Volume Title: Measuring and Accounting for Innovation in the Twenty-First Century

Volume Authors/Editors: Carol Corrado, Jonathan Haskel, Javier Miranda, and Daniel Sichel, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-72817-9 (cloth); 978-0-226-72820-9 (electronic)

Volume URL: https://www.nber.org/books-and-chapters/measuring-and-accounting-innovation-twenty-first-century

Conference Date: March 10–11, 2017

Publication Date: April 2021

Chapter Title: Measuring Moore's Law: Evidence from Price, Cost, and Quality Indexes

Chapter Author(s): Kenneth Flamm

Chapter URL: https://www.nber.org/books-and-chapters/measuring-and-accounting-innovation-twenty-first-century/measuring-moores-law-evidence-price-cost-and-quality-indexes

Chapter pages in book: p. 403 – 470

11

Measuring Moore's Law
Evidence from Price, Cost, and Quality Indexes

Kenneth Flamm

"Moore's law" in the semiconductor manufacturing industry is used to describe the predictable historical evolution of a single manufacturing technology platform ("silicon CMOS") that has been continuously reducing the costs of fabricating electronic circuits since the mid-1960s.[1] Some features of its future evolution were first correctly predicted by Gordon E. Moore (then at Fairchild Semiconductor) in 1965, and Moore's law became an industry synonym for continuous, periodic reduction in both size and cost for electronic circuit elements.

Technological innovation for this manufacturing platform was coordinated and synchronized across a variety of different engineering fields,

Kenneth Flamm is professor and Dean Rusk Chair in the LBJ School of Public Affairs at the University of Texas at Austin.

I am most grateful to Anjum Khurshid, Kevin Williams, Caroline Alexander, Pablo Cruzat, Javier Beverinotti, Manuel Chavez, Changgui Dong, and Miha Vindis for their excellent research assistance over the years this data was collected and maintained, and to financial support from the Kauffman Foundation and the National Science Foundation. This research is based in part upon work supported by the National Science Foundation under Grant No. 0830389. I would also like to thank Ana Aizcorbe, David Byrne, Carol Corrado, Stephen Oliner, James Prieger, Marshall Reinsdorf, Steve Sawyer, Dan Sichel, Neil Thompson, participants in the Conference on Research in Income and Wealth (CRIW) conference Measuring and Accounting for Innovation in the 21st Century and the International Monetary Fund (IMF) Fifth Statistical Forum, "Measuring the Digital Economy," and two anonymous referees, for their many useful comments on earlier versions of this chapter. Supplemental appendix tables referred to in the text are available online at https://www.nber.org/data-appendix/c13897/appendix.pdf. For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see https://www.nber.org/books-and-chapters/measuring-and-accounting-innovation-21st-century/measuring-moores-law-evidence-price-cost-and-quality-indexes.

1. Complementary metal oxide semiconductor (CMOS) is the most widely used "flavor" of semiconductor technology used to manufacture an integrated circuit (IC).

including materials, optical systems, ultraclean precision manufacturing, factory automation, electronic circuit design and simulation, and improved computer software for computational modeling in all of these fields. It was a self-reinforcing dynamical process, since the largest market for the semiconductor manufacturing industry's products has always been the computer industry.[2] Cheaper computing hardware meant cheaper modeling and engineering to further reduce the costs of the semiconductors manufactured for use in future computers. New public-private institutions and organizations were developed to coordinate the simultaneous arrival of the very heterogeneous technological building blocks required for this increasingly complex semiconductor manufacturing technology platform.

The result was an industrial dynamic that, since the mid-1960s, had effectively worked as a "virtual shrinking machine" for electronic circuits. On a regular basis, new "technology nodes" delivered 30 percent reductions in the size of the smallest dimension ("critical feature size," F) that could be reliably manufactured on a silicon wafer. This implied a 50 percent reduction in the area occupied by the smallest manufacturable electronic circuit feature (F^2) and a doubling in density—the number of circuit elements (e.g., transistors) per area of silicon in a chip.

Section 11.1 of this chapter develops some stylized economic facts, reviewing why this progression in manufacturing technology delivered a 20 to 30 percent annual decline in the cost of manufacturing a transistor, on average, as long as it continued. It constructs a simple economic framework that explains how improvements in manufacturing technology, which resulted in feature size reductions, created manufacturing cost reductions for all types of electronic circuits.

Section 11.2 reviews other economically significant benefits (in addition to increased density and lower cost per circuit element) that would be associated with smaller feature sizes. Some of those characteristics would be expected to have significant economic value, and historical trends for these characteristics are reviewed. Chip speed, in particular, would have major impacts on computer performance. Econometric analysis of software benchmark data provided in this section of the chapter shows that rates of performance improvement in microprocessors fell off dramatically in the new millennium, a retreat from very high rates of increase measured in the late 1990s. Lower manufacturing costs alone pose no special challenges for price and innovation measurement, but these other benefits do, and they motivate quality adjustment methods when semiconductor product prices are measured.

Section 11.3 analyzes empirical evidence of recent changes to the his-

2. This defines the computer industry expansively, to include the computer systems embedded in the smart electronic systems and mobile devices whose sales have grown most rapidly in recent decades.

torical Moore's law trajectory and finds corroborating evidence for a slowdown of Moore's law in prices for the highest-volume products: memory chips, custom chip designs outsourced to dedicated contract manufacturers (foundries), and Intel microprocessors. In this section, in addition to reviewing price indexes available in the public literature, I construct a new, high-frequency hedonic price index for Intel desktop microprocessors utilizing very detailed chip characteristics. I use a variety of data sources, including both Intel list prices and retail processor transaction prices. My results are consistent with the other public data I review and support the notion of a marked slowdown in Moore's law–driven price declines over the last decade.

Section 11.4 reviews evidence to the contrary, which relates primarily to Intel microprocessors. It analyzes Intel's own publicly released information on the topic, discusses economic reasons why Intel microprocessor prices might behave differently from prices for other types of semiconductor chips, and reviews other published studies, one of which came to the opposite conclusion: that quality-adjusted price decline for Intel processors continued at unchanged high rates in recent years. After investigating a variety of forms of evidence in detail, I conclude that the finding of an unchanged rate of price decline for Intel microprocessors is most likely an artifact of omitted variables in the estimated econometric model.

Section 11.5 dives into Intel microprocessors in even greater depth and tests the computer architecture textbook view of how a small set of specific chip characteristics affects performance of microprocessors in executing programs. I outline a simple structural model of microprocessor computing performance and then estimate that model empirically. Simple econometric models, using only a small set of explanatory chip characteristics, explain 99 percent of variance across processor models in performance on different, commonly used CPU performance benchmarks. However, the impact of different chip characteristics on performance varies quite dramatically across benchmarks.

The economic implication is that these characteristics, which determine benchmark scores, should clearly be included in any hedonic price equation. Most of these chip characteristics would also be expected to affect chip production cost and therefore have an independent rationale for inclusion in a hedonic price equation. It may seem reasonable to assume that a scalar, fixed-weight average of different benchmark scores for a chip perfectly captures the impact of changing chip characteristics on computer performance and therefore on user demand (though this is a very strong assumption given substantial heterogeneity and change over time in the mix of computer applications relevant to different computer market segments). But even if it were true that some fixed weighted average of benchmark scores was a perfect measure of changes in chip performance relevant to demand shifts, inclusion of this variable would not eliminate the need to also include cost-shifting product characteristics as additional controls in a hedonic model

of market equilibrium chip prices. This argument is actually illustrated by a simulation created to depict the impact of perfect collinearity among chip characteristics on hedonic price coefficients in section 11.3.

A sixth and final section of the chapter points to some economically important conclusions that can be drawn from this evidence. Available empirical evidence, on balance, suggests that Moore's law–related historical declines in chip manufacturing cost have clearly been greatly attenuated over the last decade, resulting in much more slowly declining quality-adjusted chip prices. If we accept earlier economic research showing a strong link between technological innovation in semiconductors and IT and productivity growth across the broader economy, then a slowdown in semiconductor manufacturing innovation, inducing slower quality-adjusted price declines for both chips and IT utilizing those chips, will affect measures of productivity growth in industrialized economies. Finally, the winding down of Moore's law means that much of the continuing hardware cost decline driving ever more intensive use of IT across the economy over the last 50 years will no longer hold and that computing costs—including energy use per computation, the principal variable cost—will decline much more slowly in the future than was true in the past. Improvement in software, rather than dramatically cheapening hardware, may well emerge as the main focus for IT innovation over the next 50 years.

11.1 Stylized Facts about Semiconductor Manufacturing Innovation

In 1965, five years after the integrated circuit's invention, Gordon E. Moore (who would shortly move on to cofound Intel) predicted that the number of transistors (circuit elements) on a single chip would double every year.[3] Later modifications of that early prediction—Moore's law—became shorthand for semiconductor manufacturing innovation.

Moore's prediction requires other assumptions in order to create economically meaningful connections to the information age's key economic variable: the cost (or price) of electronic functionality on a chip (embodied in the 20th century's supreme electronic invention, the transistor).[4] Chip fabrication requires coordinating multiple technologies, combined in very complex manufacturing processes.

The pacing technology has been the photolithographic process used to pattern chips. From the 1970s through the mid-1990s, a new "technology node"—a new generation of photolithographic and related equipment and materials required for successful use—was introduced roughly every three years or so. Starting in the mid-1970s, this three-year cycle coincided with the time interval between introductions of next-generation DRAM computer

3. G. Moore (1965).
4. Jorgenson (2001); Flamm (2003, 2004); Aizcorbe, Flamm, and Khurshid (2007).

memory chips, storing four times the bits in the previous-generation chip.[5] This observed 18-month "doubling period" became a new de facto "revised" Moore's law.[6]

The close early fit of DRAM product development cycles with leading-edge chip manufacturing technology introductions was no coincidence. DRAMs at that time were the highest-volume standardized commodity chip product manufactured, and a rapidly expanding computer market drove leading-edge chip manufacturing technology development. Moore's prediction morphed into an informal—and later, formal—technology coordination mechanism (the International Technology Roadmap for Semiconductors, or ITRS) for the entire global semiconductor industry—equipment and material producers, chip makers, and their customers.

Relationships between Moore's law and fabrication cost[7] trends for integrated circuits can be described by the following identity, giving cost per circuit element (e.g., transistor):

$$(1)\qquad \frac{\$}{\text{element}} = \frac{\dfrac{\$\text{processing cost}}{\text{area "yielded" good silicon chips}} \times \dfrac{\text{silicon wafer area}}{\text{chip}}}{\dfrac{\text{elements}}{\text{chip}}}.$$

Moore's original "law" described only the denominator—a prediction that elements per chip would quadruple every two years. Back in 1965, Moore hadn't originally anticipated rapid future advances in technology nodes. Acknowledging that an integrated circuit (IC) containing 65,000 elements was implied by 1975, Moore wrote, "I believe that such a large circuit can be built on a single wafer. With the dimensional tolerances already being employed . . . 65,000 components need occupy only about one-fourth a square inch."[8]

Rewriting this more concisely without relying on Moore's prediction about numbers of elements per chip (therefore eliminating the need for assumptions about chip size) yields

$$(2)\qquad \frac{\$}{\text{element}} = \frac{\$\text{processing cost}}{\text{area "yielded" silicon}} \times \frac{\text{silicon area}}{\text{element}},$$

5. DRAM (dynamic random access memory) was invented in 1968 by Robert Dennard at IBM and first commercialized by Moore's newly founded company, Intel, in 1970. DRAM chips are the most common type of IC used for "main" memory storage in modern computer systems and, until the early 21st century, were the type of IC semiconductor chip produced in the highest production volume. DRAMs are a type of "volatile" memory chip—information stored on the chip in binary (0,1) form disappears when electrical power is turned off.

6. A decade later, Moore himself revised his prediction to a doubling every two years. G. Moore (1975), 11–13.

7. Analysis of fabrication costs, which account for most of chip costs, ignores assembly, packaging, and testing.

8. G. Moore (1965). The largest wafer sizes in use then were comparable in diameter to a modern minipizza appetizer.

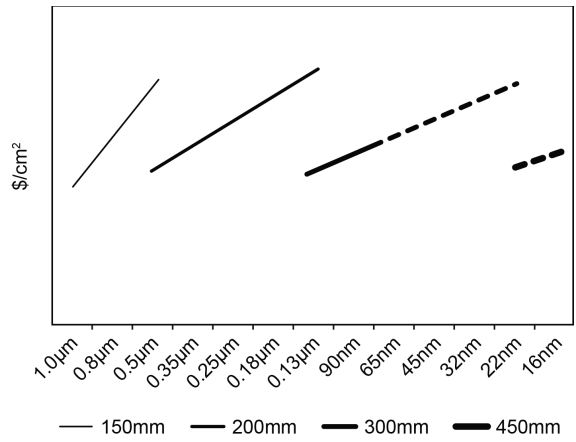

Fig. 11.1 Wafer size conversions offset Intel's increased wafer-processing cost
Source: Holt (2005).

which depends directly on the defining characteristic of a new technology node, the smallest patternable feature size, as reflected in chip area per transistor. This "Moore's law" variant came into use in the semiconductor industry as a way of analyzing the economic impact of new technology nodes. New technology nodes increased the density of transistors fabricated in a given area of silicon in a readily predictable way. Time between new nodes—and a new node's impact on wafer-processing costs—jointly determined decline rates in transistor fabrication cost.

Through 1995, new technology nodes were introduced at roughly three-year intervals. Each new node reduced the smallest planar dimension ("critical feature size," F) in circuit elements by 30 percent, implying 50 percent smaller silicon areas (F^2) per circuit element.

Completing the economic story, the cost per silicon wafer area processed, averaged over long periods, increased only slowly.[9] At new technology nodes, processing cost per silicon wafer area indeed increased. But episodically, larger wafer sizes were introduced, sharply reducing processing costs per area. The net effect was nearly constant long-run costs with only slight increases. Figure 11.1, presented in 2005 by Intel's chief manufacturing technologist, shows new wafer sizes "resetting" wafer-processing costs. Significantly, larger diameter wafer sizes (450mm) were expected at the 22 nanometer (nm) node. However, 450mm wafers were not introduced as Intel adopted 22nm technology in 2012 and had not been introduced by 2020, and even future introduction now seems highly uncertain. The most recent wafer

9. Over 1983–98, wafer-processing cost/cm^2 silicon increased 5.5 percent annually. Cunningham et al. (2000), 5. This estimate relates to total silicon area processed (including defective chips). Since defect-free chips' share of total processed area increased historically (chip fabrication yields increased), wafer-processing cost per good silicon area rose even more slowly, approximating constancy.

size "reset," adoption of 300mm diameter wafers, occurred at the 130nm technology node, around 2002.

Using these stylized trends—wafer-processing cost per area of silicon roughly constant and silicon area per circuit element halved, with new technology nodes introduced every three years—equation (2) above predicts that every three years, the cost of producing a transistor would fall by 50 percent, a 21 percent compound annual decline rate.

In reality, leading-edge computer chips—like DRAM memory (the primary product originally produced at Intel after Moore and others founded that company, which immediately became the largest-volume product in the semiconductor industry and the primary product driving Intel's initial growth)—dropped in price substantially faster than 20 percent pre-1985. The steeper decline rate in part reflected further increases in density due to circuit design improvements (e.g., reduction in memory cell footprint),[10] 3D interconnect layers enabling tighter packing of circuit elements,[11] and gradual introduction of 3D into physical designs of transistors and other circuit elements.[12] In addition, operating characteristics of a given circuit design—in particular, switching speed and power requirements—improved with new manufacturing technology and made additional contributions to quality-adjusted price. Finally, smaller and cheaper transistors made it economical to add ever greater electronic functionality to chips, and more and more of a complete electronic system was progressively integrated onto a single chip, which greatly improved system reliability.[13]

In the mid-1990s, the semiconductor manufacturing industry arrived at a significant technological inflection point.[14] New technology nodes began

10. Flamm (2010), figure 2, documents a 62 percent decline in minimum memory bit cell footprint between 1995 and 2004.

11. Anticipated by Moore back in 1965: "no space wasted for interconnection . . . using multilayer metallization patterns separated by dielectric films." G. Moore (1965).

12. Recent examples of 3D transistor structures include RCAT (recessed cell array transistor) and FinFET (fin field effect transistor) structures; 3D capacitor designs have been used in DRAM since the late 1990s.

13. Electrical interconnections between components have historically been the most frequent point of failure in electronic systems.

14. Industry road maps originally dated this transition to two-year node rollouts to 1995; post-2004 road maps revised that date to 1998. Aizcorbe, Oliner, and Sichel (2008) have persuasively argued that the turning point was closer to the mid-1990s than late in the decade.

The mid-1990s were also a technological inflection point for Intel's manufacturing capabilities. Intel had exited the DRAM business in 1985, which previously had been driving its leading-edge manufacturing technology development, and refocused its R&D on logic circuit design (Burgelman 1994, 32–46). As a consequence, by the late 1980s, Intel manufacturing capability was trailing well behind the leading edge of the manufacturing technology it had once pioneered.

In order to catch up, Intel began adopting new nodes every two years, even as the rest of the industry continued at the historical three-year pace. Comparing launch dates for Intel processors at new technology nodes with initial use of those nodes by DRAM makers, Intel was two years behind in 1989 (at 1000nm); three years behind in 1991 (800nm); and one year behind in 1995 (350nm). Intel caught up with the DRAM makers in 1997, at 250nm, and remained on a roughly two-year cycle through 2014. Author's calculations based on Intel (2008), IC Knowledge (2004), and http://ark.intel.com.

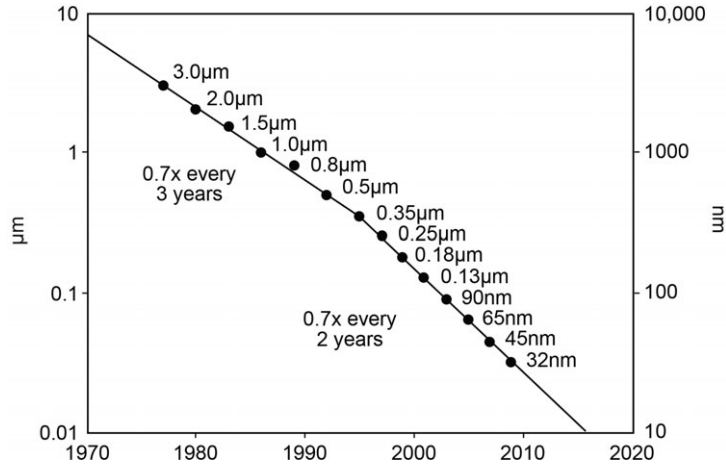

Fig. 11.2 Feature size scaling as observed by Intel in 2005
Source: Holt (2005).

arriving at two-year intervals, replacing three-year cycles. (Intel's perception of this trend, as of 2005, is documented in figure 11.2.) The origins of this change lie in the early 1990s, when the US SEMATECH R&D consortium sponsored a road map coordination mechanism in pursuit of an acceleration in the introduction of new manufacturing technology, intended to benefit the competitiveness of US chip producers. By the mid-1990s, with the increasing reliance of semiconductor manufacturing on a global industrial supply chain, the American national road map evolved into the International Technology Roadmap for Semiconductors (ITRS).[15] Explicitly coordinating the simultaneous development of the many complex technologies required to enable a new manufacturing technology node every two years apparently succeeded in raising the tempo of semiconductor manufacturing innovation for over a decade.[16]

Using equation (2) but adopting shorter two-year cycles for new technology nodes implies rates of annual decline in transistor costs accelerating to almost 30 percent. In short, if the historic pattern of 2- to 3-year technology node introductions, combined with a long-run trend of wafer-processing costs increasing very slowly, were to have continued indefinitely, a minimum floor of perhaps a 20 to 30 percent annual decline in quality-adjusted costs for manufacturing electronic circuits would be predicted due solely to these "Moore's law" fabrication cost reductions. On average, over long periods, the denser, "shrunken" version of the same chip design fabricated a year

15. Flamm (2009); Spencer and Seidel (2004).
16. The last (incomplete) official road map prepared by ITRS was released in 2012. Intel and others reportedly withdrew from ITRS around this time.

earlier would be expected to cost 20 to 30 percent less to manufacture purely because of the improved manufacturing technology.

It now appears that this two-year cycle for technology nodes definitively ended in 2014 with deployment of the 14nm node. The most historically prominent adopter of leading-edge chip manufacturing technology, Intel, currently projects a delayed introduction of its next 10nm processor products to no earlier than late 2019.[17] This means that the time between introductions of new technology nodes now is approaching *five* years for Intel, a dramatic change from its two-year cadence through 2014.[18]

At Intel, the post-1995 two-year technology development cycle had been explicitly incorporated into marketing efforts and was dubbed the Intel "tick-tock" development model in 2007.[19] Every two years, there would be a new technology node introduced ("tick"), with the existing microprocessor computer architecture ported to the new node (effectively, "die shrinks" using the new process), followed by an improved architecture fabricated with the same technology the following year ("tock"). The death of the "tick-tock" model was officially acknowledged by Intel in its 2016 annual report.[20]

Intel publicly disclosed a version of equation (2) to its shareholders in 2015, purged of sensitive cost numbers by indexing all variables to equal one at the 130nm technology node—the technology node at which the transition to a larger wafer size occurred.[21] The 2015 Intel decomposition of manufacturing cost per transistor, using equation (2), is shown as figure 11.3 and in table 11.1. Generally, Intel's average silicon area per transistor did not decline by the predicted 50 percent between technology nodes, primarily because of the increasing complexity of interconnections in processor designs.[22] If accurate, these numbers indicate that the average chip area per transistor shrank by 38 percent at each new node from 130nm through 22nm.[23] Nor did Intel's wafer-processing costs stay constant over the post-130nm period as a whole, since the adoption of 450mm wafers, and the subsequent cost reset, never happened at 22nm as had been predicted back in 2005. However, as long as the average area per transistor declined at

17. See Moammer (2017).

18. Intel chip manufacturing competitor TMC was said in early 2017 to be manufacturing a "10nm" node in volume for Apple (see Merritt 2017), but it is widely believed in the industry that its current technology is physically equivalent to a half-node advancement over the previous-generation Intel technology node. See https://www.semiwiki.com/forum/f293/intel-tsmc-samsung-10nm-update-8565.html; Pirzada (2016); Rogoway (2018); Cutress and Shilov (2018).

