from zone to zone has 0.243 K/W and 2.43 s, 8.24 times less, as Figure 1.5a shows. A smooth correction is cheap and a ragged one costs heater power, 0.5 W per kelvin against 4.12 W. The number of zones does not set the resolution a plate can deliver. The ratio of loss to coupling does [23].

A loop on one zone sees a third gain. With its neighbors held at their setpoints by their own loops, the center zone answers its heater with $1/(U + 2k_x) = 0.4\,\mathrm{K/W}$ and a time constant of 4 s. Tuned on that response, with the loop gain $K_c G = 1$ and the integral time equal to the time constant, $K_c = 2.5\,\mathrm{W/K}$ and $\tau_I = 4\,\mathrm{s}$. The tuning is right for a move of one zone, which reaches 90 % in 8.1 s without overshoot. When all five setpoints move together the loops push the uniform pattern, whose gain is five times larger and whose time constant is five times longer, so the loop gain is 5 and the integral time is a fifth of the plant's. The center zone overshoots 19.6 %, as the dash-dotted curve of Figure 1.5b shows. Tuned instead on the uniform pattern, $K_c = 0.5\,\mathrm{W/K}$ and $\tau_I = 20\,\mathrm{s}$, the loops move all five zones without overshoot in 46 s to 90 % but bring a single zone only to 56 % in a minute. Neither tuning is wrong for the step it was tested on. The relative gain of the center zone is 1.77, positive and of order one, so the one-to-one pairing stands, and it says nothing about the direction in which all loops move at once.

Chapter 5 meets exactly this on a plate of a hundred zones, where the uniform pattern has 4.43 times the single-loop gain and a tuning made on one zone overshoots 18.8 % when all zones move. Its remedy is a model of the coupling. The power that a setpoint profile needs in steady state is $\mathbf{A}\theta_{SP}$, one sparse product per scan with no inverse in it, and supplying it as feedforward leaves the loops only a trim. The plate also has a one-sided actuator. A heater cannot cool, so a zone that must come down waits for its own losses. Chapter 9 adds two features of lamp heating. Its lamps light each other's zones, so it decouples with the inverse of a five-by-five gain matrix [24]. Its loss is radiation, so the wafer's time constant falls from 13.7 to 2.4 s between 400 and 1050 °C [25, 26]. The bake plate matters because the resist integrates an Arrhenius rate, and Chapter 5 derives 3.2 nm of linewidth per kelvin of bake [27]. Chapter 11 has the same spatial structure in pressure. Eight carrier chambers act on 33 radial points through broad footprints, so a ragged radial target is out of reach rather than badly tuned.

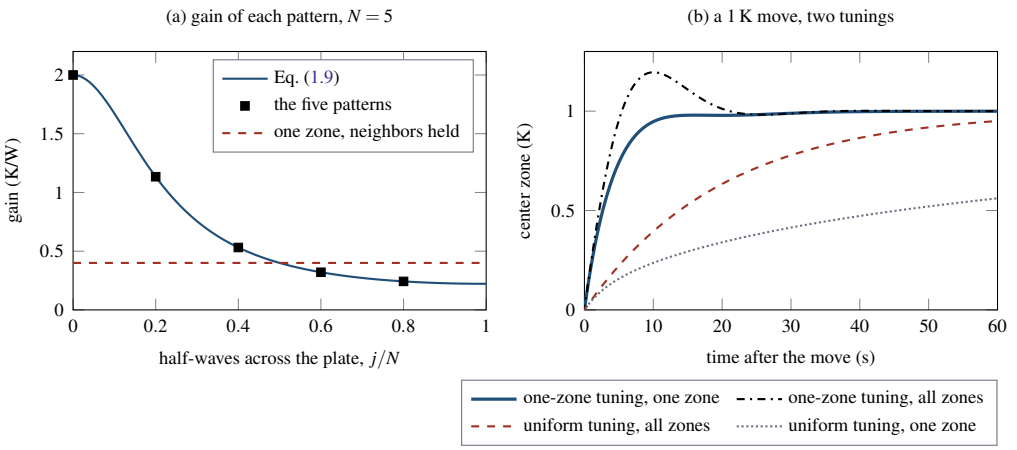


Figure 1.5 M-1's Five-Zone Bake Plate. (a) The Steady Gain of Each Spatial Pattern of Heater Power (Solid Curve, Squares at the Five Patterns) Against the Gain One Zone Sees With Its Neighbors Held (Dashed). (b) The Center Zone After a 1 K Setpoint Move of One Zone or of All Five, Under Loops Tuned on One Zone (Thick Solid, Dash-Dotted) or on the Uniform Pattern (Dashed, Dotted).

1.4.6 Endpoint Detection

M-1's etcher removes a film of $H = 150\,\text{nm}$ at a nominal rate of 3 nm/s. The rate varies by $\pm v = \pm 4\,\%$ across the wafer, so the film clears first where the rate is highest and last where it is lowest. Once a point has cleared, the etch attacks the layer underneath at a tenth of the rate. The main etch must run until the slowest point clears, and any time beyond that cuts into the underlayer where the film is already gone. The wafer passes if every point clears and the underlayer loses no more than 3 nm. From wafer to wafer the rate also moves, here between 0.94 and 1.06 of nominal as a freshly cleaned chamber seasons.

A timed recipe stops at a fixed $t_{end} = (1 + f_o)\,H/r_0$ after the nominal clearing time H/r_0. To clear the slowest wafer it needs an over-etch f_o of 10.8 %, and then the fastest wafer loses 3.32 nm of underlayer and fails. No timed recipe passes both ends. The rate band a timer can serve is $\pm 5.1\,\%$. An endpoint detector instead watches the light of an etch product, which falls as the film clears [7]. For a uniform spread of local rates between $r(1 - v)$ and $r(1 + v)$, half the wafer has cleared at H/r, and a detector that fires there and adds an over-etch f_o in proportion to its own firing time gives the conditions of Equation (1.10).

$$1 + f_o \geq \frac{1}{1 - v}, \qquad L = \frac{H}{S}\left(1 + f_o - \frac{1}{1 + v}\right) \tag{1.10}$$

where L is the underlayer loss at the fastest point and S the selectivity, 10 here. Neither condition contains the rate. The endpoint needs 4.17 % of over-etch to clear every wafer, and the non-uniformity alone sets it. M-1's detector filters the signal with a lag of 0.5 s and fires when about 62 % of the wafer has cleared. At 5 % over-etch every wafer from 0.90 to 1.10 of the nominal rate clears and loses 1.53 to 1.56 nm, as Figure 1.6 shows. The detector is not a feedback loop. It has no setpoint and no actuator, and it decides when the step ends.

