INTERACTIVE CONTROL ENGINEERING
Principles and Applications

Volume 14 Control in Semiconductor Manufacturing

CHAPTER 4

Wafer Scanner Simulator for Preview Feedforward Where No Measurement Exists

Pages 82–108

Edited by John D. Hedengren

How to cite this chapter. "Wafer Scanner Simulator for Preview Feedforward Where No Measurement Exists," in *Control in Semiconductor Manufacturing*, John D. Hedengren, Ed., vol. 14 of *Interactive Control Engineering: Principles and Applications*. Advanced Process Solutions, LLC, 2026, ch. 4, pp. 82–108.

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

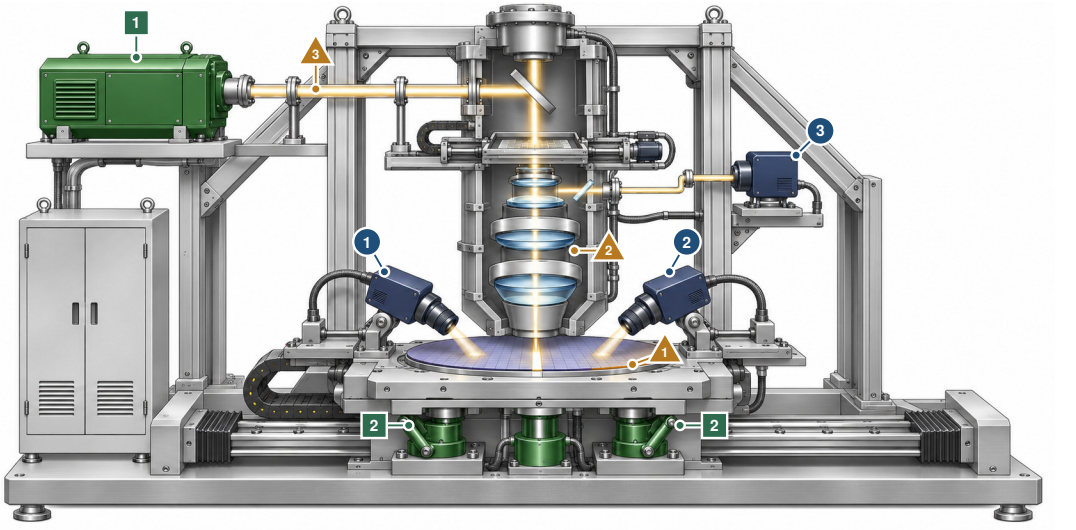
The app for this chapter
https://apmonitor.com/apps/litho-scanner/

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

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

CHAPTER 4

Wafer Scanner Simulator for Preview Feedforward Where No Measurement Exists



The controller moves 1 laser, whose pulse energy the dose loop sets
2 stage drives that lift and tilt the wafer, commanded from the sensor ahead

The controller reads 1 height sensor reading the surface ahead of the slit
2 height sensor reading the gap behind the slit, for the slow trim 3 energy monitor on the beam, read by the dose loop

What disturbs it 1 a bowed and rough wafer surface 2 a lens that warms and shifts the image size
3 laser pulses that scatter in energy

Notice that both height sensors aim beside the slit and never into it: the height where the light lands is never measured. Illustration, not to scale.

IN PLAIN TERMS **Keeping a Moving Wafer in Focus**

The main idea. Chips are printed by shining light through a narrow slot onto a silicon wafer that slides past underneath. The picture is sharp only within a very thin band of height. A stage therefore lifts and tilts the wafer as its uneven surface goes by.

The challenge. Nothing can measure the height where the light falls. A sensor reads the surface a short way ahead, and the stage moves early. How far ahead depends on how fast the wafer slides, so a faster machine quietly blurs the print.

Why it matters. Faster scanning prints more wafers each hour, and nothing on the machine warns that the focus has paid for it.

THREE WAYS TO READ THIS CHAPTER

Explore	11 min	The first page, the fact box in Section 4.1, the history box in Section 4.2, the fact and try-it boxes in Section 4.6, and the key points
Study	55 min	The whole chapter with its pause points and the twelve exercises
Practice	20 min	The first page, Section 4.6, Section 4.7, Section 4.8 with its open challenge, and the second challenge in Section 4.9

The control question Nothing can measure focus under the exposing slit of a scanner, so the stage is commanded from a level sensor that reads the wafer surface ahead of the slit. How far ahead should it read, why does that distance move whenever the scan speed is raised for throughput, and what is left for feedback to do?

Learning objectives

- Compute the group delay of a second-order stage servo and the matched preview distance at a given scan speed and bandwidth.
- Explain why a loop fed from a sensor behind the slit cannot be tuned into the depth of focus, from its transport delay, its servo lag and the frequency of the topography.
- Trace the offset of the swept optimum preview to the slope that the trailing-sensor trim reads, and set the trim slower than one field scan.
- Compute the dose floor $3\sigma_p/\sqrt{N}$ from the pulse count, and recognize a loop that has reached a statistical floor.
- Trade scan speed and alignment marks against focus, dose, overlay and wafers per hour.

Prerequisites Sections 1.4.7 (preview feedforward), 1.4.2 (what a tool can measure), 1.6.2 (velocity-form PID and loop action), 1.6.1 and 1.6.4 (runs, seed, tiles and data).

The app https://apmonitor.com/apps/litho-scanner/

See also Volume 3, Chapter 6 for feedforward from a measured disturbance with the late measurement demoted to a slow trim. Volume 5, Chapter 2 for control through a transport delay and Volume 5, Chapter 5 for qualities fixed before they can be measured. Volume 8, Chapter 10 for iterative learning on a repeated scan path and Volume 8, Chapter 9 for corrections from sampled, late results.

4.1 Introduction

A projection scanner exposes a wafer through a narrow slit, and the slit is where the light is. No sensor can occupy that space, so the one quantity the focus loop exists to regulate, the distance between the resist surface and the focal plane at the point being printed, has no instrument attached to it and never will. The measurement that does exist reads the surface a millimeter or two up the scan path, before the slit arrives. The stage is commanded from that reading. This is feedforward from a measurement of the disturbance itself, and the lead it supplies is set by geometry and by scan speed rather than by a tuning constant [1, 2]. Section 1.4.7 places this preview among the other feedforward paths of the fab.

Plants of this shape are common once the pattern is recognized. Any process that transports material past a fixed working point can measure the material before it arrives, and a controller that waits for an error at the working point is then accepting a delay it did not have to accept. Active suspensions driven by a road sensor looking ahead of the wheel are the textbook case [3], and the quarter car of Volume 7, Chapter 11 carries road-preview feedforward of this kind. The optimal-preview formulation that describes them is written for any plant whose future reference or disturbance is known a fixed interval in advance [4], including the case where the preview is wrapped around a feedback loop that is already commissioned [2]. The common feature is that the correct lead time is a property of the plant being driven, not of the disturbance, and that the correct lead *distance* is that time multiplied by the transport velocity. Change the velocity for throughput and a control parameter has moved with no indication anywhere that it did.

A student meets feedforward as a block diagram in which a disturbance transfer function is divided by a process transfer function, and the lesson usually stops at the algebra. Two things that decide whether the scheme works are missing from that picture. The first is that the lead is physical, so it is installed as a distance on a machine and is right only at one speed. The second is that the feedforward path is blind to any error in the sensor it reads, which is why a preview controller needs a slow feedback trim beside it and why that trim is a different loop with a different purpose. Neither point survives a transfer-function derivation, and both are visible in half a minute on a plant that exposes a wafer.

Simulation of lithography has been developed principally on the optical and chemical side, where aerial image, resist and process window models answer questions about linewidth [5, 6]. The mechatronic side is described in the control literature mostly at the level of a single stage axis, its resonances and its feedforward identification [1]. However, the machine-level statement that connects the two, that the wafer per hour number and the focus budget are set by one shared parameter, is difficult to demonstrate without a tool that writes a whole wafer and reports what came out of it. This chapter works with an interactive step-and-scan exposure simulator in which the focus loop is commanded from a sensor that reads ahead of the slit, so that the matched preview distance, the dose noise floor and the throughput can be driven from one panel and their coupling read off a finished wafer. Preview control and scanner physics are established subjects. What the chapter measures is how far the standard group-delay estimate of the matched preview distance sits from the swept optimum, and why.

> ✦ **DID YOU KNOW? A Heavy Stage Placed to Nanometers**
>
> A wafer scanner moves masses of around 100 kg, accelerates them at over 100 m/s^2, and places them to within nanometers in every degree of freedom in a few milliseconds [7]. A modern wafer may pass through the lithographic cycle up to 60 times [7]. The height axis of this chapter's stage is limited to 30 m/s^2.

Section 4.2 traces how the scanning slit created the focus control problem. Section 4.3 gives the eight-state model of the stage, the laser, the lens and the exposure sequence, with the hundred-field layout, the six-parameter grid fit and the step-and-scan state machine, in enough detail to rebuild the plant, and states where each assumption fails. Section 4.4 derives the matched preview distance from the servo group delay and compares it with a swept optimum. Section 4.5 describes the four corrections: preview feedforward with a trailing trim, open-loop leveling, the dose loop and the alignment fit. Section 4.6 runs the scenarios in an order that pulls apart architecture from tuning, spatial frequency from amplitude, and a statistical floor from a loop that is merely detuned, and ends with the graded challenge. Section 4.7 names the sensors, the stage and the controller behind the corrections. Sections 4.8 to 4.10 state the limitations, discuss the throughput trade and conclude.

4.2 From Contact Printing to the Scanning Slit

Early integrated circuits were printed by pressing a mask against the wafer and exposing the whole substrate at once. Contact printing gave excellent resolution and destroyed masks, proximity printing spared the mask and lost resolution to the gap, and neither had a focus control problem worth the name because the mask and the wafer moved together. Projection printing separated them and introduced the optical requirement that the resist sit within the depth of focus of a lens. Reduction step-and-repeat projection printers arrived in the late 1970s and dominated production for fifteen years, exposing one field, stepping and exposing the next [8].

A stepper can level. The field is stationary while the shutter is open, so the tool measures the surface under the stationary field, drives the stage to best focus, and only then exposes. Field-by-field leveling in exactly that form was introduced on high numerical aperture i-line steppers in the late 1980s [9] and was combined with a global alignment grid shortly afterward [10]. Both papers describe a measurement that finishes before the exposure starts, which is the defining property of the stepper era and the reason focus was a positioning problem rather than a tracking problem.

Step-and-scan began to replace step-and-repeat during the 1990s because the field outgrew what a lens could image at once. Scanning a slit across the reticle and the wafer together buys a large field from a small well-corrected image area, at the price of having to synchronize two stages while the light is on [11]. That change created the focus control problem in its present form. A scanning tool cannot level a field in advance, because the point being exposed moves across the field at hundreds of millimeters per second, so the stage must follow the surface during the exposure. Level sensing that reads ahead of the slit, and the software delay that matches the lead to the stage servo, arrived with

the architecture. The resulting focal plane budget was measured and published early in the scanner era [12], and the incoming wafer specification that supports it moved from a whole-wafer flatness number toward a site flatness number defined over one field, because a scanner cares about height change per unit of travel rather than about total thickness variation [13].

Two of the other three corrections in a scanner have equally specific origins. Exposure dose on an excimer-illuminated tool is delivered as a burst of discrete pulses, and the statistics of that burst were analyzed as a control problem as soon as excimer sources were proposed for lithography [14]. The process sensitivity those pulses feed, the dependence of linewidth on dose and on focus, is the Bossung description of the process window that has been standard since 1977 [5, 6]. Lens heating, the slow absorption of exposure energy in the projection optics and the magnification and focus drift that follows it, was characterized once tools became productive enough for the thermal transient to matter within a single lot [15]. All three of these survived the move to immersion and then to extreme ultraviolet, where the numbers tightened but the structure did not [16].

> ⧖ HISTORY **From Three Points to Every Field (1982 to 1991)**
>
> A patent application filed in 1982 described how a reduction projection printer leveled a wafer: an autofocus read the surface at three points and the whole wafer was set square to the lens. Larger wafers, new substrate materials and every round of exposure and processing made the surface less even, so each part of the wafer needed its own height, read with light sent in at an oblique angle [17]. In 1987 the autofocus of a stepper was turned into a gauge of flatness, and on new wafers it found, in the worst case, more than four micrometers of thickness variation within one field [13]. By 1991 a stepper leveled every field and still wrote more than 80 wafers of 150 mm an hour, because it no longer aligned each field separately [10]. The level sensor of this chapter reads the surface the same way, but while the wafer moves.