19. See Intel (2017).

20. Intel (2016), 14.

21. Intel actually produced microprocessors in volume on both 200mm (8″) and 300mm (12″) wafers using its 130nm manufacturing process technology. See Natrajan et al. (2002), 16–17.

22. See Flamm (2017), 34, for a more detailed explanation.

23. Absolute constancy in reported decline rates for average area per transistor over five generations of new Intel manufacturing technology is puzzling, suggesting long-run trend-based estimates rather than actual averages computed from empirical manufacturing data.

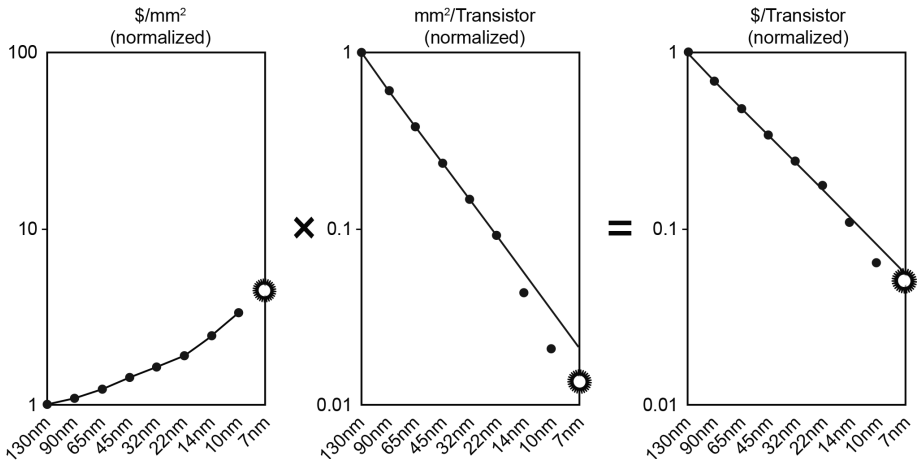

Fig. 11.3 Intel 2015 version of equation (2)
Source: Holt (2015).

Table 11.1 Decomposing Intel transistor cost declines into wafer cost and transistor size changes

Year Intel 1st shipped product at new tech node	Tech node (nm)	Wafer processing cost ($/mm²)	×	Silicon area (mm² /transistor)	=	$ cost/ transistor	Compound annual percentage change: Wafer processing cost ($/mm²)	Silicon area (mm²/ transistor)	$ Cost/ transistor
2002	130	1		1		1			
2004	90	1.09		0.62		0.68	5%	−21%	−18%
2006	65	1.24		0.38		0.47	7%	−21%	−16%
2008	45	1.43		0.24		0.34	7%	−21%	−15%
2010	32	1.64		0.15		0.24	7%	−21%	−16%
2012	22	1.93		0.09		0.18	8%	−21%	−14%
2014	14	2.49		0.04		0.11	14%	−31%	−22%

Source: Holt (2015), slide 6, graph digitized using WebPlotDigitizer. Year node introduced from ark.intel.com.

faster rates than processing costs per area increased, transistor costs would continue to decline. Intel's cost-per-transistor estimates are revisited below.

How would reductions in production cost translate into price declines? One very simple way to think about it would be in terms of a "pass-through rate," defined as dP/dC (incremental change in price per incremental change in production cost). The pass-through rate for an industry-wide decline in marginal cost is equal to 1 in a perfectly competitive industry with constant returns to scale but can exceed or fall short of 1 in imperfectly competitive industries. Assuming the perfectly competitive case as a benchmark for long-run pass-through in "relatively competitive" semiconductor product

markets, this would then imply an expectation of 20 percent to 30 percent annual declines in price due solely to Moore's law.

Historically, most semiconductor chip production ultimately seems to have migrated to more advanced technology nodes.[24] Other kinds of innovations in semiconductor manufacturing, or innovations in the design and functionality going into electronic circuits, might be expected to stimulate even greater rates of quality-adjusted price declines. Thus the 20 percent to 30 percent annual decline in manufacturing cost associated with Moore's law could be interpreted as a floor on the quality-adjusted price declines in the most competitive segments of the semiconductor market.

11.2 Other Benefits from "Moore's Law" Manufacturing Innovation

Impressive declines in transistor manufacturing cost accompanying denser chips with smaller feature sizes at more advanced technology nodes measure only a part of the economic benefits of the Moore's law innovation dynamic. With smaller transistor sizes also came faster switching times and lower power requirements.[25] The complementary benefits of speed and power improvements were highly significant for chip consumers (like computer makers) and their customers.

This was particularly true for chip makers manufacturing microprocessors. Existing computer architectures running at faster speeds run existing software faster and enable more data processing in any given time. Until 2004, computer processor clock rates increased rapidly, as did performance of computers incorporating these faster microprocessors. Figure 11.4 shows clock rates for Intel desktop microprocessors in computers tested on industry standard benchmark programs over the last 20 years as well as benchmark scores for these computers. As clock rates increased, so did performance.[26] Cheaper processors were also faster, stimulating increased demand for new computers in offices, homes, and workplaces.

The logarithmic scale used in figure 11.4 obscures a fairly dramatic slowdown in improvement in CPU performance after the millennium. Table 11.2 shows compound annual growth rates in performance over time of Intel desk-

24. At SEMATECH, the US semiconductor industry consortium (with which the author worked as a consultant in the first decade of the 2000s), the planning rule of thumb was that a fab would be a candidate for an upgrade to a new technology node no more than twice over its lifetime and then would be shut down as uneconomic.

25. The underlying theory ("Dennard scaling") suggested that a 30 percent reduction in transistor length and a 50 percent reduction in transistor area would be accompanied by a 30 percent reduction in delay (40 percent increase in clock frequency) and a 50 percent reduction in power. Esmaeilzadeh et al. (2013), 95.

26. For given software and computer architecture, time required for programs to execute is inversely proportional to processor clock rate, assuming data transfer does not constrain performance. Lower rates of performance improvement after 2004, as processor clock rates plateaued, were obvious to computer designers. See Fuller and Millett (2011), chap. 2; Hennessey and Patterson (2012), chap. 1.

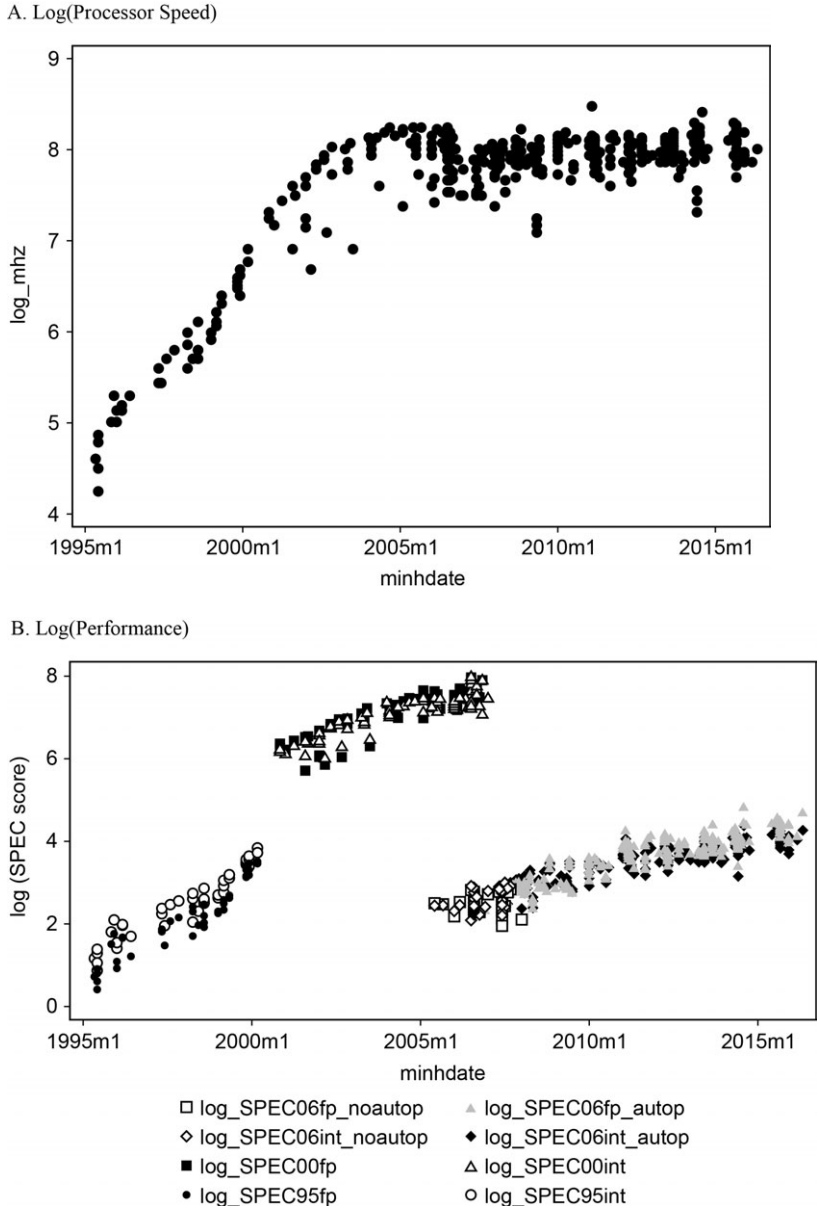

Fig. 11.4 Processor clock rate and performance for Intel desktop processors running SPEC CPU benchmarks, by first availability date of tested hardware

Source: Author's analysis of SPEC submissions, SPEC.org. Performance scores for 1995, 2000, and 2006 SPEC benchmarks have different values for same processor, and different vintage benchmark scores are not directly comparable. "minhdate" is date on which first SPEC benchmark for computer system with that processor is run. "log_SPECyyxx" is log of median SPEC year yy benchmark xx score, by processor model. SPEC06xx results include separate scores with compiler autoparallelization turned on (autop) and off (noautop) for same model, when reported.

Table 11.2 **Annual growth in processor performance improvement over different time periods and benchmarks**

SPEC CPU benchmark	Coeff. CAGR	Robust SE
1995m5–2000m3		
int95	.583	.018
fp95	.640	.023
int95_rate	.624	.027
fp95_rate	.723	.033
2000m11–2004m11		
int2000	.330	.017
fp2000	.343	.024
int2000_rate	.470	.051
fp2000_rate	.399	.035
2005m2–2007m1		
int2000	.322	.016
fp2000	.337	.022
int2000_rate	.465	.048
fp2000_rate	.399	.033
2005m6–2012m11		
int2006	.171	.007
fp2006	.247	.008
int2006_rate	.247	.013
fp2006_rate	.254	.010
2013m1–2016m5		
int2006	.169	.006
fp2006	.241	.007
int2006_rate	.242	.012
fp2006_rate	.248	.009

Source: Author analysis of SPEC benchmark performance of Intel desktop processors.

top processors on standard CPU benchmark software (the Standard Performance Evaluation Corporation [SPEC] benchmarks; see appendix 11.A1).

Three different versions of the SPEC CPU test suite were released—one around 1995, one in 2000, and the most recent in 2006. Each suite contains a selection of "integer" application tests (e.g., programming and code processing, artificial intelligence, discrete-event simulation and optimization, gene sequence search, video compression) and a set of "floating point" math-intensive application tests (e.g., solution of systems modeling problems in physics, fluid dynamics, chemistry, and biology; finite element analysis; linear programming; ray tracing, weather prediction; speech recognition). These test suites are designed to test single-process (programming task) performance on a CPU.[27]

27. The overall benchmark score is calculated as a geometric mean of scores on the individual programs within the benchmark.

In addition, so-called rate versions of these test suites, which run multiple versions of the single-process benchmarks simultaneously on a single CPU, are available. The "rate" benchmarks are intended to show how the CPU would perform as a server running multiple independent jobs or, alternatively, running an "embarrassingly parallel" programming problem—a task that could be divided up into multiple software processes not requiring any communication or coordination between processes.[28]

Changes in trends over time in the SPEC benchmark performance scores for Intel desktop processors are quite dramatic.[29] Over the 1995–2000 period, integer computing performance increased by about 58 percent annually and floating point performance by 64 percent. The suite was revised in 2000, and from the end of 2000 through 2004, both integer and floating-point performance improvement rates were almost halved, to an increase of about 33 percent to 34 percent per year.[30] Finally, over the most recent time period, after the 2006 revision of the SPEC benchmarks, from 2005 through 2016, annual performance gains were reduced substantially again, to rates of 17 percent (integer) and 25 percent (floating point) annual improvement.[31]

11.3 An End to Moore's Law?

Unfortunately, the golden age of more rapidly cheapening transistors (which were also faster and drew less power) that began in the late 1990s did not survive unchallenged past the new millennium.

2004: The End of Faster. The first casualty was the "faster thrown in for free," along with smaller, cheaper, and greener. Around 2003–4, higher clock rates stalled (see figure 11.4), as disproportionately greater power was required to run processors reliably at ever higher frequencies. With tinier transistors drawing higher power in denser chips, dissipating heat generated by higher power density became impossible without expensive cooling systems. (The highest processor speed shipped by Intel until very recently was 4 GHz; IBM's fastest z-series mainframe CPU, with advanced cooling, hit 5.5 GHz in 2012, but subsequent CPUs ran at lower frequencies.[32]) Intel and others abandoned architectures reliant on frequency scaling to achieve

28. Unfortunately, there is no SPEC rule about how many instances of the single benchmark programs should be run for the rate benchmarks on a multicore CPU. It could be as many as the number of cores in the CPU or twice that number (the number of threads that can be run simultaneously on a CPU with additional processor hardware supporting symmetric multithreading—a feature called hyperthreading by Intel) or some number of instances less than either of those bounds.

29. Pillai analyzed the apparent slowdown in microprocessor quality improvement (as measured by software benchmarks) from 2001 to 2008. See Pillai (2013), figure 1.

30. There was a statistically significant—but substantively insignificant—additional decline of under a percent per year after 2004 through 2007.

31. There was another statistically significant, but substantively insignificant, decline by a fraction of a percent in performance improvement rates after 2012.

32. Raley (2015), 23.

better processor performance after 2004. Clock rates in subsequent processor architectures actually fell, and processing more instructions per clock tick became the focus for improved computing performance.

Two-year node introductions continued to produce smaller and cheaper transistors, though. Ever-cheaper transistors were utilized to create more CPUs—"cores"—per chip, thus processing more instructions per clock at lower clock frequencies. This new "multicore" strategy's weakness was that application software required "parallelization" to run on multiple cores simultaneously, and software applications vary greatly in the extent to which they can be easily parallelized. Further, improving software was more costly than simply adopting the cheaper hardware delivered by new technology nodes: quality-adjusted prices for software historically have fallen much more slowly than quality-adjusted prices for processors.[33]

The difficulty and cost of parallelization of software is an economic factor limiting utilization of cheap multicore CPUs on hard-to-parallelize applications.[34] In addition, a fundamental result in computer architecture (Amdahl's law) maintains that if there is any part of a computation that cannot be parallelized, then there will be diminishing returns to adding more processors to the task—and in many applications, decreasing returns are noticeable fairly quickly. One widely used computer architecture textbook summarized the challenges in utilizing multicore processors: "Given the slow progress on parallel software in the past 30-plus years, it is likely that exploiting thread-level parallelism broadly will remain challenging for years to come."[35]

2012: The End of Rapid Cost Declines? Until roughly 2012, transistor fabrication costs continued falling at rapid rates. At the 22/20nm technology node, which went into volume production around 2012 (at Intel), continuing cost declines began to look uncertain. Figure 11.5 shows contract chip maker GlobalFoundries' 2015 transistor manufacturing costs at recent technology nodes.[36]

Numerous fabless chip design companies, which outsource chip production to contract manufacturing "foundries," began to publicly complain that transistor manufacturing costs had actually *increased* at the 20/22nm node.[37]

33. Economic studies of mass-market, high-volume packaged software prices have typically found quality adjusted rates of annual price decline in the 6 percent to 20 percent range. See, e.g., Gandal (1994); Oliner and Sichel (1994); White et al. (2005); Copeland (2013); and Prudhomme and Yu (2005).

34. The opposite—software problems easily divided up across processors and run with little or no interprocessor communication or management required—is described in the computer engineering literature as "embarrassingly parallel."

35. Hennessey and Patterson (2012), 411.

36. Like table 11.1, this figure probably does not include R&D costs.

37. Fabless chipmakers Nvidia, AMD, Qualcomm, and Broadcom all publicly complained about a slowdown or even halt to historical decline rates in their manufacturing costs at foundries. Shuler (2015), Or-Bach (2012) (2014), Hruska (2012), Lawson (2013), Qualcomm (2014), Jones (2014, 2015).

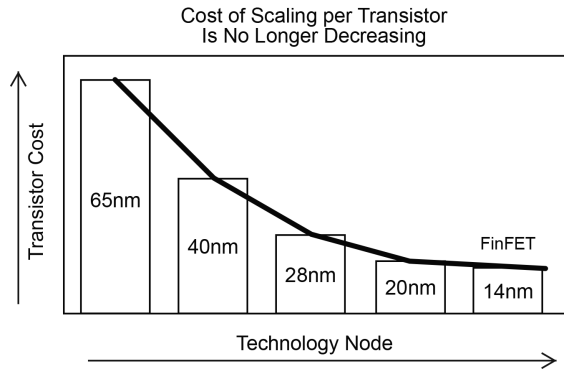

Fig. 11.5 GlobalFoundries' transistor manufacturing cost at recent technology nodes
Source: McCann (2015).

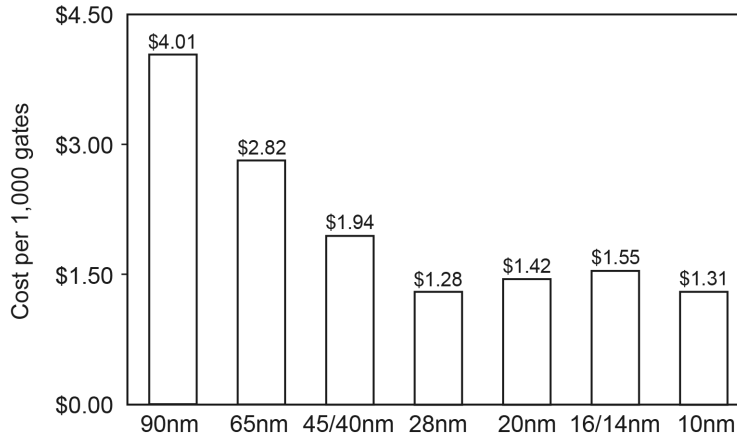

Fig. 11.6 Cost per logic gate, with projection for 10nm technology node
Source: Jones (2015).

(Fabless companies accounted for 25 percent of world semiconductor sales in 2015; foundries, which also build outsourced designs for semiconductor companies with fabs, had a 32 percent share of global production capacity.[38]) Charts like figure 11.6, showing increased costs at sub-28nm technology nodes, were frequently published between 2012 and 2016. Figure 11.6 is not inconsistent with figure 11.5, since figure 11.6 likely includes the fab-

38. Foundry share calculations based on Yinug (2016), Rosso (2016), IC Insights (2016). Charts like figure 11.6 should be viewed cautiously, as underlying assumptions about products, volumes, and costs are rarely spelled out in published sources.

less customer's nonrecurring fixed costs for designing a chip and making a set of photolithographic masks used in fabrication, while figure 11.5—the foundry's processing costs—would not.[39] These fixed costs have grown exponentially at recent technology nodes and create enormous economies of scale.[40] Some foundries have publicly acknowledged that recent technology nodes now deliver higher density or performance at the expense of higher cost per transistor.[41]

Because of these trends, fabless graphics chip specialists Nvidia and AMD actually skipped the 20/22nm technology node, waiting a high-tech eternity—five years—after launch of 28nm graphics processors in 2011 to move to a new technology node (14/16nm) for their 2016 products.

2018: "Dark Silicon" and Limits on Green? The microprocessor industry's response to the end of frequency scaling was to use ever-cheaper transistors to build more cores on a chip. Though limited by software advances in parallelizing different kinds of applications, this strategy at first seemed effective. More recently, continued future improvement of CPU performance on even easy-to-parallelize applications has been questioned.

As transistors get very small, power requirements to switch these transistors are not reduced at the same rate as transistor size. The "green," lower-power benefit of smaller transistors diminishes. Furthermore, as the power density of chips increases, heat dissipation becomes an issue. Thus the heat problem that blocked further frequency scaling returns in a new guise and prevents the increasing numbers of smaller cores squeezed into a multicore chip from simultaneously operating at a chip's fastest feasible clock rate.

The fraction of a chip's cores that must be powered off at all times in order for a chip to operate within thermal limits, dubbed "dark silicon" by researchers modeling the problem, had been projected to grow as large as 50 percent by 2018.[42] Indeed, current PC users are already seeing their multicore machines "throttling" with attempts to use all cores for intensive computations at the highest clock rates, hitting thermal limits and then either falling back to lower clock rates or idling cores. Continued reductions in power requirements are still feasible but no longer are a free benefit of Moore's law—they now come at the cost of reduced speed and additional on-chip circuitry needed to turn off power to unused portions of a processor chip.

