Has Moore's Law Been Repealed? Empirical Analysis of Innovation in Semiconductors

Kenneth Flamm

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

Very Preliminary
Early Draft
Not For Circulation

An important economics literature (Oliner and Sichel, 2000; Jorgenson, 2001; Jorgenson, Ho, and Stiroh, 2005) credited much of the marked acceleration of U.S. productivity growth in the U.S. in the late 1990s to the impacts of information technology investments on the economy, which in turn were fed by technological advances in microelectronics, upstream. This literature utilized data on quality-adjusted price declines for semiconductors, and downstream, semiconductor-using information technology (computers, communications equipment), and traced through impacts on output and productivity in the larger U.S. economy using a standard economic growth accounting framework.

This literature highlighted the fact that the late 1990s were a period of unusually rapid technological progress in the manufacture of semiconductor components responsible for a considerable portion of technological improvement in information technology. Estimates suggest, for example, that from 40 to 60 percent of the decline in quality-adjusted prices for computers around this time was attributable to improvements in price-performance for semiconductors going into computers. Similarly, a rough estimate suggests that from 20 to 30 percent of declines in quality-adjusted communications equipment prices was attributable to improved semiconductors used in building this equipment. (Aizcorbe, Flamm, and Khurshid, 2007).

This paper explores whether the more rapid pace of technical progress in microelectronics in the late 1990s, attributable to technological innovation in manufacturing, as well as other factors, continued through the first decades of the 21st century, and whether the rate of innovation, as reflected in declines in quality-adjusted semiconductor prices, has indeed slowed.

Have contributions from innovation in semiconductor manufacturing to declining semiconductor prices really declined substantially? Are price and cost-based metrics measuring semiconductor innovation constructed by statisticians and economists in contradiction with a widely held view among engineers and technologists involved in the semiconductor industry that manufacturing innovation has gotten more costly and less productive? Furthermore, if semiconductor manufacturing innovation played such an instrumental role in lowering information technology costs in earlier decades, and, indirectly, increasing productivity in the U.S. economy, in the 1990s, does it follow that a decline in the pace of semiconductor manufacturing innovation is playing a significant role in the more sluggish rates of productivity improvement currently being measured in the U.S. economy?

To discuss these questions, I begin by describing what I will characterize as a current industry "majority view" about slowing technical progress in semiconductor manufacturing, as described in engineering journals and the semiconductor trade press, and the types of empirical evidence that is mustered in these outlets to support this view. I translate this into some observable economic consequences, and compare predictions to empirical data—detailed data on prices for different types of semiconductors.

Paradoxically, for such an important industry, useful public data on semiconductor prices is quite poor, and has been getting worse over time. This is due in part to industrial consolidation and increased concentration in a globalizing semiconductor manufacturing industry, and the sharp reduction in the numbers of corporate clients willing to pay industrial consulting firms for private market intelligence on prices for specific semiconductor product niches. These data have been available sporadically and proven useful to both academic researchers and government statistical agencies. They are now less available, and less useful, then they used to be.

Technological Innovation in Semiconductor Manufacturing[1]

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 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]

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):

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

Moore's original "Law" described only the denominator—a prediction that elements per chip would quadruple every two years. In 1965, Moore didn't originally anticipate rapid future advances in technology nodes. Acknowledging that an IC containing 65,000 elements was implied by 1975, Moore

[1] This section draws heavily on Flamm, 2017 (forthcoming).

[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.

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

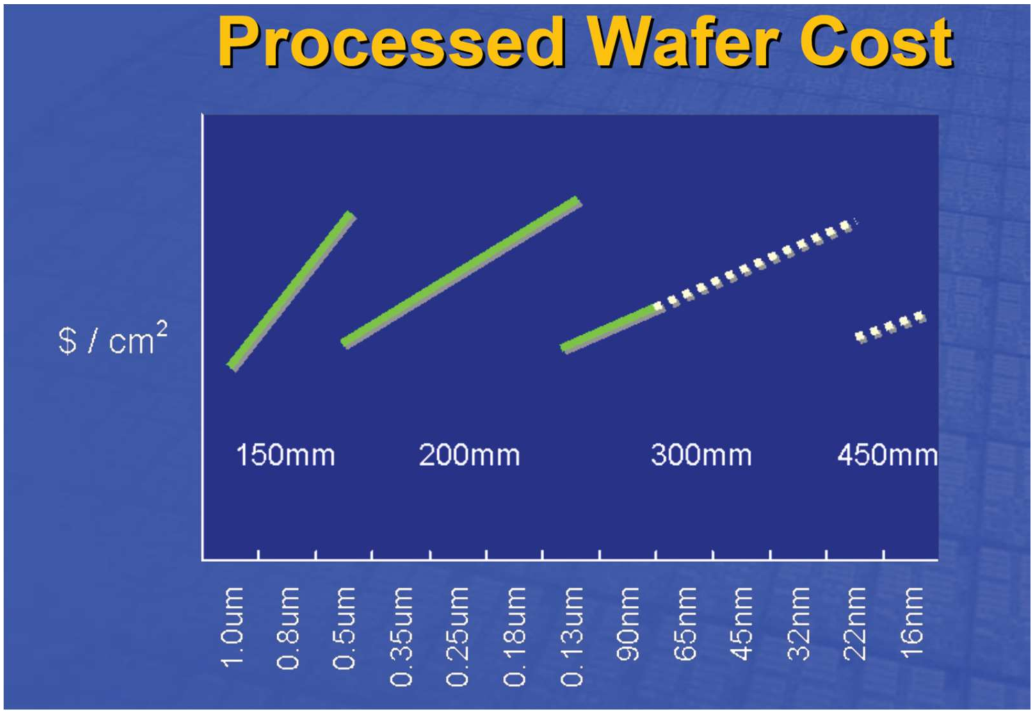
wrote: "I believe that such a large circuit can be built on a single wafer. With the dimensional tolerances already being employed…65,000 components need occupy only about one-fourth a square inch."[7]

Rewriting this more concisely without relying on Moore's prediction about numbers of elements per chip (and adding 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") in circuit elements by 30%, implying 50% smaller silicon areas per circuit element.