4.3 Process Model

The simulated tool writes a hundred 22 mm by 24 mm fields on a 300 mm wafer. Field centers sit on a 22 mm by 24 mm pitch and are selected in order of distance from the wafer center until one hundred are collected, then reordered into the serpentine the stage actually drives. The slit is 6 mm tall along the scan and a full field wide, so one field scan covers $L = H + h_s = 30\,\text{mm}$ of travel. Figure 4.1 shows a section through the slit along the scan direction, carrying every symbol the equations of this section use.

What the four corrections act on and act against is set out in Table 4.1. Working units are the ones a scanner engineer quotes, seconds for time, millimeters in the wafer plane, micrometers for stage height, microradians for tilt, nanometers for defocus, overlay and linewidth, and mJ/cm^2 for dose. Eight states are integrated. Two of the three controlled variables have no instrument on the machine, which is the subject of this chapter.

4.3.1 The Surface the Stage Has To Follow

The wafer surface is the sum of three terms of very different character, as shown in Equation (4.1).

$$z_w(x,y) = b\left(\frac{r^2}{R^2} - \frac{1}{2}\right) + a\,\sin\frac{2\pi x}{L_1}\cos\frac{2\pi y}{L_2} + \rho_t\,\varepsilon(x,y;L_m) \tag{4.1}$$

where b is the chuck clamp bow from center to edge, $r = \sqrt{x^2+y^2}$ is the radius in the wafer plane, R is the wafer radius, a is a long-wavelength shape of wavelengths L_1 and $L_2 = 0.74L_1$, ρ_t is the amplitude of the short-wavelength topography, and ε is a smooth pseudo-random field of unit amplitude and correlation length L_m built by hashing a lattice of pitch L_m and interpolating with a smoothstep. Every term is deterministic in (x,y), so the same wafer is the same wafer on every run and a sweep of a control parameter is a controlled experiment rather than a sample of a random surface.

The three terms carry the whole argument of this chapter. At 400 mm/s the bow of 1.5 µm across the 300 mm diameter presents itself to the servo at about 1.3 Hz, and over the hundred exposed fields

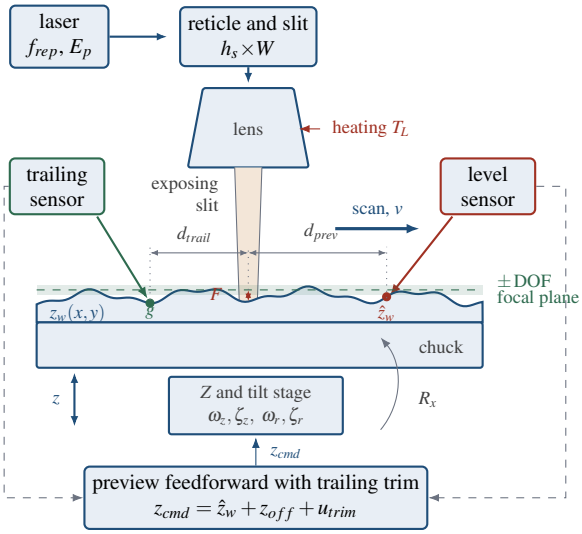


Figure 4.1 Section through the exposing slit along the scan direction. The exposure point travels left to right across the wafer at speed v, and the defocus F is the gap between the surface z_w and the focal plane at the slit, where no sensor can be placed because the slit is where the light is. The level sensor reads the surface d_{prev} ahead of the slit and supplies the feedforward command, the trailing sensor reads a gap g that is d_{trail}/v seconds old and supplies the trim, and the Z and tilt stage answers the command with natural frequencies ω_z and ω_r and damping ratios ζ_z and ζ_r. Only the incident ray of each oblique sensor is drawn, and the vertical scale of the surface is exaggerated by about four orders of magnitude against the horizontal.

the stage height command travels 1.74 µm and the stage itself 1.68 µm, while the servo's following error on it adds next to nothing to the error budget. The topography term is 0.05 µm, a thirtieth of that amplitude, and at $L_m = 7$ mm it presents itself at $v/L_m = 57.1$ Hz, which is where a 60 Hz servo lives. A spatial disturbance becomes a temporal one when it is divided by a velocity, and what a servo answers to is amplitude multiplied by the square of the resulting frequency.

The tilt stage removes the plane fitted across the slit and nothing finer, so the slope it is asked for is taken as a chord over $\pm h_c = 5.5$ mm rather than as a local derivative, as shown in Equation (4.2).

$$\theta(x,y) = 10^3 \, \frac{z_w(x+h_c,y) - z_w(x-h_c,y)}{2h_c} \tag{4.2}$$

where θ is in µrad and the factor 10^3 converts µm/mm to µrad. What no plane can follow is the curvature of the surface across the slit, given by Equation (4.3).

$$\kappa(x,y) = \frac{z_w(x+h_c,y) + z_w(x-h_c,y) - 2z_w(x,y)}{2} \tag{4.3}$$

Equation (4.3) is in µm and enters the focus budget with a weight of 0.35, which represents the root-mean-square of the second-order departure over the slit. Together with the across-slit tilt term of (4.9) it sets a floor that no servo bandwidth reaches. The curvature alone contributes 7.7 nm over the hundred fields of the nominal wafer, and the floor is measured at 8.8 nm in Section 4.6.

4.3.2 The Stages, the Laser and the Lens

The height stage is a second-order position servo specified as a bandwidth and a damping ratio, because those are the two numbers a stage engineer quotes and because the preview distance is

Table 4.1 Classification of Variables. Nominal Values Are Those of the Design Operating Point at Which Every Scenario Starts.

Symbol	Role	Nominal	Range
F defocus at slit	CV, unmeasured	0 nm	±60 nm DOF
$R_x - \theta$ tilt error	CV, unmeasured	0 µrad	±5 µrad band
$\dot{D}$ dose rate	CV, measured	400 mJ/(cm^2 s)	band 4
z_{cmd}	MV	surface	±60 µm
$R_{x,cmd}$	MV	slope	±400 µrad
E_{cmd} pulse energy	MV	0.440 mJ	0.02–1.10 mJ
v scan speed	operator input	400 mm/s	100–800 mm/s
n_m marks per wafer	operator input	20	4–40
D_{SP} dose	SP	30 mJ/cm^2	10–60 mJ/cm^2
z_{off} focus offset	SP	0 nm	±200 nm
clamp bow scale	DV, unmeasured	1×	0–3×
topography scale	DV, unmeasured	1×	0–3×
laser gain g_p	DV, unmeasured	1×	0.8–1.2×
monitor gain g_m	DV, unmeasured	1×	0.95–1.05×
level sensor bias δ	DV, unmeasured	0 nm	±150 nm
grid T_x, T_y	DV, sampled	8, −5 nm	±60 nm
grid m, ρ	DV, sampled	0.40 ppm, 1.20 µrad	±2, ±5
non-correctable A	DV, unmeasured	1.2 nm	0–15 nm
z, $\dot{z}$	states	surface, 0	±60 µm
R_x, $\dot{R}_x$	states	slope, 0	±400 µrad
D_{acc}	state	0 mJ/cm^2	≥ 0
E_p	state	0.440 mJ	0.02–1.10 mJ
T_L lens heating	state	0	0–1
ξ scan coordinate	state	0 mm	0–30 mm

written in the same two numbers. The differential equations are Equations (4.4) and (4.5).

$$\ddot{z} = \mathrm{sat}\left[\omega_z^2(z_{cmd} - z) - 2\zeta_z\omega_z\dot{z},\ \pm a_{max}\right], \quad \omega_z = 2\pi f_z \tag{4.4}$$

$$\ddot{R}_x = \omega_r^2(R_{x,cmd} - R_x) - 2\zeta_r\omega_r\dot{R}_x, \quad \omega_r = 2\pi f_r \tag{4.5}$$

where z is in µm, R_x is in µrad, and $a_{max} = 30\,\mathrm{m/s^2}$ is the acceleration limit of the height axis expressed in µm/s^2 inside the saturation. The height state is clipped to $\pm z_{lim}$ and the tilt state to $\pm R_{lim}$ after every step, so a command that exceeds the stroke moves the axis to the stroke and no further. The laser follows its energy command through one lag, as shown in Equation (4.6).

$$\tau_L \frac{\mathrm{d}E_p}{\mathrm{d}t} = E_{cmd} - E_p \tag{4.6}$$

The lens absorbs a fixed fraction of every pulse and relaxes when the shutter is closed, as shown in Equation (4.7).

$$\frac{\mathrm{d}T_L}{\mathrm{d}t} = \begin{cases} (1 - T_L)/\tau_{up}, & \text{exposing} \\ -T_L/\tau_{dn}, & \text{otherwise} \end{cases} \tag{4.7}$$

where $T_L \in [0, 1]$ is the normalized heating state. It drives a magnification error $k_L T_L$ in parts per million and is measured by nothing on the tool. The scan coordinate inside the field advances at the scan speed while the field is being written, $\dot{\xi} = v$ during exposure and zero otherwise, and the field is complete when ξ reaches L.

4.3.3 Dose and What Arrives on the Wafer

A point in the field receives the pulses that happen to pass over it while the slit crosses, and because the slit is a full field wide the energy density per pulse is $E_p/(Wh_s)$ while the point sees $f_{rep}h_s/v$ of

them. The slit height cancels, and the delivered dose is Equation (4.8).

$$D = \frac{E_p\, g_p\, f_{rep}}{W\, v}, \qquad N = \frac{f_{rep}\, h_s}{v}, \qquad \frac{\sigma_D}{D} = \frac{\sigma_p}{\sqrt{N}} \tag{4.8}$$

where g_p is the laser gain disturbance, N is the pulse count a point receives and σ_p is the pulse-to-pulse energy scatter. With $f_{rep} = 6000\,\text{Hz}$ and $h_s = 6\,\text{mm}$, N is 90 pulses at 400 mm/s and 51.4 at 700 mm/s. The state actually integrated is the dose accumulated over the whole field scan, $\dot{D}_{acc} = E_p g_p f_{rep}/(WL)$ while exposing and zero otherwise, which is equal to (4.8) at the end of the field and is available to the controller while the field is still being written. The scan ends on a controller period, so the slit can travel past L to ξ_{end} by up to one period of travel (it overshoots by 0.24 mm at 480 mm/s), and the integral is scaled back by L/ξ_{end}, because the dose a point receives does not depend on when the scan stops. Scatter is applied once when the field completes, so the dose error credited to field i is $\Delta D_i/D = D_{acc,i}(L/\xi_{end})(1+\eta_i)/D_{SP} - 1$ with η_i drawn from a normal distribution of standard deviation $\sigma_p/\sqrt{N}$.

The defocus at the slit is the difference between where the stage is and where the recipe wants it, $F = z - (z_w(x,y) + z_{off})$, and the quantity the process responds to combines it with the two terms a stage cannot remove, as shown in Equation (4.9).

$$F_{tot}^2 = F^2 + \frac{1}{3}\left(\frac{W}{2}\,\varepsilon_R\right)^2 + (0.35\,\kappa)^2 \tag{4.9}$$

where $\varepsilon_R = R_x - \theta$ is the across-slit tilt error in radians and $W/2$ is the half field width expressed in µm so that the term carries the units of F. The second term is the mean square of a linear defocus ramp across the field whose value at the edge is $(W/2)\varepsilon_R$, which is what a stage whose slit center is in perfect focus still writes when it has not picked up the surface slope. Equation (4.9) is accumulated only while the shutter is open, and the field result is $F_{rms,i} = \sqrt{\langle F_{tot}^2 \rangle}$ over that interval. Linewidth follows the standard exposure-latitude and Bossung correlation [5, 6], as shown in Equation (4.10).

$$CD_i = CD_0 + s_E \ln\left(1 + \frac{\Delta D_i}{D}\right) - c_F F_{rms,i}^2 \tag{4.10}$$

where s_E is the exposure latitude in nm per natural log of dose and c_F is the Bossung curvature in nm/µm^2. A field is counted as failed when $F_{rms,i}$ exceeds the depth of focus, or the overlay of (4.13) exceeds its specification, or $|CD_i - CD_0|$ exceeds the linewidth tolerance.

