UNIVERSITY OF CAMBRIDGE

Bennett Institute for Public Policy Cambridge

WORKING PAPER

Twenty-first century progress in computing

Authors:

Diane Coyle and Lucy Hampton

Bennett Institute for Public Policy, University of Cambridge, UK

Date:

July 2023

Twenty-first century progress in computing

Diane Coyle[1] & Lucy Hampton[2]

Abstract

In the search for explanations for slower productivity growth since the mid-2000s in many countries, one possibility is a slower pace of progress in digital technologies. In this paper we show that the cost of computation has continued to decline rapidly, taking into account innovation in chip types and cloud computing. This is a continuation of its long-run trend; the decline has slowed since the mid-2010s, but not earlier. To the extent that the productivity slowdown is linked to technology use, the explanation is more likely to lie in elements of the input bundle other than computational power, such as human and organisational capital.

JEL codes: L16, L63, O47, D24

Keywords: computation, prices, productivity, AI, cloud computing

[1] Corresponding author.

Email addresses: dc700@cam.ac.uk (Diane Coyle) and ljh92@cam.ac.uk (Lucy Hampton).

Postal address: Alison Richard Building, 7 West Road, Cambridge, CB3 9DT, UK.

[2] Bennett Institute, University of Cambridge. Coyle is funded by The Productivity Institute Grant number ES/V002740/1 & by the Economic Statistics Centre of Excellence; Hampton is funded by The Gatsby Foundation. The authors are grateful for comments on earlier drafts from David Byrne, Bill Janeway, Shane Greenstein, Lennart Heim, Jen-Chung Mei, David Nguyen, and Ali Sen. We also thank Professor William Nordhaus for sharing his data workbook. The publicly available data constructed for this paper will be lodged in the University of Cambridge data repository. Sources for the copyrighted data used are available from the authors on request.

1. Introduction

Large technical advances have been made in the suite of digital technologies during the past 15-20 years, and both businesses and consumers have extensively adopted technologies ranging from cloud computing and machine learning to multiple apps and platform services. Use metrics such as time spent online, data usage, or reported use of cloud services have grown substantially.[3] The great majority of people use digital technologies frequently and value them highly (Brynjolfsson *et al.* 2020, Coyle and Nguyen 2023, Poquiz 2023).

A question often raised in the light of this progress is why, then, aggregate productivity growth seems to have stalled. This apparent paradox was first famously highlighted by Robert Solow (1987) and seems to have re-emerged with a vengeance since the widespread productivity slowdown from the mid-2000s, despite the dramatic behavioural and organisational changes involved in use of the technology. Indeed, productivity decompositions find that – although information and communication has remained one of the sectors with the fastest productivity growth over a longer period – it is also one of the biggest contributors to the recent slowdown in the UK (Coyle and Mei 2023, Goodridge and Haskel 2023). For example, Coyle and Mei (2023) find that labour productivity growth in the IC (Information and Communication) sector more than halved between 1998-2007 and 2008-2019, although it remained throughout the fastest productivity-growth sector in the economy.

There is an extensive literature exploring the productivity puzzle which divides` broadly into two camps. One argues that the modern digital innovations are not inherently productivity- or utility-enhancing, compared to earlier 20th century waves of innovation, or that the pace of innovation has diminished (Gordon 2017, Bloom *et al.* 2020) – raising the question as to why nevertheless firms are widely adopting digital innovations. The other focuses on the time and complementary investments needed to adopt and use the new technologies, implying that the productivity payoff will arrive eventually (Brynjolfsson *et al.* 2021, Bessen 2022).

In this paper, we focus on the recent progress of productivity in the computer sector itself to understand its contribution to the broader slowdown. This complements previous work (Abdirahman *et al.* 2020, 2022) looking at the deflator (and hence real output and productivity) of the telecommunications sector, which found that – with a range

[3] For example, for the UK, https://www.ofcom.org.uk/__data/assets/pdf_file/0023/238361/online-nation-2022-report.pdf; for the US https://ourworldindata.org/grapher/daily-hours-spent-with-digital-media-per-adult-user

depending on the selection of weights used in constructing the index – the price of telecommunications services had fallen by significantly more than the official price index used at the time indicated.