Source: Holt (2005), slide 8.

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

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

Completing the economic story, cost per wafer area processed, averaged over long periods, increased only slowly.[8] At new technology nodes, processing cost per 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.

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 produced at Intel after Moore and others left to found that company, which immediately became the largest volume product in the semiconductor industry and the primary product driving Intel's initial growth—dropped in price substantially faster than 20% pre-1995. The steeper decline rate in part reflected further increases in density due to circuit design improvements (e.g., reduction in memory cell footprint), 3-D interconnect layers enabling tighter packing of circuit elements,[9] and gradual introduction of 3-D into physical designs of transistors and other circuit elements.[10] In addition, operating characteristics of a given circuit design—in particular, switching speed and power requirements—improved with new manufacturing technology, and made an additional contribution to quality-adjusted price.

In the mid-1990s, the semiconductor manufacturing industry arrived at a significant technological inflection point.[11] New technology nodes began arriving at two-year intervals, replacing three-year cycles. 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

[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.

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

[10] 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.

[11] 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.

In 1985, Intel had exited the DRAM business, which had been driving its manufacturing technology development, and refocused its R&D on logic circuit design. Burgelman (1994), pp. 32-46.

By the end of the 1980s, Intel was trailing in manufacturing technology. Intel shifted gears, and 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 DRAM makers in 1997, at 250nm, and remained on a 2 year cycle through 2014. Author's calculations based on Intel (2008), IC Knowledge (2004), http://ark.intel.com.

introduction of new manufacturing technology, intended to benefit the competitiveness of US chip producers. In 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.[12] 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.[13]

Using (2), but adopting shorter two-year cycles for new technology nodes, implies rates of annual decline in transistor cost accelerating to almost 30%. If other innovations added at least another ten or more percentage points decline in quality-adjusted price onto manufacturing cost declines (as apparently happened pre-1995), annual declines in quality-adjusted transistor prices would exceed 40% annually.

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.

At Intel, the post-1995 two-year technology development cycle was explicitly incorporated into marketing efforts, and dubbed the Intel “tick-tock” development model in 2007.[14] 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”).

Intel’s publicly disclosed version of (2), purged of sensitive cost numbers by indexing variables to equal one at 130nm, is shown in Figure 2, and in Table 1, with annualized trends. 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. [15] If accurate, these numbers indicate average chip area per transistor shrank by 38% at each new node from 130nm through 22nm.[16] Nor did Intel’s wafer-processing costs stay constant. However, as long as average area per transistor declined at faster rates than processing costs per area increased, transistor cost would continue to decline. The cost per transistor estimates are revisited below.

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

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

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

[15] See Flamm (2017) for a more detailed explanation.

[16] 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.

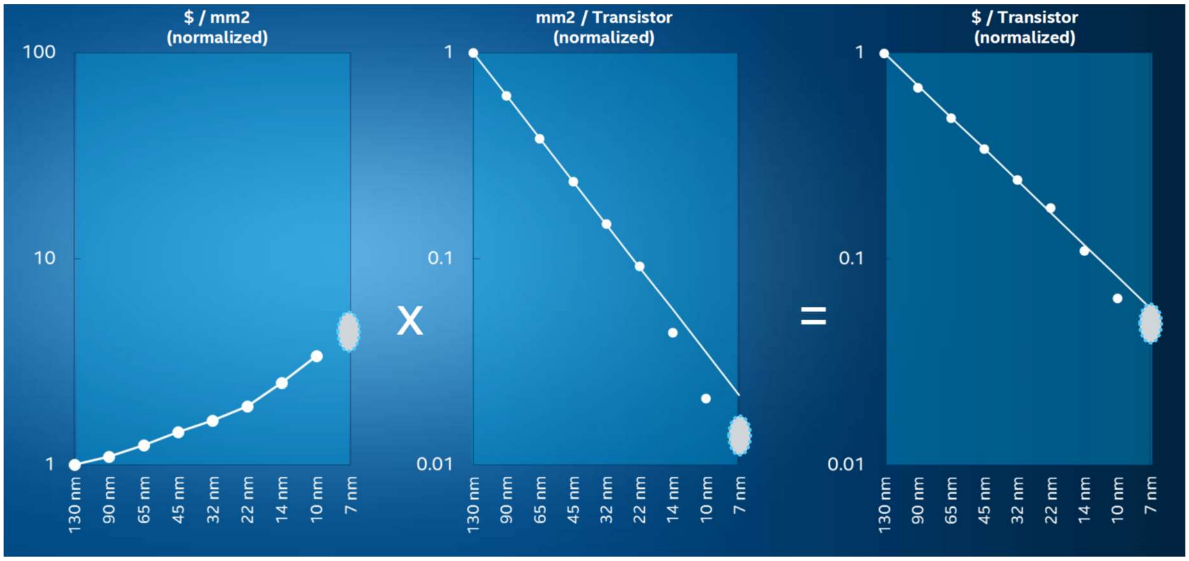


Figure 2 Intel's Version of Equation (2)

Source: Holt(2015), slide 6.