Chapter 7 measures the same contrast on a chamber after a wet clean. Three of its four chamber loops hold their setpoints to four decimal places, the etch rate rises 6 % over eight wafers as the walls season, and a timed recipe scraps the first three. With the emission endpoint all eight pass, and the main etch shortens from 93.6 to 87.8 s as the walls season. In its graded cassette the endpoint is the single most valuable change. Other chapters end a step on an event in the same way. Chapter 2 harvests its rods at a diameter, Chapter 8 repaints the part of the wafer whose dose a spark interrupted,

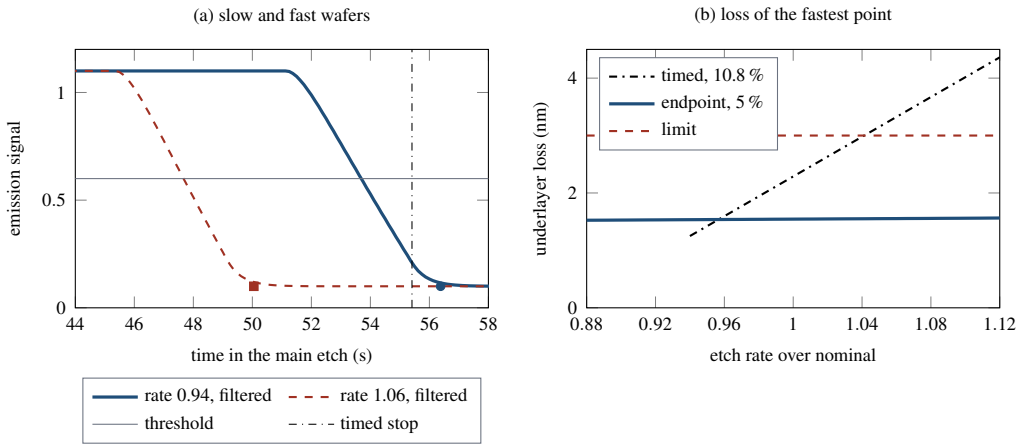


Figure 1.6 Ending M-1's Main Etch. (a) The Filtered Emission Signal of a Wafer Etched at 0.94 of the Nominal Rate (Solid) and at 1.06 (Dashed), the Detector's Threshold (Thin) and the Stop of a Timed Recipe (Dash-Dotted). The Dot and the Square Mark Where the Endpoint Recipe Stops Each Wafer. (b) Underlayer Loss Against the Etch Rate. The Timed Recipe (Dash-Dotted) Leaves Wafers Below 0.94 Uncleared and Exceeds the Limit (Dashed) Above About 1.04. The Endpoint Recipe (Solid) Clears Every Wafer With About 1.5 nm of Loss.

and Chapter 10 calls a dump of the bath when bleeding can no longer reopen its window. Batch endpoints in general are the subject of Volume 8.

1.4.7 Feedforward Between Process Steps

The developed linewidth of an M-1 lot is measured before the lot is etched, so the etcher can cancel the error that lithography left. Suppose lithography varies the developed line from lot to lot with a spread $\sigma_L = 1.0\,\text{nm}$, the microscope reads it with a spread $\sigma_M = 0.4\,\text{nm}$, and the etch adds $\sigma_E = 0.3\,\text{nm}$ of its own. Adding g times the measured deviation, divided by the trim rate, to the trim of the same lot leaves the variance of Equation (1.11).

$$\sigma_y^2 = (1-g)^2\sigma_L^2 + g^2\sigma_M^2 + \sigma_E^2, \qquad g_* = \frac{\sigma_L^2}{\sigma_L^2 + \sigma_M^2} \tag{1.11}$$

Without feedforward the finished spread is 1.044 nm. With $g = 1$ the whole lithography term leaves the expression and is replaced by the microscope's, 0.500 nm, and the best gain $g_* = 0.862$ trims it to 0.477 nm. Full feedforward helps exactly when the measurement is quieter than the process it measures. It acts only on lots that were measured, so measuring one lot in three keeps a third of the benefit, 0.900 nm. Feedback cannot match it. A white lot-to-lot error is gone before any later reading can describe it, and feedback at any weight adds spread to it, as Section 1.4.3 showed.

Chapter 6 implements this path. At an unchanged feedback weight it measures the spread fall from 1.852 to 0.850 nm, below the 1.476 nm of no run-to-run control at all. It also shows that the feedback estimator must regress on the feedback part of the recipe only, or it will undo what the feedforward did. Chapter 11 feeds the measured incoming film profile to its recipe, and Chapter 10 doses its additives in proportion to the charge passed, a feedforward from the load. Two more chapters feed forward within a step. The scanner of Chapter 4 reads the wafer surface ahead of the slit and commands the stage from it [28], with a lead distance equal to the scan speed times the stage's lag, so the right distance moves with the throughput. The bake plate of Chapter 5 adds a power pulse when a cold wafer lands, a disturbance whose time and size are known in advance, and holds the zone integrators while the pulse runs so that feedback does not undo it. Feedforward in general is the subject of Volume 3.

