The Engineering Reality of Semiconductor Manufacturing Costs: Vintages, Learning, and Scale Economies

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

June 17, 2026

Abstract

As semiconductors dominate recent industrial-policy debates, understanding their underlying economics is critical. In contrast with the existing literature that relies on indirect estimates of semiconductor manufacturing costs, this paper combines comprehensive product specification data with a commercial-grade engineering cost model, constructing a granular panel of over 77,000 CPU and GPU chip-quarter observations from 2015 to 2025. By reverse-engineering the engineering cost model through a series of regressions, we document several findings that challenge standard modeling assumptions: (i) vintage-capital effects are highly heterogeneous and can even increase costs; (ii) physical learning-by-doing confers a modest 1.9% quarterly cost reduction; and (iii) static scale economies flatten above a 70% capacity utilization threshold. We also find accrual-based depreciation accounting to be a critical source of biases that could dominate all of these estimates if not properly isolated from physical, cash-flow-based costs. Finally, we confirm the presence of persistent firm-level heterogeneity, which could be a major bottleneck for industrial subsidies.

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

1 Introduction

With recent advances in information technologies—including the massive investment boom in data centers and artificial intelligence—as well as heightened geopolitical tensions, the economic literature has been enthusiastically revisiting some of the old, perennial questions regarding the nature of semiconductor manufacturing technologies and the benefits and costs of investments and innovations. However, the semiconductor industry is notoriously complicated in terms of both technological details and institutional context, and granular data are difficult to assemble. Hence, much of the existing literature has relied on somewhat aggregate data or micro-level data with

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

limited coverage. A comprehensive analysis of production efficiency and manufacturing costs is lacking, especially one based on direct measurement of costs. This combination of a resurgence of interest in "big picture" questions and limited data availability has resulted in high-stakes policy recommendations and large-scale public funding of technological investments based on potentially unreliable measures and assumptions. This paper aims to fill this critical gap by extensively using detailed data on the actual manufacturing costs of semiconductor devices on a global scale.

We assemble a unique and highly detailed dataset from two primary sources covering the period from 2015:Q1 to 2025:Q2. First, we utilize a comprehensive product specification database containing every mainstream CPU and GPU released by major chip designers, capturing granular attributes such as die size, core counts, foundry choice, and substrate packaging technologies. Second, we input these precise physical parameters into a commercial-grade engineering cost model used extensively by industry professionals to compute cost schedules across all four primary stages of the integrated-circuit supply chain: wafer fabrication, wafer sort and test, packaging and assembly, and final chip test. This combination allows us to construct a robust panel of 50,655 CPU and 26,496 GPU chip-quarter observations with full cost decompositions.

By running parsimonious linear and saturated full-dummy regressions on these direct measures, we reverse-engineer the economic determinants of semiconductor costs and document several key findings that challenge standard modeling assumptions. Our results show: (i) the vintage-capital effect is highly heterogeneous and can even increase manufacturing costs, (ii) the physical learning-by-doing effect is a relatively modest 1.9% per quarter, (iii) static scale economies from capacity utilization largely disappear above a 70% operational threshold.

In our data handling, we make what turns out to be a critical distinction between full accounting costs, including accrual-based depreciation expenses, and a purely cash-flow-based cost measure that isolates them. The comparison between these two cost measures allows us to uncover systematic biases introduced by accrual-based depreciation expenses that are frequently larger than the effects of the economic mechanisms under investigation, such as those highlighted in (i)–(iii) above. We also document substantial and persistent firm-level heterogeneity.

Our paper builds upon and contributes to several strands of the empirical industrial organization and macroeconomics literature. Early foundational work by Irwin and Klenow (1994) and Jovanovic and Rousseau (2002) established classical benchmarks for learning-by-doing (LBD), though they relied on aggregate shipments and industry average selling prices to infer underlying marginal costs through specific market structure assumptions. By contrast, Aizcorbe and Kortum (2005) and Pillai (2013) shifted the analytical focus toward a vintage-capital framework, interpreting rapid price declines as a reflection of technological change embodied in upstream equipment rather than continuous operational learning. More recently, Bloom et al. (2020) highlighted the staggering research burden required to maintain Moore's Law, demonstrating that the exponential growth in transistor density has demanded an ever-increasing influx of research input.

As semiconductors have taken center stage in modern geopolitical strategies, a new wave of policy-oriented research has emerged to evaluate state interventions. Bown and Wang (2024) outline

the geopolitical logic behind supply-chain diversification and the enormous fixed costs associated with modern fabrication facilities. Structural econometric works by Miao (2025), Goldberg et al. (2026), and Zhuo and Tiew (2026) have attempted to evaluate the efficacy of capacity subsidies, international learning spillovers, and fab location decisions. While these papers push the frontier of policy analysis using wafer-pricing surveys, contract data, and capacity trackers, they must still infer or back out cost parameters indirectly through pricing models or bargaining games. Our paper complements this literature by examining the direct, underlying physical and accounting costs inside the fabs.

The remainder of this paper is organized as follows. Section 2 reviews some of the important papers in the literature, with a specific focus on their main data and modeling assumptions. Section 3 describes the technical supply-chain background and the construction of our highly detailed product-specification and manufacturing-cost datasets, as well as the definitions of the four cost measures. Section 4 details our linear and non-linear regression models and reports the estimation results for both CPU and GPU production. Section 5 concludes by summarizing our findings into five main takeaways and discussing their broader implications.

2 Related Literature

We have assembled an incomplete list of related papers on semiconductor manufacturing. The type and granularity of data significantly vary, but none of them seems to use detailed cost data as we do in this paper. The Appendix provides further details about the main datasets and modeling assumptions of each paper.

Irwin and Klenow (1994). Irwin and Klenow provide the classic empirical benchmark on semiconductor learning-by-doing spillovers. The paper uses unpublished Dataquest quarterly data on average industry selling prices and shipments by 32 merchant firms for seven DRAM generations from 1974–1992. Because firm-level cost data are unavailable, the authors infer marginal costs using a Cournot model and parametric demand elasticities. The central learning variables are own cumulative production, other firms' cumulative production in the same country, and other firms' cumulative production abroad. The paper assumes shipment data are a close proxy for production and that capacity constraints do not bind in the baseline. The main findings are that DRAM learning rates average about 20 percent, own experience matters about three times as much as others' experience, spillovers are international rather than purely national, Japanese firms do not learn faster than others, and spillovers across DRAM generations are weak.