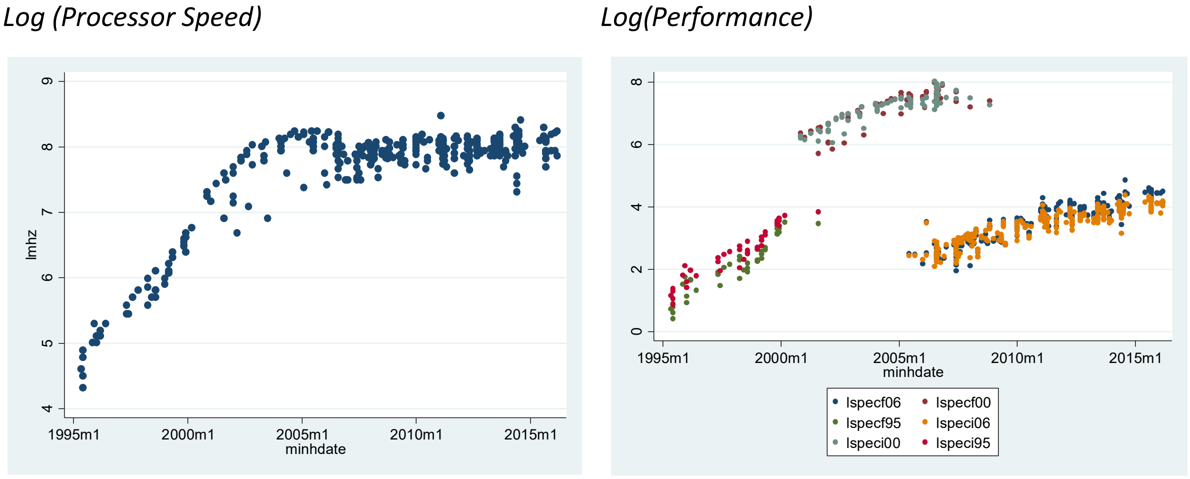
								Compound Annual Percentage Change:		
Year Intel 1st Shipped New Product at Tech Node	Tech Node (nm)	Wafer Processing Cost ($ / mm^2)	X	Transistor size (mm^2 / transistor)	=	$ Cost / Transistor		Wafer Processing Cost ($ / mm^2)	Transistor size (mm^2 / transistor)	$ Cost / Transistor
2002	130	1		1		1				
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

Smaller is Cheaper, Faster and Greener for Free

These 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.[17] 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 3 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.[18] Cheaper processors were also faster—stimulating increased demand for new computers in offices, homes, and workplaces.


Log (Processor Speed)

Log(Performance)

Figure 3. 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.

And, of course, the electronic circuits made from these ever cheaper transistors continued to evolve and improve. Ever more complex electronics made from smaller, faster, cheaper transistors enabled an explosion in electronics design creativity, that led to the PC, mobile computing communications equipment, and the ubiquitous electronic hardware infrastructure of today's Internet.

[17] 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.

[18] 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 Hennessey and Patterson (2012), chap. 1; Fuller and Millett (2011), chap. 2.

An End To Moore's Law?

Unfortunately, the golden age of more quickly 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 3), 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.[19]) 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 became the focus for improved computing performance.

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

The difficulty and cost of parallelization of software is an economic factor limiting utilization of cheap multicore CPUs on hard-to-parallelize applications.[20] 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."[21]

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 4 shows contract chipmaker GlobalFoundries' 2015 transistor manufacturing costs at recent technology nodes.[22]

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

[20] 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".

[21] Hennessey and Patterson (2012), p. 411.

[22] Like Table 1, this figure probably does not include R&D costs.

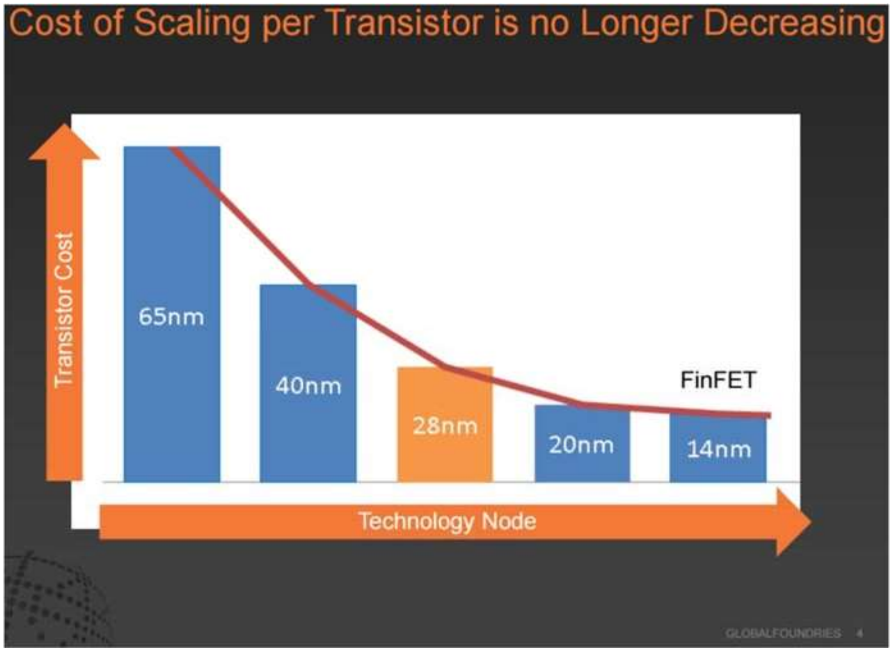


Figure 4. Transistor Manufacturing Cost at Recent Technology Nodes
Source: McCann (2015).

