INTERACTIVE CONTROL ENGINEERING
Principles and Applications

Volume 14 Control in Semiconductor Manufacturing

CHAPTER 12

Summary and Outlook

Pages 290–303

Edited by John D. Hedengren

How to cite this chapter. "Summary and Outlook," in *Control in Semiconductor Manufacturing*, John D. Hedengren, Ed., vol. 14 of *Interactive Control Engineering: Principles and Applications*. Advanced Process Solutions, LLC, 2026, ch. 12, pp. 290–303.

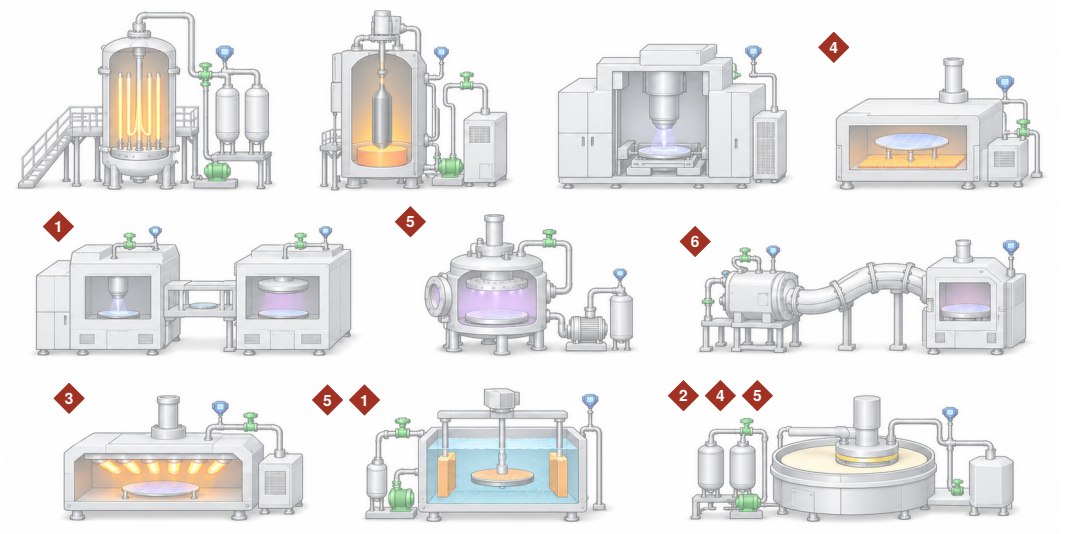
The page numbers in this file are those of the printed volume.

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

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

CHAPTER 12

Summary and Outlook



Open gaps 1 queues that move 2 metrology unsure 3 unknown film stack 4 sparse maps 5 maintenance events 6 reading validity

Notice that the polisher carries three gaps, and that no gap marks the silicon reactor, the puller or the scanner. Illustration, not to scale.

IN PLAIN TERMS **What Ten Wafer Tools Taught**

The main idea. Each chapter of this book ran one tool of a chip factory in a simulator. On almost none of them could the controller see the wafer while the work was being done. A sensor read the heater or the gas, and the wafer was measured later, often after more wafers had gone through.

The challenge. A loop can hold its reading perfectly while the wafer drifts away. A late reading also limits how hard a loop may correct.

Why it matters. The tools that did best saw the cause before it reached the wafer, ended each step on a sign from the wafer itself, and checked their readings against the finished product.

Ten chapters controlled ten tools of a wafer fab, and on almost none of them could a controller see the product while it was being made. A pyrometer read the rod surface while the limit sat at its center, a level sensor read the wafer ahead of the slit, and a microscope read the etched line two lots late. Table 1.1 of Chapter 1 listed, for each tool, what is measured in the tool, what is measured on the product and when, and how the next wafer or lot is corrected. This chapter adds what each chapter measured when the correction worked and when it failed, draws a short guide for choosing a method, looks at where practice is heading, reports how the volume itself was made with AI agents and what they got wrong, and lists the problems that remain open.

12.1 What the Chapters Showed

Tables 12.1 and 12.2 compare the chapters. The second column condenses Table 1.1: what the tool measures and when the product is measured. The other columns name a correction that worked, one

Table 12.1 What Each Tool Measured, and How the Next Wafer Was Corrected, Chapters 2 to 6

Chapter	Measured, and when	What worked	What failed	Measured result
2, polysilicon rods	Surface pyrometer, supply volts and amperes. The center limit is never measured.	A gain schedule on the diameter with an override on the center inferred from the supply.	Fixed gains on a gain that falls 126 times. The schedule alone, which melts the rod.	IAE 760.4 fixed, 43.1 scheduled and 38.8 °C h with both. Melted at 149.3 mm without the override, center peak 1323 °C with it (Table 2.5).
3, crystal puller	Camera diameter, heater power. No instrument reports the pull-rate limit.	Diameter on the pull rate with derivative action, and the heater returning the pull rate to mid-range.	The heater held fixed, so the pull rate drifts toward its limit.	Worst margin 0.221 against 0.358 mm/min, prime crystal 73.46 against 88.00 kg. On a degraded camera, lost at 483.0 min against complete (Table 3.5).
4, wafer scanner	Wafer surface ahead of the slit, before exposure. Overlay on sampled fields after.	Preview feedforward with a lead equal to scan speed times the stage's lag.	A trailing sensor with feedback at any gain. A faster scan without rematching the lead.	Focus RMS 119.0 trailing, 41.6 at the best trailing setting, 13.7 nm with preview. Fields failed 81.7 against 0 (Table 4.5).
5, bake plate	A hundred zone thermocouples every 0.1 s. Linewidth map of every wafer after.	A conductance model as feedforward, a power pulse at each landing, offsets per zone with a neighbor average.	Gains tuned on one zone. Offsets per zone without the average.	All-zone overshoot 18.8 % (Table 5.5). Landing error 4.015 against 0.110 nm. Bowed wafers 3.765 against 0.595 nm (Table 5.4).
6, litho and etch line	Developed line at once. Etched line two lots late.	Feedforward from the developed line. A matched-lot estimator.	Feedback alone at any weight. The newest-reading estimator at a weight of 0.95.	Spread 1.476 with no control, 1.852 with feedback, 0.850 nm with feedforward. Newest reading 8.002 against 1.548 nm matched (Table 6.4).

that failed, and the measured result with the table of the chapter that reports it.

