Historical Geography of the Semiconductor Industry

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

March 7, 2022

Abstract

The production and the use of solid-state electronics–devices built using semiconductors–have played a central role in economic developments in the post-war era. Most recently, disruption of the global electronics supply chain has slowed emergence of the global economy from the COVID-19 pandemic. Prior to that, concerns about a slowdown in electronics innovation–the end of "Moore's Law"–colored the outlook for productivity growth. Analysis of these and other topics requires data on the electronics sector that is fundamentally global in scope. We take a step in that direction, building a dataset for an industry central to the sector, semiconductor fabrication, covering the industry from 1996 forward, consistently measured over time and across countries. We use these data to shed light on three questions: whether the industry has entered a period of slower innovation, whether the roles of the United States, Taiwan, China, and other countries have changed in recent years, and whether a simple indicator–the industry's "transistor budget"–provides a reasonable measure of output free of the complications of quality-adjusted price indexes.

*Federal Reserve Board of Governors, Washington, D.C.: david.m.byrne@frb.gov. Additional contributors to this research, described using the CRediT taxonomy (Allen, O'Connell and Kiermer, 2019), are Andrea Garcia (data curation), Steve Young (funding acquisition). The views expressed in this paper are not represented to be the views of the Federal Reserve Board of Governors. This is preliminary work. Please do not cite without consulting the authors. Corresponding author: Dave Byrne, david.m.byrne@frb.gov.

1 Introduction

The production and the use of solid-state electronics–devices built using semiconductors–have played a central role in economic developments in the post-war era. Most recently, disruption of the global electronics supply chain has slowed emergence of the global economy from the COVID-19 pandemic. Prior to that, concerns about a slowdown in electronics innovation–the end of "Moore's Law"–colored the outlook for productivity growth. Going back further, electronics played a central role in the emergence of Silicon Valley as a global innovation hub, trade conflicts with Japan in the 1980s, the emergence of the "Asian Tigers" (South Korea, Taiwan, Singapore, and Hong Kong), the struggle of China to climb up the production value chain, the economic boom of the late 1990s, and current geopolitical tensions over the status of Taiwan.

Analysis of these topics requires data on the electronics sector that is global, distinguishes steps in the supply chain, and spans the full history of the industry. We take a step in that direction by building a dataset for an industry central to the sector—semiconductor fabrication—using disparate sources of data on the plants, known as "fabs," where circuits are created on raw silicon wafers, consistently measured over time and across countries from 1996 to 2021.[1]

Leveraging these data, we provide indicators for the pace of technical advance in the industry, the rate of growth in real output of chip fabrication, the location of the leading edge of production, and the location of basic and applied R&D.

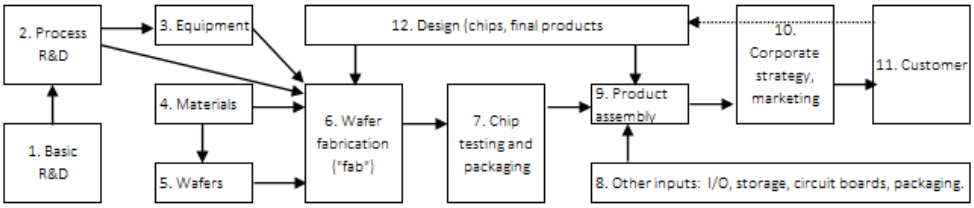
2 The Electronics Sector and the Semiconductor Fabrication Industry

The electronics sector is composed of distinct industries devoted to circuit design, wafer fabrication, chip test and packaging, and assembly of consumer and business electronic products (fig. 1).

The fabrication of semiconductors, in contrast to the earlier technology based on signal transmission via gasses in vacuum tubes, entails the creation of electronic circuits on a substrate of semiconducting material, most

[1] In research underway, we extend this dataset back to the inception of the industry.

Figure 1: The Electronics sector and its industries.



commonly silicon. These devices may be discrete components–individual resistors, capacitors, transistors, and so forth–or integrated circuits (ICs) which perform logical functions such as mathematical transformations (processors) or holding data to support processing (memory). Typically, multiple copies of the same device are created on a silicon wafer in a series of steps involving addition and removal of materials on its surface.

The manufacturing of the specialized equipment used in the fabrication of electronic devices on silicon wafers plays a central role in the pace of technical advance. Most notably, advances in the equipment that performs the photolithography to deposit a film on the surface embodying the circuit design governs the level of miniaturization achievable. Technical advances focus in the main on creating devices on a smaller and smaller scale, leading to lower cost of production, lower power consumption, and faster task execution.

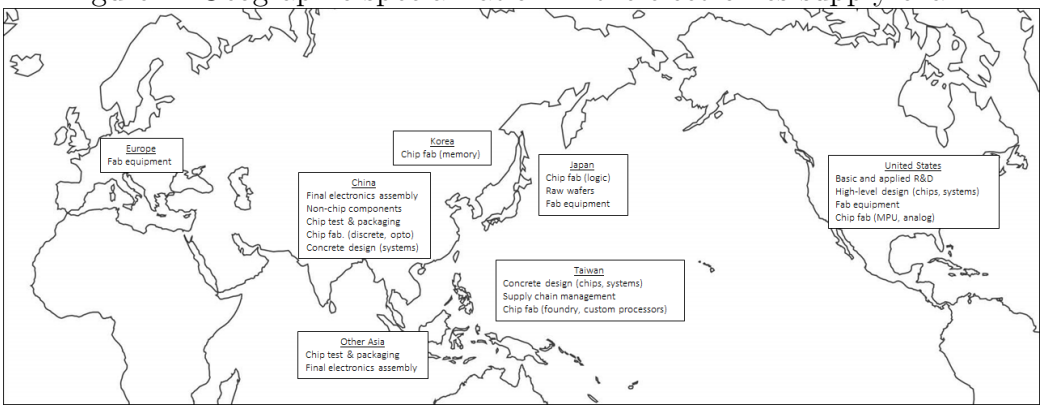
Wafer fabs house the ultra-clean environment where this equipment is used to conduct the dozens of steps needed to gradually build the device. Here, technical advance involves engineering the production line, careful calibration of the equipment, and refinement of the process to raise the yield–number of usable chips per wafer–for particular technologies and specific chip designs.