4.3.4 Overlay as a Fit to a Sample

The placement error of the printed layer relative to the one underneath it is four correctable terms and one that no linear grid model can represent, as shown in Equation (4.11).

$$\begin{aligned} dx &= T_x + MX - \rho Y + A \sin\frac{2\pi X}{\lambda_{nl}} \\ dy &= T_y + MY + \rho X + A \sin\frac{2\pi Y}{\lambda_{nl}} \end{aligned} \tag{4.11}$$

where X and Y are the field center coordinates in mm, dx and dy are in nm, $M = m + k_L T_L$ is the magnification error in ppm including the lens heating contribution, ρ is the grid rotation in µrad, so that each product with a coordinate in mm is directly in nm, A is the amplitude of the non-correctable distortion and $\lambda_{nl} = 90\,\text{mm}$ is its period. Alignment marks are measured on n_m of the hundred fields, chosen at a stride that spreads them over the wafer rather than clustering them at one radius, and each measurement carries a noise of standard deviation σ_a. Three coefficients per axis are recovered by ordinary least squares through the normal equations, as shown in Equation (4.12).

$$\hat{c}_x = (A_m^\top A_m)^{-1} A_m^\top b_x, \qquad A_m = \begin{bmatrix} 1 & X_k & Y_k \end{bmatrix} \tag{4.12}$$

where A_m carries one row per measured mark, b_x is the vector of measured x displacements at those marks, and the y axis is fitted on $\begin{bmatrix} 1 & Y_k & X_k \end{bmatrix}$ so that the same three-column solver serves both. The fit is performed once, after the marks are measured and before the first field is exposed, and is then applied open loop to all one hundred fields. The residual left on field i is Equation (4.13).

$$e_x = dx - (\hat{c}_{x0} + \hat{c}_{x1} X + \hat{c}_{x2} Y + u_L X) + \sigma_{sync} n \tag{4.13}$$

where u_L is the optional lens heating feedforward of Section 4.5 and σ_{sync} is the per-field stage and reticle synchronization error. The wafer figure of merit is $3\sqrt{\langle e_x^2 + e_y^2 \rangle / 2}$, pooled over both axes, because that is the form an overlay specification is written in.

4.3.5 Sequence and Throughput

The tool runs a state machine of five phases. A wafer exchange takes t_{load}, the alignment phase visits the n_m marks in turn at t_{mark} each, and then one hundred cycles of a step-and-settle of t_{step} followed by an exposure of L/v are executed before the wafer is unloaded. Throughput follows in Equation (4.14).

$$\text{wph} = \frac{3600}{n_m t_{mark} + 100\,(t_{step} + L/v) + t_{load}} \tag{4.14}$$

At the design point Equation (4.14) gives $3600/(6.0 + 16.5 + 0.5) = 156.5$ wph with the reduced 0.5 s exchange the running scenarios use, against a measured 155.8 wph, and 141.2 wph with the default 3.0 s exchange against a measured 140.6 wph. The difference of under half a percent is one 1 ms controller period per field, spent on the transition from step to exposure. The end of the scan is detected at the first controller execution at which the slit has traveled L, within 1 nm of rounding, so at 400 mm/s each field is exposed for exactly 75 ms. Every term in the denominator is a control decision. The alignment time is a sampling plan, the scan time is a focus and dose decision, and both are traded in Section 4.6.

4.3.6 Parameters, Initialization and Numerics

The constants appear in Table 4.2. Three of them never reach the parameter panel, and those three set dynamics a student has to infer from behavior rather than read off a slider.

Every scenario opens with the stage standing on the surface it is about to expose. The height state is initialized to $z(0) = z_w(0,0)$ rather than to the nominal plane, because a clamped 300 mm wafer sits more than a micrometer away from that plane and a stage started at zero begins twenty depths of focus out. The tilt state is initialized to the slope the servo is about to be asked for, $R_x(0) = \theta(0, d_{prev,R})$ with $d_{prev,R}$ the matched tilt lead of Section 4.5, the pulse energy to the open-loop value for the requested dose, and the lens heating and scan coordinate to zero. A shipped numerical test confirms that a stage commanded to the surface it stands on, with the laser at its commanded energy and the shutter closed, has derivatives below 10^{-9} in every state.

The engine of Section 1.6.1 carries the state vector through fourth-order Runge-Kutta on a 0.5 ms fixed step, which leaves a 60 Hz servo well inside the step. Both controllers execute every 1 ms, and 200 rows per simulated second reach the log. Every random draw traces back to seed 20260916, so a swept parameter is the only thing that differs between two runs.

4.3.7 Assumptions

Detail is concentrated on the servo and the sampling, and taken out of everything else. Each simplification has a place where it stops working.

Scalar imaging. Equation (4.10) is a dose and focus correlation, not a lithography simulation. There is no aerial image, no mask topography and no illumination shape, so nothing here reaches optical proximity correction, and the Bossung response is a single quadratic where a real process window is a two-dimensional region in dose and focus whose shape is most of the engineering [5].

Table 4.2 Constants of the Simulated Scanner. The Hidden Three Never Appear on the Parameter Panel, and the Tuning Block Is Shown at Its Shipped Values.

Symbol	Meaning	Value	Unit	Visibility
W	field width	22	mm	fixed
H	field height	24	mm	fixed
R	wafer radius	150	mm	fixed
h_s	slit height along scan	6	mm	adjustable
f_{rep}	laser repetition rate	6000	Hz	adjustable
σ_p	pulse energy scatter	3.0	%	adjustable
b	clamp bow, center to edge	1.5	µm	adjustable
a	long-wavelength shape	0.25	µm	adjustable
ρ_t	short-wavelength topography	0.05	µm	adjustable
L_m	topography correlation length	7	mm	adjustable
h_c	slit chord for slope and curvature	5.5	mm	fixed
z_{lim}	height stage stroke	60	µm	adjustable
R_{lim}	tilt stage stroke	400	µrad	adjustable
DOF	depth of focus, half-width	60	nm	adjustable
CD_0	linewidth at nominal	62	nm	adjustable
s_E	exposure latitude	−24	nm	adjustable
c_F	Bossung curvature	120	nm/µm^2	adjustable
—	linewidth tolerance	2.0	nm	adjustable
—	overlay specification, 3σ	5.0	nm	adjustable
σ_{LS}	level sensor noise	6	nm	adjustable
σ_a	alignment mark noise	1.0	nm	adjustable
σ_{sync}	stage synchronization	0.60	nm	adjustable
t_{mark}	time per alignment mark	0.30	s	adjustable
t_{step}	step and settle between fields	0.090	s	adjustable
t_{load}	wafer exchange	3.0	s	adjustable
k_L	magnification at full heating	0.15	ppm	adjustable
τ_{up}	lens heating time constant	25	s	adjustable
τ_{dn}	lens cooling time constant	180	s	adjustable
L_1	long shape wavelength	95	mm	hidden
a_{max}	height axis acceleration limit	30	m/s^2	hidden
τ_L	laser energy response	0.004	s	hidden
d_{prev}	preview distance	1.70	mm	tuning
d_{trail}	trailing sensor distance	2.0	mm	tuning
f_z, ζ_z	height servo	60, 0.80	Hz	tuning
f_r, ζ_r	tilt servo	45, 0.80	Hz	tuning
$K_{c,f}$	focus trim gain	0.30	µm/µm	tuning
$\tau_{I,f}$	focus trim reset time	0.800	s	tuning
$K_{c,D}$	dose loop gain	8×10^{-4}	—	tuning
$\tau_{I,D}$	dose reset time	0.006	s	tuning
$\hat{k}_L$	lens model coefficient	0.150	ppm	tuning

One height axis and one tilt axis. A production scanner controls six degrees of freedom on the wafer stage and six more on the reticle stage, synchronized to each other to sub-nanometer moving averages [1], and that synchronization carries a large share of the real overlay budget. Here it appears only as the scalar σ_{sync} of (4.13). The assumption removes the cross-coupling that decides how much bandwidth a real stage can be given.

The stage is a clean second order system. Equation (4.4) has no structural resonance, and on a real stage the first flexible mode is usually what limits the achievable bandwidth [1]. The consequence is that raising f_z in this application is free where in a machine shop it is not.

The surface is synthetic. Equation (4.1) is a smooth pseudo-random field rather than a measured wafer, and it carries no topography from the layers underneath. In production, process-induced topography over printed structures is most of the focus problem, and it is not smooth.

Pulses are not simulated individually. The dose integral is continuous and the scatter of (4.8) is applied once when the field completes, so within-field dose structure caused by the servo transient at scan start is absent. The $1/\sqrt{N}$ statement itself is exact, and Section 4.6 verifies it against the model to within 3.8 % at every speed tested.

Alignment is six parameters fitted once per wafer. Production systems fit twenty to forty parameters, apply a correction per exposure field, and close an overlay loop from measurements on the previous lot. The application has none of those, so it overstates how much of the residual a sampling plan alone can remove.

Lens heating drives magnification only. Equation (4.7) is one first-order state. Real lens heating moves several aberrations with different time constants, and it moves focus as well as magnification [15].

The sequence has no stage overlap. Equation (4.14) charges the alignment time to the wafer being written. A dual-stage tool measures the next wafer while the current one is exposed and hides that time entirely, which would change the marks-against-throughput trade of Section 4.6 substantially.

Two disturbances are stylized. A level sensor that acquires a 60 nm bias in one step is not what a real sensor does, but it isolates the one disturbance the feedforward path is structurally blind to. The 0.5 s wafer exchange of the running scenarios is shorter than any real load sequence, and it brings the second wafer quickly to a reader watching at a quarter of real time.

4.4 Why the Preview Distance Belongs to the Scan Speed

The command in Figure 4.1 is built from a surface height measured d_{prev} ahead of the slit, so relative to the point being exposed the command is issued d_{prev}/v seconds early. The plant that answers it is (4.4), whose transfer function is $\omega_z^2/(s^2+2\zeta_z\omega_z s+\omega_z^2)$. Differentiating its phase at the origin gives the low-frequency group delay in Equation (4.15).

$$\tau_g = -\left.\frac{\mathrm{d}\phi}{\mathrm{d}\omega}\right|_{\omega\to 0} = \frac{2\zeta_z}{\omega_z} \tag{4.15}$$

Setting the lead equal to the lag gives the matched preview distance of Equation (4.16), which is the design rule the application puts in front of the reader.

$$d_{prev}^{*} = v\,\tau_g = v\,\frac{2\zeta_z}{\omega_z} = \frac{v\,\zeta_z}{\pi f_z} \tag{4.16}$$

At the design point $\tau_g = 2(0.80)/(2\pi\cdot 60) = 4.244$ ms and (4.16) returns 1.70 mm at 400 mm/s and 2.97 mm at 700 mm/s. Two properties of (4.16) matter more than its value. It is proportional to the scan speed, which is the throughput knob, and it is inversely proportional to the bandwidth, which is the stage design. Neither of the two is a control parameter in the mind of the engineer who moves it.

🕮 **Pause.** Try Exercise 4.1 before reading on.

Equation (4.16) is an estimate and not an identity, because it equates a lead at all frequencies to a group delay that is flat only well below ω_z. The topography this servo is chasing arrives at 57.1 Hz against a bandwidth of 60 Hz, which is exactly the region where the phase is not linear in frequency. Table 4.3 measures the difference by sweeping d_{prev} in 0.05 mm steps over a full wafer at each configuration and recording the focus RMS, and Figure 4.2 plots three of those sweeps.

The first block of Table 4.3 carries the central measurement of this chapter. The measured optimum moves from 1.70 mm at 250 mm/s to 3.65 mm at 700 mm/s, a change of 1.95 mm, while the estimate moves from 1.06 to 2.97 mm, a change of 1.91 mm. The slope is 1.02, so (4.16) scales correctly with scan speed. Across the bandwidth block the estimate falls from 3.40 to 0.85 mm while the measured optimum falls from 3.90 to 1.55 mm, a slope of 0.92 in the same sense.