Table 1.1 For Each Chapter, What the Tool Measures While It Runs, What Is Measured on the Product and When, and How the Next Wafer or Lot Is Corrected

Ch.	In the tool, fast	On the product, and when	Next wafer or lot
2	Rod surface pyrometer, supply volts and amperes	None in the batch, center temperature inferred from the supply	Within the batch: gain schedule on diameter, override on inferred center
3	Camera diameter, heater power	No instrument reports the pull rate limit	Within the batch: diameter on pull rate, heater mid-ranges the pull rate
4	Surface ahead of the slit, dose in the slit	Overlay from sampled fields of the wafer	Preview feedforward with a trailing trim, grid fit applied open loop
5	A hundred zone thermocouples, 0.1 s	Linewidth map, every wafer	A run-to-run offset per zone, a pulse at each landing
6	Plate temperature	Developed line at once, etched line two lots late	Dose and trim by EWMA, trim fed forward from the developed line
7	Pressure, density, bias, wafer temperature, emission	Seven criteria on each finished wafer	Main etch ends on the emission endpoint
8	Extraction current every 1 ms, beam current	Dose from the scan ledger	Gated cascade, local repaint of missed dose
9	Five pyrometers, 10 ms	Sheet resistance after each wafer	Soak time corrected wafer to wafer
10	Charge passed, temperature	Additives every 2 h, 42 min old, thickness each wafer	Dosing by charge with a sampled trim, plating time run to run
11	Chamber pressures, pad counters	Removal map one or two lots late, incoming map	Constrained recipe, age-aware EWMA per mode, fresh-pad estimate

1.4.8 What Each Tool Measures and How the Next Wafer Is Corrected

A fab runs these techniques in layers: fast loops on tool sensors inside the step, an endpoint or a supervisor that ends the step, and run-to-run and feedforward corrections between steps on metrology results [29, 30]. Table 1.1 lists, for each chapter, what the tool measures while it runs, what is measured on the product and when, and how the next wafer, lot or run is corrected. Two chapters have no product measurement between runs because the run is one batch. Chapter 2 schedules its gain on the diameter of a rod that grows a hundredfold in area, and Chapter 3 returns its fast, limited pull rate to mid-range with the slow heater. Chapter 12 adds the measured results and compares what each tool measures, when, and how the next wafer or lot is corrected.

1.5 Instruments and Final Elements in This Volume

Every loop of this volume begins at a sensor and ends at a final element, and in a fab the two sit at two levels. Figure 1.7 lists what the controllers of each chapter read and what they move. Inside the tools a temperature is read by a thermocouple or a probe in four chapters and by a radiation pyrometer in two, and a current, a voltage or a charge serves four. Heater, lamp or rod power is moved in five chapters, and a supply that feeds a plasma, an ion source or a laser in three. Five chapters read a measurement made on the wafer after the step, and in each of them a final element is a number in the recipe of a later wafer or lot.

The temperature instruments differ in what they touch. A thermocouple is two wires of dissimilar metals joined at one end, which give a voltage roughly proportional to the temperature of the junction. It is less accurate than a resistance thermometer but rugged, and its most common failure is to burn out open [31]. Each of the hundred zones of Chapter 5 reads one, as drawn in Figure 1.8. A pyrometer, drawn in Figure 1.9, concentrates the radiation of a surface onto a small sensing element and touches nothing. It reveals only the surface, and the only certain way to know an object's emittance is to test its radiation at a known temperature [31]. That is the error of Section 1.4.2, met in Chapter 2 and Chapter 9.

The product is read elsewhere. A linewidth, a thickness or a sheet resistance is measured on a metrology tool that serves many process tools, and the correction it causes is written into a later

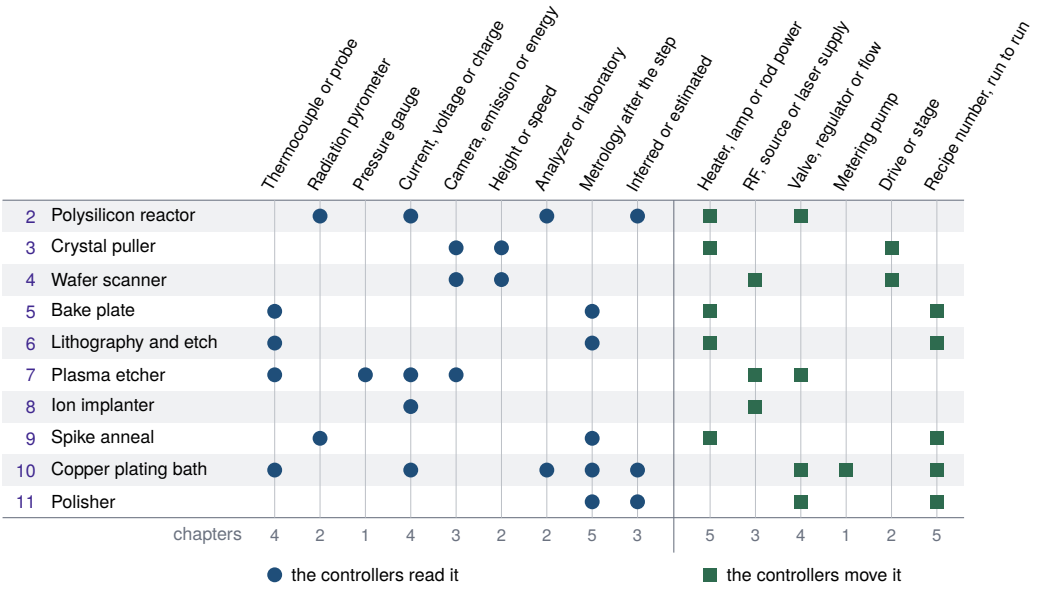
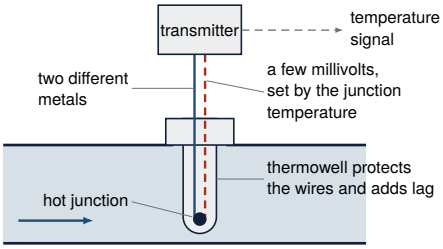


Figure 1.7 Sensors and final elements by chapter. A round mark means that the chapter's controllers read the device, a square mark that they move it. A mark is set only where the chapter's variables table, its instrument text or its section on technology in practice names the device or its reading, so a blank says that the model leaves the device out and not that a real tool lacks it. Notice that the five chapters that read metrology after the step are the five whose final element is a recipe number.



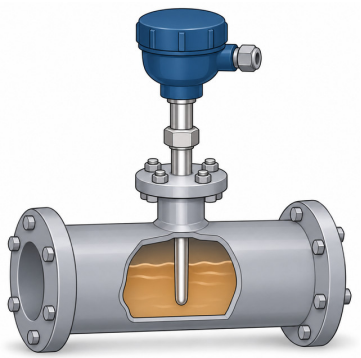


Figure 1.8 Principle of a thermocouple in a thermowell, with a view of the device as installed. Notice that the reading is the temperature of the junction, so a junction that sits in a plate reads the plate and not the wafer on it.