The subsequent testing of each individual device on the completed wafer and the packaging of viable devices is a distinct industry in itself. And, the design process, where engineers use specialized software and engage with the end-user (a computer manufacturer, for example) and the chip manufacturer to create in circuitry the functional identity of the final equipment, is a separate process as well. Finally, there is the manufacture of specialized materials, notably including the manufacture of ultra-pure raw silicon wafers.

The electronics sector is managed by firms encompassing a shifting collection of plants, often from multiple industries within the sector. We focus

on industry-level analysis, and on the fabrication stage in particular. Early in the history of the industry, the entire supply chain may have been located under one roof but gradually the stages moved to distinct locations (Bauer and Wilder, 2020). Nowadays, materials production, semiconductor equipment manufacturing, chip fabrication, test and assembly, and design most often take place at distinct establishments and often in separate countries 2. The feasibility of physical separation and the different production functions in use at each stage argue in favor of treating each stage as a distinct industry as we do. That is, there are electronics industries, not a singular industry. Our focus in this paper is on the location and the output of the semiconductor fabrication industry.

Figure 2: Geographic specialization in the electronics supply chain.



3 Data Construction

To address the drawbacks of data currently available from national statistical agencies discussed below, we construct a comprehensive, global, and consistent dataset of the fabrication industry. We primarily rely on data from two sources for the construction of our database. Gartner, a consultancy focused on high tech industries, published occasional comprehensive global lists of semiconductor fabs as part of their *Semiconductor Equipment and Materials Service* beginning in 1985. We located and digitized 22 of these lists for the period from 1996 to 2007. Many reports contain a plant identification number that remains unchanged over time. For three reports that did not, we were able to match plants based on location, ownership, and capacity

information in all but 24 cases. We fill in gaps in the data–both unmatched plants and quarters for which we have no reports–by filling values in from the most recent previous observation.[2] SEMI (Semiconductor Equipment and Materials International), a trade association focused on semiconductor manufacturing equipment, publishes a comprehensive, quarterly-frequency database of fabs, the *Wafer Fab Forecast*. We use two vintages of this report that together provide history spanning the period from 2008 to 2022.

We exclude plants included in these databases not devoted to fabricating microchips. Among these exclusions are plants producing MEMS (micro-electromechanical systems), epitaxial wafers (primarily used for solar cells), and "back end" plants (used for testing completed wafers). We also exclude plants recorded as planned or under development and plants recorded as closed or inactive. Finally, we exclude academic and government research plants, though we do discuss the location of these plants below. After this process, we have roughly 700 to 900 plant level observations at any given point in the dataset (fig. 3). We are in the process of linking individual plants between the Gartner and SEMI datasets. In the meantime, we construct aggregates for each period separately and link the data sources at the aggregate level.

Figure 3: Descriptive statistics for selected variables.

Indicator	N	Mean	Median	St. Dev.	Min	Max
Capacity (thousands of wafer starts per month, 8-inch wafer equivalent)						
1996	872	7,900	5,821	7,968	7	78,227
2006	662	20,173	11,256	25,550	11	247,626
2016	818	23,583	7,844	52,170	3	722,250
2021	910	27,387	10,000	45,963	3	382,500
Geometry (nanometers)						
1996	872	1,741	1,000	2,882	90	50,000
2006	662	1,012	500	1,110	65	3,000
2016	818	1,047	500	1,376	7	20,000
2021	910	966	500	1,327	3	20,000
Wafer diameter (millimeters)						
1996	872	129	125	38	25	200
2006	662	159	150	59	51	300
2016	818	155	150	75	25	300
2021	910	164	150	78	51	450

Source: Authors' calculations using data from Gartner, Inc. (1996, 2006) and SEMI, Inc. (2016, 2021).
Note: The fourth quarter of data is used for the calculations for each year, except for 1996, which is an annual value. 2006 values for geometry are topcoded in the raw data.

Each data source provides plant-level information on location, company ownership, technology employed, wafer capacity, and product mix.[3] Detailed

[2]This preserves the timing of the first appearance of new technologies. We will refine the interpolation in future drafts.

[3]In addition, both datasets is the provision of the data on when the plant was established. This information will provide the foundation for future efforts to construct the

information is available by plant for the scale of production and technology employed. In particular, the data provide plant capacity in the number of silicon wafers that can be processed in a month, diameter of wafer employed, and a measure of the maximum level of miniaturization achieved at the plant, known as geometry.[4] The geometry of a technology, sometimes referred to as linewidth, is a barometer of the technical limitations of a technology generation. These generations, also known as nodes, were well characterized by the width of the transistor's gate–the on-off switch used to control electron flow through circuits–through the 1990s. Since that time, the meaning of the geometry measure has varied as the focus of the industry has shifted to other features, but it remains a barometer of the industry's progress and a focus of discussion. Geometry is the measure most commonly employed to characterize the state of technology in the industry, though the dimension of the largest wafer in use plays an important role as well.

We employ wafer start capacity and wafer diameter to construct 8-inch wafer equivalent capacity by plant. The capacity of the average plant in our dataset rises from 7,900 wafer starts per month in 1996 to 27,400 wafer starts per month in 2021. Over time, the average wafer diameter rises and the average geometry falls, though only modestly in both cases. However, the maximum wafer diameter rises substantially from 200 to 450 between 1996 and 2021 and the minimum geometry employed falls dramatically from 90 nanometers to 3 nanometers.

Using geometry, diameter, and capacity, we construct an indicator of the "transistor budget" capacity for each fab, dividing the area of silicon represented by the wafer capacity by the square of the geometry to yield a measure which moves in proportion to the number of features that may theoretically be placed on a wafer. As chip fabrication is not literally an exercise in fitting as many transistors on a chip as theoretically possible, we benchmark our measure using the density of transistors on a leading edge personal computer MPU in 2006, the Intel Core 2 Duo. The average fab in 1996 was capable of producing about three trillion transistors per month by this measure. By 2021, the average had risen to 9 quadrillion.

dataset from the dawn of the industry. For example, the earliest date for plant establishment in the data is 1953.

[4]Importantly, the reports do not provide information on realized production. Most of our analysis of industry trends focuses on capacity, though when we compare our indicators to real output indicators we adjust for utilization, assuming global industry utilization rates match the Federal Reserve Board estimates for the U.S. industry.