Table 4.3 Swept Optimum Preview Distance Against the Estimate of Equation (4.16). Each Entry Is the Minimum of a 0.05 mm Sweep Over One Complete Wafer of One Hundred Fields at the Shipped Seed, Where the Surface Is Deterministic and the Sweep Is a Controlled Experiment. F_{rms} at d^* Is Read at the 0.05 mm Setting Nearest d^*, the Resolution the Preview Distance Accepts.

v (mm/s)	f_z (Hz)	ζ_z	d^*_{prev} (mm)	d^{opt}_{prev} (mm)	$d^{opt} - d^*$ (mm)	F_{rms} at d^{opt}	F_{rms} at d^*
250	60	0.80	1.06	1.70	0.64	9.80	12.49
400	60	0.80	1.70	2.35	0.65	11.29	13.76
480	60	0.80	2.04	2.70	0.66	12.39	14.64
550	60	0.80	2.33	3.00	0.67	13.42	15.53
700	60	0.80	2.97	3.65	0.68	16.08	18.05
400	30	0.80	3.40	3.90	0.50	14.92	16.01
400	90	0.80	1.13	1.85	0.72	10.63	13.40
400	120	0.80	0.85	1.55	0.70	10.43	13.51
400	60	0.50	1.06	1.80	0.74	10.43	13.89
400	60	1.20	2.55	3.05	0.50	13.29	14.53

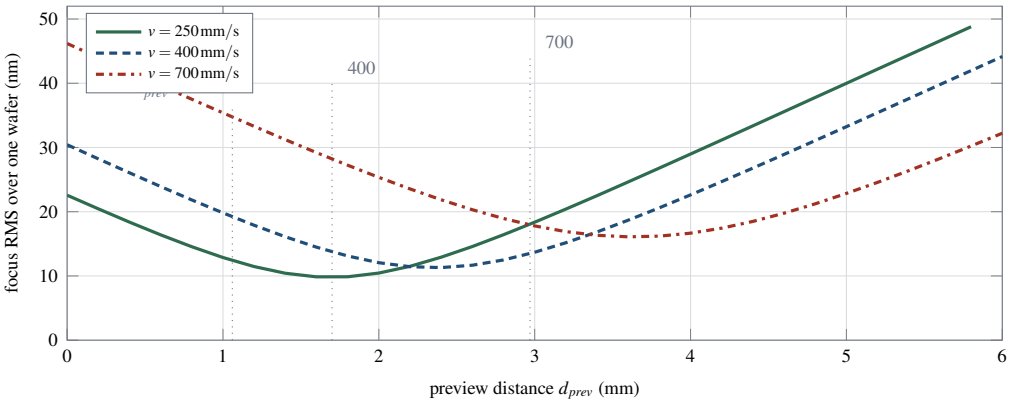


Figure 4.2 Focus RMS over one complete wafer against preview distance at three scan speeds, 250 mm/s solid, 400 mm/s dashed and 700 mm/s dash-dotted, with the estimate of Equation (4.16) marked by the thin gray dotted lines. Every plotted point is one complete wafer of one hundred fields written with the same servo, the same surface and the same seed. The minimum moves out with scan speed as the estimate predicts and sits about 0.65 mm beyond it in all three cases, the lead the trailing-sensor trim takes back, and Table 4.3 locates it on a finer 0.05 mm sweep.

What (4.16) does not give is the offset. The measured optimum exceeds the estimate by between 0.50 mm and 0.74 mm in all ten configurations, and the excess is far more nearly a constant distance than a constant fraction. Stated as a ratio the same data run from 1.15 to 1.83, which makes the ratio the less useful of the two descriptions and makes a quoted percentage correction to (4.16) misleading. The offset does not belong to the servo. It belongs to the trailing-sensor trim of Section 4.5, whose proportional term compares the stage with the surface d_{trail} behind the slit. With the stage on the surface that gap reads d_{trail} times the local surface slope, and the trim subtracts $K_{c,f} d_{trail}$ times the slope from the command, which shortens the effective preview by $K_{c,f} d_{trail} = 0.60$ mm at the shipped gain of 0.30 and distance of 2.0 mm. With the trim gain set to zero the optimum returns to within 0.16 mm of the estimate in all five configurations tried (1.80 mm against 1.70 at 400 mm/s and 3.10 mm against 2.97 at 700 mm/s), and at 400 mm/s the excess follows the trim, 0.40 mm at a gain of 0.15 or a distance of 1.0 mm and 1.15 mm at a gain of 0.60. The gap is a slope only while

d_{trail} is short against the correlation length, which is consistent with the offset falling to 0.40 mm at $L_m = 3$ mm and reaching 0.70 mm at $L_m = 20$ mm with everything else held.

Pause. Try Exercise 4.3 before reading on.

The practical consequence is that (4.16) is a starting point for a search and not an answer. Using the estimate directly at 400 mm/s gives 13.76 nm RMS against 11.29 nm at the swept optimum, and no preview at all, meaning a sensor reading directly under the slit, gives 30.43 nm. The estimate therefore captures 87 % of the available improvement and the search recovers the rest. With the trim gain at zero the estimate itself gives 10.51 nm and the search 10.35 nm, so on a wafer with no sensor bias for the trim to remove it costs 0.9 nm of focus RMS at the optimum.

4.5 Implemented Controllers

Four corrections run at once, and only one of them is a feedback loop closed on the variable it regulates.

4.5.1 Focus: Preview Feedforward With a Trailing Trim

The level sensor reports the surface ahead of the slit with a bias and additive noise, as shown in Equation (4.17).

$$\hat{z}_w = z_w(x,\ y + \sigma_d\, d_{prev}) + \delta + \sigma_{LS}\, n_k \tag{4.17}$$

where $\sigma_d = \pm 1$ is the direction of the current scan, δ is the level sensor bias disturbance and n_k is an independent standard normal draw at each controller execution. The trailing sensor reports the gap behind the slit, $g = z - z_w(x, y - \sigma_d d_{trail}) + \sigma_{LS} n_k$, which is the only true feedback available anywhere in the focus path and is d_{trail}/v seconds old. The command combines the two, as shown in Equation (4.18).

$$z_{cmd} = \mathrm{clamp}\big(\hat{z}_w + z_{off} + u_{trim},\ \pm z_{lim}\big) \tag{4.18}$$

Selecting the trailing architecture replaces $\hat{z}_w$ in (4.18) with zero, which holds the stage at the nominal plane and leaves u_{trim} as the only thing driving it. That switch is the experiment of Section 4.6 and it changes no gain.

The trim is the velocity-form PID of Section 1.6.2, running on the 1 ms controller period Δt_c with the setpoint z_{off} and the process variable g. Each scan computes increments, as shown in Equation (4.19).

$$\begin{aligned} e_k &= a\,(SP_k - PV_k), \qquad a = +1 \\ \Delta u_k^P &= K_c\,(e_k - e_{k-1}), \qquad \Delta u_k^I = \frac{K_c}{\tau_I}\, e_k\, \Delta t_c \\ s_k &= s_{k-1} + \Delta u_k^P, \qquad c_k = s_k + I_{k-1} \\ u_k &= \mathrm{clamp}(c_k,\ u_{min}, u_{max}) \\ I_k &= I_{k-1} + \Delta u_k^I + \frac{\Delta t_c}{T_t}\,(u_k - c_k) \end{aligned} \tag{4.19}$$

Equation (4.19) is Equations (1.12) to (1.15) in this chapter's symbols: s_k is the running sum S_k, c_k is the output before limiting v_k, and the output bias is zero. The action $a = +1$ makes the loop reverse acting in the convention of Section 1.6.2, which is correct here because the gap rises with the command: a stage that is too high reports a gap that is too large and must be commanded down. The integral is seeded at reset with the starting output, zero for the trim and the open-loop pulse energy for the dose loop. Back-calculation corrects only the integral, with the tracking time T_t equal to τ_I. The derivative time τ_D is zero on both loops in this application, so no derivative increment appears. Output limits are ± 20 µm on the trim and no rate limit is applied. The stroke clamp of (4.18) acts after the trim and is not fed back to it, which is immaterial because the command stays within a few micrometers of the nominal plane. Under manual operation I_k is back-solved so that c_k equals the value the operator has dialed in, which makes the transfer bumpless in both directions.

The trim defaults to $K_c = 0.30\,\mu\text{m}/\mu\text{m}$ and $\tau_I = 0.800\,\text{s}$, which is deliberately slower than one field scan of 75 ms. The reason is developed in Section 4.6. Its purpose is to remove a static bias δ in (4.17), which the feedforward path is blind to by construction because the bias is in the sensor the feedforward reads. The trim compares the stage with the surface behind the slit without removing the slope the preview has already followed, so its proportional term also reads the local surface gradient as a focus error, which Section 4.4 measures as a lead of about $K_{c,f} d_{trail}$.

4.5.2 Leveling: Feedforward Only

The tilt command is the measured slope of (4.2) evaluated at the tilt servo's own matched lead, as shown in Equation (4.20).

$$R_{x,cmd} = \text{clamp}\left(\theta\left(x,\; y + \sigma_d \frac{2v\zeta_r}{\omega_r} \right) + \tfrac{1}{4}\sigma_{LS} n_k,\; \pm R_{lim} \right) \tag{4.20}$$

The measurement noise on the slope is taken as a quarter of the level sensor figure applied directly in µrad, which is a shorthand for the slope noise a two-spot sensor produces rather than a quantity derived from the spot geometry. There is no feedback of any kind on tilt. The lead is computed directly from (4.16) applied to the tilt servo rather than exposed as a separate tuning constant, which at the design point is $400 \cdot 2(0.80)/(2\pi \cdot 45) = 2.26\,\text{mm}$. A real tool implements the two leads by delaying one copy of the same sensor signal by a different amount.

4.5.3 Dose: The One Real Loop

The dose loop is the same algorithm of (4.19) with $K_c = 8 \times 10^{-4}$ and $\tau_I = 6\,\text{ms}$, acting on the dose rate rather than on the accumulated dose so that it can correct a field while the field is still being written. The setpoint is $D_{SP} v/L$, the process variable is the delivered dose rate multiplied by the monitor gain g_m, and the output is the pulse energy command bounded to 0.02–1.10 mJ. The upper bound carries the nominal dose to the top of the scan-speed range, because 800 mm/s at 30 mJ/cm^2 needs 0.88 mJ per pulse by (4.8). The reset time is six controller periods, which is aggressive and appropriate because the plant is the single lag of (4.6) with $\tau_L = 4\,\text{ms}$ and the whole opportunity to act lasts 75 ms. Between fields there is nothing to measure, so the loop is placed in tracking on its own output rather than integrating against a measurement it cannot influence, which is the same mechanism that provides bumpless transfer in Section 1.6.2.

4.5.4 Overlay: A Fit Applied Open Loop

The grid correction of (4.12) is computed once per wafer and then applied without further measurement. When the lens heating model is enabled the correction carries one additional term, as shown in Equation (4.21).

$$u_L = \hat{k}_L \left(T_L - T_{L,align} \right) \tag{4.21}$$

where u_L is in ppm and $T_{L,align}$ is the heating state recorded at the moment the marks were measured. Equation (4.21) contains no measurement of the disturbance. It is a model driven by an exposure history the tool already knows, which is the only correct answer available when a disturbance finishes before it could be measured. The float glass line of Volume 5, Chapter 5 meets the same problem in a quality that is fixed before it can be read.

