INTERACTIVE CONTROL ENGINEERING
Principles and Applications

Volume 14 Control in Semiconductor Manufacturing

CHAPTER 11

A Chemical Mechanical Planarization Simulator for Profile Control Across a Consumable Lifecycle

Pages 265–289

Edited by John D. Hedengren

How to cite this chapter. "A Chemical Mechanical Planarization Simulator for Profile Control Across a Consumable Lifecycle," in *Control in Semiconductor Manufacturing*, John D. Hedengren, Ed., vol. 14 of *Interactive Control Engineering: Principles and Applications*. Advanced Process Solutions, LLC, 2026, ch. 11, pp. 265–289.

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

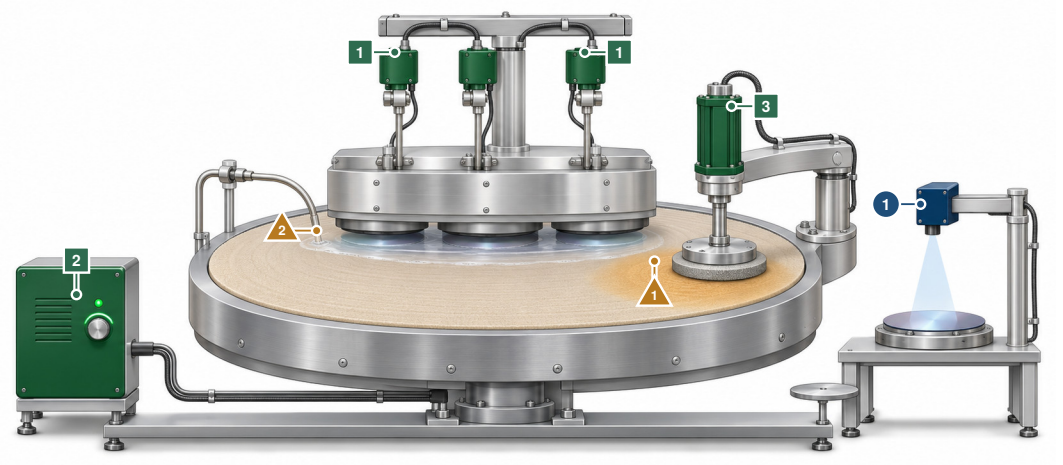
The app for this chapter
https://apmonitor.com/apps/cmp-polish/

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

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

CHAPTER 11

A Chemical Mechanical Planarization Simulator for Profile Control Across a Consumable Lifecycle



The controller moves 1 air pressures in the eight rings of the carrier head 2 the polish time of each wafer 3 how hard the diamond disc conditions the pad

The controller reads 1 a gauge that maps the polished film's thickness, one or two lots later

What disturbs it 1 the pad wearing out, and a new pad that behaves differently 2 a change in how active the slurry is

Notice that the polished film is measured only by the gauge at the right, and its map returns one or two lots after the wafer left the pad. Illustration, not to scale.

IN PLAIN TERMS **Polishing Flat on a Pad That Wears Out**

The main idea. A polisher rubs a thin film on a silicon disk down to an even thickness with a spinning pad and a gritty liquid. A head with rings of air presses harder where more film must come off.

The challenge. The pad wears with every disk and is replaced every few days by one that works a bit differently. The controller must decide what it still knows about the machine and what it must learn again.

Why it matters. Learning a new pad from one test disk, not a fixed count of six, saves silicon at every change.

THREE WAYS TO READ THIS CHAPTER

Explore	12 min	The first page, the fact boxes of the Introduction and of Section 11.6, the history boxes of Section 11.2, the try-it boxes of Section 11.6, and the key points
Study	50 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 11.6, Section 11.7, Section 11.8 with its open challenge

The control question Eight carrier head chambers drive thirty-three radial thicknesses while the pad doing the work wears out and is replaced every few days. Which estimator states should survive a pad change, and how many qualification wafers does the answer cost?

Learning objectives

- Explain why eight concentric chambers cannot reach every radial profile, and separate a profile error into its reachable, unreachable and azimuthal parts.
- Compute the relative pad speed across the wafer, the settled pad activity under a conditioning dose and the removal lost to structural wear.
- Distinguish what feedforward from an incoming map and run-to-run feedback on the removal can each correct, and choose the EWMA level and trend weights against a metrology queue.
- Compare soft, carry-all and cold resets at a pad change, and compute how much one mapped wafer says about the gain of a fresh pad.
- Judge a release on predicted probability against a fixed qualification count, and say when a surrogate measurement may be trusted.

Prerequisites Sections 1.4.3, 1.4.4, 1.4.5 and 1.4.7 of Chapter 1, and Section 1.6.4 for the scenarios; Chapter 6 for EWMA recipe control through a queue.

The app https://apmonitor.com/apps/cmp-polish/

See also Volume 8, *Batch, Run-to-Run, and Hybrid Control*: Volume 8, Chapter 9 (a run-to-run estimate that knows which command each late result answers) and Volume 8, Chapter 11 (a resin capacity that cleaning restores only in part). Volume 6, *Multivariable and Plantwide Control*: Volume 6, Chapter 6 (a regularized inverse of an ill-conditioned response). Volume 5, *Delays, Slow Measurements, and State Estimation*: Volume 5, Chapter 10 (a soft sensor whose bias a delayed laboratory corrects).

11.1 Introduction

Most control problems assume that the plant is still the same plant tomorrow. A polisher is not. The pad that removes the film is a consumable, it is conditioned back and worn down on every wafer, and after a few thousand wafers somebody takes it off the machine and puts a new one on. The tool that comes back is the same carrier head with the same eight pneumatic chambers, the same footprints, the same metrology and the same product, and it is a different process. What the control system learned about the old pad is partly still true and partly worthless, and the boundary between those two parts is where this application lives.

> **DID YOU KNOW? A Billion-Dollar Consumable**
>
> In 2005 the slurries, pads and pad conditioning materials that polishers use up were a market worth 1.1 billion US dollars, expected to grow to 1.9 billion by 2009 [1]. The pad of this chapter is one of those consumables, and it is replaced every few days.

The obvious answer is the one the industry mostly uses. Run a fixed number of sacrificial qualification wafers after every pad change, throw them away, and resume. That is safe, it is simple, and it costs scrapped silicon at every maintenance event on every polisher in the fab. The alternative is to say out loud which estimator states belong to the part that was replaced and which belong to the machine, reset only the first group, widen the covariance on what was reset, and then release production on evidence rather than on a count. The three obvious ways to handle a pad change differ by more than an order of magnitude in what they cost, and Section 11.6 measures all three on identical hidden pads.

The same shape appears outside semiconductor manufacturing. A catalyst is replaced, a membrane module is changed, an analyzer is recalibrated, a mill liner is renewed, and in every case a control system that has been adapting for months is asked what it still believes. The literature on run-to-run control is largely written about drift, where the estimator tracks a slowly moving target [2, 3], as

Section 1.4.3 sets out and Chapter 6 measures on a patterning line. A maintenance event is not drift. It is a known instant with an unknown outcome, and the correct response to it is a reset of some states and an increase of uncertainty on others rather than a change of gain.

Underneath that sits a second problem that a student meets first. Eight chambers drive thirty-three radial calculation points. Nobody gets to choose the thickness at each radius independently, the response to one chamber covers a broad band of radii, and a profile error that will not come out is often an unreachable spatial mode rather than evidence of poor tuning [4]. Equal chamber pressure does not even give a flat wafer, because relative pad speed rises toward the wafer edge and Preston behavior is linear in velocity [5, 6].

This chapter makes the cost of a maintenance event measurable on an interactive polisher. On it a retained population prior and one mapped monitor wafer release production where a fixed policy needs six, and the same single wafer applied through the wrong reset strategy puts the first production lot 41.7 nm off target. The simulator is an eight-state consumable model driving a thirty-three by eight spatial plant, with a constrained recipe optimizer, an age-aware residual estimator and a Bayesian fresh-pad update written closely enough that a reader can rebuild them. The analysis decomposes a profile error into its reachable and unreachable parts, and it shows that the information in a qualification wafer comes from the retained prior and the thirty-three sites rather than from the excitation design. On one plant, one seed and one set of hidden pad parameters, three reset strategies that are all defensible in isolation produce 2.5 nm, 41.7 nm and 59.4 nm of first-lot error. Preston's relation, the exponentially weighted moving average and Bayesian parameter estimation are all standard methods.

Section 11.2 places the problem in its domain. Section 11.3 gives the simulation model: the influence matrix of the carrier head, the removal law, the two pad states and the metrology. Section 11.4 describes the controllers as implemented, from the recipe solve to the qualification supervisor. Section 11.5 analyzes what a pad change may delete, and Section 11.6 reports the measured outcomes of the scenarios. Section 11.7 describes the hardware such a polisher runs on, Section 11.8 states the limitations and Section 11.9 concludes.

11.2 From Pad Wear to Consumable-Aware Recipe Control

Polishing as a controlled process is older than the semiconductor industry that adopted it. Its removal law came from plate glass and remains the first-order description of what a polisher does [5], with exponents near but not exactly unity [6] and the chemistry added by later models [7], as the box below tells.

> ⧖ HISTORY **From Plate Glass to Wafers (1927 to 2007)**
>
> In 1927 a study of the theory and design of plate glass polishing machines related the rate at which glass is removed to the pressure applied and to the relative velocity between the pad and the glass. Most models of chemical mechanical planarization are still inspired by that equation, although it is empirical and the effect of the abrasive's hardness and size on its coefficient is unclear [1]. A modified model of 1997 predicted that removal grows with pressure to the power five sixths and with speed to the power one half, and experiments on thermal oxide agreed [6]. By 2007 a review called the process the only technique able to give excellent local as well as global planarization [1].

Feedback on the recipe arrived in the middle 1990s, and it arrived because of the pad. Run-by-run compensation for pad wear was demonstrated on a five-hundred wafer experiment in 1996 [8], on statistical machinery developed just before [2] and analyzed just after [3], and within five years the estimator had been given the pad's age [9]. The practice of that decade is collected in a single volume [10], and the box below tells the story.

> ⧖ HISTORY **Run-by-Run Control Comes to the Polisher (1995 to 2001)**
>
> In 1995 a run-by-run controller combined statistical process control with feedback. Its rapid mode adapted to sudden shifts such as those caused by maintenance, and its gradual mode to slow drifts [2]. In 1996 a prototype system controlled a single-wafer polisher lot by lot. It measured film thickness at nine points on blanket oxide wafers of 8 inches, adapted its model with an exponentially weighted moving average, and generated recipes within bounds on the inputs and the steps of the machine settings. It compensated a substantial drift in the tool's removal rate, and its model held on more than one tool for several months [8]. By 2001 an age-based double EWMA was applied to polishing, because an EWMA alone does not suffice for a process that wears out [9].