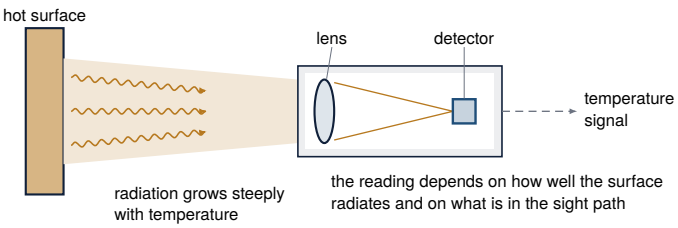


Figure 1.9 Principle of a radiation pyrometer. Notice that the instrument converts radiance to temperature with an emissivity it is given, so a film that changes the true emissivity moves the wafer while the trend still shows the setpoint.

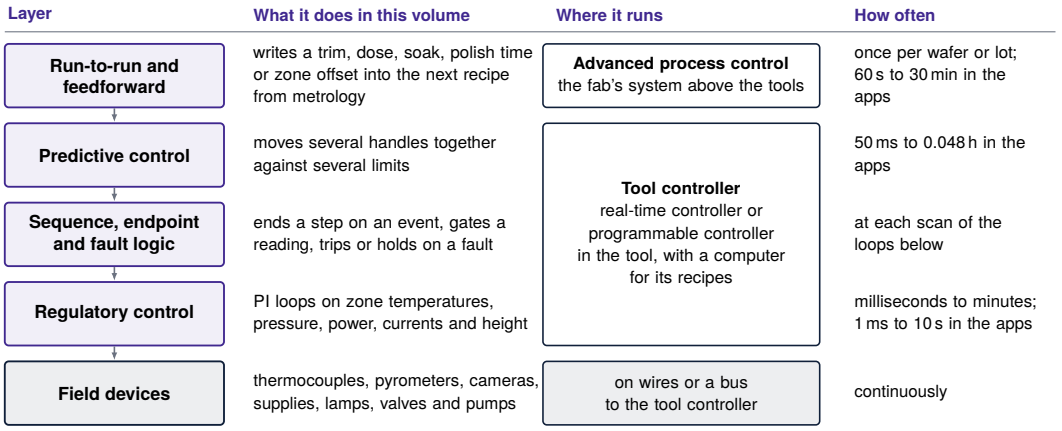


Figure 1.10 The control layers of this volume, the platform that hosts each and how often each acts. The range of loop periods is that of [32], the advanced process control system that of [10], and the ranges in the apps those of the chapters. Notice that the recipe layer acts once per wafer or lot, while the loops below it act many times within one step. In the polysilicon reactor, a plant rather than a fab tool, the predictive controller sits at the supervisory level above the loops (Section 2.5.4).

Table 1.2 What Every App Provides. Standalone Apps Are Described in Their Own Chapters.

Element	What it does
Run, Pause, Reset	Start, hold and restart the run. Reset keeps the reader's settings.
Speed	Simulated time per second of real time. It changes the pace and never the result.
Manual and Auto	Hand each actuator to the reader or to the controller, without a bump.
Scenario list	Numbered cases, each with a briefing, a completion rule and a reload button.
Control panel	Setpoints, manual outputs, tuning, options and disturbances, at a Beginner or an Advanced level.
Trends	SP and PV above the MV for each loop, with limits, event marks and a cursor.
Scorecard	IAE, overshoot, settling time, MV travel and time at a limit for each loop.
Learn card	Predict, Try It, two hints and Explain, opened by the reader in that order.
Show Model	Equations, live variable values, editable parameters and assumptions.
Export CSV	Every logged sample of the run as a text file with a header row.
Instructions	A separate page with background, an operating guide and the model's limitations.

recipe. Run-to-run control of this kind adjusts the recipe lot by lot or wafer by wafer, and most fabs combine it with fault detection and classification in an advanced process control system [10].

Figure 1.10 places the layers of the volume on their platforms. Control loops in operational technology run with cycle times from milliseconds to minutes, many on real-time operating systems [32]. The apps run their regulatory loops every 1 ms in Chapter 4 up to every 10 s in Chapter 10, their predictive controllers every 50 ms in Chapter 9 up to every 0.048 h in Chapter 2, and their recipe corrections once per wafer or lot, from every 60 s in Chapter 9 to every 30 min in Chapter 6.

1.6 The Simulator Framework

Every chapter of this book has an app that runs in a web browser, collected with the others at https://apmonitor.com/apps/. About 240 apps exist across the series. All but a few are built on one shared framework, so the engine, the PID algorithm, the trends and the data export behave the same way in every chapter. The exceptions are standalone studios, such as the cascade studio and the two Temperature Control Lab apps, which follow the same ideas in their own code. This section describes the framework once and states where the standalone apps differ. Table 1.2 lists what a reader meets.

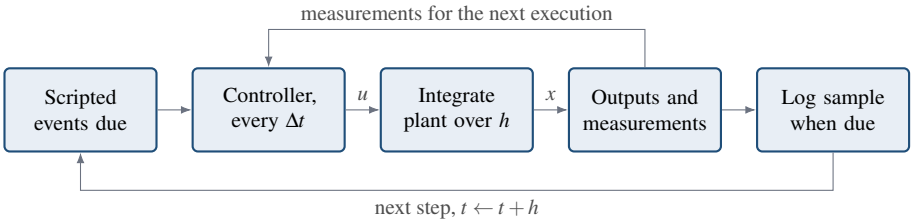


Figure 1.11 Work of the Engine in One Integration Step

1.6.1 The Simulation Engine

Each app carries a plant model written as ordinary differential equations $\frac{dx}{dt} = f(x,u,d,p)$ in the states x, the inputs u, the disturbances d and the parameters p. The engine integrates the model with the classical fourth-order Runge–Kutta method on a fixed step h set for the fastest dynamics of the app. Some apps supply their own step, such as explicit Euler for a discontinuous right-hand side or an exact update for a model driven by events.

