Measuring Moore's Law:

Evidence from Price, Cost, and Quality Indexes

November 2017 Version

Kenneth Flamm

Kenneth Flamm
University of Texas at Austin
kflamm@mail.utexas.edu

Preliminary Early Draft

Measuring Moore's Law

Kenneth Flamm

November 2017

"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. 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, including materials, optical systems, ultraclean precision manufacturing, factory automation, electronic circuit design and simulation, and improved computer software for computational modelling 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.[1] 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 1 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.

Section 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 shows rates of performance improvement in CPUs declining 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 motivate quality adjustment methods when semiconductor product prices are measured.

[1] Defining 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.

Section 3 analyzes empirical evidence of recent changes to the historical Moore's Law trajectory, and finds corroborating evidence for a slowdown in Moore's Law in prices for the highest volume products: memory chips, custom chip designs outsourced to dedicated contract manufacturers (foundries), and Intel microprocessors. Section 4 reviews evidence to the contrary, which paradoxically, also relates primarily to Intel microprocessors, and discusses economic reasons why Intel microprocessor prices might behave differently from prices for other types of semiconductor chips.

Section 5 dives into microprocessors in greater depth, and tests the computer architecture textbook view of how a small set of specific chip characteristics affect performance of microprocessors in executing programs, by outlining a structural model of microprocessor computing performance, then estimating that model empirically. This simple econometric model, using only a small set of explanatory chip characteristics, explains 99% of variance across processor models in performance on commonly used CPU performance benchmarks. 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 would therefore have an additional rationale for inclusion in a hedonic price equation quite apart from their role in determining computer performance benchmark scores.

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 co-found Intel) predicted that the number of transistors (circuit elements) on a single chip would double every year.[2] 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).[3] Chip fabrication requires coordinating multiple technologies, combined in very complex manufacturing processes.

The pacing technology has been the photolithographic processes 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, three years also happened to be the time interval between introductions of next-generation DRAM computer memory chips, storing four times the bits in the previous generation chip.[4] This observed 18-month "doubling period" became a new, *de facto*, "revised" Moore's law.[5]

[2] Moore (1965).

[3] Jorgenson (2001), Flamm (2003), (2004); Aizcorbe, Flamm, and Khurshid, (2007).

[4] The DRAM memory was invented in 1968 by Robert Dennard at IBM, and first commercialized by Moore's newly founded company, Intel, in 1970.

[5] A decade later, Moore himself revised his prediction to a doubling every two years. G. E. Moore, ''Progress in digital integrated electronics,'' in *Tech. Dig. IEEE Int. Electron Devices Meeting*, 1975, pp. 11–13.

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[6] trends for integrated circuits can be described by the following identity, giving cost per circuit element (e.g., transistor):

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

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 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."[7]

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):

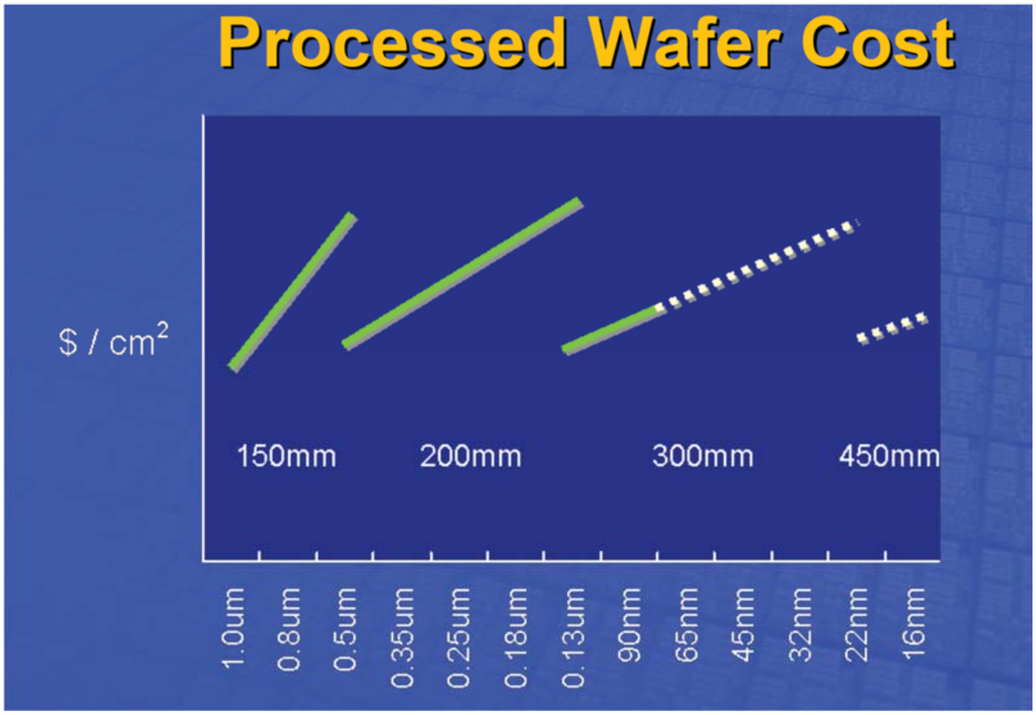
$$(2)\ \$/\text{element} = \frac{\$\text{ processing cost}}{\text{area yielded silicon}} \times \frac{\text{silicon area}}{\text{element}}$$

which depends directly on the defining characteristic of a new technology node, 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 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%, implying 50% smaller silicon areas (F^2) per circuit element.

[6] Analysis of fabrication costs, which account for most chip cost, ignores assembly, packaging, and test.

[7] Moore (1965). The largest wafer sizes in use then were comparable in diameter to a modern snack mini-pizza appetizer.



Source: Holt (2005).

Figure 1. Wafer size conversions offset Intel’s increased wafer-processing cost

Completing the economic story, cost per silicon wafer area processed, averaged over long periods, increased only slowly.[8] 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 1, presented in 2005 by Intel’s chief manufacturing technologist, shows new wafer sizes “resetting” wafer-processing costs. Significantly, larger diameter wafer sizes (450 mm) were expected at the 22 nanometer (nm) node. However, 450 mm wafers were not introduced as Intel adopted 22 nm technology in 2012, had not been introduced by 2017, and even future introduction now seems highly uncertain. The most recent wafer 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%, a 21% 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)—

[8] Over 1983-1998, wafer-processing cost/cm^2 silicon increased 5.5 percent annually. Cunningham et. al. (2000), p. 5. This estimate relates to total silicon area processed (including defective chips). Since defect-free chips’ share of total processed area increased historically, wafer-processing cost per good silicon area rose even more slowly, approximating constancy.

dropped in price substantially faster than 20% 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)[9], 3-D interconnect layers enabling tighter packing of circuit elements,[10] and gradual introduction of 3-D into physical designs of transistors and other circuit elements.[11] 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 economic 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.[12]

In the mid-1990s, the semiconductor manufacturing industry arrived at a significant technological inflection point.[13] New technology nodes began arriving at two-year intervals, replacing three-year cycles. (Intel's perception of this trend, as of 2005, is documented in Figure 2.) The origins of this change lie in the early 1990s, when the U.S. SEMATECH R&D consortium sponsored a roadmap 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 roadmap evolved into the international ITRS.[14] 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.[15]

[9] Flamm (2010), Figure 2, documents a 62 percent decline in minimum memory bit cell footprint between 1995 and 2004.

[10] Anticipated by Moore in 1965: "no space wasted for interconnection…using multilayer metallization patterns separated by dialectric films.."Moore (1965).

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

[12] Since electrical interconnections between components have historically been the most frequent point of failure in electronic systems.

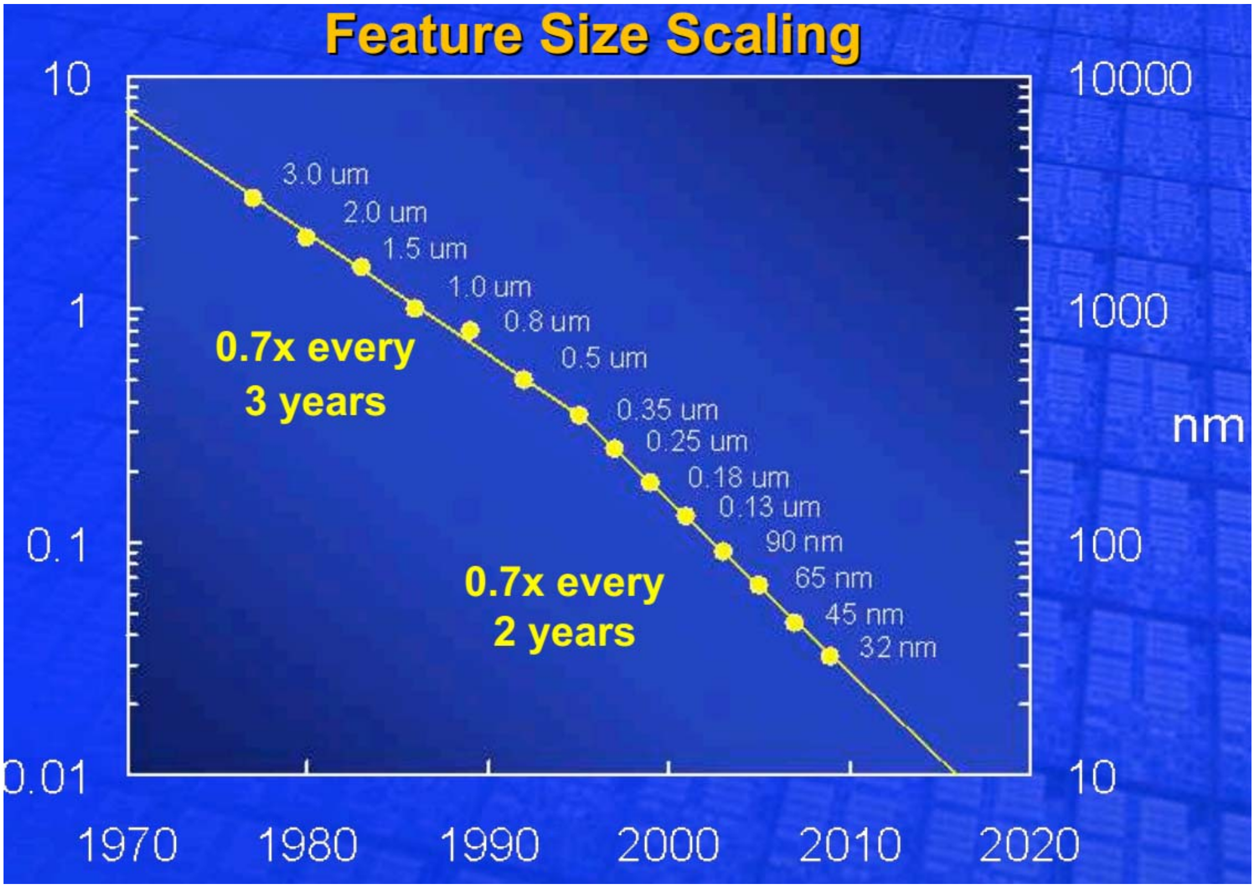
[13] Industry roadmaps originally dated this transition to two-year node rollouts to 1995; post-2004 roadmaps revised that date to 1998. Aizcorbe, Oliner, and Sichel, (2006) have persuasively argued that the turning point was closer to mid-1990s than late in the decade.