Jovanovic and Rousseau (2002). Jovanovic and Rousseau model Moore's Law as an Arrow-style learning-by-doing process in which computer producers become more efficient as cumulative production rises. The data combine quality-adjusted computer price indexes, NIPA computer sales, electricity and motor-vehicle historical price/output series, and firm-level market-to-book

data. The main cost assumption is competitive capital-goods production: the price of computers equals unit cost, so a falling quality-adjusted price reveals rising production efficiency. The model uses a learning curve in cumulative output, Cobb-Douglas production, exogenous labor growth, and calibrated capital shares. The paper estimates a much steeper learning elasticity for computers than for electricity or motor vehicles, then uses the model to argue that rapid computer price decline can imply fast growth and faster obsolescence of firms using old capital.

Aizcorbe and Kortum (2005). Aizcorbe and Kortum develop a vintage-capital model of the microprocessor industry in which price declines of individual chips are not primarily caused by learning-by-doing. Instead, newer chips enter with better embodied equipment vintages, and older chips continue to be produced for some time because chip-specific production equipment is a sunk investment. The empirical calibration uses Intel desktop microprocessor prices from Dataquest for 1985–1993 and MicroDesign Resources for 1993–1999, Intel chip characteristics and transistor counts, semiconductor price indexes, and MDR estimates of investment, capacity, and unit costs. The model treats Moore's Law as exogenous quality improvement embodied in chip-producing equipment. It assumes simple demand over quality-weighted chips, sunk setup costs, constant unit variable cost within a vintage, and old-vintage operation until price falls to variable cost. The main implication is that faster Moore's Law should shorten product lives and steepen price declines, even without learning economies.

Song (2007). Song estimates demand for Intel and AMD CPUs and measures how much consumers benefited from rapid quality improvement and falling prices in the 1990s. The core data are quarterly MDR product-level observations on prices, quantities, processor speed, cache, and related characteristics from 1993Q2 to 2000Q3. The main assumption is that CPU buyers value observed product characteristics directly, so the model omits the idiosyncratic logit error common in differentiated-products demand. Song also assumes the CPU is the key characteristic determining computer demand and that manufacturers' CPU demand reflects final consumer preferences. The paper treats older or missing CPU classes as outside options, uses total CPU units sold as the baseline market size, and uses processor-speed-by-time variables as cost proxies for price endogeneity. The main result is that consumer surplus accounts for most social surplus, especially because new high-end processors delivered large quality gains without proportional price increases.

Pillai (2013). Pillai extends the vintage-capital view by modeling technological progress in microprocessors as an optimal adoption problem. The paper builds a new Intel microprocessor dataset from 1971–2008 with product names, release dates, performance, capital-equipment vintage, and transistor counts; after excluding Celeron and server processors, the core Intel sample contains 453 products. Performance comes from SPEC and BAPCo, while transistor counts and vintages come from Intel and related online sources; prices are drawn from MicroDesign Resources and financial information from Intel annual reports. The model assumes that upstream semiconductor equipment firms generate better equipment vintages and that a microprocessor firm adopts only when the gap

between its current equipment and the frontier reaches a threshold. The paper decomposes technological progress into upstream equipment innovation and the downstream firm's efficiency in using additional transistors, arguing that the 1990s acceleration reflected faster equipment innovation while the post-2000 slowdown reflected weaker design efficiency.

Bloom et al. (2020). Bloom, Jones, Van Reenen, and Webb ask whether constant exponential growth requires ever more research effort. Their Moore's Law case measures idea output as the growth rate of transistor density and research input as semiconductor R&D spending deflated by researcher wages. The core assumption is an idea production function in which growth equals research productivity times effective researchers. In the baseline, research effort enters linearly; robustness checks allow diminishing returns. The headline semiconductor result is that the number of researchers needed to sustain the historical doubling of chip density rose by more than a factor of 18 from the early 1970s to the 2010s, implying a sharp decline in research productivity. In contrast to Jovanovic and Rousseau's learning interpretation, this paper emphasizes the rising research burden needed to keep Moore's Law moving.

Bown and Wang (2024). Bown and Wang provide a policy-oriented survey of why semiconductors have become central to modern industrial policy. The main topic is not an estimated causal model but the economic and geopolitical logic behind subsidies, export controls, and supply-chain diversification. The paper draws on industry factbooks, SIA and BCG/SIA estimates, IC Insights rankings, OECD work, government policy documents, and trade-policy evidence. The key maintained ideas are that fabs have enormous fixed costs, learning-by-doing and agglomeration may create external benefits, and geographic concentration of advanced production in East Asia creates resilience and national-security concerns. For the cost discussion, it emphasizes industry estimates that a leading-edge fab can cost around $20 billion and that U.S. fabrication may be materially more expensive than fabrication in Taiwan, South Korea, or China.

Miao (2025). Miao studies when U.S.-style capacity subsidies are optimal if semiconductor trade may be disrupted. The main data are Omdia's pure-play foundry panel, SEMI's 300mm Fab Outlook, firm reports, and complementary market-shipment data. The paper uses these data to characterize foundry concentration, node-level innovation, price declines, investment costs, and cost structure. The model is a dynamic oligopoly in which firms innovate one technology step at a time, choose where to install capacity, and then compete under either integrated trade or disrupted trade. Capacity subsidies lower investment costs; the U.S. cost of building capacity is calibrated to be higher than East Asia. Important simplifying assumptions are that capacity is reoptimized statically, trade disruption risk is fixed and known, trade costs are otherwise small, and firms that fall several generations behind no longer upgrade. The main policy result is that low disruption risk implies low optimal U.S. subsidies, while a sufficiently high disruption risk can rationalize a 25 percent investment subsidy similar to the CHIPS Act investment tax credit.