4.6 Results

Table 4.4 lists the eleven scenarios in teaching order, with a short description in place of each name the app shows. A scenario pins the opening mode, the tuning constants, the disturbance settings and whatever is scheduled to happen, and puts its expected observation in front of the reader before the run starts, as Section 1.6.4 describes. The values that follow come off the application's own scorecard and per-wafer export, read at the instant the first wafer completes so that every one of them rests on exactly one hundred fields. That convention matters for the scattered quantities, because the per-wafer

Table 4.4 Scenario Sequence and the Mechanism Each Shows

#	Scenario	What it shows
1	Stage by hand, parked	no flat wafer, manual travel against DOF
2	Servos alone, closed	second-order step response, group delay
3	One hundred fields	the whole sequence and three result maps
4	Trailing sensor only	delay in series with lag, tuning against architecture
5	700 mm/s, preview unchanged	the matched distance belongs to the speed
6	Rougher wafer, 2.2×	spatial frequency against amplitude
7	700 mm/s, 5 % laser	the $1/\sqrt{N}$ dose floor
8	Four alignment marks	fit from a sample, noise into coefficients
9	Lens heating, $k_L = 0.30$ ppm, model pinned off	disturbance finished before measurement
10	Dose monitor +2 % at $t = 8$	a healthy loop making scrap
11	Graded challenge, 480 mm/s	slow servo, short preview, fast trim

counters clear when the next wafer starts, and a run halted at a fixed time estimates a standard deviation from whatever fraction of a wafer has been written by then. Each value is one wafer at the shipped seed unless a table states a twelve-seed mean, and every setting is made before the run starts. The worst-field value is the largest of the hundred per-field values $F_{rms,i}$, not a point value. A setting changed before the run starts re-initializes the case exactly as Reset would (Section 1.6.1), so the dose loop and the laser start at the pulse energy for the speed in force and the application gives one run either way. Figure 4.4 shows the application partway through a wafer of scenario 3.

4.6.1 Manual Operation and the Flat Wafer Assumption

Scenario 1 holds the stage at a fixed height in Manual, and with the tool switched from parked to exposing it writes a wafer under that height. The surface under the slit moves over a range of 1.64 µm across the footprints of the hundred fields, which is 27 times the 60 nm depth of focus. With the stage left at the nominal plane $z = 0$ that the scenario loads, the measured focus RMS is 432.6 nm with a worst field of 955.9 nm. Ninety-nine of the hundred fields fail and the linewidth range across the wafer is 109.3 nm against a 2.0 nm tolerance. Manual mode is not a control strategy here. It is the measurement that establishes the size of the problem, and the manual helper that places the stage on the surface under the slit makes the point that the correct height is correct at exactly one point on one wafer. Parked on the surface at the wafer center, 0.805 µm below the nominal plane, the same wafer measures 684.6 nm, because only 4 % of the surface the slit crosses lies below the center of the clamped wafer.

4.6.2 Architecture Against Tuning

Scenario 4 switches the level sensor out of (4.18) and runs the stage from the trailing sensor alone. At 400 mm/s a trailing distance of 2.0 mm is 5.0 ms of transport delay, and (4.15) adds 4.24 ms of servo lag behind it, on a disturbance whose energy sits at 57.1 Hz. Table 4.5 compares the two architectures on the same wafer, the same servo and the same speed, and Figure 4.3 draws it against the depth of focus.

In the twelve-seed means of Table 4.5, the scenario tuning gives 119.0 nm RMS against a 60 nm depth of focus, a worst field of 251.7 nm and a linewidth range of 7.61 nm on a 2.0 nm specification. Gain helps, and it helps monotonically. Holding the reset time at its shipped 50 ms and raising the trim gain through 0.1, 0.3, 0.6, 1.0, 2.0 and 3.0 gives means of 234.1, 156.4, 119.0, 98.2, 74.2 and 62.8 nm, with no ringing and no reversal anywhere in the range the interface offers. Taking the reset time to its 5 ms minimum as well reaches 41.6 nm and a worst field of 83.8 nm, which is the corner of the permitted range and the best the application allows. The loop does become unstable eventually, and it happens outside the interface, because with the gain bound removed the result on the shipped

Table 4.5 Focus Performance Over One Wafer of One Hundred Fields at 400 mm/s in the Default Configuration. Each Entry Is the Mean of Twelve Seeds, and Worst Is the RMS of the Worst Field. Overlay and Dose Are Inside Specification on Every Row, So the Failure Count Is Attributable to Focus Alone.

Configuration	F_{rms} (nm)	worst (nm)	CD range (nm)	fields failed
Manual, stage at the nominal plane	432.6	955.9	109.29	99.0
Trailing, $K_c = 0.6$, $\tau_I = 50\,\text{ms}$	119.0	251.7	7.61	81.7
Trailing, $K_c = 3.0$, $\tau_I = 50\,\text{ms}$	62.8	134.8	2.26	39.6
Trailing, $K_c = 3.0$, $\tau_I = 5\,\text{ms}$	41.6	83.8	0.96	8.3
Preview at $d^*_{prev} = 1.70\,\text{mm}$	13.7	29.2	0.39	0.0
Preview at $d^{opt}_{prev} = 2.35\,\text{mm}$	11.3	21.0	0.38	0.0
200 Hz servo, noiseless sensor	8.8	19.5	0.43	0.0

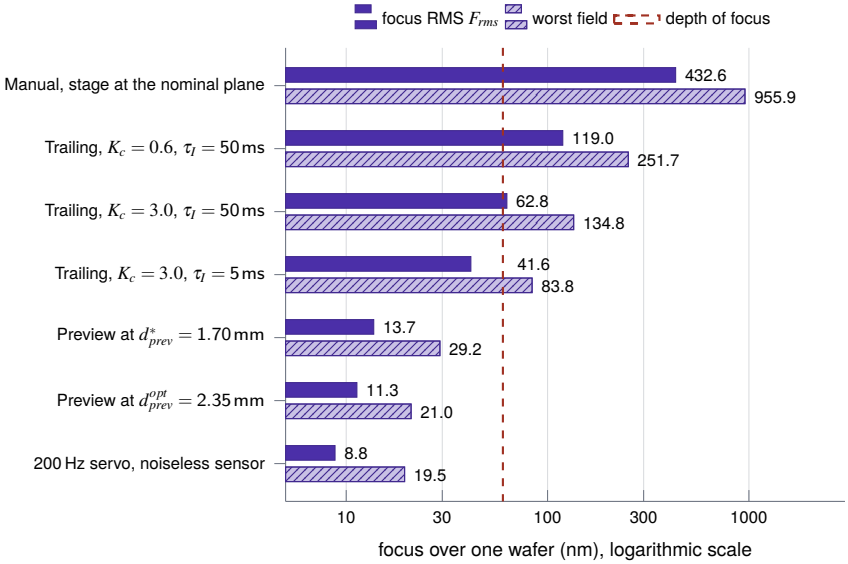


Figure 4.3 Focus RMS (solid bars) and worst field (hatched bars) of the seven configurations of Table 4.5 on a logarithmic scale, with the 60 nm depth of focus as a dashed line. Notice that no tuning of the trailing-sensor loop brings its worst field inside the depth of focus, while both preview rows do, and that the noiseless 200 Hz servo improves on preview only a little.

seed improves further to 47.1 nm at a gain of 8 and then diverges to 1682 nm at 10, while the slider stops at 3.

What gain cannot buy is enough of itself. The corner of the permitted range is 3.0 times worse than preview feedforward on the same servo, the same wafer and the same speed, its worst field of 83.8 nm lies outside the 60 nm depth of focus, and the gain-only result of 62.8 nm is more than four times the preview figure. Shortening the trailing distance helps in the same bounded way, from 119.0 to 109.1 nm at 0.5 mm instead of 2.0 mm, and to 44.6 nm with 18 fields still failing at the gain of 3.0, because the servo lag that remains is larger than the transport delay that was removed. The architecture and not the tuning sets the number that can be reached, and the reason is visible in the arithmetic before any run is made.

Preview feedforward at the estimate of (4.16) gives 13.7 nm RMS and 0.39 nm of linewidth range, and at the swept optimum 11.3 nm and a 21.0 nm worst field. The last row of Table 4.5 establishes how much of that remainder is even addressable. With the height servo at 200 Hz, the tilt servo at

Table 4.6 Throughput Against Focus and Dose. Wafers per Hour Are Computed With the 0.5 s Wafer Exchange the Running Scenarios Use. Dose 3σ Is the Mean of Twelve Seeds at the Default $\sigma_p = 3\,\%$, With the Speed Set Before the Run Starts.

v	wph	N	dose 3σ (%)	F_{rms} at 1.70 mm	F_{rms} at d^{opt}	d^{opt}_{prev} (mm)
250	130	144.0	0.77	9.8	9.8	1.70
400	156	90.0	0.92	13.8	11.3	2.35
480	164	75.0	1.02	17.3	12.4	2.70
550	171	65.5	1.12	20.6	13.4	3.00
700	181	51.4	1.25	28.2	16.1	3.65

150 Hz, the level sensor noise removed and the preview rematched to 1.25 mm, the RMS is 8.8 nm. Most of that residual is the across-slit curvature of (4.3), 7.7 nm on its own, with the tilt term of (4.9) and the slope the trim reads making up the rest, and no bandwidth reaches it. The design configuration is therefore already within 2.5 nm of the floor, which is the answer to a reader who proposes to keep raising f_z.

> ✦ **DID YOU KNOW? Twice the Bandwidth, Almost the Same Focus**
>
> From the simulator: in Table 4.3, with the preview swept to its best value each time, the 60 Hz height servo gives a focus RMS of 11.29 nm and a 120 Hz servo 10.43 nm. Doubling the bandwidth of the stage barely moves the focus, because the curvature across the slit sets a floor that no servo reaches.

4.6.3 Throughput Is a Focus Decision

Scenario 5 raises the scan speed from 400 to 700 mm/s for throughput and changes nothing else. The focus RMS goes from 13.8 to 28.2 nm and the worst field from 29.3 to 63.2 nm, which is outside the depth of focus and fails. Rematching the preview to 3.65 mm returns it to 16.1 nm and 27.6 nm. Table 4.6 carries the whole trade across five speeds.

Going from 400 to 700 mm/s buys 25 wph, an increase of 16 %, and costs 4.8 nm of focus RMS even when the preview is rematched at every step. Leaving the preview alone costs 14.5 nm instead, three times as much, for exactly the same throughput. Nothing on the tool reports that a control parameter moved when the recipe speed changed, which is the point of the scenario.

> **TRY IT IN THE APP 5 · The Preview Distance Belongs to the Scan Speed**
>
> Load the scenario and press Run. The scan speed has been raised to 700 mm/s and the preview left at 1.70 mm. Watch the Focus RMS tile when the wafer unloads: it reads 28.2 nm, against 13.8 nm at 400 mm/s. Set Level sensor preview distance to 3.65 mm and run again: the tile returns to 16.1 nm.

4.6.4 Amplitude Against Spatial Frequency

Scenario 6 multiplies the short-wavelength topography by 2.2 and leaves the clamp bow alone, taking the focus RMS from 13.8 to 21.6 nm. Doubling the same term gives 20.1 nm. Tripling the clamp bow instead, which is a change of 3.0 µm in amplitude against the 0.05 µm added by doubling the topography, gives 23.2 nm with one edge field at 64.5 nm, while the stage travels 4.3 µm during exposure. The servo is not what that wafer costs. With the trim gain at zero the tripled bow measures 13.7 nm against 10.5 nm for the nominal wafer, because the bow arrives at 1.3 Hz and the servo follows it, leaving only its curvature and tilt across the slit. The rest is the trailing sensor reading the bow's slope 2 mm behind the slit as a focus error, largest at the edge where the bow is steepest. The topography arrives at 57.1 Hz and costs about 8 nm whichever way the trim is set (18.7 nm at 2.2× with the trim gain at zero), and raising the bandwidth to 90 Hz with the preview left at 1.70 mm

Table 4.7 Measured Dose 3σ Across One Hundred Fields Against the Statistical Prediction of Equation (4.8). Each Measurement Is the Mean of Twelve Independent Seeds, Each Run Terminated at the Instant the First Wafer Completes, So That Every Estimate Rests on Exactly One Hundred Fields. Speed and σ_p Are Set in Scenario 7 Before Each Run Starts.

σ_p (%)	v (mm/s)	N	measured 3σ (%)	$3\sigma_p/\sqrt{N}$ (%)
3	200	180.0	0.645	0.671
3	400	90.0	0.919	0.949
3	700	51.4	1.250	1.255
5	200	180.0	1.075	1.118
5	400	90.0	1.532	1.581
5	700	51.4	2.083	2.092

recovers only 2.4 nm. With the trim out of the comparison, fifty times the amplitude (3.0 µm against the 0.06 µm that 2.2× adds) costs less than half as much focus. The wafer specification that matters is therefore height change per unit of travel rather than total thickness variation, which is why site flatness is defined over a field and not over a wafer [13].