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), pp. 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 2 years behind in 1989 (at 1000nm); 3 years behind in 1991 (800nm); 1 year behind in 1995 (350nm). Intel caught up with the DRAM makers in 1997, at 250nm, and remained on a roughly 2 year cycle through 2014. Author's calculations based on Intel (2008), IC Knowledge (2004), http://ark.intel.com.

[14] Flamm (2009); Spencer and Seidel (2004).

[15] The last (incomplete) official roadmap prepared by ITRS was released in 2012. Intel and others reportedly withdrew from ITRS around this time.



Source: Holt (2005).

Figure 2. Feature Size Scaling as Observed by Intel in 2005

Using (2), but adopting shorter two-year cycles for new technology nodes, implies rates of annual decline in transistor cost accelerating to almost 30%. In short, if the historic pattern of 2-3 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, "shrink" version of the same chip design fabricated year earlier would be expected to cost 20 to 30 percent less to manufacture, purely because of the improved manufacturing technology.

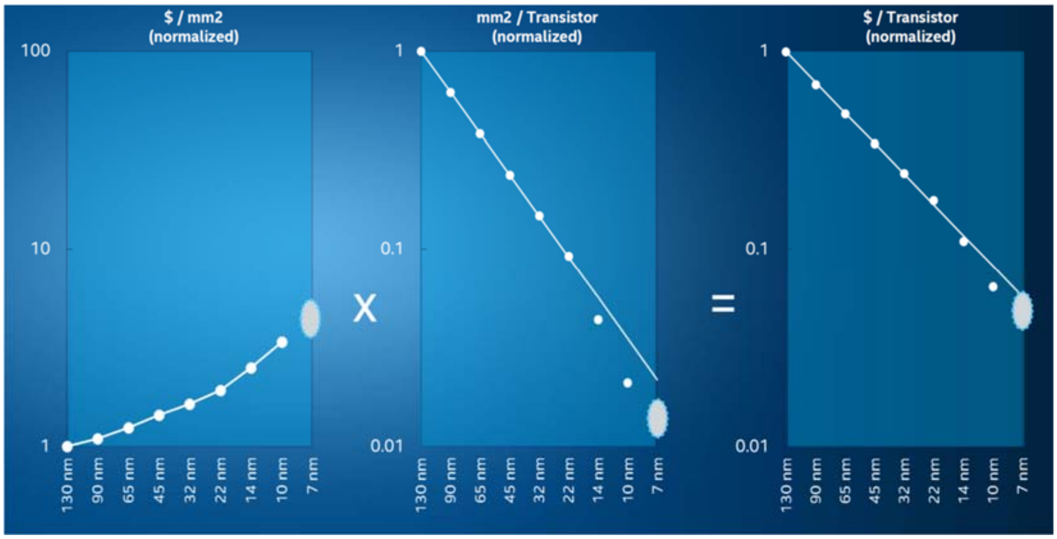
It now appears that this two-year cycle for technology nodes definitively ended in 2014, with deployment of the 14nm node. The most aggressive adopter of leading edge chip manufacturing technology, Intel, currently projects introduction of its next 10nm processor products no earlier than late 2018.[16] This means that time between introductions of new technology nodes now is approaching 4 years for Intel, a dramatic change from its two-year cadence through 2014[17]

[16] See http://wccftech.com/intel-delays-10nm-cannon-lake-cpus-end-2018/ .

[17] Intel chip manufacturing competitor TSMC was said in early 2017 to be manufacturing a "10nm" node in volume for Apple (See R. Merritt, "TSMC, Samsung Diverge at 7nm," *EE Times*, Feb. 8, 2017, (http://www.eetimes.com/document.asp?doc_id=1331324), 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

At Intel, the post-1995 two-year technology development cycle had been explicitly incorporated into marketing efforts, and dubbed the Intel “tick-tock” development model in 2007.[18] 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 formally acknowledged by Intel in its 2016 annual report.[19]

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[20]. The 2015 Intel decomposition of manufacturing cost per transistor, using equation (2), is shown as Figure 3, and in Table 1. Generally, Intel’s average silicon area per transistor did not decline by the predicted 50% between technology nodes, primarily because of the increasing complexity of interconnections in processor designs.[21] If accurate, these numbers indicate average chip area per transistor shrank by 38% at each new node from 130nm through 22nm.[22] Nor did Intel’s wafer-processing costs stay constant over the post-130nm period as a whole, since the adoption of 450mm wafers, and subsequent cost reset, never happened at 22nm, as had been predicted back in 2005. However, as long as average area per transistor declined at faster rates than processing costs per area increased, transistor cost would continue to decline. Intel’s cost per transistor estimates are revisited below.



Source: Holt (2015).

Figure 3. Intel 2015 Version of Equation (2)

technology node. See https://www.semiwiki.com/forum/f293/intel-tsmc-samsung-10nm-update-8565.html ; http://wccftech.com/intel-losing-process-lead-analysis-7nm-2022/ .

[18] See http://www.intel.com/pressroom/archive/releases/2007/20070918corp_a.htm .

[19] Intel (2016), p. 14.

[20] Intel actually produced microprocessors in volume on both 200mm (8”) and 300mm (12”) wafers using its 130nm manufacturing process technology. See Natrajan, at. al., (2002), pp. 16-17.

[21] See Flamm (2017), p. 34, for a more detailed explanation.

[22] 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.

							Compound Annual Percentage Change:			
Year Intel 1st Shipped New Product at Tech Node	**Tech Node (nm)**	**Wafer Processing Cost ($ / mm^2)**	**X**	**Transistor size (mm^2 / transistor)**	**=**	**$ Cost / Transistor**		**Wafer Processing Cost ($ / mm^2)**	**Transistor size (mm^2 / 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: Bill Holt, "Advancing Moore's Law," presentation to Intel Investor Meeting, 2015,										
Santa Clara, slide 6, graph digitized using WebPlotDigitizer. Year node introduced from ark.intel.com .										

Table 1. Decomposing Intel Transistor Cost Declines into Wafer Cost and Transistor Size Changes

In short, if the historic pattern of 2-3 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, “shrink” version of the same chip design fabricated year earlier would be expected to cost 20 to 30 percent less to manufacture, purely because of the improved manufacturing technology.

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 one 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 20-30% 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.[23] 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 20-30% annual decline in manufacturing cost associated with Moore’s Law could be interpreted as a floor on the quality-adjusted price declines that we might expect to observe in the most competitive segments of the semiconductor market.

[23] At SEMATECH, the US semiconductor industry consortium (with which the author worked as a consultant in the first decade of the 2000’s), 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.

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.[24] 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 faster microprocessors. Figure 4 shows clock rates for Intel desktop microprocessors in computers tested on industry standard benchmark programs over the last twenty years, as well as benchmark scores for these computers. As clock rates increased, so did performance.[25] Cheaper processors were also faster—stimulating increased demand for new computers in offices, homes, and workplaces.

Log (Processor Speed) *Log(Performance)*

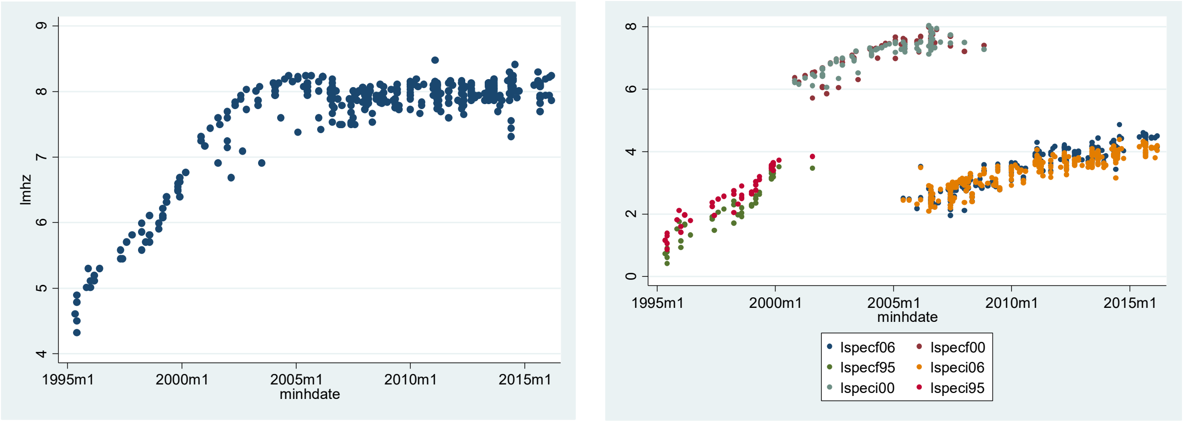


Figure 4. Processor Clock Rate and Performance for Intel Desktop Processors Running SPEC CPU Benchmarks, by First Availability Date of Tested Hardware

Source: Author's analysis of SPEC submissions, SPEC.org.

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

[25] 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.

The logarithmic scale used in Figure 4 obscures a fairly dramatic slowdown in improvement in CPU performance after the millennium. Table 2 shows compound annual growth rates in performance of Intel desktop processors on standard CPU benchmark software (the SPEC benchmarks).

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.[26]

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 which could be divided up into multiple software processes not requiring any communication or coordination between processes.[27]

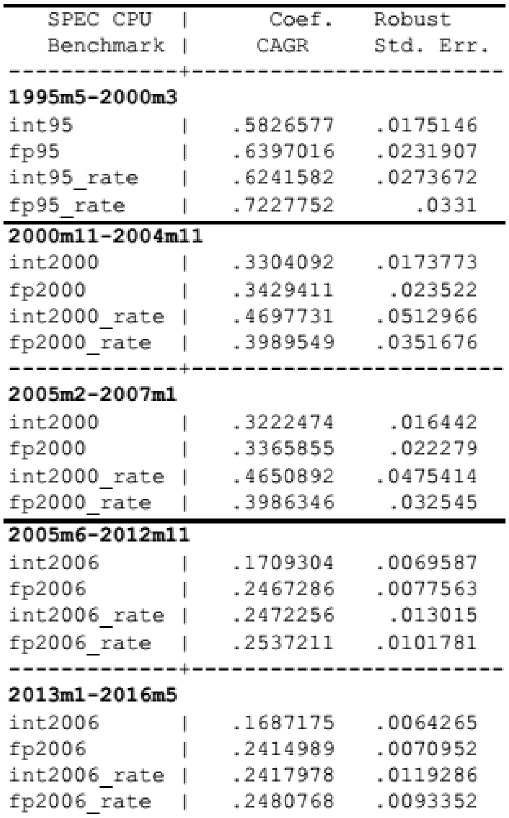
Changes in trends over time in the SPEC benchmark performance scores for Intel desktop processors are quite dramatic. Over the 1995-2000 period, integer computing performance increased by about 58 percent annually, floating point performance by 64%. The suite was revised in 2000, and from the end of 2000 through 2004, both integer and floating point performance improvement were almost halved, to an increase of about 33-34% per year.[28] 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% (integer) and 25% (floating point) annual improvement.[29]

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

[27] 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 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 multi-threading—a feature called hyperthreading by Intel), or some number of instances less than either of those bounds.