Figure 4: Descriptive statistics for measures of transistor capacity.

	N	Mean	Median	St. Dev.	Min	Max
Transistor capacity (billions, transistors per month)						
1996	872	2,797	420	7,233	0.03	110,444
2006	662	114,801	3,932	301,279	0.12	2,955,132
2016	818	3,611,768	1,216	19,846,732	0.06	310,292,690
2021	910	9,071,001	1,814	42,088,064	0.06	604,152,433
Transistors per chip (millions)						
1996	872	4.1	1.2	11.4	0.0005	151.9
2006	662	35.4	4.9	60.2	0.1368	291.3
2016	818	556.8	4.9	2,031.4	0.0031	25,117.7
2021	910	1,889.6	4.9	10,394.0	0.0031	136,752.1

Source: Authors' calculations based on data from Gartner, Inc. (1996, 2006) and SEMI, Inc. (2016, 2021).
Note: The fourth quarter of data is used for the calculations for each year, except for 1996, which is an annual value. Transistors calculated as wafer or die surface area divided by square of geometry, divided by a scaling factor to benchmark to actual transistor density in 2005 for Intel Core 2 Duo microprocessor.

For Gartner, we identify fabs as R&D by the indicator found in some of the vintages of the report. In the case of SEMI, we identify fabs as research by searching for the strings "R&D" and "research" in the fab description fields found in each database. R&D fabs are identified as government or academic using a list of keywords such as "university" in the fab description. Our analysis of R&D fab location is limited to the period from 2008 forward for the time being, though we anticipate augmenting the lab history in future iterations.

We then construct a set of aggregates for use in the analysis of the state of technology in the industry and the geography of the industry. Excluding R&D fabs, we calculate market share for each country using 8-inch wafer capacity as well as two indicators of the technology employed in each national industry: the mean technology employed and the most advanced technology employed. When calculating the most advanced technology we exclude fabs with reported wafer capacity of 2,000 wafers per month or less, a rough way to exclude small scale fabs, which focus on the production of exotic technologies, and fabs in the early stages of ramping up a new technology.

4 The Pace of Technical Change

We consider first how the dataset can shed light on the question of the pace of technical change in the semiconductor fabrication industry. The indicators we have constructed reflect the scientific and engineering advances that enable miniaturization. Other forms of innovation, most notably advances in design, are not reflected, which we view as an opportunity rather than a drawback as it offers an opportunity to contrast our indicator with a measure of real output of the final product–a completed microchip ready for

use in a computer, say–which does incorporate the value added of design.

4.1 Moore's Law

In 1965, Gordon Moore noted that as the complexity of integrated circuits increased–measured in the number of components used, such as transistors–the cost of production fell (Moore, 2006). In particular, in place of components constructed as distinct products and connected into a system (on a circuit board, say), pulling components together onto a single piece of silicon lowered costs dramatically. However, the desirability of continuing this process of cramming more features onto a chip is limited by the rising probability that a component in the integrated circuit would fail as the number of components rose. At any time, given the state of technology, there was a minimum-cost complexity, or number of components. Moore observed that "the complexity for minimum component costs has increased at a rate of roughly a factor of two per year" and speculated that technical change in semiconductor manufacturing would support a continuation of that trend–doubling of the number of transistors on an integrated circuit every year–for at least ten years. Revisiting the issue in 1975, he noted that engineers had exhausted the low hanging fruit represented by optimizing circuit layout and going forward, further advances would be more subdued, and he projected a doubling of the number of components on integrated circuits every *two* years, a prediction commonly known as "Moore's Law."

The first widely available integrated circuit to incorporate the entirety of a logic processor was Intel's 4004 microprocessor, released in 1971 and used in the electronic calculators that were the hot high-tech consumer product of the early 1970s. The size of the die–the area of the chip–was 12 square millimeters and the geometry employed was 10 microns, suggesting a value for our transistor budget indicator of 120,000. Importantly, this does not literally mean this many electronic components (transistors, capacitors, and so on) could be fit on the chip and combined to create a functioning integrated circuit. (Creating integrated circuits is not an exercise in literally packing as many transistors onto a chip as physically possible.) In practice, the 4004 chip contained 2,250 transistors. Starting from 1971, if Gordon Moore's speculation had come to pass 50 years into the future, the transistor budget indicator for an integrated circuit would be higher by a factor of 30 million in 2021.

4.2 Geometry

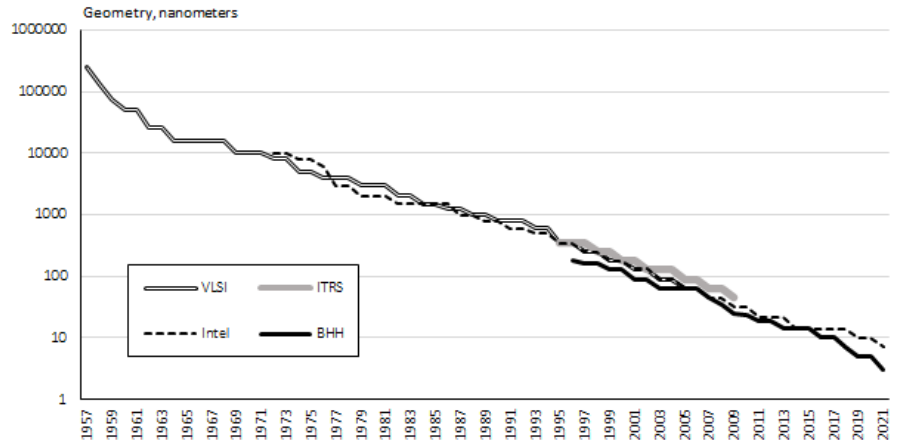
The pace of advance in geometry is the most prominent focus of the fabrication and equipment portion of the industry, serving as a benchmark for the state of technology. The industry calibrates this technical pace in terms of nodes, which serve as focal points for basic and applied research as well as development of commercially viable products. Each node represents a roughly 70 percent shrink in the linear dimension of the features on a chip, implying that a given number of features occupies half the planar area on a wafer (squaring 70 percent yields roughly 50 percent). In practice, companies develop technologies in the neighborhood of these nodes, rather than the exact geometry, so there is a bit of judgement involved in concluding a particular node has been reached.

