INTERACTIVE CONTROL ENGINEERING
Principles and Applications

Volume 14 Control in Semiconductor Manufacturing

CHAPTER 6

Lithography and Etch Virtual Laboratory for Run-to-Run Control Through a Metrology Queue

Pages 137–162

Edited by John D. Hedengren

How to cite this chapter. "Lithography and Etch Virtual Laboratory for Run-to-Run Control Through a Metrology Queue," in *Control in Semiconductor Manufacturing*, John D. Hedengren, Ed., vol. 14 of *Interactive Control Engineering: Principles and Applications*. Advanced Process Solutions, LLC, 2026, ch. 6, pp. 137–162.

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

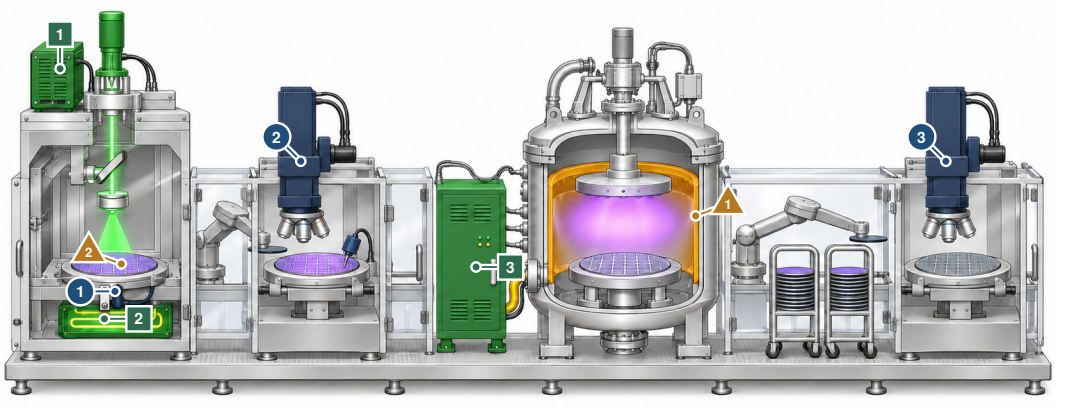
The app for this chapter
https://apmonitor.com/apps/litho-etch-r2r/

The volume online: contents, other chapters and citation records
https://apmonitor.com/apps/books/v14-semiconductor/#litho-etch-r2r

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

CHAPTER 6

Lithography and Etch Virtual Laboratory for Run-to-Run Control Through a Metrology Queue



The controller moves 1 the exposure dose of the light source 2 the heater under the bake plate 3 the trim time written into the etch recipe

The controller reads 1 the bake plate temperature 2 the microscope that measures the developed lines 3 the microscope that measures the etched lines, two lots late

What disturbs it 1 deposits on the etch chamber walls, emptied by a wet clean 2 a new resist batch that prints the lines wider

Notice the carriers of wafers waiting before the last microscope: the etched linewidth that steers the etch recipe is read only after two more lots have been made. Illustration, not to scale.

IN PLAIN TERMS **Correcting a Recipe With Late News**

The main idea. Chips are made in batches of wafers, and fine lines are printed on each batch with light and then etched. A controller reads how wide the finished lines came out and sets the etching time for the next batch.

The challenge. The answer comes back late, after two more batches have been made. As with a shower whose hot water takes a while to arrive, pushing hard on old news makes the swings grow.

Why it matters. The wait at the measuring tool, and how carefully each result is matched to its own batch, decide how hard the controller may push.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the fact and history boxes of Sections 6.1, 6.2 and 6.4, the try-it boxes of Sections 6.5 and 6.6, the fact box of Section 6.6, and the key points
Study	55 min	The whole chapter with its pause points and the twelve exercises
Practice	15 min	The first page, Section 6.6, Section 6.7, Section 6.8 and the open challenge after it

The control question A lot of wafers is etched every thirty minutes, and its etched linewidth is reported only after two more lots have run. How hard can a recipe controller push on that loop, and how much of the limit comes from the queue and the bookkeeping rather than from the process?

Learning objectives

- Write the EWMA recipe controller as a discrete integral controller and derive its closed loop through a metrology queue of m lots.
- Compute the stability boundary of the loop gain for a given queue and gain ratio, and predict the dead band and decay after a step.
- Explain why an estimator that pairs each reading with the recipe of its own lot removes the queue from the boundary.
- Compare feedforward from the developed linewidth with faster feedback on a measured spread, and price a thinner sampling plan.
- Diagnose two run-to-run loops that share one measurement and fix them by giving each loop its own.

Prerequisites Sections 1.4.3 (run-to-run control), 1.4.4 (metrology delay), 1.4.7 (feedforward between steps), 1.6.2 (velocity form and loop action) and 1.6.4 (scenarios, scorecard and CSV).

The app https://apmonitor.com/apps/litho-etch-r2r/

See also Volume 8, Chapter 9 for a correction that must know which command a late result answers, and Volume 8, Chapter 7 for correction from one part to the next. Volume 5, Chapter 3 for a loop behind a slow analyzer, and Volume 5, Chapter 9 for a sampled loop acting on a valuation a year late.

6.1 Introduction

A control loop is usually characterized by the dynamics of its plant. The loop in this chapter has a plant with no dynamics worth discussing and a measurement that arrives an hour after the material it describes has left the building. A lot of twenty-five wafers is exposed, baked, developed, measured, etched and measured again, and the second measurement is available only after two more lots have passed through the same four stations. Sampling is once per lot and the actuator moves once per lot, so the entire loop lives on a thirty-minute clock with two samples of transport delay in front of it. Nothing about the resist, the plasma or the metrology decides how hard that loop may be pushed. The queue does.

DID YOU KNOW? Half the Steps Are Measurements

In 2018 a review of semiconductor metrology reported that for some products more than 50 % of the manufacturing steps can involve measurement or characterization, and that the share grows as devices shrink and become three-dimensional [1].

The same shape appears wherever the quality measurement is an assay rather than a transmitter. A coating line whose thickness gauge sits two hundred meters downstream, a polymerization reactor whose melt index comes back from a laboratory once a shift, a mineral circuit graded by an off-line particle size analysis, and a casting line inspected after cooling are all plants where the number of units committed between the sample and the correction is the parameter that matters. Textbook treatments of dead time develop the continuous case, where the delay is a fraction of the dominant time constant and the remedy is detuning or a predictor [2]. The discrete case is different because there is no dominant time constant at all. The plant is static within one sample, the controller is a pure discrete integrator, and the delay is an integer count of samples rather than a continuous quantity. What a reader most often carries into this plant, and has to give up, is the belief that a reset rate which is calm on one of these loops will be calm on another. Removing ninety percent of the measured error before the next sample is ordinary practice on a loop whose measurement is live. The same fraction

applied once per lot to the etch recipe of this same line diverges, and the only difference between the two cases is how much material was committed before the answer came back.

Semiconductor recipe control is where that inequality was first written down and where it is still enforced hardest. The exponentially weighted moving average (EWMA) recipe controller entered the industry in the early 1990s [3], its stability and its sensitivity to process gain error were analyzed immediately after [4, 5], and the practical consequences of queued rather than inline metrology have been treated repeatedly since [6, 7]. However, a recursion written on a board as $u_{k+1} = u_k + (\lambda/\hat{b})(r - y_k)$ comes with no plant to try it on, no disturbance that outruns the loop, and no reason to discover that the boundary belongs to the bookkeeping rather than to the process. This chapter makes the delay-limited boundary measurable on an interactive lithography and etch line, on which an estimator that associates each reading with the recipe that produced it moves that boundary by a factor of three, and on which a feedforward path from an upstream measurement removes a variance source that no amount of feedback tuning can filter. The simulator is a six-state disturbance model driving a four-station lot pipeline, with three controllers on three sample rates. Two closed-loop recursions, one per estimator, are each checked against the running simulator to four decimal places in per-lot decay rate, and one plant at one weight of $\lambda_E = 0.95$ is divergent under one estimator and quiet under the other. The EWMA controller and its stability condition are standard, and Sections 1.4.3 and 1.4.4 of Chapter 1 set them out on a smaller line. Section 6.2 traces how recipe feedback replaced the control chart. Section 6.3 gives every equation and constant needed to rebuild the simulator, and Section 6.4 the three controllers with their estimator, feedforward and sampling options. Section 6.5 derives the closed loop through the queue, measures its boundary and shows that the boundary belongs to the estimator. Section 6.6 reports what each scenario measures, from the price of one lot of queue to the graded challenge, Section 6.7 names the instruments and the control layer behind the loops, Section 6.8 states the limitations and Section 6.9 summarizes.

6.2 From Control Charts to Recipe Feedback

Patterning was run open loop for its first two decades. A process engineer qualified a recipe on a chamber, released it, and defended it with a statistical process control chart on the finished linewidth. The chart answered one question, whether the process had changed, and it answered it conservatively by design, because a control chart is built to avoid reacting to common cause variation. Recipes were adjusted at a review, typically once a shift, and the adjustment was an engineering judgment rather than a computation. Every hour of drift between the sample and the review was product.

That arrangement was adequate while the tolerance was a large fraction of the feature. It stopped being adequate when the lithography and etch tools began to drift on a time scale shorter than the review interval. Chamber walls accumulate and lose polymer over hours and change the etch rate and the uniformity with them [8], a resist batch prints a different linewidth at the same dose, and the post-exposure bake plate carries a linewidth sensitivity of several nanometers per degree in a chemically amplified resist [9]. None of those sources is reported by any instrument on the tool that causes them. The only evidence that any of them moved is the linewidth itself.

Run by run control is the answer that was adopted. The controller reads the finished measurement of a lot, updates an estimate of where the process currently sits, and writes the next recipe from that estimate [3]. The estimator chosen was an exponentially weighted moving average, because it is one number of state, it needs no process model beyond a gain, and its weight λ is a single interpretable trade between chasing drift and chasing noise. How the method was proposed, analyzed and collected in its first decade is what the History box tells. A review of automatic control across microelectronics manufacturing placed recipe feedback beside the regulatory layer rather than inside it [10].

Table 6.1 Roles, Nominal Values and Operating Ranges. The Nominal Column Is the Design Recipe From Which Each Scenario Departs.

Symbol	Role	Nominal	Range
CD_{AEI}	CV, once per lot, m late	45.00 nm	38–55 nm SP
CD_{ADI}	CV, once per lot, no delay	62.00 nm	52–74 nm SP
T_{PEB}	CV, continuous	110.00 °C	95–125 °C SP
t_{trim}	MV, once per lot	10.00 s	0–40 s
E	MV, once per lot	28.00 mJ/cm^2	16–44 mJ/cm^2
P_h	MV, continuous	369.6 W	0–1200 W
θ	state, DV, unmeasured	0.5	0–1
δ	state, DV, unmeasured	0 nm	±4 nm/1000 min
g_E	state, DV, unmeasured	1.000	0.92–1.08
F	state, DV, unmeasured	0 µm	0–0.15 µm
β_M	state, DV, unmeasured	0 nm	±1.5 nm
q_{load}	DV, unmeasured	0 W	±40 W

HISTORY From Control Chart to Recipe Feedback (1991 to 2001)

Run-to-run control, a recipe adjusted lot by lot or wafer by wafer, was first proposed for chip fabrication in 1991 [11]. A 1995 framework combined statistical process control with feedback in three parts: a gradual mode for drifts, such as deposits building up in a reactor, a rapid mode for sudden shifts, such as after maintenance, and a control chart that chose between them. On an epitaxy reactor the rapid mode recovered within three runs and the gradual mode cut the variation by a factor of 2.7 [3]. A 1993 analysis gave a simple condition on the weight and the estimated process gain that keeps the controller stable [4]. Work on chemical mechanical polishing reported practical barriers: limits on machine settings, recipes rounded to the steps a tool accepts, and the choice of the weight [12]. A literature review with extensions followed in 1997 [5], and an edited volume collected the field in 2001 [13].