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

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



```
   SPEC CPU  |      Coef.   Robust
   Benchmark |      CAGR    Std. Err.
-------------+-----------------------
1995m5-2000m3
int95        |   .5826577   .0175146
fp95         |   .6397016   .0231907
int95_rate   |   .6241582   .0273672
fp95_rate    |   .7227752      .0331
2000m11-2004m11
int2000      |   .3304092   .0173773
fp2000       |   .3429411    .023522
int2000_rate |   .4697731   .0512966
fp2000_rate  |   .3989549   .0351676
-------------+-----------------------
2005m2-2007m1
int2000      |   .3222474    .016442
fp2000       |   .3365855    .022279
int2000_rate |   .4650892   .0475414
fp2000_rate  |   .3986346    .032545
2005m6-2012m11
int2006      |   .1709304   .0069587
fp2006       |   .2467286   .0077563
int2006_rate |   .2472256    .013015
fp2006_rate  |   .2537211   .0101781
-------------+-----------------------
2013m1-2016m5
int2006      |   .1687175   .0064265
fp2006       |   .2414989   .0070952
int2006_rate |   .2417978   .0119286
fp2006_rate  |   .2480768   .0093352
```

Table 2. Annual Growth in Processor Performance Improvement Over Different Time Periods and Benchmarks
Source: Author analysis of SPEC benchmark performance of Intel desktop processors.

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 2003-2004, higher clock rates stalled (see Figure 4), as disproportionately greater power was required to run processors reliably at ever higher frequencies. With tinier transistors running at 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.[30]) Intel and others abandoned architectures reliant on frequency scaling to achieve 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

[30] Raley (2015), p. 23.

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.

The difficulty and cost of parallelization of software is an economic factor limiting utilization of cheap multicore CPUs on hard-to-parallelize applications.[31] 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.”[32]

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 5 shows contract chipmaker GlobalFoundries’ 2015 transistor manufacturing costs at recent technology nodes.[33]

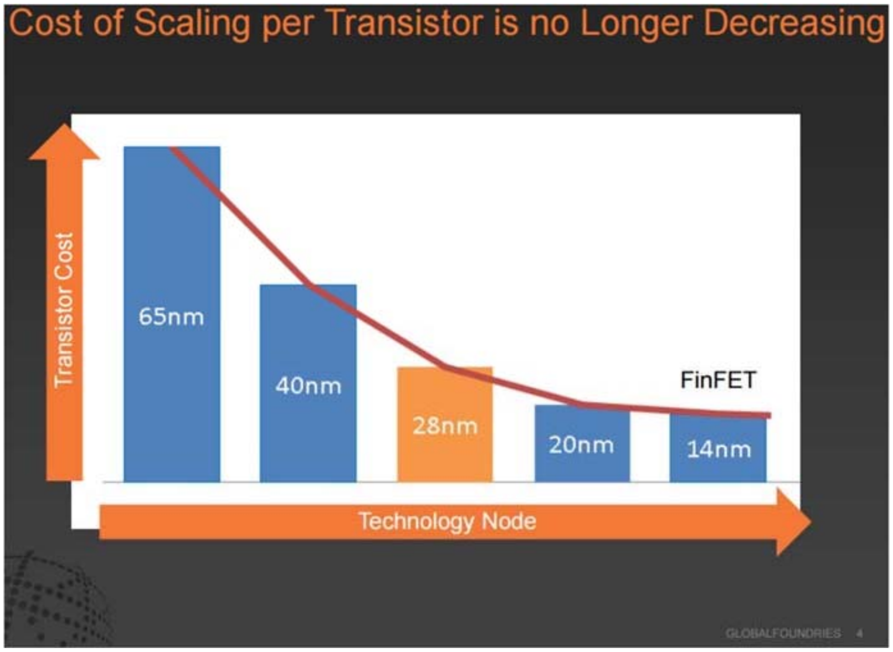


Figure 5. Global Foundries’ Transistor Manufacturing Cost at Recent Technology Nodes
Source: McCann (2015).

[31] The opposite--software problems easily divided up across processors and run with little or no inter-processor communication or management required—are described in the computer engineering literature as “embarrassingly parallel”.
[32] Hennessey and Patterson (2012), p. 411.
[33] Like Table 1, this figure probably does not include R&D costs.

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.[34] (Fabless companies accounted for 25% of world semiconductor sales in 2015; foundries, which also build outsourced designs for semiconductor companies with fabs, had a 32% share of global production capacity.[35]) Charts like Figure 6, showing increased costs at sub-28nm technology nodes, were frequently published between 2012 and 2016. Figure 6 is not inconsistent with Figure 5, since Figure 6 likely includes the fabless customer's non-recurring fixed costs for designing a chip and making a set of photolithographic masks used in fabrication, while Figure 5—the foundry's processing costs—would not.[36] These fixed costs have grown exponentially at recent technology nodes and create enormous economies of scale.[37] Some foundries have publicly acknowledged that recent technology nodes now deliver higher density or performance at the expense of higher cost per transistor.[38]

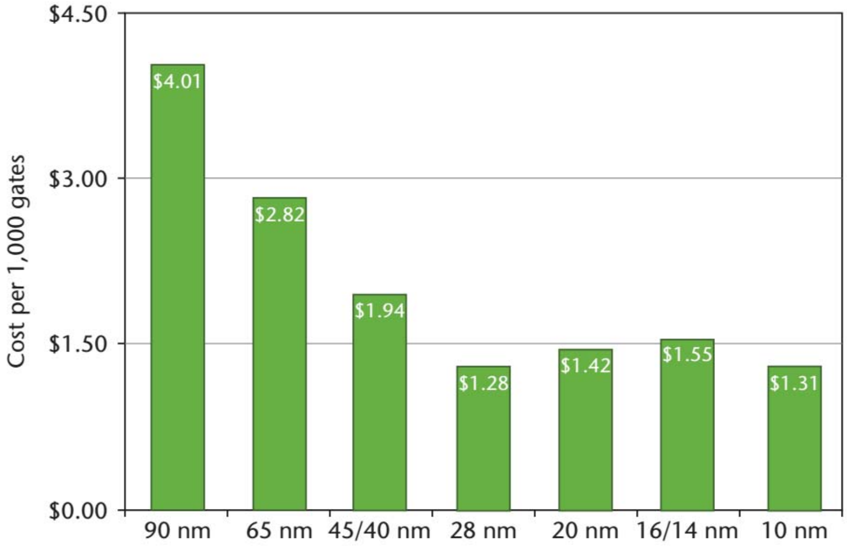


Figure 6. Cost per logic gate, with projection for 10nm technology node

Source: Jones (2015)

[34] 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).

[35] Foundry share calculations based on Yinug (2016), Rosso (2016), IC Insights (2016). Charts like Figure 4 should be viewed cautiously, as underlying assumptions about products, volumes, and costs are rarely spelled out in published sources.

[36] 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.

[37] 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.

[38] Samsung's director of foundry marketing: "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).

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 will prevent 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, has been projected to grow as large as 50% by 2018.[39] 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.

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."[40]

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.

[39] Esmaeilzadeh, et. al. (2013), pp. 93-4.
[40] Bourzac, (2016).

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.[41]

Table 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.[42] This data has no longer been available in recent years.

	Compound Annual Decline Rate					
	Flamm-Aizcorbe DRAM Composite	BoK $EPI DRAM	BoK $EPI Flash	BoK DRAM PPI	BoK Flash PPI	BoJ Chain-Wtd MOS Mem PPI
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

Table 3. Price Indexes For Memory Chips

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.[43] The Bank of Korea's export price index (based on dollar basis contracts) and the Bank of Korea's producer price index (PPI, converted to a dollar basis using quarterly average exchange rates) for DRAM and flash memory chips are available.[44]

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

[41] See http://www.icinsights.com/news/bulletins/Total-Flash-Memory-Market-Will-Surpass-DRAM-For-First-Time-In-2012/ .

[42] The data prior to 1990 is the same data used in Flamm (1995), Figure 5-2. From 1990 on, the data are taken from Aizcorbe (2002).

[43] 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 3 producers (Samsung, Hynix, and Micron) accounted for between 80 and 90% of global sales. See Competition Commission of Singapore (2013).

[44] 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).

memory, tilting more toward flash in recent years. Generally, except for the period from 1985-1995, when a string of trade disputes (between the US and Europe, and Japanese, Korean, and Taiwanese memory chip producers) had significant impacts on global chip prices,[45] prices for DRAMs and flash fell at average rates exceeding 20-30% 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 (3-D NAND) 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.[46] 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.[47]

Another "commodity-like" price in the semiconductor industry in recent years has been the cost that chip design houses face in having their chips manufactured on their behalf at so-called "foundries". The outsourced manufacturing of semiconductors designed at "fabless" semiconductor companies at foundries accounted for about 25% of world semiconductor sales in 2015. Foundries, which also build outsourced designs for semiconductor companies with fabs, held 32% of global production capacity in that year.[48]

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 (2016) portrays a slowing decline in fabricated wafer prices prior to 2012. (See Table 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.

[45] See Flamm (1995).
[46] Micron 2015 Winter Analyst Conference (2015).
[47] Micron's Raymond James Institutional Investor Conference (2016); Micron Analyst Conference (February, 2017).
[48] Foundry share calculations based on Yinug (2016), Rosso (2016), IC Insights (2016).

	Annual Index	% Rate of Change
2004	100	
2005	83.89521	-16.1048
2006	74.75891	-10.8901
2007	65.93704	-11.8004
2008	57.89118	-12.2023
2009	52.95437	-8.52774
2010	48.67003	-8.09062

Table 4. A Quality-Adjusted Price Index for Fabricated "Foundry" Wafers

Source: Byrne, Kovak, and Michaels (2016).

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, 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 U.S. Producer Price Indexes for microprocessors show annual (January-to-January) changes in microprocessor prices steadily falling from 60-70 percent peak rates during the "golden age" of the late 1990s and early 2000s, to a low of about one 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 rates of 40%, in the late 1990s, to rates mainly in the 10-20% range in the last several years.