Profile control is younger. A carrier head with several independently pressurized concentric chambers turns one manipulated variable into a small vector, and the plant it drives is tall because the film is measured at many radii. Multizone polishing has been treated as exactly that, many radial outputs driven by fewer pressures through local Preston coefficients, with the inverse taken through a pseudoinverse or a singular value decomposition [4]. The mechanical question of how chamber pressure becomes contact pressure at the film has its own literature [11, 12], and the pad surface that the pressure acts through is itself a controlled variable maintained by diamond disc conditioning [13].

The newest layer is metrology that is not metrology. Spatially resolved removal and non-uniformity can be predicted from equipment traces and consumable counters on production data sets [14], and virtual metrology has been used to close run-to-run loops on wafers that were never measured [15]. The accompanying risk has also been named. A surrogate that feeds its own corrections is a closed loop with no ground truth in it, and the practical answer has been to let the confidence of the prediction decide the physical sampling rate rather than to let the prediction replace the sample [16].

What none of that literature makes easy to see is the moment of replacement. The estimator papers treat the pad as an aging parameter, the profile papers treat the head as a fixed matrix, and the virtual metrology papers treat the tool as stationary enough to train on. The application of this chapter puts all three on one machine and then changes the pad.

11.3 Process Model

A 300 mm wafer is held face down by carrier head CH-101 against a rotating polyurethane pad on platen PL-101. Slurry is delivered to the pad surface, diamond disc conditioner CD-101 sweeps the pad to restore its texture, and the carrier head presses the wafer through eight concentric membrane chambers at pressures p_1 through p_8. The wafer is polished for a common time T_p, unloaded, and replaced by the next one. The equipment and every symbol that appears in the equations are drawn in Figure 11.1.

Table 11.1 collects what the controller acts on and what it acts against. The product factor η_{pr} is the one disturbance the controller knows, because the product identity is on the lot record. Everything else in the lower half of the table is unmeasured, and the only evidence that any of it moved is a thickness map arriving one or two lots later. Time is carried in minutes, thickness in nanometers, pressure in pounds per square inch and polish time in seconds. Eight states are integrated continuously and the per wafer arithmetic is evaluated once every $T_{takt} = (T_p + t_{ovh})/60$ minutes, which is 1.5 min at the design recipe.

11.3.1 Geometry and the Influence Matrix

The wafer is represented by $M = 33$ radial calculation points on a midpoint grid, $\rho_m = (m - 0.5)/M$ for $m = 1 \ldots M$ with $r_m = \rho_m R$ and $R = 150\,\text{mm}$. Area weights $w_m \propto \rho_m$ are normalized to sum to one, so every area-weighted mean in this chapter is an honest wafer average rather than a radial average.

Eight chambers occupy normalized radial bands with edges at 0, 0.300, 0.500, 0.645, 0.765, 0.860, 0.925, 0.968 and 1. Interior chambers are wide and edge chambers are narrow, which is how

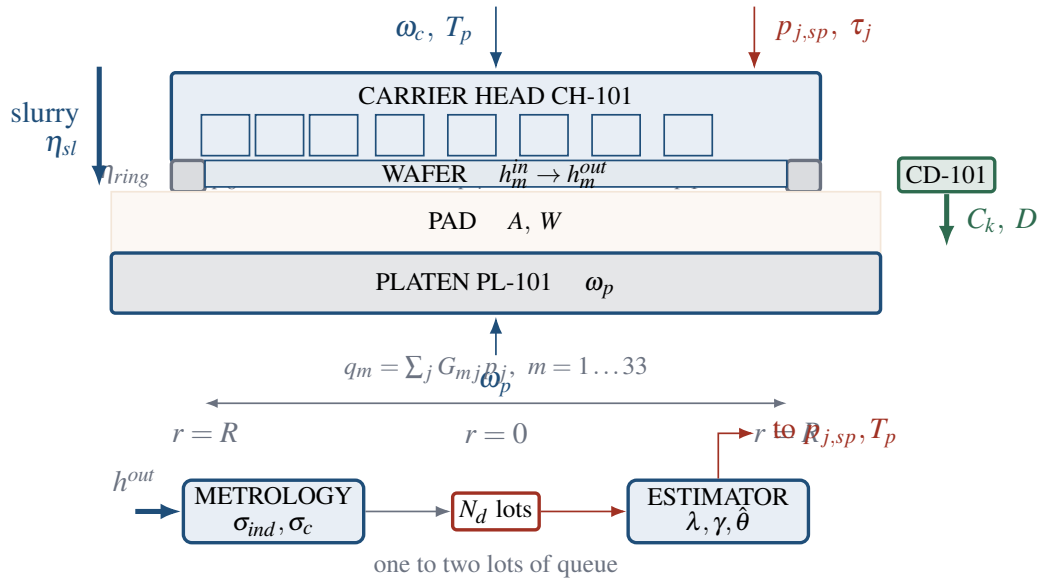


Figure 11.1 The polisher and its measurement path. Eight concentric carrier head chambers at pressures p_1 through p_8 press the wafer against the pad through an influence matrix G that spreads each chamber over a band of radii. The pad carries two states that are not the same thing, an active surface A that conditioning restores and a structural wear W that nothing restores. The post-polish thickness map is sampled and returns one to two lots late.

Table 11.1 Controlled, manipulated and disturbance variables

Symbol	Role	Design value	Range
$\bar{h}^{out}$	CV, N_d lots late	700.00 nm	400–1100 nm
h_1^{out}	CV, wafer center	703.59 nm	—
h_{33}^{out}	CV, wafer edge	696.64 nm	—
T_p	MV, once per wafer	60.0 s	35–90 s
$p_1 \ldots p_8$	MV, once per wafer	3.60 psi	2.5–4.8 psi
C_k	MV, once per wafer	0.55	0–1.5
A	state, unmeasured	0.9083	0–1
W	state, unmeasured	0 at install	0–1
D	state, unmeasured	0 at install	0–1
E	state, counter	0 at install	wafer-equivalents
η_{sl}	state, DV, unmeasured	1.000	0.80–1.20
η_{pr}	state, DV, known	1.000	0.75–1.25
η_{ring}	state, DV, unmeasured	1.000	0.70–1.30
β_M	state, DV, unmeasured	0 nm	±15 nm
θ	hidden, per pad	**0** (mean)	$\sigma_{\theta g}$, $\sigma_{\theta s}$
h^{in}	DV, measured pre-polish	1200 nm	±80 nm

multizone heads are built. Each column of the influence matrix is a Gaussian footprint centered on its chamber,

$$\tilde{G}_{mj} = \exp\left[-\frac{(\rho_m - c_j)^2}{2\sigma_j^2}\right], \qquad \sigma_j = 0.9\,\frac{b_{j+1} - b_j}{2} + 0.045 \tag{11.1}$$

where c_j is the midpoint of band j and b_j are the band edges. The matrix is then row-scaled,

$$G_{mj} = \frac{\tilde{G}_{mj}}{\sum_{\ell} \tilde{G}_{m\ell}}, \qquad q_m = \sum_{j=1}^{8} G_{mj}\, p_j \tag{11.2}$$

so that eight equal pressures reproduce a uniform contact field exactly. That scaling is a calibration of the whole matrix and not a per-column normalization, and the columns keep their different shapes and their different totals. A 0.20 psi step on interior chamber 4 raises the contact pressure by 0.095 psi at its own center, stays above half of that across 21 % of the wafer radius (the span of the calculation points above half height), and is still 38 % and 56 % of its peak at the centers of chambers 3 and 5. The same step on edge chamber 8 peaks at 0.081 psi and stays above half of it across 3 % of the radius. The edge chamber has localized authority because it is narrow rather than because it is strong. The bake plate of Section 1.4.5 has the same structure in heat, with a conductance matrix in place of G.

Pause. Try Exercise 11.3 before reading on.

11.3.2 Kinematics and Removal

The platen turns at ω_p and the carrier at ω_c about centers a distance d apart, so the azimuthally averaged relative speed at radius r_m is Equation (11.3).

$$v_m = \sqrt{(\omega_p d)^2 + (\omega_p - \omega_c)^2\, r_m^2} \tag{11.3}$$

At 80 rpm, 60 rpm and $d = 0.220$ m this is 1.8431 m/s at the innermost calculation point and 1.8689 m/s at the outermost, a ratio of 1.0140. The reference speed v_{ref} is the area-weighted mean of v_m at the nominal recipe.

Local removal follows a spatial Preston relation with multiplicative state corrections, Equation (11.4).

$$RR_m = RR_{ref}\, \eta_{pad,m}\, \eta_{sl}\, \eta_{pr}\, \eta_{ring,m} \left(\frac{v_m}{v_{ref}}\right)^{\beta} \left(\frac{q_m}{p_{ref}}\right)^{\alpha} \tag{11.4}$$

Integrated removal over a polish of T_p seconds is $r_m = RR_m T_p/60$ in nanometers, and the outgoing thickness is $h_m^{out} = h_m^{in} - r_m + \varepsilon_{proc}$ where ε_{proc} is a wafer-to-wafer process deviate of standard deviation σ_{proc} common to all radii. The retaining ring acts only at the outer edge, $\eta_{ring,m} = 1 + (\eta_{ring} - 1)\psi_e(\rho_m)$ with $\psi_e(\rho) = \exp[-((1-\rho)/0.08)^2]$.

The pressure ramp is paid for rather than discarded. Each chamber is a first order actuator of time constant τ_j that falls linearly from $\tau_{p,max}$ at chamber 1 to $\tau_{p,min}$ at chamber 8, and the time-averaged pressure over one polish has the closed form of Equation (11.5).

$$\bar{p}_j = p_{j,sp} - \frac{\tau_j}{T_p}\,(p_{j,sp} - p_{j,prev})\left(1 - e^{-T_p/\tau_j}\right) \tag{11.5}$$

Equation (11.5) is exact for the integrated removal when the pressure exponent α is unity, and is a first order approximation away from it. The contact field of Equation (11.2) is evaluated at $\bar{p}_j$.