We consider first some available indicators of progress in the industry. Beginning in the 1980s, the industry published a roadmap identifying the technical hurdles required to reach the upcoming nodes, which served as a guide for basic and applied researchers contributing to the project, the ITRS (International Technology Roadmap for Semiconductors) (Semiconductor Industry Association, 2007). The node history from 1995 to 2009, cobbled together from ITRS documents–no single document provides the full history–is shown in fig. 5 on the following page. The pace of advance was irregular, but over time, the length of time between nodes in this period averaged a little over two years. The pace of technical advance for Intel is also shown, for which a longer history is available (WikiChip contributors, 2022). Historically, nodes have sometimes been reached first by memory manufacturers, but among processor manufacturers, Intel is typically the first to achieve volume production at a particular node. Prior to the time period covered by the ITRS series, advances between nodes were less regular. Between 1971 and 1995, the average length of time between nodes was 4 years. Last, estimates from VLSI Research Inc., a consultancy with deep history in the industry, is shown back to 1957.[5] These are broadly consistent, but not identical to the ITRS and Intel node history.

Each of these measures is authoritative but has disadvantages from the point of view of economic research. The Intel measure is company-specific and, since in recent history Intel has been replaced by TSMC (Taiwan Semiconductor Manufacturing Corporation) as industry leader, is no longer a good indicator of the leading edge. The VLSI measure represents the judgement

[5]These estimates were kindly provided by G. Dan Hutcheson.

Figure 5: Semiconductor Technology Node History



Source: VLSI, Inc., International Technology Roadmap for Semiconductors, Intel, Inc., authors' calculations.
Note: Authors' calculations define a node as production using the geometry at a fab with capacity of at least 2,000 wafer starts per month.

of expert industry analysts, but is judgement nonetheless, not algorithmic. The ITRS series does provide a clear definition, but one that is somewhat difficult to implement independently. In particular, they reference the status of DRAM production technology in the 2011 roadmap, stating: "Please note that, unless otherwise specified for a particular line item, the default year header still refers (as in previous Roadmaps) to the year when product shipment first exceeds thousands of units per month of ICs from a manufacturing site using 'production tooling.' Furthermore, a second company must begin production within three months. To satisfy this timing definition, ASIC production may represent the cumulative volume of many individual product line items processed through the facility." It would be quite difficult for economists to consistently acquire the information needed to implement this definition.

Our approach is to identify a node as reached when the geometry is employed in production at a facility with capacity of at least 2,000 wafer starts per month. Below this threshold, one finds plants devoted to research, pilot lines, and boutique production. The appeal of our estimates is that we begin with widely available industry data and apply a simple formula, such that the estimates are replicable without reference to the analysis of any company, trade group, or consultancy. The average time between nodes

for our estimates is 2.1 years. The time between nodes is somewhat variable, with a notable slowdown in the early 2000s.

What do we make of Moore's Law? Extending our estimates backward to 1975 using the timing provided by VLSI, Inc., we count 20 nodes, from 10 microns to 3 nanometers, yielding 2.5 years between nodes. The average rate of linear shrinkage of (ln(10000)-ln(3))/50 = -0.16 implies an annual areal shrinkage of -0.16*2 = -0.32, which is sufficient to double the density of components on a chip of a given size every -0.69/-0.32 = 2.15 years (fig. 6). Recall that Gordon Moore's observation referred to the number of features per integrated circuit, and integrated circuits have grown larger since 1971. While the Intel 4004 chip was 12 millimeters squared, a typical leading edge Intel processor in 2005, the Intel Montecito, occupies about 600 square millimeters, a pace of increase in the chip area of approximately 8 percent per year over that period, after which the die size has remained fairly stable. Allowing for the larger size of these chips, average time for the number of potential features per integrated circuit to double is 1.73 years.

Figure 6: Moore's law calculations.

	1975-2021	1975-2003	2003-2021
Years to double components per die	1.73	1.53	2.03
Growth rate, components per die	0.40	0.45	0.34
Contribution from geometry	0.32	0.31	0.34
Contribution from die size	0.08	0.14	-
Memo:			
Years to double, geometry alone	2.15	2.20	2.03

Put differently, if one projects the number of components forward from the 1975, when the Intel 8080 was produced and Intel made his revised projection, using our measure of the transistor budget for ICs, and compares that projection to Moore's law, the industry has exceeded his prediction, as shown in figure fig. 7.

What of the uneven pace, however? It does appear that the pace of movement between nodes sped up in the mid-1990s, coinciding with the tech boom and intensified competition, most notably the battle between AMD and Intel for dominance of the processor market. Then, it slowed again in the early 2000s, remaining at remaining at 65nm for four years, leading many observers to remark on slowing innovation in the industry. At the

Figure 7: Moore's law prediction vs. technical advance.

	1975-2021	1975-2003	2003-2021
Years to double components per die	1.73	1.53	2.03
Growth rate, components per die	0.40	0.45	0.34
Contribution from geometry	0.32	0.31	0.34
Contribution from die size	0.08	0.14	-
Memo:			
Years to double, geometry alone	2.15	2.20	2.03

time, processor producers intensified focus on design, rolling out multi-core processors, for example, to complement research on miniaturization. The pace has sped up again recently, with the industry rolling out 8 nodes between 2006 and 2021, slightly faster than one every two years. On balance, it does appear that the pace of growth has slowed since 2003 for the potential features per integrated circuit. While the number of years to double from geometry alone has actually decreased from 2.20 to 2.03, in the earlier period, the die size was increasing, and adding that effect, years to double has increased from 1.53 to 2.03 since 2003 (fig. 6).

4.3 Prices

Moore's Law and the pace of miniaturization are technical measures of the pace of industry innovation. While these engineering changes are ultimately driven by market forces, they are not a direct measure of what economists are interested in: namely, the price of electronics holding quality constant. While economists have made valiant efforts to measure these prices, the rapid changes in product quality confound these efforts to a great degree. For selected products, most notably memory chips (DRAM and Flash memory, especially) and microprocessors (MPUs), academic and government research has yielded plausible price indexes. But these products account for only 40 percent of global nominal shipments of semiconductors. For the remaining products, little progress has been made. Included in the group of products for which there has been no rigorous research on quality-adjusted prices are the application-specific processors employed in many of the most advanced products, such as smartphones; the digital signal processors at bridge the divide between analog real world signals (sound, for example) and the world of digital electronics; the sensors employed in motor vehicles and many other platforms; and many others.