Table 5 shows compound annual decline rates in the PPI for microprocessors (including microcontrollers) as constructed by BLS, along with similarly defined indexes for the commodity "microprocessors". Annual decline rates slow from a rate near 50% in the late 1990s and first half decade of the new millennium, to a little over 10% in the second half of that first decade, to about 3% 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 are constructed). It is believed that these are matched model indexes based on some weighted selection of products appearing on Intel list price sheets (the same data source I utilize below),[49] 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.[50]

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

[50] The BLS web site shows 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 on a monthly basis from September 2009 through 2015. There are also two discontinued microprocessor commodity price indexes, one based in December 2004, and

		Microprocessors (including microcontrollers)			
		Commodity Price		Producer Price	
		Index (discont)	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	

Table 5. Annualized Decline Rates for Microprocessors per the BLS
Author's calculation. Middle month for quarter used, except Dec. 2007 used for 2007:4.

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 data set I have collected over the last two decades. Since the mid-1990s, Intel has periodically published, or posted on the web, current list prices for its microprocessor product line, in 1000-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 are typically updated every 4 to 8 weeks—though price updates have sometimes come at much shorter or longer intervals.[51] By combining these detailed prices with detailed attributes of different processor models, it is possible to construct a very rich data set relating processor prices to processor characteristics, over time.

This permits one to construct 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.

My 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—since mid-2004, Intel has 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 with more recent product introductions.

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.

[51] My data initially (over the 1995-1998 period) made use of compilations of this data collected by others and posted on the web; since 1998-99, most of this data was collected and archived directly off the Intel web site.

Price Indexes for Intel Desktop Processors

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.[52] A price index based on the simple geometric mean of individual product price relatives (sometimes called a Jevons price index), is chained across pairs of adjoining time periods, and depicted in Figure 7. It has the same qualitative behavior as the official government producer price index for microprocessors, falling at rates exceeding 60% in the late 1990s, and slowing to a decline rate under 10% since 2009.

This geometric mean matched model index actually falls a little more slowly than the official PPI in recent years, 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 geometric mean matched model 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.

The adjoining pairs of periods over which this regression was run were chosen to overlap. The time dummy variables in the above regression were used to construct an index of adjoining period price levels; the overlapping time period was used to link these period-to-period (on average, roughly 8-9 monthly periods per year with reported list prices) indexes into a longer chained price index. Note that typical power consumption for a processor (TDP, thermal design power) was generally unavailable for Intel processors released prior to late 1998. 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 available, and therefore the hedonic price index including TDP is the preferred index.

Figure 7 shows the price indexes produced using the above methods. The slowing of declines in price in 2004 and 2005 is quite apparent, followed by a temporary resumption of a somewhat faster rate of decline after 2006, followed by a marked and much more extreme slowdown after 2009.

The first four columns in Table 6 compare my estimated hedonic and matched model price indexes and the BLS PPIs. As expected, matched model index price declines are often close, but generally decline more slowly than those measured by the hedonic price index based on the same data. My estimates over comparable time periods are quite similar to the matched model index results of Aizcorbe, Corrado, and Doms, and to the producer price indexes. Prior to 2004, my geometric mean matched model and the PPI move quite closely, with my hedonic indexes showing a modestly higher rate of decline, as expected. From 2004 through 2006, both my geomean and hedonic price indexes decline much more slowly than the PPIs, and from 2006 through 2009 my geomean falls at about the same rate as the PPI, while my hedonic index declines more rapidly. From 2009 to 2010 both my geomean and hedonic fall more slowly than the PPI. Finally, from 2010 through 2014, both my geomean

[52] 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 prices sheets on a monthly basis, “adjoining time periods” means temporally adjacent observations.

and hedonic indexes again fall more slowly than the PPI, but all three sets of declines are in the low single digits.

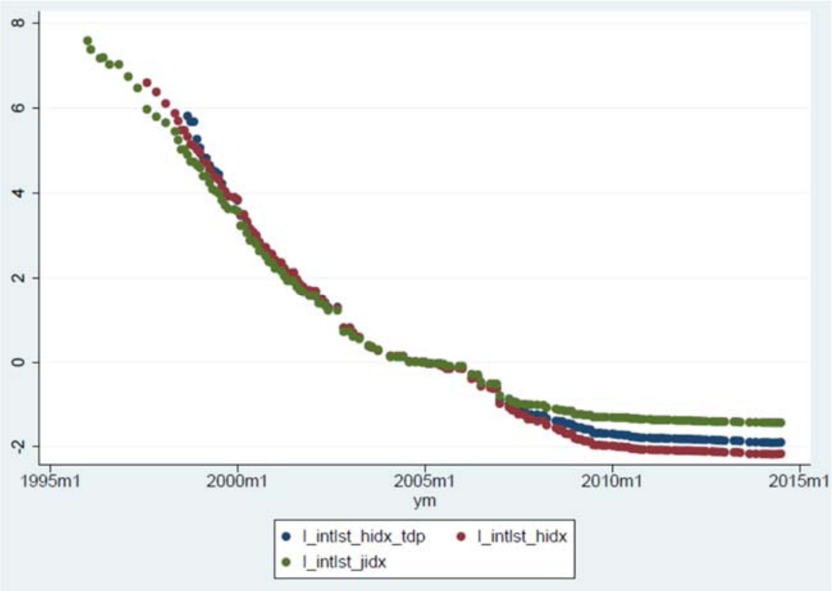


Figure 7. Geomean Matched Model and Hedonic Price Indexes for Intel Desktop Processors
Green: Geometric Mean Matched Model Index; Blue: Hedonic Index with Thermal Design Power (TDP) as included characteristic; Brown: Hedonic Index without TDP as included characteristic.

Table 6

Annualized Compound Rates of Change in Microprocessor Price Indexes

		Compound Annualized Decline Rate						
		Intel Tray Price				**Producer Price**		**Retail**
		Hedonic, no TDP	**Hedonic with TDP**	GeoMean Matched Mocel		Micropro cessor PPI		GeoMean Matched Model
1998m9-2001m10		-68.3%	**-73.0%**	-65.0%		-57.5%		
2001m10-2004m2		-50.5%	**-50.1%**	-48.2%		-46.6%		-34.0%
2004m2-2006m1		-14.4%	**-13.8%**	-10.7%		-25.2%		-11.1%
2006m1-2009m1		-42.1%	**-36.9%**	-31.5%		-29.0%		-24.2%
2009m1-2010m11		-13.7%	**-13.6%**	-6.2%		-22.7%		-11.3%
2010m11-2014m7		-2.7%	**-2.9%**	-2.2%		-3.7%		

Source: Author's dataset and calculations, except Microprocessor PPI, from BLS.

I have also constructed a geometric mean, chained monthly price index based on retail prices for processors, using data from a commercial web site that reported the lowest price for a particular processor model across a selection of internet-based retailers, over the period from 2001 through 2010. These prices are actually a relatively small subset of the much larger set of list prices for all Intel processors, and presumably represent the models that were most popular in the retail marketplace. The final column of Table 6 reports changes in this retail price index for equivalent time periods. Generally, the pattern over time is similar (steepest declines over 2001-2004 and 2006-2009, slower declines over 2004-2006 and 2009-2010).

To summarize these results, then, though there are substantial differences in the magnitude of declines across different time periods, data sources, all of the various types of price indexes constructed concur in showing substantially higher rates of decline in microprocessor price prior to 2004, a stop-and-start pattern after 2004, and a dramatically lower rate of decline since 2010.

Taken at face value, this creates a new puzzle. Even if the rate of innovation had slowed in general for microprocessors, if the underlying innovation in semiconductor manufacturing technology has 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 rates of decline in processor price that are being measured now 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 in semiconductor manufacturing has slowed markedly.

Evidence on Manufacturing Costs. Finally, 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 7.) These figures presented by Intel at its 2012 Investor Meeting seem to show accelerating cost declines in the late 1990s, 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 a larger wafer size after the 130nm technology node was accompanied by a particularly large reduction in transistor cost in the next node, using the larger size wafers.

		Transistor Cost Index, 90nm = 100		**Percent Transistor Cost Decline Rate**		**Compound Annual Decline Rate**	
		Otellini, 2012		Otellini, 2012		Otellini, 2012	
		Wafer Size		Wafer Size		Wafer Size	
Intro Date	Tech Node	200mm	300mm	200mm	300mm	200mm	300mm
1995q2	350	1575.35					
1997q3	250	1033.14		-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
Intro dates: 130nm and up from http://www.intel.com/pressroom/kits/quickreffam.htm							
	< 130nm from ark.intel.com						

Table 7. Annualized Decline Rates for Intel Transistor Manufacturing Cost, 2012
Source: Otellini (2012), digitized using WebPlotDigitizer.

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."[53] 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 plant 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.

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.[54] This, again, is consistent with a slower rate of innovation.

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 has come from Intel. Recent Intel statements about its manufacturing costs have been the primary factual evidence within the semiconductor manufacturing community against the proposition that Moore's Law is ending. Unfortunately, Intel has not been consistent in the data it has presented publicly on this issue.

The problem is illustrated by Figure 8 and Table 8, which place side by side two exhibits on manufacturing costs per transistor that Intel has 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 2). Some version of the left pane in Figure 8 has been the primary factual evidence in Intel assertions that Moore's Law continues at its historical pace. The graphics in Figure 8 have been digitized[55] and recorded in Table 8, then rebased to 100 at the 90nm technology node. Compound annual decline rates have been calculated in this table using fine-grained quarterly introduction dates for the first processors manufactured by Intel at that technology node.

[53] Li and Hall (2015), p. 13.

[54] Krzanich (2016).

[55] Using http://arohatgi.info/WebPlotDigitizer/.

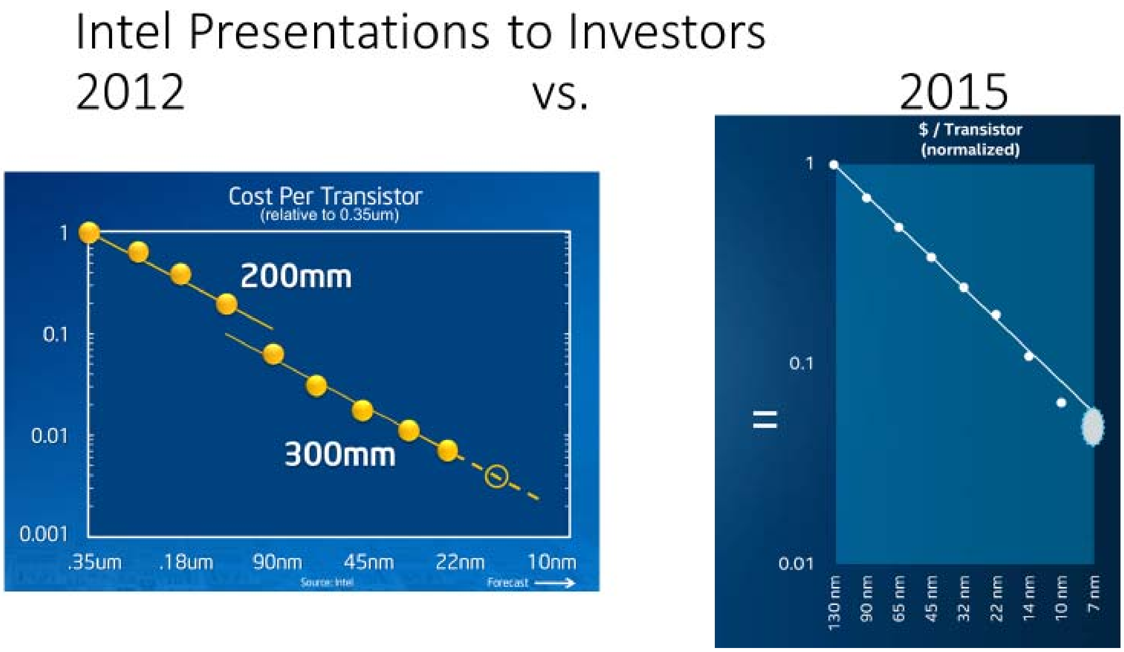


Figure 8 Intel Transistor Manufacturing Costs, 2012 vs. 2015 Versions

Source: Otellini (2012): Holt (2015).