Goldberg et al. (2026). This paper asks how to measure semiconductor industrial policy and whether subsidies can be justified when learning-by-doing and international spillovers are present. Its empirical contribution is twofold: it documents recent semiconductor policies using policy-text data and estimates a model using proprietary GSA wafer-order data. The key data are the Global Trade Alert/JLOP industrial policy database, hand-verified semiconductor policy records, supplemental estimates of Chinese national and provincial support, and the GSA Semiconductor Wafer Pricing Survey covering 2004–2019. The model treats government support as lowering marginal cost and uses the residual country cost component as a proxy for unobserved production subsidies. Its central assumptions are that demand and supply can recover marginal costs from wafer prices and quantities, that learning can be decomposed into supplier-technology learning, economies of scope across technologies, and cross-border spillovers, and that country fixed effects are informative about subsidy intensity after controlling for observed cost differences. The main takeaway is that learning-by-doing is sizable, cross-border spillovers are important, and subsidies may have positive international effects when they accelerate cost reductions that diffuse through the fabless-foundry ecosystem.

Zhuo and Tiew (2026). Zhuo and Tiew focus on the location decision of contract semiconductor manufacturers, especially whether a major foundry should locate a large fab in the United States rather than its home region. Their data combine GSA order-level wafer contracts, SEMI fab-level investment and capacity data, IC Insights revenue data, firm quarterly reports, and WSTS aggregate sales. The estimated model is a structural bargaining/auction model: buyers solicit bids from feasible manufacturers, manufacturers bid based on marginal cost, and prices reflect the best or second-best manufacturing option adjusted for regional preferences. Cost parameters capture technology, firm, region, capacity scale, proximity to buyers, and shocks. The paper assumes pre-2015 market structure is a useful baseline for current counterfactuals and then layers policy scenarios on top. Its central quantitative finding is that locating in the United States would require roughly $1.2 billion more investment than a home-region fab, while long-run flow-profit losses from higher U.S. production costs may be much larger than one-time subsidies.

3 Data

3.1 Technological Background and Datasets

We assemble the dataset from two main sources: (i) a publicly available online database containing specifications of every single mainstream CPU/GPU released by major chip designers, and (ii) the detailed estimates of the CPU/GPU manufacturing costs based on a proprietary engineering model. The former supplies critical details that enable us to fill the parameters of the latter engineering model, such as quarter of release, foundry choice, die size, number of cores, and substrate specifications to generate engineering estimates of each individual chip product at each stage of the semiconductor manufacturing supply chain.

Industry Context. We obtain cost estimates of each CPU/GPU on the four production stages in the integrated-circuit (IC) supply chain: wafer fabrication, wafer sort and test, packaging and assembly, and chip test. Integrated circuits are first patterned onto 300 mm-diameter silicon wafers during the wafer fabrication process, in which successive layers of materials are deposited, exposed to lithography, and etched to form multiple dies (pre-packaged chips) on a single wafer. Chip designers that outsource wafer manufacturing to foundries (TSMC, GlobalFoundries, Samsung Foundry) have to pay an additional margin on the raw wafer manufacturing cost that depends on the designer's volume. For integrated device manufacturers (IDMs) such as Intel, they do not pay the margin for wafers if they manufacture in-house.

The manufactured wafer then undergoes wafer sort and test to verify its quality. Each die is electrically probed while still on the wafer to screen out defective ones. The defect density of the wafer manufacturing process, along with the die size of the chip, affects the yield, which decreases with defect density and size. The larger the individual die, the more likely any given die will be exposed to random defects on the surface. The wafer is then diced into individual dies, with the functional dies proceeding to packaging and assembly. In the packaging process, the dies are attached to a package substrate or lead frame, connected through wirebond or flipchip interconnects, and encapsulated to protect the device and provide external electrical connections. Intel does its packaging in-house; fabless chip designers rely on other assembly & testing partners such as the ASE Group, Amkor, and SPIL. Finally, packaged chips undergo chip test, a final electrical and functional testing stage that verifies performance and reliability.

Database 1: Product Specifications. We obtain product attributes from the first database, a public reference database widely cited by hardware reviewers and industry analysts for CPU and GPU specifications. The original cleaned dataset contains 4,169 unique CPUs and 3,310 unique GPUs released between the late 1990s and 2025. For each chip, we observe its release date and production status; the manufacturing process (process node $\times$ foundry); the die size in millimeters squared; the number of cores; the packaging method, among other performance specs such as clock speed. After restricting to chips released during/around the coverage period of the engineering cost model (2015:Q1–2025:Q2), 2,836 CPUs and 1,788 GPUs are covered.[1]

Database 2: Manufacturing Costs. We use the product specifications from Database 1 as inputs into the engineering cost model, which is widely used as a key reference by foundries and fabless designers in the semiconductor industry. The model is a deterministic engineering-accounting workbook designed to replicate the unit cost of any chip on any process node between 2015:Q1

[1]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 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.

and 2025:Q2. As inputs, the user specifies the year and quarter, the foundry manufacturing process, die height and width, product type (number of cores), the volume tier of the design firm, the assembly location, and packaging & assembly specs: the package type, the length and width of the substrate, substrate layer counts, connection type (wirebond or flipchip), and connection count.

The model then computes the following:

1. **The wafer fabrication yield and cost**, separated by seven line items: the initial wafer, direct labor, depreciation, equipment maintenance, indirect labor, facilities, and consumables. Cost per wafer will depend on the foundry's utilization rate. The initial wafer, direct and indirect labor are variable costs that do not depend on the foundry's utilization. Depreciation is a fully fixed cost, per-wafer depreciation cost decreases with utilization. Equipment maintenance, facilities, and consumables have a fixed component and a variable component that increases with utilization. We merge the panel with the quarterly utilization rate (from yet another database) observed for each foundry and IDM to produce full wafer costs that vary according to the foundry process and utilization.

   The full wafer cost is then divided by the model's wafer yield, the probability that the whole wafer is intact after fabrication, to arrive at the yielded wafer cost, which is the expected cost per wafer for the chip designer. A foundry margin is then applied to the yielded wafer manufacturing cost to arrive at a wafer price. Margin is set to zero for IDM processes and varies by quarter, process node, and foundries for foundry processes.

