Transistors all the Way Down: Viability of Direct Volume Measurement (and Price Indexes) for Semiconductors

David M. Byrne, Adrian Hamins-Puertolas, and Molly M. Harnish

April 15, 2023

Abstract

The semiconductor (solid-state electronics) industry is a prime example of global production sharing, of the translation of basic research into productivity, and of the power of a general-purpose technology to boost macroeconomic growth. Study of these questions requires measures of industry output and prices that are complete and consistent across time, economies, and product types. Such data are not currently available. We solve this problem by measuring volume directly, relying on plant-level measures of capacity, industry utilization, and easily observed parameters for the state of technology. That is, we do not rely on estimation of constant-quality price indexes, which are the primary source of time-series, cross-sectional, and cross-product inconsistency for semiconductor industry measurement. We construct quarterly-frequency output indexes and implicit prices for key semiconductor product classes for five economies spanning the 1996-2022 period. Our results compare favorably to official measures for time periods and economies where measurement is believed to be reliable. In other cases, our output and price measures are appreciably different from official measures and suggest rather different stylized facts for the global industry.[1]

[1]All authors are employed at the Federal Reserve Board of Governors, Wash-

1 Introduction

The semiconductor (solid-state electronics) industry is a prime example of global production sharing, of the translation of basic research into productivity, and of the power of a general-purpose technology to boost macroeconomic growth. Study of these questions requires measures of industry output and prices that are complete and consistent across time, economies, and products. Such data are not currently available.

We solve this problem by measuring volume directly, relying on global data on plant-level capacity adjusted for industry utilization and the production technology in use to produce output indexes. That is, we do not rely on estimation of constant-quality price indexes, which are the primary source of time-series and cross-sectional inconsistency for semiconductor industry measurement. Specifically, we show that a simple count of transistors per chip provides a serviceable indicator of performance and, consequently, nominal output divided by transistors provides a price-performance ratio that may be used as a quality-adjusted price index.[2]

We make no claim that our simple approach—using transistors alone as a measure of quality—is preferable to approaches that leverage more detailed product characteristics for more narrowly targeted studies. Instead, we demonstrate that because transistors are highly correlated with perfor-

ington, D.C. The views expressed in this paper are not represented to be the views of the Federal Reserve Board of Governors. The authors, listed in alphabetical order, contributed equally to this research. Contributors, described using the CRediT taxonomy (Allen, O'Connell and Kiermer, 2019), are Byrne (conceptualization, formal analysis, investigation, methodology, writing–original draft, writing–review and editing), Andrea Garcia (data curation, software), Hamins-Puertolas (project administration, software, data curation), Harnish (software, data curation, writing–original draft, writing–review and editing), and Steve Young (resources). The authors are grateful for comments from Marshall Reinsdorf and participants in seminars at the Bureau of Labor Statistics and the Federal Reserve Board. The authors are grateful to the computer science community for the "world turtle cosmology" joke that inspired the title, which dates back at least to Berg and Barker (1854). *Please do not circulate or cite this preliminary work without permission from the authors.* Corresponding author: Dave Byrne, david.m.byrne@frb.gov.

[2]For simplicity, we use "transistors" as shorthand for all devices found on semiconductors, including transistors, capacitors, and other. Where appropriate, we make a distinction between transistors and other features.

mance, they provide a reasonable facsimile of full quality adjustment. And, most importantly, this approach can be used throughout the history of the industry, in all countries, and for all products. That is, we can rule out data sources, methods, and ad hoc assumptions as a cause of spurious variation that may undermine broad analysis of the industry.

We construct quarterly-frequency output indexes and implicit prices for key semiconductor product classes in the five economies that dominate the chip industry, spanning the 1996–2022 period. In particular, we provide results for microprocessors, memory chips, and other logic chips, which collectively account for roughly 70 percent of global chip production.[3] Our results compare favorably to official measures for time periods and economies where measurement is believed to be reliable. In other cases, our output and price measures are appreciably different from official measures and suggest rather different stylized facts for the global industry.

2 Existing Methods

Nearly all global production of semiconductors takes place in China, Japan, South Korea, Taiwan, and the United States, so we have focused on those economies in the initial stages of the project.[4] Official semiconductor production statistics in these economies, excepting China, are constructed as nominal output divided by a price index meant to account for the evolving quality of the chips produced.[5] Because the collection of information on nominal output is relatively straightforward, we focus our attention on issues that arise in price index construction.

Official price indexes are most often calculated using a matched-model approach: Prices for goods in a representative basket are observed repeatedly over time, and, for each period, item-level price changes (relative to the previous period) are averaged across the basket using weights reflecting their relative importance for production, when available. Under the maintained assumption that the quality of each individual item in the basket is

[3]The remaining product classes and countries, plus history back to the beginning of the industry in the early 1950s will be added to the analysis in future revisions.

[4]87 percent of non-R&D wafer capacity was located in these five economies in 2022 according to SEMI (Semiconductor Equipment and Materials International). Our sources and methods will permit us to eventually produce results for all economies with semiconductor industries.

[5]China provides a unit count of chips produced as an indicator of production.

invariant over time, the price index thus created often provides a serviceable indicator of aggregate price trends, particularly when the product market is competitive and observations are reasonably frequent. In particular, in the presence of sufficient competition to ensure that the "law of one price" applies, when a new model enters with a lower price-performance ratio—more quality for a lower price, say—the prices of incumbent models can be expected to fall, leading the workhorse matched-model index to capture the correct price trend through the growth rate of the incumbent models. Thus the matched-model index accounts for quality change indirectly.

However, in the case of IT equipment, short product life cycles and imperfect product-market competition often undermine the matched-model approach. For product markets characterized by imperfect competition—such as when the cost of switching between products or platforms creates market power—price declines for incumbent items may not capture the full difference in quality between new and old items.[6] For example, committed users of older systems may pay more for their equipment on a quality-adjusted basis because they are reluctant to switch to newer options. In such cases, supplemental analysis is needed to attribute a portion of the premium or discount for the new item relative to incumbent items to a difference in quality, leaving the residual as a measure of pure price difference (inflation).

This supplemental analysis often takes the form of imputing a price for a new product in the period just prior to its entry. The price change between the imputed price and the first observed market price is then folded into the matched-model index calculation, thus reducing the reliance on the induced price change in incumbent models for capturing quality change. This can be done with a "hedonic" regression, which leverages the correlation of item prices with observed product characteristics. Provided the qualities valued by the user, both observed and unobserved, are correlated with the characteristics observed by the statistician, this approach can be expected to perform well. The clock speed of a computer processor, say, may serve as a proxy for a set of unquantified characteristics of interest to the consumer such as the speed with which the processor can sort the rows of a spreadsheet.

Most hedonic studies closely examine the market and the technology for the product of interest and cast a wide net to consider as many observable characteristics of the product and the market as possible, hoping to hit upon a regression specification that will "explain" as much of the price variation as

[6]See the discussion of the computer market in Dulberger (2007), for example.

possible. Other studies defer to the expertise of electronics industry insiders, who have constructed benchmark tests meant to measure the performance on tasks for which the product is known to be widely used (Byrne, Oliner and Sichel, 2018; Chwelos, 2003). Although our price indexes are not hedonic in nature, they are in the spirit of the latter approach in that we use benchmark scores as a measure of quality.

2.1 Prices by Product Type

Although the producer price indexes (PPIs) produced by official statistical agencies are the natural choice for analysis of the industry, supplemental analysis by outside researchers gets significant attention for several reasons. First, official product-level indexes can be suppressed in public distributions due to concerns about disclosing sensitive information regarding specific reporting companies. Second, responding companies may not be representative of the industry as a whole. And, third, researchers may prefer to have greater visibility into sources and methods than is available from official sources. We examine approaches used for official and research price indexes for the product types that are the focus of this paper—microprocessors, memory, and other logic—restricting our attention to the United States.

2.1.1 Microprocessors

Figure 1 on page 6 presents the producer price index (PPI) published by the Bureau of Labor Statistics (BLS) for microprocessors—the black line—along with available research indexes. The PPI has some undesirable properties for economic research.[7]

- There are periods where no index is available. From 2015 onward, the MPU PPI is not disclosed, an unfortunate omission in a time when

[7]Analysis was conducted in R (R Core Team, 2021) supplemented with the packages cited here (Arel-Bundock, Enevoldsen and Yetman, 2018; Białek, 2022; Bryan, 2016; Hallman, 2018, 2020; Capital Markets section, Research and Statistics Division, Federal Reserve Board, Harkrader, Seok, Dwyer, Konzem and Deluca, 2017; Ooms, 2021; Schauberger and Walker, 2021; Wickham, 2007; Wickham, Averick, Bryan, Chang, McGowan, François, Grolemund, Hayes, Henry, Hester, Kuhn, Pedersen, Miller, Bache, Müller, Ooms, Robinson, Seidel, Spinu, Takahashi, Vaughan, Wilke, Woo and Yutani, 2019).

many observers are concerned about a technology slowdown in the industry. And, MPU production began roughly ten years before the PPI series starts in 1981.

- An index for the processing function broadly defined would be more useful for productivity analysis. Nordhaus (2007) found that the cost of computing began to fall rapidly in 1944, long before the introduction of MPUs. Flamm (1987) estimates that the price of CPUs (central processing units), the precursors to MPUs, fell an average of 28 percent per year from 1957 to 1996.[8]

- Sources used vary across the history of the index. Beginning in 1997, BLS incorporated secondary source data into the index due to the non-cooperation of some sampled firms. Although BLS cannot disclose the identity of the firm in question Byrne et al. (2018) mimic the PPI using publicly-available Intel price lists for 1,000 unit trays, suggesting Intel stopped cooperating with price collection and the PPI is based on their list prices. Whether MPUs are actually sold to major computer manufacturers at these list prices is unknown to the authors. That is, these may not be transaction prices.[9] From the mid-1990s forward, all research indexes shown suffer from the same shortcoming as they also rely on Intel list prices.

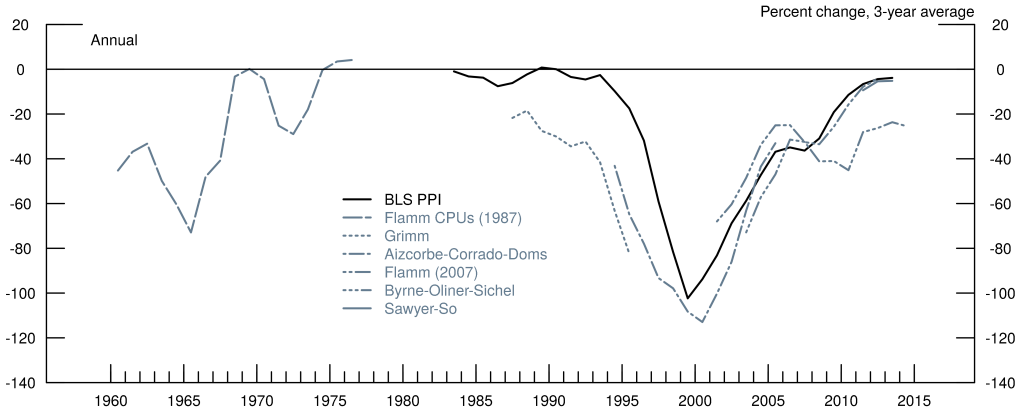
- Index methods vary across the history of the index. In 2011, BLS switched from a matched-model approach to a time-dummy hedonic methodology.