39. Historically, a set of 10 to 30 different photomasks was typically employed in manufacturing a chip design. For a low- to moderate-volume product, acquisition of a mask set is effectively a fixed cost.

40. Brown and Linden (2009), chap. 3. McCann (2015) cites a Gartner study showing design costs for an advanced system chip design rising from under $30 million at the 90nm node in 2004, to $170 million at 32/28nm in 2010, to $270 million at the 16/14nm node in 2014.

41. Samsung's director of foundry marketing said, "The cost per transistor has increased in 14nm FinFETs and will continue to do so" (Lipsky 2015). "GlobalFoundries believes the 10nm node will be a disappointing repeat of 20nm, so it will skip directly to a 7nm FinFET node that offers better density and performance compared with 14nm" (Kanter 2016).

42. Esmaeilzadeh et al. (2013), 93–94.

2021+: An End to Smaller in Conventional Silicon? Even some manufacturing technologists from Intel now believe that the Moore's law cadence of technology nodes, with ever-smaller feature sizes in conventional silicon, will end sometime in the next five years. Intel's Bill Holt put it in these terms recently: "Intel doesn't yet know which new chip technology it will adopt, even though it will have to come into service in four or five years. He did point to two possible candidates: devices known as tunneling transistors and a technology called spintronics. Both would require big changes in how chips are designed and manufactured, and would likely be used alongside silicon transistors."[43]

11.3.1 Can We See a Slowing Down of Moore's Law Cost Declines in Price Statistics?

If Moore's law has slowed or even stopped, we would expect to see it in economic metrics, like prices and manufacturing costs.[44]

11.3.1.1 Prices

An obvious place to look is in the price statistics for computer memory chips, which remained the mass-volume semiconductor product par excellence through the end of the 20th century. DRAMs were later superseded by flash memory as the technology driver for new memory manufacturing technology. After the millennium, new technology nodes were first adopted in flash memory chips before DRAMs; flash had become the highest-volume commodity chip by sales around 2012.[45]

Table 11.3 shows changes in price indexes for high-volume memory chips. The DRAM "composite" index is a matched-model, chain-weighted price index based on consulting firm Dataquest's quarterly average global sales price for different density (bits per chip) DRAM components available in the market over the years 1974–1999.[46] These data have no longer been available in recent years.

In the mid-1980s, Korean producers Samsung and Hynix entered the DRAM business and, along with US producer Micron Technology, now account for the vast bulk of current DRAM sales.[47] The Bank of Korea's export price index (based on dollar-basis contracts) and the Bank of Korea's

43. Bourzac (2016).

44. A very useful bibliography of prior matched-model and hedonic studies of semiconductor prices may be found in Aizcorbe (2014), 107–8.

45. See IC Insights (2012).

46. The data prior to 1990 are the same data used in Flamm (1995), figure 5-2. From 1990 on, the data are taken from Aizcorbe (2002).

47. Taiwanese firms entered the DRAM market in force in the early 1990s but have since largely exited, as have all Japanese producers (US producer Micron acquired Japanese DRAM fab facilities). The last remaining European producer (Qimonda) filed for bankruptcy in early 2009. By 2011, the top three producers (Samsung, Hynix, and Micron) accounted for between 80 percent and 90 percent of global sales. See Competition Commission of Singapore (2013).

Table 11.3 **Price indexes for memory chips**

	Compound annual decline rate, quarterly price indexes					
	Flamm-Aizcorbe DRAM composite	Bank of Korea, DRAM export, $ contract price index	Bank of Korea, Flash export, $ contract price index	Bank of Korea, DRAM producer price index, converted to $ at current market rate	Bank of Korea, Flash memory producer price index, converted to $ at current market rate	Bank of Japan, Chain-Wtd MOS Memory Producer Price Index, Converted to $ at Current Market Rate
1974:1–1980:1	−45.51					
1980:1–1985:1	−43.45					
1985:1–1990:1	−24.74					
1990:1–1995:1	−17.40	−10.81				
1995:1–1999:4	−46.37	−44.28		−33.26		
1999:4–2005:1*		−28.94	−31.28	−31.76		−24.04
2005:1–2011:4		−37.94	−26.92	−30.65	−29.28	−28.79
2011:4–2016:4		2.33	−12.70	−1.42	−5.76	−13.57

Source: Author's calculations from sources described in text.

*Bank of Korea Flash export price index and Bank of Japan MOS memory PPI are for 2001:1–2005:1.

producer price index (PPI, converted to a dollar basis using quarterly average exchange rates) for DRAM and flash memory chips are available.[48]

Finally, since 2000, the Bank of Japan has published a chain-weighted "MOS memory PPI" with weights that are updated annually. This index is likely to be predominantly a mix of DRAM and flash memory, tilting more toward flash in recent years. Generally, except for the period from 1985 to 1995, when a string of trade disputes (between the United States and Europe and Japanese, Korean, and Taiwanese memory chip producers) had significant impacts on global chip prices;[49] prices for DRAMs and flash fell at average rates exceeding 20 percent to 30 percent annually.

It is notable that rates of decline in memory chip prices in the last five years generally have been half or less of their historical decline rates over the previous decades. Korean price indexes (which track the majority of the DRAM manufactured and sold) have basically been flat for the last five years. US memory chip manufacturer Micron (like other flash memory manufacturers) is no longer planning to invest in new technology nodes beyond 16nm in its leading-edge flash memory production. Instead, a new device design built vertically (3D NAND[50]) using existing manufacturing process technology is more cost effective than the continued planar scaling of components at new technology nodes described by the Moore's law dynamic.[51] In DRAM, the mantra that "technology-driven growth slows due to scaling limits" ("scaling limits" being industry jargon for a slowing or ending of Moore's law manufacturing cost reductions) had become a staple in Micron's investor conferences.[52]

Another "commodity-like" price in the semiconductor industry in recent years has been the cost that chip design houses face in having their chips

48. These are not well documented but are believed to be fixed-weight Laspeyres indexes, with weights updated every five years, that have been spliced together (2010 is the current base year). The export indexes are actually measured in dollars, while the Korean won-denominated and Japanese yen-denominated producer price indexes have been converted to dollars at current exchange rates. As a practical matter, except for a brief period during the 1980s when export controls related to the US-Japan Semiconductor Trade Agreement were put in place, DRAM prices historically and through the present have been set and quoted in dollars in a highly integrated global market. See Flamm (1993), 163–64, 167–68. Flamm (1995), chapter 5, analyzes empirical evidence that regional price differentials in DRAM briefly appeared and then disappeared when restrictive trade policies were applied and then removed in the 1980s. With minuscule transport costs relative to product value, zero tariff costs globally for most countries (under the Information Technology Agreement, concluded in 1996 and bound into the WTO), and a large number of active global distributor/broker arbitrageurs, the global DRAM market has always been the poster child for the relevance of a "law of one price."

49. See Flamm (1995).

50. Since the early 21st century, the highest-volume semiconductor chips produced have been so-called flash memory chips, and in particular flash memory using Not-AND (NAND) logic (a type of logic circuit) to store binary data. Flash memory is a nonvolatile storage medium—information stored on the chip is maintained after electric power is turned off.

51. Micron 2015 Winter Analyst Conference (2015).

52. Micron's Raymond James Institutional Investor Conference (2016); Micron Analyst Conference (February, 2017).

Table 11.4 **A quality-adjusted price index for fabricated "foundry" wafers**

	Annual index	% rate of change
2004	100	
2005	83.90	−16.10
2006	74.76	−10.89
2007	65.94	−11.80
2008	57.89	−12.20
2009	52.95	−8.53
2010	48.67	−8.09

Source: Byrne, Kovak, and Michaels (2017).

manufactured on their behalf at so-called foundries. The outsourced manufacturing of semiconductors designed at "fabless" semiconductor companies at foundries accounted for about 25 percent of world semiconductor sales in 2015. Foundries, which also build outsourced designs for semiconductor companies with fabs, held 32 percent of global production capacity in that year.[53]

A recent study of quality-adjusted fabricated wafer prices (the form in which manufactured chips are sold to the semiconductor design houses that have outsourced their production) by Byrne, Kovak, and Michaels (2017) portrays a slowing decline in fabricated wafer prices prior to 2012. (See table 11.4.) While the pattern seems consistent with a slowing down of Moore's law prior to 2012, this study unfortunately ends with data from 2010 and thus cannot be used as a check against the claims of the most vocal US fabless designers (see above) that the prices they pay for having their transistors manufactured in foundries were no longer declining significantly at new technology nodes post-2012.

Price Indexes for Intel Processors. Since their invention in the 1970s, microprocessor sales have grown rapidly and since the 1980s have constituted another huge market segment. Official government statistics show a tremendous slowdown in the rate at which microprocessor prices have been falling after the millennium as well as a significant attenuation in the rate at which prices of the desktop and laptop PCs that make use of these processors have declined. The US Producer Price Indexes for microprocessors show annual (January-to-January) changes in microprocessor prices steadily falling from 60 percent to 70 percent peak rates during the "golden age" of the late 1990s and early 2000s to a low of about 1 percent annual decline for the year ending in January 2015. (The Bureau of Labor Statistics stopped reporting its PPI for microprocessors in April 2015, apparently because of confidentiality concerns.) A parallel fall in price declines for laptop and desktop computers seems also to have occurred, from peak annual decline

53. Foundry share calculations based on Yinug (2016), Rosso (2016), and IC Insights (2016).

Table 11.5 Annualized decline rates for microprocessors per the BLS

	Microprocessors (including microcontrollers)		
	Commodity price		Producer price
	Index (discontinued)	Index (current)	Index
1995:1–1999:4	−50.0		−50.5
1999:4–2004:4	−48.6		−49.2
1999:4–2005:1			−47.8
2005:1–2007:4			−37.7
2007:4–2011:4		−10.8	−10.8
2011:4–2015:1		−3.0	−3.0

Author's calculation. Middle month for quarter used, except December 2007 used for 2007:4.

rates of 40 percent in the late 1990s to rates mainly in the 10 percent to 20 percent range in the last few years.

Table 11.5 shows compound annual decline rates in the PPI for microprocessors (including microcontrollers) as constructed by the Bureau of Labor Statistics (BLS), along with similarly defined indexes for the commodity "microprocessors." Annual decline rates slow from a rate near 50 percent in the late 1990s and the first half decade of the new millennium, to a little over 10 percent in the second half of that first decade, to about 3 percent annually in recent years. This too is consistent with a substantial slowing down in the impact of Moore's law manufacturing technology innovation.

The Bureau of Labor Statistics had historically been somewhat opaque about its methodology in constructing its microprocessor price series (there is no published methodology describing precisely how these numbers were constructed).[54] It is believed that these were matched-model indexes based on some weighted selection of products appearing on Intel list price sheets (the same data source I utilize below),[55] but this is not entirely certain. There is also some evidence that the BLS may have experimented with several different methodologies for measuring its microprocessor price indexes over the 1995–2014 periods[56] before ceasing publication of the index for confidentiality reasons in 2015.

54. Ironically, the BLS is now much more open about the details of how it constructs the current (unpublished) microprocessor price index than it was about some previous (published) versions. See Sawyer and So (2017).

55. Based on a brief conversation with BLS officials, Cambridge, MA, July 2014. See also Sawyer and So (2017).

56. The BLS website showed three different "commodity" price indexes (as opposed to its single microprocessor producer price index) for microprocessors over this period. The most recent microprocessor "commodity" price index is based in December 2007 but is only reported monthly from September 2009 through 2015. There are also two discontinued microprocessor commodity price indexes, one based in December 2004 and running through June 2005 and another based in December 2000 and running from 1995 through December 2004. One might speculate that the BLS changed its methodology for measuring microprocessor prices three times during this period.

As an alternative to the BLS measure, I have previously constructed alternative price indexes for Intel desktop microprocessors, tracing the contours of change over time in microprocessor prices using a unique, highly detailed dataset I have collected over the last two decades.[57] Since the mid-1990s, Intel has periodically published, or posted on the web, current list prices for its microprocessor product line in 1,000-unit trays. These list prices are available at a very disaggregated level of detail—distinguishing between similar models manufactured with different packaging, for example—and were typically updated every four to eight weeks, though price updates have sometimes come at much shorter or longer intervals.[58] By combining these detailed prices with detailed attributes of different processor models, it is possible to construct a very rich dataset relating processor prices to processor characteristics, over time.

This permits the construction of both "matched-model" price indexes, the traditional means by which government statistical agencies measure industrial prices, and so-called hedonic price indexes, which relate processor prices to processor characteristics. It is now well understood in the price index literature that there is a close relationship between matched-model indexes and hedonic price indexes.

The Intel dataset permits measuring differences in processor characteristics down to individual models of processors, controlling for such things as processor speed, clock multiplier, bus speed, differing amounts of level 1 (L1), level 2 (L2), and level 3 (L3) cache memory, architectural changes, and particular new processor features and instructions. The latter have become particularly important recently—beginning in mid-2004, Intel dropped processor clock speed as the principle characteristic used to differentiate processors in its marketing and introduced more complex "processor model number" systems that distinguish between very small and arguably minor differences between processors that proliferated at more recent product introductions.

For comparison purposes, I begin by constructing a matched-model price index for Intel desktop processors. Since I do not have sales or shipment data at the individual processor model level, I weight each observed model equally by taking the geometric mean of price relatives for adjoining periods in which the models are observed.[59] A price index based on the simple geometric mean of individual product price relatives (dubbed the Jevons price index) is chained across pairs of adjacent time periods and depicted in figure 11.8. It has the same qualitative behavior as the official government producer

57. See Flamm (2007).

58. My data initially (over the 1995–98 period) made use of compilations of these data collected by others and posted on the web; since 1998–99, most of these data were collected and archived directly from the Intel website.

59. Since there occasionally were multiple price sheets issued within a single month, I have averaged prices by model by month. Since Intel did not issue new price sheets monthly, "adjoining time periods" means temporally adjacent observations.

price index for microprocessors, falling at rates exceeding 60 percent in the late 1990s and slowing to a decline rate under 10 percent since 2009.

This geometric mean matched-model index actually falls a little more slowly than the official US microprocessor PPI, which may be attributable to the fact that the geometric mean index weights all models equally, while the PPI probably uses a subset of the data, with some weighting scheme for models drawn (and replaced periodically) from subsets of processor types. The PPI also uses fixed weights from some base period to weight these price changes, while my Jevons index chains adjoining paired comparisons of models and therefore implicitly allows weights given to different models over pairs of adjoining time periods to evolve over time.

I have also constructed a hedonic price index, using an econometric model that utilizes more of the information available in my sample of Intel list prices. The basic hedonic price model I estimated statistically was

$$\textit{(H0)}\ \mathrm{lprice}_{it} = \mathrm{constant} + d_t + b_a \mathrm{arch\_d}_i + b_p * \mathrm{lproc}_i + b_m \mathrm{lmaxmhz}_i + b_w \mathrm{lbw}_i + b_{co} \mathrm{lcores}_i + b_h \mathrm{ht}_i + b_{ca} \mathrm{lcache}_i + b_g \mathrm{int\_graph}_i + b_{tdp} \mathrm{ltdp}_i + b_{64} \mathrm{em64t}_i + b_{st} \mathrm{eist}_i + b_v \mathrm{vt}_i + u_{it},$$

with the following covariates for chip model i, period t:

- d_t, a time dummy indicator variable for the later period in a pair of adjacent time periods
- $\mathrm{arch\_d}_i$, architecture dummy for Intel chip architecture (e.g., Haswell, Coppermine, Ivy Bridge)
- lproc_i, log of base processor clock rate
- lmaxmhz, log of maximum clock rate if processor has turbo mode, = lproc if not
- lbw_i, log of memory bandwidth (8 × memory bus clock rate if older front-side bus architecture or max memory bandwidth if reported in Intel Ark database)
- lcores_i, log of number of physical cores on chip
- ht_i, hyperthreading (additional virtual core per physical chip core) hardware support, binary indicator variable
- lcache_i, log of maximum cache memory for highest level cache on processor
- $\mathrm{int\_graph}_i$, binary indicator variable for integrated graphics, 1 if on chip graphics
- ltdp_i, log of thermal design power (watts), rating of chip
- $\mathrm{em64t}_i$, binary indicator dummy for Intel 64-bit memory architecture
- eist_i, binary indicator dummy for enhanced Intel speedstep technology (dynamic frequency scaling and power reduction) feature
- vt_i, binary indicator dummy for hardware virtualization support, 1 if virtualization hardware support
- and u_{it}, a statistical disturbance term for chip model i, time period t

Choice of Characteristics. Choice of characteristics was primarily based on a review of the computer architecture literature (discussed below). The most widely used textbook in that literature holds that computer instruction processing performance is based primarily on the *processor architecture* (which determines how many software instructions can be executed per processor clock cycle: IPC, or instructions per clock) and the computer's *clock rate*. Since the mid-2000s, desktop PC processors have further boosted performance by incorporating a *turbo* mode, increasing clock rate to some maximum above the chip's baseline frequency for short periods of time. Frequently, software performance can also depend on its on-chip *(cache) memory size* and on the sustained speed at which a computer can transfer data from its off-chip, secondary memory—its *maximum memory bandwidth*. Over the last decade, additional processor units (cores) have been added to desktop computer processors, and if software can be parallelized and run simultaneously on multiple *cores*, this too will improve performance. In addition, adding hardware support for "virtual cores," so that a hardware processor core can be time-shared simultaneously by two instruction-processing threads, can speed things up—Intel's version of this feature is called *hyperthreading*. Several other features—hardware support for *virtualization* and a *64-bit memory architecture*—can improve computer performance on particular applications, particularly when desktop processors are used in servers. Basic *graphics* are now integrated onto many processor chips, sparing the end user the need to purchase a costly discrete graphics card, which should also affect demand for a processor by consumers. Finally, power consumption is probably the major variable cost of computing (and drives use of relatively expensive cooling systems needed to dissipate heat from high-powered processors). Low thermal design power (TDP) in desktop processors is considered beneficial for this reason,[60] and processor makers like Intel have also developed hardware support for power-saving features in the chip's micro architecture (Intel's proprietary version—enhanced Intel Speedstep—is abbreviated EIST).

Note that maximum memory bandwidth, cache sizes, number of cores, and even TDP typically take on only a handful of discrete values in any two-period estimation sample interval and are often perfectly collinear with binary indicators for processor architecture, 64-bit support, hardware virtualization, and integrated graphics. In addition, as I show below, performance on different SPEC processor benchmark suites is nearly perfectly predicted by a linear combination of a subset of five of these processor characteristics (chip architecture, clock rate, number cores, hyperthreading, turbo mode).

The regression coefficients (weights) on each of these characteristics, however, vary substantially by software benchmark type. Since the mix of software programs run on computers has evolved substantially over time

60. In addition, low power consumption has the additional very important benefit of producing longer battery life in a laptop computer, irrelevant for a battery-less desktop computer processor.

(these changes have led SPEC to periodically revise its various benchmarks), using the underlying characteristics determining processor benchmark performance (rather than a particular benchmark score) seems the more flexible way to accommodate the impact of changes over time in market demand for different types of software applications running on computers.

The very same characteristics that one might expect to affect processor demand would also be expected to affect processor cost on the supply side. Faster chips supporting the highest clock rates are culled from larger numbers of chips fabricated in batches of wafers through extensive testing (a process dubbed "binning" within the industry). Slower- and faster-running chips are sorted into higher and lower performance bins and sold as distinct chip models. Processors with defects in circuitry in their memory caches and feature circuits also have their defective circuitry fused off electronically and are then sold as lower performance chips (with less memory and fewer features). Redundant circuits can be added to a chip design (at a cost, by increasing chip die area) to yield larger shares of chips on a wafer with functioning features. Every desirable feature of a processor also has some incremental cost incurred in order to increase the number of chips produced with that functioning feature—either through a bigger and therefore more costly chip footprint on a silicon wafer (driven by redundant circuitry needed to fix defects) or through the larger numbers of wafers that must be processed in order to get the desired target numbers of chips with functional features and characteristics.

Computer architectures also affect processor cost, as well as performance, since numbers of transistors on a chip, and therefore chip manufacturing cost, are directly related to the chip's architecture. In addition, since at least the early 2000s, Intel has marked the introduction of new manufacturing technology nodes by rolling out improved chip architectural designs when introducing the new node. So manufacturing technology nodes and chip architectural family will be perfectly collinear in a statistical analysis of Intel prices and costs.

In short, the chip characteristics in this hedonic regression would be expected to affect both computing performance and power consumption, as well as processor cost, and are relevant to both the demand and supply cost sides of the market. For that reason, even if a single, perfectly accurate measure of average processor computing performance (a "market average" benchmark based on the relative mix of software applications run by final computer end users in computing service markets at that particular moment in time) existed, changing in perfect lockstep with the changing mix of applications run by different end users,[61] changes in processor character-

61. It is worth noting that the SPEC benchmarks report an unweighted geometric mean of performance in a variety of applications and that these fixed (equal) weights remain fixed over long periods of time (since 2006, as of October 2018) for the SPEC benchmark composite scores.

istics would have additional impacts on price working through processor manufacturing cost and therefore need to be accounted for separately in the estimated hedonic price equation.

One potentially important pitfall in using large numbers of characteristics in a hedonic equation is that many of these characteristics are likely to be perfectly collinear with others. This is a real-world problem. For example, all the chips developed with a new architecture design may, at least initially, have a common size for their highest-level cache, may all have a 64-bit architecture, or may all have hyperthreading. Most regression software will drop perfectly collinear characteristics automatically, and the coefficients of the other covariates (the ones with which the dropped characteristics are perfectly collinear) will include the effects of the dropped covariates in their estimated values.

This can make interpretation of signs and values of hedonic characteristics problematic and liable to big jumps in value (and coefficient interpretation) in different estimation periods, depending on which characteristics are perfectly collinear and which characteristics are dropped (often automatically) by the statistical software. It also may appear at first glance to look like undesirable "coefficient instability."