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.[23] (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.[24]) Charts like Figure 5, showing increased costs at sub-28nm technology nodes, were frequently published between 2012 and 2016. Figure 5 is not inconsistent with Figure 4, since Figure 5 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 4—the foundry's processing costs—does not.[25] These fixed costs have grown exponentially at recent technology nodes and create enormous economies of scale.[26] Some foundries have publicly acknowledged that recent

[23] 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).

[24] 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.

[25] A set of 10 to 30 different photomasks is typically employed in manufacturing a chip design. For a low to moderate volume product, acquisition of a mask set is effectively a fixed cost.

[26] 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.

technology nodes now deliver higher density or performance at the expense of higher cost per transistor.[27]

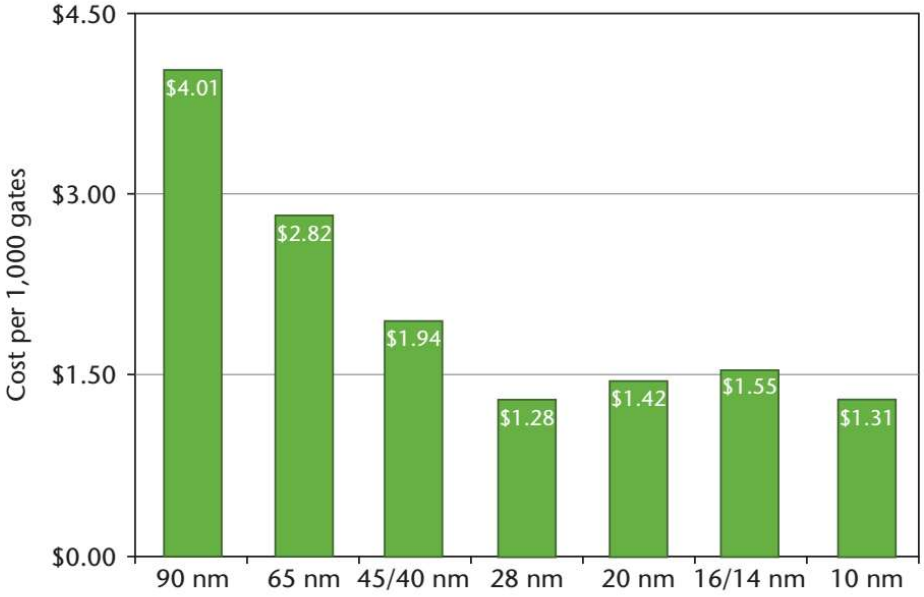


Figure 5. Cost per logic gate, with projection for 10nm technology node
Source: Jones (2015)

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.

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.[28] Indeed, current PC users are already seeing their multicore machines

[27] 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).

[28] Esmaeilzadeh, et. al. (2013), pp. 93-4.

"throttling" with attempts to use all cores for intensive computations at the highest clock rates, hitting thermal limits and then 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.

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."[29]

Do We See A Slowing Down of Moore's Law Cost Declines in Economic Statistics?

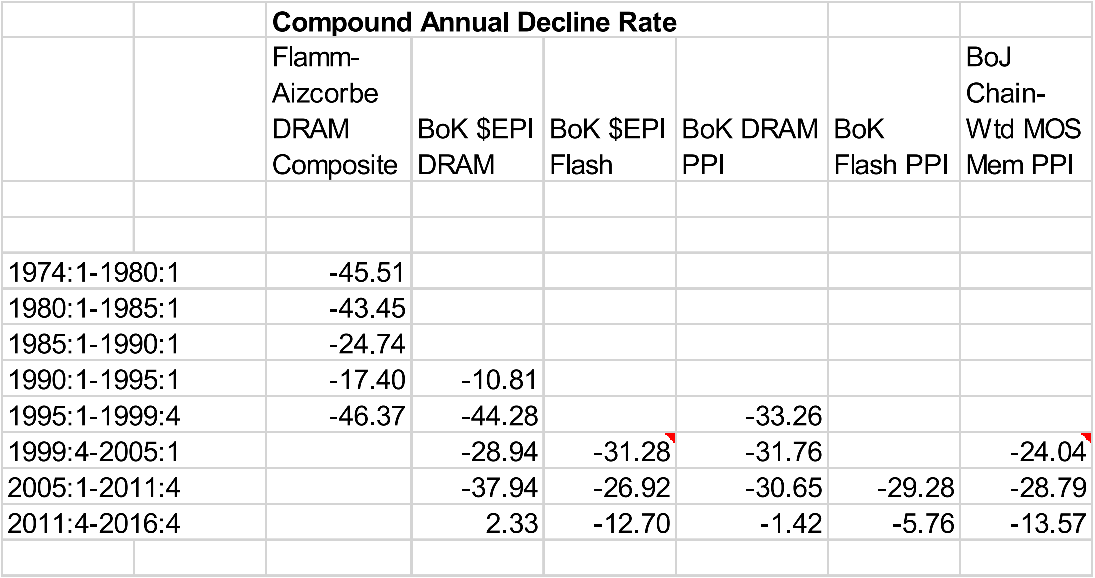
If Moore's Law has slowed or even stopped, we should see it in economic metrics. An obvious place to look is in the price statistics for computer memory chips, which remained the mass volume semiconductor product par excellence through the end of the 20th century. DRAMs were later superseded by flash 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.[30]

Table 2 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.[31] This data has no longer been available in recent years.