Two refinements matter for this application. The first is the double EWMA, which carries a drift term as well as an intercept so that the loop tracks a ramp without a constant offset, at the cost of a second tuning constant and a transient that is longer than the single EWMA transient [14]. The second is the explicit treatment of metrology delay. Once the measurement of a lot is unavailable until several later lots have run, the loop acquires a pure sample delay, the largest usable weight collapses, and the age of the measurement becomes a tuning input in its own right [6, 7]. Casting the EWMA controller in an internal model control framework makes the same trade explicit as a performance and robustness choice [15]. Integrated metrology, which places the measurement on the process tool and drives the queue toward zero, is the hardware answer to the same inequality, and the alternative software answer is to associate every reading with the tool, chamber and recipe context that produced it so that the controller is working on a matched pair [16]. This application implements the single EWMA with an intercept only, offers both estimators, and exposes the queue length as a control the reader sets.

6.3 Process Model

The line is four stations in series. Litho cell LC-101 coats, exposes at dose E, bakes on plate PEB-101 and develops. Scanning electron microscope CD-SEM-A measures the developed resist linewidth, the after-develop inspection or ADI. Plasma etcher ET-201 trims the resist for t_{trim} seconds and transfers the pattern. CD-SEM-B measures the finished linewidth, the after-etch inspection or AEI. One lot of N_w wafers enters the line every T_{takt} minutes and every station is occupied by a different lot at any instant. Every symbol that appears below is drawn on the line itself, together with the two correction paths, in Figure 6.1.

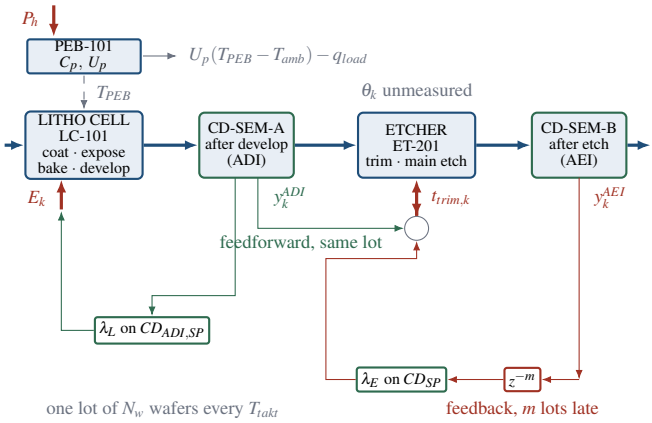


Figure 6.1 The four-station patterning line and its two recipe loops. Lot k is exposed at dose E_k on a plate held at T_{PEB}, measured as y_k^{ADI}, etched for $t_{trim,k}$ in a chamber whose wall condition θ_k is reported by no instrument, and measured again as y_k^{AEI}. The feedback path reaches the etcher only through the queue z^{-m}, while the feedforward path adds $(y_k^{ADI} - CD_{ADI,nom})/\hat{r}$ to the recipe of the same lot one station before that lot is etched.

What the three loops act on and act against is collected in Table 6.1. Every disturbance in the table is unmeasured, which is the defining property of the plant. There is no instrument anywhere on the line that reports the chamber wall condition, the lithography drift, the scanner dose calibration error, the defocus or the metrology offset, and the only evidence that any of them moved is a linewidth arriving late. Time is carried in minutes, linewidth in nanometers, dose in mJ/cm^2, trim time in seconds and plate temperature in degrees Celsius. Six states are integrated continuously and the lot arithmetic is evaluated once per T_{takt}.

🕮 **Pause.** Try Exercise 6.1 before reading on.

6.3.1 The Developed Linewidth

Each lot is a single number. The developed resist linewidth of lot k is the algebraic expression of Equation (6.1).

$$CD_{ADI,k} = CD_{r0} + s_E \ln\left(\frac{g_E E_k}{E_0}\right) - c_F F^2 - k_T\,(T_{PEB} - T_0) + \delta + \varepsilon_{L,k} \tag{6.1}$$

where CD_{r0} is the resist linewidth at nominal dose, best focus and nominal bake, s_E is the exposure latitude in nanometers per unit natural logarithm of dose, g_E is the scanner dose calibration ratio, c_F is the Bossung curvature, F is the defocus, k_T is the bake sensitivity, δ is the accumulated lithography drift and $\varepsilon_{L,k}$ is a zero mean Gaussian lot-to-lot lithography disturbance of standard deviation σ_L. Both sides carry nanometers. Dose enters logarithmically because the resist responds to the logarithm of exposure, which is why exposure latitude is quoted in nanometers per percent of dose rather than per millijoule. At the design point a 1 % dose change moves the developed linewidth by 0.2388 nm and a step from 28.00 to 29.00 mJ/cm^2 moves it by –0.8422 nm.

Focus enters as a square because the Bossung curve through focus has its extremum at best focus [17]. A defocus of either sign moves the printed line in the same direction, so a linear feedback controller can trim the mean back but can never recover the sign of the error and can never say which way to move the stage. A defocus of 0.09 µm narrows the line by 0.9720 nm whether the wafer is high or low. The bake term is linear at 3.2 nm/°C, which puts a tenth of a degree of plate error at 0.32 nm of linewidth and is the reason a bake plate is specified the way it is [9].

6.3.2 The Etcher

The etcher removes a bias from the developed line, set by a fixed component, by the resist trim time and by the condition of the chamber walls, as shown in Equations (6.2) and (6.3).

$$b_k = b_0 + r_{trim}\, t_{trim,k} + \beta\, \theta \tag{6.2}$$

$$CD_{AEI,k} = CD_{ADI,k} - b_k + \varepsilon_{E,k} \tag{6.3}$$

where b_0 is the bias at zero trim, r_{trim} is the resist trim rate in nanometers per second, β is the additional bias contributed by a fully seasoned chamber, θ is the dimensionless wall condition running from 0 for bare aluminum to 1 for fully seasoned, and $\varepsilon_{E,k}$ is a zero mean Gaussian etch disturbance of standard deviation σ_E. The trim time is the manipulated variable of the loop this chapter is about and r_{trim} is its true process gain. At the design recipe of 28.00 mJ/cm^2 and 10.00 s on a chamber at $\theta = 0.5$ with the plate at 110.00 °C, Equation (6.1) returns exactly 62.00 nm, Equation (6.2) returns exactly 17.00 nm and Equation (6.3) returns exactly 45.00 nm. The application verifies those three identities and the vanishing of every state derivative at startup, and the manual scenario holds them: after twenty lots by hand at 369.6 W of plate power the measured plate temperature is 110.000 000 °C, the developed linewidth is 62.000 000 nm and the etched linewidth is 45.000 000 nm.

The wall condition is the only slow state on the etch side and it relaxes toward whatever the chemistry is driving it to, as shown in Equation (6.4).

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\theta_\infty - \theta}{\tau_\theta}, \qquad 0 \le \theta \le 1 \tag{6.4}$$

where θ_∞ is the target condition and τ_θ is the seasoning time constant of 600 min, which is twenty lots. A wet clean sets θ to zero and θ_∞ to one in the same instant, which is a step of −2.0 nm in the bias followed by a ten-hour first-order recovery of +4.0 nm. The state is clipped to $[0, 1]$ after every integration step.

6.3.3 The Post-Exposure Bake Plate

The plate is the one part of the line with a live measurement and continuous dynamics, and it is included so that the contrast is available on the same product characteristic. Its energy balance is Equation (6.5).

$$C_p \frac{\mathrm{d}T_{PEB}}{\mathrm{d}t} = 60\,[P_h + q_{load} - U_p\,(T_{PEB} - T_{amb})] \tag{6.5}$$

where C_p is the plate heat capacity in J/K, P_h is the heater power in watts clipped to $[0, 1200\,\mathrm{W}]$, q_{load} is an unmeasured heat load, U_p is the loss coefficient in W/K and T_{amb} is the track ambient. The factor 60 converts watts and joules, which are per second, to the minute clock the rest of the model uses. The open-loop time constant is $C_p/U_p = 61.9\,\mathrm{s} = 1.032\,\mathrm{min}$ and the steady power at 110.00 °C is $U_p(T_{PEB} - T_{amb}) = 369.6\,\mathrm{W}$.

6.3.4 Drift, Calibration and Metrology

The remaining four states carry what moves underneath the line. Equation (6.6) is the lithography drift and Equation (6.7) is the common form of the three calibration states.

$$\frac{\mathrm{d}\delta}{\mathrm{d}t} = \frac{\rho}{1000} \tag{6.6}$$

$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{g_{sp} - g}{\tau_{cal}}, \qquad g \in \{g_E,\ F,\ \beta_M\} \tag{6.7}$$

where ρ is the drift rate in nanometers per thousand minutes, g_{sp} is the value the operator or a timed event has requested, and τ_{cal} is a short relaxation time of 3 min that keeps a commanded step continuous enough for a fixed-step integrator. The dose calibration state is clipped to $[0.5, 1.5]$. A

calibration of $g_E = 0.97$, meaning that the scanner delivers three percent less light than it reports, widens the developed line by 0.7310 nm.

Both scanning electron microscopes are modeled as an offset plus white noise on the lot mean, as shown in Equation (6.8).

$$\begin{aligned} y_k^{ADI} &= CD_{ADI,k} + \beta_M + \varepsilon_{M,k} \\ y_k^{AEI} &= CD_{AEI,k} + \beta_M + \varepsilon'_{M,k} \end{aligned} \tag{6.8}$$

where β_M is the tool matching offset and $\varepsilon_{M,k}$ and $\varepsilon'_{M,k}$ are independent zero mean Gaussian draws of standard deviation σ_M. The controller believes Equation (6.8) completely, so an offset goes straight into the silicon. A CD-SEM offset of +0.8 nm applied to a line running the etch loop at $\lambda_E = 0.30$ settles the true etched linewidth at exactly 44.200 nm against a 45.00 nm target, which is 0.8 nm of real material removed to satisfy an instrument error, as Section 1.4.2 of Chapter 1 sets out. Metrology sampling is by lot. Only lots with $k \bmod n_s = 0$ are measured at either station, where n_s is the sampling interval in lots.

A deterministic across-wafer signature is added to the lot mean for the hundred exposure fields of the displayed wafer map, as shown in Equation (6.9).

$$s_{k,i} = f_S\left(u_i^2 - 0.35\right) + f_R\left(\rho_i^2 - 0.42\right) + 1.7\, f_N\, \eta_{k,i} \tag{6.9}$$

where u_i is the normalized across-slit coordinate of field i, ρ_i is its normalized radius, f_S, f_R and f_N are the slit, radial and field-to-field amplitudes, and $\eta_{k,i} \in [-1, 1]$ comes from an integer hash of the lot and field indices so that the same lot always produces the same map. Equation (6.9) is display only. It does not feed back, and the run-to-run controllers see the lot mean, so this application cannot teach spatial recipe control of the kind that corrects the map itself [18].

6.3.5 Parameters, Initialization and Numerics

All twenty-five constants of the plant appear in Table 6.2. Twenty-two are adjustable from the model panel and three never appear in the interface at all.