2. **The gross and net die per wafer at the wafer test stage.** Gross dies represent the maximum number of dies that could be made on a 300mm wafer, which depends on the die size specifications. The net die is the gross die count multiplied by the die yield, which depends on the defect density of the foundry process in the quarter. In the engineering cost model, die yield generally improves within a process across time as the manufacturer gains experience.

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

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

The output of the cost model is a (chip, year-quarter)-level panel with the full cost decomposition at each of the four supply-chain stages. The CPU panel contains 50,655 chip-quarter observations; the GPU panel contains 26,496.

3.2 Definition of Four Cost Measures

We have four measures of manufacturing cost for each CPU/GPU product j at fab/process k in year-quarter t at multiple stages of production. First, the 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. As its definition clarifies, this measure removes the 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.[2] 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.

[2] In the case of integrated device manufacturers (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 IDMs.

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

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

As the definition suggests, this final cost accounts for several additional costs associated with the packaging-and-assembly step and the final classification (or "class") tests.

4 Determinants of Manufacturing Cost

What determines the cost of manufacturing semiconductor devices? Our cost data are generated by a deterministic engineering model, and we can effectively reverse-engineer its mechanics by running regressions.

4.1 Specifications of the Regression Model

For each of the four cost measures we introduced in the above (here generically denoted by C_{jkt} for product j at fab k in calendar quarter t), we estimate a linear model,

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

as well as its nonlinear version, in which we replace each RHS variable with a full set of dummies.

Definitions. We briefly explain the definition of all RHS variables here. Their economic interpretations will be discussed subsequently, alongside the coefficient estimates.

The first three regressors capture the process-level characteristics of fab k. Gen_k is the technological generation (process node).[3] Age_{kt} is the number of calendar quarters since the beginning of volume production. $Capa_{kt}$ is capacity utilization in percentage points.

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)

[3] 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.

in thousands. $Substrate_j$ is a vector of substrate (the foundational layer that connects the die to the larger motherboard) specifications, including (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 $Substrate_j$ represents the architecture and complexity of CPU/GPU packaging, it is absent from the analysis of the wafer- and die-level costs (C^{Full}_{jkt}, C^{Cash}_{jkt}, and C^{Die}_{jkt}) and becomes relevant only for 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 (manufacturers who own and operate fabs, indexed by f) and fablesses (buyers of manufacturing services who design and sell chips, indexed by l), respectively.[4] 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 additional nonlinearity and interactions.

4.2 Central Processing Units (CPUs)

Table 1 reports the results for CPU manufacturing costs based on the sample of 48,158 fab-product-time observations for which all LHS and RHS variables are available. The four measures of cost and the two specifications are organized across eight columns; the regressors are presented across rows in five groups (fab, product, manufacturer, buyer, and time).

One of the most salient results is that the fit is extremely high. Even the simple linear specification achieves the adjusted R^2 of 0.787 (column 1) or above, and all of the full-dummy specifications entail 0.952 (column 6) or above. Hence, 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 vintage-capital or learning-by-doing effect for different combinations of products and manufacturers).

Process Generation. The linear regression in column 1 shows that, on average, full wafer cost (C^{Full}_{jkt}) declines by $\left|e^{-0.095}-1\right| \times 100 = 9.1\%$ with an advance in technological generation of fab equipment. However, the estimate for cash wafer cost (C^{Cash}_{jkt}) in column 3, which entirely ignores depreciation, is an order of magnitude smaller: $\left|e^{-0.003}-1\right| \times 100 = 0.3\%$. Hence, the measurement of the vintage-capital effect is highly sensitive to the treatment of accrual-based depreciation cost. Meanwhile, columns 5 and 7 show that the vintage effect *increases* cost at the level of dies and chips, which is confusing.

The estimates from the non-linear full dummy specification in Figure 1 (A) explain why the linear estimates are mixed up: some vintages decrease costs while others increase them. Thus, unlike other manufacturing technologies in the IT hardware sector, such as those for hard disk

[4] 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.

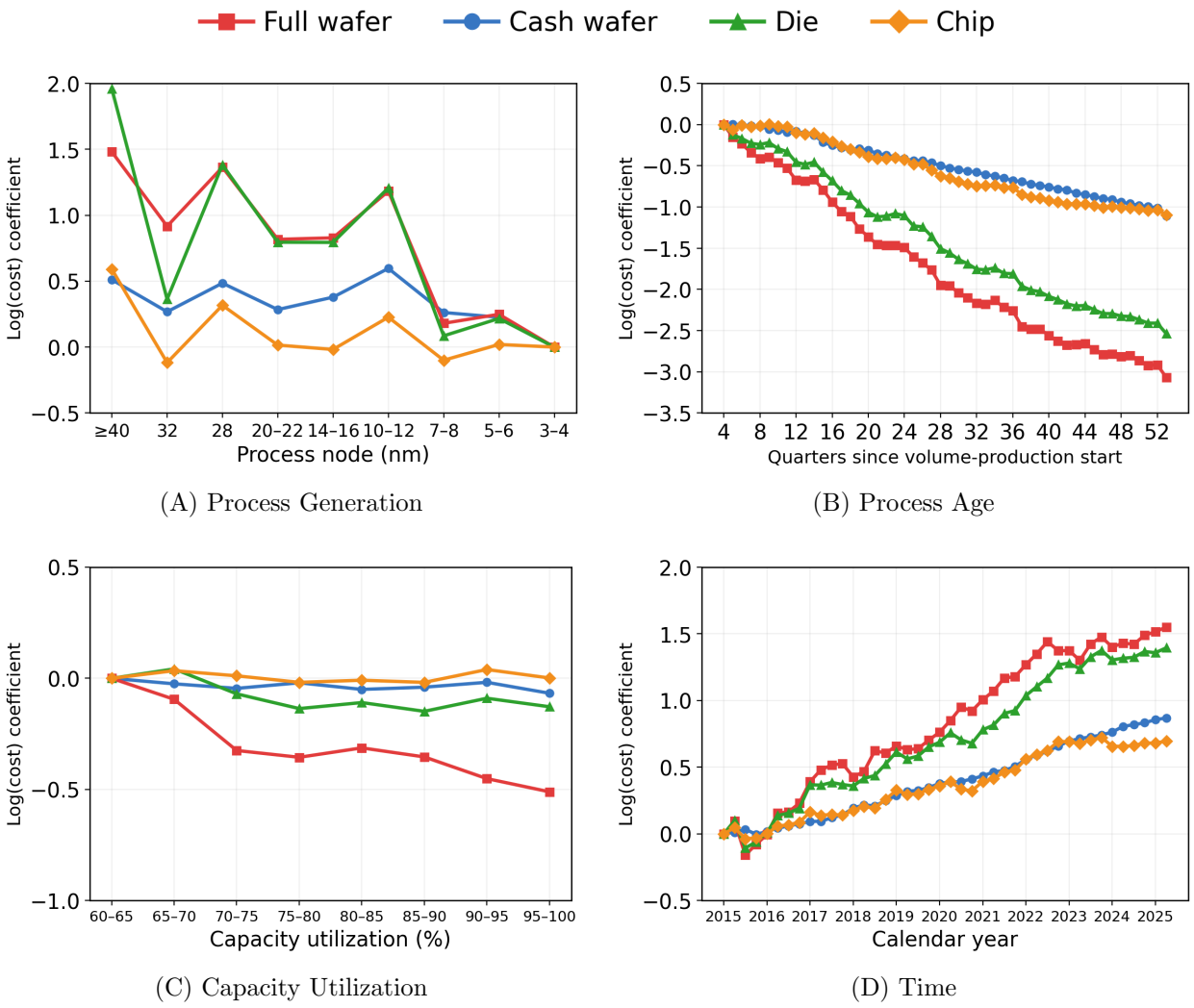
Table 1: Determinants of Manufacturing Cost (CPU)