Three clocks run at separate rates. The plant advances every step h. The controller executes every control period Δt, equal to h unless the app declares a slower controller, and holds its outputs between executions. The log takes one sample at a fixed interval of simulated time for the trends, the scorecard and the exported file. The order of work in one step is shown in Figure 1.11. The controller also executes once when a scenario is loaded, so every run starts from an initialized controller.

Drawing the screen is kept apart from these clocks. A speed of 5 means five units of the app's time base per second of real time. Speed, frame rate and pauses change how fast a run is shown and never what it computes, so a phone and a workstation produce the same trajectory and the same file.

Noise comes from a deterministic pseudorandom generator. Its seed is declared by the app, or by a scenario that needs its own noise record, and it is restored whenever a scenario is loaded or Reset is pressed. The framework gives the reader no seed setting, so *default seed* in this book means the seed the app ships with. A run repeats bit for bit when the same scenario is loaded and the same settings are made before Run. A change made by hand during a run lands on a different step each time, which is one reason numeric answers carry a tolerance. The standalone simulators draw noise from the browser's own generator, so a noisy run there repeats in its statistics and not sample for sample.

Run and Pause start and hold the clock. Reset restarts the run at $t = 0$, paused, and keeps the reader's settings: every setpoint, tuning value, option, model parameter, disturbance slider and mode the reader changed is put back before the plant and the controller are initialized. Scripted events replay, and a one-shot disturbance button pressed during the run is cleared. The reload button beside the scenario list restores the scenario's own settings instead.

1.6.2 The Velocity-Form PID Algorithm

One PID implementation serves the PID loops of every framework app. It is the velocity (incremental) form: each execution adds the change in the proportional and derivative terms to a running sum, and the integral is a separate term, as shown in Equations (1.12) to (1.15) for execution k and control period Δt.

$$e_k = a\,(\mathrm{SP}_k - \mathrm{PV}_k), \qquad P_k = K_c\,e_k, \qquad D_k = -a\,K_c\,\tau_D\,\frac{\mathrm{PV}^f_k - \mathrm{PV}^f_{k-1}}{\Delta t} \tag{1.12}$$

$$S_k = S_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \tag{1.13}$$

$$v_k = u_b + S_k + I_{k-1}, \qquad u_k = \mathrm{sat}(v_k) \tag{1.14}$$

$$I_k = I_{k-1} + \left[\frac{K_c}{\tau_I}\,e_k + \frac{u_k - v_k}{T_t}\right]\Delta t \tag{1.15}$$

where $a = \pm 1$ is the controller action, PV^f is the measurement after an optional first-order filter, u_b is the output bias, S is the running sum of proportional and derivative increments, v is the output before limiting, u is the output sent to the actuator, I is the integral term and T_t is the tracking time constant, equal to τ_I unless a loop sets another value. The function sat clamps v_k to the actuator range and then applies a rate limit if the loop declares one.

The limit acts on the emitted output u_k only. The sum S_k is never clamped, so it equals $P_k + D_k$ and the output before limiting is the position form $u_b + K_c e_k + D_k + I_{k-1}$. Anti-windup touches the integral alone. While the output is limited, the last term of Equation (1.15) drives I toward the value that reproduces the output applied, which is back-calculation [33]. With no integral action, Equation (1.15) is skipped. Clamping the running sum would instead discard each increment that crosses a limit and keep each one that returns, so symmetric noise would push the output off the limit even in a proportional-only loop.

The derivative acts on the measurement, so a setpoint step gives a proportional kick and no derivative kick. A change of K_c acts on the whole error at the next execution, as in the position form. The action $a = +1$ suits a process whose PV rises with the MV, a reverse-acting controller in the ISA convention, and $a = -1$ gives the direct-acting one.

In Manual the controller keeps executing and sets $I_k = u_{\mathrm{man}} - u_b - S_k$, so the output it would send equals the manual value u_{man}. The first automatic output differs from it by one period's change in P and D. In Auto the engine copies each controller output to the manual station, so a switch to Manual is also bumpless. The standalone studios with PID loops use velocity-form code of their own, with the limit on the output and anti-windup on the integral alone, in which a gain change applies only to later increments.

1.6.3 The Predictive Control and Estimator Options

More than 100 apps offer model predictive control (MPC) as an option beside the structure they open with, which stays the default. Volume 9 develops the theory. In most of them the controller plans with a small linear model in discrete time as shown in Equation (1.16).

$$x_{k+1} = A\,x_k + B\,u_k + E\,w_k + c, \qquad y_k = C\,x_k \tag{1.16}$$

where w is a measured or scheduled disturbance and c is a constant left by linearization. The model comes from a step test at the design point, from the app's balances linearized there or again at every decision, or from the model the reader fits. It never reads the true states or parameters of the run, so a parameter the reader changes in Show Model is model mismatch, as on a real plant.

At each decision the controller minimizes the objective of Equation (1.17) over a horizon of N steps.

$$J = \sum_{k=1}^{N} (y_k - r_k)^{\top} Q\,(y_k - r_k) + \sum_{j} \Delta u_j^{\top} R\,\Delta u_j + \sum_{i} \rho_i\, s_i^2 \tag{1.17}$$

where r is the setpoint, Q weights the tracking error, R is the move suppression on each input change Δu_j and ρ_i is the price of a slack $s_i \geq 0$ that lets an output band be broken when it cannot be held. Actuator limits are hard constraints. The quadratic program is solved in the browser by a dual active-set method [34]. The first move is applied, and the plan is made again at the next decision.

The plan starts from an estimate of the model state. About half of these apps use a steady-state Kalman filter on the model with an added disturbance state, which absorbs what the model did not predict and removes steady offset without an integral term [35]. Others correct the model output with a filtered bias to the same end. About a dozen apps also offer moving horizon estimation (MHE), which solves the estimation problem over a window of recent samples and can hold bounds that a Kalman filter cannot, such as a concentration that is never negative [36].

With the MPC selected, the tuning fields of the loops it replaces leave the Control panel. In most of these apps the Advanced level shows up to three settings, each a multiple of the shipped value: the prediction horizon, the move suppression and the band price. The trend draws the accepted plan

ahead of the present time, and a list of constraint prices shows which limits bind. A plan that is not optimal, not finite or outside its limits is rejected, and a fallback, in most apps the shipped controller, holds the actuator until a plan succeeds. The Model Fitting and Tuning Studio builds the same kind of controller from the reader's own data in its tab, MPC from the fit.