In any event, the aggregate price index for global total industry output employed by the Federal Reserve Board indicates that prices for semiconductors, employing the quality adjustments available, fell an average of 18 percent per year from 1996 to 2021.

Measured prices fell at roughly twice this average pace from 1996 to 2001, during the tech boom when IT production and use accounted for the bulk of the surge in labor productivity in the United States. Since that time, prices have slowed, most notably declining only 4 percent from 2015 to 2021. In light of the sustained pace of advance in technology, the weak decline in quality-adjusted prices is puzzling.

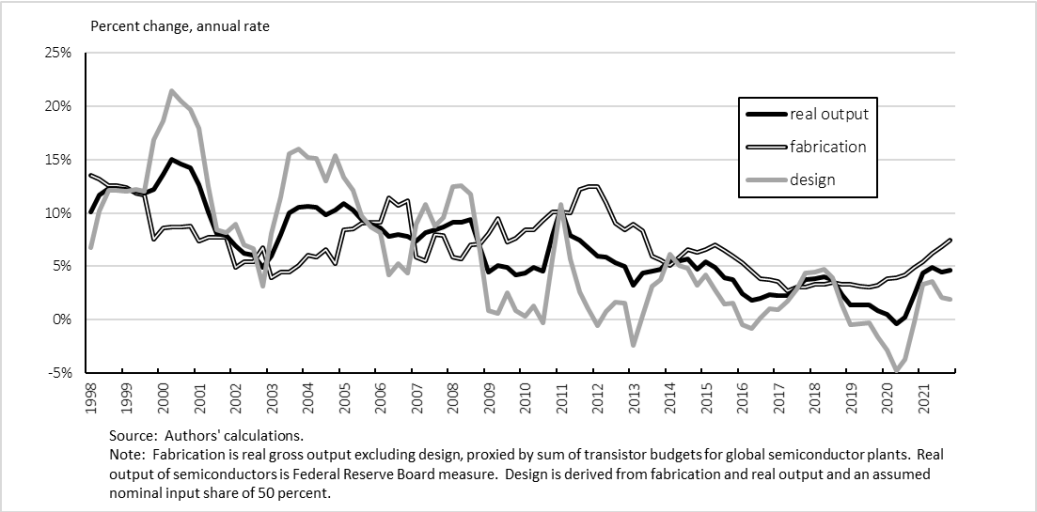
4.4 Prices and Geometry Together

Employing the price index to deflate nominal output, fig. 8 on the next page depicts the rate of growth in the real gross output of the semiconductor industry. Also shown is the rate of growth in an indicator for technical advance for the chip fabrication industry: the transistor budget for the aggregate global capacity for production of semiconductors.[6] That is, the ability of the industry to provide a greater transistor budget for circuit design in a given surface area is a measure of technical change at the fabrication stage. To arrive at a production indicator, we scale the capacity indicator by the measure of capacity utilization for the U.S. industry provided by the Federal Reserve Board, assuming that the operating rate for the global industry is the same.

Showing the indicator for gross output of the fabrication stage in combination with the real output of the semiconductor industry as currently measured provides an interesting contrast. Taking these indicators at face value–that is, ignoring measurement issues for the moment–yields interesting implications for the story of the industry. The real gross output of the semiconductor industry is the sum of the contribution from the fabrication stage and the value added contributed by design, management, and test and as-

[6]We assume that the effect of factors that one would account for in a literal measure, such as the limitation on usable surface implied by creating square chips on a circular surface and the yield, are a fixed proportion over time. That is, we are constructing an indicator, not an exact measure. More importantly, the ability of the industry to accommodate more complicated designs over time, through a greater number of layers, is omitted. This makes the growth rate of the indicator for fabrication gross output biased downward, which, for the most part, does not undermine our broad conclusions.

Figure 8: Decomposition of real output of semiconductors.



sembly. Using the gross margin of major semiconductor firms as a barometer of the nominal weight of this residual category suggests that these remaining categories account for about 50 percent of the gross output of the industry. This implies an unseen pace of real value added from these remaining categories shown in the dashed line in the figure, which we refer to as "design."

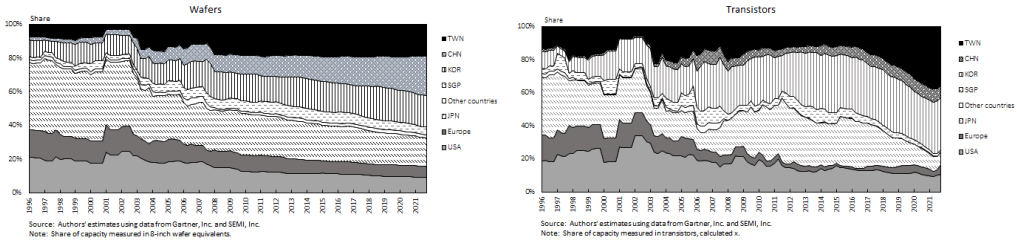
What story does this tell for the industry over the past 30 years? In the late 1990s and early 2000s, both fabrication and design accounted for a substantial share of real output gains in the industry. More recently, both fabrication and design contributions to real output growth have moderated somewhat, but the design contribution has fallen rather more. What should we conclude from this? Have advances in design in the semiconductor industry stalled? Returning to the issue of measurement, an alternative explanation is that design advances have shifted to chip categories that are relatively less well measured. The rate of decline in the price index for the categories encompassing the design embedded in smartphones and other devices that rely relatively less on MPUs and memory and relatively more on custom processors, which account for roughly half of nominal output for the global industry, has slowed substantially in recent years. Further inquiry into the performance gains of these chips is in order.