Every scenario opens at the design point. The states are initialized to $\theta(0) = 0.5$, $T_{PEB}(0) = T_0$, $\delta(0) = 0$, $g_E(0) = 1$, $F(0) = 0$ and $\beta_M(0) = 0$, and the two recipe estimators are seeded from the manual recipe so that a transfer to automatic does not step the line. Equations (6.4) through (6.7) are advanced by a fixed-step fourth-order Runge-Kutta integrator at 0.02 min (1.2 s), the three controllers execute on a 0.04 min (2.4 s) period, and five samples per simulated minute reach the historian. One generator, seeded at 20260915, supplies every random draw in the model. The engine, its clocks and its seeding are those of Section 1.6.1, so a scenario carrying noise replays to the last digit at any playback speed.

6.3.6 Assumptions

The plant is deliberately thin, because everything discarded was discarded to leave the relationship between sampling, delay and gain visible. Each simplification fails somewhere.

Each station is a static map with an additive lot term. Equations (6.1) and (6.3) have no dynamics inside a lot and no memory from one lot to the next except through the six continuous states. Real linewidth response is nonlinear in dose over a wide range, and only the focus term is kept nonlinear here. The assumption fails as soon as the dose is pushed more than about ten percent from E_0.

One number per lot. Within-wafer and within-lot variation enter only through Equation (6.9), which is display only. A production controller also corrects the shape of the across-wafer map [18], and no configuration of this application can show that.

The etcher is one scalar. Equation (6.2) reduces an entire plasma chamber to θ. A chamber wall state genuinely changes radical recombination and therefore both rate and uniformity [8], so the scalar is a legitimate reduction for the lot mean and a poor one for anything spatial.

The lot cadence is fixed and the pipeline is instantaneous. The line never blocks, starves or reorders, every lot takes exactly T_{takt}, and all four stations of a lot are evaluated at one instant rather than as the

Table 6.2 Constants of the Plant. A Hidden Constant Does Not Appear Anywhere in the Interface.

Symbol	Meaning	Value	Unit	Visibility
T_{takt}	lot cadence	30	min	adjustable
N_w	wafers per lot	25	—	hidden
CD_{r0}	resist CD at nominal dose	62	nm	adjustable
E_0	nominal exposure dose	28	mJ/cm^2	adjustable
s_E	exposure latitude	−24	nm per ln E	adjustable
c_F	Bossung curvature	120	nm/µm^2	adjustable
k_T	bake sensitivity	3.2	nm/°C	adjustable
T_0	nominal bake temperature	110	°C	adjustable
C_p	plate heat capacity	260	J/K	adjustable
U_p	plate loss coefficient	4.2	W/K	adjustable
T_{amb}	track ambient	22	°C	hidden
b_0	etch bias at zero trim	10.00	nm	adjustable
r_{trim}	resist trim rate	0.500	nm/s	adjustable
β	bias from a seasoned chamber	4.00	nm	adjustable
τ_θ	seasoning time constant	600	min	adjustable
σ_M	CD-SEM repeatability	0.45	nm	adjustable
σ_L	lot-to-lot litho noise	0.80	nm	adjustable
σ_E	lot-to-lot etch noise	0.35	nm	adjustable
τ_{cal}	calibration relaxation	3	min	hidden
—	specification half-width	2.0	nm	adjustable
f_S	across-slit signature	0.85	nm	adjustable
f_R	radial etch signature	1.10	nm	adjustable
f_N	field-to-field random	0.40	nm	adjustable
—	value of an in-specification lot	180 000	$	adjustable
—	cost of an out-of-specification lot	22 000	$	adjustable

lot walks down the line. The only transport the model carries is the metrology delay m, which the reader sets rather than inherits from dispatch, and that understates how variable the delay is in a real fab.

Metrology noise is white and zero mean within a tool. Tool-to-tool bias is offered as a switchable disturbance in Equation (6.8), but metrology drift is absent, sampling plans that skip wafers rather than lots are absent, and measurement uncertainty that depends on the feature itself is absent.

The estimators carry an intercept and no drift term. A double EWMA would remove the steady offset against a ramp at the cost of a second constant [14]. The offset is left visible on purpose and is quantified in Section 6.5.

The numerics are more reproducible than a fab. Freezing the step and the seed reproduces one trajectory everywhere, which is convenient and false.

6.4 Controllers as Implemented

Three loops run on three clocks. The bake plate is an ordinary proportional-integral controller on a 2.4 s scan. The exposure dose and the etch trim time each move once every 30 min. The lot clock is advanced inside the controller, which fires the whole of a lot at one instant whenever simulated time passes a multiple of T_{takt}.

6.4.1 The EWMA Recipe Controller

Write A_k for the etched linewidth lot k would have had with no trim at all, which by Equations (6.2) and (6.3) is $A_k = CD_{ADI,k} - b_0 - \beta\,\theta$. The measured linewidth is then $y_k = A_k - r_{trim} u_k$ where u_k is the trim time applied to that lot. The controller holds one state, an estimate α of A, and writes the recipe that would put the estimate on target using the trim rate $\hat{r}$ it believes, as shown in Equation (6.10).

$$u_{k+1} = \frac{\alpha_k - CD_{SP}}{\hat{r}} \tag{6.10}$$

The intercept form of the EWMA estimator is Equation (6.11), and it is what the application applies when the estimator option is set to the matched lot.

$$\alpha_k = \lambda_E \left(y_{k-m} + \hat{r}\, u_{k-m}^{FB} \right) + (1 - \lambda_E)\, \alpha_{k-1} \tag{6.11}$$

where m is the after-etch metrology delay in lots and u^{FB} is the feedback part of the recipe. When the estimator option is set to the newest reading the application applies Equation (6.12) instead.

$$\alpha_k = \alpha_{k-1} + \lambda_E \left(y_{k-m} - CD_{SP} \right) \tag{6.12}$$

Substituting Equation (6.10) into Equation (6.12) gives the incremental form of Equation (6.13), which is why a run-to-run controller is a discrete integral controller and why λ_E is a reset time measured in lots. It is Equation (1.3) of Chapter 1 with $\hat{b} = -\hat{r}$, $y^* = CD_{SP}$ and the reading m lots old.

$$u_{k+1} = u_k + \frac{\lambda_E}{\hat{r}} \left(y_{k-m} - CD_{SP} \right) \tag{6.13}$$

There is no proportional term and no derivative term in Equation (6.13). With a 30 min sample and two samples of delay, neither would have anything useful to contribute. Substituting Equation (6.10) into Equation (6.11) at $m = 0$ collapses it to Equation (6.12) for any $\hat{r}$ whatever, so the two estimators are the same controller when there is no queue. With a queue they are not, and Section 6.5 measures the difference.

The recipe is clipped to $0 \le t_{trim} \le 40$ s. Anti-windup is by back-calculation on the estimator rather than on the output. Whenever the clip is active the estimate is recomputed from the trim the etcher can actually deliver, $\alpha_k \leftarrow CD_{SP} + \hat{r}\, u_{clipped}$, so the integral state can never advance past what the actuator will honor. In manual mode both estimators are driven to follow the operator, with $\alpha \leftarrow CD_{SP} + \hat{r}\, t_{trim,manual}$ evaluated every controller scan, which is the discrete equivalent of the tracking that gives the continuous loop its bumpless transfer. In automatic the manual trim and dose stations are set to the applied recipe after every controller execution, so a return to manual is bumpless as well. A change of $\hat{r}$ in automatic rescales the estimate, $\alpha \leftarrow CD_{SP} + (\alpha - CD_{SP})\, \hat{r}_{new} / \hat{r}_{old}$, so the next recipe of Equation (6.10) is the one the old estimate asked for and only the loop gain changes. At the design point a move of $\hat{r}$ from 0.50 to 1.00 nm/s leaves the trim at 10.00 s on every lot. The controller is direct acting on the etched linewidth, because a linewidth above target calls for a longer trim, and the application verifies that sign explicitly. That is action $a = -1$ in the convention of Section 1.6.2. The dose loop below is direct acting for the same reason, since more dose prints a narrower line.

6.4.2 The Dose Loop

The dose loop reads the developed linewidth, which has no queue in front of it, and corrects in the logarithmic domain as shown in Equation (6.14).

$$\ln E_{k+1} = \ln E_k - \frac{\lambda_L}{s_E} \left(y_k^{ADI} - CD_{ADI,SP} \right) \tag{6.14}$$

Because Equation (6.1) depends on $\ln E$ through the same constant s_E that appears in Equation (6.14), the linewidth change produced by one update is exactly $-\lambda_L$ times the measured error. The loop is therefore linear in its own variable rather than only locally linear, one λ_L works over the whole dose range, and the gain ratio is unity by construction. The dose command is clipped to 16–44 mJ/cm^2. A reader may instead point this loop at the etched linewidth, which is offered because the consequence is worth measuring and is reported in Section 6.5.4.

6.4.3 Feedforward From the Developed Linewidth

The developed linewidth of a lot is measured before that lot is etched, so an incoming resist error can be cancelled at the etcher instead of being waited out through a loop that will not see it for m more

lots. The applied recipe is Equation (6.15).

$$t_{trim,k} = u_k^{FB} + \underbrace{\frac{y_k^{ADI} - CD_{ADI,nom}}{\hat{r}}}_{u_k^{FF}} \tag{6.15}$$

where $CD_{ADI,nom}$ is the developed-linewidth setpoint the etcher was set up to trim. The correction exists only for a lot that was measured, so on the lots a sampling plan skips the etcher receives no feedforward at all. Substituting Equation (6.15) into Equations (6.2) and (6.3) with $\hat{r} = r_{trim}$ gives Equation (6.16).

$$\begin{aligned} CD_{AEI,k} = CD_{ADI,nom} - b_0 - r_{trim} u_k^{FB} - \beta\theta \\ - \beta_M - \varepsilon_{M,k} + \varepsilon_{E,k} \end{aligned} \tag{6.16}$$

Every term carrying $CD_{ADI,k}$ has cancelled, including the whole of $\varepsilon_{L,k}$. The lithography disturbance of standard deviation σ_L has been replaced by the ADI metrology disturbance of standard deviation σ_M, so the arrangement is worth having exactly when the measurement is quieter than the process, which is the usual case. It is the full-gain case of Equation (1.11) in Chapter 1.

> **HISTORY Feedforward From Lithography to Etch (1999)**
>
> Feedforward between two process steps adjusts a recipe to compensate for a deviation measured upstream, and it promises lower run-to-run variance, less rework and less scrap. A 1999 study named two problems with it. Noise in the measuring tool can make the variance under feedforward exceed the variance with no control at all, so the recipe adjustment was detuned by the confidence in the sensor. And changing one recipe parameter can worsen others, so the controller chose among recipes already qualified [19]. Its controller adjusted the recipes of a reactive ion etcher to compensate for variations in the lithography linewidth and reduced the variance of the etch process [20].

Equation (6.11) regresses on u^{FB} and deliberately not on the applied trim of Equation (6.15). Regressing on the applied trim would add $y_{k-m}^{ADI} - CD_{ADI,nom}$ to the estimated intercept, which is a lithography deviation the feedforward path has already removed from the product. The feedback loop would then move the next recipe by λ_E times a deviation that no longer appears in any error, and would spend the following lot undoing what the feedforward did.

Pause. Try Exercise 6.3 before reading on.

