Moore's Law in the Cost Data: Vintages, Learning, and Scale in Semiconductor Manufacturing

Ming-yen Ho* Mitsuru Igami† Shao-Yu Jheng‡

September 11, 2026

Abstract

Semiconductors dominate today's industrial-policy debates, which rest on "stylized facts" about the underlying technology and economics. Few of these facts are empirically grounded, because the industry is complex and the data are limited. We combine the specifications of every mainstream CPU and GPU product with a commercial-grade engineering cost model, building a panel of 77,151 chip-quarter observations for 2015–2025 in which physical quantities and the accounting cost of capital are recorded separately. Three findings emerge. First, newer process generations ("nodes"), or capital vintages, do not necessarily reduce the cost of a wafer or a chip: they raise it for GPUs. Moore's Law must therefore live in transistor counts, which the industry discloses too selectively to measure here. Second, most of the measured impacts of the much-celebrated learning-by-doing and scale economies are artifacts of depreciation accounting. Third, cost differences across firms are large and persistent, though comparisons per wafer or per chip confound capability with product mix. These results highlight the critical importance of engineering and accounting details in measuring physical productivity growth and its technological determinants.

Keywords: cost function, production function, productivity measurement, fab investment, innovation, learning-by-doing, scale economies, semiconductors, vintage capital.

1 Introduction

Governments worldwide are committing historic sums to semiconductors. The United States, the European Union, Japan, and China have each legislated subsidy programs measured in tens of billions of dollars. The premise is that this industry features steady productivity growth from new generations of manufacturing processes ("nodes"), strong learning-by-doing, and large scale economies. The scale of these bets is staggering. A single leading-edge fabrication plant ("fab") now costs on the order of $20 billion (Bown and Wang, 2024). Sustaining the historical doubling of

*Haas School of Business, University of California, Berkeley. E-mail: mho@berkeley.edu.
†Department of Economics, University of Toronto. E-mail: mitsuru.igami@utoronto.ca.
‡Department of Economics, Harvard University. Email: shaoyujheng@fas.harvard.edu.

transistor density has required more than eighteen times as many researchers as in the early 1970s (Bloom et al., 2020).

What these premises rest on, however, is rarely measured. Firms do not disclose their manufacturing costs, so the empirical literature has had to infer them from prices, shipments, and contracts, often under unverifiable assumptions about firms' competitive conduct. Physical productivity inside the fab is observed even less often than cost. Hence some of the industry's celebrated stylized facts—which technologies cut costs, how fast fabs learn, and how strong scale economies are—have not been checked against direct measurements.

This paper directly measures the determinants of semiconductor manufacturing costs. Our object is the cost function of a fabrication process, the dual of its production function, measured from engineering data rather than inferred from the joint behavior of inputs and output. We chose this industry for three reasons. First, the industry is important in its own right, as the object of the largest industrial-policy programs of the era. Second, it can be unusually well documented, because both buyers and sellers of chip fabrication services rely on an engineering cost model that embodies a detailed tear-down cost analysis as a common reference. Enjoying this documentation requires obtaining that proprietary model, which we did. Third, it is representative of high-technology manufacturing at large: a handful of fiercely competing firms, global supply chains centered on East Asia, and technical progress embodied in successive generations of fabs and equipment. Semiconductor devices share these features with many other electronic devices, including hard disk drives (Igami, 2017, 2018; Igami and Subrahmanyam, 2019; Igami and Uetake, 2020) and liquid crystal display panels (Igami et al., 2026).

A word on the industry's structure helps in reading what follows. Chip production separates design from fabrication. "Fabless" firms, such as Nvidia and AMD, design chips and purchase fabrication services from contract manufacturers called "foundries," of which TSMC is the largest. Integrated device manufacturers (IDMs), such as Intel, do both in-house. Production itself proceeds in four stages: wafer fabrication, wafer sort and test, packaging and assembly, and final chip test.

We combine two databases. The first is a public reference containing the specifications of every mainstream CPU and GPU product ("chip") released by these firms: die size, core counts, the foundry and process employed, and packaging technology. The second is the engineering cost model, which takes those specifications as inputs and computes the cost of each chip at each of the four production stages. The output is a panel of 77,151 chip-quarter observations (50,655 CPUs and 26,496 GPUs) covering 2015–2025, with a detailed cost decomposition at every stage.

These data help us in three important ways. First, we observe physical quantities—wafer yields, good dies per wafer, defect densities—separately from costs and margins. Foster et al. (2008) emphasized the importance of distinguishing physical productivity from its revenue-based counterpart; our panel contains exactly the physical measures their distinction calls for. Physical productivity is thus observed here rather than recovered as a residual. Second, depreciation is reported as its own line item. Production-function estimates usually measure capital at deflated book value, which buries the accounting treatment of capital expenditure inside the capital stock. Here the depre-

ciation charge is visible. We can therefore compute each cost measure with and without it, and quantify how much that treatment changes the estimated technology parameters. Third, technological vintage, process age, capacity utilization, and input-cost trends are recorded separately, so we can disentangle these four drivers of costs and productivity that are nearly collinear in aggregate time-series data. In production-function terms, the first three are embodied technical change, learning-by-doing, and short-run returns to scale, the shifters that a single productivity residual would otherwise absorb together. Three main findings emerge. First, newer process generations do not necessarily reduce the cost of a wafer, a die, or a chip. Across the nine node generations we observe, the estimated vintage effects are nearly monotone within each product category but opposite in sign between them. Newer nodes make GPU wafers and chips costlier, by roughly 5% per generation in cash terms and 15% in full-cost terms, and CPU wafers and chips cheaper. The GPU estimate is the better identified of the two, because a single foundry there runs nearly every generation side by side, whereas the node bins in the CPU sample are nearly foundry-specific, so that vintage cannot be separated from manufacturer and age effects. Whatever cost reduction Moore's Law delivers must therefore come through the growing number of transistors (i.e., the efficiency unit of computing) that each new generation places on a wafer of the same size, and not through a cheaper wafer. Measuring that channel requires transistor counts that the industry discloses only selectively, so we reserve it for the next version of this paper and report what the cost data alone support.

Second, most of the measured impacts of learning-by-doing and scale economies, two distinct and famous mechanisms of cost reduction, are artifacts of depreciation accounting. Full accounting cost per wafer, which prominently features depreciation expenses, falls by 9.4% per quarter of volume production for CPUs, whereas cash cost falls at a more modest 4.7%; the GPU rates are 2.3% and 1.0%. The gap reflects the mechanical rundown of depreciation charges as equipment ages, not physical improvement in productivity. The static scale economies long invoked to explain this industry's concentration shrink in the same way. Only the depreciation-inclusive measures decline appreciably as capacity utilization rises, whereas cash wafer cost varies by no more than 7% across the 70%–100% range in which roughly 95% of our chip-quarter observations lie.

Third, cost differences across firms are large and persistent, even among the few top players. They are not, however, rankings of efficiency, because a comparison per wafer or per chip confounds a firm's capability with the mix of products it makes. Among CPUs, TSMC's cash wafer cost sits about 50% above GlobalFoundries' and Intel's about 6% above it, because TSMC fabricates the densest CPU designs on the newest nodes, whose wafers pass through more process steps. Among GPUs, where the four manufacturers overlap at 14 nm–16 nm, the order reverses: TSMC and Samsung sit about 14% below GlobalFoundries, whereas Intel sits about 8% above it. Per chip, where die size and packaging dominate, Intel's GPUs cost 65% more than GlobalFoundries' and TSMC's 29% more. A firm's rank thus depends on the unit of account and on what it makes. Ranking manufacturers by technological capability requires a common unit of output, which is the transistor-level comparison we reserve.

The mechanism behind these findings is a pair of seemingly innocuous measurement choices. The two databases described above are what allow us to identify and isolate the sources of biases. The first choice is the unit of output. A wafer is a physical object of fixed size (e.g., 300 mm in diameter), whereas the number of transistors it carries has grown by orders of magnitude. The same cost data can therefore show stagnation or rapid progress, depending on which of the two sits in the denominator. Our estimates show the first of these two faces. The second choice is the treatment of capital. Fab equipment is enormously expensive. Accrual accounting spreads its cost over a fixed schedule (e.g., five years), so the depreciation expense booked to each wafer falls mechanically as the equipment ages and as more wafers go through the fab. Every cost measure that includes depreciation therefore declines with process age and with capacity utilization even where physical efficiency and economic marginal costs are unchanged. Such a mechanical decline is most of what conventional learning and scale estimates pick up. The line-item breakdown in the engineering cost model allows us to strip it out.

The takeaway is a caution on asking and answering "big-picture" questions without proper measurement. Vintage, learning, and scale effects are exactly the premises of the current investment and subsidy programs. We show that the measured strength of each depends on seemingly obscure engineering and accounting details: the unit in which output is counted, and the schedule on which tangible assets are written off. Studies that infer costs from prices and aggregates cannot observe either detail. Hence some of the "stylized facts" they have established, and the policy premises built on them, could be artifacts of measurement rather than features of the technology or the true underlying economics of the industry. Industrial policy in general has had mixed results, and semiconductor subsidies in particular have been a big money-loser in most countries. Evidence-based investments might face better odds, even though they are still a gamble.

We have organized the rest of the paper as follows. Section 2 reviews the related literature, with a focus on the data and modeling assumptions of each study. Section 3 describes the technology, our two databases, and the cost measures. Section 4 specifies the regressions and reports the determinants of cost per wafer, die, and chip. Section 5 concludes with three facts and their implications.

2 Related Literature

This section reviews the related literature in two steps. We first describe the older measurement tradition that this paper revives, and then organize the semiconductor studies by the inference problem each must solve. The Appendix tabulates the main datasets and modeling assumptions of each paper.