The columns make five points, and Figure 12.1 draws one pair of results from each row.

What can be seen before it arrives is fed forward. The largest single gains of the volume came from corrections made before the product was affected. The scanner's preview cut the focus error to a third of the best that trailing feedback could reach, 13.7 against 41.6 nm. Feedforward from the developed line took the etched spread from 1.852 to 0.850 nm with the feedback tuning untouched, while the feedback alone gave more spread than no control at all. The copper bath dosed by charge voided no wafer on a doubled load that feedback alone met after 97, and the bake plate's landing pulse cut a 4.015 nm linewidth error to 0.110. Each of these disturbances was measured or known in advance, and the feedback loop behind it was left a trim.

The queue and the estimator set the usable gain. A late reading makes a recipe loop a loop with dead time, and Section 1.4.4 derived the boundary. In Chapter 6 the newest-reading estimator at a weight of 0.95 gave a spread of 8.002 nm, and the same weight with the matched-lot estimator gave 1.548. Through a wet clean at a weight of 0.40, a queue of four lots gave 3.812 against 1.497 at two. In Chapter 10 the 42 min age of the analysis set a trim limit of 210, and gains of 300 and 400 voided 16 and 37 wafers where 30 voided none. The limit belongs to the age of the reading and to how the estimator matches it with its lot, not to the plant.

Table 12.2 What Each Tool Measured, and How the Next Wafer Was Corrected, Chapters 7 to 11

Chapter	Measured, and when	What worked	What failed	Measured result
7, plasma etch	Pressure, density, bias, wafer temperature and emission in the step. Seven criteria on the finished wafer.	Ending the main etch on the emission endpoint.	A timed recipe after a wet clean, whether the loops are retuned, closed or both.	Wafers good: 0 of 8 for every change to the loops, 7 of 7 with the endpoint (Table 7.7).
8, ion source	Extraction current every 1 ms. Dose from the scan ledger.	A gate that marks the reading invalid during a spark and holds the command. A local repaint.	A PI loop that chases the spark. A repaint of the whole pass.	Overshoot 14.75 against 0.29 %. Dose error −0.12, +16.58 and +0.21 % for no, whole-pass and local repaint (Table 8.3).
9, spike anneal	Five pyrometers every 10 ms. Sheet resistance after each wafer.	Inverse-model feedforward, a decoupler, a film-stack emissivity table, soak time corrected wafer to wafer.	Feedback alone on the ramp. Five independent loops. One emissivity for every film stack.	Lag 65.2 against 2.9 K. Radial difference 34.8 against 2.9 K. Sheet resistance 26.31 % high on a capped wafer (Table 9.5). Graded lot 0 against 25 of 25 (Table 9.6).
10, plating bath	Charge and temperature always. Additives every 2 h, 42 min old. Thickness each wafer.	Dosing by charge with a sampled trim, a bleed on the modeled breakdown products, a dump.	Feedback alone on the analysis. A trim gain above the limit of 210.	Doubled load: 97 voided wafers on feedback alone, none with the dose by charge. Trim gain 300 and 400: 16 and 37 voids (Table 10.5).
11, polisher	Chamber pressures, pad counters. Removal map one or two lots late, incoming map.	A constrained recipe, incoming feedforward, an age-aware EWMA, a soft reset at a pad change.	Feedback or feedforward alone. Carrying every estimate across a pad change.	In specification 84.2 hybrid, 28.2 feedforward, 24.4 % feedback. After a pad change 2.5 soft against 41.7 nm carried (Table 11.3).

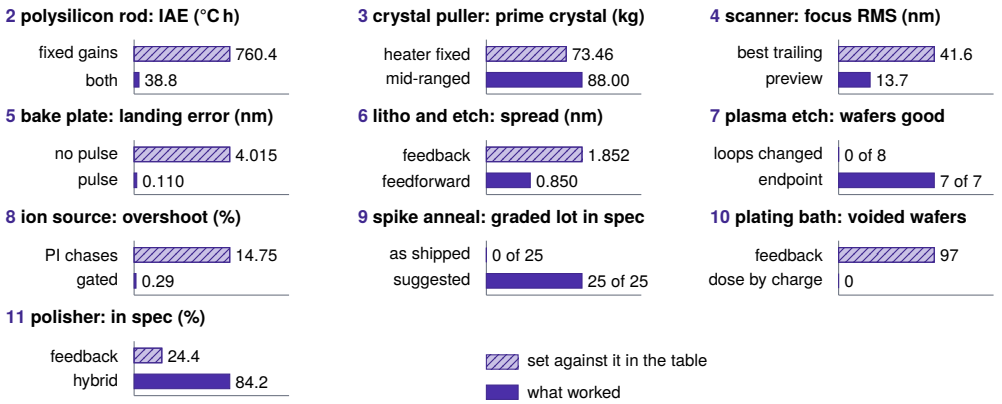


Figure 12.1 One pair of results from each chapter, with the numbers as the "Measured result" column of Tables 12.1 and 12.2 prints them. Each panel has its own scale from zero. A larger value is better in the panels of Chapters 3, 7, 9 and 11. Notice that in every panel the better result came from a change of structure or measurement, not of tuning alone.

A step ends on what the wafer does. After a wet clean the main etch a fresh chamber needs varied by about twelve percent from wafer to wafer. No change to the four chamber loops saved a wafer of the graded cassette, and the emission endpoint saved all seven. The polysilicon batch ends on its center temperature, which no instrument reads. Under the schedule alone the rod melted at 149.3 mm, and the override on the inferred center reached 150 mm with the center at 1323 °C. The ion source repaints the dose a spark interrupted. A local repaint left 0.21 % of dose error where repainting the whole pass left 16.58 %. The step's end, not its loops, decided these results.