However, as long as the key variable of substantive interest (the last period time dummy variable in a regression model spanning two adjacent time periods, the coefficient of which is used to construct a hedonic price index) is not perfectly collinear with the other included characteristics variables, there is no difficulty in interpreting the coefficient of the time dummy variable. Fortunately, it is straightforward to check that this is the case by simply running an auxiliary linear regression of the time dummy on all other explanatory covariates and verifying that it is not perfectly predicted by other regression covariates.

Perfect Collinearity in a Simple Hedonic Simulation. The problem of perfect collinearity—and its effects—is very real in my sample of Intel microprocessors. In every single pair of adjacent time periods, multiple characteristics are dropped as perfectly collinear by statistical software. The problems this can create in interpreting regression results are easily illustrated in a simple simulation model.

Consider a simplified, stylized processor market over two adjacent time periods. Suppose that half of manufacturing capacity is used to fabricate a baseline processor architecture (arch_dummy = 0) and half is dedicated to a different architectural alternative (arch_dummy = 1). Suppose that initially, half of fabricated chips from both architectures can run at a clock rate of 1,000, and half at 1,500. All chips manufactured run 500 faster in the later period (i.e., half at 1,500, half at 2,000; think of this as the result of manufacturing process improvement). Substantively, this means there will be a positive correlation between a binary time period indicator variable (first_period = 0, last_period = 1) and processor clock rates.

Let us also suppose that the only thing all processor buyers care about is processing speed on a single, common software application (so we are ignoring the problem of heterogeneity in demand—i.e., which benchmark to run). Further, let's assume that this single measure of speed (software processing performance) relevant to users is perfectly determined by a simple linear function of three processor characteristics:

$$speed = clock\_rate + 500*arch\_dummy + 200*turbo$$

(where "turbo" is a binary indicator for a functioning turbo speedup feature that is enabled in half of the chips produced for each architecture and clock combination).

Each unique combination of architecture, clock rate, and turbo capability under these assumptions can be thought of as a distinct "processor model."[62] With this setup, there are 12 distinct microprocessor models (2 processor architectures × 3 clock rates × 2 turbo values) sold over two periods. Half the models are sold in both periods (the ones running at 1,500), and half are sold only in the beginning or end periods (the models running at the 1,000 and 2,000 clock rates, respectively).[63]

Unit manufacturing cost for the chip is assumed to be given by

$$cost = 50 + 2 * clock\_rate + 2000 * turbo + 500 * arch\_dummy - 10 * end\_period.$$

End-period manufacturing costs decline by $10 for any constant quality "computer model," simulating a uniform $10 drop in manufacturing cost, given any set of fixed model characteristics, over time.

In the spirit of Pakes (2003), we write out an extremely simple hedonic price reduced-form equation:

$$price = 600 + 2 * speed + cost + random\ disturbance\ term,$$

with the first two terms on the right-hand side of the equation reflecting the further assumption that expected markup over incremental unit cost, reflecting user demand, is a linear function of speed alone. After substituting for unit cost (which we typically cannot observe in available data), this gives us a "hedonic price equation" as a function only of observable processor characteristics:

62. I draw a sample of 10 million observations, using pseudorandom draws from independent uniform distributions, to create a simulated population of processor "models," uniformly and independently distributed over architecture, clock rate and turbo feature. Another set of independent, pseudorandom draws from a uniform distribution create a mean zero disturbance term added into the realized sales price on the left-hand side of the hedonic price equation.

63. Because clock rates increase over time, a binary indicator variable for the end period is positively correlated with clock rate but uncorrelated with either architecture or the turbo feature (which are independently and randomly assigned to wafers/chips prior to fabrication).

$$(H1)\ price = 650 + 2 * speed + 2 * clock\_rate + 2000\ turbo + 500 * arch\_dummy - 10 * end\_period + random\ disturbance\ term.$$

The disturbance term in the simulation is drawn from a zero-mean uniform distribution. The assumed across-the-board $10 end-period average reduction in manufacturing cost, conditional on fixed processor characteristics, induces a $10 decline over time in quality-adjusted (constant characteristic) mean price across all computer models (since markup by assumption depends only on speed, in turn a function of the other processor characteristics we are conditioning on).

Most importantly, we cannot actually estimate (H1), because speed, architecture, frequency, and turbo characteristics, as a group, are perfectly collinear with one another (since speed is a linear function of arch dummy, clock rate, and turbo). Since these three chip characteristics exactly determine speed, any three of these four variables exactly determines the value of the fourth. If we were to substitute for speed as a function of its three determinants and so drop it from the hedonic price equation, we get

$$(H2)\ price = 650 + 4 * clock\_rate + 1500 * arch\_dummy + 2400 * turbo - 10 * end\_period.$$

If we substitute for turbo in terms of the other three variables, we get

$$(H3)\ price = 650 + 12 * speed—8 * clock\_rate - 4500 * arch\_dummy - 10 * end\_period.$$

If we substitute for clock_rate in terms of the other three characteristics, we get

$$(H4)\ price = 650 + 4 * speed - 500 * arch\_dummy + 1600 * turbo - 10 * end\_period.$$

And substituting for architecture, we get

$$(H5)\ price = 650 + 3 * speed + clock\_rate + 1800 * turbo - 10 * end\_period.$$

Table 11.6 summarizes a simple simulation demonstrating that with a large simulated sample (10 million observations), a regression model with any of the four above specifications (H2–H5) recovers the above parameters correctly.[64] A key point of substantial practical relevance is that *all four of*

64. Appendix 11.A2 contains the short Stata program giving these simulation results.

Table 11.6 **Simulation of perfectly collinear characteristics in hedonic price equation**

	(drop speed) p	(drop turbo) p	(drop clock) p	(drop arch) p	(speed only) p
time	−10.22***	−10.22***	−10.22***	−10.22***	−75.24***
	(0.258)	(0.258)	(0.258)	(0.258)	(0.677)
clock_rate	4.000***	−7.999***		1.000***	
	(0.000365)	(0.000983)		(0.000517)	
architecture_dummy	1,500.0***	−4,499.8***	−500.1***		
	(0.183)	(0.492)	(0.258)		
turbo dummy	2,399.9***		1,599.9***	1,799.9***	
	(0.183)		(0.197)	(0.197)	
speed		12.00***	4.000***	3.000***	4.130***
		(0.000913)	(0.000365)	(0.000365)	(0.000762)
constant	650.0***	650.0***	650.0***	650.0***	992.5***
	(0.492)	(0.492)	(0.492)	(0.492)	(1.281)
N	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000
R^2	0.980	0.980	0.980	0.980	0.808

Notes: Standard errors in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$. Stata code for this simulation in appendix 11.A2.

these estimable specifications are correct and produce exactly the same estimate for the coefficient of the time dummy variable, the parameter of greatest substantive interest. But the coefficients of the perfectly collinear characteristics need to be interpreted differently in each case as the joint effects of that characteristic plus the effects of the dropped, perfectly collinear characteristic. In fact, there are wild swings in coefficient values (from 12 to 3 for speed and from 1,600 to 2,400 for turbo) and even sign (from 1,500 to –4,500 for arch_dummy) as different candidates from the set of perfectly collinear variables get dropped from the estimated regression specification.

This is important because with large numbers of characteristics in a hedonic regression, particularly with binary dummies, or nominally continuous covariates that in any given time frame take on only a fixed number of discrete values, perfect collinearity among characteristics is very common. Covariates are typically dropped from the regression automatically by the econometric software. If this is happening and different subsets of the perfectly collinear covariates are used in two different time periods, then wild variation in coefficient estimates, rather than representing worrisome instability in (nonperfectly collinear) explanatory covariates selected and used in the estimated regression, should be anticipated.

A second, even more important point is that estimated coefficients for variables that are not in the set of perfectly collinear variables are not affected by which of the perfectly collinear variables is dropped. In this simulation, for example, the estimated effect of the time dummy—the variable of greatest substantive interest, since its coefficient would be used to estimate a hedonic

price index—does not change in value at all as the excluded perfectly collinear variable changes. It is likely to be relatively rare and fairly obvious when a time dummy variable is perfectly collinear with other covariates. In any event, it is easy to verify that the time dummy variable is not perfectly collinear with other included variables by simply running auxiliary regressions of the time dummy against all other explanatory variables, both those included and those dropped as perfectly collinear.

Finally, there is an important specification issue illustrated by this simulation. If one uses speed as one of the explanatory covariates, it is also important to include the full, nonperfectly collinear subset of relevant characteristics affecting cost, even if speed entirely captures the impact of these characteristics from the user demand side. Table 11.6 demonstrates that when only speed and time are used as explanatory variables (last column in the table), bias from the omitted characteristics greatly confounds the coefficient estimate for the time dummy variable, incorrectly magnifying the drop of quality-adjusted price by a factor of 7.5! I return to this point below.

A Hedonic Price Index for Intel Desktop Processors. Model (H0) above was run for each of 162 pairs of adjacent months in which I collected Intel's desktop processor list prices.[65] The first set of adjacent list prices is for January and February 1996. The last pair of adjacent price sheets is for June and July 2014.[66] Overall, R^2 was uniformly high and was not driven primarily by the inclusion of the architectural dummy variables—these were treated as fixed effects, and I also report a "within" R^2 (after demeaning all variables by their group mean), which is also quite high. (See appendix tables 11.A4 and 11.A6.)

The time dummy variables in the above regression were then exponentiated and used to construct price index relatives for adjacent time period

65. The list prices refer to per-chip prices for processors packaged in quantity 1,000 trays sold to original equipment manufacturers (OEMs). By adjacent month, I mean a month and the next month in which an updated list price was published. For example, if Intel issued a price sheet in January, March, April, August, and November of a year, there would be four adjacent month pairs: January–March, March–April, April–August, and August–November. Roughly three-fourths of the monthly observation pairs were a month apart; the next most frequent value observed was two months; the largest time gap between adjacent price lists observed was four months. A hedonic model excluding TDP produced useful estimates for price relatives over 162 adjacent pairs of months. Results for a model with TDP are shown in the appendix tables based on an initial period ending in October 1998, but the problem of a large share of observations lacking a TDP measure does not really fade away until the pair of adjacent months ending in January 2000.

66. The number of processors in early years was very small and characteristics extremely collinear; numbers of processor prices (with TDP) in adjacent month pairs more than double from under 15 to over 30 in late 1999, and estimated price relatives after that date are probably much more reliable. See appendix table 11.A4 and 11.A6 for details on numbers of observations in different adjacent month samples. Entry and exit of architecture and indicator variables from estimation period to period have been color coded in this table. After the first nonzero observation for an indicator variable occurs, blanks indicate the variable was dropped as perfectly collinear. In no case was the time dummy variable perfectly collinear with other covariates; this was checked with auxiliary regressions.

pairs.[67] The resulting price index relatives were then used to chain link these period-to-period indexes into a longer chained price index, shown in appendix table 11.A3.

In addition, I report the values of other coefficients in the hedonic regression in appendix tables 11.A5 and 11.A7, which show how large qualitative jumps in coefficient values from estimation period to period often occur as nonzero values for new characteristics, indicators, or architecture variables that enter and exit the sample, due to perfect collinearity. But there is often perfect collinearity even when there is no new architecture or indicator entering or exiting the sample—this may be seen in the many blank coefficient estimates that appear when architecture or other indicators, or even continuous covariates (which often take on only a handful of discrete values in any single estimation period), are dropped due to perfect collinearity.

The processor architecture family variables are treated as fixed effects and not reported. There were anywhere from one to seven such architecture fixed effects, depending on the pairs of adjacent months used for estimation of the hedonic equation.

Note that nominal power consumption for a processor (TDP, thermal design power) was simply unavailable for most Intel processors released prior to late 1999. I therefore estimated two versions of a hedonic index: one with TDP as a characteristic and one without. TDP is statistically significant when it is used, and therefore the hedonic price index including TDP is the preferred index from 2000 onward (the small numbers of observations with TDP reported prior to late 1999 make these pre-2000 estimates less reliable). I have linked the post-2000 index with TDP to the pre-2000 index without TDP and show this in the final column of table 11.A1 as a composite "best effort" index. The TDP-inclusive and -exclusive indexes are virtually identical from 2000 through January 2005, departing significantly from one another only afterward. Prior to 2000, the earlier the time period, the more limited the available data and the less reliable the resulting estimate.

Figure 11.7 visualizes some of the estimation model summary statistics from appendix table 11.A6 for the TDP variant of the price index (which is also the "composite" index over the period from 2000 onward). The upper panel shows an overall R^2 that across estimation periods averaged .96 and ranged from .91 to .99 from 2000 onward. "Within" R^2 (explained variance after demeaning all variables by architecture fixed effects group means) averaged .92 and ranged from .74 to .99. The lower panel, using a logarithmic

67. One-half of the coefficient's squared standard error was added to the exponentiated coefficient to produce an unbiased estimate of the price relative (the exponentiated coefficient's value). See the sources cited in Triplett (2006, 54n41) for details on the rationale for the correction. Sergio Correia's reghdfe Stata command was used to estimate the hedonic regressions, because it removes noninformative singleton observations for dummy variables from the regression, because it provides detailed reports on perfectly collinear variables, and because it also calculates a "within" R^2—that is, an explained variance of the dependent variable after demeaning all variables within fixed effect groups (in this case, the processor architecture indicator variables were treated as fixed effects).

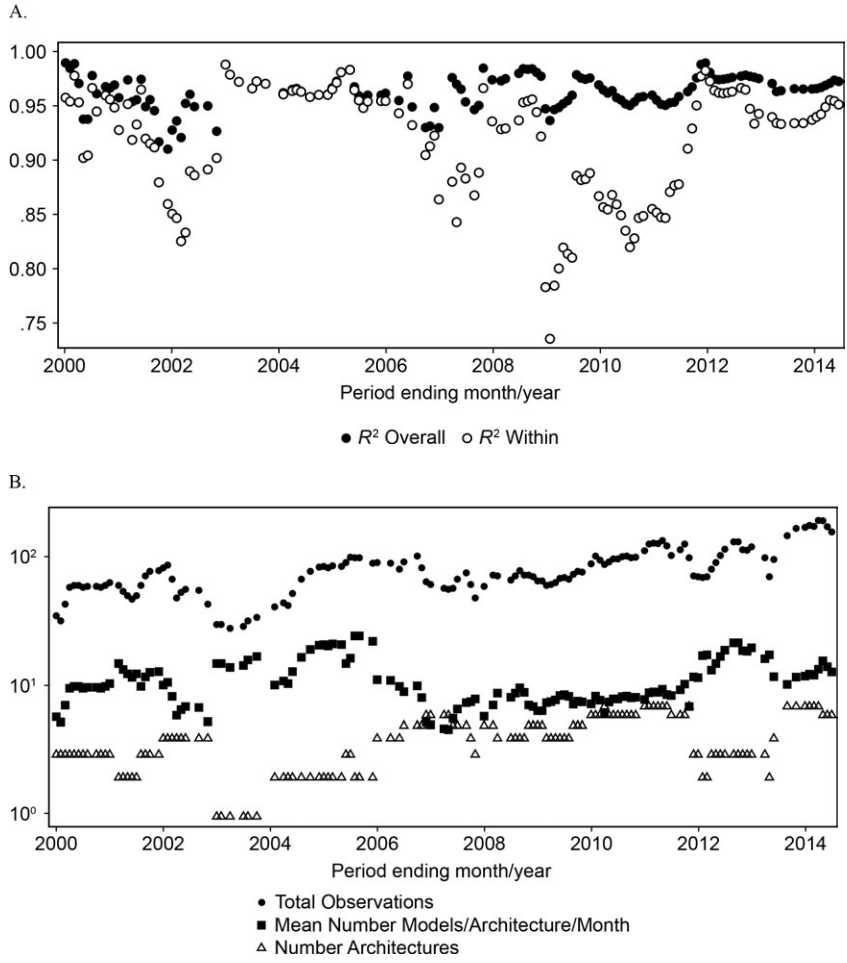

Fig. 11.7 Summary statistics for hedonic regressions
Source: Appendix table 10.A6.

scale, shows that anywhere from one to seven processor architectures were being listed for sale as Intel desktop processors during two-month adjacent estimation periods over this time frame. The number of observations used in the individual hedonic regressions after 1999 ranged from 28 to 190, averaging 82. The average number of processor models per architecture per month listed for sale during the post-1999 period ranged from 4.7 to 24.5, indicative of significant historical changes to Intel's product differentiation strategies in marketing desktop processors over time.

Some important substantive points are supported by figure 11.7. First, there is substantial variation over time in how important the processor design (architecture) dummy variables are in accounting for price varia-

tion. While the overall explained variation in price in these hedonic regressions remained uniformly high, within relatively narrow bounds (.91 to .99) throughout the sample period, the role of architectural dummies varied greatly over different subperiods. "Within" R^2 measures how much of the variation in price around architecture-specific means is explained by other covariates. The "within" R^2 coincides exactly with "overall" R^2 in the special case of their being only one "architecture" fixed effect (i.e., a single common constant intercept). The difference between overall and within R^2 can therefore be interpreted qualitatively as a measure of how important controlling for the multiple intercept levels (the processor architecture fixed effects) is in a hedonic model explaining price variation.

Figure 11.7 shows that, at times, a substantial share of overall explained variation (as much as a difference of .10 to .20 between overall R^2 and within R^2) was accounted for by the processor architecture effects prior to 2003 and from late 2006 through 2012. Processor architecture effects from 2013 onward are more modest contributors to explaining price variation, but not nil.

As is suggested visually by figure 11.7, within R^2 (measuring the role of nonarchitectural characteristics in explaining price variation) has a negative and statistically significant correlation with the number of different desktop processor architectures present on Intel price sheets.[68] Not surprisingly, perhaps, it appears that processor architectural variation is more important in explaining price during periods when Intel marketed a larger variety of processor architectural designs and less important in periods with less architectural variation. Indeed, the two measures of R^2 are virtually identical from 2003 through 2005, the heyday of the Pentium 4 series and its "Netburst" design, when only one or two design families accounted for all Intel desktop processors listed on its price sheets (compared with four architectures in 2002 and as many as seven architectures in late 2006).

Figure 11.8 visualizes the hedonic price indexes produced using these models. A dramatic slowing of declines in quality-adjusted price from 2004 through 2006 is quite apparent, followed by a temporary resumption of a somewhat faster rate of decline after 2006 and then another marked slowdown from 2010 onward.[69]

68. For the TDP-inclusive hedonic specification for adjacent periods ending after December 2000, the correlation coefficient between within R-squared and number of processor architecture dummies used is –.53. I reject the hypothesis that it is equal to zero (p-value is .0000).

69. It is not coincidental that in 2004, the Pentium 4's architecture hit its clock rate ceiling and power dissipation reached maximum limits compatible with inexpensive air cooling systems. The rollout of Intel's next-generation response—the Conroe architecture (two cores on a single die at a much lower clock rate but with more instructions per clock processed)—happened in mid-2006. To many industry observers, Intel appeared to be lagging behind its effectively duopolist rival AMD, architecturally, in the early 2000s. AMD was first to market with a 64-bit architecture and, later, the first single die dual core chip. (AMD had brought its Athlon X2 processor out in 2005, a full year before Intel's Core 2 Duo [Conroe architecture] chips.) For empirical evidence on AMD's technological challenge to Intel in the early 2000s, see Nosko (2011), Pakes (2017), and European Commission (2009).

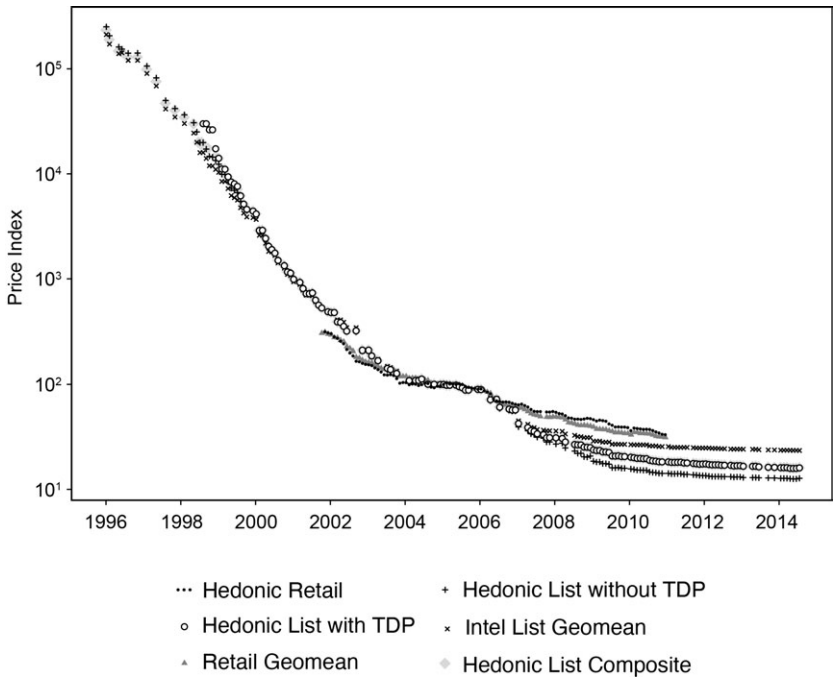

Fig. 11.8 Matched-model and hedonic price indexes for Intel desktop processors, January 2005 = 100

The first four columns in table 11.7 compare my estimated hedonic and matched-model price indexes and the BLS PPIs. As expected,[70] the matched-model geometric mean (Jevons) index price declines are mostly very close to the hedonic indexes but generally decline more slowly than those measured by the hedonic price index based on the same dataset. My estimates over comparable earlier time periods are quite similar to the matched-model indexes of Aizcorbe, Corrado, and Doms (2003) and to the US producer price indexes. Prior to 2004, my Jevons matched-model (geometric mean) index and the PPI move quite closely, while my hedonic indexes show a modestly higher rate of decline, as expected. The hedonic price indexes based on Intel list prices with and without TDP are virtually identical over the period beginning in 2000 through the beginning of 2005.