6.4.4 The Bake Plate Loop

Plate PEB-101 runs the shared velocity-form controller of Section 1.6.2 as a proportional-integral loop on the filtered measurement, with back-calculation anti-windup, written for this loop in Equation (6.17).

$$\begin{aligned} e_k &= T_{SP} - \hat{T}_k, \qquad & s_k &= s_{k-1} + K_{c,T}\,(e_k - e_{k-1}) \\ c_k &= s_k + I_{k-1}, \qquad & P_{h,k} &= \mathrm{clip}(c_k,\ 0,\ 1200\,\mathrm{W}) \\ I_k &= I_{k-1} + \frac{K_{c,T}}{\tau_{I,T}}\, e_k \Delta t_c + \frac{\Delta t_c}{T_t}\left(P_{h,k} - c_k\right) \end{aligned} \tag{6.17}$$

where $\Delta t_c = 0.04\,\mathrm{min}$ is the scan, s_k is the running sum S_k of Equation (1.13), c_k is the power the algorithm demands, which is v_k of Equation (1.14) with no bias, I is the integral state of Equation (1.15) and $T_t = \tau_{I,T}$ is the tracking time of the back-calculation. The integral is seeded with the 369.6 W that balances the plate. While the heater is at a limit only I is corrected, by $\Delta t_c / T_t = 1/25$ of the clipped excess per scan at the default tuning. There is no derivative term and no slew limit. More heater power raises the plate temperature, so the error is setpoint minus measurement and the loop is reverse acting, action $a = +1$ in the convention of Section 1.6.2. The measurement $\hat{T}$ is the plate temperature through a first-order filter of 0.1 min updated by an exact zero-order hold, which is

stable at any period. The default tuning is $K_{c,T} = 15\,\mathrm{W/^{\circ}C}$ and $\tau_{I,T} = 1.00\,\mathrm{min}$ against an open-loop time constant of 1.032 min. Transfer between manual and automatic is bumpless in both directions, by the tracking that Section 1.6.2 describes.

6.5 Stability Through the Queue

6.5.1 The Delay-Limited Boundary

Let $e_k = y_k - CD_{SP}$ and let $\xi = r_{trim}/\hat{r}$ be the ratio of the true trim rate to the one the controller believes, the gain ratio of Section 1.3. Differencing $y_k = A_k - r_{trim}u_k$ with A held constant and substituting Equation (6.13) gives the closed loop of Equation (6.18).

$$e_{k+1} = e_k - \lambda_E\,\xi\,e_{k-m} \tag{6.18}$$

which is a linear difference equation with the characteristic polynomial of Equation (6.19), writing $K = \lambda_E\xi$ for the loop gain.

$$z^{m+1} - z^m + K = 0 \tag{6.19}$$

The pair is Equation (1.6) of Chapter 1 written for this loop. Placing a root on the unit circle, $z = \mathrm{e}^{j\omega}$, gives a phase condition that fixes the crossing frequency at $\omega = \pi/(2m+1)$ with no reference to K and a magnitude condition $2\sin(\omega/2) = K$, so the boundary is that of Section 1.4.4, Equation (6.20).

$$(\lambda_E\xi)_{\max} = 2\sin\frac{\pi}{2\,(2m+1)} \tag{6.20}$$

Equation (6.20) returns 2, 1, 0.618, 0.445 and 0.347 for m = 0 through 4. The value at two lots is the reciprocal of the golden ratio. The crossing frequency also fixes the period of the marginal oscillation at $2\pi/\omega = 2(2m+1)$ lots, which is ten lots at $m = 2$ and is the ringing period a reader sees on the control chart just inside the boundary.

Three consequences are worth separating. The boundary halves between $m = 0$ and $m = 1$ and falls by only a factor of 1.28 between $m = 3$ and $m = 4$, because $2\sin[\pi/(2(2m+1))]$ tends to $\pi/(2m+1)$, so the first lot of queue is the expensive one. The gain ratio ξ and the weight λ_E enter only as a product, so believing a trim rate half the true one doubles the effective gain and can cross the boundary with no tuning change at all. Neither statement mentions the plant, because Equation (6.18) contains no plant dynamics.

Figure 6.2 verifies Equation (6.20) against the running application. The theory curves are the magnitude of the dominant root of Equation (6.19). The markers are measured on the actual etched linewidth after the 1.2 nm resist step of scenario 3, which carries no noise, with λ_E on the 0.01 grid of its slider and m set before the run. The envelope of the error is taken as $(e_k^2 - e_{k-1}e_{k+1})^{1/2}$, which for one oscillatory mode is proportional to $|z|^k$ whatever the phase, and its decay per lot is read between lots 26 and 56, where the trim stays between 10.0 and 14.5 s.

The fifteen measured points agree with the predicted decay to four decimal places in every case, the largest discrepancy being 0.9771 measured against 0.9770 predicted at $m = 4$ and $K = 0.30$. The cusp on each curve is the gain at which the two dominant roots collide and leave the real axis. Below the cusp the recovery is overdamped and slow, because one real root approaches unity as the gain goes to zero, and above it the error rings at the frequency ω.

Equation (6.20) can also be tested at the boundary itself rather than read off a curve. With λ_E set in the simulator to the predicted value, which the slider can only bracket, the envelope measured between lots 50 and 90 decays at exactly 1.0000 per lot at m = 1, 2, 3 and 4. Three percent below the predicted value the same measurement returns 0.9849, 0.9912, 0.9938 and 0.9952, and three percent above it returns 1.0149, 1.0086, 1.0061 and 1.0047, each the dominant root to four decimals. On the slider grid the envelope decays at λ_E = 0.99, 0.61, 0.44 and 0.34 and grows at 1.01, 0.62, 0.45 and 0.35, so the sign of the growth changes where Equation (6.20) says it does and nowhere else. That the weight and the gain ratio enter only as a product is a separate claim and is measured separately.

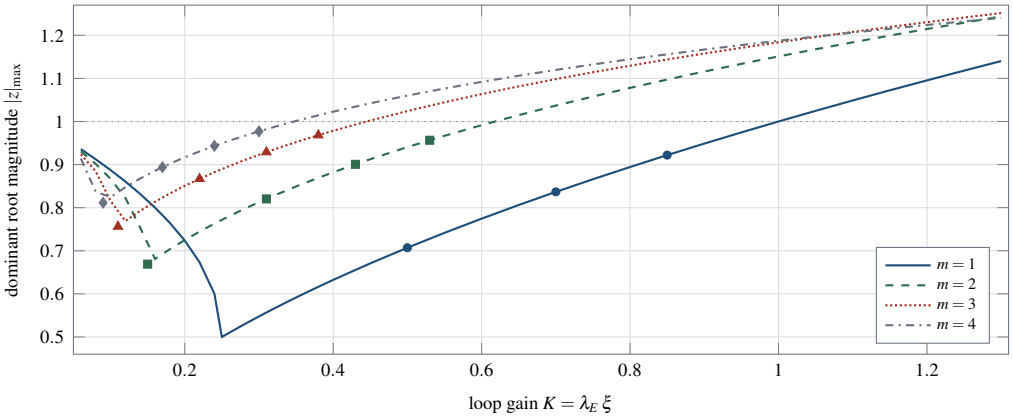


Figure 6.2 Closed-loop decay per lot against loop gain for four queue lengths. Lines are the dominant root magnitude of Equation (6.19), solid for $m = 1$, dashed for $m = 2$, dotted for $m = 3$ and dash-dotted for $m = 4$. Markers are envelope decay rates measured from the application after the resist step of scenario 3, circles for $m = 1$, squares for $m = 2$, triangles for $m = 3$ and diamonds for $m = 4$. The curves cross unity at 1.000, 0.618, 0.445 and 0.347, which are the four values Equation (6.20) predicts for $m = 1$ through 4.

Holding $\lambda_E \xi = 0.30$ at $m = 2$ while moving the believed trim rate through 1.00, 0.50 and 0.25 nm/s, which is ξ = 0.5, 1 and 2, gives a decay of 0.8127 per lot in all three runs, the predicted value. At $\lambda_E \xi = 0.44$, with weights of 0.88, 0.44 and 0.22, the three runs give 0.9066, again the predicted value, and in both sets the three runs agree lot by lot to within 3×10^{-14} nm. Three controllers whose believed trim rate differs by a factor of four recover at the same rate because the product is the same.

6.5.2 Dead Band, Decay and the Residual Against a Ramp

Equation (6.18) also fixes the shape of a step response. A disturbance appearing on lot j is not measured until lot j reaches CD-SEM-B, its reading is not delivered until lot $j + m$, and the recipe it produces is applied to lot $j + m + 1$. Exactly $m + 1$ lots are therefore produced at full error before the recipe moves at all. In the noiseless step scenario a resist batch change widens the developed line by 1.2 nm on lot 6, lots 6, 7 and 8 come out at an error of exactly 1.200 nm, and the feedback trim first moves from 10.000 to 10.720 s in time for lot 9, which is λ_E times the error divided by $\hat{r}$. The errors on lots 9, 10, 11 and 12 are 0.840, 0.480, 0.120 and −0.132 nm, which Equation (6.18) reproduces exactly at $K = 0.30$.

> **TRY IT IN THE APP 3 · One Step, No Noise**
>
> Load the scenario and press Run. Watch the control chart of the etched linewidth after the resist batch changes at 150 min. Lots 6, 7 and 8 come out 1.200 nm wide, because their readings are still in the queue. Then the trim steps from 10.000 to 10.720 s, and the error falls to 0.840, 0.480 and 0.120 nm before it overshoots to −0.132 nm on lot 12.

After the dead band the envelope decays geometrically at the dominant root magnitude, which is 0.8127 per lot at $m = 2$ and $K = 0.30$. Measured successive peaks of the same run give 0.8168, 0.7883, 0.8205, 0.8189, 0.8175 and 0.8162 per lot, spaced six or seven lots apart against the half period of 6.78 lots that the angle of the dominant root predicts. The error falls permanently below ten percent of the step 11 lots after the disturbance, which is 5.5 h of production. A useful contrast is that $(1 - \lambda_E) = 0.70$ is the decay the same controller would achieve with no queue, so two lots of delay slow the recovery by a factor of $\ln 0.70 / \ln 0.8127 = 1.72$ on top of the three lots lost to the dead band.

Against a ramp the intercept-only estimator leaves a constant offset. If the disturbance advances by d nanometers per lot, the recipe must advance at the same rate in steady state, and Equation (6.13) supplies that rate only through $\lambda_E e$, so the residual is Equation (6.21), the result of Equation (1.4) in Chapter 1.

$$e_\infty = \frac{d}{\lambda_E} \tag{6.21}$$

Equation (6.21) contains no m, so the queue costs nothing at all in steady-state ramp following. The drift scenario runs $\rho = 2.5$ nm per 1000 min, which is $d = 0.075$ nm per lot, and the measured residuals at λ_E = 0.15, 0.30, 0.45 and 0.60 are 0.5000, 0.2500, 0.1667 and 0.1255 nm against the 0.5000, 0.2500, 0.1667 and 0.1250 nm that Equation (6.21) predicts. The last of the four is the one still converging, because at $\lambda_E \xi = 0.60$ and $m = 2$ the dominant root is 0.9914 and the ring decays slowly. Lengthening the window from 50 to 200 lots moves that reading from 0.1260 through 0.1255 to 0.1249 nm, so it is approaching d/λ_E and not settling anywhere else. From lot 203 the ramp holds the trim at its 40 s limit, and from there on the residual stops being a control result at all. A double EWMA removes this term [14] and is not implemented, so the residual is available to be measured.

6.5.3 The Boundary Belongs to the Estimator

Equation (6.20) is a property of Equation (6.12), which corrects a fraction of whatever the newest reading says with no record of which lot produced it. The intercept form of Equation (6.11) regresses each reading against the recipe that actually made that lot, and it gives a different closed loop. Substituting $y_{k-m} = A - r_{trim} u_{k-m}$ into Equation (6.11), eliminating α with Equation (6.10) and writing the result in the error gives Equation (6.22).

$$e_{k+1} = (1 - \lambda_E)\, e_k + \lambda_E\, (1 - \xi)\, e_{k-m} \tag{6.22}$$

The delayed term in Equation (6.22) carries the factor $(1 - \xi)$ rather than the factor 1 of Equation (6.18). When the trim rate is known the delayed term vanishes identically, the loop collapses to the first-order recursion $e_{k+1} = (1 - \lambda_E) e_k$, and it is stable for every $0 < \lambda_E < 2$ at any queue length whatever. The queue still costs $m + 1$ lots of dead band and still costs response time against a drift. It no longer costs stability.