[29] Bourzac, (2016).

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

[31] 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).



		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 2. 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.[32] 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.[33]

Finally, since 2000, the Bank of Japan has published a chain-weighted "MOS memory PPI" with weights that are updated annually. This index is likely to be predominantly a mix of DRAM and flash memory, tilting more toward flash in recent years. Generally, except for the period from 1985-1995, 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,[34] 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 (now producing the majority of the DRAM sold) price indexes 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

[32] 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 now owns 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).

[33] 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).

[34] See Flamm (1995).

Law dynamic.[35] 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) has become a staple in Micron's investor conferences.[36]

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, had 32% of global production capacity in that year.[37]

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 3.)_While the pattern seems consistent with a slowing down of Moore's Law prior to 2012, this study unfortunately ends with data from 2012, 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.

	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 3. Quality-Adjusted Price Index for Fabricated Wafers
Source: Byrne, Kovak, and Michaels (2016).

The Intel Exception?

In contradiction to the above observations in other product segments, there is one enormously important player in the semiconductor industry—Intel—that maintains vehemently that its costs continue to come down at historical Moore's Law rates. The main exhibit used to support this point factually in public is the leftmost panel in Figure 2, which shows transistors costs declining at 14-18% annual rates after the millennium, and falling faster at the most recent technology nodes.

Interpreting the recent economic history of Moore's Law, how can Intel's description of continuing declines in manufacturing cost per transistor be consistent with reports from other chip

[35] Micron 2015 Winter Analyst Conference (2015).
[36] Micron's Raymond James Institutional Investor Conference (2016); Micron Analyst Conference (February, 2017).
[37] Foundry share calculations based on Yinug (2016), Rosso (2016), IC Insights (2016).

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 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.[38] (Intel overlooked GlobalFoundries, which by acquiring IBM's semiconductor business in 2015, substantially increased its scale.)[39] Both TSMC and 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.[40] 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.[41] 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,

[38] Krzanich (2012), slide 19.

[39] What constitutes leading edge technology in memory chips requires more of a judgment call, and several large memory specialist IDMs (Hynix, Toshiba, Micron) might also arguably be categorized as being near the leading edge.

[40] 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.

[41] 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.

photomask costs become the overwhelmingly dominant element of silicon wafer-processing cost at leading edge technology nodes.[42]

Intel, with the largest production runs in the industry (perhaps 300 to 400 million processors in 2014[43]), has huge volumes of wafers to amortize the cost of its masks, and is certainly benefitting from significant economies of scale. A single Intel processor design (and mask set) is the basis for scores of different processor models sold to computer makers. Processor features, on-board memory sizes, processor speeds, and numbers of functioning cores can be enabled or disabled in the final stages of chip manufacture, and manufacturing process parameters can even be altered to shift the mix of functioning parts in desired ways.[44]

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 advantages are increasingly shifting 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![45]

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. Further, Intel's historical public disclosures about its manufacturing costs are quite confusing. Recent Intel statements

[42] Lattard (2014), slide 6.

[43] 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.

[44] 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.

[45] Dieseldorff (2016).

about its manufacturing costs have been deployed as the primary factual evidence against the proposition that Moore's Law is slowing down, within the semiconductor manufacturing community.

Revisionist History?

The problem is illustrated by Figure x and Table x, which places 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). The graphics in Figure 6 have been digitized with the assistance of digitizing software[46] and recorded in Table 4, 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 at that technology node.

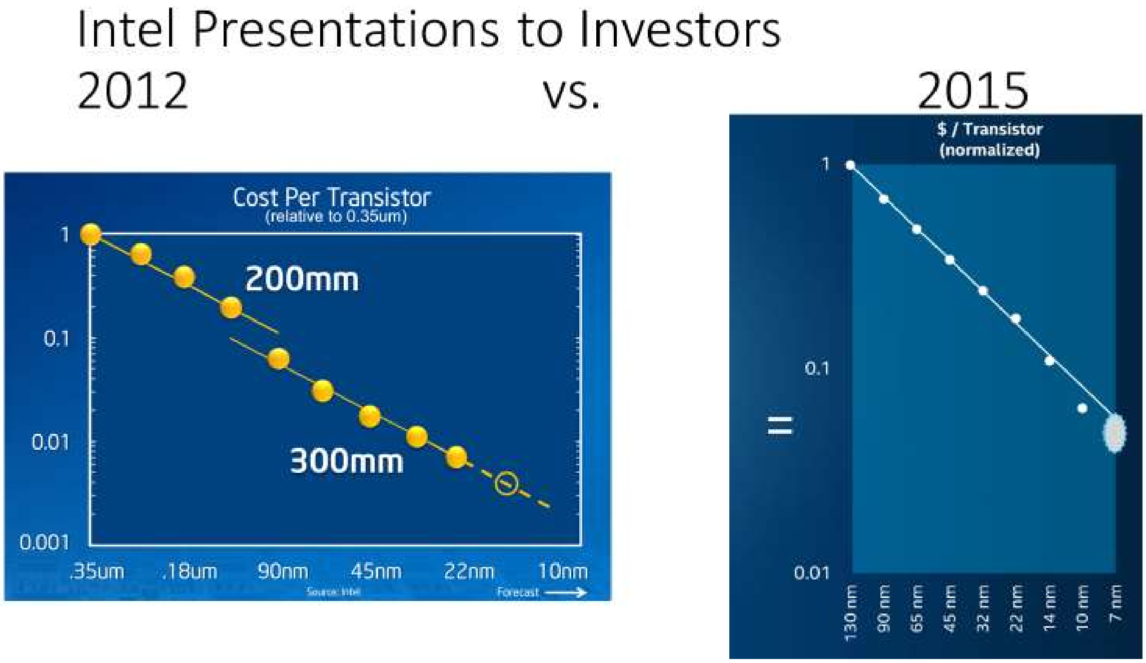


Figure 6

[46] See http://arohatgi.info/WebPlotDigitizer/.