The question is what impact technological innovation in the sector since the mid-2000s, including the examples listed in Table 1, has had on its real output and productivity: is the appearance of significant productive innovation illusory? After all, it is widely argued that Moore's Law, the regularity that the number of transistors per chip doubled every 18-24 months, is slowing due to physical limits (Theis and Wong 2017, Flamm 2017, Ramu 2018, Shalf 2020). Or alternatively are there measurement challenges (similar to those identified in telecommunications) that mean price declines and corresponding output or productivity increases have been understated in current statistics? The answer matters not only for the narrow purpose of productivity measurement but also because – like all general purpose technologies (GPTs) – cheap and eventually ubiquitous computation, including increasingly artificial intelligence (AI) applications such as generative or large language models (LLMs), would have potentially transformative effects on the economy and society. How cheap and ubiquitous is computation?

Table 1: Post-2005 major information and communications technology (ICT) innovations

Innovation	**Applications**
Smartphones	Handsets, operating systems
3G-5G networks	Data-enabled network services
GPUs, TPUs, H100	Specialist processors; also reduced instruction set chips (RISC)
Deep learning, reinforcement learning	Machine learning advances e.g., AlphaFold
Large language or foundation models	AI systems e.g., GPT4, Dall-E
Cloud computing	Storage, SaaS, IaaS, ML
Additive manufacture	3D printing of components
Robotics in distribution	Drones, delivery robots
Robotic process automation	Automates business processes
Augmented and virtual reality	Some existing applications: "Metaverse"?
Quantum computing	At the R&D stage
Wearables	Devices e.g., smart watches, clothing, medical implants
Advances in nanotechnology	Construction of miniaturised components in computer chips, sensors etc.

Source: Authors' own

We address this using an engineering-based[4] approach to the cost of computation, taking computation as the utility-giving activity in Lancaster's (1966) sense. We start from previous work looking at the price of computational activity given the evolution of the hardware needed to compute, taking Nordhaus (2001, 2007) as our starting point. We extend this broad approach to CPU (central processing unit) improvements in recent decades as Moore's Law has slowed. We then go on to discuss how to consider some key recent innovations, including GPUs (graphics processing unit), cloud computing and AI. We show that they have enabled continuing rapid declines in the price of computing. However, as the capability of the hardware has increased, and rapid price declines have continued, the demand for compute in key applications has also grown. Recent work on the US software deflator has argued that the official software price index understates the decline in quality- or performance-adjusted price (Fleming 2022). On the other hand, the latest generation of AI applications require large amounts of compute (Sevilla *et al.* 2022), over and above the phenomenon of software 'bloat'.

We establish that there is considerable under-measured progress in computing since the start of the 21st century. To the extent its growth has slowed, this has mainly occurred since 2015 (and not earlier this century) but may accelerate again with expanding use of the cloud and new AI models. We end with a discussion of the economic impact of 21st century innovations in computing, including accessible AI applications, and the implications for productivity. Our conclusion is that computational power is in no way the binding constraint on productivity improvement, and argue that the barrier to progress lies in the need for co-investment and organisational innovation. If there is a 'puzzle' it lies in the need to understand why digital technologies are so difficult to *use* productively.

2. CPU performance and the cost of computing

2.1 Background

The standard way for an economist to understand progress is in terms of productivity – either labour productivity or total factor productivity. In their survey of the evidence from growth accounting studies and production function estimation to that date, Draca *et al.* (2007) conclude that the ICT sector had experienced a productivity acceleration since

[4] By 'engineering-based approach', we mean that we view different types of processor (including CPUs and GPUs) and cloud computing services as delivering different presentations of a single type of output: computation. For other papers using this approach, see Abdirahman et al. (2020, 2022) for telecommunications and Byrne *et al.* (2023) for semiconductors.