Engineering data as a measurement tradition. The ambition to build cost and production functions from engineering relations is an old one. Chenery (1949) argued that statistically estimated cost curves are confined to the input combinations firms happened to choose, and cannot

separate technological change from price variation; engineering data are free of both problems. Walters (1963) set out with equal care what such data cannot deliver: engineering relations describe a process or a plant rather than a firm, because they omit administration, selling, and entrepreneurial capacity.[1] Within that boundary, Walters identified the analysis of technical progress as the most promising use of engineering data. This paper conducts that exercise, at a scale unavailable when the question was first posed, following the recent use of engineering cost data for the LCD panel industry by Igami et al. (2026), as well as other use cases such as Agarwal et al. (2019) and Demirer and Karaduman (2024).

Production-function estimation. The modern literature recovers productivity from the joint behavior of inputs and output, treating it as an unobservable to which input choices respond (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015); De Loecker and Syverson (2021) survey the program from an IO perspective. Within it, learning-by-doing enters as a productivity shifter, estimated from plant-level records of output and inputs in aircraft and automobile assembly (Benkard, 2000; Levitt et al., 2013). Returns to scale have a dual measure, the elasticity of cost with respect to output, which Christensen and Greene (1976) estimated from cost data. Our exercise is the dual of that program with its unobservables observed. The engineering cost function records physical output, yields, and the accounting cost of capital as separate objects, so the three shifters that must otherwise be inferred, embodied technical change, learning, and scale, can be read off cost directly. The price of that directness is Walters' boundary: we measure a process, not a firm.

The studies of semiconductor manufacturing, by contrast, differ greatly in the type and granularity of their data but share one feature: none observes manufacturing cost directly. Each must infer it from prices, contracts, or shipments, under assumptions about firms' conduct or technology.

Learning-by-doing inferred from prices and shipments. Irwin and Klenow (1994) provide the classic benchmark, using Dataquest quarterly prices and shipments for seven DRAM generations between 1974 and 1992. Because firm-level cost data are unavailable, they recover marginal cost from a Cournot model with parametric demand elasticities, and treat shipments as a proxy for production. They estimate DRAM learning rates of about 20% per doubling of cumulative output, with own experience mattering roughly three times as much as others', and spillovers that are international rather than national. Jovanovic and Rousseau (2002) model Moore's Law itself as economy-wide learning-by-doing, assuming competitive capital-goods production so that a falling quality-adjusted price reveals falling cost; they estimate a much steeper learning curve for computers than for electricity or motor vehicles. Song (2007) estimates CPU demand from product-level data for 1993–2000 and finds that consumer surplus accounts for most of the social surplus; lacking cost

[1] Returns to scale in a process therefore need not imply returns to scale in the firm, as increasing returns in fabrication may be offset by the higher administrative costs of a larger organization. We respect that boundary throughout. Our firm effects measure technological capability rather than administrative efficiency. We make no claim about firm-level returns to scale.

data, the study relies on processor-speed-by-time interactions as cost proxies. In each case, the estimates are only as good as the assumptions that convert prices into implied costs.

Vintage capital as the alternative explanation. Aizcorbe and Kortum (2005) argue that chip price declines reflect not learning but better embodied equipment vintages, with older chips remaining in production because chip-specific equipment is sunk; their calibration uses Intel microprocessor prices, characteristics, and transistor counts, and treats Moore's Law as exogenous quality improvement. Pillai (2013) extends this view by modeling vintage adoption as an optimal stopping problem on a new Intel dataset covering 1971–2008, and decomposes progress into upstream equipment innovation and the downstream firm's efficiency in using additional transistors. Our results speak directly to this debate and locate it precisely. In the units in which cost is conventionally denominated, newer vintages are not reliably cheaper, so the vintage explanation must rest on the transistors that a wafer carries rather than on the cost of the wafer itself.

The research burden behind Moore's Law. Bloom et al. (2020) ask whether constant exponential growth requires ever more research effort, measuring idea output by the growth rate of transistor density and research input by R&D spending deflated by researcher wages. They find that the number of researchers required to sustain the historical doubling of chip density rose by more than a factor of eighteen from the early 1970s to the 2010s. Their subject is the input side of technological progress; ours is what that progress did to the cost of production.

Policy-oriented structural work. A newer literature evaluates state intervention. Bown and Wang (2024) survey the economic and geopolitical logic of subsidies and export controls, drawing on industry estimates such as the roughly $20 billion cost of a leading-edge fab. Miao (2025) models a dynamic oligopoly with step-by-step innovation and capacity subsidies under trade-disruption risk, calibrating US capacity costs above East Asian ones, and finds that only a sufficiently high disruption risk rationalizes a CHIPS-Act-scale subsidy. Goldberg et al. (2026) combine policy-text data with the GSA wafer-pricing survey, recovering marginal costs from wafer prices and quantities and reading residual country cost components as unobserved subsidies; they find sizable learning-by-doing and important cross-border spillovers. Zhuo and Tiew (2026) estimate a bargaining model on GSA order-level contracts and conclude that a US fab would require roughly $1.2 billion more investment than a home-region one. These papers push the frontier of policy analysis, but each must back cost parameters out of pricing or bargaining models. Our contribution is complementary: we measure the physical and accounting costs inside the fab directly, and report what they imply about the parameters those models must otherwise assume or infer.

3 Data

This section describes our data in two steps. We first explain the technology and the two databases we combine, and then define the four cost measures we analyze in the rest of the paper.

3.1 Technological Background and Datasets

We assemble the dataset from two sources: (i) a public online database containing the specifications of every mainstream CPU and GPU released by major chip designers, and (ii) detailed estimates of their manufacturing costs from a proprietary engineering model. The former supplies the inputs that the latter requires, such as the quarter of release, foundry choice, die size, number of cores, and substrate specifications. Combining them yields an engineering estimate of cost for each chip at each stage of the supply chain.

Industry Context. Chip manufacturing proceeds in four stages: wafer fabrication, wafer sort and test, packaging and assembly, and final chip test. We obtain cost estimates for each of them.

Integrated circuits are first patterned onto silicon wafers, which are 300 mm in diameter for every process in our sample.[2] Successive layers of materials are deposited, exposed to lithography, etched, and diced to form multiple dies (pre-packaged chips) from a single wafer. Chip designers that outsource this step to foundries (e.g., TSMC, GlobalFoundries, and Samsung Foundry) pay a margin on top of the raw manufacturing cost. The size of that margin depends on the designer's order volume. Integrated device manufacturers (IDMs) such as Intel fabricate in-house and pay no such margin.

The wafer then undergoes sort and test. Each die is electrically probed while still on the wafer to screen out defective ones. Yield, the fraction of units that survive a production stage intact, is the central quantity in the economics of semiconductor manufacturing. Each of the four stages has its own. Yield at the sort-and-test stage falls with the defect density of the process (i.e., the number of fabrication defects per unit of wafer area) and with the size of the die, because a larger die is more likely to encounter a randomly located defect. After dicing, the functional dies proceed to packaging and assembly. There, the dies are attached to a package substrate or lead frame, connected through wirebond or flipchip interconnects, and encapsulated. Intel packages in-house, whereas fabless designers rely on assembly and testing partners such as the ASE Group, Amkor, and SPIL. Packaged chips finally undergo chip test, which verifies performance and reliability.

Database 1: Product Specifications. Our first database is a public reference widely cited by hardware reviewers and industry analysts. It contains 4,169 unique CPUs and 3,310 unique GPUs released between the late 1990s and 2025. For each chip we observe the release date and production status, the manufacturing process (process node $\times$ foundry), the die size, the number of cores, the packaging method, and performance specifications such as clock speed. We also observe the transistor count, where the manufacturer reports one. Restricting attention to chips released during the coverage period of the engineering cost model (2015:Q1–2025:Q2) leaves 2,836 CPUs and 1,788 GPUs.[3]

[2]The engineering cost model also covers older processes on 150 mm and 200 mm wafers, none of which appears in our CPU and GPU samples.

[3]Each chip in the first database is in one of three statuses: active, unknown, end-of-life, which indicates retired chips that are no longer supported. Manufacturing costs of active or unknown-status chips are collected

Database 2: Manufacturing Costs. We feed these specifications into an engineering cost model that foundries and fabless designers use as a reference. The model is a deterministic engineering-accounting workbook built to replicate the cost of any chip on any process node between 2015:Q1 and 2025:Q2. The user specifies the year and quarter, the foundry process, die height and width, product type, the volume tier of the design firm, and the assembly location. The packaging inputs are the package type, substrate length and width, substrate layer counts, connection type, and connection count.

The model computes four objects, one for each stage.

1. **Wafer fabrication yield and cost**, separated into seven line items: the initial wafer, direct labor, depreciation, equipment maintenance, indirect labor, facilities, and consumables. Cost per wafer depends on the foundry's utilization rate, to which the seven items respond differently. The initial wafer and both labor items are variable costs. Depreciation is entirely fixed, so the charge per wafer falls as utilization rises. Equipment maintenance, facilities, and consumables have both a fixed and a variable component. We merge the panel with the quarterly utilization rate observed for each foundry and IDM. This replaces the workbook's default assumption and lets full wafer cost vary with the process and its utilization.

   The model then divides full wafer cost by the wafer yield, the probability that a wafer survives fabrication intact. The result is the yielded wafer cost that a chip designer expects to pay. A foundry margin applied to this cost gives the wafer price. The margin is zero for IDM processes, and varies by quarter, process node, and foundry for the rest.