		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 4 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. In contrast, a more recent presentation by Intel in 2015 restates the more distant historical record to show much slower declines in cost per transistor. 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). How do government price statistics compare to these divergent portraits?

Measuring Quality-Adjusted Prices for Microprocessors

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 rates during the “golden age” of the late 1990s and early 2000s, to a low of 2.5 percent for the year ending in January 2013. 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.

The Bureau of Labor Statistics is 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),[47] but this is not

[47] Based on a brief conversation with BLS officials, Cambridge, MA, July 2014.

entirely clear. There is also some evidence that the BLS may have employed variety of different methodologies for measuring its microprocessor price indexes over the 1995-2014 periods.[48]

As an alternative to the BLS measure, I have 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.[49] 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.

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.[50] 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

[48] The BLS web site shows three different "commodity" price indexes (as opposed to its single semiconductor industry price index) for microprocessors over this period. The current microprocessor "commodity" price index is based in December 2007, but is only reported on a monthly basis from September 2009 through the present. There are also two discontinued microprocessor commodity price indexes, one based in December 2004, and running through June 2005, and another based in December 2000 and running from 1995 through December 2004. One inference that might be drawn is that the BLS changed its methodology for measuring microprocessor prices three times during the period we are discussing.

[49] 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.

[50] 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 contiguous observations.

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 months 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 5 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, and 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 and hedonic indexes again fall more slowly than the PPI, but all three sets of declines are in the low single digits.

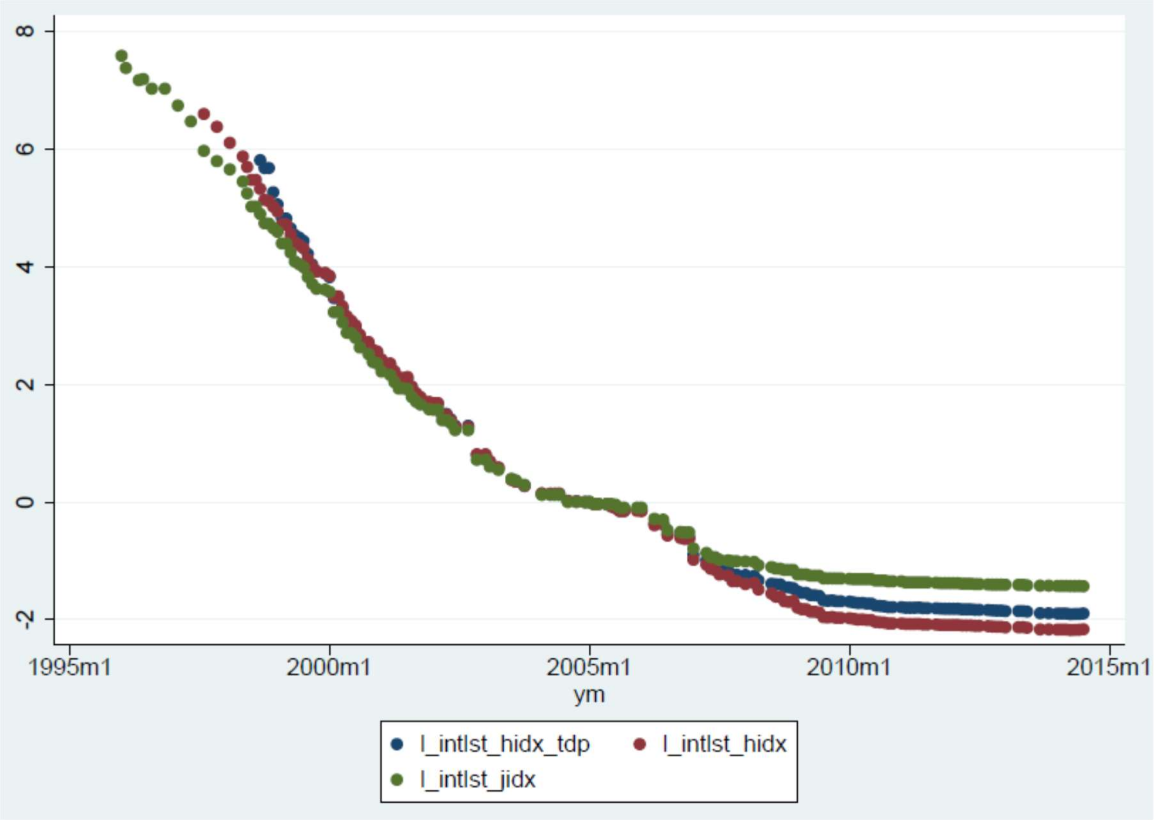


Figure 7. Geomean Matched Model and Hedonic Price Indexes for Intel Desktop Processors

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 5

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 4 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.

Is the Slowdown Real?

One recent study (Byrne, Oliner and Sichel, 2015) suggests an alternative explanation for the recent behavior of the official price indexes. This study suggests 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.[51] 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 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."[52] 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. Throwing away most of their sample of Intel products, and keeping only newly introduced models, they run an annual hedonic price model over pairs of years, and find quality-adjusted prices declining at the same rate in 2000-08 as in 2008-12, with a 39 percent annual rate of decline.[53] This is vastly higher than any of the rates shown in Table 5 for the equivalent time periods.