1.6.4 Scenarios, Progress and Data

An app teaches through scenarios, the numbered cases at the top of the Control panel, nine to thirteen in most apps. Loading one sets the model parameters, the controller settings, the disturbances, the mode, the initial state and any scripted events. The information button beside the list opens a briefing in four parts: what the case is, why it matters, what to do and one detail for the practicing engineer. The Learn card holds the lesson back, with Predict and Try It before the run, hints on request and Explain afterward.

Each scenario has a completion rule that the app judges from the run on the simulated clock: a run length, the scripted events passed, a setting made by the reader, a state held for a stated time or a target met. Choosing a case earns nothing. A completed case is marked in the list, and a badge row shows the count and a level, Explorer after one case, Practitioner at half and Completed at all. Progress is kept in the reader's browser with no account. A certificate of progress or of completion can be issued from it, as a record of self-paced work and not an assessed credential.

One case, usually the last, is the graded challenge. It states a target, such as an error band to hold through a sequence of upsets. Changing the model parameters or the disturbances voids the attempt in most apps, and a record made with Reveal suggested tuning says so.

Show Model opens the equations, live values of the variables, the editable parameters and the assumptions. The trends plot SP and PV above the MV for each loop, with actuator limits and marks at setpoint moves, disturbances, mode changes and scripted events. The scorecard reports for each loop the integral of the absolute error (IAE), overshoot, settling time, MV travel and share of time at a limit, with the app's own measures, and it does not combine them into one score. IAE and travel accumulate from Reset, so two runs are compared at the same clock reading.

Export CSV saves the log as a text file with a header row of column names that carry their units and one row per logged sample. In most apps the time column is in seconds, minutes or hours according to the run length. The log holds the most recent 60 000 samples unless the app sets another size. Apps with the MPC option begin the file with comment lines that start with #, which a script must skip.

An app exercise in this book names the scenario by its number and name as the app shows them, every setting that differs from the scenario's defaults, the run length in simulated time, the seed (the default seed unless stated), what to read and where (a trend, a scorecard tile or a column of the exported file) and a tolerance on a numeric answer. Settings are made before Run unless the exercise gives a time.

1.7 Working With AI Agents While Studying

A reader of this book is likely to have an AI assistant within reach, and the apps make its work easy to test. This section gives the editor's view of how to use one while working through a chapter. In short, the assistant may do the algebra and the scripting, and the reader keeps the checking.

What to delegate. Tasks with a clear input and a result that can be tested are good candidates. An assistant can derive the linearization of a chapter's balance at an operating point and the transfer function that follows from it. It can write a short script that reads the exported CSV file and fits a first-order-plus-dead-time model to a step test. It can propose a first tuning from that fit with a named rule, convert units, explain a term in a model or outline a report. The assistant knows nothing about the app until it is told. Give it the equations and parameter values from Show Model, the column names of the exported file and the scenario that produced the data, as described in Section 1.6.4. Ask it to state its assumptions and the units of every result.

Table 1.3 Roadmap of the Volume

Ch.	Plant	The lesson	Level	Time
2	Siemens polysilicon reactor	A gain that falls a hundredfold is scheduled on its cause, the diameter.	introductory	3 h
3	Czochralski crystal puller	A fast, limited actuator is returned to mid-range by a slow one.	intermediate	3 h
4	Wafer scanner	Where nothing can be measured, the command leads the process.	intermediate	2.5 h
5	Hundred-zone bake plate	Zone loops tuned one at a time overshoot together.	intermediate	2.5 h
6	Lithography and etch line	The queue, not the plant, limits the recipe loop.	introductory	2.5 h
7	Plasma etch chamber	The recipe and the endpoint matter more than the loops.	introductory	2.5 h
8	Ion implanter source	A measurement can stop meaning anything, and the loop must know.	advanced	2.5 h
9	Spike anneal chamber	A perfect trend can sit on the wrong wafer temperature.	advanced	2.5 h
10	Copper plating bath	A two-hour analysis limits the trim, so the dose follows the charge.	intermediate	2 h
11	Chemical mechanical polisher	A pad change decides what the estimator may still believe.	advanced	2.5 h

What to check, and how. An assistant returns a fluent answer whether or not the answer is right, so each result needs a test that it could fail. Four tests cover most of the errors a chapter will produce.

- *Units.* Every term of an equation carries the same units, and a process gain is in PV units per MV unit. A gain computed with the valve as a fraction is 100 times the gain of a loop whose valve is in percent.
- *Signs.* Decide from the physics which way the PV moves when the MV rises. Compare that with the sign of the fitted gain and with the controller action of Section 1.6.2.
- *Steady state.* Set the derivatives to zero. The model must return the operating point that Show Model reports, and a linearization must give zero deviation there.
- *A run on the app.* Load the named scenario at the default seed, enter the proposed numbers and run. The trend and the scorecard decide. A tuning that reads well on paper and holds the valve at a limit for half the run has failed.

A fitted model gets one more test: compare it with data it was not fitted to, such as a second step in the opposite direction. When a check fails, report the failure to the assistant with the evidence and ask for the cause, then check the correction the same way.

How the exercises train the habit. One exercise in each chapter is marked *Review the agent*. It presents a tuning, a model or a claim as an AI assistant might propose it, with a flaw, and asks the reader to find the flaw in about five minutes. The flaws are ordinary ones: a unit, a sign, a steady state that does not close, a result measured on the wrong scenario or at a different run length, or a limit that was ignored. Each is found with one of the tests above. The purpose is not distrust. An engineer who can say why an answer is right can hand an assistant much larger tasks and still stand behind the result. That larger picture, in which agents do most of the routine work and the engineer supervises and checks, is the subject of Chapter 12.

1.8 Chapter Roadmap

Table 1.3 lists the chapters with the lesson each adds to this one. The level says how much of Section 1.4 a chapter assumes. The time is an estimate for reading the chapter once and running each case of its app once, at about eight minutes a case.