Zones on one wafer move together. Section 1.4.5 showed that a loop tuned on one zone sees a smaller gain than the common move. The bake plate's common move had 4.43 times the single-loop gain and overshot 18.8 %. The anneal's five independent loops let the radial difference reach 34.8 K and slip the wafer, and the decoupler held it to 2.9. A hundred run-to-run offsets cut the bowed-wafer signature from 3.765 to 0.745 nm and roughened the map to 0.254, and a neighbor average reached 0.595 and 0.189. Many actuators on one wafer need a model of how they share it.

A reading can be steady and wrong. The anneal's pyrometers held their setpoint on a nitride-capped wafer while the wafer ran cold and its sheet resistance came out 26.31 % high. Its graded lot went from none of 25 in specification to all 25 once a film-stack table set the emissivity. The ion source's current reading stopped describing the source during a spark, and a gate that ignored it cut the overshoot from 14.75 to 0.29 %. The polisher's virtual metrology with an 18 nm bias reported 699.6 nm while the true film was 682.8, and the share in specification fell from 46 to 24 %. After a pad change the polisher that carried every estimate released production on a confident belief about a pad in the bin and landed 41.7 nm off target. A loop that holds its measurement has measured its sensor, and only the product says whether that was the right thing to hold.

12.2 Choosing a Method

The chapters suggest an order of questions for a tool that is new to the reader. Each question names the chapter where its answer decided the result, and Figure 12.2 draws the same questions as a chart.

1. *What does the tool measure, what does the product need, and when does each arrive?* Fill in a row of Table 1.1 before choosing anything. The answer separates loops that can correct this wafer from loops that can only correct a later one.
2. *Can the disturbance be seen or predicted before it reaches the product?* Feed it forward, from the surface ahead of the slit (Chapter 4), the step before (Chapter 6, Chapter 11), the charge passed (Chapter 10) or a known landing (Chapter 5), and leave the feedback a trim.
3. *How many runs pass before the reading returns?* Keep the weight times the gain ratio below the boundary of Equation (1.7), and match each reading to the recipe of its own lot, as Chapter 6 does. Count the age of a sampled analysis the same way (Chapter 10).
4. *Does the step end on an event?* End it on a signal from the wafer, an endpoint, a diameter or an inferred limit, rather than on a clock (Chapter 7, Chapter 2).
5. *Do several actuators act on one wafer?* Tune on the common move as well as the single zone, and model the coupling as a feedforward or a decoupler (Chapter 5, Chapter 9).
6. *Can the fast measurement be wrong while it looks right?* Validate it against the product, calibrate it for the film stack, and mark it invalid when its physical basis has gone (Chapter 9, Chapter 8).
7. *Is the fast actuator also the limited one, or does the plant grow during the run?* Return the actuator to mid-range with a slow one (Chapter 3), and schedule the gain on the variable that moves it (Chapter 2).

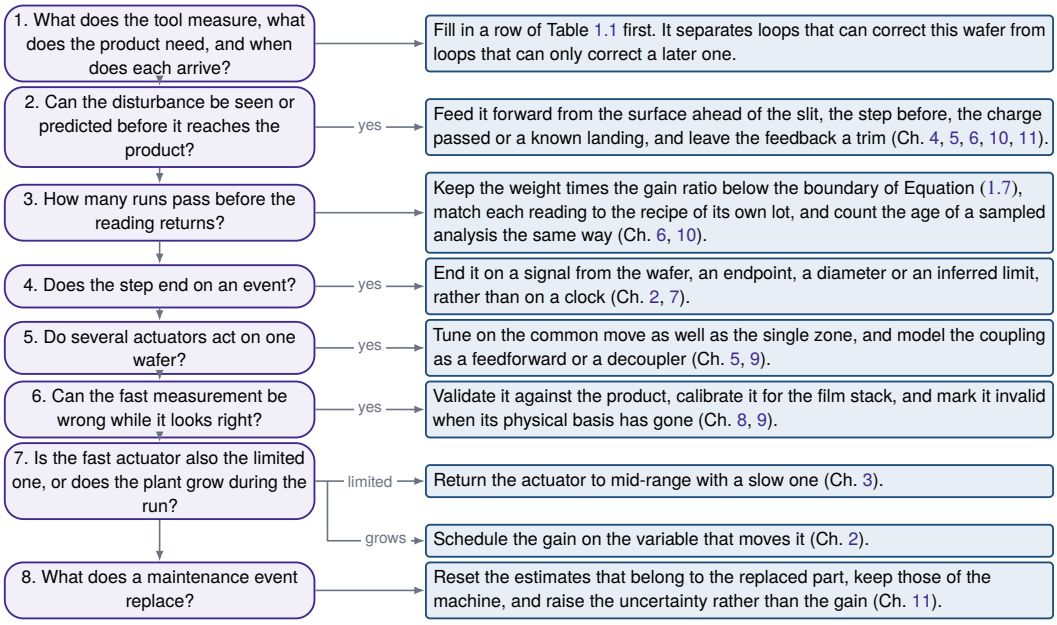


Figure 12.2 The eight questions of this section in order, with the list's answer and the chapters where it decided the result. Notice that the first three questions ask what is measured and when, and are answered before any loop is tuned.

8. *What does a maintenance event replace?* Reset the estimates that belong to the replaced part, keep those of the machine, and raise the uncertainty rather than the gain (Chapter 11).

A predictive controller paid where several actuators had to move together against limits. On the crystal puller it cut the diameter IAE of scenario 11 from 462.0 to 119.1 mm min with less heater travel, where loops retuned by the same rule reached 176.0 (Table 3.7). On the etch chamber it cut the time out of band after a recipe step from 37.08 to 3.54 s per wafer, but the loops retuned once reached 4.14 (Table 7.8). On the anneal's fast ramp it held the radial difference to 10.89 K against 17.27 for the decoupled loops, with the sheet resistance 3.34 % high, while the loops with a slip-aware ramp limit reached 6.29 K and −0.77 % (Table 9.7). On the polysilicon rod it saved time only by spending margin. With the band narrowed it harvested in 28.482 against 29.595 h with 4.4 °C left below the limit, and a retuned override reached 28.675 h with 9.1 (Table 2.6). Before reaching for prediction, try the retune, the supervisor and the feedforward. Volume 9 treats that choice in full.