11.3.3 The Pad Is Two States

The pad carries an active surface A that conditioning restores and a structural wear W that nothing restores. Conditioning is delivered in situ at dose C_k per wafer and mechanical exposure accumulates with pressure, speed and time, so both are written as rates per minute using ΔE_k from Equation (11.6).

$$\Delta E_k = \frac{T_p}{T_{ref}} \frac{\bar{v}_k}{v_{ref}} \frac{\bar{q}_k}{p_{ref}}, \qquad \dot{E} = \frac{\Delta E_k}{T_{takt}} \tag{11.6}$$

The three consumable states then obey Equations (11.7) through (11.9), in which $\dot{C} = C_k / T_{takt}$ is the conditioning dose rate.

$$\dot{A} = k_c \, \eta_{cond} \, \dot{C} \, (1 - A) - k_g \, \dot{E} \, A \tag{11.7}$$

$$\dot{W} = \frac{\dot{E} + \rho_c \dot{C}}{N_{EOL}}, \qquad N_{EOL} = N_{pad} \, (1 + \rho_c C_{nom}) \tag{11.8}$$

$$\dot{D} = \frac{\dot{C}}{N_{cond}}, \qquad \eta_{cond} = \max[\eta_{c,min}, \, 1 - (1 - \eta_{c,min}) D] \tag{11.9}$$

Equation (11.7) settles at $A^* = k_c C_{nom} / (k_c C_{nom} + k_g) = 0.9083$ under the default dose, which is the value the reference removal rate is calibrated at. The capacity N_{EOL} in Equation (11.8) is scaled so that the quoted pad life is the life obtained under the default conditioning policy rather than a polishing-only count, which matters because conditioning consumes the pad. The affinity resin of Volume 8, Chapter 11 has the same split, a capacity that cleaning restores only in part.

The pad activity multiplier assembles the two states with the hidden pad-lot correction, Equation (11.10).

$$\begin{aligned} \eta_{pad,m} = e^{\theta_g} \, f_A(A) \left(1 - d_w W^{1.4}\right) \\ \times \left[1 + \theta_t \phi_{t,m} + \theta_c \phi_{c,m} + \theta_e \phi_{e,m}\right] \end{aligned} \tag{11.10}$$

with $f_A(A) = (0.80 + 0.25A) / (0.80 + 0.25A^*)$, which is unity at the running operating point rather than at an idealized $A = 1$. The exponent 1.4 on structural wear is why the decline accelerates. On a 260 wafer pad a fixed recipe holds the mean inside the ± 14 nm band until wafer 70, the first to leave it, and then finishes 90 nm above target at end of life.

11.3.4 Profile Modes and the Hidden Pad Parameters

Three shape functions carry the pad-to-pad variation. Each is centered by subtracting its own area-weighted mean on this grid, which is below 0.001 for the two polynomials and 0.167 for the edge bump, so all three have exactly zero area-weighted mean. They move the profile without moving the mean, which is what allows the mean and the shape to be released against separate chance constraints later.

$$\begin{gathered} \phi_t = 2\rho^2 - 1 - \bar{\phi}_t, \qquad \phi_c = 6\rho^4 - 6\rho^2 + 1 - \bar{\phi}_c, \\ \phi_e = e^{-\left(\frac{1-\rho}{0.10}\right)^2} - \bar{\phi}_e \end{gathered} \tag{11.11}$$

Together with the constant they are orthonormalized under the area weights at load time, giving a basis $\{\mathbf{b}_0 \ldots \mathbf{b}_3\}$ on which a measured residual is projected by a weighted dot product rather than by a least squares solve. At every pad replacement the hidden vector $\theta = [\theta_g, \theta_t, \theta_c, \theta_e]^T$ is drawn once from the population, with standard deviations $\sigma_{\theta g}$ for the gain and $\sigma_{\theta s}$, $0.8\sigma_{\theta s}$ and $1.6\sigma_{\theta s}$ for tilt, curvature and edge, and held fixed for that pad.

11.3.5 Quality, Metrology and Disturbances

Two non-uniformity conventions are reported because the literature quotes both, and both are computed on the outgoing thickness rather than on the removal amount.

$$NU_\sigma = 100 \, \frac{\sigma_h}{\bar{h}^{out}}, \qquad WIWNU = 100 \, \frac{h^{out}_{max} - h^{out}_{min}}{2 \, \bar{h}^{out}} \tag{11.12}$$

Blanket-monitor practice quotes removal-rate non-uniformity, and for an unpatterned wafer with a flat incoming film the two agree. They do not agree here, because correcting an incoming profile requires deliberately non-uniform removal and a removal-based figure would penalize the entire purpose of the eight chambers.

A physical map carries an independent per-site deviate of standard deviation σ_{ind} and a correlated whole-map deviate of standard deviation σ_c, so its covariance is $\mathbf{R} = \sigma_{ind}^2 \mathbf{I} + \sigma_c^2 \mathbf{1}\mathbf{1}^T$ plus the process

Table 11.2 Model parameters. All are adjustable in the model dialog unless marked hidden, which means the interface does not expose them, or fixed, which marks the geometry.

Symbol	Meaning	Value	Unit
R	Wafer radius (fixed)	150	mm
M	Radial calculation points (fixed)	33	—
N_z	Carrier head chambers (fixed)	8	—
N_{lot}	Wafers per lot (hidden)	25	wafer
t_{ovh}	Load, unload and index	30	s
N_{pad}	Nominal pad life	1500	wafer
RR_{ref}	Reference mean removal rate	500	nm/min
p_{ref}	Reference chamber pressure	3.60	psi
α	Pressure exponent	1.00	—
β	Velocity exponent	1.00	—
h_0^{in}	Nominal incoming thickness	1200	nm
h^{target}	Target final thickness	700	nm
h_{min}	Minimum remaining film	640	nm
$\tau_{p,min}$	Fastest chamber time constant	0.5	s
$\tau_{p,max}$	Slowest chamber time constant	3.0	s
ω_p	Platen speed	80	rpm
ω_c	Carrier speed	60	rpm
d	Carrier offset	0.220	m
k_c	Conditioning recovery	0.18	/dose
k_g	Glazing coefficient	0.010	/wafer-eq
d_w	End-of-life removal loss	0.18	—
ρ_c	Pad consumed per dose	0.25	wafer-eq/dose
N_{cond}	Conditioner disc life	2500	dose
$\eta_{c,min}$	Conditioner efficiency floor	0.50	—
C_{nom}	Dose the pad life is quoted at	0.55	—
t_{pre}	Required break-in conditioning	8	min
$\sigma_{\theta g}$	Fresh-pad gain spread	0.05	1σ
$\sigma_{\theta s}$	Fresh-pad shape spread	0.025	1σ
σ_{ind}	Independent site noise	2.0	nm
σ_c	Correlated map uncertainty	3.0	nm
σ_{VM}	Virtual metrology uncertainty	5.0	nm
σ_{proc}	Wafer-to-wafer process noise	1.2	nm
σ_{in}	Incoming variation	6.0	nm
τ_{dist}	Disturbance relaxation (hidden)	1.0	min
e_h^{max}	Mean thickness tolerance	14	nm
$RMSE^{max}$	Profile error limit	12	nm
$WIWNU^{max}$	Range non-uniformity limit	1.8	%

term. Maps are taken on the first wafer of each lot and on every $\lfloor N_{lot}/n_s \rfloor$-th wafer after it, up to n_s maps per lot, so two maps per lot are wafers 1 and 13. Maps arrive N_d lots later, except during qualification and the first released lot, when measurement is immediate. Section 1.4.4 gives the cost of such a queue in general. Virtual metrology predicts the same map from the controller model calibrated at the last physical anchor, with its own uncertainty σ_{VM} and a switchable bias.

Four disturbances relax to their slider values on a time constant τ_{dist}, which keeps a step continuous enough for a fixed-step integrator. The incoming profile is $h_m^{in} = h_0^{in} + \delta_0 + a_t\phi_t + a_c\phi_c + \varepsilon_{in}$ with a_t scaled to 40 nm at full slider travel and a_c to 30 nm. The azimuthal disturbance is separate. It has zero radial mean by construction, contributes its own root-mean-square of $a_{az}/\sqrt{2}$ in quadrature with whatever the radial loop achieves, and no concentric chamber can touch it.

11.3.6 Initialization and Numerics

The eight states initialize to $A = A^* = 0.9083$, $W = 0$, $D = 0$, $E = 0$, $\eta_{sl} = \eta_{pr} = \eta_{ring} = 1$ and $\beta_M = 0$, which is a freshly qualified pad. A scenario may pin any of them, and the pad replacement action sets W, E to zero and A to 0.15, because a new pad is smooth rather than conditioned. Integration is fixed-step fourth order Runge-Kutta at $\Delta t = 0.05$ min with the controller on the same scan and the historian logging twice per minute, so every trend and every exported record reproduces exactly on any machine (Section 1.6.1). Matrix work is done with Cholesky factorizations and solves, including the four by four posterior covariance, which is formed column by column from the factor. The eight by eight recipe solve and the four by four estimator update both run once per wafer, which is once every 1.5 min of simulated time at the design recipe.

11.3.7 Assumptions

The head is a fitted influence matrix with Gaussian footprints and not a finite element model of membrane, wafer, pad and retaining ring. A real identified matrix will not be as smooth as this one, and its columns will not be symmetric about their chamber centers.

Preston behavior is the nominal relationship and the exponents default to unity. The pressure dose of Equation (11.5) is exact only at unit pressure exponent, so away from $\alpha = 1$ the integrated removal carries a first-order approximation of the ramp.

Pad activity, structural wear and conditioner efficiency are low-order latent states chosen to reproduce break-in, glazing, slow decline and a pad-to-pad reset. Their coefficients are not fitted to any named pad, slurry, conditioner or tool, and the quoted pad life is a configurable teaching value and not a prediction.

The consumable states integrate continuously at the per wafer average rate rather than stepping once per wafer. For in situ conditioning that is the physics. For ex situ conditioning between wafers it smooths a sawtooth the application never shows.

The wafer is thirty-three radial points. Azimuthal and die-scale variation enter only as an uncorrectable contribution to the graded profile error, so dishing, erosion, defects, scratches and electrical yield are all outside the model.

The fresh-pad prior assumes a repeated pad family with stable metadata. It should be widened or abandoned after a material, supplier, conditioner, head or recipe change, and the application offers no mechanism for detecting that such a change has happened.

11.4 Controllers as Implemented

Four layers run on four clocks. Chamber pressure settles in seconds inside one wafer, the recipe is written once per wafer, the run-to-run residual updates when a measurement arrives one or two lots later, and the fresh-pad estimator updates once per pad. The recipe authority control selects which of the middle two layers is active, and the header MANUAL and AUTO switch selects whether any of them is.

11.4.1 What the Controller Believes

The controller reads the pad counters and the product identity and nothing else. Its model state is A and W from the counters, η_{pr} from the lot record, unity for η_{sl} and η_{ring}, and $\hat{\theta}$ from the estimator. The gap between that and the true state of Table 11.1 is exactly what the run-to-run residual has to absorb, and naming it explicitly is what makes the residual interpretable.