2. **Gross and net dies per wafer at the wafer test stage.** Gross dies are the maximum number that fit on a 300 mm wafer, which the die size determines. Net dies are gross dies multiplied by the die yield, which depends on the defect density of the process in that quarter. Die yield improves within a process over time as the manufacturer gains experience.

3. **Yield and cost at the packaging-and-assembly stage**, which depend on the packaging method, the substrate size, and the type and number of interconnects.

4. **Cost at the class-test stage**, which increases with the number of cores.

The output is a chip-quarter panel with the full cost decomposition at each of the four stages. The CPU panel contains 50,655 chip-quarter observations and the GPU panel 26,496.

3.2 Definition of Four Cost Measures

This subsection defines the cost measures we analyze. We construct four of them for each chip j at fab or process k in year-quarter t, corresponding to successive stages of production. First, the

from max(release, 2015:Q1) to 2025:Q2. For end-of-life chips, we collect estimated manufacturing costs until min(release Q + 5 years, 2025:Q2), assuming that the median chip will not be in production after five years. Chips with missing release dates or release dates before 2015:Q1 are dropped.

foundry's unit cost of manufacturing a wafer is

$$\begin{aligned} C_{jkt}^{\text{Full}} &= \text{Yielded Full Wafer Cost}_{jkt} \\ &= \frac{\text{Full Wafer Cost}_{jkt}}{\text{Wafer Yield}_{jkt}}, \end{aligned} \tag{1}$$

where Full Wafer Cost_{jkt} includes the accrual-based accounting cost of depreciation of fab equipment. Its division by Wafer Yield_{jkt} accounts for the cost of waste from defective wafers.

Second, we also construct a purely cash-flow-based measure,

$$\begin{aligned} C_{jkt}^{\text{Cash}} &= \text{Yielded Cash Wafer Cost}_{jkt} \\ &= \frac{\text{Full Wafer Cost}_{jkt} - \text{Depreciation}_{jkt}}{\text{Wafer Yield}_{jkt}} \qquad (2) \\ &= \frac{\text{Cash Wafer Cost}_{jkt}}{\text{Wafer Yield}_{jkt}}, \qquad (3) \end{aligned}$$

by using Cash Wafer Cost_{jkt} instead of Full Wafer Cost_{jkt} in equation 1. This measure removes accrual-based depreciation expenses from the full wafer cost.

Third, once the wafer is sold and shipped from the foundry, the foundry's profit margin becomes part of the cost for the downstream industry. The foundry margin is a gross margin expressed as a fraction of the selling price.[4] Hence, the price per wafer charged by the foundry—and paid by the fabless chip-design firms—is

$$P_{jkt}^{\text{Wafer}} = \frac{C_{jkt}^{\text{Full}}}{1 - \text{Foundry Margin}_{jkt}}, \tag{4}$$

where the industry convention is to use the full cost, including depreciation. The wafer then undergoes a sorting and testing process, which entails a per-wafer Wafer Sort Cost_{jkt}, before being sliced and diced into dies. Thus, the cost at the level of dies is

$$\begin{aligned} C_{jkt}^{\text{Die}} &= \text{Cost per Good Tested Die}_{jkt} \\ &= \frac{P_{jkt}^{\text{Wafer}} + \text{Wafer Sort Cost}_{jkt}}{\text{Number of Good Tested Dies per Wafer}_{jkt}} \\ &= \frac{P_{jkt}^{\text{Wafer}} + \text{Wafer Sort Cost}_{jkt}}{\text{Number of All Tested Dies per Wafer}_{jkt} \times \text{Die Yield}_{jkt}}, \end{aligned} \tag{5}$$

which accounts for the costs of both the sorting-and-testing process and the wasted inputs used for defective dies.

[4] In the case of IDMs, such as Intel, the vertical separation between manufacturing and design does not apply. The engineering cost model sets Foundry $\text{Margin}_{jkt} = 0$ for them.

Fourth, the cost of the chip product (CPU or GPU) is

$$C_{jkt}^{\text{Chip}} = \left(\frac{C_{jkt}^{\text{Die}} + \text{Cost of Packaging \& Assembly}_{jkt}}{\text{Assembly Yield}_{jkt}} + \text{Cost of Class Test}_{jkt} \right) \div \text{Class Test Yield}_{jkt}. \tag{6}$$

This final measure adds the costs of packaging and assembly and of the final classification ("class") tests.

Why We Do Not Report Cost per Transistor. A natural fifth measure divides each of the four by the number of transistors. It is the measure this paper would prefer. The transistor is the unit in which both the cost and the capability of a fabrication process are denominated, and hence the level at which the engineering economics of Moore's Law becomes visible. We nevertheless report no per-transistor estimates in the current version of our paper, because some of the relevant data issues have not been fully resolved yet.

Transistor counts are directly disclosed for approximately 62% of the CPU and 90% of the GPU chip-quarter observations. The disclosure is selective in exactly the dimensions our regressions exploit. Among CPUs with disclosed counts, TSMC's products sit at 8 nm and below whereas Intel's and GlobalFoundries' sit at 10 nm and above, so manufacturer and vintage effects cannot be isolated. Old process ages are populated by old, transistor-poor chips, so comparisons across chip models confound age with product mix. Within a chip model, where that confound disappears, process age advances one-for-one with calendar time, generating perfect collinearity between learning and input-cost trends. Imputing the missing counts would not rescue the exercise.[5] We therefore leave the transistor-level analysis to the next version of this paper, in which obtaining reliable counts on a broader set of products is the first item on our agenda.

4 Determinants of Cost per Wafer, Die, and Chip

What determines the cost of manufacturing a semiconductor device? Our cost data come from a deterministic engineering model, whose mechanics we can reverse-engineer by regression. This section studies the four cost measures defined in Section 3: (i) yielded full wafer cost, C_{jkt}^{Full}; (ii) yielded cash wafer cost, C_{jkt}^{Cash}; (iii) cost per good tested die, C_{jkt}^{Die}; and (iv) cost per chip, C_{jkt}^{Chip}. We first specify the regressions and then report the estimates.

[5] Missing counts would be filled by a regression on the process node, die area, core count, and foundry and fabless-firm indicators. Because the process node is among the predictors, imputation would contaminate the vintage estimates exactly where the question is sharpest. Imputed values are also model output rather than engineering measurement, so their use would partly defeat the purpose of building the analysis on directly measured costs.

4.1 Specifications of the Regression Model

For each cost measure defined in Section 3, generically denoted by C_{jkt} for chip j at fab k in calendar quarter t, we estimate a linear model,

$$\ln C_{jkt} = \underbrace{\beta_1 Gen_k + \beta_2 Age_{kt} + \beta_3 Capa_{kt}}_{\text{fab/process}} + \underbrace{\beta_4 DieArea_j + \beta_5 Cores_j + \beta_6 Substrate_j}_{\text{product}} + \underbrace{\delta_{f(k)} + \delta_{l(j)}}_{\text{foundry/fabless firms}} + \underbrace{\beta_7 t}_{\text{calendar time}} + \varepsilon_{jt}, \quad (7)$$

as well as its nonlinear version, in which we replace each right-hand-side (RHS) variable with a full set of dummies.

Definitions. We define the regressors here in three groups, and discuss their economic interpretations later, alongside the coefficient estimates.

The first three regressors capture the process-level characteristics of fab k. Gen_k is the technological generation (process node).[6] Age_{kt} is the number of calendar quarters since the beginning of volume production. $Capa_{kt}$ is capacity utilization in percentage points. In production-function terms, the three are embodied technical change, learning-by-doing, and short-run returns to scale, given installed capacity.

The next three regressors are the die/chip-level characteristics of product j. $DieArea_j$ is the surface area of the die in $100\,\text{mm}^2$. $Cores_j$ is the number of cores (individual processing units) in thousands. $Substrate_j$ is a vector of specifications for the substrate, the foundational layer connecting the die to the motherboard. It contains (i) substrate size in $100\,\text{mm}^2$, (ii) the number of core layers, (iii) the number of build-up layers, (iv) connection count, and (v) an indicator for the more advanced "flipchip" connection technology as opposed to traditional "wirebond." Because $Substrate_j$ describes packaging rather than fabrication, it enters only the chip-level cost (C^{Chip}_{jkt}).

The last three regressors control for firms and time. $\delta_{f(k)}$ and $\delta_{l(j)}$ are firm-type dummies for foundries, which own and operate fabs, and for fabless firms, which design and sell chips.[7] In the linear specification, t is calendar time; in the nonlinear specification, it is replaced by year-quarter dummies, δ_t. Finally, the error term ε_{jkt} represents the remaining variation in cost arising from

[6] We categorize process nodes into nine bins ($\geq$40 nm, 32 nm, 28 nm, 20 nm–22 nm, 14 nm–16 nm, 10 nm–12 nm, 7 nm–8 nm, 5 nm–6 nm, and 3 nm–4 nm), which Gen_k maps into an integer index $(1, 2, \ldots, 9)$ in this order. A cautionary technical note is warranted regarding the precision of these node labels. Because node names simply reflect the manufacturers' own designations and not objective physical standards, they are not strictly comparable across firms and are often regarded as cosmetic labeling for marketing purposes. For example, Intel's 10 nm process is widely regarded as comparable to TSMC's 7 nm process in terms of transistor density. Thus, Gen_k and its coefficient estimates should be interpreted with this context in mind.

[7] More precisely, the latter is a set of order-volume dummies for high, average, and low volumes. Each fabless firm is classified into one of these tiers based on their typical order size, which is the sense in which $\delta_{l(j)}$ works as firm-type dummies.

additional nonlinearity and interactions.