12.3 Where Practice Is Heading

More data, used with process knowledge. A review of case studies in semiconductor manufacturing describes an industry that uses large data sets and analytics to improve fault detection and to support newer capabilities such as predictive maintenance. It finds that data quality is the most important factor for these solutions, and that subject matter expertise is often required to make them work on line [1]. In the editor's view the chapters of this volume show why. The anneal's pyrometers and the polisher's virtual metrology both produced clean, steady data that was wrong about the wafer, and only a model of the physics said which reading to distrust.

Virtual metrology between measurements. Virtual metrology estimates a product's quality from process data without measuring it, and a systematic review of 199 papers names machine control and

sampling decisions among its applications [2]. On an industrial plasma etch chamber, models updated over a moving window gave useful estimates of etch rate across several chamber maintenance events and many thousands of wafers [3]. That is the seasoning problem of Chapter 7, seen from the data. In the editor's view the estimate will increasingly sit inside the run-to-run loop, and Chapter 11 shows the cost when it does: a bias the loop cannot see becomes a bias in the product.

> ⧖ HISTORY **1991 to 2024: From a Proposal to the Fab's Default**
>
> The timeline of Chapter 1 placed the first run-to-run controller for chip fabrication in 1991. By 2008 a dissertation called run-to-run control the preferred technique for model-based process control, with the recipe adjusted lot by lot or wafer by wafer, and found most fabs combining it with fault detection and classification as advanced process control [4]. A 2018 review of metrology found that for some products more than half of the manufacturing steps can involve measurement [5]. A 2024 review of plasma etching named three-dimensional devices, critical dimensions approaching the atomic scale and new materials among its challenges, and digital twins of etch chambers built from hybrid models among its opportunities [6]. The loop of 1991 has become the fab's default, and the gaps of Section 12.5 are what it still lacks.

Sampling decided by risk. A review of inspection sampling classifies the techniques as static, adaptive and dynamic, and finds significant improvements from static toward dynamic sampling, which it judges more appropriate for modern plants [7]. Chapter 6 adds a control argument. With every lot measured, feedforward gave 1.050 nm of spread against 1.525 without it, so a sampling plan justified against the feedback loop alone undervalues each measurement.

Maintenance predicted, and models of the whole tool. Machine learning has been used to set maintenance decisions by the trade between unexpected breaks and unexploited lifetime, demonstrated on a semiconductor maintenance problem [8]. The pad of Chapter 11, the bath of Chapter 10 and the chamber walls of Chapter 7 are consumables of exactly that kind. Digital twins have been used for years in model-based process control and predictive maintenance, and a framework for them has been proposed with requirements of reuse, interoperability and autonomy over the twin's life [9]. In the editor's view each app of this volume is a small twin of a tool, and the record below shows the discipline a twin needs: it must keep the plant's truth on one side and the controller's knowledge on the other.

Anneals move to milliseconds. Flash lamp annealing heats on a sub-second time scale, away from equilibrium, and has been used in microelectronics mostly after ion implantation, to activate dopants, recrystallize amorphous layers and anneal out defects [10]. On a prototype flash system, a thermometer that compensates for emissivity read the wafer through a layer of flowing water that suppressed the flash, and its peak temperatures tracked the sheet resistance whatever the emissivity of the surface [11]. In the editor's view such a reading would answer the capped wafer of Chapter 9, which ran 31.4 K cold behind a steady trend.

Plating baths are read online. A portable microfluidic workstation reported in 2025 measured the additives of an acid copper bath online in small volumes, with a mean relative error below ten percent, in place of the manual stripping analysis of the laboratory [12]. A model-based approach to industrial electroplating takes the bath's composition from the data of the manufacturing execution and planning systems, because an exact analysis typically cannot be made within a plating company, and it was validated on two electrolytes [13]. In the editor's view both attack the 42 min age that set the trim limit of Chapter 10, one by shortening it and one by modeling across it.

Polishers learn from fab data. A typical fab generates millions of data points a year for process control and fault detection, and models trained on the data of a high-mix, high-volume fab predicted

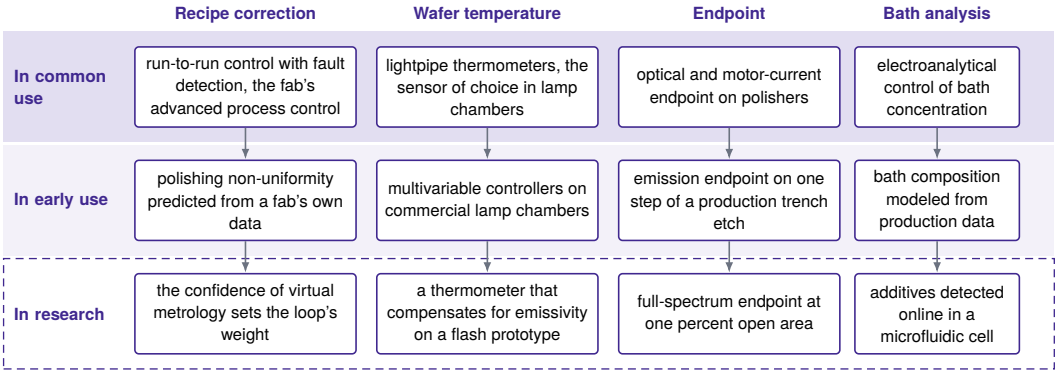


Figure 12.3 Developments in four threads. The top lane holds what the cited source calls standard or common practice, the middle lane what it reports in service at some sites or on plant data, and the bottom lane what it reports from simulation, laboratory work or as a need, in the editor's reading. Sources by thread: recipe correction [4, 14, 16], wafer temperature [11, 17, 18], endpoint [19–21] and bath analysis [12, 13, 22]. Notice that three of the four threads move the measurement closer to the wafer.

the non-uniformity and the spatially resolved removal rate of a polisher [14]. A 2025 review of machine learning in polishing finds that it can make control more adaptive and reduce the demand for metrology, and names data leakage, sparse labels of the target and hybrid modeling among its challenges [15]. In the editor's view the 18 nm bias that the virtual metrology of Chapter 11 carried unseen is the risk such a model must state. Figure 12.3 sets these developments on three lanes.