- The PPI displays a markedly slower rate of secular decline than almost all the research indexes shown—the grey lines—for reasons unknown to the authors. One possibility is that low-performance trailing edge processors produced for use equipment other than computers are heavily weighted in the PPI (Holdway (2000)).

[8]Prior to the advent of MPUs, which contain all elements of the CPU on a single microchip—arithmetic logic unit, controller, and cache memory—the functions were performed by a set of components, known collectively as the CPU.

[9]Some online sales of single units take place at these prices, but this is a minor portion of the market.

Figure 1: Microprocessor Price Indices for the U.S. Industry



Note: Indexes differ with respect to price data source, methodology, and scope (see text).
Source: Bureau of Labor Statistics, Grimm (1998), Aizcorbe, Corrado, and Doms (2003), Flamm (2007), Holdway (2000), Byrne, Oliner, and Sichel (2018), Sawyer and So (2017).

2.1.2 Memory

The BLS PPI for memory chips is shown in fig. 2 on the next page along with available alternative indexes. It has shortcomings for use in the study of broad research questions as well.

- There are periods where no index is available to the public. Memory chip production for commercial use began in the late 1960s, roughly ten years before the PPI series starts.[10] And, from 2004 forward, no index is released to the public. In the U.S. industry, Micron, Inc. accounted for almost all memory production from the mid-2000s forward—a likely explanation for the PPI not meeting requirements for disclosure.

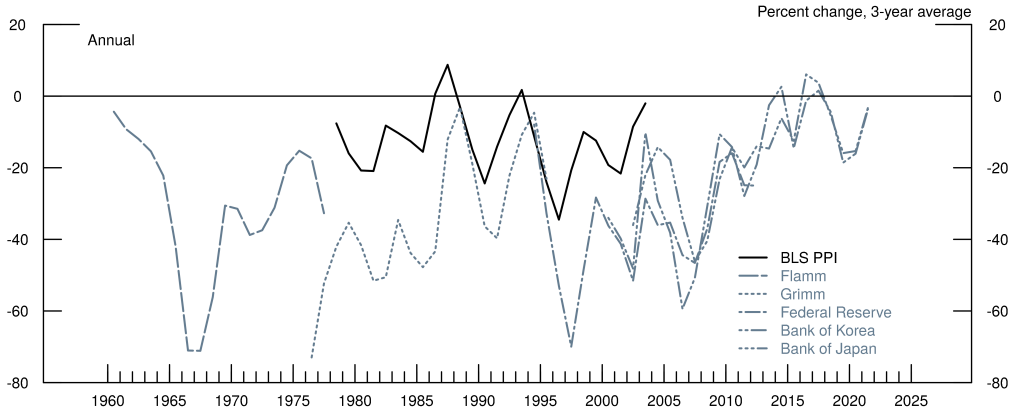
- Research price indexes and the PPIs from Korea—the center of the global memory industry—and Japan have found substantially faster rates of decline.

The alternatives to the BLS PPI have shortcomings as well. The Federal Reserve Board (FRB) research index shown is based on commercially available memory market data. These data provide average prices for product

[10] In fig. 2 on the following page, prices prior to the advent of memory ICs correspond to ferrite memory cores.

classes, not model-level data. In addition, the weights provided in the commercial sources correspond to sales, not production. And, of course, the Bank of Korea and Bank of Japan indices correspond to the Korean and the Japanese memory industry, respectively. That being said, under an assumption that memory prices are set globally, these indexes are reasonable proxies for U.S. production prices. Following that logic, the Bank of Korea price is used to extrapolate the FRB price index in the construction of the industrial production index.

Figure 2: Memory Price Indices for the U.S. Industry



Note: Indexes differ with respect to price data source, methodology, and scope (see text).
Source: Bureau of Labor Statistics, Flamm (1980), Grimm (1998), Federal Reserve Board, Bank of Korea, Bank of Japan.

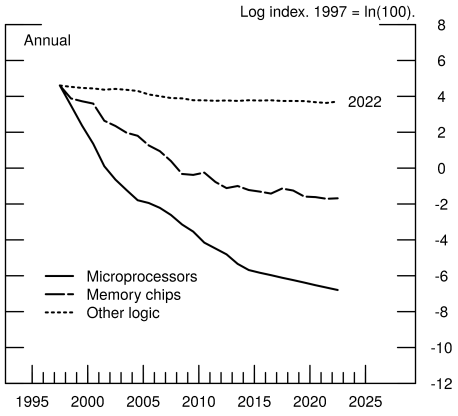
2.1.3 Other Logic

"Other logic," as defined here, includes processors designed for narrow purposes, in contrast to MPUs, which are designed to be used for a wide array of calculations. Among these chips are hardware accelerators—which specialize in graphics (GPUs), artificial intelligence (TPUs), signal processing (DSPs), and so on—and application specific integrated circuits (ASICs), which are customized for a particular device—an entertainment device, a motor vehicle, etc.[11] The BLS does not release a PPI for other logic. A price

[11] Also included are standard cell logic, FPGAs (field programmable gate arrays), and general purpose logic, both MOS (metal oxide semiconductor) and bipolar types.

index constructed using average prices from the Semiconductor Industry Association for other logic product classes is shown in fig. 3. In contrast to the rapidly falling FRB price indexes for microprocessors and memory chips, the other logic price index only inches down over time. In light of the highly innovative and competitive nature of other logic production, this contrast is implausible. Acknowledging this tension, the FRB uses a geometric mean of the three indexes shown to deflate nominal output of other logic chips in the industrial production index. An index built directly from data on these chips would be more defensible.

Figure 3: Federal Reserve Board Price Indexes by Chip Type



Note: MPU and memory price indexes are used in the industrial production index as shown. Other logic chip price index is adjusted for bias before use.
Source: Federal Reserve Board.

2.2 Semiconductor Quality by Country

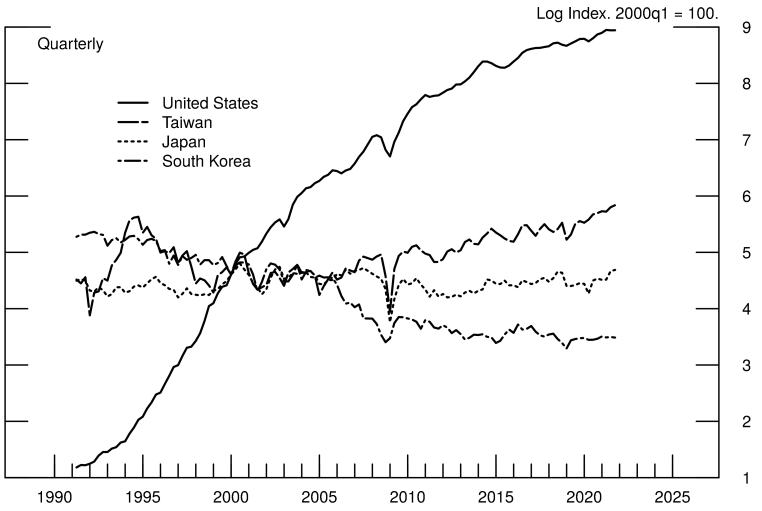
Figure 4 on the following page illustrates tension among national measures from a different perspective. The apparent quality of chips produced has a wildly different trajectory in four of the leading national semiconductor industries. Real output, measured by industrial production, divided by a measure of unit production—the amount of silicon used in the industry measured in wafers—provides an indicator of the quality of industry output.[12]

[12]The wafer index is constructed using data from Semiconductor Industry Materials International and Gartner, Inc., as explained in section 5 on page 20.

U.S. chip quality climbs at a rate of roughly 20 percent per year, on average from 1991 to 2022. Quality in Taiwan, long a rival of the U.S. industry for semiconductor innovation, climbs at only a modest pace, Japanese quality moves roughly sideways, and quality in the Korean industry shows an outright decline. In light of the differing sources and methods used for industrial production in these four economies, one cannot take this result at face value.

The highly implausible contrast between national industries shown in fig. 4 highlights the need for statistics for the global industry created with a consistent approach. With that goal in mind, we anchor our measures in data available for all economies over the full history of the industry, which are primarily measures of plant capacity and technology in use. To translate these data into economic output requires an understanding of the production techniques employed by the industry.

Figure 4: Implied Chip Quality by Country



Note: Natural log of industrial production index divided by (eight-inch equivalent) wafers shipped.
Source: National statistical agencies, SEMI (Semiconductor Equipment and Materials International), Gartner, Inc., authors' calculations.

3 Semiconductor Manufacturing

The fabrication of semiconductors, in contrast to the predecessor circuitry technology based on electrical relays, entails the creation of *electronic* devices

Figure 5: Moore's Framework

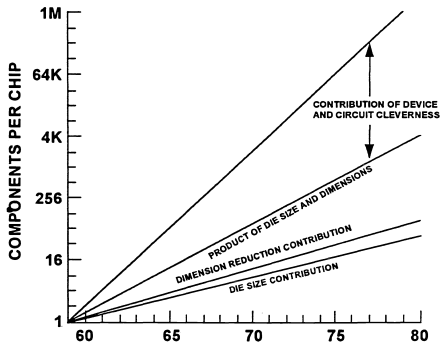


Figure 8. Resolution of the increase in complexity into die size, dimension reduction and "cleverness" factors. Proceedings IEEE IEDM 1975.

on a substrate of semiconducting material, most commonly silicon at the time of writing. These devices may be discrete components—individual transistors, capacitors, resistors, and so forth—or integrated circuits (ICs) which perform logical functions such as mathematical transformations (processors) or hold data to be processed (memory). In semiconductor fabrication plants (fabs), multiple copies of the same device are created on a silicon wafer in a series of steps involving addition and removal of materials on the surface. The individual devices are then cut out of the wafer, tested, packaged, and shipped to customers for use in final electronics, such as computers.