From 2004 through 2006, both my Jevons and hedonic price indexes decline much more slowly than the PPIs, while from 2006 through 2009, my Jevons and hedonic indexes fall at rates a little faster than the PPI. From 2009 to 2010, the Jevons and hedonic bracket the PPI. Finally, from 2010

70. Since if there were no entering or exiting processor models (all sampled processor models were observed in both time periods) and all hedonic coefficients were the same in the two adjacent periods (assumed by the time dummy method), the time dummy hedonic price index would be equal to the Jevons price index. See De Haan (2010), equation (23), and Triplett (2006), 55.

Table 11.7 **Annualized compound rates of change in microprocessor price indexes**

	Intel OEM list prices			Intel retail		BLS
	Hedonic w/TDP	Hedonic w/o TDP	Jevons matched model	Hedonic	Jevons matched model	Microproc PPI
1998m9–2001m12	−71.5%	−66.2%	−64.0%			−56.8%
2001m12–2004m4	−49.6%	−49.6%	−48.9%	−40.2%	−35.5%	−47.1%
2004m4–2006m1	−9.6%	−10.1%	−10.7%	−4.6%	−11.1%	−25.2%
2006m1–2009m1	−35.4%	−40.3%	−31.5%	−19.9%	−24.2%	−29.0%
2009m1–2010m11	−13.3%	−13.5%	−6.2%	−15.9%	−11.3%	−10.7%
2010m11–2014m7	−3.5%	−2.9%	−2.3%			−4.2%

Source: Author's dataset and calculations, except Microprocessor PPI, from BLS. See appendix table 11.A3.

through 2014, my hedonic indexes fall more slowly than the PPI, but all decline rates are in the low single digits. These are not the only hedonic price indexes for Intel processors available over this time span, and below I discuss alternative estimates that others have constructed.

Using almost the exact same hedonic regression model,[71] I also estimated a hedonic index using weekly data on retail internet pricing for desktop processor models that I had collected over the same time span. The data came from a now-defunct website (sharkyextreme.com) that published the minimum weekly price quoted by a selection of national US internet retailers over the period from the end of 2001 through the end of 2010. Similarly, I calculated a Jevons index based only on matched models in adjacent periods. These prices are a relatively limited subset of the much larger set of list prices for all Intel desktop processors and presumably are more representative of lower-end models most popular in the retail marketplace. Generally, the pattern over time is similar (steepest declines over 2001–4 and 2006–9 and slower declines over 2004–6 and 2009–10).

One interesting observation that emerges from these results is that except for the period from 2006 through the end of 2007, all the Intel list price indexes, including both hedonic and geometric mean matched-model (Jevons) indexes, move together in a fairly tight formation. This can be seen by comparing the original index (with January 2005 = 100) to rebased indexes with January 2010 = 100. (See figure 11.9.) This is consistent with 2006–7 being a highly atypical period, with many more older, exiting models (from now obsolete Pentium 4–branded architecture families) and new

71. With one additional characteristic—a binary "OEM" indicator variable, indicating whether the product sold by the retailer came in a "boxed" retail package with heatsink and fan or it came in "OEM" packaging without a fan, heat sink, and retail box. Monthly average prices were calculated from published weekly reports. The published weekly price was the reported minimum in a sample of larger internet component retailers.

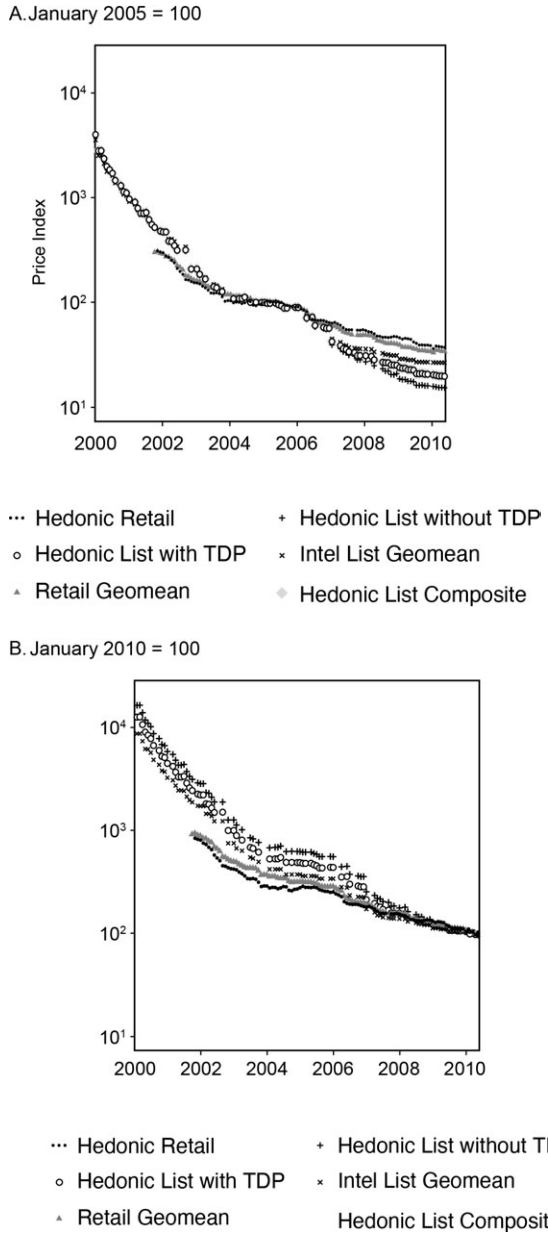

Fig. 11.9 Jevons (geometric mean) and hedonic price indexes with alternative base periods

entering models (from its new Core 2 Duo–branded architecture families) than has generally been the case for Intel historically, before or after this period. The change in Intel's product design strategies from 2006 through 2007, in responding to AMD's earlier technological challenge, has been commented upon by researchers[72] and appears to have had impacts that are visible in these price indexes.

Although there are substantial differences in the magnitude of declines across different time periods and data sources, all the various price indexes I have constructed concur in showing substantially higher rates of decline in desktop microprocessor price prior to 2004, a stop-and-start pattern after 2004, and a dramatically lower rate of decline after 2010.

Taken at face value, this creates a puzzle. Even if the rate of innovation had slowed in particular for microprocessors, if the underlying innovation in semiconductor manufacturing technology had continued at the late 1990s pace (i.e., a new technology node every two years and roughly constant wafer-processing costs in the long run), then manufacturing costs would continue to decline at a 30 percent annual rate, and the recent rates of decline in processor price just measured fall well short of that mark. Either the rate of innovation in semiconductor manufacturing must also have declined, or the declining manufacturing costs are no longer being passed along to consumers to the same extent, or both. The semiconductor industry and engineering consensus seems to be that the pace of innovation derived from continuing feature-size scaling in semiconductor manufacturing has slowed markedly. I next examine what other direct evidence is available.

11.3.1.2 Costs

Evidence on Manufacturing Costs. Microprocessors are a semiconductor product sold in truly large volumes. The overwhelmingly dominant player in this market, Intel, released a slide in a presentation to its stockholders in 2012 that supports the narrative of a slowing down in Moore's law cost declines (table 11.8). The figures from Intel's 2012 Investor Meeting seem to show accelerating cost declines in the late 1990s and rapid declines near a 30 percent annual rate around the millennium, followed by substantially slower declines in cost per transistor after the 45nm technology node (introduced at the end of 2007). As discussed previously, the transition to use of

72. "Note that in June 2006 there was intense competition for high performance chips with AMD selling the highest priced product at just over $1000. Seven chips sold at prices between $1000 and $600, and another five between $600 and $400. July 2006 saw the introduction of the Core 2 Duo and Fig. 2 shows that by October 2006; (i) AMD no longer markets any high performance chips (their highest price chip in October is just over two hundred dollars), and (ii) there are no chips offered between $1000 and $600 dollars and only two between $600 and $400 dollars. Shortly thereafter Intel replaces the non-Core 2 Duo chips with Core 2 Duo's.

"Nosko goes on to explain how the returns from the research that went into the Core 2 Duo came primarily from the markups Intel was able to earn as a result of emptying out the space of middle priced chips and dominating the high priced end of the spectrum." From Pakes (2017), 251–54; see also Nosko (2011), 8–9.

Table 11.8 Annualized decline rates for Intel transistor manufacturing cost, 2012

		Transistor cost index, 90nm = 100 Otellini 2012 Wafer size		Percent transistor cost decline rate between nodes Otellini 2012 Wafer size		Compound annual decline rate Otellini 2012 Wafer size	
Intro date	Tech node	200mm	300mm	200mm	300mm	200mm	300mm
1995q2	350	1,575.35					
1997q3	250	1,033.14		−34.4		−17.1	
1999q2	180	616.10		−40.4		−22.8	
2001q1	130	311.09		−49.5		−32.3	
2004q1	90		100.00		−67.9		−31.5
2006q1	65		48.87		−51.1		−30.1
2007q4	45		27.54		−43.6		−27.9
2010q1	32		17.69		−35.8		−17.9
2012q2	22		11.23		−36.5		−18.3

Source: Otellini (2012), digitized using WebPlotDigitizer. Intro dates: 130nm and up from http://www.intel.com/pressroom/kits/quickreffam.htm. < 130nm from ark.intel.com.

a larger wafer size after the 130nm technology node was accompanied by a particularly large reduction in transistor cost at the next node, using the larger-size wafers.

11.3.1.3 Other Economic Evidence

Depreciation Rates for Semiconductor R&D. Another innovation metric in semiconductors is the depreciation rate for corporate investments in semiconductor R&D. As the rate of innovation increases (decreases), the stock of knowledge created by R&D should be depreciating more rapidly (less rapidly). One recent economic study estimates R&D depreciation rates in a number of high-tech sectors, including semiconductors. The authors conclude that "the depreciation rate of the semiconductor industry shows a clear declining trend after 2000 in both datasets, albeit imprecisely measured."[73] This is consistent with a slowing rate of innovation.

Semiconductor Fab Lives. Faster (slower) technological change in semiconductor manufacturing should presumably shorten (lengthen) fab lifetimes. There are no recent studies of economic depreciation rates for semiconductor plants and equipment, but the anecdotal evidence on the 200mm fab capacity "reawakening" (detailed below) strongly suggests that fab lives have increased, consistent with a slowing rate of innovation in semiconductor manufacturing.

In August 2018, GlobalFoundries (one of four remaining firms that had committed to the development of leading-edge logic manufacturing pro-

73. Li and Hall (2015), 13.

cess technology) announced that it was abandoning its effort to move to its next targeted technology node (7nm) and would stick instead with its current-generation technology: "'The lion's share of our customers . . . have no plans for' 7nm chips. Industry-wide demand for the 14/16 node was half the volume of 28nm, and 7nm demand may be half the level of the 14/16nm node, Caulfield said. 'When we look out to 2022, two-thirds of the foundry market will be in nodes at 12nm and above, so it's not like we are conceding a big part of this market,' he added."[74] This left only three remaining semiconductor manufacturing firms (Samsung, Intel, and TSMC) developing sub-10nm manufacturing technology going forward into 2019.

A slowing pace of innovation in semiconductor manufacturing was even undeniable at Intel. Intel had introduced its 14nm technology node back in 2014 but ran into difficulties bringing its next-generation 10nm technology to market. In August 2018, Intel acknowledged that it was now delaying volume manufacturing of 10nm technology products until late 2019, over five years after its last technology node (i.e., almost triple its previous two-year "tick-tock" cadence between new technology nodes) and almost three years after its initial projection (see table 11.9 below).[75]

Personal Computer Replacement Cycles. One reason for businesses and consumers replacing computers more frequently (less frequently) is if the rate of innovation in key components in computers, like microprocessors, increases (decreases), so performance improvements associated with replacement are more (less) economically compelling. While published studies of PC replacement cycles are scarce, Intel monitors replacement cycles for PCs, a major market for its desktop processors. In 2016, Intel CEO Brian Krzanich noted that PC replacement cycles had extended from four years, the previous average, to five or six years, the current average.[76] This, again, is consistent with a slower rate of innovation.

11.4 Is Moore's Law Still Alive? Intel's Perspective in Microprocessors

The most significant evidence against any current slowdown in semiconductor manufacturing cost reduction from Moore's law had come from Intel. Fairly recent Intel statements about its manufacturing costs had been the primary factual evidence within the semiconductor manufacturing community countering the proposition that Moore's law is ending. Unfortunately, Intel had not been consistent in the data it had presented publicly on this issue. Since late 2017, Intel appears to have refrained from releasing any new public information on its manufacturing costs.

The problem with Intel's previous statements is illustrated by figure 11.10

74. Merritt (2018); see also S. Moore (2018).
75. Rogoway (2018); see also Cutress and Shilov (2018).
76. Krzanich (2016).

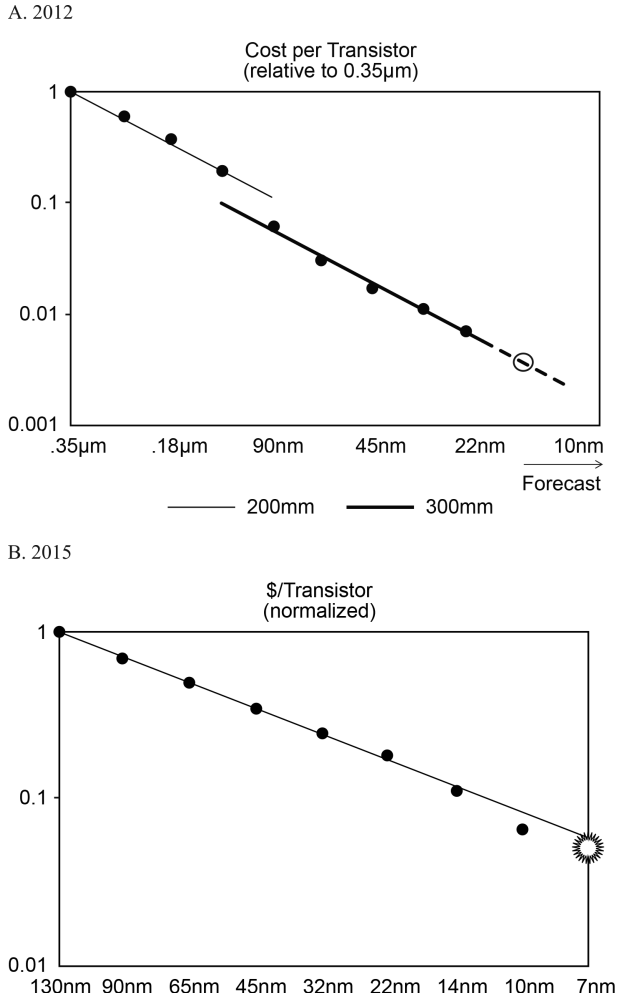

Fig. 11.10 Intel transistor manufacturing costs, 2012 vs. 2015 versions
Source: Otellini (2012); Holt (2015); Intel.

and table 11.9, which contrast two exhibits on manufacturing costs per transistor that Intel had presented at its annual investor meetings—one in 2012 (by then-CEO Paul Otellini) and one in 2015 (by its top manufacturing executive, Bill Holt; see figure 11.2). Some version of the bottom pane in figure 11.10 had been the primary factual evidence in Intel's assertions that Moore's law continues at its historical pace. The graphics in figure 11.10 have been digitized[77] and recorded in table 11.9, then rebased to 100 at the 90nm

77. Using http://arohatgi.info/WebPlotDigitizer/.

Table 11.9 Comparison of Intel cost per transistor at various technology nodes, 2015 vs. 2012

		Transistor cost index, 90nm = 100			Percent transistor cost decline rate between nodes			Compound annual decline rate between nodes		
		Otellini 2012		Holt 2015	Otellini 2012		Holt 2015	Otellini 2012		Holt 2015
		Wafer size			Wafer size			Wafer size		
Intro date	Tech node	200mm	300mm	300mm?	200mm	300mm	300mm?	200mm	300mm	300mm?
1995q2	350	1,575.35								
1997q3	250	1,033.14			−34.4			−17.1		
1999q2	180	616.10			−40.4			−22.8		
2001q1	130	311.09		146.93	−49.5			−32.3		
2004q1	90		100.00	100.00		−67.9	−31.9		−31.5	−12.0
2006q1	65		48.87	71.26		−51.1	−28.7		−30.1	−15.6
2007q4	45		27.54	50.30		−43.6	−29.4		−27.9	−18.1
2010q1	32		17.69	35.64		−35.8	−29.1		−17.9	−14.2
2012q2	22		11.23	26.03		−36.5	−26.9		−18.3	−13.0
2014q3	14			16.13			−38.0			−19.2
2017q4?	Intel 2015 10nm estimate			9.46			−41.4			−21.1
2019q4?	Intel actual 10nm			9.46			−41.4			−9.7

Notes: "300mm?" assumed by author, based on Intel using both 200mm and 300mm wafers with its 130nm tech node. Natarajan et al. (2002). "2017q4?" assumes 2015 Intel forecast of 3 years to next tech node intro date, for 10nm, and 2015 projections of transistor cost decline. "2019q4?" estimated cost decline rate uses Holt (2015) projections of 10nm transistor cost declines at 10nm, but with actual 10nm ship date. Intro dates: 130nm and up from http://www.intel.com/pressroom/kits/quickreffam.htm. < 130nm from ark.intel.com.

technology node. Compound annual decline rates have been calculated in this table using quarterly introduction dates for the first processors manufactured by Intel at that technology node.

The figures presented by Intel to shareholders in 2012 seem to show rapid declines in the 30 percent range around the millennium, then substantially slower declines in cost per transistor after the 45nm technology node (i.e., after 2007). In contrast, a more recent presentation by Intel in 2015 restates the more distant history to show very much slower declines in cost per transistor at earlier technology nodes. Intel has a stock disclaimer that numbers it presents are subject to revision, but in this case the revisions to the historical record are quite dramatic.

The 2015 graphic substantially revises what in the semiconductor industry would be considered the distant historical past (i.e., five technology nodes back from the 22nm node that was in production at the time the earlier 2012 presentation was given). Intel's most recent version of its history now shows transistor costs declining at 12 percent to 18 percent annual rates after the millennium rather than the 30 percent annual declines it showed to its investors in 2012. Its transistor cost decline rate accelerates, rather than slowing further, at the most recent couple of technology nodes.

It now seems likely that one important reason for Intel's restatement of its historical cost declines in 2015 was a definitional change in technical information made public by Intel. Instead of reporting transistor density (transistors per die area) based on actual die area and the number of transistors processed on an actual microprocessor die (which allows one to calculate an actual average of transistors fabricated per die area), Intel apparently began using an entirely theoretical measure of area per designed transistor that appears not to take into account the increasingly relaxed (from design rules) layout of transistors in actual die designs, imposed in part by the need to allow for additional area between transistors needed to fabricate increasingly complex interconnections.[78] (For die designs released prior to 2010, Intel had previously disclosed both actual die size and the number of transistors processed on the die for many of its chip models.)

Most interestingly, assume Intel's 2015 forecast of 10nm transistor manufacturing costs was correct and simply postpone its use in shipped processors from 2017 by an additional two years (2019 was the actual ship date for Intel's first commercial 10nm processors). This delay slows the annual decline rate for its transistor manufacturing costs from 21 percent to 9.7 per-

78. See Flamm (2017, 34) for a brief explanation of this issue. Intel's latest redefinition of its publicly disclosed "transistor density metric" is entirely theoretical: .6 × (transistors in a NAND logic cell / area of a NAND logic gate) + .4 × (transistors in a complex scan logic flip-flop cell / area of complex scan logic flip-flop cell) = # transistors/mm^2. Such a definition does not allow for the practical effects of relaxation (from theoretical design rules) in actual cell layout needed, for example, to accommodate metal interconnections between logic cells. On Intel's new transistor density definition, see Bohr (2017).

cent and implies a marked attenuation of Moore's law–driven cost declines, consistent with the other evidence discussed previously.

11.4.1 An Intel Exception?

Interpreting the recent economic history of Moore's law, how can Intel's description of accelerating declines in manufacturing cost per transistor (as recently as September 2017[79]) be consistent with reports from other chip manufacturers, and their customers, of stagnating cost declines or even cost increases? Increasingly important scale economies provide one plausible and coherent explanation.

Scale economies at the company level are obvious. The cost of a production scale semiconductor fab has increased dramatically at recent technology nodes, and only the very largest chip IDMs (integrated device manufacturers) can depend on their internal demand to justify a fab investment. Intel made this case quite accurately at its 2012 Investor Meeting, predicting that only Samsung, TSMC, and itself would have the production volumes required to economically justify investment in leading-edge fab technology for logic chips by 2016.[80] (Intel overlooked GlobalFoundries, which, by acquiring IBM's semiconductor business in 2015, substantially increased its scale.)[81] Both TSMC and GlobalFoundries are "pure" foundries and achieve their volumes entirely by aggregating the demands of external chip design customers.

Many US-based semiconductor companies have exited chip manufacturing (e.g., AMD, IBM) or stopped investing in leading-edge fabrication while continuing to operate older fabs (Texas Instruments pioneered this so-called fab-lite strategy). Other "pure play" US foundries (e.g., TowerJazz, On Semiconductor) operate mature foundry fabs that remain cost effective for lower volume chips. Long-established American chip companies, such as Motorola, National Semiconductor, and Freescale, disappeared in the course of mergers or acquisitions that continue to reshape the industry.