Chapters 2, 6 and 7 form a first course: a schedule on a growing rod, a recipe loop through a metrology queue, and a step that ends on an endpoint. They rest on Sections 1.4.3, 1.4.4 and 1.4.6. Chapters 3, 4, 5 and 10 add mid-ranging, preview feedforward, spatial modes and a sparse late analysis. Chapters 8, 9 and 11 combine a measurement that can mislead with fast dynamics or a consumable that is replaced, and are best left until last. A reader who follows the wafer can read the chapters in their order, which is the order of Figure 1.2.

KEY POINTS

- Most steps of chipmaking leave nothing that can be measured while they run, so the property that matters is measured afterward, late and on a sample.
- A run-to-run controller is a discrete integral controller sampled once a run, and its weight trades noise against drift.
- A metrology queue halves the usable gain at its first lot, but the boundary belongs to the estimator, not to the plant.
- Zone loops tuned one at a time can overshoot when all zones move together.
- An endpoint needs an over-etch that the non-uniformity alone sets, while a timer must also cover the spread of the etch rate.

Where else this applies. Run-to-run and batch control in general return in Volume 8, dead time and estimation in Volume 5, feedforward in Volume 3 and interaction in Volume 6.

References

[1] J. Czochralski, "Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle," *Zeitschrift für Physikalische Chemie*, vol. 92U, no. 1, pp. 219–221, 1918.

[2] K. S. Gokhale and B. M. Moudgil, "Particle technology in chemical mechanical planarization," *KONA Powder and Particle Journal*, vol. 25, pp. 88–96, 2007.

[3] H. Schweickert, K. Reuschel, and H. Gutsche, "Production of high-purity semiconductor materials for electrical purposes," U.S. Patent 3,011,877, filed 11 June 1957 (priority Germany, 25 June 1956), 1961, granted 5 December 1961.

[4] J. H. Freeman, "A new ion source for electromagnetic isotope separators," *Nuclear Instruments and Methods*, vol. 22, pp. 306–316, 1963.

[5] M. Steinbuch, T. Oomen, and H. Vermeulen, "Motion control, mechatronics design, and Moore's law," *IEEJ Journal of Industry Applications*, vol. 11, no. 2, pp. 245–255, 2022.

[6] T. Kimoto, A. Wachi, S. Sakurai, and M. Mikami, "Automatic diameter control of Czochralski single crystal rods," *Transactions of the Society of Instrument and Control Engineers*, vol. 9, no. 5, pp. 595–600, 1973, in Japanese, with an English abstract.

[7] K. Hirobe and T. Tsuchimoto, "End point detection in plasma etching by optical emission spectroscopy," *Journal of The Electrochemical Society*, vol. 127, no. 1, pp. 234–235, 1980.

[8] M. Uehara, T. Sudo, and F. Kanatani, "Horizontal position detecting device," U.S. Patent 4,558,949, filed 17 September 1982, 1985.

[9] R. Singh, "Rapid isothermal processing," *Journal of Applied Physics*, vol. 63, no. 8, pp. R59–R114, 1988.

[10] A. V. Prabhu, "Performance monitoring of run-to-run control systems used in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, 2008. [Online]. Available: http://hdl.handle.net/2152/17721

[11] D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, T. McDevitt, W. Motsiff, A. Simon, J. Dukovic, R. Wachnik, H. Rathore, R. Schulz, L. Su, S. Luce, and J. Slattery, "Full copper wiring in a sub-0.25 µm CMOS ULSI technology," in *International Electron Devices Meeting. IEDM Technical Digest*, Washington, DC, USA, 1997, pp. 773–776.

[12] D. J. Chivers, "Freeman ion source: An overview (invited)," *Review of Scientific Instruments*, vol. 63, no. 4, pp. 2501–2506, 1992.

[13] A. Ingolfsson and E. Sachs, "Stability and sensitivity of an EWMA controller," *Journal of Quality Technology*, vol. 25, no. 4, pp. 271–287, 1993.

[14] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: Combining SPC and feedback control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 1, pp. 26–43, 1995.

[15] H. Ito, "Chemical amplification resists: Laboratory curiosity to paradigm," *Journal of Photopolymer Science and Technology*, vol. 20, no. 3, pp. 319–332, 2007.

[16] G. S. May and C. J. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control*. Hoboken, NJ: Wiley, 2006.

[17] S. J. Qin, G. Cherry, R. Good, J. Wang, and C. A. Harrison, "Semiconductor manufacturing process control and monitoring: A fab-wide framework," *Journal of Process Control*, vol. 16, no. 3, pp. 179–191, 2006.

[18] A. A. Khan, J. R. Moyne, and D. M. Tilbury, "An approach for factory-wide control utilizing virtual metrology," *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 4, pp. 364–375, 2007.

[19] P. J. Timans, "Emissivity of silicon at elevated temperatures," *Journal of Applied Physics*, vol. 74, no. 10, pp. 6353–6364, 1993.

[20] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[21] R. P. Good and S. J. Qin, "On the stability of MIMO EWMA run-to-run controllers with metrology delay," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 78–86, 2006.

[22] A. Chen and R.-S. Guo, "Age-based double EWMA controller and its application to CMP processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 1, pp. 11–19, 2001.

[23] B. Bamieh, F. Paganini, and M. A. Dahleh, "Distributed control of spatially invariant systems," *IEEE Transactions on Automatic Control*, vol. 47, no. 7, pp. 1091–1107, 2002.

[24] C. D. Schaper, T. Kailath, and Y. J. Lee, "Decentralized control of wafer temperature for multizone rapid thermal processing systems," *IEEE Transactions on Semiconductor Manufacturing*, vol. 12, no. 2, pp. 193–199, 1999.

[25] C. D. Schaper, M. M. Moslehi, K. C. Saraswat, and T. Kailath, "Modeling, identification, and control of rapid thermal processing systems," *Journal of The Electrochemical Society*, vol. 141, no. 11, pp. 3200–3209, 1994.

[26] F. Roozeboom and N. Parekh, "Rapid thermal processing systems: A review with emphasis on temperature control," *Journal of Vacuum Science & Technology B*, vol. 8, no. 6, pp. 1249–1259, 1990.