12.4 Engineering With AI Agents

An AI agent is a language model that works toward a goal in a loop. It plans, calls tools such as a compiler, a simulator or a browser, reads the results and continues [23]. In the editor's view, agents can now carry out much of the work that used to fill the hours of a control project. They derive a model from balances or fit one to plant data. They write the simulator and its tests, code the controller, tune it against stated criteria, script commissioning checks, write the documentation and watch running loops for degradation. The apps and chapters of this series were produced that way under the editor's direction, and the account that follows this section reports what the agents got wrong in this volume and how it was caught.

In the editor's view, this changes where an engineer's value lies. When a model, a controller and a report can each be produced in minutes, producing them is no longer the scarce skill. The work moves to five things.

Quality control. An agent's output arrives as a claim: this model fits, this tuning settles in four minutes, this code implements the equation. The engineer checks claims against the plant or the simulator. A reported number is measured again on a fresh run at the stated seed. The code is read for what it clamps, what it initializes and what it reads, because a display that shows the controller's model in place of the plant looks correct until the two differ. A tuning is tried on a disturbance the agent did not test. The most useful checks are independent of the author, made by a second agent or a person who has not seen the first answer, and they are checks that can fail.

Supervision. The engineer sets the scope of each task and decides what an agent may touch. Reading a historian and proposing a tuning is one level of access. Writing a setpoint, downloading a configuration or changing an interlock is another, and it passes through the plant's management of

change with a named approver. Supervision also fixes the order of work: a baseline before a change, one change at a time and a comparison that shows what moved.

Creativity. An agent answers the question it is given. Choosing the question remains the engineer's task: which loop is worth improving, whether a disturbance calls for feedforward, a cascade or a change to the process, which step test the operation can tolerate, and which comparison would show that a predictive controller earns its place over a retuned PID loop. The experiment that separates two explanations is designed by someone who can imagine both.

Process insight. A model contains what its author put in. The equipment does more. A valve sticks, the dead time of an analyzer grows when its sample flow drops, a thermowell lags, a heat exchanger fouls and an operator intervenes at the same hour each night. The engineer who knows the unit recognizes a result that is too good, asks what the model left out and knows which measurement to distrust first.

Accountability. Responsibility does not transfer to a tool. Functional safety standards assign the activities of the safety lifecycle to persons and organizations of stated competence, and they require an assessment that includes someone independent of the design team [24]. An agent may draft the hazard analysis or the test procedure. A person owns the decision to accept it, to start up and to keep running.

These five are the professional form of the habit introduced in Section 1.7: check before trusting, and know enough to check.

12.4.1 How This Volume Was Made

Each of the ten apps was built and tested by agents. An agent then wrote the paper on it, keeping verification notes that trace every number to a run, and a referee report was written on each paper. Six of the ten reports call themselves self-reviews. Defects went into a log for each app. For this book one agent per chapter reran the numbers its exercises use and checked the app's helpers, completion steps and figures. The editor set the scope of each round and the rules on what each agent could change, and statements that runs contradicted were left for the editor to decide.

The defect logs. Eight of the ten logs give each entry a severity, 129 entries in all, from 10 in the copper log to 26 in the bake plate's. Of these, 51 are wrong behavior or output of an app, 59 are documentation, 16 are cosmetic or display and 3 are other kinds. The scanner and etch logs number their items without a severity and are not added in. Five cases show the errors that matter for a volume about what a tool measures.

A timer that knew the etch rate. The etch app offered an operator aid, "Time the etch for this chamber". It set the main etch from the chamber's true etch rate, which no instrument on the tool measures. With a hidden etch-rate parameter raised by 20 % the aid shortened the etch from 86 to 73 s, as if it knew. A second aid set the actuators from hidden parameters of the chamber. In the anneal app the supervisor, the feedforward and the soak correction read the plant's own constants. Raising a parameter by 20 % moved the lamp commands by 18 to 150 W in one execution with no measurement changed. A truth audit found both by moving one hidden value at a time and asking whether the controller moved while the measurements did not. Both now compute from the declared model, and no shipped run moved. In a volume about endpoints, an aid that knew the rate had quietly removed the reason for an endpoint.

A field exposed one scan too long. The scanner ended a field when its scan coordinate reached the field length. At 400 mm/s the coordinate after 150 steps was 29.999999999999925 mm, so the test failed and every field ran one controller period longer. Each field received 1.33 % more dose, and the nominal mean linewidth was 61.665 nm instead of about 61.98. A dose bias of about 1.3 % had been noted earlier and not investigated. A consistency pass that re-measured every number traced it to the comparison. Every exposing run moved when it was fixed, and the paper was updated the same day.

Helpers pressed twice. Nine chapter records report a check of the app's helpers, and three report pressing a helper repeatedly. The ion source's "Set the bias current for the extraction setpoint" moved the bias by 0.06 A per milliampere of error where the source's sensitivity calls for about 0.0085. One press from the nominal state set 2.80 A and the extraction current went to 57.25 mA, a second set 1.77 A and 3.26 mA, and a third tripped the cathode. The polisher's "Set the polish time" scaled the time by the last wafer's removal, which had been made at the old time. The film came out at 699.92 nm after one press and at 754.33 after four. The bath's dosing aid, pressed six times, set 93.214 mL/h every time. A helper that answers once can still be a controller with the wrong gain when a person presses it again. Both helpers have since been corrected.