11.4.2 The Constrained Recipe

With $\alpha = 1$ the predicted removal is linear in the pressure vector, so the recipe solve is a small regularized least squares problem. Writing a_m for the removal per unit contact pressure per sixty seconds at the controller's model state and $\hat{\mathbf{e}}$ for the age residual of Section 11.4.4, the target is

$\mathbf{t} = \mathbf{h}^{in,est} - h^{target}\mathbf{1} - \hat{\mathbf{e}}$ and the solve is Equation (11.13).

$$\min_{\mathbf{p}} \quad \sum_m w_m \left(a_m \frac{T_p}{60} \sum_j G_{mj} p_j - t_m \right)^2 + \lambda_p \|\mathbf{p} - p_{ref}\mathbf{1}\|^2 + \lambda_\Delta \|\mathbf{p} - \mathbf{p}_{prev}\|^2 + \lambda_s \|\mathbf{D}_2 \mathbf{p}\|^2 \tag{11.13}$$

where $\mathbf{D}_2$ is the second difference operator on the eight chamber pressures and $\lambda_p = 0.05\lambda_s$, $\lambda_\Delta = 0.10\lambda_s$. The smoothness weight is what prevents an ill-conditioned inverse from answering a high-frequency radial target with a sawtooth, the problem that Volume 6, Chapter 6 meets at the scale of a ninety-six by ninety-six response. Its scale matters more than it looks. The diagonal of the data term of the normal-equation matrix runs from about 360 for the edge chamber to about 1280 for chamber 2 at the default removal rate and a 60 s polish, so a second difference weight in the hundreds does not smooth the recipe, it replaces it, and the optimizer then returns eight equal pressures for every target it is given. The shipped default is $\lambda_s = 8$.

Polish time carries the level and the chambers carry the shape, so the solve alternates. Four iterations of solving Equation (11.13) at the current T_p and then rescaling T_p by the ratio of desired to predicted mean removal, followed by a final pressure solve, put the predicted mean within 0.1 nm of target on 95 % of the production wafers of scenarios 3, 6 and 11. The pressure box, the neighboring-chamber gradient limit $\Delta p_{ad\,j}$ and the run-to-run move limit Δp_{move} are enforced by three sweeps of projection.

The minimum film is a chance constraint and not a nominal one. With z_α the normal quantile at the release probability and $\sigma_{h,m}$ the predicted output standard deviation of Section 11.4.5, the polish time is reduced until Equation (11.14) holds at every radius.

$$h_m^{in,est} - \hat{r}_m - z_\alpha \sigma_{h,m} \geq h_{min} \tag{11.14}$$

The film constraint is hard, so it overrides the target.

11.4.3 Feedforward

What the optimizer is allowed to know about the wafer in front of it is a control with three settings, a case of the feedforward between process steps of Section 1.4.7. The product nominal gives $h_m^{in,est} = h_0^{in}$. The lot-representative mean gives $h_m^{in,est} = h_0^{in} + \delta_0$, one number for the whole lot. The inline map gives the measured h_m^{in} at all thirty-three points. The difference between the last two is the entire value of inline pre-polish metrology, and it appears in the first wafers after a product change, as fast as the run-to-run move limit lets the recipe step, and nowhere else.

11.4.4 The Age-Aware Residual

When a measured removal map for origin wafer i arrives, the controller reconstructs the model as it stood when that wafer was polished, using the stored pad states, product factor, estimate vector, applied pressure dose and polish time. The residual $\mathbf{e}_i = \mathbf{r}_i^{meas} - \hat{\mathbf{r}}_i$ is projected onto the four orthonormal profile modes and each coefficient is updated by Equation (11.15), the EWMA of Equation (1.2) applied to each mode with a slope term in exposure added.

$$\ell_k^+ = \lambda\, c_{k,i} + (1-\lambda)\ell_k^-, \qquad s_k^+ = \gamma \frac{\ell_k^+ - \ell_k^-}{E_i - E_{prev}} + (1-\gamma) s_k^- \tag{11.15}$$

and replayed forward to the present exposure by $\hat{e}_k = \ell_k^+ + s_k^+ (E - E_i)$. Two implementation details are not optional. The slope update is skipped when the exposure interval is below one wafer-equivalent, because a rate needs a real interval to be a rate. Without that guard, and with the exposure recorded before rather than after the wafer, the first measurement of a run arrives with an exposure interval of zero and the slope goes straight to its bound. Both the level and the slope are bounded, the level at half

the reference removal and the slope at 2 nm per wafer-equivalent, which is 24 times the steepest wear slope of a default 1500 wafer pad (0.084 nm per wafer-equivalent at end of life) and still finite. On the 260 to 400 wafer pads of the scenarios the steepest wear slope is 0.32 to 0.48 nm per wafer-equivalent, four to six times below the bound.

A delayed result is evidence about the model that produced it and not about today's model, the comparison with the model of its time that Section 1.4.4 asks for. Replaying it against the controller's current product factor, pad state and estimate rather than the ones in force at the originating wafer charges whatever changed during the queue to the measurement. In the challenge scenario with the suggested settings, the first map delivered after the 10 % product change at 150 min then reports a removal residual 20 nm smaller than the true one and the second 28 nm smaller, and the loop under-corrects the earlier slurry step by that amount.

An arriving map is gated. The normalized innovation is computed against the full innovation covariance, which is the measurement covariance plus the contribution of the parameter posterior, and not against the measurement noise alone. On the qualification wafers of scenarios 7, 8 and 11 a fresh pad whose gain the controller does not yet know produces an innovation of 15 to 34 nm root-mean-square against a measurement standard deviation of 3.8 nm. Scaled on measurement noise only, that is 3.9 to 8.9 standard deviations and the gate shrinks the one wafer that was supposed to describe the pad, while scaled on the full covariance it is 0.4 to 1.2 and the wafer is used at full weight. Above three standard deviations the map is shrunk toward the prediction in Huber fashion, above nine it is rejected, and five consecutive rejections place the tool on hold, because a gate that rejects every map is an open loop wearing a safety badge.

11.4.5 The Fresh-Pad Posterior

At a qualification wafer the four fresh-pad parameters are updated in information form. With $\mathbf{H}$ the sensitivity of predicted removal to θ, whose columns are the base removal multiplied by the four basis functions, and $\mathbf{z}$ the residual against the base model with $\theta = \mathbf{0}$, the update is Equation (11.16).

$$\begin{aligned} \left(\mathbf{P}^{+}\right)^{-1} &= \left(\mathbf{P}^{-}\right)^{-1} + \mathbf{H}^{T}\mathbf{R}^{-1}\mathbf{H} \\ \mu^{+} &= \mathbf{P}^{+}\left[\left(\mathbf{P}^{-}\right)^{-1}\mu^{-} + \mathbf{H}^{T}\mathbf{R}^{-1}\mathbf{z}\right] \end{aligned} \tag{11.16}$$

The rank-one structure of $\mathbf{R}$ inverts analytically,

$$\mathbf{R}^{-1} = \frac{1}{s^{2}}\left(\mathbf{I} - \frac{\sigma_{c}^{2}}{s^{2} + M\sigma_{c}^{2}}\mathbf{1}\mathbf{1}^{T}\right), \qquad s^{2} = \sigma_{ind}^{2} + \sigma_{proc}^{2} \tag{11.17}$$

which keeps the whole update inside four by four arithmetic and guarantees that the posterior never widens on a valid measurement. That rank-one term is also the reason thirty-three radial sites on one wafer are not thirty-three wafers. It absorbs most of the information about global gain while leaving the three shape modes sharply determined.

The predicted output standard deviation used by Equation (11.14) and by the release gate separates the mean from the shape, because mode zero is a global gain and modes one to three have exactly zero area-weighted mean. Lumping them into one number makes the profile error marginal carry the gain uncertainty a second time and puts the release gate out of reach of any qualification wafer.

11.4.6 The Qualification Supervisor

The supervisor is a state machine over PAD INSTALLED, PRECONDITIONING, QUALIFICATION, RELEASED WITH ENHANCED SAMPLING, PRODUCTION, END OF LIFE and HOLD. A pad replacement holds production, zeroes W and E, sets A to 0.15, draws a new hidden θ, discards every queued result from the previous pad, and resets the production baseline recipe to the qualified nominal. What it does to the estimator is the subject of Section 11.5.

Preconditioning runs conditioning for t_{pre} minutes with no wafer consumed, which raises A from 0.15 to about 0.90 and charges the clock and the pad for it. A model-based controller can reduce the

number of qualification wafers. It cannot replace the physical work that gives a new pad a usable surface, and the application keeps the two separate on purpose.

Each qualification wafer is deliberately under-polished, at 86 % of the nominal removal demand computed against the controller's own model, and carries a balanced excitation of the three shape modes (tilt, curvature and edge) at 0.22 psi amplitude. The excitation pattern is chosen from a small safe candidate set, the nominal recipe and four sign patterns, by maximizing the log determinant of the posterior information matrix. The under-polish is not a detail. A fresh pad can be more active than the model believes (one population standard deviation of gain is 5 % of removal), so a monitor wafer run at the nominal recipe can land below target and one carrying a profile excitation on top of that can cross the minimum film limit while still being called safe. The qualification wafers of scenarios 7 and 8 keep 116 and 66 nm of film above h_{min}.

Release requires three things at once. No chamber may be faulted or saturated, the recipe must be within its limits, and the predicted joint probability of meeting the mean, profile error and non-uniformity limits must reach $P_{release}$. The joint probability is the product of three normal marginals. They do not fail independently, because a gain error moves all three together, so the number is optimistic in exactly the case where it matters most and it supplements process qualification rather than replacing it. When the qualification budget is spent without release the tool holds for engineering review, which is the correct outcome and not a failure of the policy.

Production recipes are limited to Δp_{move} per run, measured from the last production recipe rather than from the last command, because a qualification wafer is a deliberate excitation and not a recipe anybody intends to run. A pad replacement resets that baseline to the flat qualified recipe. Qualification and the first released lot carry a larger post-maintenance allowance of 0.7 psi, because the evidence for that step is a mapped wafer taken deliberately under a hold with the whole following lot measured. In scenario 7 the allowance raises the predicted pass probability at the release decision from 0.9985 to 0.9999.

11.5 What a Maintenance Event May Delete

Three strategies are offered and all three are defensible in isolation.

A *soft* reset zeroes the age residual level and slope, sets μ to the population mean, restores $\mathbf{P}$ to its prior, a standard deviation of 0.05 on each of the four parameters (the population spread of the gain, and wider than that of the three shape parameters), and keeps the influence matrix, the measurement model and the product models. It is the only one of the three under which a mapped monitor wafer has anything to update.

A *carry-all* reset keeps the worn pad's converged estimate and narrows the covariance, on the reasoning that the tool is the same tool and the estimate was hard won. The estimator is then confident about a pad that is in the bin. The release gate is computed from the controller's own belief, so it passes immediately, and the error is the difference between the two pads.