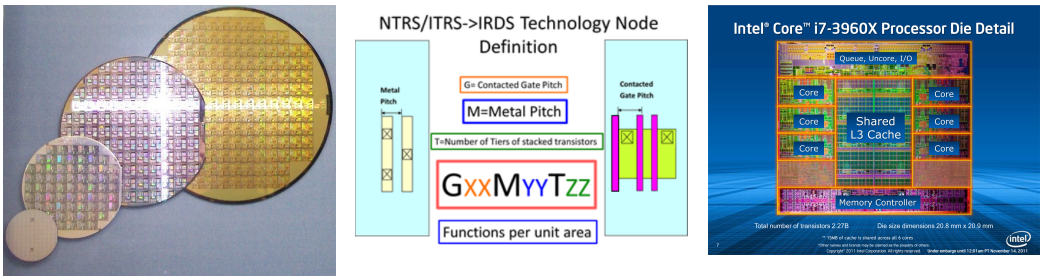
In great part, as Moore (1965, 2006) noted, technical progress in the industry consists in "cramming more components on integrated circuits" to lower production costs, increase reliability, and raise the speed of the devices.[13] As shown in fig. 5, this is done by increasing chip (die) size, miniaturization (dimension reduction), and improving layout (cleverness).

Increasing the size of the wafer used drives down production costs as well. Over time, the industry has introduced silicon wafers of progressively greater diameters, lowering the cost of production through scale economies and by reducing the wasted silicon at the edge of the wafer caused by placing rectangular chips on a circular surface (fig. 6a on the following page). Die size has risen as well, enabling the integration of more logic onto a single chip, reducing costs and raising reliability, albeit at the cost of greater wasted sili-

[13]By combining on one chip the elements of a system, such as a computer, formerly realized on discrete chips and connected together on a circuit board, an integrated circuit advances all three of these objectives.

Figure 6: Factors Affecting Transistor Count

(a) Wafer & Chip Size (b) Minaturization (c) Layout



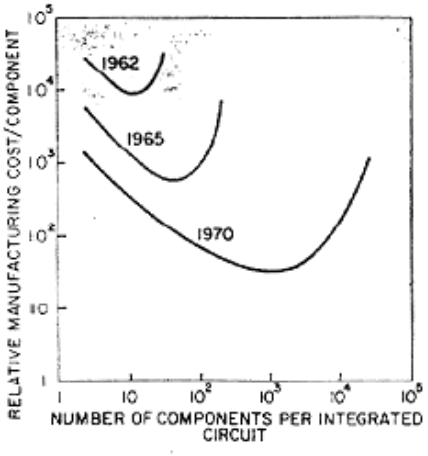
con. Miniaturization, achieved through advances in photolithography—using fine wavelengths of light to create tiny circuit elements on silicon—and other technologies raises the number of components achievable in a given surface area and in some cases increases the clock rate—number of electronic pulses per second—of the device, a key contributor to performance. The state of miniaturization—referred to as the "node," "geometry," or "line width"—at a semiconductor plant is measured by the distance between certain elements on the chip, typically between two lines used for interconnections between gates (metal pitch) or between gates themselves (contacted gate pitch) (fig. 6b). Finally, a given abstract circuit design can be laid out on the silicon substrate in a variety of ways, some of which are more compact than others—more "clever" in Moore's description (fig. 6c). Very recently, the industry has turned to stacking gates on top of one another—another form of "cleverness"—and layers of stacking has become a relevant parameter in characterizing the state of technology in addition to contacted gate pitch and metal pitch (fig. 6b).

As Moore (1965, 2006) observed, while pulling more logical elements onto an integrated circuit lowers production costs—relative to the alternative of using discrete components arrayed on a circuit board—cramming more components on a chip also lowers the yield of useable chips from a wafer as each component has a small probability of being defective and more components entails more chances for the chip as a whole to be defective, lowering the yield of good chips per wafer (fig. 7a on the following page). Consequently, there is an optimal number of components to put on an integrated circuit at any point in time. Looking ahead in 1965, he anticipated that this optimal number would double every year. Reevaluating the state of the industry, Moore

(1975) concluded that the contribution of increasing cleverness—more compact circuit designs—had diminished and predicted that the optimal number of components on an IC would take two years to double from 1980 forward, a prediction known as Moore's Law (fig. 7b).

Figure 7: Moore's Law

(a) Optimal Production



(b) Early Track Record

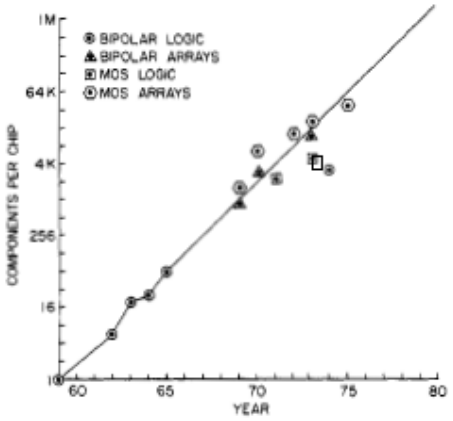


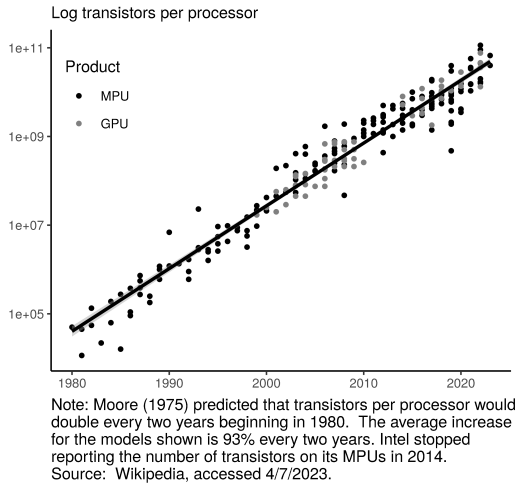
Fig. 1. Approximate component count for complex integrated circuits vs. year of introduction.

Source: Left figure is from Moore's 1965 article. Right figure is from Moore's 1975 reassessment.

Moore's predicted outcome has come to pass for nearly 50 years. That is, the complexity of ICs—measured by the number of components—has doubled roughly every two years, on average. The question of whether Moore's Law holds is not easily answered precisely. As shown in fig. 8 on the following page, at each point in time, a range of complexities (component counts) is on offer in the market, not a single complexity. Moore provided a useful but highly stylized depiction of the supply side of the market—an approach economists will appreciate (fig. 7a)—but, without a fully specified model, including a richer supply side that accounts for product location in characteristics space as well as a demand side, Moore's Law is not a testable hypothesis. Nevertheless a log-linear regression of transistor counts on time—shown in the figure—yields a coefficient on the time of 0.3282, impling a doubling of

transistors per chip every 2.11 years. It seems Moore was off by 6 weeks.

Figure 8: Transistor Counts for MPUs and GPUs

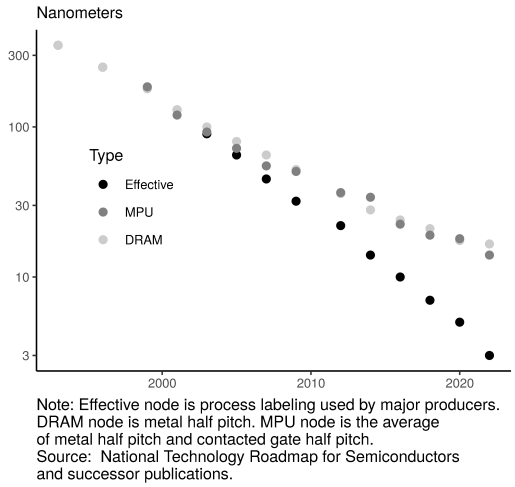


Note: Moore (1975) predicted that transistors per processor would double every two years beginning in 1980. The average increase for the models shown is 93% every two years. Intel stopped reporting the number of transistors on its MPUs in 2014.
Source: Wikipedia, accessed 4/7/2023.

Discussions of Moore's Law have at times been mistakenly focused on the dimension reduction contribution alone. The pace of miniaturization has, in fact, slowed down in recent years, as shown in fig. 9 on the following page, a development somewhat obscured by industry communication strategy: In the early 2000s, leading semiconductor manufacturers adopted the misleading practice of reporting "effective" rather than actual linewidth, in an effort to convey their continued success in pushing performance forward despite a slowdown in miniaturization. Moore's Law as originally stated (about component counts) has held despite this development due to rising contributions from the "cleverness" factor included in Moore's analysis (fig. 5 on page 10). In addition to the more compact circuit layout Moore had in mind, examples of cleverness include technology to stack memory cells vertically and transistors with "fins" that allow for multiple gates in the same location.

The components Moore was referencing are primarily transistors, though capacitors, diodes, and resistors are found on ICs as well. Most notably, DRAM (dynamic random access memory) relies heavily on capacitors. We will refer to transistors—the most important components, except in the case of DRAM and optoelectronics—in our component counts for simplicity. Figure 10a on page 15 shows a stylized transistor. Electric current travels from the source (S) to the drain (D) when the gate (G) made of semiconducting

Figure 9: Definitions of Semiconductor Technology Nodes



Note: Effective node is process labeling used by major producers.
DRAM node is metal half pitch. MPU node is the average
of metal half pitch and contacted gate half pitch.
Source: National Technology Roadmap for Semiconductors
and successor publications.

material (in pink) is caused to switch from an insulator to a conductor by current in the black connector. This allows the transistor to act as an on-off switch and to represent a Boolean variable. Using transistors, any Boolean logic can be represented in a solid-state electronic form. For example, fig. 10b on the next page shows the use of semiconductor transistors (the thick vertical green lines) to create a NAND ("not and") Boolean function, also known as a gate. Two electrical lines enter the gate. If each has positive voltage, the exiting line will not conduct current; otherwise it will.[14] Figure 10c on the following page shows a microprocessor—an integrated circuit for general-purpose calculation—built from millions of gates which are, in turn, built primarily from transistors. Thus the number of transistors found on an IC are a measure of its complexity.

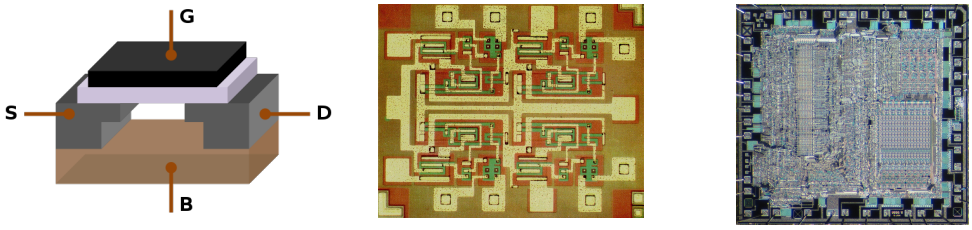
4 Transistors as a Barometer of Quality

Because the IC transistor count rises with circuit complexity, it serves as an indicator of quality from an engineering perspective. Whether this corre-

[14]NAND gates have the useful property of being a "universal gate," meaning that any Boolean logic can be represented entirely in NAND gates. The use of universal gates simplifies the semiconductor production process.