Grading against the learner's own number. The anneal challenge asks for every true center peak within 3 °C of 1050 °C. Its scoring judged the peak against the learner's Peak temperature setting instead. With that setting at 1048 °C and a soak of 1.10 s the lot was credited while its worst true peak was 4.23 K below 1050. The polisher's target thickness slider never reached its controller. With it at 720 nm the mean stayed at 699.92, yet three scenarios and the challenge judged against the slider, so lowering it to 685 credited a wear step on a fresh pad. Both were found by trying each rule with a value the learner can set. Both are the volume's lesson in another form: a grade on a setpoint measures the setpoint, not the wafer. Both rules now judge against the fixed target.

The book round. The ten chapter records list 147 rows of numeric verification, each a stated value or group of values set beside its reproduction. Eight report no drift. The polysilicon record found a table cell printed 1258.1 where the app gives 1258.03, and the ion source record found a −9.15 % species shift read at the first logged row, where the beam settles 14.4 % lower. Nine records report walking every completion step against its control's step and range. The book round's defect list holds eleven entries for this volume. Two apps label a direct-acting loop reverse acting: the puller's diameter and mid-ranging loops and the bath's bleed loop. The chapters name the action as Section 1.6.2 does, and the puller's and the bath's apps have since been corrected. Chapter 1 said that four of the etch chamber's loops held their setpoints to four decimal places. The etch agent found that three do, and the orchestrating agent corrected it and added a notation note from the bath chapter. The introduction's writer reported that a bake plate reference lacked an author, and the chapter agent's Crossref query showed the entry complete. Another session then changed the papers and apps of the polysilicon and puller chapters, and both chapters were synced again. Their app changes were text only, and no exercise answer moved. A later repair round changed six more of the volume's apps, and one answer of the etch chapter moved with them.

Every error above was caught by a check that could fail: a hidden value moved, a number traced to its run, a helper pressed twice, a rule tried with a setting the learner can change, a reference queried. None was caught by rereading. Four working rules follow. Give a controller and an aid only what the tool can measure. Press every aid twice. Grade against the product, never against a setting. Re-measure every number after every change.

12.5 Gaps in Semiconductor Process Control

Six gaps belong to this volume's topic, and the gaps common to agent-era practice follow them in Section 12.5.1. In the editor's view none is closed. A sentence on difficulty or progress that carries no

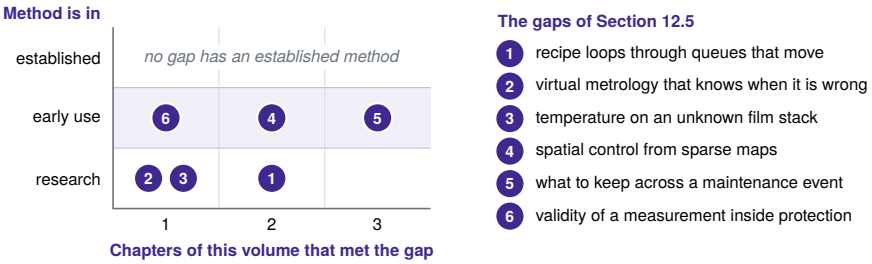


Figure 12.4 The six gaps of this section. Across is the number of chapters that each item names as evidence: two for gap 1 (Chapters 6 and 10), one for gap 2 (11), one for gap 3 (9), two for gap 4 (5 and 11), three for gap 5 (7, 10 and 11) and one for gap 6 (8). Up is the maturity of the method available, the editor's reading of the sources cited in each item. Notice that the gap most chapters met has a method in early use.

citation is the editor's view. Figure 12.4 places the gaps by the chapters that met them and by the maturity of the method available, and the same numbers mark the tools in the picture that opens this chapter.

1. *Recipe loops through queues that move.* The boundary of Chapter 6 assumes a fixed queue and a known gain ratio, and a gain believed at half its value took the spread to 7.712 nm. The age of the bath's analysis set the trim limit of Chapter 10 the same way. In a fab the queue is set by dispatch and changes from lot to lot. A design that states its margin against a queue and a gain that both vary is not routine. A 2008 dissertation reports that wafers wait in a backlog for metrology and are measured by priority, so the delay varies from run to run, mostly from 0 to 5 runs with a mean of 1 or 2 [4]. It is hard because the loop does not see the dispatch that sets its delay. Progress would be a weight chosen against the spread of the delay, with each reading's delay logged beside it.

2. *Virtual metrology that knows when it is wrong.* The polisher's estimate kept its loop on target while the product drifted 18 nm away (Chapter 11). An estimate needs a statement of its own error that a loop can act on, and a rule for when the loop should stop believing it. A 2018 study used the variance of a Gaussian-process prediction as its confidence and adjusted the weight of an EWMA run-to-run controller with it, on a polishing case study [16]. It is hard because a bias need not raise the spread that a model reports about itself.

3. *Temperature on an unknown film stack.* The anneal's pyrometers held a wafer 31.4 K cold without a sign on the trend, and a table or a calibration wafer corrected it only for stacks known in advance (Chapter 9). A 2005 study of a lamp-heated chamber reports that films on the measured side change the emissivity, that such wafers are screened out by hand, and asks for a systematic way to alarm the tool [25], and a lightpipe thermometer cannot give the true temperature unless the effective emissivity is known [26]. It is hard because the wrong emissivity leaves a steady, plausible trend. A measurement of the wafer's own emissivity, robust enough for every layer, would remove a whole class of silent error.

4. *Spatial control from sparse maps.* The bake plate's hundred offsets traded signature for roughness, and the polisher's eight chambers reach 33 radial points only through broad footprints (Chapter 5, Chapter 11). Methods that choose which spatial patterns a sparse map can correct, and leave the rest alone, are still tuned by hand. A 2004 study of a three-zone polisher read 60 points of film thickness, could not hold all 60 at their setpoints with three inputs, and controlled them through a singular value decomposition [27]. It is hard because the patterns that no actuator reaches still show in the map and pull the estimate.