🕮 **Pause.** Try Exercise 11.2 before reading on.

A *cold* reset discards the fresh-pad model along with the pad. Nothing is left for an informative wafer to update, the posterior never contracts, and the qualification policy reverts to counting wafers. The only mechanism still running is the delayed age residual, which needs several lots to converge.

The distinction that makes this tractable is between states that belong to the part that was replaced and states that belong to the machine. The influence matrix, the neighboring-chamber coupling, the measurement covariance and the population of fresh pads survive a pad change. The wear state, the accumulated residual trend and the pad-specific correction do not. Writing that as a rule gives three instructions in order: reset what belongs to the part that was replaced, keep what belongs to the machine, and raise the uncertainty rather than the gain. Figure 11.2 sorts the estimator's states by owner and shows what each strategy does with them.

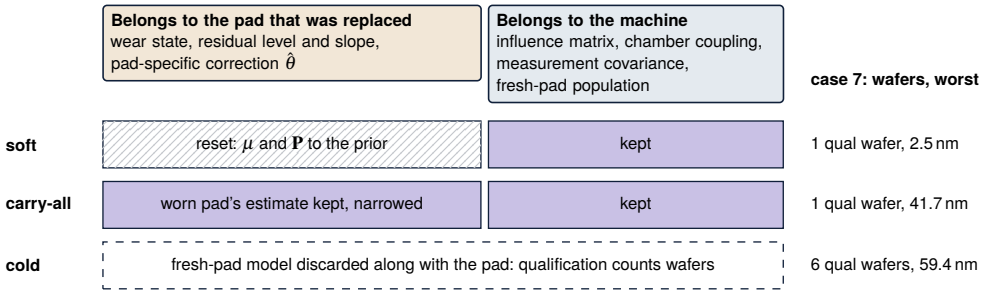


Figure 11.2 What a pad change may delete. The states of the estimator sorted by the part they belong to, and what the soft, carry-all and cold resets do with each, beside the outcome of case 7 in Table 11.3. Tinted cells are kept, hatched cells are reset, and the dashed cell is discarded. Notice that only the soft reset both clears what belongs to the pad and keeps the population that one mapped wafer can update.

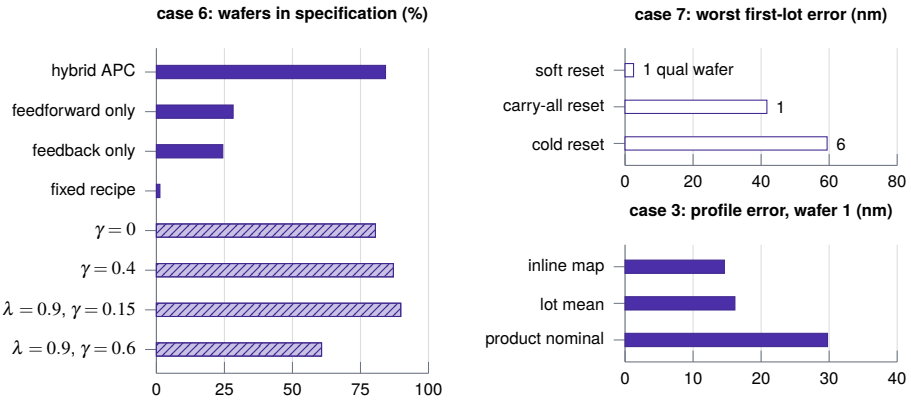


Figure 11.3 Rows of Table 11.3 drawn as bars. Left, the wafers in specification across the pad life of case 6 for the four architectures (solid bars) and for other trend and level weights of the hybrid loop (hatched bars). Right, the worst of the first twelve production wafers after the pad change of case 7, labeled with the qualification wafers spent, and the profile error of the first wafer after the product change of case 3. Notice that the two halves of the hybrid loop hold far fewer wafers alone than together, and that the cold reset spends the most wafers and lands farthest off.

The third instruction is the one that is usually skipped. A controller that resets the correct states and leaves the covariance alone is as confident after the change as before it, which is not a statement anybody can defend.

11.6 Results

Every number in this section is read from the running application at its default seed, on the shipped settings of the scenario named beside it unless a change is stated, with each change made before the run starts. A change made before the run and the same change carried through Reset give one run. Wafer counts and yields are the application's own tallies, and an event time is the instant of the controller execution at which it happens. The eleven scenarios run from manual chamber mapping to a graded lifecycle with two pad replacements, and Section 1.6.4 describes how a scenario is briefed, credited and reset. Table 11.3 collects the measured outcomes, and Figure 11.3 draws the rows of cases 3, 6 and 7 as bars.

Table 11.3 Measured outcomes at the default seed

Case	Configuration	Metric	Value
2	Eight at 3.60 psi	mean thickness	700.00 nm
2	—	center to edge	6.94 nm
2	—	*WIWNU*	0.50 %
3	inline map	*RMSE*, wafers 1–3	14.6, 4.4, 0.35 nm
3	lot mean	*RMSE*, wafer 1	16.15 nm
3	product nominal	*RMSE*, wafer 1	29.76 nm
3	feedback only	mean error	+25.0 nm, all
4	fixed recipe, 260 wafers	leaves band at wafer	70
4	—	end-of-life error	+90 nm
5	$\lambda = 0.08$	bias, spread	−15.0, 6.4 nm
5	$\lambda = 0.30$	bias, spread	−10.3, 6.4 nm
5	$\lambda = 0.90$	bias, spread	−9.0, 9.2 nm
5	$\lambda = 0.90$	pressure travel	2.64 psi
6	hybrid APC	wafers in spec	84.2 %
6	feedforward only	wafers in spec	28.2 %
6	feedback only	wafers in spec	24.4 %
6	fixed recipe	wafers in spec	1.3 %
6	$\gamma = 0, 0.15, 0.4$	wafers in spec	80.5, 84.2, 87.1 %
6	$\lambda = 0.9$, $\gamma = 0.15, 0.6$	wafers in spec	89.8, 60.7 %
7	soft reset	qual wafers, worst	1, 2.5 nm
7	carry-all reset	qual wafers, worst	1, 41.7 nm
7	cold reset	qual wafers, worst	6, 59.4 nm
8	adaptive policy	wafers, release	1, 11.5 min
8	fixed policy	wafers, release	6, 18.4 min
8	after one wafer	$\sigma_{\hat{\theta}_g}$	0.0500 to 0.0070
9	VM only, 18 nm bias	surrogate, true	699.6, 682.8 nm
9	—	wafers in spec	46 % to 24 %
10	26 nm azimuthal	radial *RMSE*	0.14–4.8 nm
10	—	total *RMSE*	18.4–19.0 nm
11	as given	wafers in spec	13 of 890
11	suggested	wafers in spec	862 of 914
11	as given, suggested	qual wafers	12, 2
11	as given, suggested	film violations	333, 0
11	as given, suggested	worst mean error	140.4, 7.2 nm

11.6.1 The Plant Is Tall

A profile error is not a list of eight errors. Projecting a radial ripple of four full periods across the radius, $\sin(8\pi\rho)$, onto the column space of G by unregularized least squares leaves 66 % of it unreachable, and a ripple of two periods, $\sin(4\pi\rho)$, leaves 10 %. The right question about a residual is therefore not which chamber to move but whether the shape is in the reachable subspace at all, and the application reports the radial error and the azimuthal floor separately.

> **+ DID YOU KNOW? Equal Pressure, Unequal Removal**
>
> From the simulator: in Table 11.3, case 2 polishes with all eight chambers at 3.60 psi and hits the mean of 700.00 nm, yet the center finishes 6.94 nm thicker than the edge. The pad sweeps the wafer's edge faster than its center (Section 11.3), so equal pressure is not equal removal.

The unreachable case with the cleanest answer is azimuthal. Scenario 10 adds 26 nm of asymmetry with zero radial mean, a floor of $26/\sqrt{2} = 18.38$ nm. Over the 240 min run the radial residual stays between 0.14 and 4.8 nm while the total profile error is 18.4 to 19.0 nm against a 12 nm limit. Raising

λ to 0.8 does not move the floor. It lets the radial part reach 9.0 nm and the total 20.5 nm, because a loop tuned harder against an azimuthal error only adds radial noise to it. Control authority has a shape, and head gimbal, wafer bow, retaining ring wear and pad groove orientation are fixed in the machine shop rather than in the control system.

11.6.2 Feedforward and Feedback Answer Different Questions

Scenario 3 changes product at 90 min, flipping the incoming center-to-edge shape from 0.6 to −0.7 units of tilt (from 24 nm edge-thick to 28 nm center-thick at 40 nm per unit) on a film that stays 25 nm thick. On the first wafer after the change the inline map leaves a profile error of 14.6 nm, the second 4.4 nm and the third 0.35 nm, and after the second change at 180 min 18.9, 7.3 and 1.4 nm. The run-to-run move limit of 0.15 psi per wafer, which exists to guard against a noisy map, also meters out the feedforward correction, and raising it to 0.50 psi brings the first wafer to 1.08 nm. The lot-representative mean leaves 16.15 nm on every wafer and the product nominal 29.76 nm. The lot mean corrects mode zero and nothing else, so what survives it is the shape. Run-to-run feedback alone never removes the 25 nm level error at all, because the residual it learns from is the measured removal and an incoming film error is not a removal error. That is not a tuning failure. The information that would have fixed those wafers existed before they were polished and the architecture had no path for it.

Pause. Try Exercise 11.1 before reading on.

> **TRY IT IN THE APP 3 · Feedforward Sees the Wafer in Front of the Tool**
>
> Load the scenario and press Run. With Incoming profile known at Inline map, as shipped, watch the profile error on the first three wafers after the product change at 90 min: 14.6, 4.4 and 0.35 nm. Press Reset, set it to Lot mean before Run, and run again: 16.15 nm of shape stays on every wafer.

Feedback earns its place against everything upstream metrology cannot see. In scenario 6 a hybrid architecture holds 84.2 % of wafers in specification (255 of 303) across a whole pad life containing a slurry step and a product change. Feedforward alone holds 28.2 % and feedback alone holds 24.4 %, and the two middle numbers are close for entirely different reasons. Feedforward misses the slurry step completely, so every wafer after it is out of specification, and feedback never corrects the incoming shape change at 260 min, after which all 127 of its wafers are out. A fixed recipe runs the flat qualified recipe and holds 1.3 %.