1995, and that there was reasonable evidence of a firm-level association between ICT use and productivity. A more recent literature review (Biagi 2013, 59) also concluded that the sector had seen faster productivity growth – more in the US than the EU – and this had contributed to economy-wide productivity: "From the empirical evidence analyzed in this literature review it seems quite safe to conclude that ICT had a major role in the US productivity acceleration observed in the period 1995-2005, both in terms of TFP growth in ICT-producing sectors and capital-deepening in ICT-using sectors,". In the EU the effect varied by country but in general experienced less TFP growth in the ICT sector and beyond. Cardona *et al*. (2013) reached the same conclusion in their review of the empirical literature. Van Ark (2016) found that despite rapidly declining ICT prices, investment in ICT hardware had slowed; although investment in ICT services had increased, the contribution of the sector (hardware plus software and services) to aggregate productivity had declined. One potential reason for this, he argued, was the deceleration of the decline in ICT prices compared with the 1995-2005 period.

As Nordhaus (2001, 2007) notes, the pace of price decline is a key statistic, because for productivity calculation a constant-quality deflator must be used. He constructed very long run price indices based on the computational performance of the hardware available in each era, and cost per unit of performance. Byrne (2022, 16) notes that although some relevant prices for computer equipment are adequately adjusted for quality change, many are not: "Despite extensive research on the subject, roughly 50% of US investment, 60% of US consumption, and some 80% of US production of electronic equipment is measured using price indexes with little to no supporting research verifying their ability to separate quality change from inflation." In a paper inspired by the same engineering-based approach to ours, Byrne *et al*. (2023) develop a volume-based semiconductor output index based on the number of transistors on chips of different types, finding that the resulting implicit price indices differ in significant ways from official indices, with a less pronounced (quality-adjusted) output boom in the 1990s and a less pronounced slowdown from the late 2000s.

Nordhaus describes four possible approaches to adjusting the deflators for quality improvements. He rejects hedonic regression, used by statistical agencies for some technology goods, on the grounds that it adjusts for a fixed set of input (not performance) characteristics in a linear way, whereas the features that affect performance change over time and their marginal prices will sometimes change in non-linear ways.

The three alternatives Nordhaus (2001) presents involve cost per metric of computing performance. He opts for an information-theoretic measure of performance: standardised operations per second. This begins with a standard measure, millions of instructions per

second, taking account of the nature of the instruction, and standardising by converting to 32 bit units. This calculation gives a physical, hardware performance measure, millions of standardised operations per second (MSOPS). A user cost of capital is applied to this to give a cost per unit of computation per second, based on an assumed 10% a year constant real interest rate, exponential depreciation at 10% a year, and 2,000 hours per year usage. His results show a decline in price far greater than in the official US statistics.

Our first contribution here is to extend the Nordhaus performance and price-performance series from 2007 to 2022, using performance and price data for CPUs. Several changes to the determinants of CPU performance occurred during this period. Most notable is the end of Dennard scaling around 2004 (Lotfi-Kamran and Sarbazi-Azad 2018, Johnsson and Netzer 2016), that is, the relationship that power consumption scales in proportion to chip area rather than number of transistors. Combined with Moore's Law, Dennard scaling implied that performance could be improved by doubling transistor counts without increasing power consumption. Post-2006, however, constraints on power and heat dissipation resulted in chips' clock speeds stalling around this time (Leiserson *et al*., 2020), slowing the rate of performance improvements of single-core systems. While chip performance continued to improve – in part due to Moore's Law and increasing the number of cores – the earlier rate of performance improvements may not be sustainable, as there are already signs of Moore's Law slowing down. (A glossary of relevant terms is given in the appendix.)

2.2 Data and methodology

For performance data, we used the Standard Performance Evaluation Corporation (SPEC) CPU benchmarks CPU2000, CPU2006 and CPU2017.[5] SPEC benchmarks, often considered the industry standard, have the advantage of being based on real-world applications which are updated in each version. Each version provides separate scores for floating point (FP) and integer (INT) computing, which are further divided into *speed* (time to complete a single task) and *rate* (throughput, or tasks completed in a time) scores. This produces a total of four types of benchmark: INT RATE, INT SPEED, FP RATE, FP SPEED. The distinction between speed and rate scores has become particularly important over the last two decades since rate better captures the benefits of modern multi-core systems, which outperform single-core systems at performing many parallelisable tasks but not necessarily single tasks. It is worth noting, however, that the SPEC sample may not be perfectly representative of the entire population of purchased computers, as their

[5] https://www.spec.org/benchmarks.html#cpu

benchmarks are primarily used to measure higher-end server performance in industries such as finance and telecommunications rather than personal computers or mobiles.