> **TRY IT IN THE APP** **6 · A Rougher Wafer**
>
> Press Run and read the Focus RMS tile when the wafer unloads: 21.6 nm with the short-wavelength topography at 2.2×, against 13.8 nm for the nominal wafer. Reset, set Chuck clamp bow to 3× and Short-wavelength topography back to 1×, and run again: 23.2 nm, with one edge field at 64.5 nm. Set Focus trim gain to 0 as well, and the bowed wafer gives 13.7 nm.

4.6.5 A Floor No Loop Reaches

Scenario 7 runs at 700 mm/s with a 5 % laser. Table 4.7 compares the measured dose 3σ over one wafer against $3\sigma_p/\sqrt{N}$ from (4.8), averaged over twelve seeds at each condition.

The measured values track the prediction within 3.8 % at every condition in Table 4.7, and the residual is the sampling error of a standard deviation estimated from one hundred fields. The dose loop has no feedforward from the scan speed, so a speed changed during a run reaches the next field with the old pulse energy. Changed from 700 to 400 mm/s after the first wafer, as scenario 7 asks, the first field after the change prints 13 % hot and that wafer reads 4.24 %, which is why the scenario asks for a second wafer after the speed change. Raising the dose loop gain tenfold, to 0.008, moves none of the floor, and the 5 % laser reads 1.49 % at 400 mm/s at either gain on the shipped seed, because by the time a pulse could be corrected it has been delivered to the point and the point has left the slit. The gain range ends at 0.0095, below the clipped oscillation that appears at 0.01 on the 400 mm/s case. The loop is doing its real work elsewhere. When the laser gain drops to 0.93, the loop sees the change on the energy monitor within one field and raises the command, and only the first field written after the drop receives a small part of the change while the loop catches up. Recognizing that a loop has reached a statistical floor is the lesson, and the ways out of it are repetition rate, slit height, laser quality and scan speed rather than tuning [14].

Pause. Try Exercise 4.2 before reading on.

The companion failure is scenario 10. The dose monitor begins reading 2 % high at $t = 8$ s. The loop lowers the pulse energy by the same 2 % and reports the dose exactly on setpoint. Running the whole wafer with the monitor gain at 1.02 moves the mean linewidth from 61.976 to 62.450 nm, a shift of 0.474 nm on every field of every wafer at every seed tried, produced by a fast and correctly tuned loop doing exactly what it was told. In the scenario as loaded the drift arrives as the stage steps to field 10, and the first wafer's mean reads 62.407 nm with no field failed, so nothing on the tool flags it.

Table 4.8 Overlay Residual and Throughput Against the Alignment Sampling Plan. Each Residual Is the Mean of Twelve Seeds, With the Observed Range in Parentheses. The Specification Is 5.0 nm.

n_m	overlay 3σ (nm)	alignment (s)	wph
4	8.84 (4.30–21.76)	1.2	197
8	4.06 (3.31–4.48)	2.4	185
12	3.65 (3.30–4.34)	3.6	174
20	3.22 (2.95–3.45)	6.0	156
32	3.21 (3.05–3.40)	9.6	135
40	3.16 (2.97–3.36)	12.0	124

4.6.6 A Trim Faster Than One Product Unit

The trim reads the physical gap behind the slit, surface slope included, and the preview is a separate feedforward path, so the trim must stay slow enough to correct offsets without chasing the topography. It exists to remove the sensor bias δ of (4.17), which the feedforward path cannot see. Firing a 60 nm level sensor drift at the start of the run with the trim gain at zero gives a focus RMS of 60.8 nm, because the whole wafer is now exposed one full depth of focus away from best focus. Restoring the default gain of 0.30 returns it to 13.6 nm. That is the entire justification for the loop. The coagulant dose of Volume 3, Chapter 6 divides the work the same way, with a fast feedforward from the raw water and the late measurement demoted to a slow trim.

Its reset time then decides whether it helps or hurts, because it is the only element of the focus path with a memory longer than one sample. A field scan at 400 mm/s lasts 75 ms. Setting $\tau_I = 30$ ms makes the trim fast enough to track the surface within a field, and because the trailing sensor sits 2.0 mm behind the slit, driving its gap to the offset inside a field drives the slit itself off best focus by the local surface gradient multiplied by that distance. The measured cost at the design point is a focus RMS of 26.3 nm and a worst field of 51.2 nm, against 13.8 nm and 29.3 nm at the default 800 ms. The same effect decides the graded challenge. With the servo and the preview already repaired, the challenge configuration gives 24.2 nm RMS and a 47.8 nm worst field at $\tau_I = 30$ ms and 11.1 nm and 22.4 nm at 800 ms. Same loop, same gain, a factor of 2.2 in the result, and the rule that produced it is that a feedback trim on a batch process must average over several units rather than act inside one.

Taken from its given state, the challenge scenario runs at 480 mm/s on a 28 Hz servo at 0.45 damping, with the tilt servo at 30 Hz and 0.60, 0.50 mm of preview, a 30 ms trim and 2 nm of level sensor noise. Its first wafer measures 44.0 nm RMS with a worst field of 80.8 nm and 13 fields failed. The challenge is graded over a 45 s run, two wafers at this speed, on the defocus at the center of the slit as logged at 200 Hz, and every sample in the last third of the run (from $t = 29.7$ s, inside the second wafer) must lie within ± 60 nm. The given tuning reaches 131.7 nm and fails. The suggested configuration of 110 Hz at 0.85 damping, the tilt servo at 85 Hz and 0.85, 1.6 mm of preview and an 800 ms trim measures 11.1 nm and 22.4 nm on the first wafer and a worst graded sample of 34.1 nm, inside the band with 25.9 nm to spare. Repairing the servo alone gives 27.9 nm RMS and a worst sample of 81.5 nm, and slowing the trim alone gives 38.5 nm and 116.8 nm, so neither fault can be cleared on its own. The suggested tuning carried through a Reset gives the same run.

4.6.7 Sampling a Spatial Field

Scenario 8 fits the six parameters of (4.12) from four marks. Table 4.8 reports the overlay residual and the throughput against the number of marks in the default configuration, where the lens heating model is on and the sampling effect is therefore the only thing moving. Each entry is the mean over twelve seeds, because a single realization of the mark noise is exactly what a sparse fit amplifies.

Four marks give 8.84 nm against a 5.0 nm specification, and the spread across seeds runs from 4.30 to 21.76 nm. Three parameters fitted from four points leave one degree of freedom, so the mark noise lands almost entirely in the coefficients, and because two of the three coefficients multiply

position the resulting error grows with radius. Eight marks remove most of it at 4.06 nm, and the curve then flattens hard. From twelve marks on, the means of 3.65, 3.22, 3.21 and 3.16 nm differ by less than the spread across seeds, while throughput falls from 174 to 124 wph. The flattening is the non-correctable term A of (4.11) together with the synchronization error σ_{sync}, and no sampling plan reaches below it. Setting A to zero takes twenty marks from 3.22 to 2.06 nm, and setting σ_{sync} to zero as well takes it to 1.04 nm, which leaves only the mark noise and decomposes the floor directly. The optimum sampling plan therefore depends on how much of the real distortion is linear, and the application exposes A as a slider so that the answer can be computed rather than assumed.

4.6.8 A Disturbance That Finishes Before It Is Measured

Scenario 9 lets the lens absorb a fraction of every pulse with $k_L = 0.30$ ppm and $\tau_{up} = 18$ s, so the magnification drifts through the twenty seconds in which the hundred fields are written. The marks were measured before the first field and are not measured again, so no feedback is available at all, and a loop closed on the next wafer would be a wafer too late. The overlay residual is 12.72 nm with the model off and 6.98 nm with the model of (4.21) on, averaged over twelve seeds, and matching the model coefficient $\hat{k}_L$ to the true 0.30 ppm returns 3.22 nm, the residual of a wafer whose lens drift is fully corrected. Nothing measured the disturbance in any case. Lens heating is a first-order rise driven by an exposure history the tool knows exactly, so predicting it is easier than measuring it [15], and the quality of the correction becomes the quality of the model rather than the quality of any tuning.

The same mechanism decides the nominal result, which is why the model is on in the default configuration and scenario 9 is the one place that deliberately pins it off. At twenty marks with the default $k_L = 0.15$ ppm the twelve-seed mean overlay residual is 3.22 nm against the 5.0 nm specification with no field failing. Switching the model off on the same wafers takes the mean to 5.68 nm and fails seven of the hundred fields on overlay alone at the shipped seed (1 to 10 across twelve seeds), while the focus and dose results do not change by a measurable amount. A correction with no measurement anywhere in it is worth 2.46 nm of overlay on a specification of five.

4.7 Technology in Practice

The corrections of Section 4.5 run on two height sensors, an energy monitor, a stage, a laser and the scanner's own controller, drawn together in Figure 4.5.

What Measures It

The feedforward reads the level sensor of Equation (4.17), which the model reduces to a reading of the surface d_{prev} ahead of the slit with a noise of 6 nm and a bias δ.

> **TECHNOLOGY Sensor: An Oblique Beam on the Wafer**
>
> A level sensor sends a light beam onto the wafer at a grazing angle, of the order of 80° from the normal, and images the spot it makes onto a detector of two cells. When the surface lies in the image plane the returned spot falls symmetrically and the difference of the two cells is zero. A surface too high or too low moves the spot, and the sign and size of the difference give the direction and size of the error [18]. The wafer itself can fool the sensor. A wafer under several layers of resist reads at a different height from one under a single layer, though both sit at the same height, an error called process dependency [19]. On a focus trend it is a steady offset, invisible to a feedforward that reads the same sensor.

Process dependency is the kind of error the bias δ stands for, and the stylized 60 nm step of the level sensor drift is the case the trailing-sensor trim exists to remove. The dose loop reads an energy monitor whose gain g_m is the disturbance of scenario 10.

What Moves It

The height and tilt axes are second-order servos of 60 Hz and 45 Hz at 0.80 damping, with a stroke of 60 µm, an acceleration limit of 30 m/s^2 and no structural resonance (Equation (4.4)).



Figure 4.4 Scenario 3 partway through a wafer at $t = 14.1$ s, with field 46 of 100 being exposed at 400 mm/s and the preview distance matched at 1.70 mm, cropped to the Scan view and the first two trend strips. The Scan view is a section through the slit at a vertical scale of 10 nm per pixel, with the level sensor spot ahead of the slit (1), the trailing spot behind it (2) and the ±60 nm depth of focus drawn as a band (3). The stage height command (5) sits at −0.177 µm against a stage at −0.185 µm while the defocus at the slit reads +8.6 nm (4), and the scorecard chips above the view report a focus RMS of 14.2 nm, an overlay 3σ of 3.28 nm and a dose 3σ of 0.82 % over the forty-five fields written so far, with no field failed. Throughput reads zero because no wafer has completed yet.

TECHNOLOGY **Final Element: A Levitated Wafer Stage**

The stage is driven by linear and planar motors on the Lorentz principle. Their permanent magnets give a field of about 0.7 T at the moving coils, whose current density is pushed to several tens of amperes per square millimeter by water cooling [7]. Moving without contact, by magnetic levitation, avoids friction and improves reproducibility, and in vacuum it avoids the wear and lubricants that would pollute the tool. Throughput rises as the moving mass falls, but a lighter stage bends, and its flexible modes then reach into the frequencies the loop must control [20]. Each degree of freedom is held by its own PID loop, with notch filters on the resonances that endanger stability [7]. An acceleration feedforward, the estimated mass times the reference acceleration, leaves the feedback little to correct [20].

The second final element is the laser, which follows its energy command through a lag of 4 ms within 0.02 to 1.10 mJ per pulse (Equation (4.6)). Both PI loops use $a = +1$ and are reverse acting in the convention of Section 1.6.2, because each process variable rises with its output.