11.6.3 What the Weight Trades, and What the Trend Term Costs

Scenario 5 drifts slurry activity upward in four steps while a poorly matched metrology tool adds 7 nm of correlated error to every map. Over the 229 production wafers polished from the first step at 60 min to 400 min, at $\lambda = 0.08$ the loop carries a standing bias of −15.0 nm with a spread (standard deviation) of 6.4 nm and moves the mean chamber pressure a total of 0.49 psi. At $\lambda = 0.30$ the bias is −10.3 nm at a spread of 6.4 nm and 0.96 psi. At $\lambda = 0.90$ the bias reaches −9.0 nm and costs 9.2 nm of spread, 2.64 psi of actuator travel and a worse peak error, 34.6 nm against 28.5 nm. The bias is the lag against the ramp and the spread is map noise being copied into silicon, the trade of Section 1.4.3. The bias shrinks monotonically across the whole slider, from −20.8 nm at $\lambda = 0.02$ to −9.0 nm at 1.0. The spread is smallest, 6.1 nm, near $\lambda = 0.12$ and grows on both sides, below it because a loop that far behind the ramp lets the four steps through as variance.

The trend weight is worth a few points here, and it turns on its owner only when the level weight is fast as well. Across a pad life at the shipped $\lambda = 0.30$, $\gamma = 0.15$ holds 84.2 % of wafers in specification against 80.5 % at $\gamma = 0$, twelve more good wafers, and the worst mean error in the second half of the pad falls from 12.3 nm to 6.9 nm. Larger trend weights keep paying at this level weight, 87.1 % at $\gamma = 0.4$ and 89.4 % at 0.6, because the two maps of each lot are twelve wafers apart and every slope is read across about twelve wafer-equivalents of exposure. At $\lambda = 0.9$ the level jumps most

of the way to the slurry step on the first map that sees it, the trend term reads that jump as the start of a ramp and extrapolates it across about 23 wafer-equivalents of metrology delay, and the mean overshoots the step. With $\gamma = 0.6$ the overshoot reaches 35.6 nm and the yield falls to 60.7 % (185 of 305), against 89.8 % (273 of 304) at $\gamma = 0.15$. A level step and the start of a ramp look the same across a metrology delay, and the challenge's suggested settings keep γ at 0.05.

11.6.4 The Price of Three Reset Strategies

Scenario 7 replaces a pad on identical hidden parameters and runs each strategy from the same seed. A soft reset releases production after one mapped monitor wafer and the worst of the first twelve production wafers is 2.5 nm off target. A carry-all reset releases just as fast, on a confident belief about a pad that is no longer on the machine, and the worst of the first twelve is 41.7 nm. A cold reset spends six qualification wafers and is still 59.4 nm out, because counting wafers is not the same as learning from them. The soft case at the end of that first production lot is shown in Figure 11.4.

> **TRY IT IN THE APP** **7 · The Pad Change, and What the Estimator Keeps**
>
> Load the scenario and press Run. Watch the Qual wafers tile after the pad is replaced at 30 min. Under the shipped Soft reset one mapped wafer releases production, and the worst of the first twelve production wafers is 2.5 nm off target. Press Reset, choose Carry all for Pad-change reset before Run, and run again: the release is as fast and the worst wafer is 41.7 nm off.

The single mapped wafer of the soft reset takes the standard deviation of the global gain estimate from 0.0500 to 0.0070 and the predicted joint pass probability to 0.99989. On the fresh pad of scenario 8 the adaptive policy therefore releases on the wafer polished at 11.5 min, where the fixed count releases at 18.4 min, one wafer against six.

Two qualifications are owed to that result. The correlated whole-map term caps how much any one wafer can say about global gain. Without it the same wafer would take the gain standard deviation 6.5 times lower, while the tilt, curvature and edge spreads (0.0050, 0.0041 and 0.0110 after the wafer) would hardly change. And the D-optimal choice of excitation, which is the right method and costs nothing, is worth well under one percent of the posterior spread on this plant (0.6 % in the geometric mean of the four standard deviations, at most 2.4 % on any one). A multiplicative pad correction is seen almost as well by any recipe that removes material, so the evidence is in the retained prior and in the thirty-three sites rather than in the experiment design. That is a measured result about this class of correction and not a general statement about experimental design.

The gate is also a gate. Triggering a chamber tracking fault before release causes the supervisor to refuse release whatever the probability says, because a chamber that is not tracking makes the whole prediction meaningless. In scenario 8 the tool spends its cap of four qualification wafers and goes to hold at 15.7 min.

11.6.5 A Surrogate With No Anchor

Scenario 9 removes physical metrology entirely and gives the virtual model an 18 nm bias at 40 min. By 100 minutes the surrogate reads 699.6 nm and the wafers are 682.8 nm. The bias enters the run-to-run residual as if it were a process error, the recipe removes it, and the true thickness moves by the whole of it while the measured mean on the top trend stays on target. In a fab the yield record would be the only symptom, and here the share of wafers in specification falls from 46 % at 100 min to 24 % at 190 min. The simulator also draws the actual wafer thickness as a dashed trace, which no real tool has. Physical metrology returns at 200 min with a one-lot delay, so the first physical map arrives at 239.4 min and finds the removal estimate carrying the whole bias. Its gain-mode residual differs from the estimator's prediction for that wafer by 21.5 nm, more than four times the 3.8 nm noise of a physical map, and the application raises its anchor alarm. That map moves the level estimate from −18.4 nm to −12.0 nm, and the actual mean is back above 692 nm at 241 min.

The return is that fast, and then too fast, because of the trend term of Section 11.4.4 at its default $\gamma = 0.12$. The level recovering from −18.4 nm looks exactly like the start of a ramp, the slope is 0.37 nm per wafer-equivalent after the first map and still 0.19 nm per wafer-equivalent near 387 min, and the estimator extrapolates it across the metrology delay. The actual mean overshoots to 711.0 nm at 350 min and is 2.4 to 10.4 nm high between 350 and 400 min. With $\gamma = 0$ the peak is 702.4 nm and the mean from 350 to 400 min is 699.0 nm, at the price of a slower return, above 692 nm only at 270.5 min.

The surrogate in the application is deliberately anchored at the last physical map rather than carrying the controller's live residual estimate. Letting it carry the live estimate makes the loop feed on its own output, the level then grows by λ times the bias at every update, the minimum film limit is crossed eleven wafers after the bias appears, and the mean settles near 450 nm once the level reaches its bound. That is a different and much less instructive failure than the one the scenario is about. The soft sensor of Volume 5, Chapter 10 is kept honest the same way, by a sampled and delayed laboratory that corrects its bias.

11.6.6 The Graded Lifecycle

The challenge runs 1400 minutes containing a slurry step, two product changes, one corrupted metrology map and two pad replacements, starting from a fixed recipe, a carry-all reset, a fixed six-wafer qualification policy and an EWMA weight of 0.06. It is scored on actual production thickness, not on the measured trend. The worst deviation of the production-wafer mean from 700 nm from 924 min to the end must stay within ±10 nm, and no more than six qualification wafers may be spent across both replacements. At least one full lot of 25 production wafers must be completed in the graded window, and the tool must finish released for production. This prevents a held last-wafer value from satisfying the thickness target without making product. The challenge is judged on the case as given, so changing a model parameter or a disturbance, including firing an extra disturbance button, voids the attempt. The suggested run completes 323 production wafers in that window. As given, the worst deviation is 140.4 nm, 13 wafers of 890 meet specification, twelve qualification wafers are consumed and the minimum film limit is crossed 333 times. With the suggested settings, which correct the architecture, the reset strategy, the qualification policy and the estimator weights, the same seeded run gives a worst deviation of 7.2 nm, 2.8 nm inside the bound, 862 of 914 wafers in specification, two qualification wafers in total, no film violations and 39 min of downtime. In the same graded window the measured mean on the top trend strays as far as 7.3 nm from target, which is metrology noise and delay rather than product. The corrupted map is delivered at 754.1 min with a normalized innovation of 9.34 and is rejected, just above the threshold of nine.

Most of that difference is the architecture and the estimator weights. The suggested settings with the shipped carry-all reset and fixed six-wafer qualification put back still hold 829 of 906 wafers in specification, but they spend twelve qualification wafers and miss the mean by 19.7 nm. The shipped architecture with only the reset and the qualification policy corrected holds 13 of 900. The two replacements decide the qualification budget, and the architecture decides the yield. No single change from the shipped setup passes. A pass in production takes the inline map read by Feedforward or Hybrid APC, a soft reset and at most three qualification wafers per pad, and with those alone the shipped estimator weights and metrology settings pass, at 4.8 nm and 85 % in specification.

11.7 Technology in Practice

The layers of Section 11.4 run on a thickness gauge, eight pressure regulators, the polish timer and the conditioner, and a recipe controller above the tool. Figure 11.5 draws them around the polisher of Figure 11.1.

What Measures It

The model reduces the gauge to a map of thirty-three radii on two wafers a lot, with 2.0 nm of independent noise at each site and 3.0 nm of correlated error on the whole map, delivered one or two



Figure 11.4 The application during scenario 7 at 80 min, 50 min after a pad replacement under the soft reset strategy and at the end of the first production lot on the new pad, cropped to the plant view and the top trend. (1) The badges read one qualification wafer and a standard deviation of the gain estimate of 0.0070. (2) The scene shows the platen from above with the slurry feed and the conditioner arm, and the carrier head chambers colored by the pressure they are holding. (3) The outgoing radial profile of the last wafer is drawn against the model prediction and the last delivered map. (4) The bars give the pad active surface (0.91), structural wear (0.075) and conditioner life. (5) The trend over the last hour carries the measured mean thickness with the actual wafer beside it (705.4 and 701.8 nm at this instant) and the minimum film limit. The single qualification wafer polished at 40.5 min stands out as the deliberately under-polished excursion to 792 nm.

lots late. Inside a tool, endpoint systems watch the polish while it happens. The motor current method is the less expensive, but its slow response limits what it can show of the process [17]. The chapter's polisher has no such sensor, so it learns the removal only from a map.

> **TECHNOLOGY** **Sensor: An Optical Map of Film Thickness**
>
> Optical methods that measure a film's thickness, the light it scatters or the reflectance of the wafer surface are the most common way to see a polished wafer. They give direct information about the surface, because the signal depends little on the other polishing conditions, while the pad itself is hard to watch this way [17]. The 1996 run-by-run controller read nine points on each blanket wafer [8], where this chapter reads thirty-three radii. No open, vendor-neutral source gives a production gauge's range or accuracy, so none is given here. A gauge that is poorly matched to its fleet adds the same error to every site of a map, as the 7 nm of case 5 does.

The inline pre-polish map is also a choice with a price: feedforward conventionally needs every lot measured, which most operations cannot afford [18].

What Moves It

The model's eight chamber pressures run from 2.5 to 4.8 psi, each a first order lag of 3.0 s at the center chamber falling to 0.5 s at the edge, beside a polish time of 35 to 90 s and a conditioning dose of 0 to 1.5.