To construct the performance index, different versions of rate and speed scores are chained together using reference machines present in multiple benchmark versions. While speed and rate scores are not directly comparable, we use the fact that they are scalar multiples of one another for single-core systems (Munafo, 2022). Hence, we transform the data to make single-core rate scores more comparable with speed scores while maintaining comparability between multi-core and single-core rate scores. We identify outliers (often corresponding to systems with hundreds of chips) using the Mahalanobis distance[6] based on scores and the date and remove them from the sample; although including them does not alter the main conclusions of the paper (see appendix tables D.2, D.3 and D.4). Scores for systems with the same processor, date, clock speed, number of chips and number of cores are averaged to give chip-date-level scores which can be matched to chip-level price data. Appendix B sets out the methodological detail.

As we are interested in the pace of change in computing performance, we use a log-linear spline regression to estimate the rate of change in different decades for chips. This technique was also used in Nordhaus (2001, 2007) and has the property that average annual decline rates in different periods can be obtained as the cumulative sum of the coefficients. The model specification is as follows:

$$log10(perf_t) = \beta_0 + \beta_1 Year_t + \beta_2 DUM2005_t + \beta_3 DUM2015_t + \varepsilon_t \qquad (1)$$

where $perf_t$ is an index of computing performance, $Year_t$ is the year of hardware introduction, ε_t is the error term, and $DUM[YEAR]_t$ is a truncated power basis function defined as follows[7]:

$$DUM[YEAR]_t = \begin{cases} 0, & Year_t < [YEAR] \\ Year_t - [YEAR], & Year_t \geq [YEAR] \end{cases} \qquad (2)$$

[6] The Mahalanobis distance is a measure of the distance between a point and the mean of a distribution in a multidimensional space. In our case, the two dimensions are the SPEC scores and the date. Unlike the Euclidian distance, the Mahalanobis distance accounts for correlations between variables (e.g., increasing SPEC scores over time). Outliers are selected using this distance in the BACON algorithm implemented in Stata as described in Billor, Hadi, and Velleman (2000) and Weber (2010).

[7] t is the number of days between the date of hardware introduction and 1 January 1960 (i.e., it is a Stata datetime variable). To convert the coefficients to annual rates, the 'YEAR' and 'DUM[YEAR]' variables are divided by 365.25.

The years 2005 and 2015 are chosen as the values for the breakpoints ('knots'). 2005 is chosen as it is generally considered that Dennard scaling ended around 2004, and 2015 allows approximately equal spacing between the knots.[8]

The coefficients on the DUM[YEAR] variables capture the effect of increasing 'year' by 1 on the logarithm of the computing performance index (and thus the average annual logarithmic rate of growth) relative to the rate of growth in the previous decade. The cumulative sum of these coefficients up to the decade can be interpreted as average annual rates of growth for that decade.

To obtain a price-performance measure, we match the chip-level performance data to chip-level price data. Release prices were obtained from Passmark,[9] TechPowerUp,[10] and CPUWorld.[11] As Passmark does not report prices at introduction, we used the Wayback Machine[12] to obtain past versions of the CPU 'mega page' at yearly intervals and matched to the performance data by year of computer system hardware availability.[13] Passmark price data appears to be scraped from Amazon and eBay, while TechPowerUp and CPUWorld data is uploaded manually by website contributors. Where sources disagree on prices, the average is taken. For multi-chip systems we take the price of a single chip and multiply by the number of chips, although this is likely to produce an overestimate. Prices are deflated using the 2021 GDP deflator to obtain constant prices.

To obtain decadal price decline rates, we use a similar specification to the one used above for the performance data, with the same assumptions for the knots. i.e.:

$$log10(price_perf_t) = \beta_0 + \beta_1 Year_t + \beta_2 DUM2005_t + \beta_3 DUM2015_t + \varepsilon_t \qquad (3)$$

where all regressors are defined as before.