4.2 Estimates

This subsection reports the estimates, one regressor at a time. Tables 1 and 2 contain the results for CPUs and GPUs, based on the 48,158 and 22,860 chip-quarter observations for which all variables are available.[8] The four cost measures and the two specifications occupy eight columns. The regressors appear in five groups of rows: fab, product, manufacturer, buyer, and time. Figures 1 and 2 plot the full-dummy estimates of the three fab-level variables and calendar time. The two samples show broadly similar patterns, so we discuss them together and note the exceptions where they arise.

Two features of the results deserve emphasis at the outset. First, the fit is extremely high: the adjusted R^2 is at least 0.801 in every linear specification and at least 0.927 in every full-dummy specification. Practically all relevant factors seem captured in these relatively parsimonious regressions, without additional dummies for cross-category interactions (e.g., to allow for heterogeneous trajectories of the vintage-capital or learning-by-doing effect across combinations of products and manufacturers).

Second, wherever the linear and full-dummy estimates disagree, we rely on the latter and treat the disagreement as evidence in its own right. It signals collinearity among variables that move together, which is precisely the problem afflicting studies that work with more aggregated data. For the three variables that advance together—vintage, process age, and calendar time—the dummy profiles should also be read jointly rather than one at a time. Only the sum of the three profiles is well identified if processes of different vintages do not coexist within a node bin. That caution binds in the CPU sample, for a reason we document below. There we still report the first-to-last averages of the dummy profiles as our preferred estimates, read jointly, with the linear estimates as a secondary check. It does not bind in the GPU sample.[9]

Process Generation. Within each sample the vintage profile is nearly monotone, whereas between the two samples it runs in opposite directions.[10] For CPUs, newer nodes appear cheaper. In the linear specification, full wafer cost (C^{Full}_{jkt}) falls by $\left|e^{-0.088} - 1\right| \times 100 = 8.4\%$ per generation, as do cash and chip cost. The full-dummy profile in Figure 1 (A) also declines toward the frontier.

For GPUs, by contrast, newer nodes are costlier at every level of aggregation. All four linear

[8] Because each chip is manufactured on a specific fab and process, a (chip, quarter) pair identifies a (product, fab, quarter) triple.

[9] The CPU node, age, and calendar-time profiles reach 3.7, −5.0, and +3.4 log points of full wafer cost at their extremes. No single variable can plausibly carry such magnitudes. The three profiles largely offset one another. The linear coefficients on the same three variables are −0.088 per generation, −0.047 per quarter of age, and 0.034 per calendar quarter. The GPU profiles stay within 1.2, 1.2, and 0.5 log points, in line with their linear counterparts.

[10] We use process generations ("nodes") as our measure of technological vintage, in the sense of Aizcorbe and Kortum (2005). The correspondence with equipment vintages is close but not one-to-one. Adjacent nodes often share or upgrade existing tools, whereas the major transitions in transistor architecture and patterning replace critical equipment and change the overall equipment mix. Comparisons across broadly separated bins therefore capture embodied equipment advances; a single step between adjacent bins need not.

coefficients are positive, from $\left|e^{0.051}-1\right| \times 100 = 5.2\%$ per generation for cash wafer cost to 15.4% for full wafer cost. The full-dummy profiles in Figure 2 (A) rise monotonically from 20 nm–22 nm to the frontier, with small reversals only among the older bins. On a full-cost basis, a 3 nm–4 nm GPU wafer costs about 2.7 times as much as one at 40 nm or older.

The two samples differ in their identifying variation for the vintage effect. The GPU estimate is the one we trust. In the GPU sample TSMC alone spans eight of the nine generations, several of them in production at once. Four manufacturers share the 14 nm–16 nm node. Each of the three largest node bins therefore contains processes of different vintages in most quarters, which is what separates vintage from process age and calendar time. The GPU results suggest that a frontier wafer is more expensive to process than a mature one, before any adjustment for what it carries (i.e., the number of transistors). The cost increase is consistent with the lengthening process flow: in the engineering model, a TSMC wafer passes through 51 mask layers at 40 nm and 78 at 7 nm.

The CPU sample has no such overlap. Its node bins are nearly foundry-specific: 32 nm, 20 nm–22 nm, and 10 nm–12 nm contain only Intel (i.e., a single Intel process each, P1269, P1271, and P1275), 28 nm is almost entirely GlobalFoundries, and the three frontier bins are almost entirely TSMC (i.e., TSMC is the only manufacturer with sizable CPU volume at the frontier nodes). Its largest bins each rest on a single process online date, or on two that fall in the same quarter, so that within them age moves one-for-one with calendar time. This collinearity generates coefficient estimates for the node, age, and time dummies with offsetting values of implausible size. Hence the apparent CPU cost reduction toward the frontier cannot be separated from the foundry effects (see below) and the age profile.

Technological advances in semiconductor manufacturing thus do not reduce the cost of a wafer, a die, or a chip; they raise it in the GPU sample, where the identifying variation is more reliable. This lack (or reversal) of vintage effects, however, does not necessarily mean the death of Moore's Law, because the findings are specific to these units of account. A new process generation packs more transistors onto a wafer of the same size, so cost *per transistor* can fall where cost *per wafer* does not. The counts needed to estimate that margin are disclosed too selectively to support it (Section 3). Finally, one robust pattern holds in both samples: cash wafer cost is far less sensitive to vintage than the three measures that embed depreciation.

Process Age. The number of calendar quarters since a fabrication process came online is our measure of accumulated experience, or learning-by-doing (LBD).[11] It captures the continuous effort of process engineers to fine-tune the equipment and the production process through trial and error. It is exactly this variable that determines yield improvement inside the engineering cost model underlying our data. The same holds in the LCD panel industry, where the engineering cost model makes yield a deterministic function of the fab's equipment generation and its elapsed quarters of

[11] Age counts quarters from the online date of the exact process, so processes that came online before our 2015 window enter the sample at correspondingly high ages.

volume production (Igami et al., 2026).[12] In two different industries, the cost engineers who serve both buyers and sellers treat experience as a function of time.

Typical LBD studies measure experience by cumulative output instead. Cumulative output is also a proxy, and one adopted because shipments are more widely available. It is the weaker proxy here, for two reasons. First, any well-managed fab in this industry runs near capacity: roughly 95% of our chip-quarter observations lie at or above 70% utilization. Output therefore accumulates at a rate that varies little across fabs and quarters, whereas the time a process has been running varies a great deal. Second, cumulative output conflates the size of a fab with its experience. A larger fab runs more machines in parallel; it does not learn faster, because the opportunities to fine-tune a process arrive with time rather than with the number of wafers in the line at once. Hence age is the dimension along which experience actually differs in these data. It is also the one the engineering models use.

Figure 1 (B) plots the CPU age profiles.[13] All four decline almost linearly with age. The full-cost and die-cost profiles fall much faster than the cash profile. Between quarters 4 and 52, C^{Cash}_{jkt} falls by 2.30 log points, an average of 0.048 per quarter, or $\left|e^{-0.048}-1\right| \times 100 = 4.7\%$. The linear estimate in column (3) of Table 1 is $\left|e^{-0.027}-1\right| \times 100 = 2.7\%$ per quarter. The full-cost profile falls by 4.75 log points over the same range, or 9.4% per quarter, against 4.6% in the linear estimate of column (1). Our preferred estimate is the dummy average, 4.7% per quarter for cash cost, with the caution that in the CPU sample the age, node, and time dummies are identified only jointly, so the age profile carries part of what belongs to the other two. The linear estimate, about half as steep, is a secondary check that disagrees. We read the disagreement as collinearity rather than as evidence for either number.

Figure 2 (B) and Table 2 tell a more consistent story for GPUs, where the identifying variation is better: TSMC runs several generations side by side, so age is not tied to vintage and calendar time within a node bin. The cash profile falls by 0.46 log points between quarters 4 and 52, an average of 1.0% per quarter, close to the linear estimate of $\left|e^{-0.012}-1\right| \times 100 = 1.2\%$ in column (3). The full-cost profile falls by 1.12 log points, or 2.3% per quarter, against 2.8% in the linear estimate of column (1). Both specifications thus put the physical learning rate for GPUs at about 1% per quarter of volume production, and the depreciation-inclusive rate at two to three times as much.

The gap between the full- and cash-cost profiles is the robust bottom line, common to both samples and both specifications. The gap reflects depreciation accounting, which exaggerates the apparent effect of age because depreciation mechanically declines over time.[14] We therefore read

[12] The full list of yield determinants there is the equipment generation, the number of quarters since volume production began, the product's size, and the manufacturer's technological tier. Calendar time enters only through the last of these, and negligibly.

[13] The age dummies pool observations at four quarters or less into the reference bin, and those at 53 quarters or more into a top bin. Hence the plotted profiles start at quarter 4, where they are zero by construction, and end with the top bin at 53. The average slopes below are computed between quarters 4 and 52, the last single-quarter bin.

[14] More specifically, the depreciation cost becomes zero after a certain number of years, regardless of the exact schedule chosen. In addition, any constant-rate depreciation schedule results in costs that strictly decrease with time. Curiously, C^{Chip}_{jkt} shows a trajectory similar to C^{Cash}_{jkt} despite its inclusion of depreciation cost.

the cash-based profile as the closest available measure of the physical LBD effect: 4.7% per quarter for CPUs, read with the caution above, and 1.0% for GPUs.