Figure 10: From Transistors to Integrated Circuits

(a) Transistor (b) Gate (c) Integrated Circuit



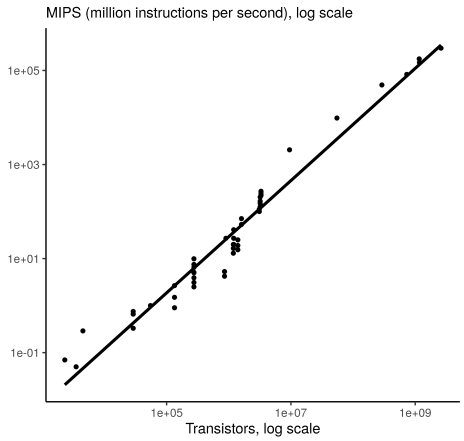
Note: In the transistor figure, elements are source (S), drain (D), gate (G), and base (B); Pink material is semiconducting material. The gate depicted is a NAND gate. The integrated circuit depicted is a microprocessor.
Source: Wikimedia Commons.

sponds to more performance as valued by the ultimate user of the device—the "quality" economists have in mind—is a separate question. We test this proposition using benchmark performance scores. Device benchmarks mimic common use cases either with a suite of complete programs (sorting strings, searching for prime numbers, etc.) or a set of fundamental computing tasks (loading data, addition, etc.). As Triplett (2005) notes, "benchmark measures have the advantage that they measure machine performance, rather than measuring some proxy for machine performance, or some input that may influence machine performance." That is, although much of the literature on quality-adjusting output and prices has focused on physical or engineering characteristics, this is only a suitable approach to the extent that device performance from a user perspective is correlated with those characteristics. Ohta and Griliches (1976) express the same view, noting that "ideally, quality adjustments should be based on performance variables, which presumably enter the utility function directly, not physical characteristics."[15] Unfortunately, benchmarks are not available for all chip types in all time periods, so they cannot be the foundation for a complete measurement system. Where coverage is available, though, we find that performance is highly correlated with transistor counts, suggesting they serve as a reasonable indicator of quality-adjusted production.

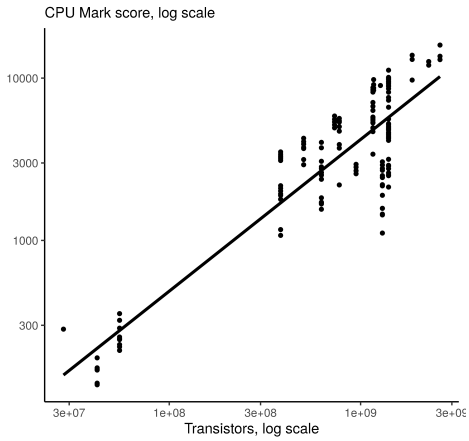
[15]In some cases, there has been little difference between quality adjustment based on physical characteristics and based on benchmark performance (Chwelos, 2003), and in others, that has not been the case (Byrne, Oliner and Sichel, 2018).

Figure 11: Microprocessor Benchmark Performance v. Transistors

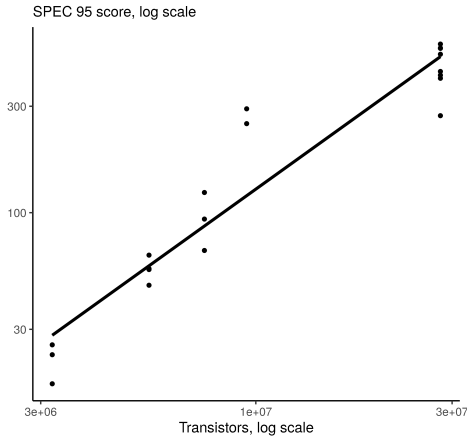
(a) MIPS Benchmark



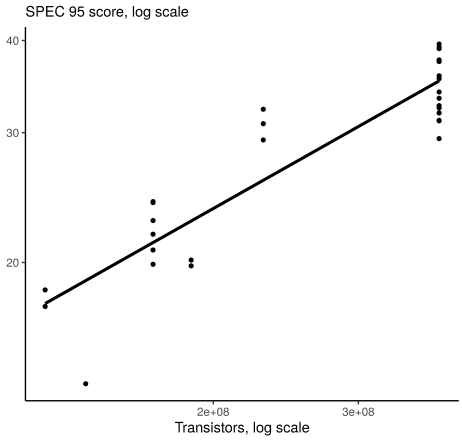
(b) Passmark Benchmark



(c) SPEC 1995



(d) SPEC 2000



Note: Panel a uses Intel processors released from 1971 to 2014. Panel b includes Intel Pentium III, Pentium 4, Core i3, Core i5, Core i7, and Celeron MPUs. Panels c and d are primarily high-performance MPUs. SPEC scores are geometric mean of integer and floating point "rate" scores.
Source: For panel a, Intel Inc., Wikipedia, Computer History Museum. For panel b, Passmark, Inc. For panels c and d SPEC (System Performance Evaluation Corporation).

Table 1: Processor Benchmark Alternatives

Benchmark	Coverage	Dates	Access
Microprocessors			
MIPS	broad	1954–	public
PassMark	broad	2006–	license
SPEC	Server MPUs	1989–	public
GeekBench	PC MPUs	TBD	TBD
GPUs			
PassMark	PC GPUs	TBD	licensed
SPEC	Server GPUs	1989–	public
GeekBench	PC MPUs	TBD	TBD
Google TPU	TPUs	TBD	TBD
RAM Memory			
PassMark RAM	broad	2003–	licensed

Note: SPEC is System Performance Evaluation. GPU is graphics processing unit. TPU is Tensor processing unit.
Sources: Passmark: http://www.passmark.com/; SPEC: http://spec.org/; GeekBench: http://www.geekbench.com/; MIPS: "Computer Benchmarks, Stress Tests and Performance History," Roy Longbottom, (http://www.roylongbottom.org.uk/); Google TPU: http://cloud.google.com/blog/products/ai-machine-learning/cloud-tpu-v4-mlperf-2-0-results.

We employ results from several sources to test the relationship between benchmarks and transistor counts (table 3 on page 19). For MPUs, we compare transistor counts to three benchmarks—MIPS (Million Instructions per Second), Passmark CPU Mark, and SPEC CPU (System Performance Evaluation Corporation)—each of which has its advantages.

Figure 11a on the preceding page shows that for Intel MPUs released between 1971 and 2014, there is a tight relationship between transistor count and performance. (The correlation between log transistor count and log MIPS score is 0.97.) MIPS has the appeal of coverage over a long period of time, but has the drawback of using of representative programming instructions rather than full programs.[16]

[16]Because MPUs from different architecture families—Intel's x86, IBM's PowerPC—use

Figure 11b on page 16 shows that for Intel MPUs released between 2010 and 2014 for the PC (personal computer) market, there is a strong relationship as well, albeit less so than for MIPS. The Passmark CPU Mark benchmark is a composite of the time taken to complete a set of common program types. These include data compression, data encryption, integer and floating point math, searching for prime numbers, sorting strings, and physical system simulation.[17] Benchmarks based on complete programs are considered more reliable than MIPS (Mashey, 1990). A drawback of Passmark—as well as Geekbench, not in use in the analysis at the time of writing—is that scores first become available in 2007. SPEC scores—not in use in the analysis at the time of writing—are available beginning in 1996, but are only available for MPUs designed for use in servers and high-performance PCs.

Acknowledging the substantial variation in Passmark performance scores given the number of transistors is essential to fairly characterizing the merits of our approach. For illustration, table 2 on the following page provides a list of all processors shown in fig. 11b on page 16 with the same number of transistors (1.16 billion). PassMark scores range from 3,399 to 8,727, more than twice as high. The MPUs shown differ with respect to clock speed, power consumption, number of simultaneous processing threads supported, integrated graphics, and amount of cache memory, all of which play a role in processor performance. For the processors in fig. 11b on page 16, 46 percent of the variation in performance is explained by transistor count. As shown in table 3 on the next page, an additional 32 percent of variation can be explained when controlling for clock speed and power consumption in a linear regression.

We examine the relationship between transistors and benchmark scores for GPUs as well. The PassMark GPU benchmark suite tests rendering of 2D images, image filtering, rendering of dynamic 3D scenes, and a set of computationally-intensive parallel processing tasks. The computational tasks are meant to account for the common use of GPUs in non-graphics tasks, such as supercomputing. The correlation between log transistor count and log GPU benchmark score is 0.98.

We have assumed for the moment that memory chip performance is equal to the number of storage cells used on the chip. Benchmarks are available

differing combinations of instructions to achieve the same end-user task, comparisons using MIPS can be problematic.

[17] In particular, PassMark uses the Bullet Physics package.

Table 2: Sources of Performance Score Variation across MPUs with Matching Transistor Counts

Processor	Score	Speed	Power	Intended platform use; other features
i5-2450M	3,399	2.5	35	Mobile
i5-2500T	4,680	2.3	45	Desktop; very low power
i5-2405S	4,928	2.5	65	Desktop; low power
i5-2400S	4,941	2.5	65	Desktop; low power
i5-2500S	5,299	2.7	65	Desktop; low power
i5-2300	5,346	2.8	95	Desktop
i5-2310	5,523	2.9	95	Desktop
i5-2380P	5,679	3.1	95	Thin mobile; GPU disabled
i5-2320	5,755	3.0	95	Desktop
i5-2500	6,322	3.3	95	Desktop
i5-2550K	6,700	3.4	95	Desktop; high-performance, unlocked
i7-2600S	7,078	2.8	65	Desktop; low power
i7-2600	8,201	3.4	95	Desktop
i7-2600K	8,457	3.4	95	Desktop; high-performance, unlocked, advanced GPU
i7-2700K	8,727	3.5	95	Desktop; high-performance, unlocked, advanced GPU

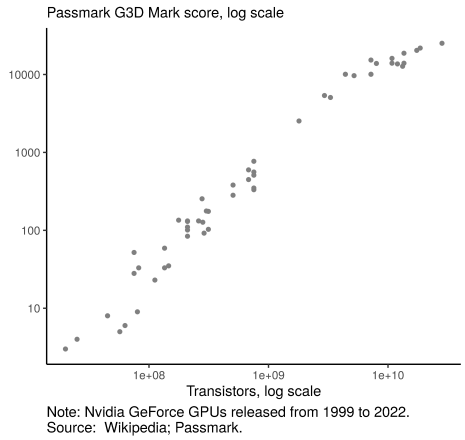
Note: Performance score is Passmark CPU Mark. Speed is clockspeed in GHz. Power is TDP (thermal design power). All processors shown are second generation i5 or i7 models. i5 and i7 models differ with respect to level 3 cache memory (6MB vs. 8MB), the number of processor threads (4 vs. 8), and the quality of graphics processing integrated on the chip.
Sources: Passmark, Wikipedia, Intel, Inc.