[8] It is possible to allow the breakpoints to be selected automatically, for example based on percentiles of the data, or by using an algorithm such as the multivariate adaptive regression spline (MARS) (Friedman, 1991). However, this approach was not taken in this paper as the same knots must be chosen in multiple regressions to allow comparability of different results.

[9] https://www.cpubenchmark.net/CPU_mega_page.html

[10] https://www.techpowerup.com/cpu-specs/?sort=name

[11] https://www.cpu-world.com/

[12] https://web.archive.org/

[13] Year of computer system hardware availability in the performance data is not necessarily the same as the year of CPU availability in the Passmark data, as the former is the year in which *all* of the computer hardware is available, including other components such as memory.

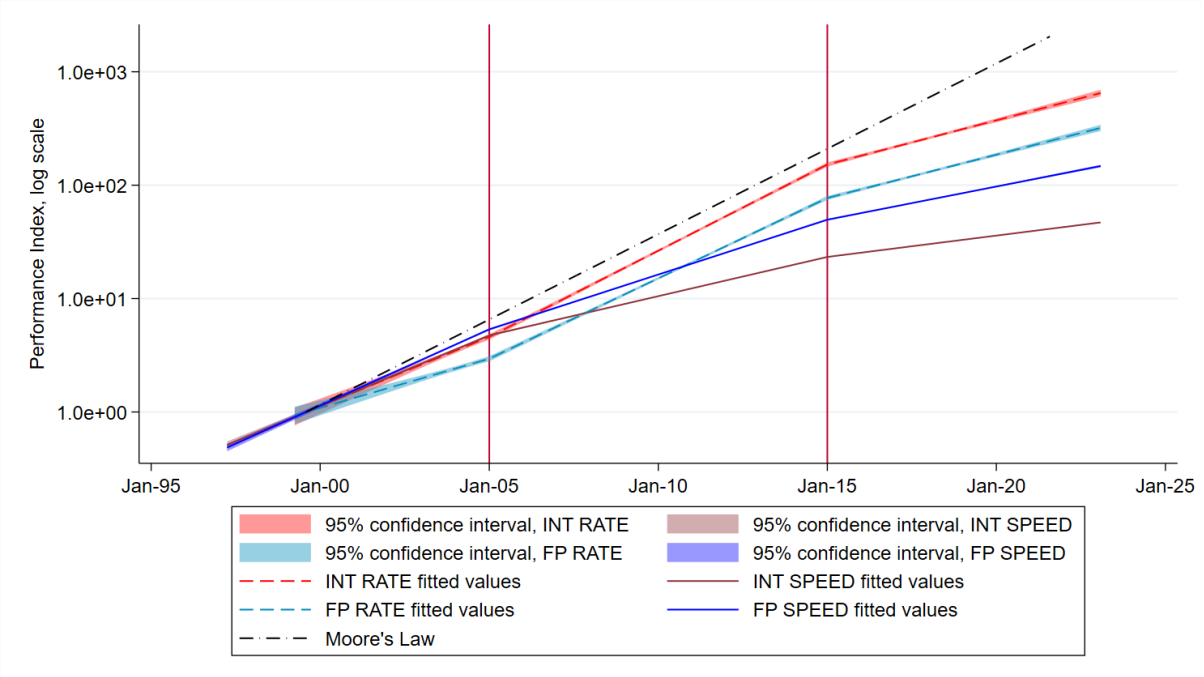
2.3 CPU results

Figure 1 plots the fitted values of the performance regression based on equation (1) for the period 1998-2023 for the four benchmarks. The logged performance index for each benchmark is normalised to 0 at the start of the period. Actual scores and regression tables are shown in Appendix D. Figure 1 demonstrates that hardware performance as measured by SPEC scores has improved by around 2 orders of magnitude since 2000. Rates of improvement for different measures of performance also appear to diverge after 2005: while floating point and integer speed performance improvements slow down around this time, rate scores improve at approximately the same rate as before until slowing down around 2015. Although Moore's Law is not directly about performance, it has been shown in the figure for comparison. Performance improves less rapidly than Moore's Law in all periods.