Capacity Utilization. Figure 1 (C) plots the CPU utilization profiles. The reference bin pools every observation at or below 65% utilization, which is only 1.4% of the CPU sample.[15] Moving from the reference bin to 70–75% lowers C_{jkt}^{Full} by 0.43 log points, or $\left|e^{-0.425}-1\right| \times 100 = 35\%$, but C_{jkt}^{Cash} by only 0.07 log points, or 7%. Above 70%, the range in which these fabs normally operate, the two profiles part ways. The full-cost profile keeps falling, by another 0.18 log points at 95–100%, whereas the cash profile is flat: its values between 70% and 100% lie within 0.06 log points of one another. The linear estimates in Table 1 summarize the same contrast. One percentage point of utilization reduces full wafer cost by $\left|e^{-0.010}-1\right| \times 100 = 1.0\%$ but cash wafer cost by only 0.2%.

Figure 2 (C) and Table 2 show the same pattern for GPUs. The step out of the reference bin is larger, 0.54 log points (42%) for full wafer cost against 0.11 log points (10%) for cash wafer cost. Above 70% the full-cost profile is no longer even monotone. The cash profile stays within 0.02 log points of its 70–75% value. The linear slopes are 0.7% and 0.3% per percentage point.

The gap between the full- and cash-cost profiles is again depreciation: a fixed expense spread over more wafers falls mechanically as utilization rises, which inflates the full-cost estimate without any change in physical efficiency. We therefore take the cash estimate as the physical scale economy. Given installed capacity, utilization is output. The utilization profile is therefore the short-run average cost curve itself. Its slope is the textbook economy of scale, the dilution of a fixed cost over more wafers (Christensen and Greene, 1976). A flat cash profile says that the physical part of that economy is small.[16] The only sizable cash saving comes from leaving the lowest bin, which is consistent with a fab that must staff continuous shifts regardless of load. Beyond 70% there is little physical efficiency left to gain.

Product Specs. Die/chip-level product characteristics affect costs in mostly predictable ways. Die size has a small impact on wafer-level costs. In the linear estimates of Tables 1 and 2, the largest of the four wafer-level coefficients implies $\left|e^{-0.023}-1\right| \times 100 = 2.3\%$ per $100\,\text{mm}^2$ (GPU full wafer cost), which suggests that how to slice and dice a wafer is not a first-order issue in wafer processing. By contrast, die size is a critical factor at the level of dies and chips for two reasons. First, it mechanically alters the number of dies per wafer, as larger dies mean fewer of them on a wafer of a given size. Second, the larger the die, the higher the defect probability, because defects arise stochastically at a certain density per unit of surface area. A $100\,\text{mm}^2$ increase in die size raises C_{jkt}^{Die} by $\left|e^{0.520}-1\right| \times 100 = 68.2\%$ for CPUs, with the chip-level effect and both GPU counterparts of the same order. One pattern resists quick interpretation: larger dies are correlated with slightly

[15] The regressions clip utilization at 60 before binning, so the reference bin is labeled 60–65% in the figure but pools everything at or below 65%: 1.4% of the CPU and 1.5% of the GPU chip-quarter observations. Roughly 95% of the observations in each sample lie at or above 70%.

[16] Installed capacity does not enter the regressions directly. The industry data behind our panel record each manufacturer's installed capacity by process node and quarter, so a fuller specification can enter capacity and its utilization separately, and their product as the level of output. We leave that extension to the next version.

lower wafer costs in both samples. A hypothesis is that more dies per wafer mean more work, or other costs, at the wafer stage, for example in handling the blank space between them; verifying it requires engineering facts rather than theorizing, which we leave to the next version.

The number of cores is measured in 1,000 units so that CPU and GPU core counts can be read on the same scale. Additional cores come at a cost. GPU cores are one to two orders of magnitude cheaper than CPU cores, because they are numerous and small. The five substrate specifications enter only the chip-level cost. We do not discuss each in detail, but four of the five have intuitive positive coefficients in each sample, reflecting the extra cost of extra components or quality; the exception is the number of build-up layers for CPUs and the flipchip indicator for GPUs. In the full-dummy specification these five specifications are subsumed in the substrate-type dummies anyway.

Manufacturers. Firm effects are large. Their sign depends on the specification and the unit of account. In the CPU sample, the full-dummy estimates place TSMC's cash wafer cost $\left|e^{0.406} - 1\right| \times 100 = 50.1\%$ *above* GlobalFoundries (the omitted category), whereas the linear cash estimate is close to zero; Intel sits below GlobalFoundries on a full-cost basis and above on a cash basis. In the GPU sample, the full-dummy estimates put TSMC and Samsung each about 14% *below* GlobalFoundries in cash wafer cost, but TSMC flips above at the die and chip levels. Our interpretation is that foundries operating frontier processes specialize in large, high-specification products, so that comparisons denominated in wafers, dies, or chips confound a firm's technological capability with the mix of products it manufactures. Separating the two requires a common unit of output, which is the transistor-level comparison we reserve in the current version of the paper.

Buyers. Manufacturers tend to offer discounts on high-volume orders, which the engineering cost model incorporates. The estimates recover it for GPUs, where high-volume purchases translate into lower costs throughout, but not for CPUs, where high-volume buyers appear to face *higher* cash wafer costs. Our interpretation is the same as for manufacturers: high-volume buyers such as Apple and Nvidia tend to order high-specification products, and product mix contaminates the comparison.

Time. The full-dummy estimates in Figures 1 (D) and 2 (D) show clearly positive calendar-time effects in both samples, consistent with the secular industry-wide inflation in the cost of materials and intermediate inputs. The CPU magnitudes overstate that inflation for the reason we mentioned at the beginning of this subsection: the time, age, and node profiles are collinear, and only their sum is well identified in the CPU data.

For CPUs, the cash profile in Figure 1 (D) rises by 1.91 log points between 2015:Q1 and 2025:Q2, an average of 0.047 per quarter, or $\left|e^{0.047} - 1\right| \times 100 = 4.8\%$. The full-cost profile rises by 3.36 log points, or 8.6% per quarter, with the die and chip profiles at 7.0% and 3.2%. Taken alone, these figures would imply input-cost inflation of nearly 5% per quarter in cash terms, which is not credible. The CPU time dummies absorb part of what belongs to the age and node dummies. Only

the sum of the three profiles is identified. The linear trend for cash cost, $\left|e^{0.027} - 1\right| \times 100 = 2.7\%$ per quarter in column (3) of Table 1, is lower but suffers from the same collinearity.

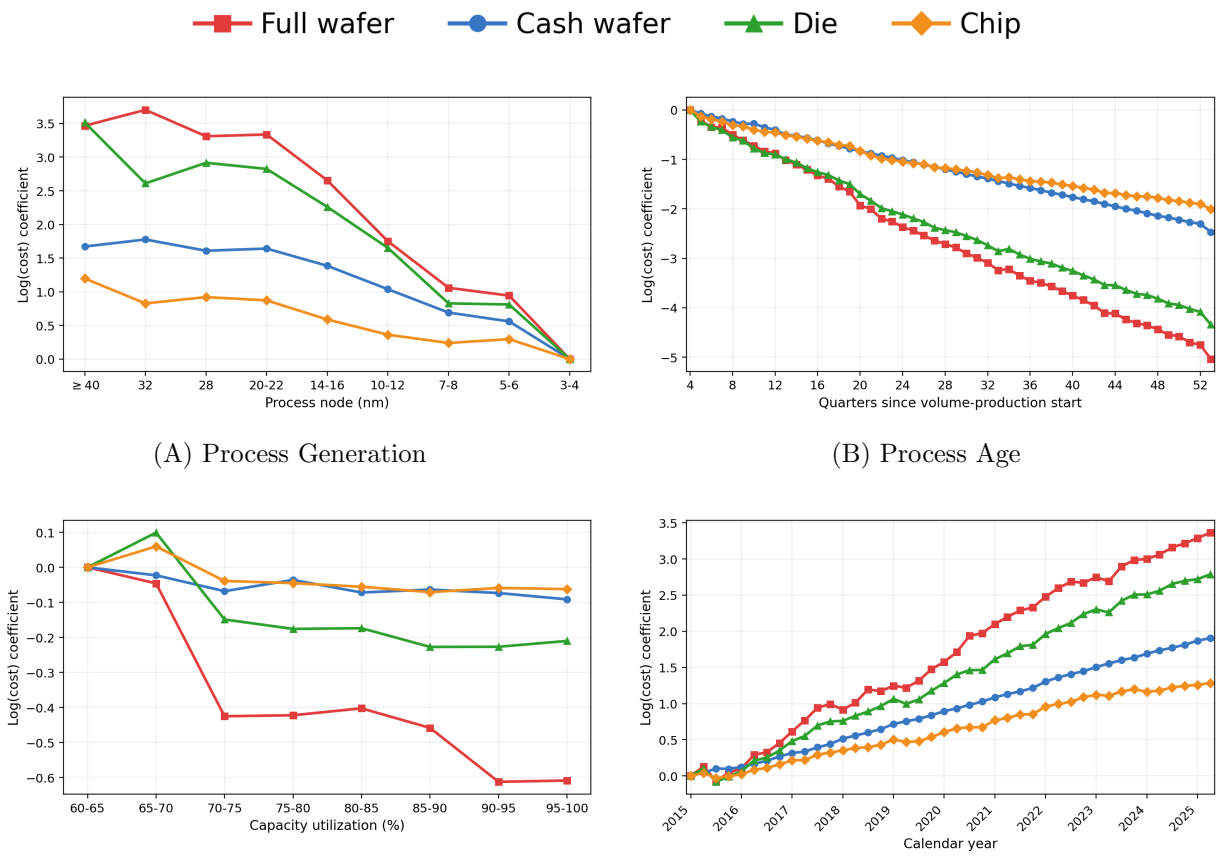
For GPUs, where the three time-like variables are separately identified, the cash profile in Figure 2 (D) rises by 0.44 log points over the same 41 quarters, an average of 1.1% per quarter. The full-cost profile rises by only 0.31 log points, or 0.8% per quarter, with the die and chip profiles at 2.4% and 1.1%. The cash figure is our preferred estimate of input-cost inflation: about 1% per quarter, or 4% per year. The linear trend for cash cost, 1.4% per quarter in column (3) of Table 2, confirms it, whereas the linear trends for the other three measures are negative or close to zero, contradicting their own full-dummy counterparts. This sign instability is the clearest evidence in our data of collinearity among the regressors that vary over time, and a caution against reading linear trend coefficients as structural objects.