Table 3: Sources of Performance Score Variation across MPUs with Matching Transistor Counts

Variable	Coefficient	Standard Error
Transistors (million)	3.29	0.29
Clock speed (GHz)	794.87	268.49
Power (TDP in watts)	40.27	5.07

Note: All coefficients are significant at the 99 percent level. R-squared for the regression is 0.78. R-squared for a regression of performance on transistors alone is 0.46.

Figure 12: Graphics Processors



for additional dimensions of memory chip performance which we will explore as the project progresses.

Reiterating the goal of our project, we endeavor to construct measures of prices and volumes for all time periods, all chip types, and all countries using a consistent methodology. Performance benchmarks have the appeal of directly measuring quality from the perspective of the typical user. However, they are only available for a small subset of chip types and for limited time periods. Where the data are available, there is a strong positive relationship between performance and transistor counts, suggesting using transistors as an imperfect but reasonably compelling indicator of quality. The crucial distinction between our approach and approaches using model-level data is that we have the ability to measure transistor counts for all chip types in all countries for the full history of the industry.

5 Volume and Price Index Construction

In the construction of our price and volume indexes, the intuition we exploit is that real output is proportional to performance which in turn is proportional to transistor counts:

$$\text{real output} \propto \text{performance} \propto \text{transistors}. \tag{1}$$

Thus, our first objective is to estimate the volume of transistor production. We approach the task at the plant level, calculating transistor volume, T, for each product, p, produced plant i in period t. Volume is the product of transistors per chip, chips per wafer, and wafers shipped:

$$T_{i,p,t} = \left(\frac{\text{transistors}}{\text{chip}}\right)_{i,p,t} \cdot \left(\frac{\text{chips}}{\text{wafer}}\right)_{i,p,t} \cdot \text{wafers}_{i,p,t} \tag{2}$$

The first term, transistors per chip, depends on the size of the chip, technology employed, and chip layout. We allow chip size to vary over time and across product types: $S_{i,t} = f_S(p_{i,t}, t)$.[18] Transistors per unit is inversely related to the geometry (line width) employed in production, $G_{i,t}$. (Here we ignore that plants typically produce using multiple geometries at once, a fact we account for as discussed in section 6 on the next page.) The chip layout factor, $L_{i,t} = f_L(p_{i,t}, t)$, accounts for increases in density not accounted for by geometry, achieved by circuit design improvements—Moore's cleverness factor (fig. 5 on page 10)—which vary over time and across products.

$$\left(\frac{\text{transistors}}{\text{chip}}\right)_{i,t} = S_{i,t} \cdot G_{i,t}^{-2} \cdot L_{i,t}. \tag{3}$$

For the second term, chips per wafer, we require a measure of wafer diameter, $D_{i,t}$, and chip area, which allows us to account for the silicon lost at the edge of the wafer. We use the approximation of the number of chips per wafer provided by (De Vries, 2005), which assumes the chips are square:

$$\Delta_{i,t} = \frac{\pi D_{i,t}^2}{4S_{i,t}} - \frac{\pi D_{i,t}}{\sqrt{2S_{i,t}}}. \tag{4}$$

The third term, wafers shipped per plant, is the product of plant capacity in wafers, $N_{i,t}$, and plant utilization, $U_{i,t}$:

$$\text{wafers}_{i,t} = N_{i,t} \cdot U_{i,t}. \tag{5}$$

Suppressing plant, time, and product subscripts, production can be written as a simple product of these factors:

$$T = (S \cdot G^{-2} \cdot L) \cdot (\Delta) \cdot (N \cdot U). \tag{6}$$

[18] We explain the data limitations that lead us to ignore the fact that chip size employed varies across models within product types at a point in time discussed in section 6 on the following page.

Production of a given product is simply the sum of output, T, across all plants producing that product. Implicit price indexes by product are calculated at the global level, since there are no consistently available product-country-level nominal output measures. Or, to put it differently, we assume there is a single global market where product prices are set:

$$\pi_{p,t} = \frac{R_{p,t}}{\sum_{i \in I_{p,t}} T_{i,t}}, \tag{7}$$

where $R_{p,t}$ is nominal global production of chip type p in period t and $I_{p,t}$ is the set of plants producing product p in period t.

The industry price index for each country is constructed by weighting global product price indexes for all product, P, using country-level production by product, $T_{p,t,c}$, using a chained Fischer formula, where the index level is the cumulative product of the geometric mean of Laspeyres and Paasche price indexes constructed in each pair of consecutive periods since the base period.

$$\pi_{c,t} = \prod_{\tau=1}^{t} F_{c,\tau}, \text{ where} \tag{8}$$

$$F_{c,t} = \sqrt{\text{Laspeyres}_{c,t} \cdot \text{Paasche}_{c,t}}, \text{ and} \tag{9}$$

$$\text{Laspeyres}_{c,t} = \frac{\sum_{p \in P}(\pi_{p,t} \cdot T_{p,t-1,c})}{\sum_{p \in P}(\pi_{p,t-1} \cdot T_{p,t-1,c})}, \text{ and} \tag{10}$$

$$\text{Paasche}_{c,t} = \frac{\sum_{p \in P}(\pi_{p,t} \cdot T_{p,t,c})}{\sum_{p \in P}(\pi_{p,t-1} \cdot T_{p,t,c})}. \tag{11}$$

6 Data Construction

Fundamentally, our measure is an aggregate of output measures for individual plants which accounts for capacity, technology, product mix, and utilization. As explained below, we rely on quarterly-frequency, exhaustive global lists of fabs (semiconductor fabrication plants) which include key technical details of production. Combining capacity with industry utilization rates yields a raw production flow in silicon wafers. Refining this production measure to account for technology, we construct an estimate of transistors shipped as explained in section 5 on page 20.

6.1 Plant Characteristics

We rely heavily on data from two sources for the construction of our database: Gartner, a consultancy focused on information technology, and SEMI (Semiconductor Equipment and Materials International), a trade association focused on semiconductor manufacturing equipment. Gartner published comprehensive global lists of semiconductor fabs as part of their *Semiconductor Equipment and Materials Service* beginning in 1985 (Gartner, 2007). We located and digitized 22 of these lists published from 1996 to 2007. We constructed a quarterly-frequency panel of fabs for this period, filling in gaps with values from the most recent previous observation. SEMI currently publishes a comprehensive, quarterly-frequency database of fabs with history beginning in 2008, the *Wafer Fab Forecast* (SEMI, 2022). We merged three vintages of this report to build a panel for 2008-2022. Combining the two sources, we have a comprehensive record of global capacity, by plant, from 1996 to 2022.[19]

Each data source provides plant-level information on location, company ownership, technology employed, capacity, and product mix.[20] In particular, the data provide plant capacity (measured in the number of silicon wafers that can be processed in a month), diameter of the wafers processed, and a measure of the greatest level of circuit miniaturization achieved at the plant.

In addition, data is provided on the product mix of the plant, which we use to classify plants as producers of analog chips, DRAM chips, Flash or other memory chips, discrete chips and optoelectronics, microprocessors, and other logic chips. Foundries, which fabricate designs under contract, are identified in the data and are assumed to produce chips from the other logic category. For plants with products listed that fall into more than one of these six product classes, capacity is allocated equally between the categories in most cases. In cases of plants with MPU production and production of other product classes as well, 90 percent of capacity is assigned to MPU production and the remaining 10 percent is divided equally among the relevant other classes, in keeping with our understanding of the strategy used by MPU

[19]In the work presented in the current draft, we have linked the SEMI and Gartner databases at the economy-product-technology level. Plant-level matching is underway for future work.

[20]In addition, both datasets contain information on when the plant was established. This information will provide the foundation for future efforts to expand the dataset to earlier years. The earliest date for plant establishment in the data is 1953.

companies.

We exclude from our scope plants producing MEMS (microelectromechanical systems) or epitaxial wafers (primarily used for solar cells), and "back end" plants (used for testing completed wafers); these plants use rather different production processes than plants producing microchips do. We also exclude plants recorded as planned or under development and plants recorded as closed or inactive. Finally, we set aside industry, academic, and government research plants for separate analysis. For the current stage of the project we focus our detailed analysis on DRAM, flash and other memory, MPUs, and other logic.

After this cleaning process, we have between 200 and 350 plant-level observations for each quarter in the dataset (table 4). Using 8-inch (200 mm) equivalent wafers—a coarse measure of capacity—capacity shifted fairly steadily toward China, South Korea, and Taiwan over the period of study—each of which accounted for roughly one-quarter of the industry in 2022—and shifted away from Japan and the United States.[21] Importantly, wafer capacity does not account for differences in technology employed. From 1996 to 2022, the mix of technologies in use varied widely across the countries studied. In 2022, Taiwan and South Korea had capacity at the most advanced geometry in use, 3 nanometers. The U.S. industry was just behind the frontier and the most advanced plants in China and Japan were several technology generations behind. In contrast, in 1996, the U.S., Japan, and Taiwan industries were at the leading edge. As shown in fig. 13 on page 26 each economy produces chips using a diverse set of technologies. Plants using technologies many generations behind the frontier continue to operate.

6.2 Utilization

Importantly, the reports from SEMI and Gartner do not provide information on realized production. To construct a measure of production, we multiply capacity by an estimate of the operating rate for the industry. We rely on the operating rate reported by the U.S. Census Bureau's Quarterly Survey of Plant Capacity (SPC) for the industry in the United States. While we would prefer a measure global in scope, no such measure is consistently available.

[21] The 8-inch equivalent wafer measure divides the surface area of wafers with other diameters by the surface area of an 8-inch equivalent wafer. The most common diameters in use during the study period were 150 mm, 200 mm, and 300 mm. In calculations of unused edge silicon, the true wafer diameter is used.