[27] J. P. Cain, P. Naulleau, and C. J. Spanos, "Critical dimension sensitivity to post-exposure bake temperature variations in EUV photoresists," in *Emerging Lithographic Technologies IX*, ser. Proc. SPIE, vol. 5751, 2005, p. 1092.

[28] M. Tomizuka, "Optimal continuous finite preview problem," *IEEE Transactions on Automatic Control*, vol. 20, no. 3, pp. 362–365, 1975.

[29] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. S. Balakrishnan, and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, vol. 36, no. 11, pp. 1567–1603, 2000.

[30] E. Del Castillo and A. M. Hurwitz, "Run-to-run process control: Literature review and extensions," *Journal of Quality Technology*, vol. 29, no. 2, pp. 184–196, 1997.

[31] T. R. Kuphaldt, *Lessons in Industrial Instrumentation*, version 2.33 ed. Open textbook, Creative Commons Attribution 4.0 license, ibiblio.org/kuphaldt/socratic/sinst, 2024.

[32] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[33] K. J. Åström and L. Rundqwist, "Integrator windup and how to avoid it," in *Proc. 1989 American Control Conference*, 1989, pp. 1693–1698.

[34] D. Goldfarb and A. Idnani, "A numerically stable dual method for solving strictly convex quadratic programs," *Mathematical Programming*, vol. 27, no. 1, pp. 1–33, 1983.

[35] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[36] C. V. Rao, J. B. Rawlings, and D. Q. Mayne, "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations," *IEEE Transactions on Automatic Control*, vol. 48, no. 2, pp. 246–258, 2003.

1.9 Exercises

1.1 Theory and calculation · 4 min

A run-to-run controller of the form of Section 1.4.3 runs at $\lambda = 0.25$ with the process gain known and no metrology delay.

(a) The only disturbance is white lot-to-lot noise with a spread of 0.8 nm. Find the spread of the controlled output. *Checkpoint:* 0.855 nm.

(b) The only disturbance is instead a drift of 0.04 nm per lot. Find the settled offset. *Checkpoint:* 0.160 nm.

1.2 Concept check · 2 min

M-1's after-etch reading of a lot arrives two lots after it, $m = 2$. A step disturbance appears on lot 20. Why are lots 20, 21 and 22 all produced at the full error, whatever the weight λ?

(a) the EWMA averages three readings before it acts on any of them

(b) no reading of lot 20 or later has reached the controller before lot 22 is etched

(c) the trim rate is unknown until three lots have been measured

(d) the integral action of the controller needs three lots to wind up

1.3 Theory and calculation · 5 min

A run-to-run loop adds a fraction of the newest reading's error to its estimate, as in Equation (1.6).

(a) Find the largest stable loop gain K for a queue of one lot and for a queue of three lots. *Checkpoint:* 1.000 and 0.445.

(b) With three lots of queue, the controller runs at $\lambda = 0.4$ but believes a process gain 20 % lower than the true one. Is the loop stable? *Checkpoint:* $K = 0.50$, unstable.

1.4 Theory and calculation · 4 min

M-1's controller uses the matched-lot estimator of Section 1.4.4 with the gain known, $\lambda = 0.5$ and a queue of $m = 3$ lots. A step of 1 nm appears on lot 10. Give the error on lots 13, 14 and 15.

1.5 Concept check · 2 min

Two run-to-run controllers on the same line have the same weight, the same believed gain and the same two-lot metrology queue. One adds a fraction of the newest reading's error to its estimate, and the other regresses each reading on the recipe that made that lot. What separates their stability?

(a) the matched-lot controller samples more lots, so its queue is shorter

(b) the newest-reading controller uses a smaller weight, so it is slower

(c) the matched-lot controller filters the metrology noise, so it is less excited

(d) the matched-lot controller keeps the record of which recipe made each reading, so the delayed term cancels

1.6 Theory and calculation · 5 min

A small plate has three zones in a row with insulated ends. Each zone loses $U = 1\,\mathrm{W/K}$ to the surroundings and exchanges $k_x = 2\,\mathrm{W/K}$ with each neighbor.

(a) Find the steady gain of the uniform pattern, of the pattern with one half-wave and of the alternating pattern. *Checkpoint:* 1.000, 0.333 and 0.143 K/W.

(b) A loop on the center zone is tuned with its neighbors held at their setpoints, to a loop gain of one. What loop gain does it have when all three zones move together? *Checkpoint:* 5.

1.7 Theory and calculation · 3 min

A total-radiation pyrometer is set to an emissivity, but the wafer's true emissivity is 5 % lower. The loop holds the reading at 600 °C. Find the true wafer temperature.

1.8 Theory and calculation · 4 min

A film of 200 nm is etched with a local rate that varies by $\pm 5\,\%$ across the wafer, over an underlayer etched 25 times more slowly. The main etch ends on an endpoint detector that fires when half the wafer has cleared, followed by an over-etch proportional to the main etch.

(a) Find the smallest over-etch that clears every point. *Checkpoint:* 5.26 %.

(b) Find the underlayer loss at the fastest point with 8 % over-etch. *Checkpoint:* 1.02 nm.

1.9 Theory and calculation · 5 min

Lithography leaves a lot-to-lot spread of 1.5 nm in the developed linewidth, which a microscope reads with a spread of 0.5 nm. The etch adds no spread of its own. The etcher's trim is corrected on the same lot from the developed reading, as in Section 1.4.7.

(a) Find the finished spread with full feedforward, $g = 1$. *Checkpoint:* 0.500 nm.

(b) Find the best feedforward gain and the spread it gives. *Checkpoint:* $g_* = 0.900$, 0.474 nm.

(c) With $g = 1$, only one lot in two is measured. Find the spread over all lots. *Checkpoint:* 1.118 nm.

1.10 Concept check · 2 min

A cold wafer lands on a hot bake plate at a known instant, and the plate dips before any thermocouple loop can respond. Which control element answers this disturbance best?

(a) a power pulse triggered by the landing, with the zone integrators held while it runs
(b) a faster integral time on every zone loop
(c) a run-to-run correction of the zone setpoints from the next wafer's linewidth
(d) a filter on the thermocouples to keep the loops from reacting to the dip