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)	−0.095	–	−0.003	–	0.165	–	0.013	–
	(0.005)	(–)	(0.002)	(–)	(0.004)	(–)	(0.003)	(–)
Process age (quarters)	−0.066	–	−0.031	–	−0.030	–	−0.025	–
	(0.001)	(–)	(0.001)	(–)	(0.001)	(–)	(0.001)	(–)
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)	0.014	–	0.014	–	0.546	–	0.266	–
	(0.001)	(–)	(0.000)	(–)	(0.002)	(–)	(0.001)	(–)
Number of cores (1,000)	0.682	–	0.250	–	1.323	–	2.851	–
	(0.086)	(–)	(0.040)	(–)	(0.124)	(–)	(0.147)	(–)
Substrate size (100 mm^2)	–	–	–	–	–	–	0.008	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.000)	(–)
Substrate core layers	–	–	–	–	–	–	0.169	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.003)	(–)
Substrate buildup layers	–	–	–	–	–	–	−0.028	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.002)	(–)
Connections (1,000)	–	–	–	–	–	–	0.142	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.002)	(–)
Flipchip Connection	–	–	–	–	–	–	−0.004	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.004)	(–)
C. Manufacturer								
Foundry: Intel	0.308	0.563	0.317	0.346	−0.185	0.584	−0.107	0.106
	(0.009)	(0.007)	(0.005)	(0.003)	(0.009)	(0.010)	(0.006)	(0.007)
Foundry: TSMC	−0.445	0.424	−0.318	0.098	−0.425	0.913	−0.014	0.267
	(0.009)	(0.026)	(0.004)	(0.011)	(0.009)	(0.046)	(0.006)	(0.053)
D. Buyer specs								
Volume: High	0.060	−0.030	0.291	0.015	−0.114	0.028	−0.112	0.197
	(0.006)	(0.012)	(0.002)	(0.003)	(0.005)	(0.015)	(0.005)	(0.025)
Volume: Low	0.734	−0.461	0.767	−0.127	0.800	−0.828	−0.079	−0.129
	(0.020)	(0.036)	(0.009)	(0.012)	(0.033)	(0.051)	(0.017)	(0.060)
E. Time and constant								
Time	0.054	–	0.033	–	0.019	–	0.021	–
	(0.001)	(–)	(0.000)	(–)	(0.001)	(–)	(0.001)	(–)
Constant	−425.537	7.908	−255.819	6.884	−153.223	0.852	−166.385	2.785
	(7.522)	(0.043)	(3.972)	(0.017)	(6.099)	(0.082)	(4.708)	(0.067)
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	0.787	0.963	0.888	0.991	0.860	0.952	0.933	0.972
Adjusted R^2	0.787	0.963	0.888	0.991	0.860	0.952	0.933	0.972

Note: The CPU sample consists of 48,158 fab-product-time 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.

drives (HDDs) and liquid crystal display (LCD) panels, technological advances in semiconductor manufacturing do not necessarily reduce costs in every new generation.

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



(A) Process Generation

(B) Process Age

(C) Capacity Utilization

(D) Time

Process Age. The elapsed time since the start of volume production at each fab/process is a good measure of yield engineering, encapsulating continuous efforts of process/production engineers to fine-tune the "parameters" of the equipment and overall production processes through trial and error. Because any reasonably well-managed fab in this industry would operate at a relatively high rate of capacity utilization, the age becomes the primary indicator of accumulated experience or learning-by-doing (LBD). It is exactly this variable (age) that completely determines the yield improvement in the inner working of the engineering cost model underlying our data.

The linear estimate of -0.066 in column 1 implies that C_{jkt}^{Full} decreases by $\left|e^{-0.066}-1\right| \times 100 = 6.4\%$, but column 3 reports a much smaller coefficient estimate (in absolute value) of -0.031 for

C^{Cash}_{jkt}. This discrepancy again reflects depreciation accounting, which exaggerates the impact of age because depreciation cost mechanically declines over time.[5] Thus, $\left|e^{-0.031}-1\right| \times 100 = 3.1\%$ is closer to the physical reality of the LBD effect at the wafer level. Figure 1 (B) reports the full-dummy version of the age effects, in which C^{Full}_{jkt} and C^{Die}_{jkt} exhibit much faster decline than C^{Cash}_{jkt}.[6] The coefficient estimate of the quarter-52 dummy is approximately -1 for both C^{Cash}_{jkt} and C^{Chip}_{jkt}, which implies an average quarterly cost reduction of 1.9%.