Measurement on the application at $m = 2$ confirms both statements. With $\lambda_E = 0.95$ and the newest-reading estimator the error after the resist step reaches 11.76 nm by lot 38, with the trim driven against its 0 s floor from lot 24, and the run-to-run loop is divergent, while the matched-lot estimator on the identical plant, the identical queue and the identical weight drives the error to numerical zero. Successive lot-to-lot error ratios measured with the matched-lot estimator at $m = 2$ are 0.5000 at $\lambda_E = 0.50$, 0.0500 at 0.95, 0.5000 at 1.50 and 0.9000 at 1.90, each constant from lot to lot and each equal to $|1 - \lambda_E|$ to four decimal places. Those are the ratios of a first-order recursion

with no delayed term in it at all, which is what Equation (6.22) predicts at $\xi = 1$, and they hold at $m = 4$ as well.

Nothing about the plant, the metrology or the queue is different between the two cases. The difference is whether the controller records which recipe produced the reading it just received, which is a database join rather than a control law and is exactly what a production advanced process control system maintains as tool, chamber and recipe context [12, 16]. Volume 8, Chapter 9 meets the same bookkeeping on a vial filler whose checkweigher results arrive while later doses are already committed. The factor between the two boundaries at $m = 2$ is $2/0.618 = 3.24$.

6.5.4 Two Integrators and One Measurement

The dose loop and the etch loop may both be pointed at the etched linewidth. Each is a competent controller and each corrects the whole of the error it sees. The dose update of Equation (6.14) moves the linewidth by exactly $-\lambda_L e$ and the trim update of Equation (6.13) moves it by exactly $-\lambda_E \xi e$, so the two loops superpose into Equation (6.23).

$$e_{k+1} = e_k - (\lambda_E \xi + \lambda_L)\, e_{k-m} \tag{6.23}$$

Two integrators on one measurement are one integrator whose gain is the sum, and Equation (6.20) applies to that sum. After the resist step of scenario 3 the measured envelope decay at $m = 2$ is 0.8822 per lot for $\lambda_E = \lambda_L = 0.20$ and 0.8822 per lot for $\lambda_E = 0.40$ with the dose loop switched off, the dominant root at a gain of 0.40, and the two etched-linewidth sequences agree lot by lot to 10^{-13} nm. At $\lambda_E = \lambda_L = 0.30$ the decay is 0.9914 per lot, just inside the 0.618 boundary, and at $\lambda_E = \lambda_L = 0.40$ the sum of 0.80 is outside it and the loop diverges until the trim reaches its 0 s floor on lot 43 and the dose its 16 mJ/cm^2 floor on lot 53. A configuration in which each loop is individually conservative is therefore unstable as a pair, and no single-loop tuning rule reveals it.

🕮 **Pause.** Try Exercise 6.2 before reading on.

The remedy is not detuning but partitioning. With the dose loop returned to the developed linewidth and the etch loop left on the etched linewidth, each has a measurement the other does not touch. Scenario 9 runs both loops at $\lambda_E = \lambda_L = 0.40$ against a 1.5 nm per 1000 min drift, and over its first 40 lots the shared-measurement configuration gives a spread of 12.012 nm, a capability index of 0.03 and 6 lots in specification. The dose and the trim move together in growing swings about nine lots apart. The trim first reaches zero on lot 14, the dose first reaches its 16 mJ/cm^2 floor on lot 32, and on lot 37 the trim sits at its 40 s ceiling with the etched line at 19.35 nm. Returning the dose loop to the developed linewidth and changing nothing else gives 1.846 nm, a capability index of 0.34 and 31 lots in specification over the same 40. Over five further noise seeds the shared configuration gives 4.57 to 8.98 nm with 13 to 17 lots in specification and the partitioned one 0.90 to 1.46 nm with 33 to 39, so the comparison does not rest on one realization.

The scenario carries the feedforward path switched off, and it has to. Equation (6.16) removes CD_{ADI} from the etched linewidth entirely whenever $\hat{r} = r_{trim}$, and with it the only route the dose loop has to the shared error. With feedforward on, the etched linewidth, the spread, the capability index and the in-specification count are identical to four decimal places whichever measurement the dose loop is pointed at, and the interaction this subsection is about becomes invisible in the controlled variable while remaining plainly visible in the dose.

6.6 Results

The eleven scenarios of Table 6.3 are ordered the way they are taught. Each pins a starting mode, a run-to-run configuration, both weights and the believed trim rate, the noise levels and any timed events, and each asks for a prediction before it is run, in the scenario format of Section 1.6.4. The exported record carries the true linewidth beside the measured one, the feedback half of the trim recipe beside the applied trim, all six continuous states, the loop gain $\lambda_E \xi$, and both stability boundaries side by side, the $2\sin[\pi/(2(2m+1))]$ of Equation (6.20) and the 2 of Equation (6.22), so that the

Table 6.3 The Eleven Scenarios and the Result Each One Isolates

#	Scenario	What it shows
1	The line by hand (MANUAL)	three static gains and three delays
2	On target, nothing moving	a fixed recipe on a plant that is not moving
3	One step, no noise	dead band of $m+1$ lots, geometric decay
4	A slow drift, no run-to-run	ramp offset, Equation (6.21)
5	The chamber clean	a step in an unmeasured state
6	Choosing λ_E at $m = 2$	the boundary, and the estimator switch
7	What the queue costs, $m = 4$	λ_E fixed, only m moved
8	Lithography noise, $\sigma_L = 1.3$ nm	the λ trade and feedforward
9	Two integrators, one measurement	Equation (6.23)
10	One lot in three measured	sampling, and what it does to feedforward
11	Graded challenge, three upsets	clean, resist step and drift together

estimator comparison can be plotted without switching the controller and Equation (6.18) can be rebuilt in a spreadsheet.

Scenario 1 establishes the architecture before any recipe controller runs. A dose change reaches the developed linewidth and the actual etched linewidth on the very next lot, and the measured etched linewidth two lots after that, when CD-SEM-B reports it. A trim change reaches the actual etched linewidth on the next lot and its measurement two lots later, and never touches the developed linewidth at all, which is the structural zero in the gain matrix and the reason the two loops can be partitioned. The plate is the third input and it is on a different clock. Moving the setpoint 0.5 °C in automatic puts the plate inside 0.1 °C of its new value after 0.32 min and inside 0.01 °C after 0.56 min, and the developed linewidth of the next lot moves by exactly −1.600 nm.

6.6.1 What the Queue and the Weight Buy

Table 6.4 collects the application scorecard for the configurations that matter, each the shipped scenario with the stated change made before the run. Every figure is the on-screen spread over the most recent 40 lots, the capability index against a specification half-width of 2.0 nm, and the fraction of all lots produced that fell inside that specification, read at the stated lot count. Figure 6.3 draws the same scorecard as bars.

The queue block reads across in one direction. Holding λ_E at 0.40 and moving only the metrology delay takes the spread from 1.652 to 3.812 nm and the in-specification fraction from 78 to 47 % between three lots of queue and four. Nothing else changed, and over five further noise seeds four lots of queue gives 2.78 to 3.80 nm against 1.24 to 2.10 nm at three. At $m = 3$ the gain of 0.40 is inside the boundary of 0.445 and the loop is merely slow. At $m = 4$ it is outside the boundary of 0.347 and the loop diverges until the trim command reaches a limit. One lot of queue, with no change to the controller whatever, is the difference between a line that works and a line that does not, and that pair of numbers is the business case for integrated metrology.

> **TRY IT IN THE APP** **7 · What the Queue Costs**
>
> Press Run and let 60 lots pass through the wet clean with four lots of after-etch queue and a weight of 0.40. The scorecard reads a spread of 3.812 nm and 47 % of lots in specification. Reset, set After-etch metrology delay to 3, and run again: 1.652 nm and 78 %, with nothing else changed.

The gain mismatch block reads the same way with the plant held fixed and only the believed trim rate changed. The three configurations have loop gains of 0.25, 0.50 and 1.00 against the 0.618 boundary, and the spread rises from 1.167 to 7.712 nm across them. A controller commissioned correctly and then left alone while the etcher changes is the practical route into the last row, which is the sensitivity the earliest analyses of the EWMA controller identified [4].

Table 6.4 Measured Scorecard. σ Is the Spread of the Most Recent 40 Lots and the In-Specification Fraction Counts Every Lot Produced. Read at the Lot Count in the First Column.

Lots	Configuration	σ (nm)	C_{pk}	In spec
Lithography noise, $\sigma_L = 1.3$ nm				
60	no run-to-run	1.476	0.45	80 %
60	feedback $\lambda_E = 0.15$	1.599	0.41	77 %
60	feedback $\lambda_E = 0.30$	1.852	0.34	67 %
60	feedback $\lambda_E = 0.60$	2.663	0.24	47 %
60	$\lambda_E = 0.30$ + feedforward	**0.850**	**0.76**	100 %
Wet clean of the etch chamber				
50	no run-to-run	0.782	0.59	100 %
50	etch loop $\lambda_E = 0.40$	0.411	1.43	100 %
Lithography drift, 2.5 nm per 1000 min				
60	no run-to-run	0.877	−0.37	45 %
60	etch loop $\lambda_E = 0.30$	0.001	—	100 %
Choosing the weight at $m = 2$, drift and full noise				
60	$\lambda_E = 0.30$, newest reading	1.265	0.50	82 %
60	$\lambda_E = 0.95$, newest reading	8.002	0.00	15 %
60	$\lambda_E = 0.95$, matched lot	1.548	0.41	82 %
Queue length at $\lambda_E = 0.40$ through a wet clean				
60	$m = 2$	1.497	0.39	80 %
60	$m = 3$	1.652	0.37	78 %
60	$m = 4$	3.812	0.08	47 %
Gain mismatch at $\lambda_E = 0.50$, $m = 2$				
60	$\hat{r} = 1.00$ nm/s, $\xi = 0.5$	1.167	0.53	83 %
60	$\hat{r} = 0.50$ nm/s, $\xi = 1$	1.907	0.34	60 %
60	$\hat{r} = 0.25$ nm/s, $\xi = 2$	7.712	−0.00	13 %
Metrology sampling with feedforward on				
50	every lot measured	1.050	0.61	96 %
50	one lot in two	1.270	0.42	86 %
50	one lot in three	1.092	0.44	88 %
50	every lot, feedforward off	1.525	0.43	82 %

✦ DID YOU KNOW? A Wrong Model That Helps

From the simulator: in Table 6.4, at a weight of 0.50 behind a two-lot queue, a controller that believes the trim rate is twice its true value holds a spread of 1.167 nm with 83 % of lots in specification. The controller that knows the true rate holds 1.907 nm and 60 %.

The weight block contains the chapter's sharpest comparison. At $\lambda_E = 0.95$ and $m = 2$ the newest-reading estimator produces a spread of 8.002 nm and 15 % of lots in specification. The matched-lot estimator, with the same weight, the same queue, the same noise and the same drift, produces 1.548 nm and 82 %. That is the in-specification fraction of the properly detuned newest-reading loop at $\lambda_E = 0.30$, with a spread 22 % larger because a weight near one passes most of the metrology noise into the recipe. Over five further noise seeds the newest reading at 0.95, the matched lot at 0.95 and the newest reading at 0.30 keep 20 to 27, 83 to 95 and 90 to 100 % of lots in specification.