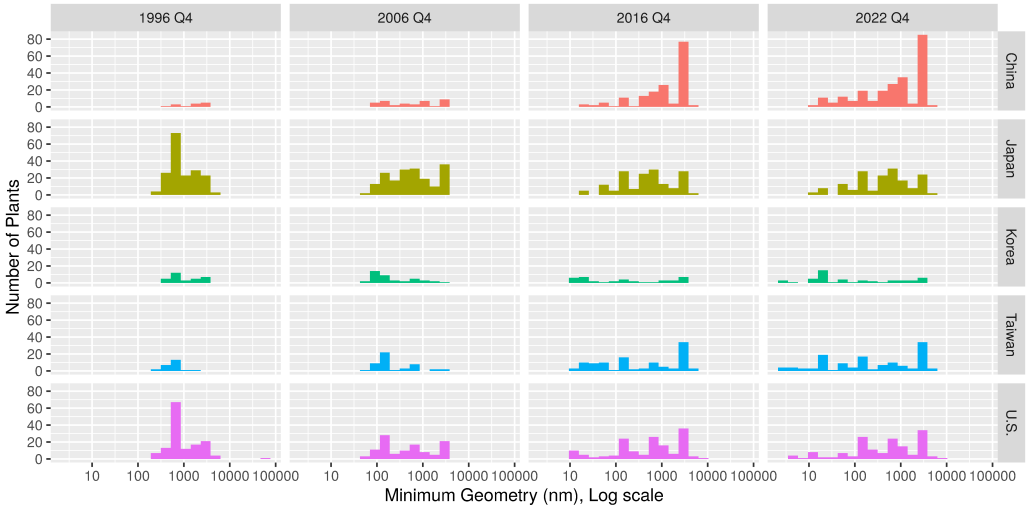
Table 4: Database Descriptive Statistics

	1996	2006	2016	2022
Fab Count	321	286	209	273
Capacity Share (percent), by Economy				
China	1.4	8.4	14.3	26.3
Japan	44.9	23.7	19.8	13.1
S. Korea	10.8	20	25.2	25.4
Taiwan	12.1	25.2	28.5	26.1
United States	30.8	22.7	12.1	9.1
Capacity Share (percent), by Product				
DRAM	34.3	17.4	23.4	31.3
Oth. Memory	18.7	34.9	24.8	13.9
MPU	18.5	8.2	3.2	2.8
Oth. Logic	28.5	39.5	48.6	52.1
Minimum Geometry (nm)				
China	800	90	22	14
Japan	250	65	22	14
S. Korea	350	65	10	3
Taiwan	250	65	10	3
United States	250	65	10	5

Note: Fabs producing microprocessors, memory, and other logic. Capacity share is measured in 8-in equivalent wafers. Minimum demonstrated geometry shown; high-volume production may not have been achieved at that node. Source: Gartner, SEMI, and authors' calculations.

However, a global industry utilization report from Semiconductor Industry Statistics (SICAS), now discontinued, covered 1994-2010 (U.S. Census Bureau, 2022). The contour of global utilization shown in SICAS is closely aligned with the SPC, so we are comfortable using the SPC as an indicator of the global utilization rate (fig. 14a on page 27). Because the SICAs data show that fabs using the most advanced (largest) wafer size had higher utilization rates we vary utilization by wafer diameter (fig. 14b on page 27). We assume plants processing the largest wafers in each quarter operate at the SPC operating rate. Plants behind the wafer size frontier are assumed to operate with a 10 percent lower utilization rate, the rough average difference

Figure 13: Minimum Geometry Distribution



in the SICAS data.[22]

6.3 Chip/Die size

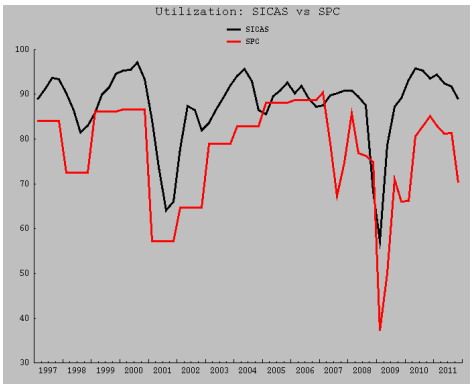
We further refine the measure to deduct an estimate of the unused wafer area that naturally arises from cutting rectangular chips out of circular wafers. To do so, we use an estimate of die (chip) size by year for microprocessors (MPUs) developed in Burg and Ausubel (2021) and make the coarse assumption, loosely consistent with industry history recorded in the industry press and on company websites, that memory and logic chip die are half as large as MPU die and that other chip types are one-tenth the size. Accounting for wafer diameter by plant as well, we use this formula, found in De Vries (2005): $\Delta = \pi D^2/4S - \pi D/\sqrt{2S}$, where Δ is die (chips) per wafer, D is wafer diameter, and S is the surface area of the die.[23]

[22] When 300mm wafers were introduced by the industry in 2002, they became the largest wafer size. We assume that from that point forward, the operating rate of 200mm fabs (previously the largest) declined by 10 percent over a five-year period.

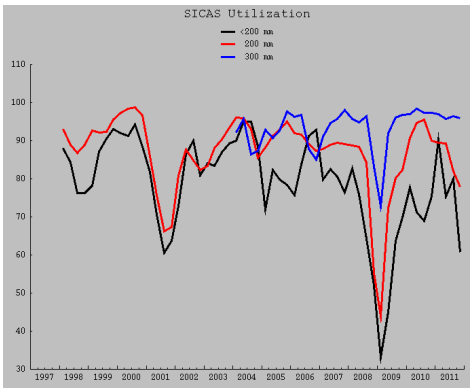
[23] We are refining this time-series for die size using information from the *International Technology Roadmap for Semiconductors*. Analysis was not complete at the time of writing.

Figure 14: Semiconductor Industry Utilization Rates

(a) U.S. and Global Semiconductor Industry Utilization Rates



(b) Global Semiconductor Industry Utilization Rates by Wafer Size



6.4 Transistor-Density "Cleverness" Factor

The formula in eq. (6) on page 21 for transistors produced implicitly assumes that transistors have no space between them when placed on chips. In practice, transistor density is affected by the need for other features on the chip—to conduct current for example. Chip designers adjusting the layout to minimize wasted space (with specialized software)—the cleverness Moore referred to in his analysis (fig. 5 on page 10). We adjust output for each of the four product classes to account for evolving cleverness.

For MPUs, we benchmark our transistor production measure using the density of transistors on leading edge MPUs reported by Burg and Ausubel (2021) through 2013 and extended by us using information on more recent MPUs available from Intel (Intel, Inc., 2022). In particular, we calculate the ratio of reported density to theoretical density for these leading edge chips and apply that ratio to both MPU, DRAM, and other logic production.[24]

In recent years, expanding chip circuitry vertically has provided a new tactic for increasing density (for a given geometry). The Flash memory industry has concluded that further miniaturization is no longer the optimal approach to technical advance, and has focused on stacking multiple transistors on top of one another. We construct a time-varying density factor to

[24]Work in progress will construct a distinct density adjustment for DRAM, which does not face the same design challenges as processors.

account for stacking based the ratio of flash memory measured in bytes to flash memory measured in chips reported in the *World Semiconductor Trade Statistics* published by the Semiconductor Industry Association (SIA).

6.5 Standardizing Minimum Geometry

Each plant in the SEMI and Gartner database reports the minimum geometry employed at the plant for each quarter. In practice, the technology used on production lines has small variability around standard geometry nodes. We standardize the reported geometries by assigning them to the next largest standard node.[25] Next, we ensure that no fab in our dataset has an anachronistic minimum geometry. If the assigned geometry of any fab in a given quarter is smaller than the industry cutting-edge technology for that quarter, we replace it with the industry cutting-edge node for that quarter.[26]

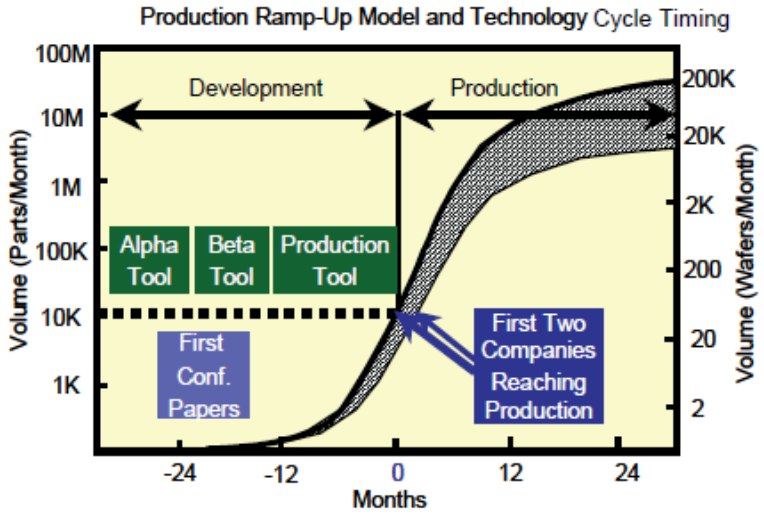
6.6 Technology Ramp-up Process

Semiconductor plants are typically fabricating wafers using multiple technologies at any point in time. Because we only observe the most advanced technology in use, we model the process distribution following Byrne, Kovak and Michaels (2017) such that at any given time, the two previous geometry nodes behind the most advanced one in use are also employed in production at each plant. In particular, we loosely follow information from the *International Technology Roadmap for Semiconductors* (fig. 15 on the next page). We assume that when a new technology is introduced at a plant—as indicated by its first appearance in the fab's listing in the databases used—5 percent of production takes place using that technology in the first quarter. Over the next five quarters, the new technology's share of production increases linearly until it reaches 50 percent. Over the same time period, the second previous (largest) node's share of production decreases linearly from 50 to 0 percent. Meanwhile, the first previous (second largest) node's share of production holds steady at 50 percent until another new technology is introduced, at which point it becomes the largest node and the cycle begins

[25] These notes are available on Wikichip and Wikipedia. See https://en.wikichip.org/wiki/technology_node for nodes before 1971 and https://en.wikipedia.org/wiki/Semiconductor_device_fabrication#Semiconductor_node for nodes after 1971.

[26] We speculate that reporters may provide technology information for processes that are under development, rather than in production.

Figure 15: Technology Generation Life Cycle



Note: Described as "a typical production 'ramp' curve" in the original source.
Source: International Technology Roadmap for Semiconductors, 1997.

again. In effect, we impute a distribution of production across geometries for each plant in the dataset at each point in time.

6.7 Final Dataset

After applying the calcuation in eq. (6) on page 21 and the adjustments described for cleverness and for technology ramp-up, we arrive at a transistor-based indicator of production for each product produced in each plant in each quarter. We aggregate from the plant level to create country-level and global production by product. Product prices are calculated by dividing global nominal production (from the *World Semiconductor Trade statistics*) by our global production indicator for each product. These global price indexes are aggregated using economy-level production weights to arrive a semiconductor industry price index for each economy. Our final dataset will consist of output for analog, optoelectronic and discrete, MPU, DRAM, other memory, and other logic chips produced in each economy with a semiconductor industry from 1953 to 2022. For the present analysis, we have restricted the dataset to focus on MPU, memory, and other logic production in the United States,

Japan, Korea, Taiwan, and China for the 1996-2022 period. By applying the same data cleaning methods and calculations across time, economies, and product types, our approach allows us to examine output and prices consistently on those dimensions.

7 Results

7.1 Output: Transistors

Our production measure, while broadly consistent with capacity measures, adds color to the capacity and geometry statistics above. Similar to the capacity shares in table 4 on page 25, our production measure shows China, South Korea, and Taiwan gaining ground while the U.S. and Japan's shares of production decrease (table 5 on the following page).