5. *What to keep across a maintenance event.* The polisher's soft reset was right for a pad, the etch chamber's walls reseason after every clean, and the bath is dumped when bleeding cannot reopen its window (Chapter 11, Chapter 7, Chapter 10). A general method that splits an estimator's state by the part it describes and sets the uncertainty after the event is not available. The Open challenge box after this list takes the gap further.

6. *Validity of a measurement inside protection.* The ion source's gate decided in milliseconds whether a reading still described the source (Chapter 8). Each tool now carries its own such rules. A 1996 qualification study found an implanter highly susceptible to arcing, which can damage its hardware [28], and a 2021 system detects an implanter's faults in real time from tool data, because its wafers cannot be reworked [29]. It is hard because the rule must decide before the evidence is complete. A shared way to state, test and certify when a reading is valid, and what the loop does when it is not, would help every chapter of this volume.

OPEN CHALLENGE What Does a Maintenance Event Replace?

Gap 5 is the gap most chapters met. A run-by-run controller of 1995 gave maintenance its own rapid mode, which recovered an epitaxy process within three runs of a disturbance [30]. A 2008 dissertation reports a step in bias whenever a tool or chamber is maintained, and seemingly identical tools that process identical wafers differently with the number of lots since the last maintenance event [4]. In a production etch chamber the first one or two wafers of each lot differed considerably [20]. The event is hard because an estimator sees only the sum of its parts. Progress would be an estimator split by the part each state describes, with the maintenance record setting each part's uncertainty after the event.

12.5.1 Gaps Common to Agent-Era Practice

The final chapter of each volume lists gaps specific to its topic. Seven more are common to every topic in the series. In the editor's view none of them is closed.

1. *Verifying agent-written control code and models before they reach a plant.* Tests show that a controller behaves on the cases tested. For controllers learned from data, guarantees of stability and constraint satisfaction rest on assumptions about the model and its uncertainty, and open problems remain [31]. In the editor's view, code written by an agent has no comparable body of method yet. Practice needs a routine path from an agent's draft to code a plant can trust: independent derivation, tests of what the code reads and clamps, and trials with hardware in the loop.

2. *Trustworthy models from sparse or poor plant data.* Plant records are short, taken in closed loop, noisy and marked by unmeasured disturbances. A fitting routine returns a model from any record. What is missing is a dependable statement, delivered with the model, of where it can be trusted, and a method that says so when the data cannot identify a parameter.

3. *Safety cases and standards for agents in commissioning and operation.* Functional safety standards define a lifecycle of specified requirements, verification and assigned human roles [24]. In the editor's view, how an agent that writes configuration, runs commissioning checks or advises an operator is qualified, limited and audited under them is not yet settled.

4. *Security of agent access to control systems.* An agent with credentials to a control system is a new path into it. Industrial security standards state the requirements a control system must meet at each security level [32]. An agent that reads documents, messages or web pages can also be steered by text placed there. In the editor's view, least privilege, read-only access by default and human approval of every write are a starting point and not a solution.

5. *Benchmarks and reproducibility.* An agent asked twice may deliver two different controllers, and a result reported without its plant, seed and scoring cannot be checked. Control needs shared

benchmark plants with fixed disturbances and published measures, on which agent-built and conventional designs are compared like for like. The apps of this series, with a default seed and an exported record of every run, are a start.

6. *Transfer from simulation to hardware.* A controller developed on a simulator inherits what the simulator omits. The robotics literature documents the gap between simulated and real performance and the methods used to narrow it, such as randomizing the simulated plant and adapting on the real one [33]. Process plants add slow dynamics, rare upsets and little freedom to experiment, so each transfer still needs a staged plan from the simulator to a bench rig to the unit.

7. *Keeping engineers' judgment sharp when agents do the routine work.* Automation removes the routine work through which skill is built and leaves the person the hard cases, an irony described for process operators in 1983 [34]. The same now applies to engineers. An engineer who has never tuned a loop by hand or closed a balance on paper is poorly placed to review an agent that has. Deliberate practice, of the kind the exercises in this book provide, has to be designed into education and into work.

KEY POINTS

- What can be seen before it reaches the wafer is fed forward, and the feedback behind it is left a trim.
- A late reading sets the usable gain of a recipe loop, and so does how the estimator matches each reading with its lot.
- A step that ends on a signal from the wafer absorbs a drift that no loop on the chamber can see.
- Zones on one wafer move together and need a model of how they share it.
- A loop that holds its measurement has measured its sensor. Only the product says whether that was the right thing to hold.

References

[1] J. Moyne and J. Iskandar, "Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing," *Processes*, vol. 5, no. 3, p. 39, 2017.

[2] P.-A. Dreyfus, F. Psarommatis, G. May, and D. Kiritsis, "Virtual metrology as an approach for product quality estimation in Industry 4.0: A systematic review and integrative conceptual framework," *International Journal of Production Research*, vol. 60, no. 2, pp. 742–765, 2022.

[3] S. A. Lynn, J. Ringwood, and N. MacGearailt, "Global and local virtual metrology models for a plasma etch process," *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 1, pp. 94–103, 2012.

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

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

[6] G. S. Oehrlein, S. M. Brandstadter, R. L. Bruce, J. P. Chang, J. C. DeMott, V. M. Donnelly *et al.*, "Future of plasma etching for microelectronics: Challenges and opportunities," *Journal of Vacuum Science & Technology B*, vol. 42, no. 4, p. 041501, 2024.

[7] J. Nduhura-Munga, G. Rodriguez-Verjan, S. Dauzère-Pérès, C. Yugma, P. Vialletelle, and J. Pinaton, "A literature review on sampling techniques in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 26, no. 2, pp. 188–195, 2013.

[8] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 812–820, 2015.

[9] J. Moyne, Y. Qamsane, E. C. Balta, I. Kovalenko, J. Faris, K. Barton, and D. M. Tilbury, "A requirements driven digital twin framework: Specification and opportunities," *IEEE Access*, vol. 8, pp. 107 781–107 801, 2020.