The two disturbance blocks in the middle of Table 6.4 are the reason the layer exists at all. A wet clean is a step in a variable no instrument reports, and closing the etch loop at $\lambda_E = 0.40$ takes the spread over 50 lots from 0.782 to 0.411 nm and the capability index from 0.59 to 1.43. A 2.5 nm per 1000 min lithography ramp is worse, because a fixed recipe simply walks out of the band. The

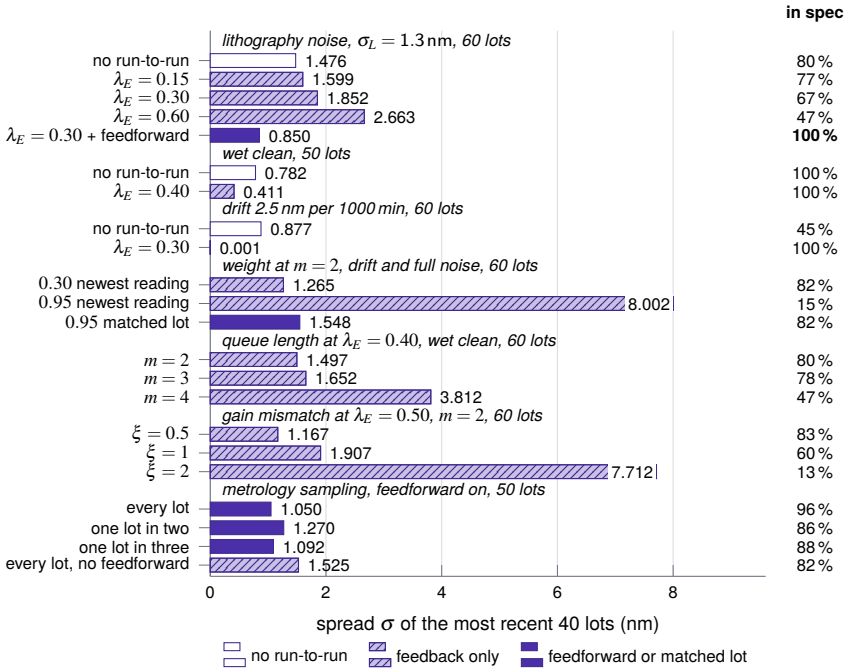


Figure 6.3 The scorecard of Table 6.4 as bars: the spread over the most recent 40 lots, with the share of all lots in specification in the column at the right. Open bars run with no run-to-run loop, hatched bars with feedback alone, solid bars with feedforward or the matched-lot estimator. Notice that the three longest bars, 8.002, 7.712 and 3.812 nm, are the three configurations beyond the delay-limited boundary, by too large a weight, too small a believed trim rate or too long a queue, and that each solid bar is shorter than the hatched bar with the same weight and sampling.

first lot outside specification arrives at $t = 810$ min, against the 800 min the ramp rate and the 2.0 nm tolerance predict, and by 60 lots the capability index is -0.37 with 45 % of lots inside the band. The same loop at $\lambda_E = 0.30$ holds every lot inside it at a spread of 0.001 nm, which makes the capability index meaningless and puts the whole of the error into the 0.250 nm mean offset that Equation (6.21) predicts.

6.6.2 Feedforward Is Not on the Weight Trade Curve

The first block of Table 6.4 is the classical run-to-run trade and then something that is not on it. With 1.3 nm of lot-to-lot lithography scatter the etcher inherits in full, raising λ_E from 0.15 to 0.60 raises the spread from 1.599 to 2.663 nm and drops the in-specification fraction from 77 to 47 %, because a fast discrete integrator writes measurement noise into the recipe and the recipe writes it into the product. The uncontrolled line is better than any of the three feedback tunings, at 1.476 nm and 80 %, which is the honest answer whenever a loop is asked to reject a disturbance that is white at its own sample rate. The margin over $\lambda_E = 0.15$ is small. Over five further noise seeds the uncontrolled spread is below the $\lambda_E = 0.15$ spread in every case but only by 1 to 8 %, inside the scatter of a 40-lot estimate, while $\lambda_E = 0.60$ is the worst of the four in every seed.

Switching feedforward on at $\lambda_E = 0.30$, with the feedback tuning untouched, gives 0.850 nm, a capability index of 0.76 and every one of the 60 lots in specification. Over five further noise seeds the spread is 0.57 to 0.89 nm. The mechanism is Equation (6.16) and not a filter. With λ_E set to zero so that only the feedforward path is active, the measured spread is 0.639 nm against the $\sqrt{\sigma_M^2 + \sigma_E^2} = 0.570$ nm that Equation (6.16) predicts for a 40-lot sample (0.51 to 0.62 nm over

five further seeds), and the entire 1.3 nm lithography term has left the expression rather than being attenuated inside a loop. The remaining difference between 0.639 and 0.850 nm is what the feedback loop adds back by chasing its own after-etch metrology noise.

6.6.3 Sampling Costs Twice

Metrology is expensive and slow, so a production line measures a fraction of its lots. The expected consequence is that the feedback loop corrects once every $n_s T_{takt}$ minutes instead of once every T_{takt}, so the wall-clock recovery from a disturbance lengthens by n_s with no change to λ_E. The unexpected consequence is in Equation (6.15). A lot with no after-develop reading has no feedforward correction available, so $n_s - 1$ lots in every n_s are etched with no knowledge of their own incoming linewidth.

The sampling block of Table 6.4 separates the two effects. Measuring every lot with feedforward on gives 1.050 nm and 96 % in specification. Turning the feedforward path off while still measuring every lot gives 1.525 nm and 82 %, which is the value of the feedforward path alone. Measuring one lot in three with feedforward nominally on gives 1.092 nm and 88 %, which for this seed sits between the two because the path is delivering on a third of the material. The one-in-two row at 1.270 nm is worse than the one-in-three row rather than better, and that ordering should not be read as a result. Each row is one seeded realization of 50 lots, and the standard error of a spread estimated from 40 samples is $\sigma/\sqrt{2 \times 39} = 11\,\%$ of the value, so the two rows are about one and a half standard errors apart. What is outside that band is the 1.525 nm of the last row against the 1.050 nm of the first, although not by a wide margin. Over five further noise seeds feedforward lowers the every-lot spread in four of them, to 0.85 to 1.28 nm from 0.93 to 1.50 nm without it. A sampling plan justified against the feedback loop alone is therefore the wrong calculation, and the error is in the direction of sampling too little.

6.6.4 The Same Integral Action With a Live Measurement

The bake plate is the control in this experiment. An unmeasured −25 W draft across the plate moves its open-loop steady state by $q_{load}/U_p = 5.95\,^{\circ}\mathrm{C}$, which through k_T is 19.0 nm of linewidth, and with the heater held at the 369.6 W that balances the plate the measured temperature falls from 110.0000 to 104.0476 °C over the 50 min that follow. With Equation (6.17) running at the default tuning the same draft produces a peak deviation of 1.199 °C at 0.44 min and the plate is permanently inside 0.1 °C 2.96 min after the draft and inside 0.01 °C after 5.18 min. The peak is 3.84 nm of linewidth equivalent and it lasts three minutes out of a thirty-minute lot.

Compare that with the etch loop meeting the 1.2 nm resist step of scenario 3 at $\lambda_E = 0.30$, which commits three lots at full error and needs 11 lots, or 5.5 h, to fall permanently inside ten percent of the step. Both are integral controllers, both act on the same product characteristic, and the ratio of recovery times is about 110. A wet clean is slower still for the same loop, which needs 26 lots to fall permanently inside ten percent of its 2.0 nm step because the walls go on seasoning underneath it. The entire difference is that one of them is measured every 2.4 s and the other is measured every 30 min with the answer two lots old.

A view of that wet clean recovery in progress appears in Figure 6.4. The disturbance driving all of it shows up in the wall-condition trend strip and in no instrument reading anywhere else, which is the situation run-to-run control was invented for.

6.6.5 The Graded Challenge

Scenario 11 combines a wet clean at 150 min, a resist change at 450 min and a 1.2 nm per 1000 min drift. It is passed when the measured etched linewidth stays within ±1.2 nm of 45 nm from 594 min (about lot 20) to 900 min, with the model parameters and disturbances as given. With the shipped configuration, the etch loop alone at $\lambda_E = 0.08$ without feedforward, the worst measured error from 594 min is 1.89 nm, 0.69 nm outside the band. With the suggested configuration, both recipe loops with $\lambda_E = 0.30$ and $\lambda_L = 0.45$ and feedforward on, it is 0.88 nm, inside with 0.32 nm to spare. Over five further noise seeds the suggested configuration passes every time (0.45 to 0.77 nm) and the

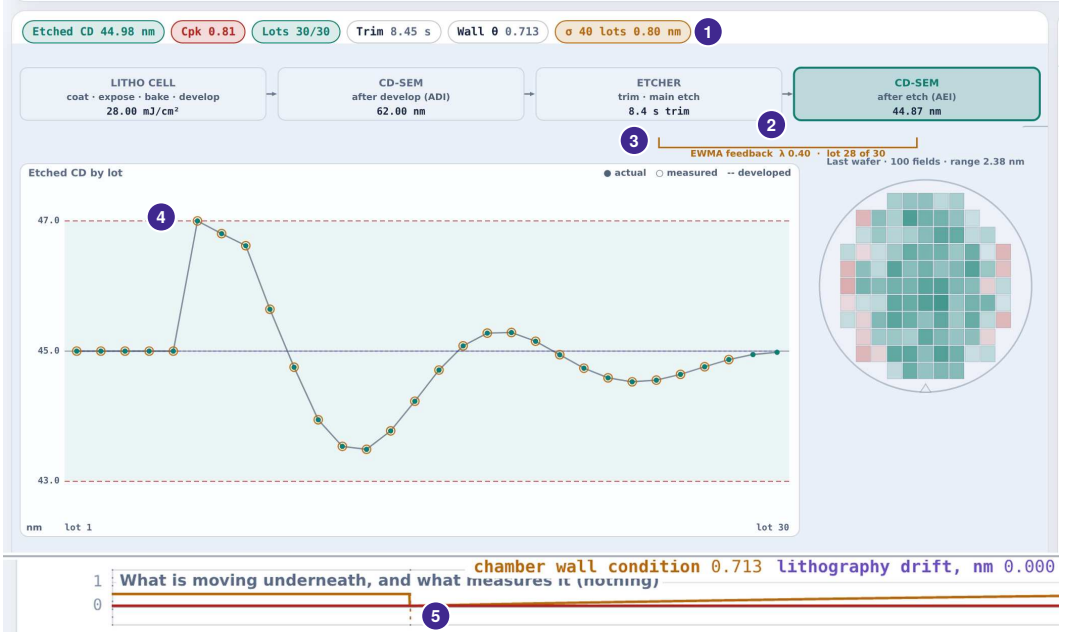


Figure 6.4 Recovery from a wet clean of the etch chamber, scenario 5 with the etch run-to-run loop set to $\lambda_E = 0.40$ before the run and a two-lot metrology queue, at $t = 15.00$ h and lot 30 of 30. The chamber is wet cleaned at $t = 150$ min, the instant lot 6 is etched, so lot 6 is the first lot on bare walls and comes out at exactly 47.00 nm, on the upper specification limit, which the application counts as inside it. The control chart shows the recipe pulling the line back through an undershoot to 43.49 nm at lot 13 before it settles. The trim time has been driven from 10.00 to 8.45 s to follow the wall condition back to 0.713, the reading being acted on is 44.87 nm from lot 28, and the scorecard reads $\sigma = 0.80$ nm with $C_{pk} = 0.81$ and 30 of 30 lots in specification. The view is cropped to the line, the control chart and, below the gray rule, the trend strip of what moves underneath the line. Callout 1 marks the scorecard, 2 the reading CD-SEM-B delivers, 3 the lot that reading belongs to, 4 lot 6, the first lot on bare walls, and 5 the wall condition, which is measured by nothing on the line.

shipped one fails three times. The feedforward path carries most of the improvement, because the resist change is a lithography disturbance that CD-SEM-A reads before the lot is etched, and the etch loop is left with the wet clean, which nothing on the line measures.