		Transistor Cost Index, 90nm = 100			**Percent Transistor Cost Decline Rate Between Nodes**			**Compound Annual Decline Rate**		
		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	1575.35								
1997q3	250	1033.14			-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?	10			9.46			-41.4			-21.1
Intro dates: 130nm and up from http://www.intel.com/pressroom/kits/quickreffam.htm										
	< 130nm from ark.intel.com									

Table 8. Comparison of Intel Cost per Transistor at Various Technology Nodes, 2015 vs. 2012

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 transistors costs declining at 12-18% annual rates after the millennium, rather than the 30% 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 the restatement by Intel 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 average actual transistors fabricated per die area), Intel apparently began using an entirely theoretical measure of area per designed transistor that may not 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.[56] (For die designs released prior to 2010, Intel had previously reported both actual die size, and the number of transistors processed on the die, for many of its chip models.)

An Intel Exception? Interpreting the recent economic history of Moore's Law, how can Intel's most recent description of accelerating declines in manufacturing cost per transistor 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 by 2016.[57] (Intel overlooked GlobalFoundries, which by acquiring IBM's semiconductor business in 2015, substantially increased its scale.)[58] Both TSMC and

[56] See Flamm(2017), p. 34, for a brief explanation of this issue. Intel's latest redefinition of its publicly disclosed "transistor density metric" is entirely theoretical: .6 x (transistors in a NAND logic cell/area of a NAND logic gate) + .4 x (transistors in a complex scan logic flip-flop cell/area of complex scan logic flip-flop cell) = # transistors/mm2. 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. See Mark Bohr, "Moore's Law Leadership," March 2017, available at https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Mark-Bohr-2017-Moores-Law.pdf .

[57] Krzanich (2012), slide 19.

[58] 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.

GlobalFoundries are “pure” foundries, and achieve their volumes entirely by aggregating the demands of external chip design customers.

Many U.S.-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” U.S. foundries (e.g., TowerJazz, On Semiconductor) operate mature foundry capacity that remains 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.[59] In Asia, there are arguably only Toshiba in Japan, Samsung and Hynix in Korea, and foundry TSMC in Taiwan. Firm level scale economies explain why fewer firms can afford leading edge fabs, but can’t explain why Intel’s cost per transistor would have declined much faster than 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.[60] At 14nm, (updating wafer production costs using Intel costs in Table 1 implies 150% 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 5000 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.[61]

Intel, with the largest production runs in the industry (perhaps 300 to 400 million processors in 2014[62]), 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

[59] 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.

[60] Both 130 nm 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.

[61] Lattard (2014), slide 6.

[62] 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), p. 1.

chip manufacture, and manufacturing process parameters can even be altered to shift the mix of functioning parts in desired ways.[63]

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 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.

Scale-driven cost disadvantages are increasingly pushing low volume chip production to older, 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% by 2020![64]

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 what make it economically attractive for fabricating 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.

Hedonic Price Indexes for Microprocessors. The second major piece of evidence arguing against a slowdown in Moore's Law is a study by Byrne, Oliner, and Sichel (BOS, 2015), which also is focused on data from Intel for its argument The BOS study puts forward an alternative explanation for the recent behavior of the official price indexes, arguing that the Intel posted list prices that are being used by all analysts of microprocessor pricing trends are not in fact representative prices, and raise the possibility that the post-2004 slowdown is a spurious artifact of changes in Intel pricing practices.[65] Their argument is that "[b]y 2006, the company had moved to a business model that featured more active management of its product offerings below the [technological] frontier…by setting list prices that were relatively

[63] 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, etc., 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.

[64] Dieseldorff (2016).

[65] D.M. Byrne, S.D. Oliner, and D.E. Sichel, "How fast are semiconductor prices falling,," AEI Economic Policy Working Paper 2014-06, revised 2015, available at www.aei.org/publication/how-fast-are-semiconductor-prices-falling/ .

stable over a chip's life cycle, Intel may have been attempting to extract more revenue from less price-sensitive buyers while offering discounts on a case-by-case basis."[66] Arguing that new products get little discount from the posted list price, while older products are heavily discounted from list, they argue that a hedonic price index based only on newly introduced products is the correct measure of quality-adjusted price trends for Intel microprocessors. DIscarding most of their sample of Intel products, and keeping only 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 2000-08 as in 2008-12, at about a 40 percent annual rate of decline.[67] This is vastly higher than any of the rates shown in Table 6 for the equivalent time periods.

The BOS observation that Intel seems to have changed its advertised list prices much less frequently after 2006 than before 2006 certainly seems true, based on the public Intel price list data. But they also assert that actual transaction prices for recently introduced chips are not significantly discounted from list, while transaction prices for older chips after 2006 are heavily discounted, with a discount that increases with age. Unfortunately, this behavior is both unobservable and untestable, since no data on Intel transaction prices for its wholesale sales to large buyers are publicly available. Indeed, evidence produced in the AMD-Intel antitrust investigation seems to show that even new chips sold to large customers were heavily discounted from list prices prior to 2006, at times with conditional rebates that were not publicly reported by Intel or its customers.[68]

An alternative hypothesis to the one put forth by BOS is that Intel's diminished propensity to alter its list prices in fact reflects its ***actual*** pricing behavior. Figure 9 shows the fraction of incumbent (i.e., omitting newly introduced products) desktop processor prices that changed from one list price sheet to the next one issued. It is evident that while its propensity to alter list prices on existing processors diminished over time, Intel never stopped changing list prices after introduction of a new processor. Further, there clearly was no sharp dividing line between its behavior before and after 2006. In 2008 and 2009, for example, there were price sheets on which anywhere from 35 to 40 percent of already introduced desktop processor prices changed from the previous sheet.

[66] Ibid., pp. 8.

[67] Ibid, Table 7, p. 34. Note that, with very much smaller sample sizes, the researchers use only two processor characteristics—performance on a single software benchmark, and power draw—in their hedonic regression.

[68] See European Commission, "Non-confidential Version of the Commission Decision of 13 May 2008, COMP/37.990 Intel," available at
http://ec.europa.eu/competition/antitrust/cases/dec_docs/37990/37990_3581_18.pdf .

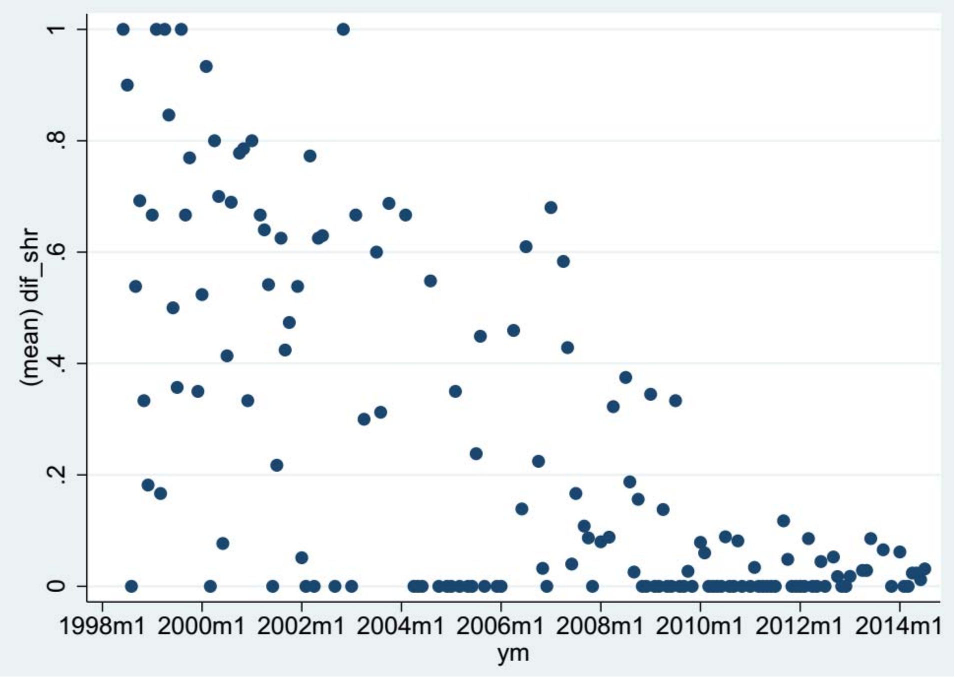


Figure 9. Fraction of Intel Desktop Processor Prices Changing From One Price List to the Next.
Source: Author's tabulation from Intel list price dataset.

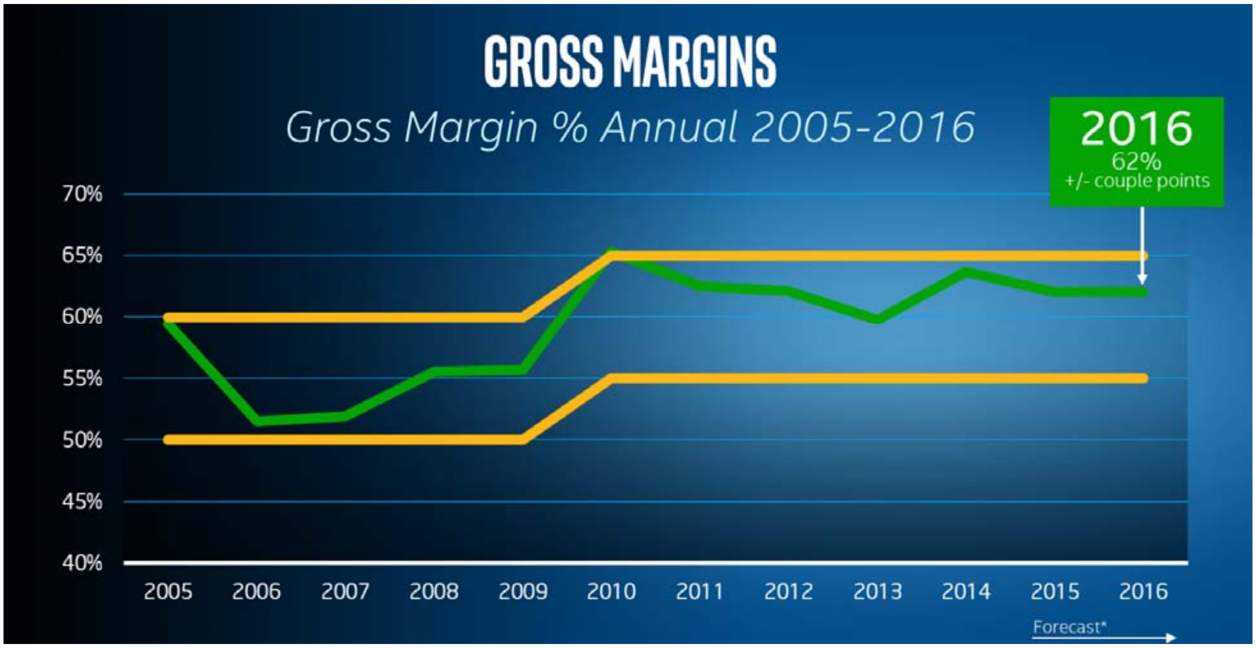


Figure 10. Intel's Post-2010 Gross Margin Elevation Objective
Source: Smith (2015).