Where the Controller Runs

The focus and dose loops run on the scanner's own real-time controller. Control loops in operational technology run with cycle times from milliseconds to minutes, many on real-time operating systems,

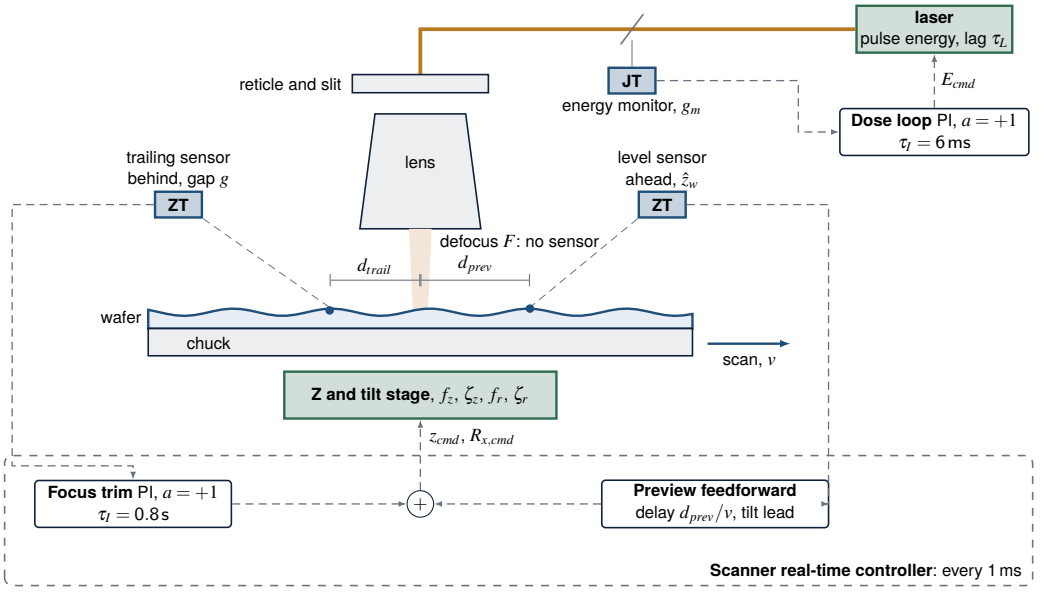


Figure 4.5 The focus and dose loops of the simulated scanner drawn with their hardware. The level sensor reads the surface d_{prev} ahead of the slit, and its reading, delayed by the scan, becomes the height command z_{cmd}. The trailing sensor reads the gap d_{trail} behind the slit and feeds the slow focus trim, a PI loop with $a = +1$. The energy monitor feeds the fast dose loop, which sets the pulse energy of the laser. Both loops run every 1 ms. Notice that no instrument looks at the slit itself, where the defocus F is decided.

and what counts as real time depends on the application and has to be stated for each [21]. The stage that executes the height command is held by one PID loop with notch filters for each degree of freedom [7]. The app executes both of its loops every 1 ms against a physics step of 0.5 ms, and a field scan at 400 mm/s lasts 75 ms. In the controller the structure is three blocks: a delay on the level-sensor signal that turns the preview distance into a lead time, a slow PI block on the trailing sensor, and a fast PI block on the energy monitor. Corrections that need measurements of the printed wafer run outside the scanner, lot by lot, in the run-to-run loops of Section 1.4.3 and Chapter 6.

> **IN PRACTICE** **Commissioning a Stage Loop**
>
> A stage loop is tuned on measurements. The engineer measures the frequency response of the stage, decouples it so that each degree of freedom can be handled alone, and shapes a PID controller with notch filters by hand on that measured response. The feedforward, mass times the reference acceleration, is then tuned by removing the correlation between the measured error and the acceleration profile, and learning control may refine it afterward [20]. The procedure holds while the stage moves as a rigid body over the frequencies that matter. This chapter's height servo, a bandwidth and a damping ratio, is what such a procedure hands to the focus layer above it.

What Is Modernizing It

Three developments bear on this tool. The first is overactuation: more actuators and sensors than the stage has rigid-body degrees of freedom, used to stiffen and damp its flexible modes. It has entered industrial wafer stages over the last five to ten years [7], and on a prototype stage one added actuator and sensor pair raised the bandwidth by 35 % [20]. The second is the reuse of identified models in real time, as digital twins that detect and isolate faults and support predictive maintenance [7]. The third attacks the surface itself. The unflatness of the wafer table is one of the main contributors to

focus variation, and a table that deforms the wafer edge with over 1600 embedded piezoelectric actuators improved from over 200 nm to under 15 nm in simulation [7].

4.8 Limitations

The application is not an imaging simulator and cannot be used to choose a process. Equation (4.10) is a two-term correlation where a real process window is a surface in dose and focus whose shape decides the exposure latitude and the usable depth of focus [5, 6]. The wafer surface is synthetic and carries no process-induced topography, which in production is most of the focus budget. One height axis and one tilt axis stand in for twelve controlled degrees of freedom across two stages, so the synchronization that dominates a real overlay budget appears only as a scalar noise term [1]. The stage has no structural resonance, so the bandwidth in this model is free where in a machine it is bought with mass, stiffness and a great deal of identification work [1]. Alignment is six parameters fitted once per wafer with no per-field correction and no lot-to-lot loop of the kind Chapter 6 closes on linewidth, which makes the sampling trade of Table 4.8 cleaner than the real one. Pulses are not simulated individually, so within-field dose structure from the scan-start transient is absent even though the $1/\sqrt{N}$ floor is reproduced to 3.8 %. Throughput charges the alignment time to the wafer being written, and a dual-stage tool hides it. Exact repeatability is itself a fiction of the simulator, because a fixed step and a seeded generator hand the same trajectory to every machine, which is not what a production tool does. What survives the simplification is the structure of the problem rather than the machine. The matched preview relation, the $1/\sqrt{N}$ dose floor, the spatial frequency argument and the fit-from-a-sample trade are statements about any process that transports material past a fixed working point, and they are what transfers [2, 3]. Section 4.3.7 says of each assumption in turn where it stops holding.

> **OPEN CHALLENGE** **The Error Where the Light Falls**
>
> On a wafer stage the accuracy that matters is at the spot of exposure, while the position sensors measure the edge of the stage. A lighter stage would raise throughput, but it would bend within the control bandwidth, and the error at the exposure spot would no longer follow from the error at the sensors. Inferring that unmeasured error, from more sensors, more actuators and a model, is an open problem, and the lack of control methods for it is the main reason such lightweight stages are not yet feasible [20]. Progress would be a stage whose error at the exposure spot is estimated and held as well as its sensor reading is today.

4.9 Discussion

The coupling this application makes visible has a commercial name. A scanner is the most expensive tool in a fab and its output is counted in wafers per hour, so the scan speed is under continuous pressure from the side of the organization that measures throughput. Table 4.6 shows what that pressure buys and what it spends. Between 400 and 700 mm/s the tool gains 25 wph and spends 4.8 nm of focus RMS with the preview rematched and 14.5 nm without, while the dose 3σ of Table 4.7 rises from 0.92 to 1.25 % at a 3 % laser and from 1.53 to 2.08 % at a 5 % one. None of those three costs appears on a throughput report, and two of the three are recoverable only by changing a control parameter that the throughput decision never mentioned. The alignment sampling plan of Table 4.8 is the same trade written the other way around, where 10.8 s of extra measurement buys 5.7 nm of overlay and costs 73 wph.

> **OPEN CHALLENGE** **Wafers That Will Not Lie Flat**
>
> Deposition, thermal processing and polishing leave stress that warps a wafer, and memory chips of 150 to 300 exposed layers already bring up to 0.5 mm of warpage. How a warped wafer settles on its table depends on friction that changes with wear and contamination, so the result does not repeat and leaves overlay errors [7]. Progress would be a clamped surface that repeats from wafer to wafer.

The model of Section 4.3 supports control methods beyond the loops implemented here. The preview command of (4.18) is the simplest member of a family in which the whole measured surface ahead of the slit, rather than one point of it, is the input to an optimization over the stage dynamics and its acceleration limit [4]. A model predictive formulation on the same eight states, of the kind Section 1.6.3 describes, would enforce the stroke and acceleration constraints of Equation (4.4) explicitly and tell the reader directly what the constraint costs. The wafer is written a hundred times per lot with the same stage and a similar surface, which is the setting iterative learning was built for, as in the powder bed of Volume 8, Chapter 10, and the reason feedforward on these machines is identified from repeated trials rather than tuned [1]. The alignment fit of (4.12) generalizes to a regularized estimator in which the mark noise and the expected distortion enter as covariances, which is what turns Table 4.8 from a sweep into a design.

4.10 Conclusion

This chapter studied a step-and-scan exposure sequence as a laboratory for control at a point that cannot be instrumented. The simulator integrates eight states and writes one hundred fields on a 300 mm wafer in about twenty seconds. Focus is commanded from a level sensor reading d_{prev} ahead of the slit, and the matched distance $v\,2\zeta_z/\omega_z$ is 1.70 mm at 400 mm/s on a 60 Hz servo at 0.8 damping. A wafer exposed from a trailing sensor instead measures 119.0 nm of focus RMS and 7.61 nm of linewidth range, and raising the trim gain improves that monotonically to 62.8 nm at the top of its slider and 41.6 nm at the corner of the permitted range, while the same servo run from the sensor ahead measures 13.7 nm and 0.39 nm against a floor of 8.8 nm that the across-slit curvature sets. A sweep places the true optimum between 0.50 and 0.74 mm beyond the group-delay estimate in all ten configurations tested, while the estimate tracks the scaling with speed at a slope of 1.02 and with bandwidth at 0.92. The excess is the lead the trailing-sensor trim takes back, $K_{c,f}d_{trail} = 0.60$ mm, and with the trim gain at zero the optimum falls to within 0.16 mm of the estimate. Raising the speed to 700 mm/s without rematching takes the focus RMS to 28.2 nm and the dose 3σ from 0.92 to 1.25 % because a point receives 51.4 pulses instead of 90. Slowing the trailing-sensor trim from a reset time of 30 to 800 ms takes the design point from 26.3 to 13.8 nm, because a feedback trim tuned faster than one product unit converts a surface gradient into a focus error. A fuller model would add a structural resonance to the stage, a per-field overlay correction, process-induced topography from a measured wafer, and a second stage so that alignment time can be overlapped. The scanner shows that feedback is a choice wherever the disturbance can be seen before it arrives.

KEY POINTS

- No sensor can sit where the light falls, so the stage is commanded from a level sensor that reads the surface ahead of the slit.
- The matched preview distance, $v\,2\zeta_z/\omega_z$, is 1.70 mm at 400 mm/s and moves whenever the scan speed is raised for throughput.
- Fed from a trailing sensor, the same servo measures 119.0 nm of focus RMS and 41.6 nm at the corner of the permitted range, against 13.7 nm with preview.
- The swept optimum sits 0.50 to 0.74 mm beyond the estimate, the lead the trailing-sensor trim takes back.
- A feedback trim tuned faster than one product unit converts a surface gradient into a focus error.

Where else this applies. Preview from a sensor ahead of the working point in Volume 7, Chapter 11 and Volume 20, Chapter 12, and a late measurement demoted to a slow trim in Volume 3, Chapter 6.

References

[1] H. Butler, “Position control in lithographic equipment [applications of control],” *IEEE Control Systems*, vol. 31, no. 5, pp. 28–47, 2011.

[2] M. Tomizuka and D. H. Fung, "Design of digital feedforward/preview controllers for processes with predetermined feedback controllers," *Journal of Dynamic Systems, Measurement, and Control*, vol. 102, no. 4, pp. 218–225, 1980.

[3] A. Hać, "Optimal linear preview control of active vehicle suspension," *Vehicle System Dynamics*, vol. 21, no. 1, pp. 167–195, 1992.

[4] M. Tomizuka and D. E. Rosenthal, "On the optimal digital state vector feedback controller with integral and preview actions," *Journal of Dynamic Systems, Measurement, and Control*, vol. 101, no. 2, pp. 172–178, 1979.

[5] C. Mack, *Fundamental Principles of Optical Lithography*. Chichester, UK: John Wiley & Sons, 2007.

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

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

[8] J. H. Bruning, "Optical lithography: 40 years and holding," in *Optical Microlithography XX*, ser. Proceedings of SPIE, vol. 6520. SPIE, 2007, p. 652004.

[9] K. Suwa, K. Ushida, and B. J. Lin, "The optical stepper with a high numerical aperture i-line lens and a field-by-field leveling system," in *Optical/Laser Microlithography*, ser. Proceedings of SPIE, vol. 0922. SPIE, 1988, pp. 270–276.