6.7 Technology in Practice

The loops of Section 6.4 run on two linewidth microscopes, two recipe numbers, the plate's controller and the fab's run-to-run system, drawn together in Figure 6.5. For a run-to-run loop the sensor is a metrology tool and the final element is a number in a recipe.

What Measures It

Both recipe loops read a critical-dimension scanning electron microscope, CD-SEM-A after develop and CD-SEM-B after etch, which Equation (6.8) reduces to a shared offset β_M and a repeatability of 0.45 nm on the lot mean.

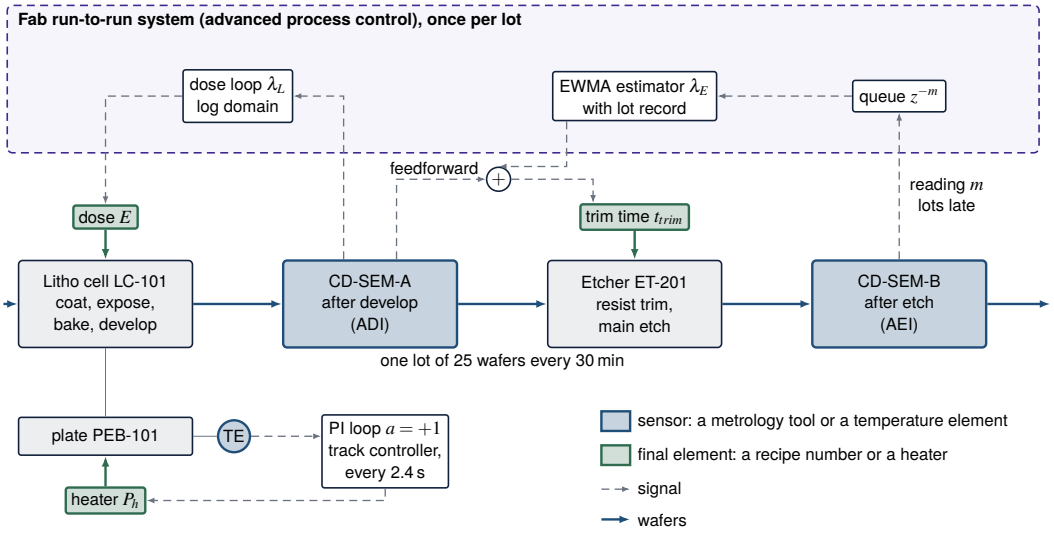


Figure 6.5 The three loops of the simulated line drawn with their hardware. CD-SEM-A reads the developed linewidth of every sampled lot, feeds the dose loop and, through the feedforward sum, the trim of the same lot. CD-SEM-B reads the etched linewidth, and its reading reaches the EWMA estimator only through the queue z^{-m}. Both recipe loops run once per lot in the fab's run-to-run system, while the plate loop runs in the track's own controller every 2.4 s. Notice that the two final elements of the recipe loops are numbers written into a recipe before a lot runs, not devices that move during it.

> **TECHNOLOGY** **Sensor: A Critical-Dimension Scanning Electron Microscope**
>
> A scanning electron microscope sweeps a finely focused electron beam over the wafer, and the secondary and backscattered electrons it releases give the shape and size of a feature with sub-nanometer resolution. Microscopes built for the fab land their electrons at low energy, typically 300 to 800 electronvolts, for repeatable imaging that does not destroy the feature. Their performance is limited less by the electron optics than by drift, vibration, beam damage, charging and contamination. For most work, precision, stability and matching across the fleet of tools count for more than absolute accuracy. Two different techniques can each repeat to better than a tenth of a nanometer and still disagree by more than three nanometers on the same line [1].

The offset β_M stands for the matching error of such a tool, and a reading 0.8 nm high settles the true line at 44.200 nm.

What Moves It

The final element of the etch loop is the trim time of ET-201, clipped to 0–40 s and applied exactly as written. The dose loop writes a second recipe number at the scanner, clipped to 16–44 mJ/cm^2.

> **TECHNOLOGY** **Final Element: A Number in the Etch Recipe**
>
> An etch chamber runs on setpoints for pressures, component temperatures and gas flows. They are developed by experiment for each product and compiled into recipes that are sent to the fabs, where for a long time they were applied open loop [20]. In gate patterning a photoresist trim step narrows the resist line before the main etch, while keeping as much resist thickness as possible for the steps that follow [11]. The same recipe etches differently over time as material deposits on the chamber walls between scheduled maintenance, and the first one or two wafers of a lot often differ from the rest [20].

The wall condition θ of Equation (6.4) is the model's form of that wall deposit, and the wet clean is its maintenance event.

Where the Controller Runs

The plate loop runs in the track's own controller. The recipe loops run above the tools in the fab's advanced process control system, which combines run-to-run control with fault detection and classification [11]. Supervisory and lower-level loops run with cycle times from milliseconds to minutes, and what counts as real time has to be stated for each application [21]. The app executes the plate loop every 2.4 s, while each recipe moves once per lot, every 30 min. In the run-to-run system the structure of this chapter is three blocks: an estimator per loop, a record that pairs each reading with the recipe of its own lot, and a sum that adds the feedforward term of Equation (6.15) to the feedback trim.

> **IN PRACTICE** **Waiting for Metrology, Starting a Thread**
>
> Metrology tools are expensive and shared, so wafers wait to be measured, the most critical first. The delay in front of a run-to-run loop changes from run to run, in most cases between zero and five runs with a mean of one or two, and a longer wait usually raises an alarm for an operator to measure a wafer [11]. A fab runs a separate moving-average controller for each thread, one combination of product and prior processing path. A new thread's first estimate is set by hand or by a pilot run, which costs, or set automatically at the risk of outliers in its first runs [16].

What Is Modernizing It

Three developments bear on this loop. Integrated metrology adds a measuring tool to each process tool so that product can be measured during or after the step, and by 2011 its wide adoption had been held back by set-up cost, limited cooperation between chip makers and tool makers, and longer cycle times [20]. Scatterometry, a variant of ellipsometry, fits the light scattered from an array of lines to a library of simulated results and returns the linewidth and height without an image [1]. Virtual metrology predicts post-process metrology variables from process and wafer state information [1], such as the light emitted by an etch plasma, for every wafer, while real metrology covers only some wafers and arrives hours or even days late [20].

6.8 Limitations

The application is not a lithography simulator. Equation (6.1) is four terms, and a real linewidth model resolves the aerial image, the resist chemistry, the development kinetics and the mask through a physical simulation, so no question about illumination, pitch or mask bias can be asked here. The etcher of Equation (6.2) is one scalar, and questions about the discharge, the neutral chemistry or the profile belong to the plasma etch chamber of Chapter 7. The across-wafer signature of Equation (6.9) is decorative. It does not respond to the controller, so spatial recipe control of the kind that corrects the map itself [18] is outside reach. The estimators carry an intercept and no drift term, so the residual of Equation (6.21) is larger than a production controller would show. Both estimators also assume that the gain ratio is constant and known to whatever accuracy $\hat{r}$ provides, while a production system re-identifies the gain as a second and slower loop that can argue with the first one. The lot cadence is fixed, the line never blocks or reorders, and every lot is one number, so nothing about dispatch, lot splitting or wafer-level sampling can be studied. The economics are a flat value per lot in specification and a flat cost per lot outside it, which is not how a rework decision is actually made. Reproducibility itself is unrealistic, because freezing the seed and the step gives one trajectory on every machine while a fab gives a distribution. Section 6.3.6 names the failure mode of each assumption separately.

Three elements of the scenarios are stylized on purpose. A wet clean that empties the chamber wall state instantly idealizes a maintenance event that takes hours, but it isolates a step in a variable no instrument reports. A metrology queue set with a slider would in practice be an outcome of

dispatch and tool availability, although making it a setting is what allows the boundary to be measured against Equation (6.20) one lot at a time. A believed trim rate set to half the true one is unlikely in a commissioned line, while it is the cleanest available way to separate λ_E from ξ in Equation (6.18).

OPEN CHALLENGE Measuring Devices That Have Gone Three-Dimensional

As transistors shrink and turn three-dimensional, the industry approaches the point where the position and type of each atom in a device need to be known, for billions of devices on each chip. No single instrument has the resolution, speed and low uncertainty to measure every parameter, and different techniques can disagree on the same feature, because their error sources and their definitions of what is measured differ [1]. Progress would be measurements traceable to a common reference, with agreed definitions of each quantity [1]. The loops of this chapter trust one linewidth number per lot completely.

6.9 Conclusion

This chapter studied control on a measurement that arrives late, on a four-station lithography and etch line. The simulator integrates six continuous states covering the etch chamber wall condition, the bake plate temperature, a lithography drift and three calibration errors, and evaluates the lot arithmetic once every 30 min at a design recipe of 28.00 mJ/cm^2 and 10.00 s that puts the developed line at 62.00 nm, the etch bias at 17.00 nm and the etched line at 45.00 nm.

The exponentially weighted moving average recipe controller is a discrete integral controller whose closed loop is $e_{k+1} = e_k - \lambda_E \xi e_{k-m}$, and the stability boundary $2\sin[\pi/(2(2m+1))]$ that follows returns 2, 1, 0.618, 0.445 and 0.347 for zero through four lots of queue. Fifteen envelope decay rates measured on the running application agree with the dominant root of that recursion to four decimal places. Regressing each reading against the recipe that produced that lot replaces the recursion with $e_{k+1} = (1-\lambda_E)e_k + \lambda_E(1-\xi)e_{k-m}$, which with the trim rate known is stable for any weight below 2 at any queue length, and the same plant that diverges at $\lambda_E = 0.95$ with the newest-reading estimator holds 1.548 nm of spread with the matched-lot one.

Feedforward from the developed linewidth removes the lithography term from the etched linewidth outright rather than filtering it, taking the spread from 1.852 to 0.850 nm and the in-specification fraction from 67 to 100 % with the feedback tuning untouched. Two integrators pointed at one measurement add their weights, so a pair that is individually conservative at 0.40 each is unstable at a sum of 0.80 and takes the spread over 40 lots from 1.846 to 12.012 nm as the recipes swing between their limits.

A fuller model would add a drift term to the estimator, a gain re-identification loop, a dispatch model that makes the queue an outcome rather than a setting, and across-wafer control of the signature that is presently display only. On this line the cost of a queued measurement is something the reader measures rather than accepts.

KEY POINTS

- A lot is measured after it leaves the etcher and the answer arrives two lots later, so three lots are made at full error before the recipe moves.
- The largest usable loop gain, $2\sin[\pi/(2(2m+1))]$, is 1, 0.618, 0.445 and 0.347 for one to four lots of queue.
- Pairing each reading with the recipe of its own lot keeps the loop stable for any weight below 2 at any queue length when the trim rate is known.
- Feedforward from the developed linewidth removes the lithography term, taking the spread from 1.852 to 0.850 nm.
- Two integrators on one measurement add their weights: 0.40 each is unstable as a pair.

Where else this applies. Late results matched to their commands in Volume 8, Chapter 9, a slow analyzer in Volume 5, Chapter 3, and run-to-run channels in Chapter 5.