Indeed, if one had to choose a date based on this chart for a climacteric in Intel pricing practices, 2010 would be as good a choice as any other. That year does indeed seem to coincide with 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 10.) Raising its average sales prices was a key element of this strategy (See Figure 11.)

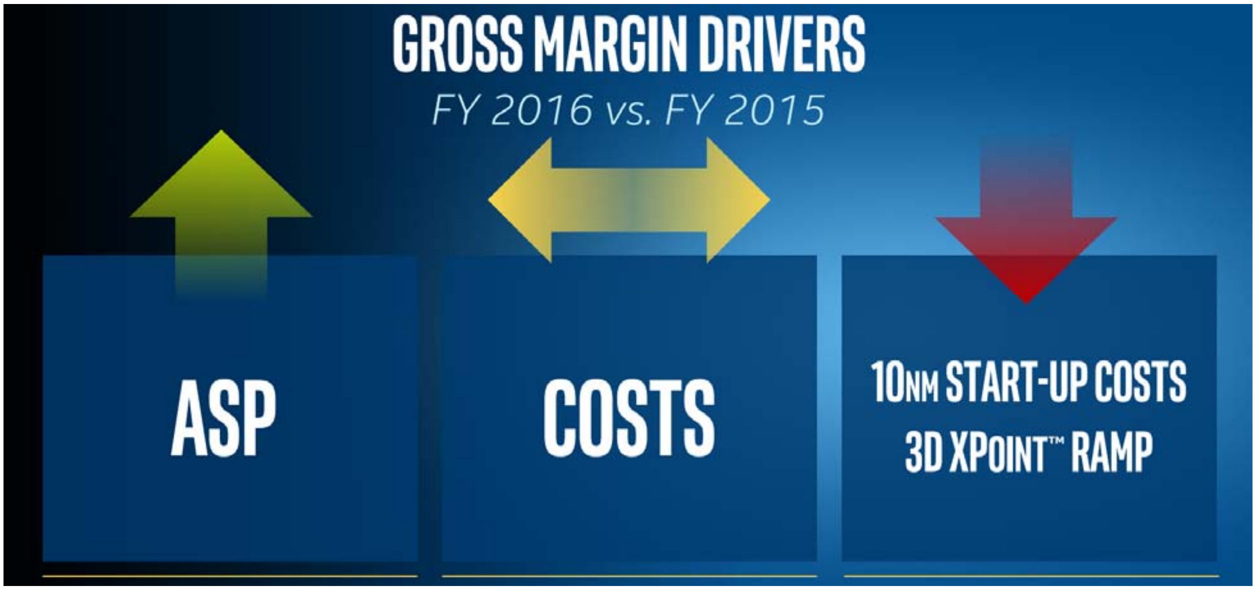


Figure 11. Intel's 2015 Explanation to Its Shareholders for Success in Maintaining High Profit Margins
Smith (2015).

Finally, there is one source of processor price data that is real, observed, and does not require maintaining 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% of Intel processors in recent years, by volume, were sold directly as boxed processors, primarily to small computer makers and electronic retailers.[69]) Computer original equipment manufacturers (OEMs), electronics system manufacturers, and electronic parts distributors also can purchase processors directly from Intel, and 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 box and OEM packaged processors are sold by retailers and brokers, and have the great virtue of having a price that is advertised publicly and directly observable in the marketplace. (The retail data use in constructing my matched model price index include both OEM and retail packaged chips sold by

[69] "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." B.Wieland, P. Mastrantonio, S. P. Willems, and K. G. Kempf, "Optimizing Inventory Levels Within Intel's Channel Supply Demand Operations," *Interfaces*, Vol. 42, No. 6, Nov–Dec 2012, pp. 517–18.

internet retailers.) The retail data used in Table 6 seem to clearly point to a deceleration in microprocessor price declines after 2004.

It is reasonable to presume that retail transaction prices (which are at least observable in the market), in the long run, should have some stable stochastic relationship to wholesale producer transactional prices. Indeed, previous studies have estimated such linkages between OEM contract prices and retail prices for high volume chips sold in the semiconductor industry.[70]

Both 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 inventories of chips are absorbed in those markets. 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 ins 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.

BOS hypothesize that a systematic change in the relationship between Intel list prices and unobserved OEM (large buyer) contract prices occurred after 2006. If true, we would then also expect to see a change in the stochastic relationship between observed 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 a total 1580 monthly observations on average retail and posted list price covering 163 distinct Intel desktop processor models sold by Internet retailers over the years 2000 through 2010.[71] The fixed effects regression model (which permits a particular low-end Celeron model, for example, to be related to Intel list price with a different retail margin than a high end Core i7 model) that I estimated specified that the log of retail price for model i in month t was given by

$$(3)\ \ln(R_{it}) = a_i + b\ \ln(I_{it}) + c\ Age_{it} + d\ OEM_{it} + After2006 + e\ After2006 \times \ln(I_{it})$$
$$+ f\ After2006 \times Age_{it} + u_{it}\ ,$$

with R_{it} an observation on average retail price for model i in month t; I_{it} the average posted Intel list price in a month in which list price had been posted at least once; Age_{it} the number of elapsed months since the month the model's price had been first posted on a published Intel price sheet; After2006 a binary

[70] See Flamm (1993), for a study documenting linkages between retail prices and OEM contract prices for DRAM memory chips.
[71] My retail price data actually end in January 2011.

indicator variable with value 1 in 2006 and thereafter, zero before; OEM a binary indicator for whether the product sold was the retail boxed version, or the bare chip in OEM packaging; and u_{it} a random disturbance term. If the Byrne, Sichel, and Oliner assumption is correct, and post-2006 transaction prices contain age discounts from Intel list price that pre-2006 prices did not, we would expect to find a statistically significant shift coefficient on the interaction of After2006 and Age.

Table 9 shows the results of estimating this model.[72] The After2006 shift variable, and all of its interactions, including the interaction with processor model Age, are close to zero and statistically insignificant individually, and jointly.[73] The relatively flat Intel list prices after 2006 are mirrored in the behavior of retail prices for the same chips.

Interestingly, there does seem to be small but statistically significant age effect, with retail price declining by about .58 percent for every additional month after the product is first sold by Intel. But this relationship holds throughout the 2000-2010 period, and we cannot reject the hypothesis that there was no change in 2006 and after. The model also suggests that on average, products originally sold unboxed to OEMs were resold by retailers in OEM packaging at a 5 percent discount from the equivalent retail boxed product. A point estimate of the elasticity of retail price with respect to a decline in Intel list price was about -.77, i.e., a ten percent decline in list price was associated with about a 7.7% decline in retail price.[74]

Based on the only evidence on actual transaction prices that is publicly available, i.e., advertised retail prices from Internet-based vendors, then, we find no evidence to support the suggestion that there was some structural change after 2006 in the relationship between observed Intel list price 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 argues against the assumption that this must have been the case.

SPEC scores vs. chip characteristics. It has recently become clear that the BOS results of no slowdown are driven primarily by their use of SPEC benchmark scores as a substitute for a more extensive set of chip characteristics in a hedonic price equation.[75] The plausibility of exclusion of chip characteristics other than SPEC scores from the hedonic price equation is simply maintained as an assumption, and never actually tested econometrically by BOS. BOS' substitution of SPEC scores for actual chip characteristics is based on the argument that direct performance measures are easier to get right than relevant chip characteristics. But this overlooks three fundamental reasons underlying chip characteristics are superior choices for a hedonic price equation.

[72] Robust standard errors clustered on processor model are shown in Figure 8.

[73] The Wald F(3,162) test statistic for the joint hypothesis that all After2006 terms were zero was .82, the p-value .49.

[74] Very similar results are produced if a model that is linear in price, rather than the logarithm of price, is used.

[75] Sawyer and So (2017).