Turning to the price-performance index, Table 2 shows the regression results, and table 3 shows the corresponding decadal decline rates. An arithmetic mean of the decline rates in each period is shown in the second-last column. Table 3 also compares decadal declines to the official BEA (Bureau of Economic Analysis) Producer Price Index (PPI) for semiconductors (Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks).[14] Figure 3 shows the resulting fitted model for the cost of computing. This is not directly comparable to the Nordhaus series (which we show for reference as Figure 2) as CPU and computer system price trends may diverge.

[14] https://fred.stlouisfed.org/series/PCU334413334413A

Figure 1: Fitted values of the performance index of computations per second (averaged by CPU), by benchmark



Notes: This figure shows fitted values of the relative performance index of computations per second at the CPU level over the date of hardware availability. Moore's Law (stating that the number of transistors in an integrated circuit doubles every two years) is added for comparison. The performance index is constructed using SPEC INT RATE (red), INT SPEED (brown), FP RATE (light blue) and FP SPEED (blue) scores converted to their estimated values on the SPEC CPU 1992 benchmarks. Outliers are excluded from the sample. Fitted values are obtained using equation (1). Fitted values and 95% confidence intervals are shifted such that the first logged fitted value is 0. Actual values are shown in figures D.1, D.2, D.3 and D.4 in the appendix.
Sources: SPEC (https://www.spec.org/benchmarks.html#cpu), authors' calculations.

As in the performance series, there is a statistically significant slowdown in the rate of decline of cost of computation post-2005 and post-2015 for all benchmarks other than INT SPEED. The cost of computation as measured by FP (floating point computation) SPEED scores is estimated to have increased post-2015. As well as the slowdown in performance improvements shown in Figure 1, and so in price declines shown in Figure 3, this may be because of the growing demand for computers with good floating point performance as derived demand for high-performance-computing applications such as AI. There have also been specific adverse supply shocks since 2020 due to the pandemic and other disruptions. The decelerating pace of decline predates the pandemic and associated chip shortages, but may have been exacerbated by these.

Comparing the penultimate to the final column of Table 3, we find that the decline rates in cost of computation nevertheless far exceed that of the official Chip PPI up to 2015, as in Nordhaus (2001, 2007). After 2015, however, the difference between the average

cost of computation decline rates and the official PPI ceases to be statistically significant (p-value 0.7372), with both declining logarithmically by 1.4% per year on average. If the floating point/speed benchmark is omitted, the estimated price declines are consistently much faster than the official PPI estimates but show the same broad trajectory over time. Considering the whole period 2000-2022, the cost of computation declines by approximately two orders of magnitude (almost 3 for rate benchmarks and just over 1 for speed), compared to Nordhaus' 4 between 1990 and 2010.

Figure 2: Nordhaus cost of computation series (2006$)

Notes: This figure depicts Nordhaus' cost of computation series. Y axis is computations per second per 2006 USD, and x axis is date of observation or hardware availability. Nordhaus' computations per second measure is defined such that the speed of manual computation is equal to 1.
Source: Nordhaus (2007) appendix.

Table 2: Regression results for CPU cost of computation series

VARIABLES	(1) INT, RATE	(2) INT, SPEED	(3) FP, RATE	(4) FP, SPEED	(5) PPI
Year	-0.254***	-0.110***	-0.260***	-0.155***	-0.026***
	(0.027)	(0.028)	(0.029)	(0.027)	(0.001)
DUM2005	0.145***	0.048	0.156***	0.071**	0.004***
	(0.035)	(0.037)	(0.037)	(0.036)	(0.001)
DUM2015	0.058***	0.038	0.063***	0.144***	0.009***
	(0.013)	(0.025)	(0.013)	(0.019)	(0.001)
Constant	11.270***	5.083***	11.517***	6.983***	3.196***
	(1.167)	(1.168)	(1.261)	(1.156)	(0.024)
Observations	299	204	281	208	276
R-squared	0.839	0.456	0.815	0.641	0.993
Breusch Pagan statistic	14.63	8.849	7.535	4.660	0.133
p-value	0.000131	0.00293	0.00605	0.0309	0.716

Robust standard errors in parentheses for columns (1)-(4), non-robust for (5)
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table shows regression results for the logged (base 10) cost of computation index based on equation (3). YEAR is calendar year. DUM2005 is equal to 0 up to 2005 and increments by 1 after 2005. DUM2015 is equal to 0 up to 2015 and increments by 1 after 2015. Outliers have been removed. Results with outliers are in table D.3 in the appendix. The Breusch Pagan statistics test the null of homoskedasticity.

Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data. The PPI is the BEA PPI for Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks (https://fred.stlouisfed.org/series/PCU334413334413A).

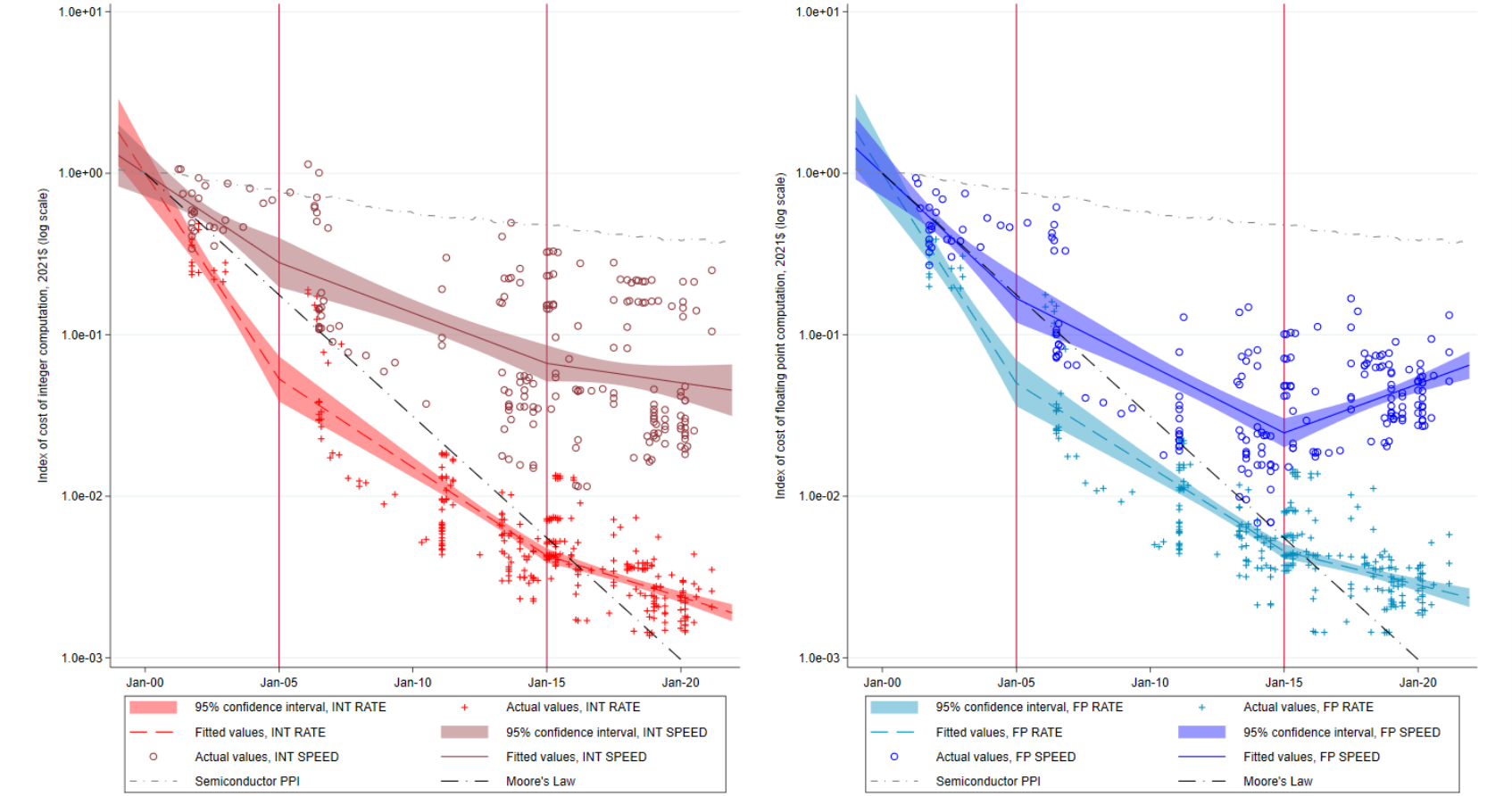
Table 3: Average annual rates of decline in CPU cost of computation in three periods, %