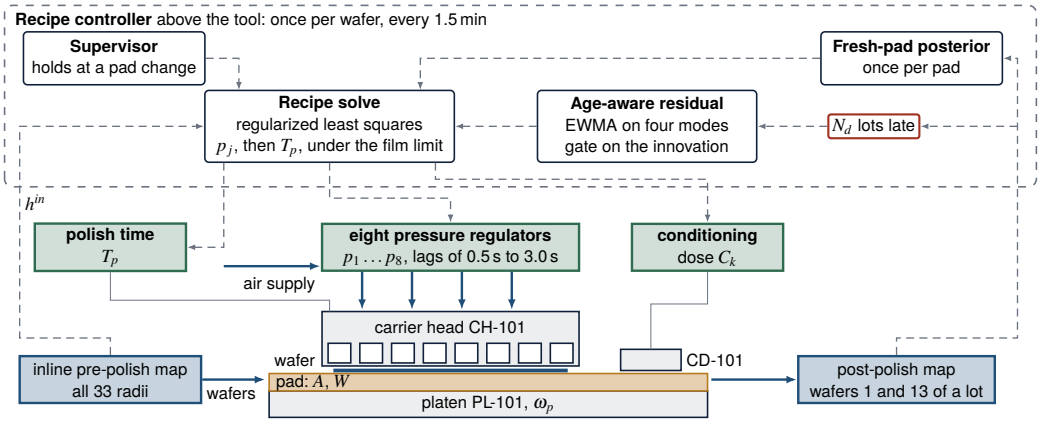


Figure 11.5 The layers of the simulated polisher drawn with their hardware. Eight pressure regulators feed the chambers of carrier head CH-101 and settle in seconds inside one wafer. The recipe solve writes their setpoints, the polish time T_p and the conditioning dose C_k once per wafer, from the inline pre-polish map and the controller's model. The post-polish map returns one to two lots late to the age-aware residual, and the fresh-pad posterior updates once per pad. The supervisor holds production at a pad change. Notice that the polished film is measured only after the next lots are already polished.

TECHNOLOGY **Final Element: An Air Pressure Behind Each Ring**

Each chamber of a membrane carrier head is a volume of air whose pressure presses one ring of the wafer against the pad. An electronic controller sets such a pressure through a current-to-pressure transducer, which takes the 4 to 20 mA output signal and acts as an electrically controlled air pressure regulator, reducing the supply pressure to the commanded level [19]. A study of a three-zone head gave each chamber a dynamic element set by the gas it holds [4], and the first-order lag of this chapter's chambers stands for it. A chamber that leaks shows as a pressure that no longer tracks its setpoint, and in case 8 the supervisor then refuses release.

The model's lag, paid for by Equation (11.5), stands for the regulator and the chamber volume. It has no dead band and no supply limit, and its only failure is the tracking fault of case 8.

Where the Controller Runs

In a fab the recipe layer belongs to advanced process control, which combines model-based run-to-run control with fault detection and classification, and run-to-run control adjusts the recipe lot by lot or wafer by wafer [18]. Supervisory and lower-level loops run with cycle times from milliseconds to minutes [20]. The app integrates every 0.05 min with the controller on the same scan, while the recipe and the estimator act once per wafer, every 1.5 min, the residual when a map arrives, and the posterior once per pad.

IN PRACTICE **Breaking In a New Pad**

A new pad is broken in before it polishes product. In one study about 50 copper wafers were polished to break in a pad, the radius of its surface asperities grew from 16 to 93 µm, and scratching became less severe [21]. In another, the surface texture that diamond disc conditioning cut into new pads settled within 30 minutes in four cases of five and within 60 minutes in all [22]. Maintenance shifts a tool's bias in a step [18], and the industry's usual answer is a fixed count of sacrificial qualification wafers after every pad change (Section 11.1).

In such a system the chapter's structure is a constrained least squares recipe block, an estimator for each of four profile modes behind an innovation gate, a Bayesian update of the fresh-pad parameters, and the supervisor's state machine.

What Is Modernizing It

The first development is virtual metrology with a stated confidence. A regression whose predictive variance sets the weight of the run-to-run EWMA did better on a polishing case study than schemes that ignore the reliability of the prediction [15], and models trained on production data from a high-volume fab predict non-uniformity and the removal across the wafer, with polishing time and carrier rotation among the strongest factors [14]. The second is sensing in the tool: a force transducer, the motor current and an acoustic emission sensor see different bands of the process and complement one another for monitoring and endpoint detection [17]. A recent review names prediction driven by artificial intelligence, sensing in the tool and cleaner consumables as what is turning polishing into a data-informed process [23]. Case 9 of this chapter shows the condition attached to the first.

11.8 Limitations

The influence matrix is a compact empirical representation. It omits membrane and pad mechanics, slurry hydrodynamics, pad groove transport, wafer bending, retaining ring contact mechanics and detailed relative velocity trajectories, and a real identified matrix must be loaded rather than generated.

The application grades a radial blanket-film profile. It does not predict die scale erosion, dishing, defects, scratches, residue, corrosion or electrical yield, and the eight chambers and their boundaries are supplied values for this project rather than an industry standard.

The release probability is only as meaningful as the covariance calibration behind it, and the joint probability multiplies three marginals that do not fail independently. It supplements process qualification and statistical process control and replaces neither.

Virtual metrology here is a model output with an explicit uncertainty, a switchable bias and an anchor at the last physical map. A real deployment needs data governance, product and tool matching, drift detection and independent validation, none of which is modeled.

OPEN CHALLENGE **A Virtual Measurement Worth Trusting**

A fab generates millions of data points a year for process control, and models trained on polisher data already predict non-uniformity and the removal at each point of a wafer [14]. A 2025 review of machine learning in polishing names what is still hard: data leakage, sparse target labels and hybrid modeling. It calls for benchmark data sets and for models of dishing and within-wafer non-uniformity [24]. Case 9 shows the stake. A surrogate with no physical anchor carries its own bias into the product. Progress would be a prediction whose confidence stays honest across a pad change, so that it can set how often wafers are measured [16].

The D-optimal choice of qualification recipe is measured at under one percent of the posterior spread on this plant. That is a property of a multiplicative pad correction observed through a shared aperture and should not be carried to a plant where the parameters enter differently.

The economic values are teaching assumptions. Wafer value, scrap cost, qualification cost, pad cost and downtime cost are round numbers chosen so the trade-offs are visible and are not fab cost data.

11.9 Conclusion

This chapter studied a 300 mm polisher as an interactive simulator in which eight carrier head chambers drive thirty-three radial outputs while the pad doing the work wears out and is eventually replaced. Eight continuous states carry the consumable and the unmeasured disturbances, a constrained least squares recipe solve carries the spatial inverse with its smoothness and gradient limits, an age-aware residual estimator carries the drift, and a four parameter Bayesian update carries the pad-to-pad variation.

The measured outcome that organizes the chapter is the price of a maintenance event. On identical hidden pads, a soft reset releases production after one mapped monitor wafer and lands 2.5 nm off target, a carry-all reset releases just as quickly on a confident belief about a pad that is in the bin and lands 41.7 nm off, and a cold reset spends six qualification wafers and is still 59.4 nm out. The rule that separates them is short enough to remember. Reset what belongs to the part that was replaced, keep what belongs to the machine, and raise the uncertainty rather than the gain.

KEY POINTS

- A polishing pad wears with every wafer and is replaced every few days, and the controller must then decide what it still believes.
- Eight concentric chambers cannot reach every radial profile, and no concentric chamber can touch an azimuthal error.
- Feedforward and feedback answer different questions: hybrid control held 84.2 % of wafers in specification, feedforward alone 28.2 % and feedback alone 24.4 %.
- Reset what belongs to the part that was replaced, keep what belongs to the machine, and raise the uncertainty rather than the gain.

Where else this applies. A regularized inverse of an ill-conditioned response in Volume 6, Chapter 6, a capacity that cleaning restores only in part in Volume 8, Chapter 11, and a soft sensor kept honest by a delayed laboratory in Volume 5, Chapter 10.

References

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

[2] 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.

[3] 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.

[4] S.-J. Shiu, C.-C. Yu, and S.-H. Shen, “Multivariable control of multizone chemical mechanical polishing,” *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 22, no. 4, pp. 1679–1687, 2004.

[5] J. Luo and D. A. Dornfeld, “Material removal mechanism in chemical mechanical polishing: theory and modeling,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 2, pp. 112–133, 2001.

[6] W.-T. Tseng and Y.-L. Wang, “Re-examination of pressure and speed dependences of removal rate during chemical-mechanical polishing processes,” *Journal of The Electrochemical Society*, vol. 144, no. 2, pp. L15–L17, 1997.

[7] E. Paul, “A model of chemical mechanical polishing,” *Journal of The Electrochemical Society*, vol. 148, no. 6, p. G355, 2001.

[8] D. S. Boning, W. P. Moyne, T. H. Smith, J. Moyne, R. Telfeyan, A. Hurwitz, S. Shellman, and J. Taylor, “Run by run control of chemical-mechanical polishing,” *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part C*, vol. 19, no. 4, pp. 307–314, 1996.

[9] 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.

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

[11] Y. H. Sun, R. K. Kang, and D. M. Guo, "Modeling of back pressure distribution on the wafer loaded in a multi-zone carrier in chemical mechanical polishing," *Materials Science Forum*, vol. 532-533, pp. 233–236, 2006.

[12] M. Tsujimura, "Chemical mechanical polishing (CMP) removal rate uniformity and role of carrier parameters," in *Advances in Chemical Mechanical Planarization (CMP)*. Elsevier, 2016, pp. 417–432.

[13] Z. C. Li, E. A. Baisie, X. H. Zhang, and Q. Zhang, "Diamond disc pad conditioning in chemical mechanical polishing," in *Advances in Chemical Mechanical Planarization (CMP)*. Elsevier, 2016, pp. 327–357.

[14] M. Breidung, T. Rothe, A. Lauff, P. Thieme, J. Langer, M. Günther, and H. Kuhn, "Process data-driven machine learning for non-uniformity prediction and virtual metrology in chemical mechanical planarization," *Journal of Intelligent Manufacturing*, 2025.

[15] J. Wan and S. McLoone, "Gaussian process regression for virtual metrology-enabled run-to-run control in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 1, pp. 12–21, 2018.

[16] G. A. Susto, "A dynamic sampling strategy based on confidence level of virtual metrology predictions," in *2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. IEEE, 2017, pp. 78–83.

[17] H. Jeong, H. Kim, S. Lee, and D. Dornfeld, "Multi-sensor monitoring system in chemical mechanical planarization (CMP) for correlations with process issues," *CIRP Annals*, vol. 55, no. 1, pp. 325–328, 2006.

[18] 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

[19] 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.

[20] 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.