Table 9
Fixed Effects Model of Log Retail Price For Intel Desktop Processors

```
               (Full Model)   (Constrained Model)

                   lp_ret          lp_ret

--------------------------------------------
lp_tray              0.763***        0.768***
[log Intel        (15.37)         (17.93)
Tray Price]
oem               -0.0497***      -0.0496***
                  (-6.70)         (-6.77)

age              -0.00676***     -0.00582***
                  (-3.70)         (-4.91)

1.aft2006          0.0204
                   (0.13)

1.aft2006#age     0.00162
                   (0.83)

1.aft2006#lp_tray -0.0108
                  (-0.39)

_cons                1.347***        1.303***
                   (4.87)          (5.55)
--------------------------------------------
N                    1580            1580
R-sq                0.987           0.987
adj. R-sq           0.986           0.986
--------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

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 simple nonlinear function of a small set of chip characteristics, including numbers of cores and threads, a set of dummy variables for 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.[76] 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 show up there either because either they affect user demand, or they affect supplier marginal cost, or they affect both demand and cost.[77] At best, SPEC scores might correctly serve as a summary measure of quality on the demand side. But there is no reason, technological or economic, why a measure of chip performance relevant to demand should be perfectly collinear with chip cost. Omitting processor

[76] Hennessey and Patterson(2003), in the Third Edition of their classic computer architecture textbook, pp. 59-60, do exactly this to compare the Pentium III with a Pentium 4 operating at the same clock rate.

[77] Pakes (2003), 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 both cost and demand functions.

characteristics relevant to chip cost will induce omitted variable bias in the hedonic coefficient estimates if the omitted characteristics are correlated (but not perfectly collinear) with included variables.

That is, assume for the sake of argument that the mix of user demand for various types of computer applications was fixed over time, and that processor performance on this fixed mix of computer applications was correctly captured in a fixed weight mean of various SPEC benchmarks. Even with the heroic assumption that this index of SPEC benchmarks correctly captured everything relevant to chip quality on the demand side (and it is clear it does not[78]), there is no plausible technological or economic reason why variations across chip models in production costs should perfectly mirror 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, non-unique combinations of numbers of cores, threads, cache memory, clock frequency, etc. But variation in each of these characteristics—cores, threads, on-chip memory, and clock frequencies—may have very different impacts on production cost for the processor than it does on SPEC scores.

Third, if benchmark scores are determined by chip characteristics, using chip characteristics directly in the hedonic equation—instead of a single benchmark score —effectively allows coefficients in the hedonic equation to change if the mix of tasks run by computer users changes 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.[79]

For all these reasons, use of the SPEC score as the sole characteristic in a hedonic price equation should be viewed as a highly implausible economic assumption. Recent work by economists at the Bureau of Labor Statistics confirms that this assumption is rejected when tested statistically. After reproducing the BOS results qualitatively in a similar (though not identical) sample, Sawyer and So (2017) show that standard statistical tests decisively reject the exclusion of processor characteristics from a hedonic price equation which also includes SPEC scores.[80] When other processor characteristics are not excluded, estimates of decline rates for quality-adjusted processor prices over time are dramatically smaller than those estimated by BOS, consistent with much lower annual decline rates in recent years. At a minimum, processor prices seem to be declining at a significantly slower pace now than in earlier epochs.

[78] Since power 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.

[79] 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 all users have a demand to run b1 applications 50% of the time, b2 applications the other 50%. Then we can represent performance on the “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% and 75% on the new “market workload,” workload performance will be incorrectly captured by the original performance index (50% weights) even if scaled by some arbitrary coefficient. However, performance on “market workload” is still correctly captured by a linear function of the two underlying chip characteristics (though coefficients of characteristics in this function change). The specification in the underlying characteristics is simply more flexible.

[80] Sawyer and So (2017), pp.

Finally, because SPEC scores are only available for a reduced 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. With Intel list price data, the number of Intel desktop processors with SPEC scores available for analysis is a fraction of all Intel desktop processor list prices in any time period. With publicly available retail or distributor processor price data, many additional desktop processor models not used in server applications may get dropped. [81]

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 essentially no additional information. 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 some particular choice of benchmark scores.

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 economists agree should be an important determinant of chip demand-- and chip characteristics, BOS have a 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: clock ticks per instruction and the seconds per clock tick in the processor's clock.[82] 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

Performance ~ IPC x 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 estimate the impact of raising clock rate on processor performance.[83]

[81]. Sample size may be further diminished if retail prices rather than Intel list prices were used to measure prices, because the selection of processors commonly sold to consumers for use in desktops may be an even smaller subset.

[82] See Hennessey and Patterson (2012), section 1.9, pp. 48-52.

[83] Hennessey and Patterson(2003), in the Third Edition of their classic computer architecture textbook, pp. 59-60, do exactly this to compare Pentium III performance with a Pentium 4 operating at the same clock rate.

IPC will depend on both the design (architecture) of the computer processor and the particular mix of instructions being executed in the computer. The specified clock rate of a processor model is typically set after testing it at the end of the manufacturing process. Random variation in a highly complex semiconductor manufacturing process leads to a distribution of functional chips by the 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. "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 non-defective, more valuable fast processors, and less valuable slow processors, on a fabricated wafer containing hundreds or thousands of these processors, determines manufacturing costs.

Speed is not the only chip processor characteristic that is 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 which require less power to perform correctly may be identified through testing, and sold as low power models of the processor.[84] Microprocessor chips generally have on-chip caches of fast local memory which can also affect the execution time for given software. The portion of on-chip cache memory which 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 can 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.[85]

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 been given the codename "Coffee Lake".[86]) 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 the many dozens of microprocessor models for each of its microarchitectures were based on somewhere between one and three basic die designs.[87] That is, the dozens of different processor models corresponding to a

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

[85] The design of a chip will segment the circuitry into functional blocks which 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 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.

[86] https://gizmodo.com/intels-latest-coffee-lake-processors-are-fast-as-hell-1819129322 .

[87] 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

single microarchitecture product family were manufactured from just one to three basic chip designs on a wafer.

It is straightforward to analyze the relationship between SPEC scores and microprocessor characteristics. Table 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, log of a highest clock rate achievable by a single core on the chip (which will differ from the base processor clock rate if a "turbo" feature is enabled on the chip, log of on-chip memory cache size,[88] log of number of physical processor cores on the chip, and log of additional multithreaded "virtual" logical cores available, if any, on a chip.[89] 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.[90]

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 96 to 98 percent of the variation in SPEC2006 benchmark scores for thousands of computer models using Intel x86 processors over the 2005-2017 period. (Table 10.) 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 a chip 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 to 4 percent of observed variation in SPEC scores. This analysis utilizes individual tested computer system data; i.e., on average there are 4 to 5 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 BOS procedure for deriving a single SPEC score for each chip model. Using that as the basis for our SPEC2006 performance regression model, we get an even higher R2, of about .99.[91] (Table 11.) It is clear that computer architecture dummies and five processor characteristics, together, essentially perfectly predict SPEC benchmark scores.

area, and numbers of processing transistors can be assumed to be derived from a single die design. Analysis of this data shows anywhere from 1 to 3 unique microarchitecture/die size/processing transistor combinations were being used to produce many dozens of processor models.

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

[89] Hyperthreading is Intel's name for multithreading capability, additional circuitry added to the processor which 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 which can speed up performance when a computer program spawns multiple threads.

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

[91] We 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

Table 10 Log of SPEC 2006 Benchmark as Function of Processor Characteristics
Six Characteristics Model

Dependent variable is log of

	SPECf06	SPECi06	SPECfr06	SPECir06
lproc	0.171***	0.102**	0.366***	0.417***
	(0.0313)	(0.0359)	(0.0710)	(0.0736)
lcache	0.103**	0.0896**	0.151**	0.126***
	(0.0327)	(0.0250)	(0.0491)	(0.0332)
lcores	0.117**	0.0190	0.566***	0.709***
	(0.0358)	(0.0321)	(0.0434)	(0.0466)
lvcore	0.0407***	0.0190*	0.0840***	0.132***
	(0.00886)	(0.00799)	(0.0113)	(0.0105)
lmaxmhz	0.567***	0.750***	0.141	0.367***
	(0.0593)	(0.0457)	(0.123)	(0.0685)
autop	0.0656*	0.00220	0.00394	-0.0175
	(0.0266)	(0.0538)	(0.0239)	(0.0374)
Microarchitecture dummies	Y	Y	Y	Y
Observations	1160	1190	2207	2417
R-squared	0.965	0.960	0.981	0.973
N_clust	31	31	29	30

Cluster robust standard errors in parentheses, clustered on Intel microarchitecture.
* p<0.05, ** p<0.01, *** p<0.001
lproc: log base processor clock rate
lmaxmhz: log maximum single core clock rate, not equal to base clock rate if turbo feature available
lcores: log of number of physical cores in processor
lvcore: log of additional “virtual” logical cores if multithreading available
lcache: log of amount of last level cache memory on processor chip
autop: autoparallelization enabled in compiler when SPEC software was compiled, dummy variable

Table 11 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
lproc	0.265***	0.150***	0.497***	0.439***
	(0.0351)	(0.0376)	(0.0840)	(0.0672)
lcache	0.0788**	0.0582**	0.164**	0.137***
	(0.0254)	(0.0191)	(0.0591)	(0.0295)
lcores	0.143***	0.0446	0.559***	0.678***
	(0.0258)	(0.0263)	(0.0527)	(0.0297)
lvcore	0.0603***	0.0315***	0.0963***	0.149***
	(0.0152)	(0.00451)	(0.0152)	(0.00787)
lmaxmhz	0.453***	0.692***	0.0151	0.334***
	(0.0652)	(0.0551)	(0.114)	(0.0644)
Microarchitecture dummies	Y	Y	Y	Y
Observations	331	340	449	454
R-squared	0.988	0.985	0.990	0.994
N_clust	30	30	28	28

Cluster robust standard errors in parentheses, clustered on Intel microarchitecture.
* p<0.05, ** p<0.01, *** p<0.001

calculating medians or means in Table 11. Table 11 reports results using logs of medians; using logs of means would give almost identical results.

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 SPEC score with a time-invariant SPEC aggregator function.

Second, these characteristics also will affect cost. Every distinct Intel microarchitecture is manufactured using a single fabrication technology node, so the microarchitecture dummies also capture variation in microprocessor manufacturing cost that is unique to particular chip microarchitectures and manufacturing technology. As previously described, different quality grades (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.

Finally, in addition to the chip characteristics determining SPEC performance, there are 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. 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. SPEC performance regressions tell us that desktop and server performance should be modelled 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 tilted toward performance on embarrassingly parallel workloads (with 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, but will be much more limited and less useful for mobile processors, with even more limited availability, and therefore least useful, for hedonic measurement of server processor prices.

6. Conclusion

There is considerable evidence that semiconductor manufacturing innovation has historically been responsible for perhaps a 20-30% 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 chip applications. 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 pressure now seems to have lessened, in part the result of rapidly increasing fixed costs sunk into the design of applications using the most advanced manufacturing technology, and, perhaps more controversially, 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 poor quality public data, speculations about Intel pricing behavior, and most likely, omitted variables in hedonic price models.

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 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 capability that both cost less and used less energy, year after year. The winding down of Moore's Law means that the technological scaling that drove these historical declines, and implicitly underlie 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; they won't decline, or even stay constant as computing capacity increases, as they have in the past. Investments in entirely new technologies will be needed, as will a renaissance of creativity and innovation in software, the neglected sibling living in the shadow of dramatically cheapening hardware for the last 50 years.

Appendix Table A1.