[10] L. Rebohle, S. Prucnal, Y. Berencén, V. Begeza, and S. Zhou, "A snapshot review on flash lamp annealing of semiconductor materials," *MRS Advances*, vol. 7, no. 36, pp. 1301–1309, 2022.

[11] Y. Yamada, T. Aoyama, H. Chino, K. Hiraka, J. Ishii, S. Kadoya, S. Kato, H. Kiyama, H. Kondo, T. Kuroiwa, K. Matsuo, T. Owada, T. Shimizu, and T. Yokomori, "In situ Si wafer surface temperature measurement during flash lamp annealing," *Japanese Journal of Applied Physics*, vol. 49, no. 4S, p. 04DA20, 2010.

[12] Y. Zhao, J.-X. Zeng, J.-Q. Yang, T. Song, R. Hu, J.-J. Su, B. Zhang, F.-Z. Yang, D. Zhan, and L. Han, "On-line detection of additive concentrations in acidic copper plating solution for metal interconnection by an electrochemical microfluidic workstation," *Industrial Chemistry & Materials*, vol. 3, no. 5, pp. 607–617, 2025.

[13] A. Leiden, S. Kölle, S. Thiede, K. Schmid, M. Metzner, and C. Herrmann, "Model-based analysis, control and dosing of electroplating electrolytes," *The International Journal of Advanced Manufacturing Technology*, vol. 111, no. 5-6, pp. 1751–1766, 2020.

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

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

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

[17] C. W. Meyer, D. P. DeWitt, K. G. Kreider, F. J. Lovas, and B. K. Tsai, "ITS-90 calibration of radiation thermometers for RTP using wire/thin-film thermocouples on a wafer," *AIP Conference Proceedings*, vol. 550, pp. 254–258, 2001.

[18] A. Emami-Naeini, J. L. Ebert, D. de Roover, R. L. Kosut, M. Dettori, L. M. L. Porter, and S. Ghosal, "Modeling and control of distributed thermal systems," *IEEE Transactions on Control Systems Technology*, vol. 11, no. 5, pp. 668–683, 2003.

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

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

[21] S. Lee, H. Choi, J. Kim, and H. Chae, "Spectral clustering algorithm for real-time endpoint detection of silicon nitride plasma etching," *Plasma Processes and Polymers*, vol. 20, no. 6, p. e2200238, 2023.

[22] R. Carpio and A. Jaworski, "Review—management of copper damascene plating," *Journal of The Electrochemical Society*, vol. 166, no. 1, pp. D3072–D3096, 2019.

[23] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[24] International Electrotechnical Commission, "Functional safety – Safety instrumented systems for the process industry sector – Part 1: Framework, definitions, system, hardware and application programming requirements," IEC 61511-1:2016, Edition 2.0, 2016.

[25] W. Cho, "Temperature control and modeling of rapid thermal processing," Ph.D. dissertation, The University of Texas at Austin, Austin, TX, 2005. [Online]. Available: https://hdl.handle.net/2152/1527

[26] Y. Qu, "Silicon wafer surface temperature measurement using light-pipe radiation thermometers in rapid thermal processing systems," Ph.D. dissertation, The University of Texas at Austin, Austin, TX, 2006. [Online]. Available: https://hdl.handle.net/2152/2794

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

[28] S. K. Joung, "Qualification of a medium current ion implantation system in a semiconductor production environment," S.M. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996. [Online]. Available: https://hdl.handle.net/1721.1/40220

[29] S. C. Horng and Y. L. Lin, "Fuzzy tree classification system for fault diagnosis on ion implanter," *Journal of Physics: Conference Series*, vol. 2020, no. 1, p. 012017, 2021.

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

[31] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[32] International Electrotechnical Commission, "Industrial communication networks – Network and system security – Part 3-3: System security requirements and security levels," IEC 62443-3-3:2013, Edition 1.0, 2013.

[33] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: A survey," in *Proc. 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 737–744.

[34] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

12.6 Exercises

12.1 Discussion · 8 min

An agent proposes to raise the etch loop's EWMA weight on the line of Chapter 6 from 0.30 to 0.60, because "the line drifts and the loop is slow". Its evidence is a simulation with no metrology queue in which the corrected line returns to target twice as fast. Using Chapter 6 and Section 1.4.4, explain why the evidence does not support the change, and say what you would do instead.

12.2 Discussion · 7 min

An agent reports that it passed the graded lot of Chapter 9 with the suggested structure, a Peak temperature of 1048 °C and a soak of 1.10 s. The Lot target tile reads met, 25 of 25 in specification. Using Chapter 9 and Section 12.4.1, say what the report does and does not show, and what you would check before accepting it.

12.3 Discussion · 8 min

A new single-wafer deposition tool has five heater zones under the wafer. Film thickness is measured after deposition on one wafer in five, and the reading returns three wafers late. The incoming wafer's prior film is measured on every wafer at the step before. The chamber walls are cleaned every few days and the rate drifts after each clean. Use the questions of Section 12.2 to propose a control structure, and name the chapter that supports each element.

12.4 Discussion · 7 min

To avoid qualification wafers after every pad change, an agent proposes that the polisher of Chapter 11 carry every estimate across the change. It argues that the estimator is confident and that the release rule will hold production if anything is wrong. Evaluate the proposal using Chapter 11, and compare it with how Chapter 7 and Chapter 6 treat a wet clean.

12.5 Discussion · 8 min

A team must decide whether to add a predictive controller to the bake plate of Chapter 5 or to the plating bath of Chapter 10. Using the four comparisons of Section 12.2 (crystal puller, etch chamber, anneal, polysilicon rod), state where prediction paid in this volume and where it did not, and argue for one of the two candidates or for neither.

12.6 Discussion · 7 min

An agent has written three new operator aids for a fab's tools and reports that each sets its tool correctly when pressed. Using the cases of Section 12.4.1, write a short acceptance test for agent-written aids and say which failure of this volume each step would have caught.