5 Industry Geography

5.1 Market Share

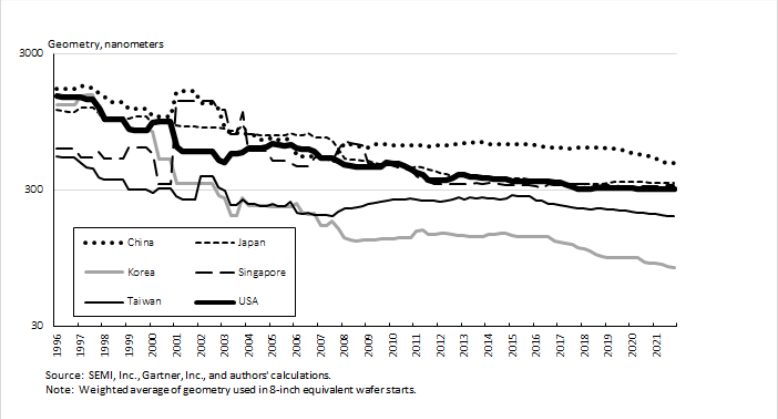
The geographic distribution of the semiconductor fabrication industry has changed dramatically over time. fig. 9 depicts the market share of key players in the industry. In terms of the number of wafers processed, Japan was the largest semiconductor fabricator in 1996. Over the 26 years covered in our data, Taiwan, South Korea, and China pushed into the market, reducing the share of Japan, especially, but also pushing down the share of Europe and the United States.

Figure 9: Country market share, semiconductor wafer fabrication.



Using our measure of transistor capacity–thus accounting the node of technology employed–the story is rather different for some countries. The South Korean fabrication industry has historically employed more advanced technology, on average, than most countries, leading them to have a more substantial market share and accounting for roughly a third of the market in 2021 fig. 10 on the next page. Taiwan uses highly advanced technology in the main as well, and increasingly so, leading to a dramatic increase in market share in recent years, to one-third as well. China and Japan, in contrast, focus on nodes that trail the leading edge. (In the case of Japan, this is somewhat misleading, as many of the specialty devices they produce are most appropriately produced at higher geometries.) In terms of transistors, the U.S. market share is little different than its market share in terms of wafers; the U.S. industry produces at a range of nodes, including leading edge production for MPUs and the trailing nodes used for many military and aerospace applications.

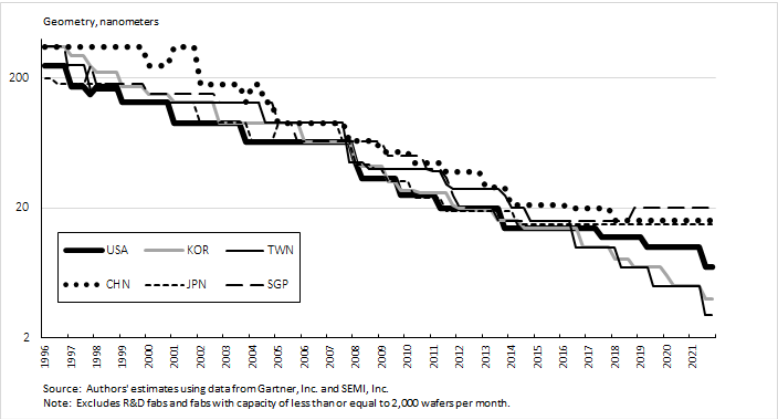
Figure 10: Average semiconductor fabrication geometry by country.



5.2 The Leading Edge

Where have the technologies shown in fig. 5 on page 9 first emerged? From at least 1996 until the late 2010s, the United States was the first location for the leading edge (fig. 11). Since then, the Taiwan industry has played this role.

Figure 11: Most advanced technology in volume chip fabrication by country.



More broadly, the United States, Taiwan, and Korea have wrangled over the leading edge in recent years, led by their national champions–Intel, TSMC (Taiwan Semiconductor Manufacturing Corporation), and Samsung, respectively. Japan and Singapore had kept up with the leaders through the early

2000s but have since fallen behind. China, hampered by trade restrictions that preclude acquisition of leading edge manufacturing equipment, has been well behind the peloton throughout this period, but has fallen even further behind in recent years.

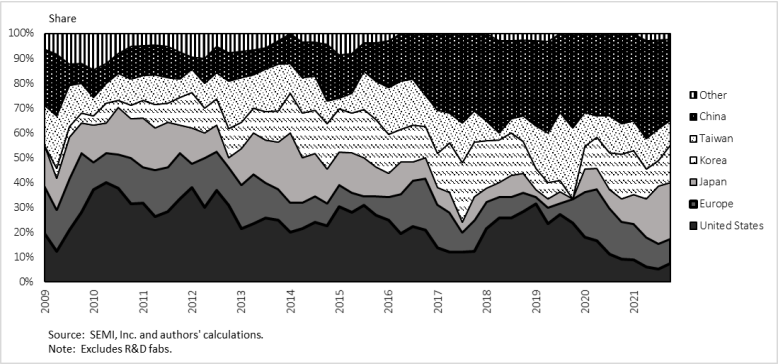
5.3 Location of New Fabs

In addition to the location of the leading edge, the location chosen for fabrication is an indicator of the direction the industry is headed as well. Addition of capacity to the industry and adoption of new technologies may take place through installation of equipment in clean room space at existing fabs. Eventually, though, companies face the decision of where to locate new fabs, a commitment of $10 billion or more and consequently a revealing choice about the anticipated comparative advantage of countries for industry activity in years to come. Here the trend is decidedly against location in Europe and the United States (fig. 12). These locations accounted for 50 percent of the new plants in 2011 and less than 20 percent in 2021. The role of China has more than doubled over that same time period, while the role of Japan, South Korea, and Taiwan has been little changed, on balance. Whether the recent flurry of new plant announcements will affect this distribution remains to be seen, as illustrated by the plans of the companies accounting for the lion's share of investment: Intel, TSMC, and Samsung. In a notable shift, all three companies have announced or broken ground on new fabs in the United States. In addition, Intel also has construction projects in Europe, TSMC has plans for new plants in Taiwan and Japan, and Samsung has expansion underway in South Korea. This wave of new plants announced in response to the semiconductor shortages experienced in the COVID-19 pandemic and the massive subsidies announced in the United States and Europe may well change the industry landscape in coming years.