Capacity Utilization. The linear estimates in columns 1 and 3 suggest a one-percentage-point increase in capacity utilization reduces C^{Full}_{jkt} and C^{Cash}_{jkt} by 1% and 0.2%, respectively. As we discussed regarding process generation and age, the inclusion of depreciation costs likely inflates the full-cost result. Hence, we take the latter, cash-cost estimate as a better reflection of the physical reality of scale economy. The full-dummy estimates in Figure 1 (C) show qualitatively similar patterns with granular details.

Product Specs. Die/chip-level product characteristics affect costs in mostly predictable manners. Die size has a relatively small impact on wafer-level costs ($\left|e^{0.014}-1\right| \times 100 = 1.4\%$ per $100\,\text{mm}^2$ for both C^{Full}_{jkt} and C^{Cash}_{jkt}), which suggests how to slice and dice wafers 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, die size mechanically alters the number of dies/chips per wafer, as larger dies mean fewer dies/chips per wafer of a given size. Second, the larger the die size, the higher the defect probability because defects arise stochastically at a certain probability (or density) per unit surface area. The linear estimates suggest a $100\,\text{mm}^2$ increase in die size increases C^{Die}_{jkt} and C^{Chip}_{jkt} by $\left|e^{0.546}-1\right| \times 100 = 72.6\%$ and $\left|e^{0.266}-1\right| \times 100 = 30.5\%$, respectively. The number of cores is measured in 1,000 units to ensure readability of the results for both CPUs and GPUs on the same scale. The estimates in Table 1 confirm that additional cores come at a cost. For example, C^{Chip}_{jkt} increases by $\left|e^{2.851/1000}-1\right| \times 100 = 0.285\%$ per core.

Manufacturers. The linear estimates suggest TSMC enjoys significant cost advantages over GlobalFoundries (the omitted category) across all measures, but the full-dummy estimates show the opposite. Our interpretation is that a foundry firm with frontier processes, such as TSMC, specializes in high-end products, which may lead to counter-intuitive results due to interactions and collinearities. Broadly similar patterns emerge for Intel as well, even though the linear estimates for C^{Full}_{jkt} and C^{Cash}_{jkt} suggest consistent cost *dis*advantages relative to GlobalFoundries at the wafer level.

Buyers. Manufacturers tend to offer discounts to high-volume orders, and our engineering cost model incorporates this element. The estimates do not always confirm this mechanism, however,

[5] 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.

[6] Curiously, C^{Chip}_{jkt} shows a trajectory similar to C^{Cash}_{jkt} despite its inclusion of depreciation cost.

and our interpretation is that many of the high-volume buyers (e.g., Apple and Nvidia) tend to order high-spec products, creating nuanced interactions and collinearities.

Time. Both the linear and full-dummy versions of the estimates show generally positive time trends, consistent with the secular industry-wide inflation in the cost of materials and intermediate inputs. Based on the linear estimate, the rate is higher at the wafer level (C_{jkt}^{Full} and C_{jkt}^{Die}) with $\left|e^{0.054}-1\right|\times 100 = 5.5\%$ and $\left|e^{0.033}-1\right|\times 100 = 3.4\%$ per quarter, respectively. The die/chip-level inflation is slower at 1.9% and 2.1% for C_{jkt}^{Die} and C_{jkt}^{Chip}, respectively. This ranking partially reverses in the non-linear estimates of Figure 1 (D), where C_{jkt}^{Full} and C_{jkt}^{Die} exhibit faster inflation than C_{jkt}^{Cash} and C_{jkt}^{Chip}, which mirrors the rank order in Figure 1 (B).

4.3 Graphics Processing Units (GPUs)

The cost-function estimates for GPUs in Table 2 are broadly similar to the ones for CPUs, but there are a few exceptions.

Process Generation. The linear estimates in columns 1, 3, 5, and 7 are all positive—between 0.033 and 0.149, or 3.4% and 16.1% per generation. Taken at their face value, these numbers suggest the manufacturing costs for GPUs tend to increase with newer generations of fab equipment on average. However, the full-dummy estimates in Figure 2 (A) show highly non-monotonic patterns reminiscent of their CPU counterparts. Hence, drawing a general conclusion about the sign of the vintage effects seems difficult. The only robust pattern common to both CPUs and GPUs is that cash wafer cost (C_{jkt}^{Cash}) is much less sensitive to capital vintages than the other three cost measures that incorporate depreciation.

Process Age. The impact of process age on GPU costs is uniformly smaller than that on CPU costs. However, the full-dummy estimates in Figure 2 (B) suggest the effects are roughly comparable in magnitude after we flexibly control for all major factors. Another common pattern between CPUs and GPUs is that cash wafer cost (C_{jkt}^{Cash}) exhibits a lower rate of LBD than the other three measures at $\left|e^{-0.011}-1\right|\times 100 = 1.1\%$ per quarter.

Capacity Utilization. The impact of capacity utilization is similar to the CPU estimates and virtually identical for cash wafer cost at -0.003.