This consolidation in leading-edge IC fabrication is global. In Europe, there are no manufacturers currently investing in leading-edge technology.[82] In Asia, there are arguably only Toshiba in Japan, Samsung and Hynix in Korea, and foundry TSMC in Taiwan. Firm-level scale economies explain

79. See Smith (2017), slide 6, "Is Moore's Law Dead? No!" Interestingly, since September 2017, Intel has not—to the best of my knowledge—published a claim that its manufacturing cost per transistor continues to decline at rates exceeding previous historical decline rates or is even falling at new technology nodes.

80. Krzanich (2012), slide 19.

81. What constitutes leading-edge technology in memory chips is somewhat more nebulous, and several large memory specialist IDMs (Hynix, Toshiba, Micron) might also arguably be categorized as being near the leading edge. Global Foundries has since announced that it is dropping out of future development of new manufacturing technology nodes.

82. The last remaining leading-edge chipmaker headquartered in Europe, ST Microelectronics, announced in 2015 that it will be relying on foundries for future advance manufacturing needs.

why fewer firms can afford leading-edge fabs but can't explain why Intel's cost per transistor would have declined much faster than that at other producers still investing in leading-edge fabs, particularly the foundries. It's possible that Intel has unique, proprietary technological advantages. A more mundane explanation is that product-level scale economies drive these differences.

In particular, there has been an exponential increase in the costs of the ever more-complex photomasks needed to pattern wafers using lithography tools—a set of masks cost $450,000 to $700,000 back in 2001, at 130nm, compared with a wafer production cost of $2,500 to $4,000 per wafer.[83] At 14nm (updating wafer-production costs using Intel costs in table 11.9 implies 150 percent increases), wafer production cost would be $6,225 to $9,960. By contrast, costs for a mask set at 14nm are estimated to run from $10 million to $18 million, a 22- to 40-fold multiple of 130nm mask costs![27] Lithography cost models suggest that with 5,000 wafers exposed per photomask set (a relatively high-volume product at recent technology nodes), mask costs per unit of output will exceed both average equipment capital cost and average depreciation cost. With smaller production runs for a product, photomask costs become the overwhelmingly dominant element of silicon wafer–processing cost at leading-edge technology nodes.[84]

Intel, with the largest production runs in the industry (perhaps 300 to 400 million processors in 2014[85]), has huge volumes of wafers to amortize the cost of its masks and is certainly benefitting from significant economies of scale. A single Intel processor design (and mask set) is the basis for scores of different processor models sold to computer makers. Processor features, on-board memory sizes, processor speeds, and numbers of functioning cores can be enabled or disabled in the final stages of chip manufacture, and manufacturing process parameters can even be altered to shift the mix of functioning parts in desired ways.[86]

For Intel, this creates average manufacturing costs per chip that are vastly smaller than costs for fabless competitors running much smaller product volumes using the same technology node at foundries. Foundries recoup those much higher per-unit mask costs through one-time charges or through high finished wafer prices charged to its fabless designer-customers. The customer

83. Both 130nm mask and wafer cost estimates were presented by an engineer in Intel's in-house Mask Operation unit (Yang 2001). Mask set cost estimates at 14nm are taken from Black (2013), slide 6.

84. Lattard (2014), slide 6.

85. Based on the fact that Intel publicly revealed that it had shipped 100 million processors a quarter, a record-setting event, in the third quarter of 2014. Intel (2014), 1.

86. When chips are tested after manufacture, the speed, power consumption, and functioning memory and feature characteristics are used to "bin" the processor into one of many different part numbers. As process yields improve over time with experience, new part numbers with faster speeds or lower power consumption are introduced. VanWagoner (2014) is a concise discussion by a former Intel manufacturing engineer of how a large variety of processor models are manufactured from a single unique processor design.

directly bears the much higher design costs per unit if the latest technology node is chosen for the product.

Exponentially growing design and mask costs at leading-edge nodes now make older technology nodes economically attractive for lower-volume products. Higher variable wafer-processing costs per transistor at older nodes are more than offset by much lower fixed design and photomask costs.

Such scale-driven cost disadvantages are increasingly pushing low-volume chip production to older chip-making technology running in depreciated fabs. This is reshaping the economics of chip production, extending the economic lives of aging fabs. Older 200mm wafer fab capacity is now growing rapidly, forecast to expand almost 20 percent by 2020![87]

Historically, this is unprecedented. The additional 200mm capacity coming into service cannot use more-advanced process technologies designed for 300mm wafer-processing equipment. Much lower fixed design and photomask costs with older technology are the primary factor making it economically attractive to fabricate low-volume products. As inexpensive computing penetrates into everyday appliances, "Internet of Things" chip designers are generating low-volume foundry orders for chip designs tailored to market niches, filling these old fabs with chip orders that don't require the greatest possible density.

Is Intel an exceptional case in the semiconductor industry? Is its portrait of recently accelerating manufacturing cost declines reflected in the actual behavior of its product prices? The problem is, Intel does not disclose data on its product pricing to either the public or government statistical agencies, so analysis of what an economist would call a quality-adjusted price is quite difficult.

Alternative Hedonic Price Indexes for Microprocessors. Apart from Intel's pre-2018 declarations of optimism, a second piece of evidence arguing against a slowdown in Moore's law is a study by Byrne, Oliner, and Sichel (2018), which also utilizes the same list price data from Intel (that I used) in making its argument. Using only the first four quarters of prices for recently introduced models, they run an annual time dummy hedonic price model over adjoining pairs of years and find quality-adjusted prices declining at the same rate in 2000–4 as in 2009–13, at about a 42 percent annual rate of decline, and an even more impressive 46 percent decline over 2004–9.[88] This is higher than any of the rates shown for 2004–9 and very much higher than the decline rates post-2009 in table 11.7.

The key differences between my hedonic price indexes and the Byrne, Oliner, and Sichel (2018) hedonic price indexes are that (1) Byrne, Oliner, and Sichel use only a subset of the desktop processors for which their chosen software benchmark scores are available (vs. all desktop processors listed on

87. Dieseldorff (2016).

88. Ibid. Byrne, Oliner, and Sichel (2018) use only the first four quarterly average prices for individual processors and a single explanatory characteristic—performance on a software benchmark—in their hedonic regression.

Intel's current price sheets); (2) Byrne, Oliner, and Sichel include quarterly average list prices for individual processors only during the first four quarters after their introduction onto the market (vs. using all available monthly average list prices); and (3) Byrne, Oliner, and Sichel use only a single processor characteristic (geometric mean of benchmark software performance scores[89]) in their hedonic model (vs. using a much larger set of processor characteristics that I argue is likely to be relevant to both demand and unit cost).

Sample Selection: SPEC Benchmark vs. No SPEC Available. Byrne, Oliner, and Sichel (2018) acknowledge that there are some differences between chips that have benchmark (SPEC) scores available and chips without (SPEC scores are primarily used to compare processor performance in servers and technical computing workstations, which generally use higher-end processors than the consumer market).[90] They report that a matched-model price index using only the SPEC chips generally falls faster than an index using the non-SPEC chips in all time periods. They also report that their matched-model indexes produce a qualitative pattern in price declines over time that is very similar to what is shown in table 11.7 for all Intel desktop processors. Thus these results suggest that the restriction of the price sample to higher-performance processors with SPEC scores may bias estimates of quality-adjusted price declines toward higher rates of price decline but is not responsible for the very different qualitative behavior over time (relatively constant vs. dramatic reductions in rates of decline after 2004).

First Four Quarters Only vs. All Prices. Byrne, Oliner, and Sichel (2018) observe that individual Intel processor list prices very rarely change over time on price sheets after 2011, in contrast to the prior decade. They identify two scenarios they believe may explain this. In one scenario, "Intel offers progressively larger [but unobserved] discounts to selected purchasers as models age,"[91] producing a measurement error for older processors but not recently introduced models. This would complicate estimation of hedonic price indexes using list price data. "The introduction period index would be unbiased even if there are unobserved discounts at the time of introduction provided that these discounts do not vary systematically over time or across models,"[92] while an index using all periods would presumably be biased.

Alternatively, they argue that even if the posted list prices are actual transactional prices, the older chips must be getting progressively more expensive in quality-adjusted terms if their nominal prices do not change, so relative demand for these models must be falling: "By focusing on prices [only]

89. They take the geometric mean of processor performance on industry consortium SPEC's benchmark scores on single program integer and floating-point software test suites. Their procedure for splicing the two or three distinct sets of benchmarks used over their sample period (SPEC2000 and SPEC2006, and possibly SPEC95) over their 2000–2013 sample period is not explicitly described. See figure 11.4 above for evidence that both levels and slopes of these benchmarks change over time when they are compared.

90. Byrne, Oliner, and Sichel (2018), table 2.

91. Byrne, Oliner, and Sichel (2018), 690.

92. Ibid.

at the beginning of each model's life cycle, a regression that applies equal weights to all observations avoids over-weighting models whose quantities have dropped off."[93] These arguments are used to justify using only prices observed during the first four quarters after a model's introduction, discarding the majority of their sample of Intel list prices.

However, in a recent study, Sawyer and So (2017) replicate the substance of the Byrne, Oliner, and Sichel (2018) results over the period after 2009 in a sample utilizing only "early" (first four quarters after introduction) Intel list prices.[94] However, when processor characteristics are added to SPEC scores as explanatory covariates, Sawyer and So show that standard statistical tests decisively reject the exclusion of processor characteristics from a hedonic price equation that also includes SPEC scores.[95] When these other processor characteristics are not excluded, estimates of recent decline rates for quality-adjusted processor prices over time are dramatically smaller than those estimated by Byrne, Oliner, and Sichel.[96] We can reasonably conclude that it is the restriction of hedonic characteristics to benchmark scores only, and not the restriction to early prices, that is producing the pattern of unremittingly high price declines found in Byrne, Oliner, and Sichel over the post-2004 time period.

Sawyer and So (2017) also note that Intel processors are typically sold in their largest volumes only after the first four quarters in which they are available for sale.[97] Intel's own economic expert made this point in its antitrust case before the European Commission, noting that processor production begins with a "ramp-up" phase that "begins with low volumes and typically lasts three to five quarters."[98] Therefore, using price data for a processor only during the first four quarters following its introduction likely would place relatively high weights on products actually being sold in relatively low volumes compared to other products.

It seems reasonable to suggest that this may be a real-world example of omitted variable bias, akin to that created in the last column of the perfect collinearity simulation in table 11.6. However, Byrne, Oliner, and Sichel (2018) articulate some real concerns about use of Intel list price data to measure processor pricing trends. They note "a sharp change over the course of the 2000s in the life-cycle properties of Intel's posted prices . . . In the early period prices fell steeply over a model's life cycle. However, by 2011–2012, price paths are flat or nearly so, with only a few instances of sizable price declines."[99] These observations are spot on.

Figure 11.11 shows the fraction of incumbent (i.e., omitting newly intro-

93. Ibid.
94. Sawyer and So (2017), 8.
95. Ibid., 11.
96. Ibid., 10.
97. Ibid., 14–15.
98. European Commission (2009), 326.
99. Byrne, Oliner, and Sichel (2018), 687.

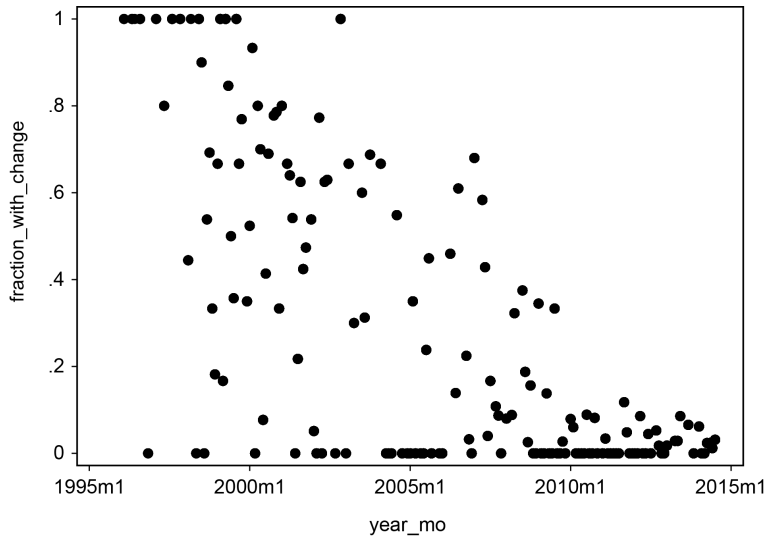

Fig. 11.11 Fraction of Intel desktop processor prices changing from one price list to the next

Source: Author's tabulation from Intel list price dataset.

duced products) desktop processor prices that changed from one list price sheet to the next one issued, from 1998 through mid-2014. Through mid-2014, it is evident that Intel's propensity to alter list prices on existing processors diminished over time, though it never entirely stopped adjusting list prices on its existing product line through mid-2014. In 2008 and 2009, for example, there were price sheets on which anywhere from 35 percent to 40 percent of already introduced desktop processor prices changed from the previous sheet.[100] Since 2014, however, existing processor prices rarely if ever change from one price sheet to the next.

Indeed, if one had to choose a date based on this chart for a climacteric in Intel pricing practices, 2010—the year after its antitrust cases were settled—would seem a promising choice. That year also apparently coincides with the beginning of a determined campaign by Intel to raise its profit margins, an effort that seems to have had some success (aided at that point by a greatly diminished competitive threat from its historical rival, AMD; see figure 11.12). Raising its average sales prices (ASP) was a key element of this strategy. (See figure 11.13.)

In earlier versions of their research, Byrne, Oliner, and Sichel (2018) focused on the evident change in Intel pricing strategies during the first decade of

100. Byrne, Oliner, and Sichel (2018), figure 4, show a similar set of patterns over time in the share of Intel desktop processors with a list price decline within four quarters of introduction.

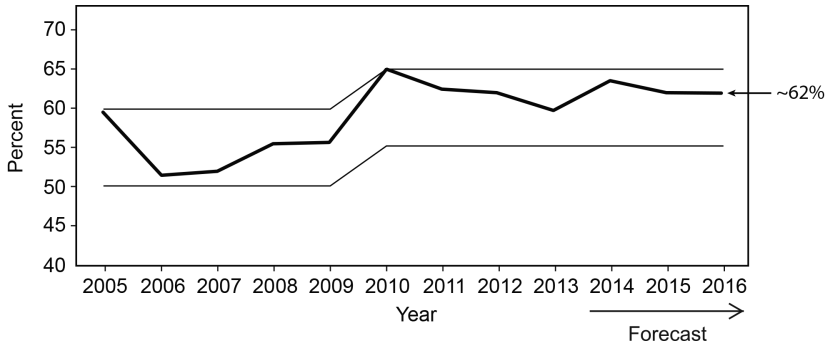

Fig. 11.12 Intel's post-2010 gross margin elevation objective
Source: Smith (2015).

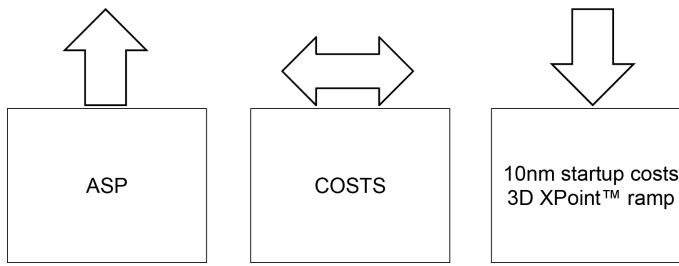

Fig. 11.13 Intel's 2015 explanation to its shareholders for success in maintaining high profit margins
Source: Smith (2015).

the 2000s as the motivation for restricting their Intel prices to "early" initial processor prices.[101] Their hypothesis, that Intel may have changed its pricing strategy during the first decade of the new millennium, actually seems quite plausible given that the European Commission launched a major antitrust case against Intel over its processor price discounting practices during the 2002–6 period, culminating in a preliminary decision against Intel in 2007 and a final decision in 2009.[102] A related private US antitrust case by AMD was filed and then settled in 2009.

The Byrne, Oliner, and Sichel (2018) scenario of "progressively larger discounts to selected purchasers as models age" is difficult to test, since no

101. In the earlier 2017 Federal Reserve working paper version of their study, BOS speculated that "it is possible that Intel actually changed its life-cycle pricing strategy to extract more revenue from older models, with the posted prices reflecting this change." Byrne, Oliner, and Sichel (2017), 8.

102. See European Commission (2009). The same antitrust concerns also resulted in government antitrust actions in Japan and Korea and by the US Federal Trade Commission. Acting on an appeal by Intel, the European Court of Justice sent the EU case back to a lower court for further consideration in 2017, so this seems destined to be litigated for years to come.

data on Intel transaction prices for its wholesale sales to large buyers are publicly available. We do know that evidence produced in the EU antitrust investigation seems to show that even the newest chips sold to large original equipment manufacturer (OEM) customers were heavily discounted from list prices prior to 2006, at times with conditional exclusivity rebates that were not publicly reported by Intel or its customers.[103]

However, there is one public source of Intel transactional price data that is real and observed and does not require any assumptions about unobserved behavior. Retail prices in the electronics industry are linked to wholesale prices, directly and indirectly. Most directly, the very largest retailers can purchase boxed processors directly from Intel or, like smaller retailers, from distributors. (Approximately 20 percent of Intel processors in recent years, by volume, were sold directly as boxed processors, primarily to small computer makers and electronic retailers.[104]) Computer OEMs, electronics system manufacturers, and electronic parts distributors who purchase processors directly from Intel can resell excess inventories to other distributors, resellers, and retailers, and these actually show up on the retail market labeled as "OEM package" (vs. "Retail Box" packaging).

Both boxed and OEM-packaged processors are sold by retailers, distributors, and brokers with a price that is advertised publicly and is directly observable in the marketplace. (The retail data used in constructing my matched-model price index include both OEM and retail-packaged chips sold by internet retailers.) The retail data used in table 11.7 also seem to clearly point to a deceleration in microprocessor price declines after 2004.

It seems reasonable to presume that retail transaction prices (which are observable in the market), at least in the long run, should have some stable stochastic relationship to wholesale producer transactional prices. Indeed, at least one previous study found such linkages between OEM contract transactional prices and retail prices for high-volume chips sold in the semiconductor industry.[105]

There are market-driven economic reasons behind this linkage. Both

103. See European Commission (2009). See also SEC v. Dell Inc. et al. Complaint (US Securities and Exchange Commission 2010), which asserts that unreported exclusivity rebates given by Intel to Dell had climbed to about three-fourths of Dell's operating income by 2006.

104. "Although it sells microprocessors directly to the largest computer manufacturers, such as Dell, Hewlett Packard, and Lenovo, its Channel Supply Demand Operations (CSDO) organization is responsible for satisfying the branded boxed CPU demands of Intel's vast customer network of distributors, resellers, dealers, and local integrators. Intel's boxed processor shipment volume represents approximately 20 percent of its total CPU shipments . . . Processors ship from CW1 to one of four CW2 'boxing' sites, which kit the processors with cooling solutions (e.g., fan, heat sink) and place them in retail boxes and distribution containers. Such boxing sites are typically subcontracted companies that ship the boxed products to nearby Intel CW3 finished-goods warehouses where they are used to fulfill customer orders. Channel customers range in size and need; they are mostly low-volume computer manufacturers and electronics retailers" (Wieland et al. 2012).

105. See Flamm (1993) for a study documenting linkages between retail prices and OEM contract prices for DRAM memory chips.

semiconductor manufacturers and their OEM customers sell their excess inventories of chips to brokers and distributors during industry downturns, pushing small buyer spot prices down in distributor and retail sales channels as excess OEM inventories of chips are absorbed in those sales channels. In tight markets, conversely, when semiconductor manufacturers are capacity constrained, wholesale contract prices to large OEMs rise. To meet surging demand, OEMs may even try to purchase additional volumes of chips, beyond the volumes negotiated in contracts with chip manufacturers, in retail and distribution markets. As both large OEMs and smaller buyers compete fiercely over the remaining unallocated output, upward pressure on retail and distributor prices is felt. In short, both direct and indirect linkages between small buyer (retail and distributor) markets and large buyer (contracts with OEMs) markets, as well as arbitrage across distribution channels, would lead an economist to expect to observe a structural relationship between observed retail processor prices and unobserved large OEM wholesale prices.

In a still earlier version of their research, Byrne, Oliner, and Sichel (2015) had speculated that the change in Intel pricing behavior (resulting in a systematic change in the relationship between Intel list prices and unobserved OEM contract prices) may have occurred after 2006.[106] This is actually an interesting and plausible choice of dates for a change in Intel pricing behavior, since it coincides approximately with the end of the exclusivity rebates that had been the subject of the government and private antitrust actions mentioned earlier. There is also a significant drop in the maximum fraction of Intel list prices changing between adjacent price sheets evident after 2006 visible in figure 11.11 (the last occasions on which 60 percent of prices for existing processors were changed at the end of 2006 and early 2007). If there was a structural shift in Intel pricing practices that caused list prices to diverge more sharply from actual transactional prices after 2006, we might then also expect to see a change in the relationship between movements in observed transactional prices in the retail market and Intel list prices after 2006. This is testable using observational data.

I explored the possibility that there was some detectable change in the relationship between Intel list (posted wholesale) prices and observed retail prices after 2006 by constructing a panel of monthly observations on average retail price and posted list price covering 163 distinct Intel desktop processor models sold by internet retailers over the years 2000 through 2010.[107] I allow for model fixed effects (which permits a particular low-end Celeron model, for example, to be related to Intel list price with a different retail margin

106. "By 2006, this pattern had completely changed; the posted price of a specific model tended to remain constant, even after a new, higher performance model became available at a similar price" (Byrne, Oliner, and Sichel 2016, 9).