6 Location of Innovation

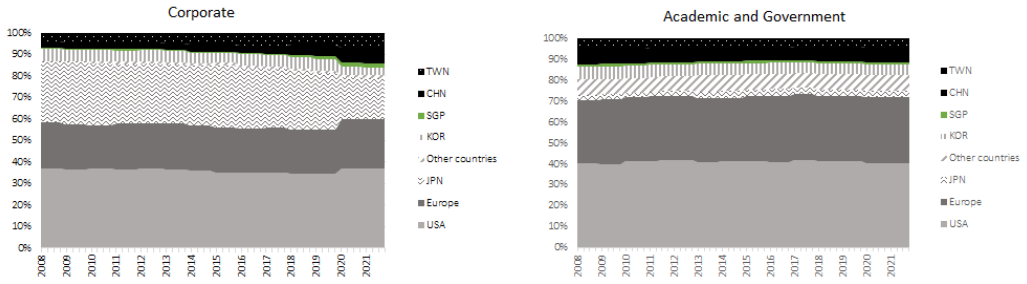
Semiconductor companies often devote fabs to research into new fabrication technologies and development of the process that allows those technologies to be employed at scale. These distribution of these corporate R&D fabs differs markedly from the distribution of production. In particular, the Asian locations accounting for roughly 60 percent of global wafer capacity

Figure 12: Location of new wafer fabrication plants.



in 2021–China, Taiwan, Singapore, and South Korea–accounted for less than 20 percent of the corporate R&D fabrication plants, though this share has increased in recent years (fig. 13). The United States, which accounted for only 12 percent of global wafer starts in 2021, accounted for more than 40 percent of corporate R&D labs, and Japan and Europe accounted for another 40 percent.

Figure 13: R&D fab locations.



Academic and government fabs are an important part of R&D for the industry as well, with a particular focus on basic research. Here the role of Asia is disproportionately small as well: China, Taiwan, Singapore, South Korea, and Japan collectively are the location for roughly one in five of these fabs. The United States is the most prominent location, and Europe, despite having relatively little production capacity, is an important location as well.

7 Comparison of the Databases to Official Statistics by Country

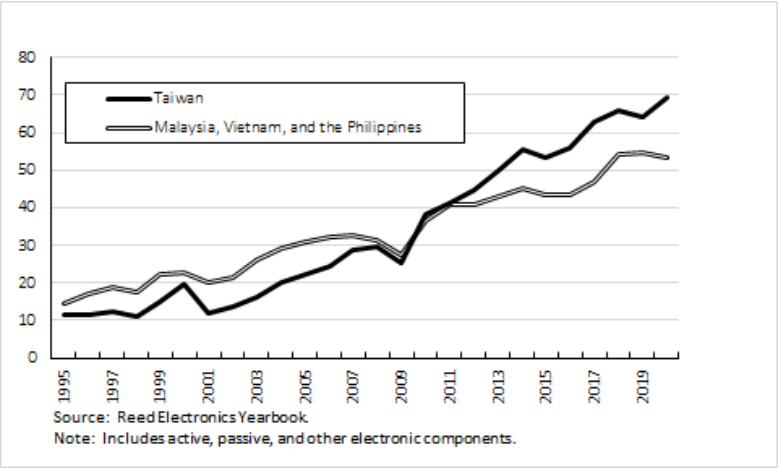
As shown in fig. 8 on page 13, the pace of growth in global real output, as measured using the Federal Reserve's global price index price index, is broadly similar to the indicator for fabrication growth–the transistor budget measure–over the period of study, though the contours of the series differs noticeably. Making this comparison at the national level is more challenging. National economies and their official statistics differ widely in two respects.

First, the mix of activity varies widely across countries. Some national semiconductor industries, such as Taiwan's, specialize in fabrication, while others, such as Malaysia, Vietnam, and the Phillipines, specialize in test and assembly. Yet, national accounts do not distinguish between the two activities, recording both as semiconductor production. Official nominal semiconductor output for Taiwan and for the three other countries combined is of a similar magnitude (fig. 14 on the next page). In contrast, our data indicates that Malaysia, Vietnam, and the Philippines processed only a few hundred thousand wafers per month in 2021, in contrast to several million per month for Taiwan. Thus, we are able to distinguish between the fabrication activity that takes place in Taiwan, and the test and assembly activity in those three countries. National statistics record both wafers fabricated but not tested and assembled (Taiwan), and the gross output of tested and assembled chips (Malaysia and others) as the same product, undermining their usefulness for analysis of supply chains. And, the creation of chip designs, a specialty of the U.S. electronics sector, which are then fabricated by contractors offshore, especially in Taiwan, leaves no discernible mark on national chip production statistics at all (Bayard, Byrne and Smith, 2015). In short, production and trade statistics for these three distinct industries–design, fabrication, and test and assembly–are needed to make sense of the electronics sector.

In our measures, the difference in product mix is evident in the average geometry statistics shown in fig. 10 on page 15. The center of mass of the Korean industry is at a far higher level of miniaturization. Our transistor measures capture the greater complexity of the leading edge memory chips in Korea. However, they also struggle with the distinction between bipolar and MOS technologies. We leave the distinction between transistors from the two to future work.

Second, semiconductor production entails noticeably different mix of prod-

Figure 14: Nominal Semiconductor Production, Selected Countries.



ucts across countries. For example, the lion's share of output in the South Korean industry is memory chips and the Japanese industry specializes in analog, power, and optoelectronic chips. In light of the fact that efforts to build real measures of memory output–which boils down to getting quality-adjusted price indexes right–have been far more successful, comparisons between Korea and Japan, for example, are distorted by measurement issues. That is, the degree to which the difference in the pace of semiconductor industry price declines between these two countries is the result of more successful measurement efforts in South Korea is unknown, but likely non-trivial. Likewise, the difference between price trends for the United States, where MPUs (microprocessors) play a major role, and Taiwan, where custom processors are far more important, is likely due in part to better measurement of MPU price indexes.