[21] T. Eusner, N. Saka, and J.-H. Chun, "Breaking-in a pad for scratch-free, Cu chemical-mechanical polishing," *Journal of The Electrochemical Society*, vol. 158, no. 4, p. H379, 2011.

[22] J. McAllister, C. Stuffle, Y. Sampurno, D. Hetherington, J. Sierra Suarez, L. Borucki, and A. Philipossian, "Effect of conditioner type and downforce, and pad break-in time, on pad surface micro-texture in chemical mechanical planarization," *ECS Journal of Solid State Science and Technology*, vol. 7, no. 11, pp. P677–P688, 2018.

[23] H. Je, S. Kang, and S. Kim, "Challenges and innovations in chemical mechanical polishing in the more-than-moore era," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 13, no. 5, pp. 2717–2756, 2026.

[24] G. Winkler, T. Rothe, M. A. Sayyed, L. Jäckel, J. Langer, H. Kuhn, and M. Stoll, "Machine learning in chemical-mechanical planarization: A comprehensive review of trends, applications, and challenges," *Advanced Engineering Informatics*, vol. 68, p. 103663, 2025.

11.10 Exercises

11.1 ⌕ Concept check · 2 min

In scenario 3 the incoming film is 25 nm thick. With Recipe authority at EWMA R2R the 25 nm level error stands on every wafer of the run, while Hybrid APC with the inline map removes it. Why does run-to-run feedback alone never correct it?

(a) Its residual is measured removal minus predicted removal, and an incoming film error is not a removal error.

(b) The metrology delay holds every map back until after the run ends, so no correction can arrive in time.

(c) The run-to-run move limit of 0.15 psi per wafer meters each correction down to nothing.

(d) An EWMA weight of 0.30 is too small to follow a step of 25 nm within the run.

11.2 Concept check · 2 min

In scenario 7 the carry-all reset releases production after one qualification wafer, as fast as the soft reset, yet the worst of the first twelve production wafers is 41.7 nm off target against 2.5 nm for the soft reset. Why does the release gate pass under carry-all?

(a) Carry-all switches the qualification policy to a fixed count of one wafer.

(b) The gate is computed from the controller's own belief, which keeps the worn pad's estimate with a narrowed covariance.

(c) The qualification wafer under carry-all carries no excitation, so its map looks like a normal wafer.

(d) The first production wafers are measured two lots late, so the gate cannot see their error.

11.3 Concept check · 1 min

A step of 0.20 psi on interior chamber 4 stays above half its peak across 21 % of the wafer radius, while the same step on edge chamber 8 peaks lower, at 0.081 psi, and stays above half of it across 3 %. What gives the edge chamber its localized authority?

(a) It has a wider pressure range than the interior chambers.

(b) It has the fastest pressure time constant of the eight.

(c) Its radial band is narrow, so its footprint in G is narrow.

(d) It acts together with the retaining ring, which doubles its effect.

11.4 Theory and calculation · 6 min

The platen turns at 80 rpm, the carrier at 60 rpm, the centers are $d = 0.220\,\mathrm{m}$ apart and $R = 150\,\mathrm{mm}$. The innermost calculation point is at $\rho = 0.5/33$ and the outermost at $\rho = 32.5/33$. Use Equation (11.3) and give speeds to four decimals.

(a) Find the relative speed at the innermost and outermost points and their ratio. *Checkpoint:* 1.8431 m/s, 1.8689 m/s, ratio 1.0140

(b) With $\beta = 1$, eight equal pressures, a uniform pad and a mean removal of 500 nm, removal at each point is $500\,v_m/v_{ref}$ nm. Take $v_{ref} = 1.8564\,\mathrm{m/s}$ and find how much thicker the center finishes than the edge, to 0.01 nm. Compare with scenario 2. *Checkpoint:* 6.94 nm

(c) Which carrier speed makes the relative speed the same at every radius? What does scenario 2 then show for the center and edge thickness? *Checkpoint:* 80 rpm; both 700.00 nm

11.5 Theory and calculation · 8 min

Pad wear lowers removal by the factor $1 - d_w W^{1.4}$ with $d_w = 0.18$ (Equation (11.10)). Scenario 4 runs a fixed recipe that removes 500 nm on a pad of $N_{pad} = 260$ wafers at the default dose $C_{nom} = 0.55$, so wafer k is polished at $W = (k-1)/260$. Ignore the slow loss of conditioner efficiency.

(a) Find the removal factor at end of life, $W = 1$, and how far above target the mean then finishes. *Checkpoint:* 0.82, 90 nm

(b) Find the wear at which the mean is 14 nm above target, to four decimals, and the first wafer outside the ± 14 nm band. *Checkpoint:* $W = 0.2647$, wafer 70

(c) The dose is raised to $C = 1.00$ on a new conditioner ($\eta_{cond} = 1$). With one wafer-equivalent of exposure per wafer, find the settled active surface A^* at both doses (four decimals) and the pad life in wafers at $C = 1.00$ from Equation (11.8), to one decimal. *Checkpoint:* 0.9083 and 0.9474; 236.6 wafers

11.6 Theory and calculation · 7 min

A qualification wafer is mapped at $M = 33$ sites with $\mathbf{R} = s^2\mathbf{I} + \sigma_c^2\mathbf{1}\mathbf{1}^T$, $s^2 = \sigma_{ind}^2 + \sigma_{proc}^2$, $\sigma_{ind} = 2.0\,\text{nm}$, $\sigma_{proc} = 1.2\,\text{nm}$ and $\sigma_c = 3.0\,\text{nm}$. Treat the gain θ_g alone, with a prior standard deviation of 0.05 and a sensitivity $\mathbf{H} = r\mathbf{1}$ with $r = 430\,\text{nm}$ at every site (86 % of the nominal 500 nm removal).

(a) Use Equation (11.17) to show that $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} = Mr^2/(s^2 + M\sigma_c^2)$. Evaluate it and the posterior standard deviation of θ_g to four decimals. *Checkpoint:* about 20 175; 0.0070

(b) Repeat with $\sigma_c = 0$, and give the ratio of the two standard deviations to one decimal. *Checkpoint:* 0.00094; 7.4

(c) How many independent sites, each of variance s^2, would carry the same information about θ_g as the 33 correlated ones (two decimals)? What standard deviation does a wafer with infinitely many sites reach (four decimals)? *Checkpoint:* 0.59 sites; 0.0069

11.7 App challenge · 8 min

Load scenario 7, "The pad change, and what the estimator keeps", at the default seed. Run each case to 80 min at the default speed of 20 min per second and export the CSV, whose time column is `Time (h)`.

(a) As shipped (Pad-change reset at Soft), read the Qual wafers tile, the `est_sigma_gain` column at 0.70 h, and the largest $|\texttt{mean_prod_nm} - 700|$ in the rows from 0.68 to 0.99 h, to 0.1 nm. *Checkpoint:* 1 wafer; 0.0070; 2.5 nm

(b) Press Reset, choose Carry all before Run, and read the same three values. *Checkpoint:* 1 wafer; 0.0100; 41.7 nm

(c) Press Reset, choose Cold before Run, and read them again. Explain the three values of `est_sigma_gain`. *Checkpoint:* 6 wafers; 0.0500; 59.4 nm

11.8 App challenge · 8 min

Load scenario 8, "The qualification wafer sprint", at the default seed. Run each case to 30 min at the default speed and export the CSV (time column `Time (min)`). In each part find the first row whose `supervisor_state` is `enhanced` or `hold`, and read its time, `qual_total`, `est_sigma_gain` (four decimals) and `pass_prob` (five decimals).

(a) As shipped, with Qualification policy at Adaptive. *Checkpoint:* 12.0 min, enhanced; 1; 0.0070; 0.99992

(b) Press Reset and choose Fixed count before Run. *Checkpoint:* 18.5 min, enhanced; 6; 0.0029

(c) Press Reset, set the policy back to Adaptive and Release probability (in Advanced) to 0.9999, its editable maximum, before Run. *Checkpoint:* 12.0 min, enhanced; 1

(d) Reload the scenario, press Run, and press Zone 6 develops a leak at any time before 11.5 min. *Checkpoint:* 16.0 min, hold; 4

11.9 App challenge · 8 min

Load scenario 10, "The error concentric zones cannot reach", at the default seed, run it to 240 min at the default speed of 20 min per second, and export the CSV. Give values to 0.01 nm.

(a) Read the smallest and largest `rmse_nm` (the radial part), the largest `rmse_total_nm`, and `uncorrectable_nm`. Check the largest total against the sum in quadrature. *Checkpoint:* 0.14 and 4.81 nm; 19.00 nm; 18.38 nm

(b) Reload, set EWMA level weight to 0.80 before Run, and read the largest radial and total values again. *Checkpoint:* 9.03 nm; 20.48 nm

(c) The floor is $a/\sqrt{2}$ for an Azimuthal asymmetry of a nm peak, set in steps of 1 nm. Using the largest radial value of part (a), predict the largest a for which every wafer meets the 12 nm limit. Reload, set that value and one more before Run, and read the Wafers in spec tile at 240 min. *Checkpoint:* 15 nm: 100 %; 16: 99 %

11.10 Application area · 4 min

A fab runs pads of 1500 wafers. Its qualification policy is a fixed count of six monitor wafers after every pad change.

(a) What percentage of all wafers polished go to qualification under the fixed count, and under an adaptive policy that releases after one mapped wafer, as in scenario 8? Give two significant figures. *Checkpoint:* 0.40 % and 0.067 %

(b) The fab changes pad supplier. What should happen to the fresh-pad prior and to the release decision, and what does the simulator not do for the engineer? *Checkpoint:* widen or set aside the prior; nothing detects the change

11.11 Application area · 3 min

In scenario 9 virtual metrology is the only measurement and develops an 18 nm bias at 40 min. By 100 min the surrogate reads 699.6 nm while the wafers are 682.8 nm.

(a) Why does the measured mean on the top trend stay on target? *Checkpoint:* the loop removes the bias as if it were a process error

(b) The first physical map disagrees with the estimator by 21.5 nm. The physical map noise is 3.8 nm and the alarm threshold is four times that. Find the ratio to one decimal, and say how a production deployment should use such an alarm. *Checkpoint:* 5.7; raise the physical sampling rate

11.12 Review the agent · 5 min

An AI assistant reviews scenario 6, “Hybrid APC across a pad life”, and writes: “At the shipped level weight $\lambda = 0.30$, raising the trend weight γ from 0.15 to 0.6 lifts the wafers in specification from 84 to 89 %. Raising λ to 0.90 at $\gamma = 0.15$ lifts it to 90 %. Both changes help, so set $\lambda = 0.90$ and $\gamma = 0.60$ together for the best tool.” The two measurements are correct. Find the flaw and say how to check it on the app.