References

[1] N. G. Orji, M. Badaroglu, B. M. Barnes, C. Beitia, B. D. Bunday, U. Celano, R. J. Kline, M. Neisser, Y. Obeng, and A. E. Vladar, "Metrology for the next generation of semiconductor devices," *Nature Electronics*, vol. 1, no. 10, pp. 532–547, 2018.

[2] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] T. W. Kim and E. S. Aydil, "Effects of chamber wall conditions on Cl concentration and Si etch rate uniformity in plasma etching reactors," *Journal of The Electrochemical Society*, vol. 150, no. 7, p. G418, 2003.

[9] 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. Proceedings of SPIE, vol. 5751. SPIE, 2005, p. 1092.

[10] 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.

[11] 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

[12] D. Boning, W. Moyne, T. Smith, J. Moyne, and A. Hurwitz, "Practical issues in run by run process control," in *Proceedings of SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 1995, pp. 201–208.

[13] J. Moyne, E. Del Castillo, and A. M. Hurwitz, Eds., *Run-to-Run Control in Semiconductor Manufacturing*. Boca Raton, FL: CRC Press, 2001.

[14] E. Del Castillo, "Long run and transient analysis of a double EWMA feedback controller," *IIE Transactions*, vol. 31, no. 12, pp. 1157–1169, 1999.

[15] S. Adivikolanu and E. Zafiriou, "Extensions and performance/robustness tradeoffs of the EWMA run-to-run controller by using the internal model control structure," *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 23, no. 1, pp. 56–68, 2000.

[16] F. Harirchi, T. L. Vincent, A. Subramanian, K. Poolla, and B. Stirton, "On the initialization of threaded run-to-run control of semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 27, no. 4, pp. 515–522, 2014.

[17] J. W. Bossung, "Projection printing characterization," in *Developments in Semiconductor Microlithography II*, ser. Proceedings of SPIE, vol. 0100. SPIE, 1977, pp. 80–85.

[18] Q. Zhang, K. Poolla, and C. J. Spanos, "One step forward from run-to-run critical dimension control: Across-wafer level critical dimension control through lithography and etch process," *Journal of Process Control*, vol. 18, no. 10, pp. 937–945, 2008.

[19] S. Ruegsegger, A. Wagner, J. S. Freudenberg, and D. S. Grimard, "Feedforward control for reduced run-to-run variation in microelectronics manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 12, no. 4, pp. 493–502, 1999.

[20] S. Lynn, "Virtual metrology for plasma etch processes," Ph.D. dissertation, National University of Ireland Maynooth, 2011. [Online]. Available: https://mural.maynoothuniversity.ie/id/eprint/2657/

[21] 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.

6.10 Exercises

6.1 Concept check · 1 min

Which of these quantities moves the etched linewidth of the patterning line but is reported by no instrument anywhere on it?

(a) The chamber wall condition θ of etcher ET-201
(b) The developed linewidth read by CD-SEM-A
(c) The plate temperature of bake plate PEB-101
(d) The etched linewidth read by CD-SEM-B

6.2 Concept check · 2 min

In scenario 9 the dose loop and the etch loop each run at a weight of 0.40, and both are pointed at the etched linewidth behind the two-lot queue. Each loop alone would be stable. Why does the pair diverge?

(a) The dose loop has no queue, so it outruns the etch loop and drags it along
(b) Both correct the same delayed error, so they act as one integrator of gain 0.80
(c) The dose stops at its 16 mJ/cm^2 floor and its integrator winds up there
(d) The feedforward path adds the dose correction to the trim a second time

6.3 Concept check · 2 min

With 1.3 nm of lot-to-lot lithography scatter, which element of the line removes that scatter from the etched linewidth instead of filtering it?

(a) Raising λ_E so the etch loop corrects each lot faster
(b) The PI loop that holds the bake plate at its setpoint
(c) Feeding the measured developed linewidth forward to the trim of the same lot
(d) Pointing the dose loop at the etched linewidth as well

6.4 Theory and calculation · 6 min

Use Equations (6.18) and (6.20) for the newest-reading estimator.

(a) Compute the boundary $(\lambda_E \xi)_{\max}$ for $m = 0$ to 4 to three decimals, and the ratio of the $m = 0$ value to the $m = 1$ value and of the $m = 3$ value to the $m = 4$ value, to two decimals. *Checkpoint:* 2.000, 1.000, 0.618, 0.445 and 0.347; ratios 2.00 and 1.28.

(b) The controller believes $\hat{r} = 0.25\,\text{nm/s}$ while the true trim rate is 0.50 nm/s, and $m = 2$. Find the largest λ_E on the slider's grid of 0.01 that keeps the loop stable. *Checkpoint:* $\xi = 2$, so $\lambda_E < 0.309$: 0.30.

(c) Find the period of the marginal oscillation at $m = 2$, in lots and in hours at a lot cadence of 30 min. *Checkpoint:* 10 lots, 5 h.

6.5 Theory and calculation · 7 min

Scenario 3 has no noise. A resist change widens the developed line by 1.2 nm from lot 6, the queue is $m = 2$, $\lambda_E = 0.30$, $\hat{r} = r_{trim} = 0.50\,\text{nm/s}$, and the feedback trim is 10.00 s before the step.

(a) With the newest-reading estimator, use Equation (6.18) to find the error on lots 6 to 12, to three decimals, and use Equation (6.13) for the trim applied to lot 9. *Checkpoint:* 1.200, 1.200, 1.200, 0.840, 0.480, 0.120 and −0.132 nm; 10.720 s.

(b) With the matched-lot estimator, use Equation (6.22) to find the error on lots 9, 10 and 11, to three decimals. *Checkpoint:* 0.840, 0.588 and 0.412 nm.

(c) Explain why both estimators give 0.840 on lot 9 and part afterward. *Checkpoint:* The first correction answers the first reading in both. From lot 10 the newest reading keeps correcting errors that earlier corrections have already answered.

6.6 Theory and calculation · 6 min

Scenario 4 drifts the lithography at 2.5 nm per 1000 min. The lot cadence is 30 min and lot k is etched at $30(k-1)$ minutes, the tolerance is ± 2.0 nm, the trim rate is 0.50 nm/s, and the line starts at the design point with 10.00 s of trim and the newest-reading estimator.

(a) With a fixed recipe, which lot is the first outside specification, and at what time is it etched? *Checkpoint:* Lot 28, at 810 min.

(b) With the etch loop on, find the settled offset at λ_E = 0.15, 0.30 and 0.45, to three decimals. *Checkpoint:* 0.500, 0.250 and 0.167 nm.

(c) Once the offset has settled the trim must rise at a steady rate. Find that rate in seconds per lot, and the first lot at which $\lambda_E = 0.60$ asks for the 40 s ceiling. *Checkpoint:* 0.15 s per lot; lot 203.

6.7 App challenge · 8 min

Load scenario 6, "Choosing λ, and the boundary the queue sets" (default seed). It runs the etch loop at $\lambda_E = 0.95$ behind a two-lot queue, with full noise and a 1.5 nm per 1000 min drift. Run it as loaded at 20 min/s and press Pause while the Lots run tile reads 60. Lot 60 runs from 1770 to 1800 min, and the tiles hold their values for the whole lot. Then press Reset, set EWMA weight · etch loop to 0.30, and run to lot 60 again. Last, press Reset, set the weight back to 0.95, set The etch estimator uses to The matched lot (in Advanced it sits under Metrology and sampling), and run to lot 60 a third time.

(a) Read the σ (40 lots) and Lots in spec tiles after each run, to 0.01 nm and 1 %. *Checkpoint:* 8.00 nm and 15 %; 1.27 nm and 82 %; 1.55 nm and 82 %.

(b) Give the loop gain $\lambda_E \xi$ of each run and the boundary of its estimator, and say which run is outside it. *Checkpoint:* 0.95 against 0.618, outside; 0.30 against 0.618; 0.95 against 2.

(c) The third run matches the second in lots in specification but not in spread. Explain the larger spread. *Checkpoint:* A weight near one passes most of the after-etch metrology noise into the recipe.

6.8 App challenge · 9 min

Load scenario 9, "Two integrators, one measurement" (default seed). Both recipe loops run at a weight of 0.40, feedforward is off, and The dose loop controls (in Advanced) is set to Etched CD. Run at 20 min/s, press Pause while the Lots run tile reads 40, and export the CSV.

(a) Read the σ (40 lots), Cpk and Lots in spec tiles. *Checkpoint:* 12.01 nm, 0.03 and 15 %.

(b) In the CSV, find the first `lot_no` whose `trim_s` is 0 and the first whose `dose_mJ` is 16. *Checkpoint:* Lots 14 and 32.

(c) Press Reset, set The dose loop controls to Developed CD, change nothing else, and run to lot 40 again. Read the same three tiles and explain the change. *Checkpoint:* 1.85 nm, 0.34 and 78 %. Each loop now has its own measurement.

6.9 App challenge · 8 min

Load scenario 8, "Lithography noise, and what feedforward does to it" (default seed). It has $\sigma_L = 1.3$ nm and the etch loop at $\lambda_E = 0.30$. Run at 20 min/s, press Pause while the Lots run tile reads 60, and read the σ (40 lots) and Lots in spec tiles. Press Reset, set Feedforward ADI → etch to On, and repeat. Press Reset again, keep feedforward On, set EWMA weight · etch loop to 0, and repeat.

(a) Give σ and Lots in spec for the three runs, to 0.01 nm and 1 %. *Checkpoint:* 1.85 nm, 67 %; 0.85 nm, 100 %; 0.64 nm, 100 %.

(b) Compute the spread Equation (6.16) predicts with feedforward alone, to three decimals, and judge the third run against it, given that a spread estimated from 40 lots has a standard error of about 11 %. *Checkpoint:* $\sqrt{\sigma_M^2 + \sigma_E^2} = 0.570\,\text{nm}$, and 0.64 is about 12 % high, about one standard error.

(c) Explain why the second run is wider than the third. *Checkpoint:* The feedback loop chases its own after-etch metrology noise.

6.10 Application area · 3 min

Table 6.4 runs the etch loop at $\lambda_E = 0.40$ through a wet clean. With three lots of queue the line holds a spread of 1.652 nm and 78 % of lots in specification. With four it holds 3.812 nm and 47 %.

(a) Explain from Equation (6.20) why one lot of queue makes this difference. *Checkpoint:* 0.40 is inside 0.445 at $m = 3$ and outside 0.347 at $m = 4$.

(b) Name two changes that would let the line run a weight of 0.40 with confidence, other than lowering the weight. *Checkpoint:* Shorten the queue (integrated metrology), or use the matched-lot estimator.

6.11 Application area · 3 min

A metrology group proposes measuring one lot in three on this line to free CD-SEM time, and supports it with the feedback loop alone: the loop will simply correct less often.

(a) Name the two costs of the plan on this line. *Checkpoint:* The feedback loop corrects once every 90 min, and two lots in three get no feedforward.

(b) Which of the two measurements would you keep on every lot, and why? *Checkpoint:* The developed linewidth, because it feeds forward to the same lot with no queue.

6.12 Review the agent · 5 min

An AI assistant looks at scenario 4 and writes: “The newest-reading loop leaves $d/\lambda_E = 0.25\,\text{nm}$ against the drift at $\lambda_E = 0.30$, and it cannot go above 0.618 at $m = 2$. Switch The etch estimator uses to The matched lot, which is stable up to $\lambda_E = 2$ at any queue when the trim rate is known, and set $\lambda_E = 0.75$. The residual then falls to $d/\lambda_E = 0.10\,\text{nm}$.” Find the flaw, and say how you would check it in the app.