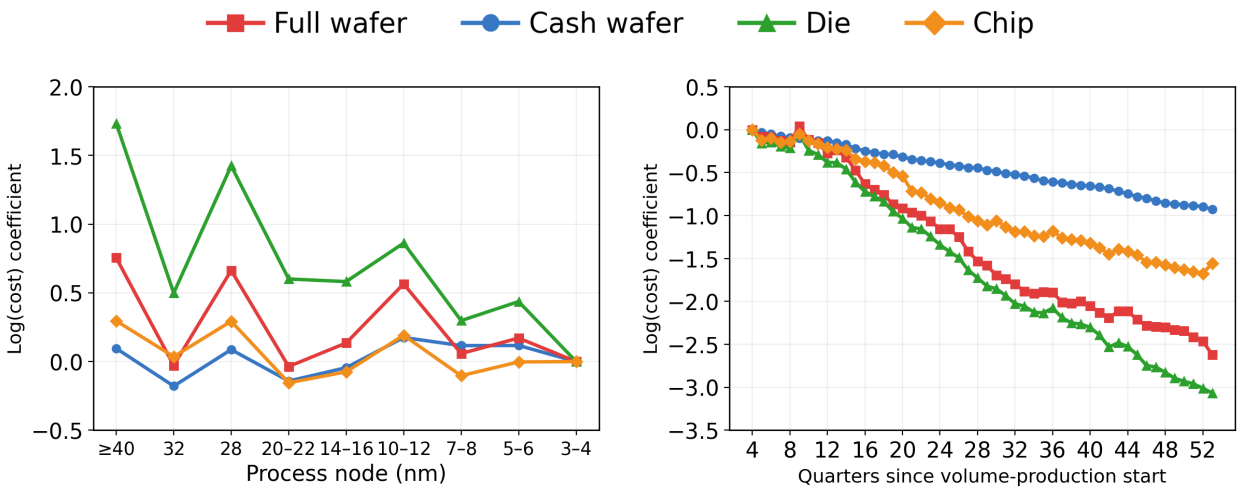
Product Specs. Improved product characteristics usually entail higher costs, but larger die size is correlated with lower wafer costs (C_{jkt}^{Full} and C_{jkt}^{Cash}) in the case of GPUs, which warrants further investigation. Meanwhile, die and chip costs for GPUs scale with die size at approximately the same rates as in CPUs at $\left|e^{0.487}-1\right|\times 100 = 62.7\%$ and $\left|e^{0.276}-1\right|\times 100 = 31.8\%$ per $100\,\text{mm}^2$, respectively. The number of cores is positively correlated with three of the four cost measures, with C_{jkt}^{Die} showing the opposite tendency (-1.3% per thousand cores). Because GPUs carry many small

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)	0.118	–	0.033	–	0.108	–	0.149	–
	(0.005)	(–)	(0.002)	(–)	(0.009)	(–)	(0.006)	(–)
Process age (quarters)	−0.027	–	−0.011	–	−0.028	–	−0.008	–
	(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 mm^2)	−0.023	–	−0.004	–	0.487	–	0.276	–
	(0.001)	(–)	(0.000)	(–)	(0.004)	(–)	(0.003)	(–)
Number of cores (1,000)	0.027	–	0.008	–	−0.013	–	0.017	–
	(0.001)	(–)	(0.000)	(–)	(0.003)	(–)	(0.002)	(–)
Substrate size (100 mm^2)	–	–	–	–	–	–	0.024	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.001)	(–)
Substrate core layers	–	–	–	–	–	–	0.036	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.003)	(–)
Substrate buildup layers	–	–	–	–	–	–	0.004	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.002)	(–)
Connections (1,000)	–	–	–	–	–	–	0.080	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.006)	(–)
Flipchip Connection	–	–	–	–	–	–	−0.095	–
	(–)	(–)	(–)	(–)	(–)	(–)	(0.004)	(–)
C. Manufacturer								
Foundry: Intel	0.052	0.440	0.246	0.308	0.427	1.036	0.200	0.966
	(0.013)	(0.020)	(0.005)	(0.006)	(0.022)	(0.022)	(0.016)	(0.044)
Foundry: Samsung	−0.315	−0.087	−0.050	−0.093	0.106	0.231	0.062	0.075
	(0.007)	(0.011)	(0.002)	(0.003)	(0.012)	(0.013)	(0.010)	(0.015)
Foundry: TSMC	−0.071	−0.102	−0.048	−0.132	0.480	0.331	0.373	0.249
	(0.004)	(0.007)	(0.002)	(0.002)	(0.008)	(0.008)	(0.006)	(0.012)
D. Buyer specs								
Volume: High	−0.007	−0.021	−0.032	−0.010	−0.090	−0.065	−0.061	−0.044
	(0.003)	(0.002)	(0.001)	(0.001)	(0.006)	(0.004)	(0.004)	(0.006)
Volume: Low	0.090	0.025	0.028	−0.006	0.535	0.466	0.335	0.294
	(0.013)	(0.007)	(0.001)	(0.001)	(0.039)	(0.016)	(0.029)	(0.013)
E. Time and constant								
Time	−0.004	–	0.015	–	−0.002	–	−0.017	–
	(0.001)	(–)	(0.000)	(–)	(0.001)	(–)	(0.001)	(–)
Constant	37.392	8.464	−112.713	7.336	16.571	1.238	139.196	3.325
	(6.227)	(0.062)	(2.310)	(0.019)	(11.151)	(0.071)	(7.720)	(0.088)
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	0.785	0.929	0.954	0.989	0.888	0.976	0.912	0.974
Adjusted R^2	0.785	0.928	0.954	0.989	0.888	0.976	0.912	0.974

Note: The GPU sample consists of 22,860 fab-product-time 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

cores for massively parallel computation, the cost of each core is an order or two smaller for GPUs relative to CPUs.

Manufacturers. TSMC exhibits cost advantages over its rivals at the wafer level, as exemplified by the full-dummy estimate for C_{jkt}^{Cash} of $\left|e^{-0.132}-1\right| \times 100 = 12.4\%$ relative to GlobalFoundries (the reference category). Samsung is closely following TSMC with $\left|e^{-0.093}-1\right| \times 100 = 8.9\%$. However, these leading manufacturers' costs are significantly *higher* than GlobalFoundries at the die and chip levels, presumably because they specialize in higher-end products.

Buyers. High-volume purchases translate into lower costs across all results. By the same token, low-volume buyers face higher costs; column 4 is the only exception (-0.006).

Time. The linear estimates feature predominantly negative time trends, but Figure 2 (D) shows clearly positive trends in the full-dummy estimates. The latter, positive time effects are more consistent with the known industry context of general cost inflation. This reversal of the sign between the two functional forms suggests the presence of potentially severe collinearities between multiple factors that vary across time.

5 Findings, Implications, and Comparison with the Literature

Our comparative analysis of CPU and GPU manufacturing costs generates at least five insights into the engineering reality and the economics of semiconductor technology. While many of the concepts highlighted below have been extensively discussed in the existing literature, our findings suggest the need to recalibrate the assumptions researchers impose when modeling and estimating every one of them in an effort to measure the benefits and costs of innovation in this industry.