```
   SPEC CPU  |      Coef.   Robust         z    P>|z|     [95% Conf. Interval]     N        R2 #
   Benchmark |     CAGR     Std. Err.                                                        CPUs
-------------+--------------------------------------------------------------------------------
1995m5-2000m3
int95        |   .5826577   .0175146    33.27   0.000     .5483296    .6169857    152     .92 41
fp95         |   .6397016   .0231907    27.58   0.000     .5942486    .6851546    142     .90 41
int95_rate   |   .6241582   .0273672    22.81   0.000     .5705194     .677797     54     .87 20
fp95_rate    |   .7227752      .0331    21.84   0.000     .6579003    .7876501     47     .83 18
2000m11-2004m11
int2000      |   .3304092   .0173773    19.01   0.000     .2963503    .3644681    215     .80 76
fp2000       |   .3429411    .023522    14.58   0.000     .2968389    .3890433    203     .81 73
int2000_rate |   .4697731   .0512966     9.16   0.000     .3692337    .5703125    160     .77 59
fp2000_rate  |   .3989549   .0351676    11.34   0.000     .3300276    .4678822    162     .84 59
-------------+--------------------------------------------------------------------------------
2005m2-2007m1
int2000      |   .3222474    .016442    19.60   0.000     .2900217    .3544732
fp2000       |   .3365855    .022279    15.11   0.000     .2929195    .3802515
int2000_rate |   .4650892   .0475414     9.78   0.000     .3719098    .5582686
fp2000_rate  |   .3986346    .032545    12.25   0.000     .3348476    .4624217
2005m6-2012m11
int2006      |   .1709304   .0069587    24.56   0.000     .1572916    .1845691    689     .84 254
fp2006       |   .2467286   .0077563    31.81   0.000     .2315266    .2619306    690     .87 254
int2006_rate |   .2472256    .013015    19.00   0.000     .2217167    .272734     728     .62 278
fp2006_rate  |   .2537211   .0101781    24.93   0.000     .2337725    .2736698    711     .76 261
-------------+--------------------------------------------------------------------------------
2013m1-2016m5
int2006      |   .1687175   .0064265    26.25   0.000     .1561218    .1813133
fp2006       |   .2414989   .0070952    34.04   0.000     .2275926    .2554053
int2006_rate |   .2417978   .0119286    20.27   0.000     .2184181    .2651774
fp2006_rate  |   .2480768   .0093352    26.57   0.000     .2297801    .2663735
```

Notes:
intxx and fpxx are SPEC CPU integer and floating point base scores (no special compiler optimizations used) when single instance of benchmark run on CPU.

intxx_rate and fpxx_rate are SPEC CPU scores with multiple instances of benchmark programs run simultaneously; number of instances is entirely at discretion of entity running benchmark—may be as high as maximum number of threads, but may also be maximum number of cores, or any number less than that (on processors with symmetric multithreading capability—Intel version is branded as "hyperthreading"—additional program execution hardware in a CPU core allows as many as two threads to simultaneously share a single core's remaining hardware).

Model estimated was
ln(SPEC CPU benchmark) = a + b * monthly date of initial CPU availability in any manufacturer's computer hardware + c * autoparallelization indicator + d * time shift indicator x monthly date of initial CPU availability in tested hardware.

where
autoparallelization = 1 if autoparallelization turned on at compile time for 2006 benchmark, 0 otherwise
time shift indicator = 1 if year > 2004 for SPEC 2000 benchmarks, 0 otherwise
= 1 if year > 2012 for SPEC 2006 benchmarks, 0 otherwise
Annualized growth rate estimated as exp(b + d* timeshift indicator)^12 -1

Time shift indicators were statistically significant, as were autoparallelization indicators.

References (Under Construction)

A. Aizcorbe, K. Flamm, and A. Khurshid, “The Role of Semiconductor Inputs in IT Hardware Price Decline: Computers versus Communications,” *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches*, E. Berndt and C. Hulten, eds., Univ. Chicago, 2007, pp. 351-381.

A. Aizcorbe, S.D. Oliner, and D.E. Sichel, “Shifting Trends in Semiconductor Prices and the Pace of Technological Progress,” Finance and Economics Discussion Paper 2006-44, Federal Reserve Board, 2006.

R. Black, “Rambus, Bring Invention to market,” July 2013, available at http://www.iesaonline.org/downloads/IDC_Presentation_to_IESA_Thought_Leadership_Forum.pdf .

M. Bohr, “14nm Process Technology: Opening New Horizons,” presentation to Intel Developer Forum, San Francisco, 2014, available at http://www.intel.com/content/dam/www/public/us/en/documents/pdf/foundry/mark-bohr-2014-idf-presentation.pdf .

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

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

R. Burgelman, “Fading Memories: A Process Theory of Strategic Business Exit in Dynamic Environments,” *Administrative Science Quarterly*, vol. 39, no. 1, 1994, pp. 24-56.

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

A. Copeland, “Seasonality, Consumer Heterogeneity and Price Indexes: The Case of Prepackaged Software,” *J. Productivity Analysis*, vol. 39, no. 1, 2013, pp. 47-59.

C. Cunningham et al., “Silicon Productivity Trends,” Int’l Sematech SEMATECH Tech. Transfer #00013875A-ENG, 29 Feb. 2000.

C. Dieseldorff, “Watch out for 200mm Fabs!”, October 19, 2016, available at http://www.semi.org/en/watch-out-200mm-fabs-fab-outlook-2020-0 .

H.E. Esmaeilzadeh et al., “Power Challenges May End the Multicore Era,” *Comm. ACM*, vol. 56, no. 2, 2013, pp. 93-102.

K. Flamm, “Measurement of DRAM Prices: Technology and Market Structure,” in M. Foss, M. Manser, and A. Young, *Price Measurements and Their Uses*, NBER and University of Chicago Press, 1993.

K. Flamm, *Mismanaged Trade? Strategic Policy in the Semiconductor Industry*, Brookings, 1995..

K. Flamm, “Moore’s Law and the Economics of Semiconductor Price Trends,” *Int’l J. Technology, Policy and Management*, vol. 3, no. 2, 2003, pp. 127–141.

K. Flamm, “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*, Nat’l Academies Press, 2004.

K. Flamm, “Economic Impacts of International R&D Coordination: SEMATECH and the International Technology Roadmap,” K. Flamm and S. Nagaoka, eds., *21st Century Innovation Systems for Japan and the United States: Lessons from a Decade of Change: Report of a Symposium*, Nat’l Academies Press, 2009.

K. Flamm, “"The Impact of DRAM Design Innovation on Manufacturing Profitability," *Future Fab International* 35 (November 2010), available at ww.future-fab.com/documents.asp?d_ID=4763 .

K. Flamm, “Causes and Economic Consequences of Diminishing Rates of Technical Innovation in the Semiconductor and Computer Industries,” presented at APPAM Fall Research Conference, 2014.

S. Fuller and L. Millett, eds., *The Future of Computer Performance: Game Over or Next Level*, Nat’l Academies Press, 2011.

N. Gandal, “Hedonic Price Indexes for Spreadsheets and an Empirical Test for Network Externalities,” *RAND J. Economics*, vol. 25, no. 1, 1994, pp. 160-170.

J. Hennessey and D. Patterson, *Computer Architecture: A Quantitative Approach*, 5th ed., Morgan Kaufmann, 2012.

B. Holt, “Facing the Hot Chip Challenge (Again),” presented at Hot Chips 17, 2005, http://www.hotchips.org/wp-content/uploads/hc_archives/hc17/2_Mon/HC17.Keynote/HC17.Keynote1.pdf.

B. Holt, “Advancing Moore’s Law,” presented at Intel Investor Meeting, Santa Clara, 2015, available at http://files.shareholder.com/downloads/INTC/0x0x862743/F8C3E42B-7DA9-4611-BB51-90BED3AA34CD/2015_InvestorMeeting_Bill_Holt_WEB2.pdf .

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

J. Hruska, “Nvidia deeply unhappy with TSMC, claims 20nm essentially worthless,” posted March 2012, http://www.extremetech.com/computing/123529-nvidia-deeply-unhappy-with-tsmc-claims-22nm-essentially-worthless .

IC Insights, “Global Wafer Capacity 2016-20 Product Brochure,” 2016, available at http://www.icinsights.com/data/reports/4/0/brochure.pdf?parm=1454865474 .

IC Knowledge, “DRAM Trends,” 2004, available at https://web.archive.org/web/20041210172733/http://www.icknowledge.com/trends/dram.html .

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

Intel, *Microprocessor Quick Reference Guide*, 2008, available at http://www.intel.com/pressroom/kits/quickreffam.htm .

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

Intel, *2015 Intel Annual Report*, 2016.

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

H. Jones, “10nm Chips Promise Lower Costs,” *EETimes*, June 15, 2015, available at http://www.eetimes.com/author.asp?section_id=36&doc_id=1326864 .

D. Jorgenson, “Information Technology and the US Economy,” *Am. Economic Rev*., vol. 91, no. 1, 2001, pp. 1-32.

D. Jorgenson, M.S. Ho, and K.J. Stiroh, “A Retrospective Look at the US Productivity Growth Resurgence,” *J. Economic Perspectives*, vol. 22, no. 1, 2008, pp. 3-24.

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

B. Krzanich, “BIG or small…It’s All About the Details,” presentation at Intel Investor Meeting, 2012, available at http://www.cnx-software.com/pdf/Intel_2012/2012_Intel_Investor_Meeting_Krzanich.pdf .

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

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

S. Lawson, “The Moore's Law blowout sale is ending, Broadcom's CTO says,” *PC World*, Dec. 5, 2013, available at http://www.pcworld.com/article/2069740/the-moores-law-blowout-sale-is-ending-broadcoms-cto-says.html .

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

J. Lipsky, “Samsung Describes Road to 14nm,” *EETimes*, April 16, 2015, available at http://www.eetimes.com/document.asp?doc_id=1326369 .

D. McCann, “Silicon Interconnect, Packaging and Test Challenges from a Foundry Viewpoint,” June 2015, available at http://www.swtest.org/swtw_library/2015proc/PDF/SWTW2015_Keynote_McCann_GlobalFoundries.p df .

G. Moore, “Cramming More Components onto Integrated Circuits,” *Electronics*, vol. 38, no. 8, 1965, pp. 114-117. Reprinted in *Proceedings of the IEEE*, vol. 86, no. 1, 1998, pp. 82-85.

Z. Or-Bach, “Is the cost reduction associated with IC scaling over?,” *EE Times*, July 16 2012.

Z. Or-Bach, “Moore’s Law has stopped at 28nm,” *Solid State Technology*, March 2014, available at http://electroiq.com/blog/2014/03/moores-law-has-stopped-at-28nm/ .

M. Prud’homme, D. Sanga, and K. Yu, “A Computer Software Price Index Using Scanner Data,” *Canadian J. Economics*, vol. 38, no. 3, 2005, pp. 999-1017.

Qualcomm, “Qualcomm Snapdragon Integrated Fabless Manufacturing,” January, 2014, p.4, available at https://www.qualcomm.com/documents/qualcomm-snapdragon-integrated-fabless-manufacturing .

T. Raley, “IBM z13 Overview and Related Tidbits,” presentation, March 2015, available at 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 .

D. Rosso, “Global Semiconductor Sales Top $335 Billion in 2015,” February 2016, available at http://www.semiconductors.org/news/2016/02/01/global_sales_report_2015/global_semiconductor_sales_top_335_billion_in_2015/ .

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

W. Spencer and T. Seidel, “International Technology Roadmaps: The US Semiconductor Experience,” National Research Council, *Productivity and Cyclicality in Semiconductors: Trends, Implications, and Questions: Report of a Symposium*, Nat’l Academies Press, 2004.

J. VanWagoner, “How does Intel design and produce so many models of CPUs?”, 2014, available at https://www.quora.com/How-does-Intel-design-and-produce-so-many-models-of-CPUs

A.J. White et al., “Hedonic Price Indexes for Personal Computer Operating Systems and Productivity Suites,” *Annales D’Economie et de Statistique*, vol. 79/80, 2005, pp. 787-807.

C. Yang, “Challenges of Mask Cost & Cycle Time,” October 2001, available at http://www.sematech.org/meetings/archives/litho/mask/20011001/K_Mask_cost_Intel.pdf .

Z. Yeraswork, “Intel Cancels Fab 42,” *EETimes*, January 16, 2014, available at http://www.eetimes.com/document.asp?doc_id=1320670 .

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