Table 1: Determinants of Manufacturing Cost (CPU)

Regressor	ln(Full Wafer Cost)		ln(Cash Wafer Cost)		ln(Die Cost)		ln(Chip Cost)	
	Linear (1)	Dummies (2)	Linear (3)	Dummies (4)	Linear (5)	Dummies (6)	Linear (7)	Dummies (8)
A. Fab specs								
Process generation (index)	−0.088	–	−0.027	–	0.061	–	−0.047	–
	(0.002)	(–)	(0.000)	(–)	(0.003)	(–)	(0.002)	(–)
Process age (quarters)	−0.047	–	−0.027	–	−0.036	–	−0.026	–
	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)
Capacity utilization (%)	−0.010	–	−0.002	–	−0.008	–	−0.002	–
	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)
B. Product specs								
Die size (100 mm^2)	−0.016	–	−0.003	–	0.520	–	0.245	–
	(0.001)	(–)	(0.000)	(–)	(0.002)	(–)	(0.001)	(–)
Number of cores (1,000)	1.925	–	0.939	–	2.039	–	2.824	–
	(0.085)	(–)	(0.032)	(–)	(0.121)	(–)	(0.132)	(–)
Substrate size (100 mm^2)	–	–	–	–	–	–	0.008	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.000)	(–)
Substrate core layers	–	–	–	–	–	–	0.189	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.003)	(–)
Substrate buildup layers	–	–	–	–	–	–	−0.037	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.002)	(–)
Connections (1,000)	–	–	–	–	–	–	0.135	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.001)	(–)
Flipchip Connection	–	–	–	–	–	–	0.066	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.004)	(–)
C. Manufacturer								
Foundry: Intel	−0.166	−0.141	0.093	0.060	−0.368	−0.029	−0.307	−0.194
	(0.003)	(0.004)	(0.001)	(0.001)	(0.005)	(0.005)	(0.003)	(0.005)
Foundry: TSMC	0.137	0.964	−0.010	0.406	−0.058	1.345	0.252	0.427
	(0.006)	(0.039)	(0.001)	(0.018)	(0.008)	(0.059)	(0.005)	(0.054)
D. Buyer specs								
Volume: High	−0.026	−0.023	0.239	0.000	−0.141	0.010	−0.141	0.188
	(0.005)	(0.010)	(0.002)	(0.004)	(0.004)	(0.012)	(0.004)	(0.025)
Volume: Low	−0.273	−0.657	0.238	−0.259	0.214	−1.009	−0.563	−0.182
	(0.017)	(0.048)	(0.004)	(0.018)	(0.024)	(0.062)	(0.011)	(0.061)
E. Time and constant								
Time	0.034	–	0.027	–	0.024	–	0.021	–
	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)
Constant	−265.390	6.029	−211.578	6.005	−188.812	−0.333	−168.916	2.501
	(1.684)	(0.073)	(0.384)	(0.033)	(2.679)	(0.116)	(1.944)	(0.076)
Process gen dummy	No	Yes	No	Yes	No	Yes	No	Yes
Process age dummy	No	Yes	No	Yes	No	Yes	No	Yes
Capacity util dummy	No	Yes	No	Yes	No	Yes	No	Yes
Product-type dummy	No	Yes	No	Yes	No	Yes	No	Yes
Substrate-type dummy	No	No	No	No	No	No	No	Yes
Time dummy	No	Yes	No	Yes	No	Yes	No	Yes
R^2	0.848	0.955	0.963	0.990	0.883	0.953	0.944	0.973
Adjusted R^2	0.848	0.954	0.963	0.990	0.883	0.953	0.944	0.973

Note: The CPU sample consists of 48,158 chip-quarter observations for which all four LHS (cost) variables, as well as all RHS variables, are available. Heteroskedasticity-robust standard errors are in parentheses. The omitted category for manufacturer dummies is GlobalFoundries; the omitted category for buyer specs is average volume.

Figure 1: Determinants of Manufacturing Cost (CPU), Full-Dummy Estimates



(A) Process Generation

(B) Process Age

(C) Capacity Utilization

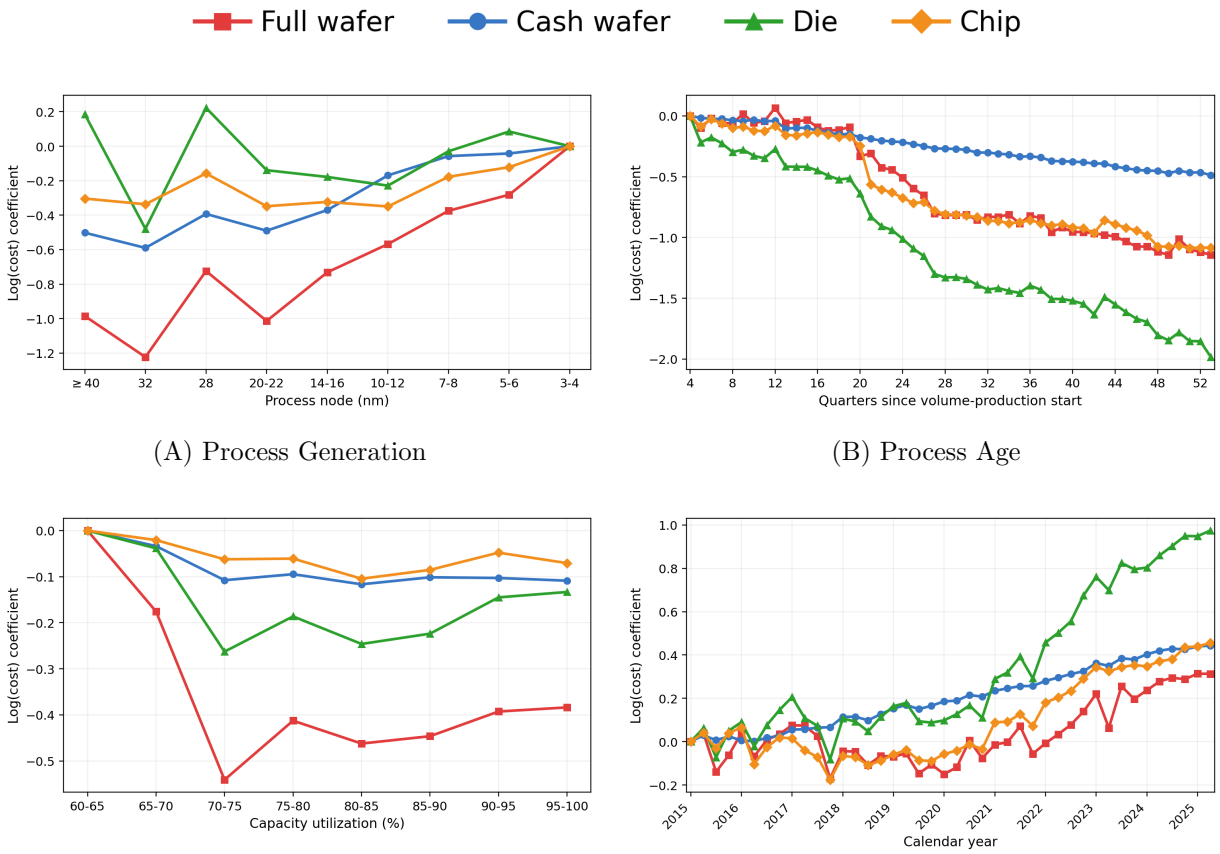
(D) Time

Table 2: Determinants of Manufacturing Cost (GPU)