While the 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, the assertions 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, is essentially unobservable and untestable, since no data on Intel transaction prices for its wholesale sales 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.[54]

Also arguing against this claim is the behavior of the BLS computer price indexes. Changes in BLS producer price indexes for computers, constructed using hedonic methods, seem to mirror the decelerating declines in list price indexes for Intel microprocessors, before and after 2006, as would be expected given the significant role of microprocessor price in computer cost.

An alternative hypothesis to the one put forth in this study is that Intel's diminished propensity to alter its list prices in fact reflects its actual pricing behavior. Figure 7 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.

[51] 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/ .

[52] Ibid., pp. 8.

[53] 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.

[54] 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 .

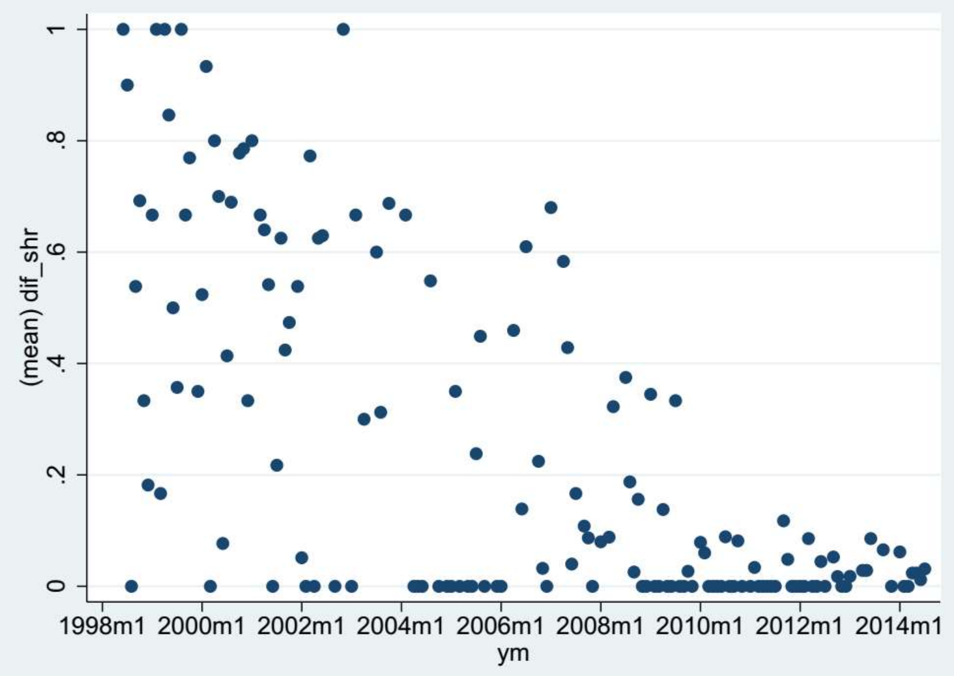


Figure 7. Fraction of Intel Desktop Processor Prices Changing From One Price List to the Next.
Source: Author's tabulation from dataset.

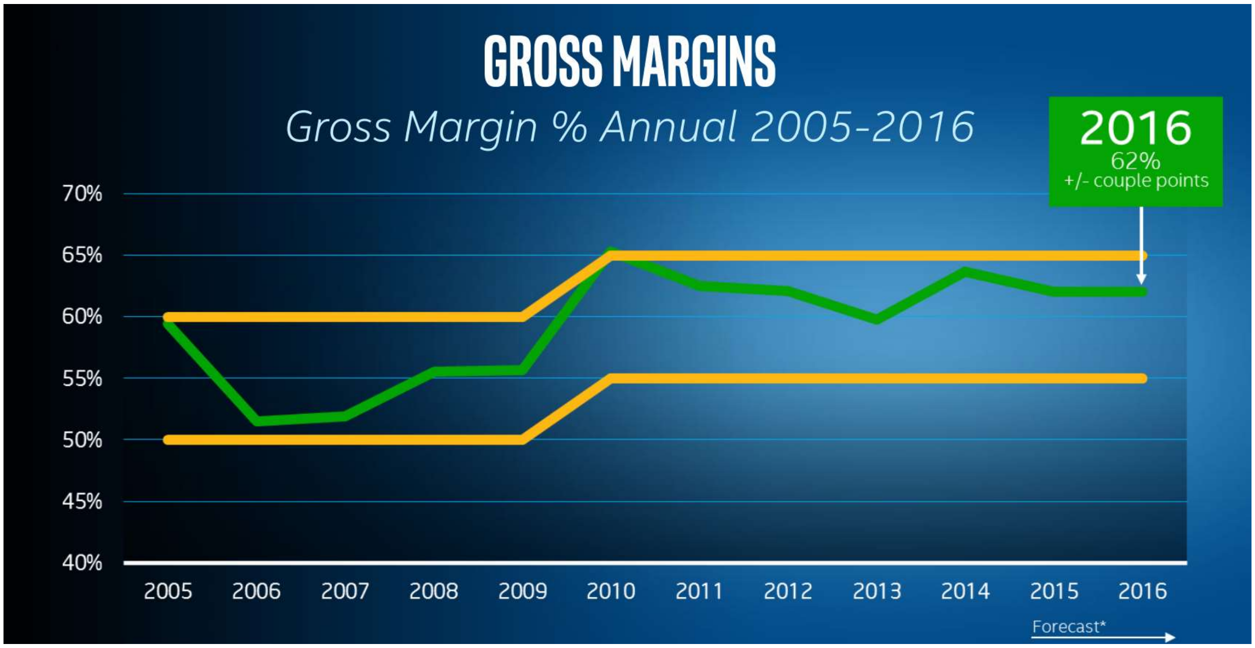


Figure 8. Intel's Post-2010 Gross Margin Elevation Objective
Source: Smith (2015).