[10] M. A. van den Brink, B. A. Katz, and S. Wittekoek, "New 0.54 aperture i-line wafer stepper with field-by-field leveling combined with global alignment," in *Optical/Laser Microlithography IV*, ser. Proceedings of SPIE, vol. 1463. SPIE, 1991, pp. 709–724.

[11] K. Suzuki, S. Wakamoto, and K. Nishi, "KrF step-and-scan exposure system using higher-NA projection lens," in *Optical Microlithography IX*, ser. Proceedings of SPIE, vol. 2726. SPIE, 1996, pp. 767–779.

[12] P. K. Govil, J. G. Tsacoyeanes, R. P. Eron, and D. Walters, "Contributors to focal plane nonuniformity and their impact on linewidth control in DUV step-and-scan system," in *Optical Microlithography XI*, ser. Proceedings of SPIE, vol. 3334. SPIE, 1998, p. 92.

[13] L. Liauw, A. Muray, and M. Chen, "Impact of wafer flatness on submicron optical lithography," in *Optical Microlithography VI*, ser. Proceedings of SPIE, vol. 0772. SPIE, 1987, p. 232.

[14] D. H. Tracy, F. Y. Wu, and B. J. Lin, "Exposure dose control techniques for excimer laser lithography," in *Optical/Laser Microlithography*, ser. Proceedings of SPIE, vol. 0922. SPIE, 1988, pp. 437–443.

[15] Y. Cui, "Fine-tune lens-heating-induced focus drift with different process and illumination settings," in *Optical Microlithography XIV*, ser. Proceedings of SPIE, vol. 4346. SPIE, 2001, p. 1369.

[16] C. Wagner and N. Harned, "Lithography gets extreme," *Nature Photonics*, vol. 4, no. 1, pp. 24–26, 2010.

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

[18] J. E. van der Werf, M. A. van den Brink, H. F. D. Linders, and J. M. M. Beltman, "Imaging apparatus having a focus-error and/or tilt detection device," U.S. Patent 5,191,200, filed 16 December 1991, 1993.

[19] P. A. A. Teunissen, R. M. G. J. Queens, P. A. M. Dekkers-Rog, and A. C. F. A. van Well, "Level sensor for lithographic apparatus," U.S. Patent 7,265,364, filed 10 June 2004, 2007.

[20] T. Oomen, "Advanced motion control for precision mechatronics: Control, identification, and learning of complex systems," *IEEJ Journal of Industry Applications*, vol. 7, no. 2, pp. 127–140, 2018.

[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.

4.11 Exercises

4.1 ⌕ Concept check · 2 min

Scenario 5 raises the scan speed from 400 to 700 mm/s and leaves the level sensor preview at 1.70 mm. The focus RMS rises from 13.8 to 28.2 nm, although the servo and the wafer are unchanged. Which explanation is right?

(a) The level sensor noise grows with scan speed, so the command carries more noise at the higher speed

(b) The stage lags its command by a fixed time, which now spans more wafer travel than the preview supplies

(c) The pulse count falls with speed, and the larger dose scatter is converted into a focus error

(d) The servo bandwidth falls when the stage moves faster, so it no longer follows the clamp bow

4.2 Concept check · 2 min

Scenario 7 runs a laser with 5 % pulse-to-pulse energy scatter. At 400 mm/s the Dose 3σ tile reads 1.49 % for a wafer with the dose loop gain at its default of 0.0008 and again with the gain at 0.008. Why does ten times the gain change nothing?

(a) The pulse energy sits on its upper limit, so the extra gain has no room to act

(b) Between fields the loop is held in tracking, so it never sees the pulse scatter

(c) Each point averages the pulses it receives, and each pulse lands before it can be corrected

(d) The energy monitor reads high, and no gain can remove a bias in the measurement

4.3 Concept check · 2 min

At 400 mm/s on the 60 Hz servo at 0.80 damping, a sweep finds the best preview at 2.35 mm, 0.65 mm beyond the group-delay estimate of 1.70 mm. With the focus trim gain at zero the best preview is 1.80 mm. What produces most of the 0.65 mm?

(a) The servo phase near 57.1 Hz departs from its low-frequency group delay and calls for extra lead

(b) The tilt servo is slower than the height servo, so the height command must lead to cover the tilt lag

(c) The stage filters the level sensor noise, and the filter adds a fixed lag that the preview must cover

(d) The trim reads the surface slope behind the slit and shortens the effective lead by about $K_{c,f} d_{trail}$

4.4 Theory and calculation · 5 min

The height servo of the design point has $f_z = 60\,\text{Hz}$ and $\zeta_z = 0.80$, and the tilt servo has $f_r = 45\,\text{Hz}$ and $\zeta_r = 0.80$. The Level sensor preview distance slider moves in steps of 0.05 mm.

(a) Find the low-frequency group delay of the height servo. *Checkpoint:* 4.244 ms.

(b) Find the matched preview distance at 400 and at 700 mm/s, and the slider setting nearest each. *Checkpoint:* 1.698 mm, set as 1.70 mm, and 2.971 mm, set as 2.95 mm.

(c) Find the lead the app computes for the tilt servo at 400 mm/s. *Checkpoint:* 2.26 mm.

4.5 Theory and calculation · 6 min

The laser fires at $f_{rep} = 6000\,\text{Hz}$ through a slit $h_s = 6\,\text{mm}$ tall along the scan, and the field is $W = 22\,\text{mm}$ wide.

(a) Find the number of pulses a point receives at 400 and at 700 mm/s. *Checkpoint:* 90 and 51.4.

(b) For a pulse-to-pulse scatter $\sigma_p = 5\,\%$, find the dose floor $3\sigma_p/\sqrt{N}$ at both speeds and the ratio of the second to the first. *Checkpoint:* 1.58 % and 2.09 %, ratio 1.32.

(c) Find the pulse energy that delivers 30 mJ/cm^2 at 400 and at 800 mm/s, and compare the second with the 1.10 mJ limit of the laser. *Checkpoint:* 0.44 mJ and 0.88 mJ.

4.6 Theory and calculation · 6 min

Linewidth follows Equation (4.10) with $CD_0 = 62\,\text{nm}$, $s_E = -24\,\text{nm}$ and $c_F = 120\,\text{nm}/\mu\text{m}^2$. A field fails if F_{rms} exceeds the 60 nm depth of focus or $|CD - CD_0|$ exceeds 2.0 nm.

(a) In scenario 10 the dose monitor reads 2 % high and the loop holds the measured dose on setpoint. Find the relative dose error delivered and the linewidth shift it causes. *Checkpoint:* −1.96 % and +0.475 nm.

(b) Find the linewidth loss of a field whose F_{rms} equals the depth of focus, and the F_{rms} at which focus alone would use the whole 2.0 nm tolerance. *Checkpoint:* 0.432 nm and 129 nm.

(c) The stage stays tilted 5 µrad from the surface across a whole field. Find the contribution of the tilt term of Equation (4.9) to F_{rms}. *Checkpoint:* 31.8 nm.

4.7 App challenge · 8 min

Load scenario 4, Feedback only, and why it is not enough, with the default seed, and set the playback speed to 1 s/s. Each run below is one wafer. Run until the Throughput tile first shows a value, at $t = 23.1\,\text{s}$, pause between 23.5 and 29 s, and read the Focus RMS tile within 0.1 nm. Make each change before Run, or press Reset after making it.

(a) Run as loaded: Focus is commanded from The sensor behind, Focus trim gain 0.60 and Focus trim integral time 0.050 s. *Checkpoint:* 119.0 nm.

(b) Set Focus trim gain to 3.00, the top of its slider, and run. Then also set Focus trim integral time to 0.005 s, its minimum, and run again. *Checkpoint:* 62.9 nm, then 41.7 nm.

(c) Set Focus is commanded from to The sensor ahead, the trim gain to 0.30 and the trim integral time to 0.800 s, and run. Find the ratio of the second reading of part (b) to this one, and say what it shows. *Checkpoint:* 13.8 nm, ratio 3.0.

4.8 App challenge · 8 min

Load scenario 5, The preview distance belongs to the scan speed, with the default seed, and set the playback speed to 1 s/s. Each run below is one wafer. Run until the Throughput tile first shows a value, at $t = 19.9\,\text{s}$, pause between 20 and 26 s, and read the Focus RMS tile within 0.1 nm. Make each change before Run, or press Reset after making it.

(a) Run as loaded, at 700 mm/s with the preview at 1.70 mm. Read Focus RMS and Throughput. *Checkpoint:* 28.2 nm and 181 wph.

(b) Set Level sensor preview distance to 2.95 mm, the slider step nearest the matched distance $v\tau_g$, and run. Then set it to 3.65 mm and run again. *Checkpoint:* 18.1 nm, then 16.1 nm.

(c) Set the preview back to 2.95 mm and Focus trim gain to 0, and run. Explain why the matched distance now does as well as 3.65 mm did with the trim on. *Checkpoint:* 16.1 nm.

4.9 App challenge · 9 min

Load the graded challenge, Challenge · 480 mm/s on a slow, soft stage, with the default seed and playback speed 1 s/s. It scans at 480 mm/s on a 28 Hz height servo at 0.45 damping with a 0.50 mm preview and a Focus trim integral time of 0.030 s.

(a) Find the group delay and the matched preview at 480 mm/s for the given servo and for a repaired one at 110 Hz and 0.85 damping. *Checkpoint:* 5.12 ms and 2.46 mm, then 2.46 ms and 1.18 mm.

(b) Before Run, set Z servo bandwidth 110 Hz, Z servo damping 0.85, Tilt servo bandwidth 85 Hz, Tilt servo damping 0.85 (in Advanced) and Level sensor preview distance 1.60 mm. Run, pause between 22 and 28 s and read the Focus RMS tile, which then holds the first wafer. Run on past 45 s and read the challenge checklist. *Checkpoint:* 24.2 nm within 0.1 nm. Not met, worst defocus 69.3 nm.

(c) Set Focus trim integral time to 0.800 s, press Reset and repeat. Pause between 45 and 50 s, export the CSV file, and find the largest magnitude of column `defoc_nm` in the rows from 29.7 to 45 s (the time column is in seconds). *Checkpoint:* 11.1 nm. Met, with 34.1 nm within 0.1 nm.

4.10 Application area · 4 min

Equation (4.14) charges the alignment time to the wafer being written. A dual-stage tool measures the marks of the next wafer while the current wafer is exposed.

(a) With 20 marks at 0.30 s each, a 90 ms step and settle, 400 mm/s and the 0.5 s exchange, find the throughput from Equation (4.14), and again with the alignment term removed as a dual-stage tool would remove it. *Checkpoint:* 156.5 wph and 211.8 wph.

(b) On this tool, 40 marks cost 73 wph against 4 marks (Table 4.8). What would they cost on the dual-stage tool, and what then limits the number of marks? *Checkpoint:* Nothing, while 12.0 s of alignment fits inside the 16.5 s of stepping and scanning.

4.11 Application area · 3 min

A wafer supplier offers two upgrades at the same price: halve the 1.5 µm clamp bow, or halve the 0.05 µm short-wavelength topography of 7 mm correlation length. The scanner runs at 400 mm/s with preview feedforward.

(a) Find the frequency at which each term reaches the servo, taking the bow as one wavelength across the 300 mm wafer, and the product of amplitude and frequency squared for each. *Checkpoint:* 1.33 Hz and 57.1 Hz, then 2.67 and 163 µm Hz^2.

(b) Which upgrade buys focus, and what kind of flatness specification should the purchase order name?

4.12 Review the agent · 5 min

An AI assistant is asked for the preview distance after a recipe change from 400 to 700 mm/s on the design servo, and for a new 120 Hz stage at 400 mm/s. It writes: “The swept optimum at 400 mm/s is 2.35 mm, which is 1.38 times the group-delay estimate of 1.70 mm. The formula therefore underestimates the lead by 38 %. Apply the same factor everywhere: use $2.97 \times 1.38 = 4.10$ mm at 700 mm/s and $0.85 \times 1.38 = 1.17$ mm, set as 1.20 mm, on the 120 Hz stage.” Find the flaw, give better settings, and say how you would check them in the app.