Regressor	ln(Full Wafer Cost)		ln(Cash Wafer Cost)		ln(Die Cost)		ln(Chip Cost)	
	Linear	Dummies	Linear	Dummies	Linear	Dummies	Linear	Dummies
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Fab specs								
Process generation (index)	0.143	–	0.051	–	0.134	–	0.137	–
	(0.003)	(–)	(0.001)	(–)	(0.007)	(–)	(0.004)	(–)
Process age (quarters)	−0.028	–	−0.012	–	−0.029	–	−0.012	–
	(0.001)	(–)	(0.000)	(–)	(0.001)	(–)	(0.001)	(–)
Capacity utilization (%)	−0.007	–	−0.003	–	−0.007	–	−0.005	–
	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)	(0.000)	(–)
B. Product specs								
Die size ($100\,\mathrm{mm}^2$)	−0.023	–	−0.005	–	0.487	–	0.275	–
	(0.001)	(–)	(0.000)	(–)	(0.004)	(–)	(0.002)	(–)
Number of cores (1,000)	0.028	–	0.009	–	−0.012	–	0.018	–
	(0.001)	(–)	(0.000)	(–)	(0.003)	(–)	(0.002)	(–)
Substrate size ($100\,\mathrm{mm}^2$)	–	–	–	–	–	–	0.023	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.001)	(–)
Substrate core layers	–	–	–	–	–	–	0.039	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.003)	(–)
Substrate buildup layers	–	–	–	–	–	–	0.002	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.002)	(–)
Connections (1,000)	–	–	–	–	–	–	0.085	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.006)	(–)
Flipchip Connection	–	–	–	–	–	–	−0.099	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.004)	(–)
C. Manufacturer								
Foundry: Intel	−0.284	−0.243	0.085	0.075	0.078	0.226	0.093	0.498
	(0.007)	(0.009)	(0.002)	(0.002)	(0.012)	(0.007)	(0.010)	(0.035)
Foundry: Samsung	−0.306	−0.279	−0.061	−0.158	0.114	0.035	0.080	−0.015
	(0.006)	(0.009)	(0.001)	(0.003)	(0.012)	(0.010)	(0.009)	(0.011)
Foundry: TSMC	−0.088	−0.190	−0.055	−0.155	0.462	0.225	0.377	0.251
	(0.004)	(0.007)	(0.001)	(0.002)	(0.007)	(0.007)	(0.005)	(0.010)
D. Buyer specs								
Volume: High	−0.009	−0.004	−0.033	−0.005	−0.092	−0.052	−0.065	−0.051
	(0.003)	(0.002)	(0.001)	(0.001)	(0.006)	(0.004)	(0.004)	(0.006)
Volume: Low	0.042	0.033	0.010	−0.003	0.486	0.468	0.304	0.295
	(0.013)	(0.009)	(0.002)	(0.001)	(0.039)	(0.016)	(0.029)	(0.013)
E. Time and constant								
Time	−0.004	–	0.014	–	−0.002	–	−0.014	–
	(0.001)	(–)	(0.000)	(–)	(0.001)	(–)	(0.001)	(–)
Constant	41.014	9.556	−102.590	7.815	20.676	2.453	113.239	3.762
	(4.949)	(0.041)	(1.878)	(0.016)	(9.020)	(0.049)	(5.494)	(0.062)
Process gen dummy	No	Yes	No	Yes	No	Yes	No	Yes
Process age dummy	No	Yes	No	Yes	No	Yes	No	Yes
Capacity util dummy	No	Yes	No	Yes	No	Yes	No	Yes
Product-type dummy	No	Yes	No	Yes	No	Yes	No	Yes
Substrate-type dummy	No	No	No	No	No	No	No	Yes
Time dummy	No	Yes	No	Yes	No	Yes	No	Yes
R^2	0.801	0.928	0.956	0.986	0.891	0.979	0.913	0.982
Adjusted R^2	0.801	0.927	0.956	0.985	0.890	0.979	0.913	0.982

Note: The GPU sample consists of 22,860 chip-quarter observations for which all four LHS (cost) variables, as well as all RHS variables, are available. Heteroskedasticity-robust standard errors are in parentheses. The omitted category for manufacturer dummies is GlobalFoundries; the omitted category for buyer specs is average volume.

Figure 2: Determinants of Manufacturing Cost (GPU), Full-Dummy Estimates



(A) Process Generation

(B) Process Age

(C) Capacity Utilization

(D) Time

5 Facts and Implications

Three facts emerge from our analysis of CPU and GPU manufacturing costs. Most of the underlying concepts are familiar from the existing literature. Our results suggest that the assumptions researchers impose when modeling them deserve careful recalibration.

Fact 1: Newer Vintages Do Not Reliably Reduce the Cost of a Wafer, a Die, or a Chip. Measured per wafer, per die, or per chip, the vintage-capital effect runs in opposite directions in our two samples. For CPUs, the estimates point toward cheaper wafers at the frontier, but that sample cannot separate the decline from foundry identity and process age, because its node bins are nearly foundry-specific. For GPUs, where the comparison has cleaner variation, newer nodes are costlier at every level of aggregation, by roughly 5% per generation in cash wafer cost and 15% in full wafer cost. The wafer itself does not become cheaper; at the frontier it becomes dearer. Moore's Law is a statement about transistors rather than about wafers. Our estimates locate it accordingly. Whatever cost reduction it delivers must come through the growing number of transistors that a new generation places on a wafer of the same size. Measuring that channel is the exercise that Section 3 defers.

Two implications follow. The first is methodological. The measured vintage effect depends on the unit of account and on whether the sample can separate vintage from foundry and age composition. Hence both the choice of output unit and the data construction are substantive issues rather than matters of convenience.

The second concerns the literature. Aizcorbe and Kortum (2005) and Pillai (2013) interpret historical price declines as technological change embodied in newer equipment vintages rather than as learning within a vintage. Our estimates do not support that interpretation in the units in which manufacturing cost is conventionally denominated, which places the weight of the vintage explanation on transistor density. Igami et al. (2026) use similarly detailed engineering data from the LCD panel industry and find that most new generations of fab equipment reduced cost per panel. The contrast with our results cautions against cross-industry generalization, even within IT hardware manufacturing.

Fact 2: The Measured Impacts of Learning-by-Doing and Scale Economies Are Largely Artifacts of Depreciation Accounting. Learning-by-doing (LBD) and economies of scale are the two mechanisms most often invoked to explain falling costs and concentrated market structure in this industry; both serve as premises for state intervention. The literature has had to measure them from outside the fab: LBD from aggregated prices and shipment volumes, combined with assumptions about firms' competitive conduct that are neither tested nor verified, and scale economies typically as a cost function decreasing in output. Shipment volumes stand in for experience because they are what those data record, whereas the engineering cost models used inside this industry (and inside the LCD panel industry) make yield improvement a function of elapsed time. We measure experience as the industry's own cost engineers do. Depreciation is a separate

line item in our data, which allows us to ask how much of each measured mechanism is physical and how much is bookkeeping. The answer is that the bookkeeping share is large for both. A fixed depreciation expense declines mechanically as equipment ages and as output expands. That one convention shows up as learning along the time dimension and as scale economies along the output dimension.

Start with learning. In our preferred full-dummy estimates, cash cost per wafer falls by 4.7% per quarter of volume production for CPUs and by 1.0% for GPUs, whereas the depreciation-inclusive full cost falls by 9.4% and 2.3% (the CPU profiles are identified only jointly with node and time, as Section 4 cautions). The gap is the mechanical rundown of depreciation charges as equipment ages, not physical improvement. Between one-half and three-fifths of the apparent learning effect is therefore bookkeeping. Because our regressors span the characteristics of fabs, products, firms, and time, and because the fit is high, we regard the cash rate as a reasonably clean estimate of the physical learning effect per wafer. We also find visible discrepancies between the linear and full-dummy specifications. These discrepancies indicate collinearity among the time-related variables that move together, which no study working with more aggregated data could separate.

Scale economies tell the same story. Without detailed cost data it is difficult to separate static scale economies from their dynamic counterpart, LBD. Our data separate them. The static component turns out to be small. Given installed capacity, utilization is output, so the utilization profile is the short-run average cost curve; its slope is the textbook economy of scale, the dilution of fixed cost, measured directly. Only the depreciation-inclusive measures decline appreciably as capacity utilization rises. In our preferred full-dummy estimates, moving out of the lowest utilization bin cuts full wafer cost by 35% for CPUs and 42% for GPUs but cash wafer cost by 7% and 10%. Across the 70%–100% range in which roughly 95% of our chip-quarter observations lie, the cash profile varies by no more than 7%. The linear estimates agree: a one-percentage-point increase reduces full wafer cost by 1.0% for CPUs but cash wafer cost by only 0.2%. What looks like a scale economy is in substantial part a fixed depreciation expense spread over larger output. Maintaining a reasonable rate of utilization still matters, because a fab runs continuous work shifts, but beyond that rate there is little physical efficiency left to gain.

The lesson for research practice is simple. When capital is measured at book value and output is counted in physical units, the accounting convention can be larger than the mechanism under investigation. Separating the two requires a line-item decomposition of the kind used here.

Fact 3: Firm Heterogeneity Is Large and Persistent, but Per-Wafer Rankings Confound Capability with Product Mix. Firm effects are large, and comparable to the effects of the three process-level cost drivers combined. They are not, however, rankings of efficiency. They are also unstable enough to reverse sign between specifications. In the CPU sample the full-dummy estimates place TSMC's cash wafer cost about 50% *above* GlobalFoundries, whereas in the GPU sample TSMC and Samsung each sit about 14% *below* it. Our interpretation is that these comparisons confound a firm's technological capability with the mix of products it manufactures,

since frontier foundries concentrate on large, dense products. A wafer of 3 nm logic and a wafer of 28 nm logic are not the same object. Ranking manufacturers by technological capability requires a common unit of output, which is the comparison we reserve in the current version of the paper.

Historically, this industry has witnessed the demise of many heavily subsidized laggards that eventually dropped out of the global technological and investment race. Unless the core technological capabilities of the few leading firms can be reliably replicated, the renewed interest in industrial policy risks sowing the seeds of disappointment as history repeats itself.

Concluding Remarks

The unique strength of this study lies in the granularity of our cost data. To keep the findings as transparent as possible, we have deliberately kept the analysis simple, using descriptive regressions rather than relying on imputations or a structural model. The results suggest that direct measurement can settle questions that the literature has had to approach through pricing models and conduct assumptions.

For the estimation of production functions, three lessons follow. First, the distinction between physical and revenue-based productivity is not the only one that matters. Within physical productivity, the unit in which output is counted, a wafer or a transistor, decides whether technical change appears as stagnation or as Moore's Law. Second, the treatment of capital is a substantive modeling choice rather than an accounting detail. When capital enters at book value, the mechanical rundown of depreciation with age and with utilization is of the same order as the learning and scale effects under study and loads onto them. Third, even data that record vintage, age, and calendar time separately identify their effects only where processes of different vintages coexist. The GPU sample satisfies that condition and the CPU sample does not, which is a warning for any panel in which a technology and its adopters advance in lockstep.