To mitigate differences in measurement technique and coverage across countries, we switch to a single set of price indexes, provided by the Federal Reserve. Aggregate prices for four broad categories are employed, combined for each country with the coarse weights shown in fig. 15 on the next page: MPUs, memory chips, discretes–including optoelectronics, and other ICs–including ASICs and analog chips. The weights by country are our best assessment of the specialization of national industries after consulting our data and the product level information available for some countries. We deflate nominal output by country using the country price indexes and compare these real output measures to our indicator for fabrication–industry transis-

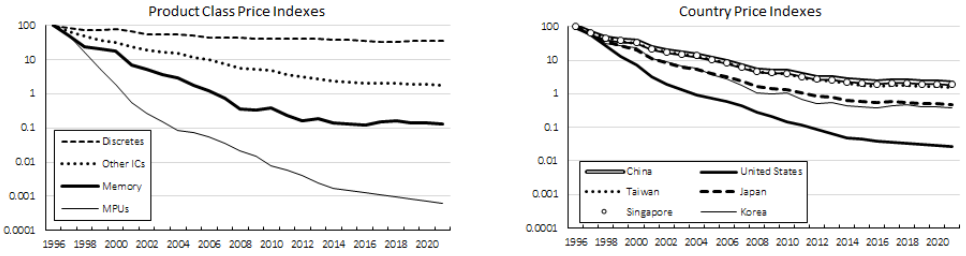
tor budget. As shown in fig. 16, price trends differ markedly across these product classes. Thus, the difference in product specialization across countries leads to noticeably different price indexes for the six national industries shown.

Figure 15: Mix of chip type output by country.

	Memory	Microprocessors	Optoelectronics & Discretes	Other Chips
China	25.2%	0.0%	27.9%	46.9%
Japan	29.2%	17.7%	30.0%	23.1%
Korea	66.3%	0.0%	6.7%	26.9%
Singapore	24.8%	0.0%	23.3%	51.9%
Taiwan	12.6%	0.0%	6.3%	81.1%
United States	3.9%	54.4%	5.9%	35.7%

Source: Authors' estimates based on data from Federal Reserve Board, Gartner, Inc., SEMI, Inc., and Semiconductor Industry Association.

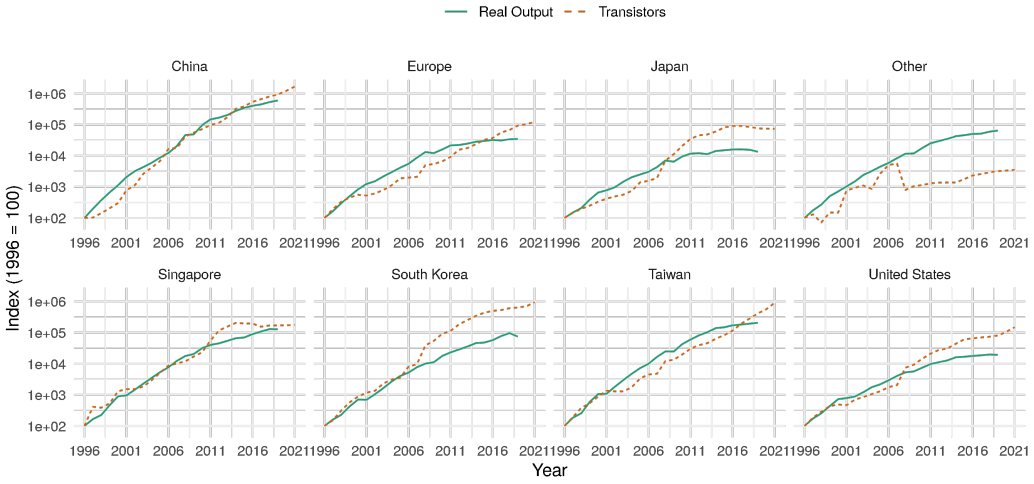
Figure 16: Official and alternative chip price indexes.



Using these mutually consistent country-specific price indexes to deflate official nominal output we arrive at the real output measures shown in fig. 17 on the next page. Comparing the path for real output to our country-specific transistor indicators, we find remarkably similar paths, especially for China and Tawian. Both measures are best considered to be indicators of underlying true real output. To the extent that price indexes fail to capture recent innovation, the real output measures for the United States, South Korea, and Japan may be understated. Conversely, our transistor measure may be a less reliable indicator for real output in recent years. Importantly, our measure

is best viewed as an indicator of the real gross output of the semiconductor fabrication industry, whereas the real output measure is meant to capture the value of completed microchips, including the value of design. So, in effect, both measures could be accurate; if so, the contribution to growth from the value of design would appear to have stepped down appreciably in recent years.

Figure 17: Transistor Indicator vs. Real Output.



Source: Authors' calculations based on data from Reed Electronics Research, Federal Reserve Board, Gartner, Inc., and SEMI, Inc.

8 Conclusion

The historical pace of technical advance in the semiconductor industry predicted by Gordon Moore has been sustained for fifty years by the metric he proposed: the number of components one can fit on an integrated circuit. Whether this measure remains a sufficient statistic for innovation in the industry is an open question–ICs come in a greater variety of forms and a produced using a wider array of technologies than was the case in 1971–but the ability to miniaturize devices continues to advance at a rapid pace. We offer a simple, well-defined measure to use to assess this question built from data on individual plant locations scrutinized by industry consultants and trade groups, which provides a useful counterpoint to measures that rely on judgemental appraisals by experts that are difficult to reproduce.

The industry's progress is to the credit of a wide array of contributors. Whereas corporate, government, and academic R&D continues to be located primarily in the United States, Europe, and Japan, as of 2021, the practical implementation of leading edge technology in volume production is dominated by Taiwan, with South Korea not far behind; the U.S. industry has fallen behind. Production in the United States as well as in Japan takes place using a broad array of technologies, with the average technology in use incorporating appreciably less miniaturization than is the case for South Korea and Taiwan. And, in recent years, construction of new plants in the global industry has shifted toward China, though recent announcements point to a substantial rise in the location of new advanced manufacturing locations in the United States.

The dataset used to establish these stylized facts about the industry, containing a long consistent history of plant-level data, will support future work on these production and R&D location decisions and the role played by such factors as the location of corporate headquarters and such policies as government subsidies and trade restrictions.

References

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.

Bauer, Luc Olivier and E. Marshall Wilder, *The Microchip Revolution*, Independently published, 2020.

Bayard, Kimberly, David Byrne, and Dominic Smith, "The Scope of US Factoryless Manufacturing," 2015.

Moore, Gordon E, "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.

Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, Austin, Texas: International SEMATECH, 2007.

WikiChip contributors, "Process Technology History - Intel," 2022. [Online; accessed 28-February-2022].