Importantly, however, our transistor-based output measure accounts for technology and product mix as well as capacity. For example, while over one-quarter of the five major economies' wafer capacity is located in China in 2022, its production makes up only an eighth of production as measured by our transistor indicator. This discrepancy can be explained by China's concentration at the high end of the minimum geometry scale (fig. 13 on page 26), which limits the number of transistors that can fit on a wafer, and the fact that 62% of China's capacity is devoted to other logic at the end of 2022, which tends to use larger chips, fewer of which can fit on a wafer. In addition, Taiwan's production share is far more consistent than its capacity share, and South Korea's rise in production is more drastic than its rise in capacity. Adding transistor counts to the analysis combines our knowledge of capacity and technology for a more robust picture of output.

Splitting output by product type has a similar effect. While DRAM's share of capacity holds roughly constant between 1996 and 2022, its share of production drops dramatically. MPU's decline in production, meanwhile, is not as dramatic as its drop in capacity. The rise in the production of other memory and other logic relative to DRAM and MPU underlines the need for further study of other memory and other logic chips, which together have grown from 27.3 to 86 percent of output over the last 26 years.

Table 5: Output Shares by Economy and Product Type

	1996	2006	2016	2022
By Economy				
China	0.1	6.2	6.4	12.0
Japan	38.6	19.5	21.2	13.4
S. Korea	11.0	24.2	37.8	36.9
Taiwan	20.2	27.4	15.5	28.2
United States	30.0	22.7	19.1	9.4
By Product				
DRAM	57.0	14.1	24.2	8.7
Oth. Memory	12.9	49.9	44.4	47.6
MPU	15.7	8.6	7.7	5.3
Oth. Logic	14.4	27.4	23.8	38.4

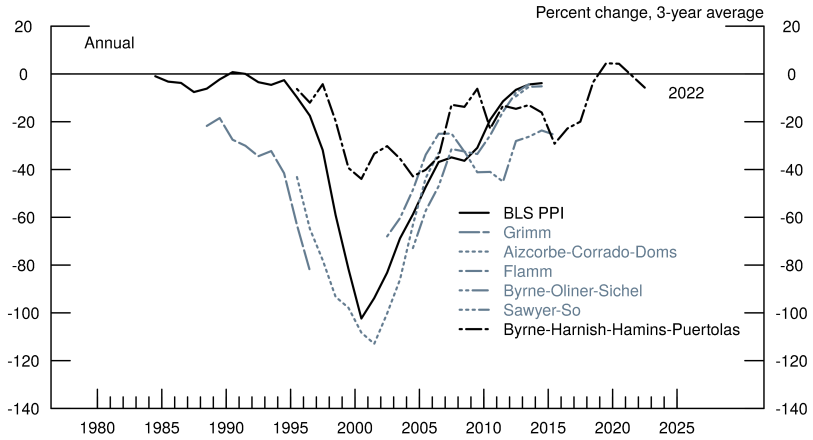
Note: Excludes fabs not producing MPUs, memory, or other logic.
Source: Gartner, SEMI, and authors' calculations.

7.2 Prices: Microprocessors

The price index for MPUs constructed using our approach tells a rather different story for the industry than existing indexes (fig. 16 on the following page). The qualitative features of the contour of price declines match between our index and the others, but the magnitude of the variation is markedly different. As in existing indexes, our index indicates that price declines accelerated substantially in the 1990s, the acceleration reversed in the late 2000s, and price declines moved toward zero very recently. The timing of these moves is somewhat different, with our index accelerating very late in the 1990s relative to the others.

The stark contrast between the acceleration in our index and the four other seemingly independent indexes is surprising. However, there is in fact little independence among these indexes. Though the methods employed by the BLS PPI, Aizcorbe, Corrado and Doms (2003), Flamm (2017), and Byrne et al. (2018) differ, the price information employed in all four cases is the same: model level list prices, *not transaction prices*, for Intel MPUs. As discussed in Byrne et al. (2018) at length, the price trajectories for Intel MPU models during the period of 60 to 100 percent annual price declines in these indexes are extraordinary and the plausibility of the indexes using these

Figure 16: Alternative Price Indexes for MPUs

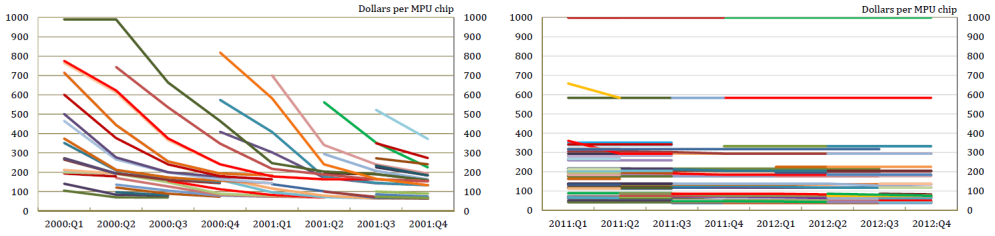


Note: Indexes differ with respect to price data source, methodology, and scope (see text).
Source: Bureau of Labor Statistics, Grimm (1998), Aizcorbe, Corrado, and Doms (2003), Flamm (2007), Holdway (2000), Byrne, Oliner, and Sichel (2018), Sawyer and So (2017).

prices depends critically on the interpretation of these model price declines (fig. 17a).

Figure 17: Intel MPU Model Price Trajectories

(a) 2000-2001 (b) 2011-2012



Note: Intel list prices for 1,000 unit trays of personal computer MPUs.
Source: Byrne et al. (2018).

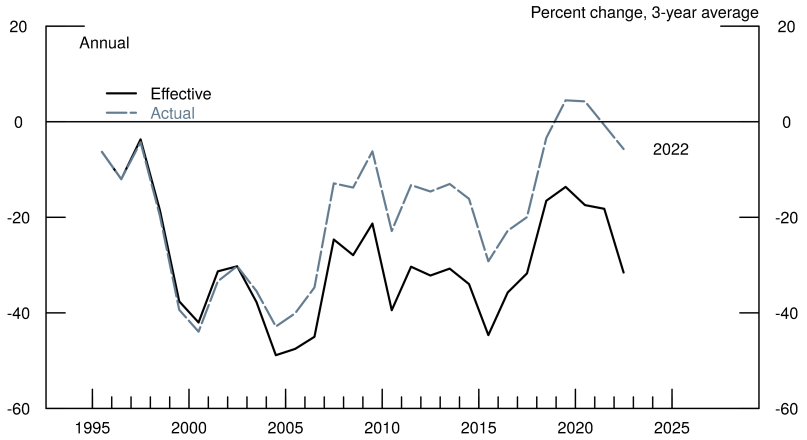
The most straightforward interpretation is that these model level prices reflect the falling costs of producing a chip of a given quality as miniaturization advances, circuit design improves, and yields (the share of chips produced that perform as intended). An alternative explanation hinges on market power and price discrimination. In a time when the downstream market

for computers employing the very fastest MPU was intense, a plausible explanation for these price trajectories is that Intel—with noteworthy market power—was able to price discriminate and charge early adopters a substantial premium when new chips were introduced, followed by a much lower price for trailing-edge consumers within a short period of time. If that is the case, linking prices paid by two very different consumer types together as is done in all four cases (using distinct methods) may create inflated price index declines (Williams and Sager (2019); Aizcorbe, Bridgman and Nalewaik (2010)).

As with the other indexes, the contour of our MPU price index includes a substantial slowdown around 2005. Again, two factors are in play in this development. On the supply side, engineering developments retarded the industry's delivery of faster and cheaper chips at that time. In particular, the industry shifted from successive rounds of miniaturization of the same processor architectures (primarily Intel's x86 architecture) toward design innovation, including a focus on parallel processing using multiple computing cores on the same chip. This shift was spurred in part by the realization that further miniaturization could not be translated into a higher clock rate because the consequent increase in heat generation could not be adequately offset by the cooling system in the computer. On the demand side, the premium paid by early adopters appears to have disappeared as Intel model price trajectories changed from steep to shallow (fig. 17b on the preceding page).

From the time of the slowdown forward, recent work by Byrne et al. (2018) and Sawyer and So (2018) (as reflected in the BLS PPI) tells rather different stories about MPU prices. The former shows sustained rapid declines of roughly 32 percent, albeit at a slower pace than during the tech boom, and the latter shows a pronounced slowdown to single digit price declines, averaging 18 percent over the same period. In the context of concerns about a slowdown in innovation in the semiconductor industry—myriad stories about the end of Moore's Law, and so forth—and stubbornly low aggregate productivity growth, the lack of consensus is unfortunate. In effect, our results contradict both stories; average annual price declines from 2005 to 2014, beyond which the BLS PPI is no longer disclosed, are 19 percent in our index, roughly the same as in the PPI on average and far smaller than Byrne-Oliner-Sichel. However, the contour of price declines is quite different: Our index enters the period falling substantially slower than the other two and shows no slowdown during the period.

Figure 18: MPU Price Indexes using "Effective" and Actual Nodes



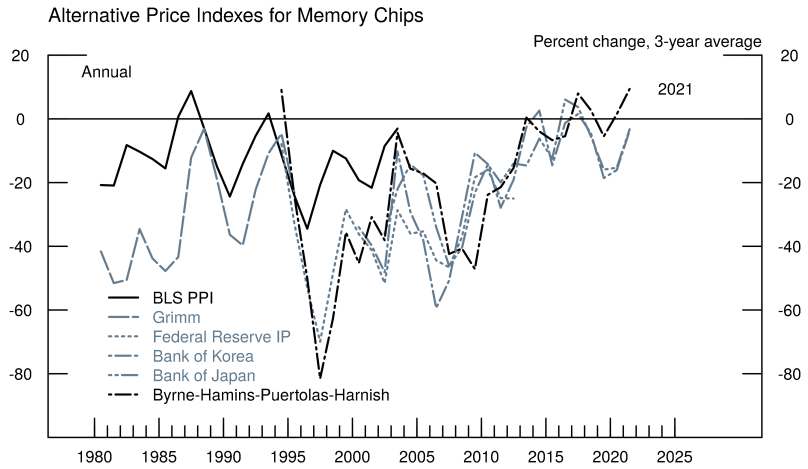
Note: Effective geometry refers to node labeling employed by the industry. Actual geometry is the average of metal half-pitch and contacted gate half-pitch reported in the International Technology Roadmap for Semiconductors.
Source: Authors' calculations.

Beyond the end of existing price indexes, we find that price declines remain in the neighborhood of 20 percent until 2017 before slowing markedly to single digit declines on average. Data construction is crucial for this point. The slowdown would not be apparent if we were to treat the "effective node" labels used by the industry as measures of the actual geometry used in semiconductor fabrication (fig. 18. Instead, as discussed above, we use the actual dimensions of the transistors as shown in fig. 9 on page 14.