Such an approach has obvious limitations. One is that our cost estimates come from an engineering model rather than from firms' accounts, so their accuracy rests on the model's standing in the industry. Another is that the transistor counts (needed to denominate cost in efficiency units) are disclosed too selectively to support estimation, which is why this version stops at the wafer, die, and chip levels. Obtaining reliable counts is the first item on our revision agenda. A third is that engineering data describe a process rather than a firm, because they omit administration, selling, and managerial capacity (Walters, 1963). Our firm effects therefore measure technological capability rather than overall organizational efficiency. We make no claim about returns to scale at the level of the firm. A fourth is that the CPU sample cannot cleanly separate vintage effects from foundry identity and process age, because its process nodes are nearly foundry-specific; the GPU sample can and carries our vintage result. Finally, our sample covers 2015 to 2025, a decade in which the frontier moved from 28 nm to 3 nm. We cannot speak to earlier eras in which the economics of scaling may have differed.

The natural next step is to connect these costs to the markets they feed. We are assembling product-level data on personal computers and on the chip market itself, which will allow us to trace

cost changes into prices, to estimate demand, and to measure the welfare gains from innovation in the manner of Igami et al. (2026). The facts reported here are a prerequisite for that exercise. They also stand on their own, as the assumptions about vintage effects, learning, and scale that they revise are already embedded in the models used to evaluate tens of billions of dollars of public investment.

Appendix

The following table lists the main datasets and modeling assumptions of the papers we covered in the literature review.

Paper	Main data	Main assumptions / modeling choices
Irwin and Klenow (1994)	• Dataquest quarterly shipments for 32 merchant firms • Average industry selling prices • Seven DRAM generations, 1974–1992	Learning is measured through cost declines with cumulative production; Cournot competition is used to infer costs without firm cost data; demand elasticity is parametric; capacity constraints are assumed not to bind in the baseline; shipment data proxy for production.
Jovanovic and Rousseau (2002)	• Quality-adjusted computer price indexes • NIPA computer sales • Historical electricity and motor-vehicle price/output data • BEA, Census, BLS, Compustat, CRSP, and Moody's data	Capital-goods producers are competitive, so price equals cost; efficiency improves with cumulative experience according to a learning curve; the growth model uses Cobb-Douglas production, exogenous labor growth, and calibrated technology shares; quality adjustment is central and imperfect.
Aizcorbe and Kortum (2005)	• Intel desktop MPU prices from Dataquest and MDR • Intel chip characteristics and transistor counts • Semiconductor price indexes from Grimm and Aizcorbe • MDR investment, capacity, and cost estimates	Price declines are generated by vintage capital rather than learning-by-doing; Moore's Law is embodied in better chip-producing equipment; chip-specific equipment investment is sunk; unit variable cost is constant within a vintage; old chips remain in production while price exceeds variable cost; results are compared under competitive, monopoly, and dominant-firm market structures.

Paper	Main data	Main assumptions / modeling choices
Song (2007)	• MicroDesign Resources CPU data • Intel and AMD quarterly prices and quantities • Product characteristics, 1993Q2–2000Q3 • 30 markets and 320 product observations	Utility is quasilinear and based only on observed CPU characteristics, with no idiosyncratic logit error; 386/486, Cyrix, and Apple CPUs are treated as outside options; baseline market size is total CPU units sold; processor speed interacted with time proxies for cost shifters.
Pillai (2013)	• New Intel microprocessor dataset, 1971–2008 • SPEC and BAPCo performance measures • Intel website and online reports for transistor counts and vintages • MDR prices and Intel annual reports	Technological progress is driven by repeated adoption of higher-quality equipment vintages from upstream semiconductor equipment firms; equipment innovations arrive stochastically; the microprocessor firm adopts when its lag behind the frontier reaches a threshold; demand has constant elasticity in price per quality unit; the firm chooses transistor counts and performance to manage marginal cost.
Bloom et al. (2020)	• Transistor density for Moore's Law • Semiconductor R&D data • Aggregate US productivity and R&D data • Compustat and Census firm data	Idea output is the growth rate of the relevant technology, and research productivity equals idea output divided by effective researchers; R&D spending is deflated by wages to measure effective research labor; baseline assumes no diminishing returns to research effort, with robustness checks allowing diminishing returns.
Bown and Wang (2024)	• SIA and BCG/SIA studies • IC Insights and OECD sources • Government documents and policy announcements • Trade-policy evidence	Primarily a narrative policy essay rather than an estimated model; treats learning-by-doing, agglomeration, national security, and geographic concentration as central rationales for policy; uses industry estimates such as leading-edge fabs costing around $20 billion and US fabs being more costly than East Asian fabs.

Paper	Main data	Main assumptions / modeling choices
Miao (2025)	• Omdia Pure-Play Foundry Market Tracker • SEMI 300mm Fab Outlook • Firm quarterly reports • Gartner, IDC, and Canalys (three sources) for the shipment data	Dynamic oligopoly with step-by-step innovation, capacity investment subsidies, and stochastic trade disruption; five main foundries are modeled; firms far behind the frontier stop upgrading; capacity is a static choice without adjustment costs; disruption risk is known and fixed; trade costs are abstracted from or treated as low.
Goldberg et al. (2026)	• Global Trade Alert/JLOP policy data, 2010–2022 • OECD subsidy evidence • Supplemental estimates of Chinese support • Proprietary GSA wafer-pricing survey, 2004–2019	Production subsidies are inferred as residual country cost advantages in a structural supply model; learning may occur within supplier-technology pairs, across technologies within a foundry, and across borders; country fixed effects can also absorb non-subsidy cost advantages, so subsidy estimates are interpreted cautiously.
Zhuo and Tiew (2026)	• Quarterly GSA wafer-order data, 2004–2015 • SEMI fab-level capacity and investment data, 1995–2015 • IC Insights contract manufacturing revenue • Firm quarterly reports and WSTS sales	Structural contracting model in which buyers solicit bids from manufacturers with available capacity; prices reflect marginal costs and the second-best alternative; costs vary by technology, firm, region, scale, proximity to buyers, and shocks; baseline uses pre-2015 data, before the largest recent policy shifts.

References

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Agarwal, N., Ashlagi, I., Azevedo, E., Featherstone, C. R., and Karaduman, Ö. (2019). Market failure in kidney exchange. *American Economic Review*, 109(11):4026–4070.

Aizcorbe, A. and Kortum, S. (2005). Moore's law and the semiconductor industry: A vintage model. *Scandinavian Journal of Economics*, 107(4):603–630.

Benkard, C. L. (2000). Learning and forgetting: The dynamics of aircraft production. *American Economic Review*, 90(4):1034–1054.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144.

Bown, C. P. and Wang, D. (2024). Semiconductors and modern industrial policy. *Journal of Economic Perspectives*, 38(4):81–110.

Chenery, H. B. (1949). Engineering production functions. *Quarterly Journal of Economics*, 63(4):507–531.

Christensen, L. R. and Greene, W. H. (1976). Economies of scale in U.S. electric power generation. *Journal of Political Economy*, 84(4):655–676.

De Loecker, J. and Syverson, C. (2021). An industrial organization perspective on productivity. In Ho, K., Hortaçsu, A., and Lizzeri, A., editors, *Handbook of Industrial Organization*, volume 4, pages 141–223. Elsevier.

Demirer, M. and Karaduman, Ö. (2024). Do mergers and acquisitions improve efficiency? evidence from power plants. NBER Working Paper 32727.

Foster, L., Haltiwanger, J., and Syverson, C. (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review*, 98(1):394–425.

Goldberg, P. K., Juhasz, R., Lane, N. J., Lo Forte, G., and Thurk, J. (2026). Industrial policy in the global semiconductor sector. Working Paper 32651, National Bureau of Economic Research. Revised February 2026.

Igami, M. (2017). Estimating the innovator's dilemma: Structural analysis of creative destruction in the hard disk drive industry, 1981–1998. *Journal of Political Economy*, 125(3):798–847.

Igami, M. (2018). Industry dynamics of offshoring: The case of hard disk drives. *American Economic Journal: Microeconomics*, 10(1):67–101.

Igami, M., Kusaka, S., Qiu, J., and Tran, T. L. (2026). Welfare gains from product and process innovations: The case of lcd panels, 2001–2011. Manuscript, University of Toronto.

Igami, M. and Subrahmanyam, J. (2019). Patent statistics as an innovation indicator? evidence from the hard disk drive industry. *The Japanese Economic Review*, 70(3):308–330.

Igami, M. and Uetake, K. (2020). Mergers, innovation, and entry-exit dynamics: Consolidation of the hard disk drive industry, 1996–2016. *Review of Economic Studies*, 87(6):2672–2702.

Irwin, D. A. and Klenow, P. J. (1994). Learning-by-doing spillovers in the semiconductor industry. *Journal of Political Economy*, 102(6):1200–1227.

Jovanovic, B. and Rousseau, P. L. (2002). Moore's law and learning by doing. *Review of Economic Dynamics*, 5(2):346–375.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, 70(2):317–341.

Levitt, S. D., List, J. A., and Syverson, C. (2013). Toward an understanding of learning by doing: Evidence from an automobile assembly plant. *Journal of Political Economy*, 121(4):643–681.

Miao, W. (2025). Technology rivalry and resilience under trade disruptions: The case of semiconductor foundries. Job market paper, Stanford University.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297.

Pillai, U. (2013). A model of technological progress in the microprocessor industry. *Journal of Industrial Economics*, 61(4):877–912.

Song, M. (2007). Measuring consumer welfare in the CPU market: An application of the pure-characteristics demand model. *RAND Journal of Economics*, 38(2):429–446.

Walters, A. A. (1963). Production and cost functions: An econometric survey. *Econometrica*, 31(1/2):1–66.

Zhuo, R. and Tiew, A. (2026). Industrial policy, location choice, and firm performance in high-tech manufacturing. Working paper.