Indeed, if one had to choose a date based on this chart for a climacteric in Intel pricing practice, 2010 would be as good a choice as any other choice. 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 8.) Raising its average sales prices was a key element of this strategy (See Figure 9.)

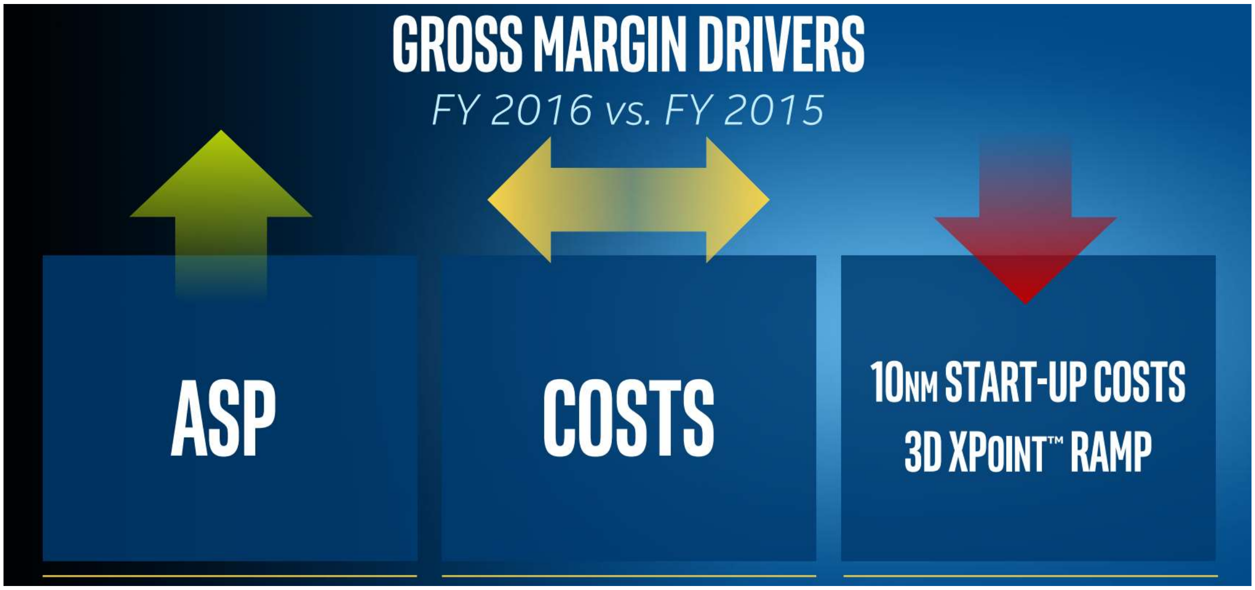


Figure 9. Intel’s 2015 Explanation to Its Shareholders for Success in Maintaining High Margins
Smith (2015).

Finally, there is one source of processor price data that is real, observed, and does not require hypothetical 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.[55]) 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 show up on the retail market labeled as “OEM package” (vs. “Retail Box” packaging). These products 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 internet retailers.) The retail data used in Table 4 seem to clearly point to a deceleration in microprocessor price declines after 2004.

[55] “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.

If one presumes that retail transaction prices (which are observable in the market), at least in the long run, should have some stable stochastic relationship to wholesale producer transactional prices (which, though not directly observable, are linked because of the direct and indirect linkages between retail and wholesale markets, and the impact of arbitrage and competitive market forces in distribution channels), then one would expect to observe a systematic change in the 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.[56] (A larger sample is in the works, but not yet complete!) The fixed effects regression model (permitting a particular low-end Celeron model, for example, to be related to Intel list price with a different retail margin than a high end i7 model) that I estimated specified that the log of retail price for model i in month t was given by

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

with R_{it} an observation on average retail price for model i in month t; Iit 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 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.

Figure 10 shows the results of estimating this model.[57] 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.[58]

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. The elasticity of

[56] My retail price data actually end in January 2011.

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

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

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.[59]

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.

Figure 10
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
```

Conclusion

There is some 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

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

technology, would be additional factors explaining even higher long run rates of decline in 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, more controversially, in part due to a slackening in the rate of cost decline at the technological frontier of semiconductor manufacturing.

While Moore's Law may not yet be entirely repealed, it clearly is undergoing significant revision, with broad implications for our society. A substantial economic literature connects 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 US economy, and higher rates of labor productivity growth in the US economy. If this is correct, it implies that a slowdown in semiconductor manufacturing innovation, and attenuation of price declines in both chips and IT, play a role in current stagnation in labor productivity growth in the US.

In the national security domain, access to superior electronics and IT capabilities historically created a qualitative, strategic technological advantage offsetting numerical inferiority in soldiers and systems, for militaries in advanced industrial societies. An end to Moore's Law would mean that the technical distance between leaders and laggards quickly shrinks, challenging the technological foundation of geopolitical strategic advantage.

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 end soon. 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 in the face of increasing computing capacity 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.

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, *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, “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-%2006101 6%20-%20Final.pdf