107. This is the same sharkyextreme.com data I previously used to construct Jevons and hedonic retail price indexes.

Table 11.10 Fixed effects model of log retail price for Intel desktop processors

	(Full model) lp_ret	(Constrained model) lp_ret
Log Intel Tray Price	0.763***	0.768***
	(15.37)	(17.93)
OEM dummy	−0.0497***	−0.0496***
	(−6.70)	(−6.77)
Age	−0.00676***	−0.00582***
	(−3.70)	(−4.91)
After2006 dummy	0.0204	
	(0.13)	
After2006 × age	0.00162	
	(0.83)	
After2006 × log Intel Tray Price	−0.0108	
	(−0.39)	
Constant	1.347***	1.303***
	(4.87)	(5.55)
N	1,580	1,580
R^2	0.987	0.987
Adj. R^2	0.986	0.986

Notes: *t* statistics in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

than a high-end Core i7 model). The model that I estimated specified the log of retail price as

$$\ln(R_{it}) = a_i + b \ln(I_{it}) + c\, Age_{it} + d\, OEM + e\, After2006 + f\, After2006 \times \ln(I_{it}) + g\, After2006 \times Age_{it} + u_{it},$$

with R_{it} as an observation on average retail price for model i in month t; I_{it} as the average posted Intel list price in a month in which list price had been posted at least once; Age_{it} as the number of elapsed months since the month the model's price had been first posted on a published Intel price sheet; After2006 as a binary indicator variable with value of 1 in 2006 and thereafter and 0 before; OEM as a binary indicator for whether the product sold was the retail boxed version or the bare chip in OEM packaging; and u_{it} as a random disturbance term. If post-2006 transaction prices reflect age discounts from Intel list prices that pre-2006 prices did not, we would expect to find a statistically significant shift coefficient on the interaction of After2006 with Age.

Table 11.10 shows the results of estimating this model.[108] The After2006 shift variable and all of its interactions, including interactions with processor model Age, are close to zero and statistically insignificant individually

108. Robust standard errors clustered on processor models are shown in figure 11.8.

and jointly.[109] The relatively flatter trajectories over time for Intel list prices after 2006 are mirrored in the behavior of flatter retail price trajectories for the same chips.

Therefore, based on the only evidence on actual transaction prices that is publicly available—that is, advertised retail prices from internet-based vendors—there is no evidence of some structural change occurring after 2006 in the relationship between observed Intel list prices and observed retail market prices. Of course, this does not directly prove that there was no change in the relationship between Intel list prices and (unobserved) discounted OEM contract prices for processors, but it certainly weighs against it.

Figure 11.11 and our earlier discussion suggests that 2010–11 is another candidate time period in which to search for a shift in Intel pricing practices. Unfortunately, the retail data analyzed in table 11.10 do not extend past this date.

SPEC scores vs. chip characteristics. As previously remarked, Sawyer and So (2017) have shown that the Byrne, Oliner, and Sichel (2018) results showing no slowdown in quality-adjusted Intel processor price declines since 2000 are not the result of using only "early" Intel list prices but instead are driven primarily by use of SPEC benchmark scores as the sole characteristic in a hedonic model in lieu of a more extensive set of chip characteristics.

The use of SPEC scores instead of actual chip characteristics is based on the argument that direct performance measures are easier to get right than relevant chip characteristics. But this argument overlooks three fundamental reasons why chip characteristics should still be included in a hedonic price equation.

First, there is a computer architecture literature that tells us that benchmark scores of a CPU on any given task should be well explained by a small set of chip characteristics, including numbers of cores and threads, computer architectural design, chip clock rate, and on-chip memory cache sizes. This literature actually identifies the chip characteristics that are relevant and even uses them to model computer CPU performance out of sample.[110] As I next show, scores on various SPEC processor benchmarks are almost perfectly predicted by a linear function of the small set of chip characteristics that the computer design literature predicts are its determinants.

Second, economics tells us that the characteristics that belong in a hedonic price equation are there because they are relevant to user demand and that they have an additional effect on price if they alter supplier marginal cost.[111]

109. The Wald F(3,162) test statistic for the joint hypothesis that all After2006 terms were zero was .8 and the p-value .49.

110. Hennessey and Patterson (2003), in the third edition of their classic computer architecture textbook (59–60) do exactly this to compare the Pentium III with a Pentium 4 operating at the same clock rate.

111. Pakes (2003, 1581, equation 3) notes that the hedonic price function can be interpreted as the sum of the expected marginal cost, conditional on characteristics, and expected markup (derived from the demand function), conditional on characteristics. The key point is that the product characteristics are arguments in the separate cost and demand function terms in the hedonic price equation.

At best, software benchmark scores might correctly serve as a perfect summary measure of quality perceived by users on the demand side. But there is no reason, technological or economic, why a measure of chip performance relevant to demand should also perfectly capture the separate effects of underlying characteristics that determine performance on chip cost. Omitting variation in processor characteristics that affects chip cost will induce omitted variable bias in the hedonic coefficient estimates if the omitted characteristics' effects on cost are correlated (but not perfectly collinear) with the included benchmark scores.

That is, assume for the sake of argument that the mix of user demands for various types of computer applications was fixed over time and that processor performance on this fixed-weight mix of computer applications was correctly captured in some SPEC benchmark. Even with the heroic assumption that this aggregated benchmark correctly captured everything relevant to chip quality on the demand side (and it is clear it does not[112]), there is no plausible technological or economic reason why variations across chip models in marginal production costs related to chip characteristics that determine benchmark scores should be perfectly mirrored by variation in SPEC benchmark scores.

Indeed, the computer architecture literature teaches us that a variety of chip characteristics can affect performance and that, therefore, the same SPEC score can potentially be produced with diverse, nonunique combinations of numbers of cores, threads, cache memory, clock frequency, and so on. In fact, if we look at actual SPEC scores, multiple distinct chip models can produce approximately the same score. But variation in each of these chips' characteristics—cores, threads, on-chip memory, and clock frequencies—may have very different impacts on production cost for the processor compared with impact on SPEC scores.

Third, if benchmark scores are determined by chip characteristics, using chip characteristics directly in the hedonic equation—instead of, or in addition, to a single benchmark score—effectively allows coefficients in the hedonic equation to change to mirror changes in the average mix of tasks run by computer users over time. Use of a single benchmark or fixed-weight index of benchmarks effectively assumes the mix of tasks relevant to performance for users is fixed over time.[113]

112. Since power draw minimization, graphics, and hardware virtualization capabilities clearly are desirable to large subsets of computer users yet will have no direct impact on SPEC scores if missing or disabled in a processor.

113. That is, assume we have two benchmarks, b1 and b2, and two processor characteristics, c1 and c2. Assume b1 = a1 c1 + a2 c2, while b2 = e1 c1 + e2 c2. Assume users in the aggregate run b1 applications 50 percent of the time and b2 applications the other 50 percent. Then we can represent performance on the "average market workload" with a performance index that looks like .5 b1 + .5 b2, or equivalently, .5 (a1 c1 + a2 c2) + .5 (e1 c1 + e2 c2) = [.5 (a1+e1)] c1 + [.5 (a2 + e2)] c2. That is, the benchmark index is equal to a simple linear function of the two characteristics. Now if the weights of b1 and b2 change to 25 percent and 75 percent on the new "market workload," workload performance will be incorrectly captured by the original performance index (50 percent weights) even if scaled by some arbitrary constant. However,

For all these reasons, use of the SPEC score as the sole characteristic in a hedonic price equation is not a highly plausible economic assumption. In addition, because SPEC scores are only available for the subset of Intel desktop processors used by OEMs in servers, the use of SPEC scores in a desktop processor hedonic price regression will considerably reduce sample size compared with statistical models using chip characteristics but not SPEC scores. In the Intel list price data, the number of Intel desktop processors with SPEC scores available for analysis is a fraction of all Intel desktop processors with list prices available in any time period. When using other publicly available retail or distributor desktop processor price data, an even larger fraction of the available data may not have SPEC scores available.[114]

To support this point, I next demonstrate that SPEC processor benchmark scores are almost perfectly predicted by a small number of underlying chip characteristics and provide little or no additional information. In making this claim, I note that I make use of a set of processor microarchitecture dummy variables in the set of chip characteristics used. Neither Sawyer and So nor Byrne, Oliner, and Sichel (2018) use processor architecture dummy variables (which I have shown make an important contribution to the explanatory power of a hedonic price model) in the set of characteristics they employ when estimating a chip characteristic–based hedonic model. It is quite possible that adding a software benchmark score to a set of chip characteristics that excludes the architectural dummies has the effect of capturing much of the effect of these dummy variables in the hedonic price model.

The role of different chip characteristics on different SPEC benchmarks, however, varies greatly across different types of SPEC benchmarks, which argues for direct use of the underlying characteristics in a hedonic equation. It is an argument for letting the data decide what the correct weights on processor characteristics in a hedonic price equation are rather than adopting the implicit weights embedded within a time-invariant weighted average benchmark score.

11.5 Chip Characteristics and Computer Performance: Building Blocks for a Hedonic Analysis

By forcing us to focus on the relationship between performance of microprocessors on representative software benchmarks—which all agree should

performance on "market workload" is still correctly captured by a linear function of the two underlying chip characteristics (though the coefficients of the characteristics in this function change). The specification that is linear in the underlying characteristics is simply more flexible in representing shifts in demand.

114. This is because the selection of processors commonly sold to consumers for use in desktop PCs may include relatively fewer desktop processors used in servers (the ones that would have SPEC scores available).

be an important determinant of chip demand—and chip characteristics, Byrne, Oliner, and Sichel (2018) have done us a great service in providing focus for a discussion of what chip characteristics should be used when estimating a hedonic price equation for microprocessors.

The theoretical computer architecture literature makes use of a *processor performance equation* to predict processor performance. Effectively, this relationship models the execution time a computer processing unit takes to perform some given software benchmark program (i.e., a given sequence of programming instructions) as the product of two parameters: average clock ticks per instruction and the seconds per clock tick in the processor's clock.[115] Since a processor performance benchmark score is proportional to the inverse of time required to run a benchmark program on a particular computer processor, we can invert the processor performance equation and then have

$$\text{Performance} \sim \text{IPC} \times \text{clock rate},$$

where IPC is processed instructions per clock tick, clock rate is measured in ticks per second, and the performance index basically compares benchmark instructions executed per unit time across processors. Indeed, given a particular computer architecture, computer engineers simply scale measured performance linearly by clock rate in order to model the approximate impact of raising clock rate on processor performance.[116]

IPC will depend on both the design (architecture) of the computer processor and the particular mix of instructions being executed in the benchmark software. The specified clock rate of a processor model is typically fixed after testing, at the end of the chip fabrication process.[117] "Binning" during testing of finished chips creates different speed grade bins, which are subsequently sold as different processor models to computer manufacturers and other consumers. The effective, yielded mix of nondefective, more-valuable fast processors and less-valuable slow processors on a fabricated wafer containing hundreds or thousands of these processors is a determinant of processor manufacturing costs.

Speed is not the only chip processor characteristic affected by random fabrication process variation. There may also be random manufacturing variation affecting the voltage needed to run the chip properly, varying from die to die on the same wafer. Chips that require less power to perform cor-

115. See Hennessey and Patterson (2012), section 1.9, 48–52.

116. Hennessey and Patterson (2003), in the third edition of their classic computer architecture textbook (59–60), do exactly this to compare Pentium III performance with a Pentium 4 operating at the same clock rate.

117. Random variation in a highly complex semiconductor manufacturing process leads to a distribution of functional chips by the maximum clock rate at which they can successfully execute some test suite. A "fast" processor can operate at a higher-than-average clock frequency, while a "slow" processor can only operate correctly at a slower-than-average clock rate.

rectly may be identified through testing and sold as low-power models of the processor.[118]

Microprocessor chips generally have on-chip caches of fast local memory that can also affect the execution time for given software. The portion of on-chip cache memory that is defect-free and therefore usable by the chip can also vary with the incidence of manufacturing defects during the fabrication process, and testing then leads to additional binning of finished chips by usable, functional cache memory.

Similarly, particular sections of chip circuitry associated with some advanced features of the chip may not be fully functional due to random processing defects. In order to maximize revenue from all usable products yielded from a finished silicon wafer, a complex system of testing "bins" based on speed, memory, power requirements, and working feature functionality is used to define distinct processor models sold as different chips to final consumers. Indeed, chips are generally designed with some redundant circuitry and electrical "fusing" options intended to maximize saleable product, and revenues, from a processed wafer with dies that may not be perfect. A dozen processor models may be derived from a single, artfully designed die manufactured in the thousands on a single wafer.[119]

At Intel, microprocessor designs are identified with a "microarchitecture," which historically is associated with a publicly available codename. (For example, the processor microarchitecture launched by Intel in October 2017 was given the codename "Coffee Lake."[120]) Prior to 2010, Intel also made public information on its processors' die sizes and the number of transistors on the die processed in its manufacture. Based on this information (which is no longer publicly released), it appears that the many dozens of microprocessor models for each of its microarchitectures were based on somewhere between one and three basic die designs.[121] That is, the dozens of different processor models corresponding to a single microarchitecture product family were manufactured from just one to three basic chip designs fabricated on silicon wafers.

118. And processing of the wafer can be optimized to produce relatively more chips requiring less power.

119. The design of a chip will segment the circuitry into functional blocks that can be disabled electronically (e.g., with programmable "fuses") during the manufacture and testing process. Some redundant circuitry is typically made part of the design, to maximize yield of usable parts after test. A more capable chip can generally be made less capable by disabling portions of its circuitry at the final stages of manufacture. This may be done deliberately by manufacturers to create additional supplies of lower-end chips when customer demand for lower-end parts exceeds the portion of output physically binned into low-end chip models on the basis of test results.

120. Cranz (2017).

121. Prior to 2010, Intel publicly released the exact die area and number of "processing transistors" used in manufacturing most of its microprocessor models. All processors with exactly the same microarchitecture, die area, and numbers of processing transistors can be assumed to be derived from a single die design. Analysis of this data shows anywhere from one to three unique microarchitecture / die size / processing transistor combinations were being used to produce many dozens of processor models.

It is straightforward to analyze the relationship between SPEC scores and microprocessor characteristics. Table 11.10 shows the results from estimating a linear regression model explaining log SPEC scores with a set of explanatory variables suggested by the computer engineering literature: a full set of microarchitecture dummy variables (since IPC is going to depend on computer microarchitecture), log of the base processor clock rate, a dummy variable indicating a "turbo" feature is enabled on the chip (the highest clock rate achievable by a single core on the chip will differ from the base processor clock rate if this feature is available), log of on-chip memory cache size,[122] log of the number of physical processor cores on the chip, and a dummy variable indicating that multithreaded "virtual" logical cores are available on a chip.[123] In addition, a binary indicator variable for use of "autoparallelization" in compiling the SPEC benchmark software code is included, since that can enable a speedup on multicore processors or on processors with multithreading.[124]

A simple log linear regression model that explains SPEC benchmark performance as a function of six processor characteristics (and a full set of 29 to 31 dummy variables for different Intel x86 processor microarchitectures) accounts for a remarkable 97 percent to 98 percent of the variation in SPEC2006 benchmark scores for thousands of computer models using Intel x86 processors over the 2005–17 period (table 11.11). Note that this regression utilizes all Intel x86 desktop, server, and mobile processors in the SPEC2006 database and, further, that it is estimated using every different individual computer making use of an included processor as the underlying set of observations used in estimating the model.

That is, variation in chipsets, motherboards, configured memory, and other components in the computer systems from different manufacturers making use of any particular chip model, which is reflected in the residual, accounted for no more than 2 percent to 4 percent of observed variation in SPEC scores. This analysis utilizes individual tested computer system data—that is, on average there are four to five different computer systems using a specific processor model.

We can alternatively calculate a median or mean score across all computer systems utilizing each processor chip model to more closely resemble the Byrne, Oliner, and Sichel (2018) procedure for deriving a single SPEC score for each chip model. Using that as the basis for our SPEC2006 performance

122. Actually, I am using the size of the "last level cache," since microprocessors can have a hierarchy of successively larger (and slower) caches onboard.

123. Hyperthreading is Intel's name for multithreading capability, additional circuitry added to the processor that creates two logical (or "virtual") processors that can access every physical core. One logical processor can begin processing the next instruction while the other logical processor is actually executing an instruction in a core, thus allowing a form of chip-level parallelism that can speed up performance when a computer program spawns multiple threads.

124. Indeed, after a short number of months at the beginning of the SPEC2006 suite in 2006, almost all the single-process SPEC benchmark scores have autoparallelization turned on.

Table 11.11 Log of SPEC 2006 benchmark as function of processor characteristics

	Six characteristics model			
Dependent variable is log of	SPECf06	SPECi06	SPECfr06	SPECir06
Log base processor speed	0.196***	0.115**	0.383***	0.429***
	(0.0401)	(0.0396)	(0.0590)	(0.0746)
Log cache memory size	0.0965**	0.0861***	0.140**	0.109***
	(0.0283)	(0.0232)	(0.0442)	(0.0208)
Log number physical cores	0.157***	0.0385	0.642***	0.826***
	(0.0284)	(0.0285)	(0.0357)	(0.0249)
Hyperthreading dummy	0.0644**	0.0318**	0.132***	0.201***
	(0.0179)	(0.0111)	(0.0169)	(0.0130)
Log max speed w/turbo	0.514***	0.722***	0.101	0.328***
	(0.0651)	(0.0560)	(0.103)	(0.0747)
Autoparallelization dummy	0.0649*	0.00310	0.0107	−0.0134
	(0.0262)	(0.0534)	(0.0211)	(0.0362)
Microarchitecture dummies	Y	Y	Y	Y
Observations	1,160	1,190	2,207	2,417
R^2	0.966	0.960	0.982	0.974
N_clusters	31	31	29	30
R^2 within	0.687	0.697	0.896	0.893

Cluster robust standard errors in parentheses, clustered on Intel microarchitecture.

* $p < .05$, ** $p < .01$, *** $p < .001$

Log base processor speed is processor base clock rate

Log of max speed is log of maximum clock rate if turbo mode available

Log cache memory is log of amount of last level cache memory on processor chip

Autoparallelization dummy =1 if feature enabled in compiler when SPEC software was compiled

regression model, we get an even higher R^2, of about .99[125] (table 11.12). It is clear that computer architecture dummies and five processor characteristics together essentially perfectly predict SPEC benchmark scores.

Two points are significant. First, the coefficients of (weights assigned to) different processor characteristics in determining SPEC scores are very different for different SPEC benchmarks. The clear implication is that different processor characteristics can have very different effects on performance for different types of workloads. A flexible hedonic price model, reflecting a changing distribution of chip consumers across distinct types of workloads, would best let the empirical data decide the weights users place on particular characteristics rather than aggregating the characteristics into a single benchmark score with the time-invariant weights implicitly used to perform the aggregation into a performance metric.

125. I drop all chips shown as underclocked or overclocked by computer system maker (having reported clock rate more than 10Mz slower or faster than the Intel-specified base clock rate) and ignore autoparallelization in calculating medians or means in table 11.12. Table 11.12 reports results using logs of medians; using logs of means would give almost identical results.

Table 11.12 **Log of median SPEC 2006 benchmark as function of processor characteristics**

Five characteristics model				
Dependent variable is log of median computer system score for particular processor model	SPECf06	SPECi06	SPECfr06	SPECir06
Log base processor speed	0.279***	0.156***	0.507***	0.460***
	(0.0347)	(0.0338)	(0.0767)	(0.0565)
Log cache memory size	0.0783**	0.0575**	0.155**	0.122***
	(0.0259)	(0.0194)	(0.0531)	(0.0184)
Log number physical cores	0.190***	0.0697*	0.644***	0.810***
	(0.0254)	(0.0274)	(0.0513)	(0.0167)
Hyperthreading dummy	0.0721***	0.0371***	0.134***	0.211***
	(0.0133)	(0.00727)	(0.0132)	(0.00788)
Log max speed w/turbo	0.421***	0.677***	−0.0109	0.286***
	(0.0716)	(0.0526)	(0.105)	(0.0575)
Microarchitecture dummies	Y	Y	Y	Y
Observations	331	340	449	454
R^2	0.988	0.985	0.990	0.994
N_clusters	30	30	28	28
R^2_within	0.843	0.853	0.941	0.975

Notes: Cluster robust standard errors in parentheses, clustered on Intel microarchitecture.
* $p < .05$, ** $p < .01$, *** $p < .001$

Second, these characteristics also will affect cost. Every distinct Intel microarchitecture is manufactured using a single fabrication technology node, so in addition to representing the processor's design architecture, the microarchitecture dummies also capture variation in microprocessor manufacturing cost that is induced by variation in chip microarchitectures and manufacturing technology. As previously described, different quality grades (measured by processor clock rates, amounts of on-chip cache memory, and chip features) produced by testing and binning are also associated with cost differences. Coefficients on these characteristics in a hedonic reduced-form price equation should be regarded as reflecting both demand and cost effects.

Finally, in addition to the chip characteristics determining SPEC performance, there is a small set of additional chip characteristics that we would certainly want to include in a hedonic price equation for microprocessors. Power dissipated by a chip determines whether expensive cooling solutions are required, shifting demand for that processor; power requirements are also important (for battery life) in mobile applications. Electricity use, the principle variable cost of computing, will vary with power consumed. Further, power dissipation varies with random manufacturing process variations, so the power rating of a chip is also going to be related to chip cost. Whether or not a graphics processor is integrated into the microprocessor

will also affect both demand and cost for that chip. Support for hardware virtualization will have no practical effect on processor performance on SPEC benchmarks but is a valuable feature for business customers wishing to increase server efficiency by running numerous "virtual machines" on their servers simultaneously.