7.2.1 Prices: Memory

Our memory price index aligns remarkably well with existing indexes throughout its history with the exception of the BLS PPI, which is an outlier for much of its history (section 7.2.1 on the next page). In contrast to the MPU index results, we find that our index matches the extraordinarily fast price declines found in existing indexes in the late 1990s. The slowdown in memory price declines that occurs in existing indexes from 2010 forward is also found in our index. The only noteworthy tension is found from roughly 2018 forward, when our index roughly moves sideways while the other indexes fall 10-20 percent per year.

Figure 19: Alternative Price Indexes for Memory Chips



Note: Indexes differ with respect to price data source, methodology, and scope (see text).
Source: Bureau of Labor Statistics, Federal Reserve Board, Grimm (1998), Bank of Korea, Bank of Japan.

7.2.2 Prices: Other Logic

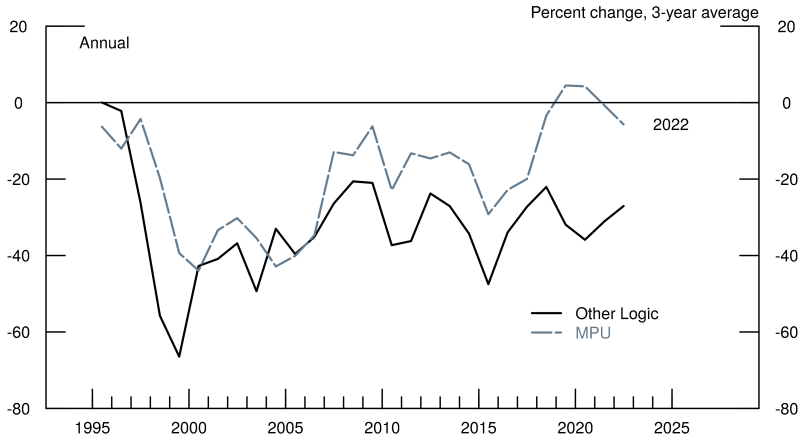
Our prices index for other logic—GPUs, TPUs, ASICs, etc.—is the only research index available for this semiconductor type, to our knowledge.[27] It falls 20-40 percent per year on average in the period of our study. From 2005 forward the index for other logic consistently falls significantly faster than MPU prices (fig. 20 on the following page). In light of the rising importance of this type of chip, which is used in place of MPUs in some high-performance computing applications and is found embedded in a wide array of products, this index sheds important light on the present pace of technical change in electronics and what may be expected in the near future.

8 Conclusion

Consistency in measures of output and prices is crucial for the study of economic questions that unfold in an integrated fashion across the global economy, that play out over long time periods, or that involve substantial product heterogeneity. For the broad electronics sector and for semiconductor production in particular, such questions include the contribution of IT to

[27] Coyle et al. have one under development.

Figure 20: Price Indexes for Other Logic and for MPUs



Note: Other logic includes hardware accelerators (e.g. GPUs), ASICs (application specific integrated circuits), and miscellaneous other non-memory, non-MPU logic chips.
Source: Authors' calculations. (See text for details.)

productivity growth, the role of supply chains in shock propagation, how the differing roles of economies in the global value chain have evolved over time, and the diffusion of innovation. The current IT measurement landscape is an incomplete patchwork quilt of measures for particular products, time periods, and markets that cannot be used to study such questions without unwieldy layers of caveats about differing methodologies, data sources, and necessary interpolations for data gaps.

We demonstrate a direct volume measurement approach that meets these consistency conditions and is feasible with readily available data. We appeal to the strong relationship between the number of transistors on a chip and its performance and use transistors as an indicator for real output. We construct approximate transistor counts using plant capacity adjusted for utilization and detailed information on the technology employed in each plant. These measures are constructed separately by chip type and aggregated using a Fischer ideal index formula to create country industry level output and price indexes. Ultimately, the data sources will support estimates for all product types in all relevant economies from the dawn of the industry to the present. In the present version, we present estimates for MPUs, memory chips, and other logic for five key economies—China, Japan, South Korea, Taiwan, and the United States—for the thirty-year period ending in 2022.

More refined approaches are feasible, but after 50 years of work on the subject in the measurement field, to continue to approach the problem in a piecemeal fashion is to let the perfect be the enemy of the good. For products and time periods where conventional measures are available, our estimates are broadly consistent in most cases, providing additional corroboration that our approach is reasonable. For memory chips, our results are little different from other estimates. However, there are cases where our results tell a rather different story.

For MPUs, a sudden acceleration in price declines in the mid-1990s and reversal in the mid-2000s is present in our index, a contour that matches existing estimates. However, the magnitude of the regime change is much smaller. In recent years, we find a further slowing in MPU price declines, but one that arrives much later than in the official PPI; and, we find a lower average rate of decline than Byrne et al. (2018).

For "other logic," we find that prices fall much faster than is indicated by the price index used for industrial production at the Federal Reserve Board. And, we find evidence that this product class, which includes some of the most technologically dynamic products in the electronics sector—hardware accelerators like GPUs and TPUs as well as the custom logic chips (ASICs) that embody the features of smartphones and many other devices—has substantially faster falling prices than MPUs, as one might expect for a world where cutting-edge computing is increasingly relying on these alternatives to MPUs.

References

Aizcorbe, Ana, Benjamin Bridgman, and Jeremy Nalewaik, "Heterogeneous Car Buyers: A Stylized Fact," *Economics Letters*, 2010, *109* (1), 50–53.

Aizcorbe, Ana M., Carol A. Corrado, and Mark E. Doms, "When do Matched-Model and Hedonic Techniques Yield Similar Measures?," *FRB of San Francisco Working Paper*, 2003, (2003-14).

Allen, Liz, Alison O'Connell, and Veronique Kiermer, "How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship," *Learned Publishing*, 2019, *32* (1), 71–74.

Arel-Bundock, Vincent, Nils Enevoldsen, and CJ Yetman, "countrycode: An R Package to Convert Country Names and Country Codes," *Journal of Open Source Software*, 2018, *3* (28), 848.

Berg, Joseph Frederick and Joseph Barker, *Great Discussion on the Origin, Authority, and Tendency of the Bible: Between Rev. JF Berg, DD, of Philadelphia, and Joseph Barker, of Ohio*, G. Turner, 1854.

Białek, Jacek, *PriceIndices: Calculating Bilateral and Multilateral Price Indexes* 2022. R package version 0.0.8.

Bryan, Jennifer, *cellranger: Translate Spreadsheet Cell Ranges to Rows and Columns* 2016. R package version 1.1.0.

Burg, David and Jesse H Ausubel, "Moore's Law revisited through Intel chip density," *PloS one*, 2021, *16* (8), e0256245.

Byrne, David M., Brian K. Kovak, and Ryan Michaels, "Quality-Adjusted Price Measurement: A New Approach with Evidence from Semiconductors," *Review of Economics and Statistics*, 2017, *99* (2), 330–342.

_ , Stephen D. Oliner, and Daniel E. Sichel, "How Fast are Semiconductor Prices Falling?," *Review of Income and Wealth*, 2018, *64* (3), 679–702.

Capital Markets section, Research and Statistics Division, Federal Reserve Board, Collin Harkrader, Brian Seok, Holt Dwyer, Scott Konzem, and Alex Deluca, *policyPlot: Functions to Make Easy Tealbook Style Plots* 2017. R package version 0.6.9.

Chwelos, Paul, "Approaches to Performance Measurement in Hedonic Analysis: Price Indexes for Laptop Computers in the 1990's," 2003.

Dulberger, Ellen R., "Sources of Price Decline in Computer Processors: Selected Electronic Components," in "Price measurements and their uses," University of Chicago Press, 2007, pp. 103–124.

Flamm, Kenneth, *Targeting the computer: Government support and international competition*, The Brookings Institution, 1987.

_ , "Has Moore's Law been repealed? An economist's perspective," *Computing in Science & Engineering*, 2017, *19* (2), 29–40.

Gartner, "Fab Database: Worldwide," Technical Report, Stamford, CT 2007.

Hallman, Jeff, *fame: Interface for FAME Time Series Database* 2018. R package version 2.23.

_ , *tis: Time Indexes and Time Indexed Series* 2020. R package version 1.38.

Holdway, Michael, "An Alternative Methodology: Valuing Quality Change for Microprocessors in the PPI," in "Unpublished paper presented at Issues in Measuring Price Change and Consumption Conference, Bureau of Labor Statistics, Washington DC, June" Citeseer 2000, pp. 5–8.

Intel, Inc., "Intel ARK Database," Technical Report 2022. [Online; accessed 30-August-2022].

Mashey, John, "SPEC results help normalize vendor Mips-ratings for sensible comparison," *SPEC Newsletter*, 1990, *2* (3), 1.

Moore, Gordon E., "Cramming more components onto integrated circuits," *Electronics Magazine*, 1965, *4*, 114–117.

Moore, Gordon E, "Progress in digital integrated electronics," in "Electron devices meeting," Vol. 21 Washington, DC 1975, pp. 11–13.

_ , "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp. 114 ff.," *IEEE solid-state circuits society newsletter*, 2006, *11* (3), 33–35.

Nordhaus, William D, "Two centuries of productivity growth in computing," *The Journal of Economic History*, 2007, *67* (1), 128–159.

Ohta, Makoto and Zvi Griliches, "Automobile Prices Revisited: Extensions of the Hedonic Hypothesis," in "Household production and consumption," NBER, 1976, pp. 325–398.

Ooms, Jeroen, *pdftools: Text Extraction, Rendering and Converting of PDF Documents* 2021. R package version 3.0.1.

R Core Team, *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing 2021.

Sawyer, Steven D. and Alvin So, "A New Approach for Quality-Adjusting PPI Microprocessors," *Monthly Lab. Rev.*, 2018, *141*, 1.

Schauberger, Philipp and Alexander Walker, *openxlsx: Read, Write and Edit xlsx Files* 2021. R package version 4.2.4.

SEMI, "World Fab Forecast," Technical Report, Milpitas, CA 2022.

Triplett, Jack E., "Performance Measures for Computers," *Deconstructing the Computer, National Academy Press, Washington, DC*, 2005, pp. 99–143.

U.S. Census Bureau, "Quarterly Survey of Plant Capacity Utilization," Technical Report 2022.

Vries, Dirk K De, "Investigation of gross die per wafer formulas," *IEEE Transactions on Semiconductor Manufacturing*, 2005, *18* (1), 136–139.

Wickham, Hadley, "Reshaping Data with the reshape Package," *Journal of Statistical Software*, 2007, *21* (12), 1–20.

_ , Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache,

Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani, "Welcome to the tidyverse," *Journal of Open Source Software*, 2019, *4* (43), 1686.

Williams, Brendan and Erick Sager, "A New Vehicles Transaction Price Index: Offsetting the Effects of Price Discrimination and Product Cycle Bias with a Year-Over-Year Index," *Economic Working Papers*, 2019.