Finding 1: Vintage Capital Effects Are Highly Heterogeneous across Generations and Can Be Positive or Negative. How capital vintage affects productivity is highly heterogeneous across technological generations. Even the sign of the vintage-capital effect varies, meaning that while some upgrading of fab equipment reduces costs, other upgrades actually increase them. The immediate implication of this finding is that we cannot impose any a priori assumptions on either the sign or the magnitude of this effect across multiple generations—doing so will instantly detach the model from reality and push all results into the realm of science fiction.

This finding forces us to carefully contextualize the vintage literature. First, regarding the semiconductor industry, Aizcorbe and Kortum (2005) and Pillai (2013) proposed to interpret the historical declines of prices as a reflection of technological change embodied in newer equipment vintage, rather than learning-by-doing within a given vintage. This view is difficult to reconcile with our detailed cost data from 2015 to 2025, even though it might have been true in the periods preceding our sample.

Second, Igami et al. (2025) use similarly detailed engineering cost data from the LCD panel industry between 2001 and 2011 and find that most of the new generations of fab equipment led to cost reductions. The contrast between their work and our results cautions against any cross-industry generalizations, even within the same overall sector of IT hardware manufacturing that shares many institutional and technological features.

Finding 2: Physical Learning-by-Doing Effects Are Approximately 1.9% per Quarter. The literature has placed considerable emphasis on learning-by-doing (LBD), but its empirical measurement has mostly relied on highly aggregated data on prices and shipment volume, frequently combined with assumptions on firms' competitive conduct that are neither tested nor verified. Our results suggest LBD contributes to lower physical/cash costs at the average quarterly rate of 3.1% (CPUs) and 1.1% (GPUs) in the linear specification, and approximately 1.9% in the full-dummy specification. Based on the commercial-grade engineering cost model and the extremely high fit of our regressions, we control for virtually everything that affects these engineering cost estimates, including the characteristics of fabs, products, firms, and time.

The fact that we find visible discrepancies between the linear and full-dummy estimates suggests the presence of complicated interactions and collinearities that are difficult to address unless a researcher uses similarly detailed cost data. These observations lead us to regard the estimates from the existing literature strictly on a best-effort basis under severe data limitations.

Finding 3: Static Scale Economies Are Capped and Flatten above 70% Capacity Utilization. In the existing literature, economies of scale are invoked to explain the concentrated market structure of the semiconductor industry as well as to motivate state interventions. Formally, it is modeled as a cost function that decreases with output. Note that this concept often serves as an umbrella category, nesting LBD as its dynamic manifestation and subcategory. Without detailed cost data, however, it is difficult to isolate the effects of static (e.g., capacity utilization) versus dynamic (e.g., LBD) economies of scale.

Our data allow us to control for LBD (see above) and identify the cost implications of capacity utilization. The results suggest a relatively limited role for capacity utilization; while it is important to maintain a certain threshold (e.g., 70%) to make the most of the prevailing work-shift schedules (i.e., 24 hours a day split into three 8-hour shifts and assigned to four teams in rotation), its impact on costs becomes mostly flat above that threshold. This flat profile is particularly conspicuous for cash costs, whereas depreciation accounting tends to exaggerate this form of scale economies.

Finding 4: Accrual-Based Accounting of Depreciation Expenses Can Materially Bias All of These Findings. Across all three process-level determinants of cost, depreciation expenses emerge as a major source of systematic biases that affect estimates of vintage effects, learning-by-doing, and scale economies. Let us take the example of CPU manufacturing in Figure 1. The current frontier process nodes (3 nm–4 nm) would appear $\left|e^{-1.5} - 1\right| \times 100 = 77.7\%$ more cost-effective than the oldest nodes ($\geq$40 nm) according to the full wafer cost. However, this vintage

effect shrinks by half to $\left|e^{-0.5}-1\right| \times 100 = 39.3\%$ once we ignore depreciation and focus on cash wafer cost.

Likewise, the approximate cumulative effect of LBD after 52 quarters (13 years) is a 95.0% and 63.2% drop, respectively. In terms of quarterly averages, these improvements translate into 5.6% and 1.9% cost reductions, respectively. Similar discrepancies exist for the impact of capacity utilization and calendar time, and the estimates for GPUs follow the same patterns.

In short, the impact of depreciation accounting is frequently bigger than the economic effect under investigation. Hence, clearly identifying these mechanisms is empirically impossible without detailed data that clearly isolate depreciation as a separate line item from other cost items that are physical and cash-flow-based.

Finding 5: Persistent Firm Heterogeneity May Dominate Short-Run Policy Interventions. Firm-level heterogeneity is quantitatively significant, reflecting persistent gaps in technological capabilities. Even though the estimates of firm fixed effects substantially vary across cost measures and model specifications, the overall magnitude is comparable to the effects of the three process features (vintages, LBD, and capacity utilization). Historically, this industry has witnessed the demise of many heavily subsidized laggards that eventually dropped out of the global technological and investment race. Thus, unless the core technological capabilities of the few leading firms could be reliably replicated, the renewed interest in industrial policy would only seem to sow the seeds of disappointment as history repeats itself.

As semiconductors remain at the center of global industrial policy and massive public investments, the economic models used to evaluate these interventions must be strictly grounded in engineering realities. By moving beyond aggregate industry data to directly measure the physical and accounting costs inside the fab, our analysis demonstrates that standard assumptions regarding vintage effects, learning-by-doing, and scale economies often obscure as much as they reveal. Properly isolating accrual-based depreciation and recognizing persistent firm-level heterogeneity is not merely an empirical measurement exercise; it is a necessary step to ensure that future economic research—and the high-stakes industrial policies it informs—is built upon a reliable foundation.

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 U.S. 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 U.S. 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 cells, 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

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

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.

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., Kusaka, S., Qiu, J., and Tran, T. L. (2025). Welfare gains from product and process innovations: The case of lcd panels, 2001–2011. Manuscript, University of Toronto.

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.

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

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.

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