In conclusion, we should remember that SPEC scores are maintained by organizations that sell servers, processors used in servers, and the largest server customers, so a SPEC-selected sample will be skewed toward the models of chips that perform best as server processors. The SPEC performance regressions in tables 11.11 and 11.12 would then seem to tell us that desktop and server performance should be modeled separately, with different weights placed on different chip characteristics.

This suggests a natural segmentation of microprocessors for purposes of price measurement. A desktop segment oriented toward single software program application performance, a mobile (laptop and tablet) segment tilted toward both performance and low power, and a server segment with a greater emphasis on performance on embarrassingly parallel workloads (servers running a mix of uncoordinated applications with performance more like the SPEC "rate" benchmarks). In terms of finding public data useful in estimating a hedonic price equation, retail/distribution prices will be most readily observable and useful in estimating desktop microprocessor prices. Retail data will be much more limited and less useful for mobile processors and even more limited, and therefore least useful, for hedonic measurement of server processor prices.

The absence of a reliable source of producer transactional data for microprocessors, for use in government price indexes, is a serious and increasingly formidable barrier to measuring prices and innovation correctly in the semiconductor industry.

11.6 Conclusion

There is considerable evidence that semiconductor manufacturing innovation has historically been responsible for perhaps a 20 percent to 30 percent annual decline in the cost of manufacturing transistors on a chip. One would expect that this predictable cost decline would be transformed into a similar price decline in a competitive industry, at least in the long run, and therefore that a decline of this magnitude would serve as a floor on the long-run trajectory of semiconductor prices for high-volume semiconductor products. Innovations in the architecture and designs being manufactured on the chip, new kinds of chip designs, and superior performance characteristics of existing designs fabricated using more-advanced fabrication technology would be additional factors explaining even higher long run rates of decline in quality-adjusted semiconductor prices.

Historically, most high-volume semiconductor applications ultimately migrated to more-advanced manufacturing technology nodes, pulled there by the simple economics of continuing declines in cost using more-advanced fabrication technology. This migration pressure now seems to have lessened, in part the result of rapidly escalating fixed costs that must be sunk into the design of new chips using the most-advanced manufacturing technology and in part due to an apparent slackening in the rate of cost decline at the technological frontier of semiconductor manufacturing.

The available empirical evidence, on balance, suggests that Moore's law–related historical declines in chip manufacturing cost have clearly been attenuated over the last decade. For chips where market price data are collected, decline rates in chip prices over time seem to have greatly diminished. The evidence for exceptionality in Intel microprocessor price declines is shaky, indicative primarily of the increasingly poor quality of publicly available processor price data, changing Intel policies on public release of meaningful list prices for its older processors, and likely, omitted variables in hedonic price models using Intel list price data.

A substantial economic literature has connected faster innovation in semiconductor manufacturing to rapidly improving price performance for semiconductors, to larger price declines for information technology, to increased uptake of IT across the economy, and to higher rates of labor productivity growth. If correct, this implies that a slowdown in semiconductor manufacturing innovation and attenuation of price declines in both chips and IT may play an important role in current stagnation in labor productivity growth.

Finally, it is now almost an article of faith in high-tech industry that an expanding cloud of computing and machine intelligence is in the process of transforming our economy and society. Much of this faith is built on projection into the future based on past experience with increasingly powerful and pervasive computing capabilities that both cost less and use less energy year after year. The winding down of Moore's law means that the technological scaling that drove these historical declines and implicitly underlies the most optimistic assumptions about the spread of ubiquitous computing in the future may no longer hold. Both cost and energy use now seem more likely to increase in lockstep with the scale of cloud computing in the future. Unless there are continuing, significant improvements in software technology, computing costs—and energy use per computation—are unlikely to decline, or even stay constant as computing capacity increases, as was true in the past. Investments in entirely new technologies will be needed, as will a renaissance of creativity and innovation in software. Software, the neglected sibling living in the shadow cast by Moore's law—and dramatically cheapening hardware—for the last 50 years, must increasingly shoulder the burden of delivering comparable economic benefits from continuing technological innovation in information technology.

References

Aizcorbe, A. 2014. *A Practical Guide to Price Index and Hedonic Techniques*. Oxford: Oxford University Press.

Aizcorbe, A. 2002. "Why Are Semiconductor Prices Falling So Fast? Industry Estimates And Implications For Productivity Measurement." Finance and Economics Discussion Series 2002-20, Board of Governors of the Federal Reserve System.

Aizcorbe, A., C. Corrado, and M. Doms. 2003. "When Do Matched-Model and Hedonic Techniques Yield Similar Price Measures?" Federal Reserve Board of San Francisco Working Paper 2003-14.

Aizcorbe, A., K. Flamm, and A. Khurshid. 2007. "The Role of Semiconductor Inputs in IT Hardware Price Decline: Computers versus Communications." In *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches*, edited by E. Berndt and C. Hulten, 351–81. Chicago: University of Chicago Press.

Aizcorbe, A., S. D. Oliner, and D. E. Sichel. 2008. "Shifting Trends in Semiconductor Prices and the Pace of Technological Progress." *Business Economics* 43 (3): 23–39.

Black, R. 2013. "Rambus, Bring Invention to Market." July. http://www.iesaonline.org/downloads/IDC_Presentation_to_IESA_Thought_Leadership_Forum.pdf.

Bohr, M. 2017. "Moore's Law Leadership." Intel Newsroom, March 2017. https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Mark-Bohr-2017-Moores-Law.pdf.

Bourzac, K. 2016. "Intel: Chips Will Have to Sacrifice Speed Gains for Energy Savings." *MIT Technology Review*, February. https://www.technologyreview.com/s/600716/intel-chips-will-have-tosacrifice-speed-gains-for-energy-savings/.

Brown, C., and G. Linden. 2009. *Chips and Change, How Crisis Reshapes the Semiconductor Industry*. Cambridge, MA: MIT Press.

Burgelman, R. 1994. "Fading Memories: A Process Theory of Strategic Business Exit in Dynamic Environments." *Administrative Science Quarterly* 39 (1): 24–56.

Byrne, D., B. Kovak, and R. Michaels. 2017. "Quality-Adjusted Price Measurement: A New Approach with Evidence from Semiconductors." *Review of Economics and Statistics* 99 (2): 330–42.

Byrne, D., S. Oliner, and D. Sichel. 2015. "How Fast Are Semiconductor Prices Falling?" AEI Economic Policy Working Paper 2014-06, revised Nov. 2015. https://www.aei.org/wp-content/uploads/2015/03/Byrne_Oliner_Sichel_Nov-16-2015.pdf.

Byrne, D., S. Oliner, and D. Sichel. 2017. "How Fast Are Semiconductor Prices Falling?" Finance and Economics Discussion Series 2017-005. Washington, DC: Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/econresdata/feds/2017/files/2017005pap.pdf.

Byrne, D., S. Oliner, and D. Sichel. 2018. "How Fast Are Semiconductor Prices Falling?" *Review of Income and Wealth* 64 (3): 679–702.

Competition Commission of Singapore. 2013. "Grounds of Decision Issued by the Competition Commission of Singapore in Relation to the Notification for Decision of the Proposed Acquisition by Micron Technology Inc. of Elpida Memory Inc. Pursuant to Section 57 of the Competition Act." Case number CCS 400/009/12.

Copeland, A. 2013. "Seasonality, Consumer Heterogeneity and Price Indexes: The Case of Prepackaged Software." *Journal of Productivity Analysis* 39 (1): 47–59.

Cranz, A. 2017. "Intel's Latest Coffee Lake Processors Are Fast as Hell." Gizmodo, October 5, 2017. https://gizmodo.com/intels-latest-coffee-lake-processors-are-fast-as-hell-1819129322.

Cunningham, C., et al. 2000. "Silicon Productivity Trends." International Sematech SEMATECH Tech. Transfer #00013875A-ENG, 29 Feb.
Cutress, I., and A. Shilov. 2018. "Intel Server Roadmap: 14nm Cooper Lake in 2019, 10nm Ice Lake in 2020." August 8. https://www.anandtech.com/show/13194/intel-shows-xeon-2018-2019-roadmap-cooper-lakesp-and-ice-lakesp-confirmed.
Dieseldorff, C. 2016. "Watch out for 200mm Fabs!" October 19. http://www.semi.org/en/watch-out-200mm-fabs-fab-outlook-2020-0.
De Haan, J. 2010. "Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and 'Re-pricing' Methods." *Jahrbucher fur Nationalokonomie und Statistik* 230 (6): 772–91.
Esmaeilzadeh, H. E., et al. 2013. "Power Challenges May End the Multicore Era." *Communications of the ACM* 56 (2): 93–102.
European Commission. 2009. *Non-Confidential Version of the Commission Decision of 13 May 2009, Case COMP/37.990 Intel*. May.
Flamm, K. 1993. "Measurement of DRAM Prices: Technology and Market Structure." In *Price Measurements and Their Uses*, edited by M. Foss, M. Manser, and A. Young, 157–206. NBER Studies in Income and Wealth 57. Chicago: University of Chicago Press.
Flamm, K. 1995. *Mismanaged Trade? Strategic Policy in the Semiconductor Industry*. Washington, DC: Brookings Institution.
Flamm, K. 2003. "Moore's Law and the Economics of Semiconductor Price Trends." *International Journal of Technology, Policy and Management* 3 (2): 127–41.
Flamm, K. 2004. "Moore's Law and the Economics of Semiconductor Price Trends." National Research Council, *Productivity and Cyclicality in Semiconductors: Trends, Implications, and Questions: Report of a Symposium*. Washington, DC: National Academies Press.
Flamm, K. 2007. "The Microeconomics of Microprocessor Innovation." Presented at Productivity Workshop, NBER Summer Institute 2007, Cambridge, MA, July. http://conference.nber.org/confer/2007/si2007/PRB/flamm.pdf.
Flamm, K. 2009. "Economic Impacts of International R&D Coordination: SEMATECH and the International Technology Roadmap." In *21st Century Innovation Systems for Japan and the United States: Lessons from a Decade of Change: Report of a Symposium*, edited by K. Flamm and S. Nagaoka. Washington, DC: National Academies Press.
Flamm, K. 2010. "The Impact of DRAM Design Innovation on Manufacturing Profitability." *Future Fab International* 35, November.
Flamm, K. 2017. "Has Moore's Law Been Repealed? An Economist's Perspective." *Computing in Science and Engineering* 19 (2): 29–40.
Fuller, S., and L. Millett, eds. 2011. *The Future of Computer Performance: Game Over or Next Level*. Washington, DC: National Academies Press.
Gandal, N. 1994. "Hedonic Price Indexes for Spreadsheets and an Empirical Test for Network Externalities." *RAND Journal of Economics* 25 (1): 160–70.
Hennessy, J., and D. Patterson. 2003. *Computer Architecture: A Quantitative Approach*, 3rd ed. Cambridge, MA: Morgan Kaufmann.
Hennessey, J., and D. Patterson. 2012. *Computer Architecture: A Quantitative Approach*, 5th ed. Cambridge, MA: Morgan Kaufmann.
Holt, B. 2005. "Facing the Hot Chip Challenge (Again)." Presented at Hot Chips 17. http://www.hotchips.org/wp-content/uploads/hc_archives/hc17/2_Mon/HC17.Keynote/HC17.Keynote1.pdf.
Holt, B. 2015. "Advancing Moore's Law." Presented at Intel Investor Meeting, Santa Clara. http://files.shareholder.com/downloads/INTC/0x0x862743/F8C3E42B-7DA9-4611-BB51-90BED3AA34CD/2015_InvestorMeeting_Bill_Holt_WEB2.pdf.

Howse, B., and R. Smith. 2015. "Tick Tock on the Rocks: Intel Delays 10nm, Adds 3rd Gen 14nm Core Product 'Kaby Lake.'" Anandtech, July. http://www.anandtech.com/show/9447/intel-10nm-and-kaby-lake.

Hruska, J. 2012. "Nvidia Deeply Unhappy with TSMC, Claims 20nm Essentially Worthless." Posted March. http://www.extremetech.com/computing/123529-nvidia-deeply-unhappy-with-tsmc-claims-22nm-essentially-worthless.

IC Insights. 2012. "Total Flash Memory Market Will Surpass DRAM for First Time in 2012." December 19, 2012. http://www.icinsights.com/news/bulletins/Total-Flash-Memory-Market-Will-Surpass-DRAM-For-First-Time-In-2012/.

IC Insights. 2016. "Global Wafer Capacity 2016–20 Product Brochure." http://www.icinsights.com/data/reports/4/0/brochure.pdf?parm=1454865474.

IC Knowledge. 2004. "DRAM Trends." https://web.archive.org/web/20041210172733/http://www.icknowledge.com/trends/dram.html.

Intel. 2007. "Intel Demonstrates Industry's First 32nm Chip and Next-Generation Nehalem Microprocessor Architecture." Press release, September. http://www.intel.com/pressroom/archive/releases/2007/20070918corp_a.htm.

Intel. 2014. "Intel Reports Record Quarterly Revenue of $14.6 Billion." News Release. http://files.shareholder.com/downloads/INTC/2751719461x0x786397/D4904F61-2F5F-48CC-82E2-21A4D0C49583/Earnings_Release_Q3_2014_final.pdf.

Intel. 2016. 2015 *Intel Annual Report*. https://s21.q4cdn.com/600692695/files/doc_financials/2015/annual/2015_Intel_Annual_Report_web.pdf.

Jones, H. 2014. "Why Migration to 20nm Bulk CMOS and 16/14nm FinFETS Is Not Best Approach for Semiconductor Industry." Los Gatos, CA: International Business Strategies, January, p. 1.

Jones, H. 2015. "10nm Chips Promise Lower Costs." *EE Times*, June 15. http://www.eetimes.com/author.asp?section_id=36&doc_id=1326864.

Jorgenson, D. 2001. "Information Technology and the US Economy." *American Economic Review* 91 (1): 1–32.

Kanter, D. 2016. "GlobalFoundries Offers 7nm Roadmap." http://www.linleygroup.com/newsletters/newsletter_detail.php?num=5592.

Krzanich, B. 2012. "Big or Small . . . It's All about the Details." Presentation at Intel Investor Meeting. http://www.cnx-software.com/pdf/Intel_2012/2012_Intel_Investor_Meeting_Krzanich.pdf.

Krzanich, B. 2016. "Intel Corporation's (INTC) CEO, Brian Krzanich Presents at Sanford C Bernstein Strategic Decisions Conference 2016 - Brokers Conference Transcript." June 1. http://seekingalpha.com/article/3979164-intel-corporations-intc-ceo-brian-krzanich-presents-sanford-cbernstein-strategic-decisions?part=single.

Lattard, L. 2014. "Mask Less Lithography for Volume Manufacturing." SEMICON Europa. http://semieurope.omnibooksonline.com/2014/semicon_europa/SEMICON_TechARENA_presentations/TechARENA1/Lithography/02_Ludovic%20Lattard,%20Cea-Leti.pdf.

Lawson, S. 2013. "The Moore's Law Blowout Sale Is Ending, Broadcom's CTO Says." *PC World*, December 5. http://www.pcworld.com/article/2069740/the-moores-law-blowout-sale-is-ending-broadcoms-cto-says.html.

Li, W., and B. Hall. 2015. "Depreciation of Business R&D Capital." Working paper, November. https://eml.berkeley.edu/~bhhall/papers/LiHall16_bus_rnd_depreciation.pdf.

Lipsky, J. 2015. "Samsung Describes Road to 14nm." *EE Times*, April 16. http://www.eetimes.com/document.asp?doc_id=1326369.

McCann, D. 2015. "Silicon Interconnect, Packaging and Test Challenges from a

Foundry Viewpoint." June. http://www.swtest.org/swtw_library/2015proc/PDF/SWTW2015_Keynote_McCann_GlobalFoundries.pdf.
Merritt, R. 2017. "TSMC, Samsung Diverge at 7nm." EE Times, February 8, 2017. http://www.eetimes.com/document.asp?doc_id=1331324.
Merritt, R. 2018. "GlobalFoundries Halts 7nm Work." *EE Times*, August 28. https://www.eetimes.com/document.asp?doc_id=1333637.
Moammer, K. 2017. "Intel Delays 10nm Cannon Lake CPUs to End of 2018." Wccftech, September 20, 2017. http://wccftech.com/intel-delays-10nm-cannon-lake-cpus-end-2018/.
Moore, G. 1965. "Cramming More Components onto Integrated Circuits." *Electronics* 38 (8): 114–17. Reprinted in 1998 in *Proceedings of the IEEE* 86 (1): 82–85.
Moore, G. 1975. "Progress in Digital Integrated Electronics." *Technical Digest. International Electron Devices Meeting*, 11–13.
Moore, S. 2018. "GlobalFoundries Halts 7-Nanometer Chip Development." *IEEE Spectrum*, August 28. https://spectrum.ieee.org/nanoclast/semiconductors/devices/globalfoundries-halts-7nm-chip-development.
Natarajan, S., et al. 2002. "Process Development and Manufacturing of High-performance Microprocessors on 300mm Wafers." *Intel Technology Journal* 6 (2), May.
Nosko, C. 2011. "Competition and Quality Choice in the CPU Market." Harvard University, June.
Oliner, S., and D. Sichel. 1994. "Computers and Output Growth Revisited: How Big Is the Puzzle?" *Brookings Papers on Economic Activity* 25 (2): 273–334.
Or-Bach, Z. 2012. "Is the Cost Reduction Associated with IC Scaling Over?" *EE Times*, July 16.
Otellini, P. 2012. "Investor Meeting 2012." Presentation to Intel Investor Meeting, Santa Clara, CA.
Pakes, A. 2003. "A Reconsideration of Hedonic Price Indexes with an Application to PCs." *American Economic Review* 93 (5): 1578–96.
Pakes, A. 2017. "Empirical Tools and Competition Analysis: Past Progress and Current Problems." *International Journal of Industrial Organization* 53(C).
Pillai, U. 2013. "A Model of Technological Progress in the Microprocessor Industry." *Journal of Industrial Economics* 61 (4): 877–912.
Pirzada, U. 2016. "Exclusive: Is Intel Really Starting to Lose Its Process Lead? 7nm Node Slated for Release in 2022." Wccftech, September 10, 2016. http://wccftech.com/intel-losing-process-lead-analysis-7nm-2022/.
Prudhomme, M., and K. Yu. 2005. "A Price Index for Computer Software Using Scanner Data." *Canadian Journal of Economics* 38 (3): 999–1017.
Qualcomm. 2014. "Qualcomm Snapdragon Integrated Fabless Manufacturing." January 2014, 4. https://www.qualcomm.com/documents/qualcomm-snapdragon-integrated-fabless-manufacturing.
Raley, T. 2015. "IBM z13 Overview and Related Tidbits." Presentation, March. https://www.ibm.com/developerworks/community/wikis/form/anonymous/api/wiki/33d270cb-c060-40f6-99f3-956c3cb452a3/page/a3b86697-49c1-4be0-b247-805276033049/attachment/f49e69a1-fb8d-4710-a23e-0318bbf76e83/media/IBM%20z13%20Overview%20for%20DFW%20System%20z%20User%20Group_2015Mar.pdf.
Rogoway, M. 2018. "Intel Splits Up Manufacturing Group amid Production Delays." *The Oregonian*, October 17. https://www.oregonlive.com/silicon-forest/index.ssf/2018/10/intel_manufacturing_vp_sohail.html.
Rosso, D. 2016. "Global Semiconductor Sales Top $335 Billion in 2015." Febru-

ary. http://www.semiconductors.org/news/2016/02/01/global_sales_report_2015/global_semiconductor_sales_top_335_billion_in_2015/.

Sawyer, S., and A. So. 2017. "A New Approach for Quality Adjusting PPI Microprocessors." Presented at NBER Summer Institute, July 18. http://www.nber.org/conf_papers/f97472/f97472.pdf.

Shuler, K. 2015. "Moore's Law Is Dead: Long Live SoC Designers." February. http://www.design-reuse.com/articles/36150/moore-s-law-is-dead-long-live-soc-designers.html.

Smith, S. 2015. "Investor Meeting 2015 Santa Clara." Presentation to Intel Investor Meeting, Santa Clara, CA.

Smith, S. 2017. "Strategy Overview." Presented at Intel Technology and Manufacturing Day China, September 19. https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/09/stacy-smith-on-milestones-in-intels-process-technology-roadmap.pdf.

Spencer, W., and T. Seidel. 2004. "International Technology Roadmaps: The US Semiconductor Experience." National Research Council, *Productivity and Cyclicality in Semiconductors: Trends, Implications, and Questions: Report of a Symposium*. Washington, DC: National Academies Press.

Triplett, J. 2006. *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes*. Paris: OECD.

US Securities and Exchange Commission. 2010. *SEC v. Dell Inc. et al. Complaint*. https://www.sec.gov/litigation/complaints/2010/comp21599.pdf.

VanWagoner, J. 2014. "How Does Intel Design and Produce So Many Models of CPUs?" https://www.quora.com/How-does-Intel-design-and-produce-so-many-models-of-CPUs.

White, A. J., et al. 2005. "Hedonic Price Indexes for Personal Computer Operating Systems and Productivity Suites." *Annales d'economie et de Statistique* 79/80: 787–807.

Wieland, B., P. Mastrantonio, S. P. Willems, and K. G. Kempf. 2012. "Optimizing Inventory Levels Within Intel's Channel Supply Demand Operations." *Interfaces* 42 (6): 517–18.

Yang, C. 2001. "Challenges of Mask Cost and Cycle Time." October. http://www.sematech.org/meetings/archives/litho/mask/20011001/K_Mask_cost_Intel.pdf.

Yinug, F. 2016. "Made in America: The Facts about Semiconductor Design." June. http://www.semiconductors.org/clientuploads/Industry%20Statistics/White%20Pape%20Profile%20on%20the%20U.S.%20Semiconductor%20Design%20Industry%20-%20061016%20-%20Final.pdf.