PERIOD	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED	Average	PPI
2000-2005	25.449***	11.046***	25.979***	15.479***	19.488***	2.615***
	(30.773 - 20.125)	(16.398 - 5.694)	(31.705 - 20.253)	(20.767 - 10.190)	(22.201 - 16.776)	(2.725 - 2.506)
2005-2015	10.945***	6.246***	10.412***	8.333***	8.984***	2.220***
	(12.567 - 9.324)	(8.325 - 4.166)	(12.070 - 8.755)	(10.265 - 6.400)	(9.900 - 8.068)	(2.270 - 2.170)
2015-2022	5.114***	2.404	4.158***	-6.065***	1.403**	1.358***
	(6.292 - 3.936)	(5.706 - -0.897)	(5.421 - 2.894)	(-3.870 - -8.260)	(2.484 - 0.322)	(1.442 - 1.273)

Notes: This table shows cumulative sums of the estimated coefficients in Table 2, i.e., average annual logarithmic (base10) decline rates, for each benchmark. The arithmetic mean and BEA chip PPI are displayed in the penultimate and last columns respectively. 95% confidence intervals are in parentheses. Sums may differ from those in Table 2 due to rounding. Outliers have been removed.

Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data. The PPI is the BEA PPI for Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks (https://fred.stlouisfed.org/series/PCU334413334413A).

Figure 3: Cost of CPU computation in 2021$ (2000-2022), by benchmark



Notes: This figure shows the CPU cost of computation in 2021$ by benchmark, with INT (integer computation) scores shown on the left panel and FP (floating point computation) scores shown on the right. Lines are the fitted values from equation (3) and shaded areas are 95% confidence intervals. Fitted values, confidence intervals and the index are shifted such that fitted values are equal to 1 in 1 Jan 2000. Knots are indicated with red lines. Moore's Law and the BEA semiconductor PPI are shown in black and grey respectively for comparison. *Source:* SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data. The Chip PPI is the BEA PPI for Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks (https://fred.stlouisfed.org/series/PCU334413334413A).

3. Extended measures

3.1 Data and methodology

Although there seems no doubt that Moore's Law, as originally stated, and Dennard scaling involve physical limits, there are new types of chip – originally designed for specialised purposes such as graphics and gaming, and now for use in data centres and for large language models (LLMs) – that may be helping to extend progress (and sustain a version of Moore's Law) in computational power. In addition, access to more powerful hardware has become available through cloud computing services.

Graphics processing units (GPUs) were designed to perform the calculations needed for video rendering, but are increasingly used (since around 2010) for the parallel processing needed for machine learning calculations. Tensor Processing Units (TPUs) were introduced by Google specifically to perform matrix calculations required in machine learning, and have been made available for third party use since 2018. A similar trend towards specialisation also exists within GPUs, with the use of lower precision common in AI training runs[15].

Many organisations use machine learning systems, either directly or via other applications or platforms, in the cloud. Cloud computing is the provision of remote access to computing services including storage and software. It is being rapidly adopted by business and other users, offering cheaper and more flexible access to computing than the alternative of in-house investment and provision, and including access to the most up-to-date software and equipment (Byrne *et al.* 2018, Coyle and Nguyen 2018). The two leading providers of cloud services are Amazon Web Services (launched for public use in 2006) and Microsoft's Azure (2010), followed by Google (2011). While there are no definitive statistics on the scale of the cloud market or either the extensive or intensive margin of use, surveys indicate that in 2020 53% of all UK businesses were purchasing some cloud services, compared with an average of 41% in the EU in 2021;[16] while cloud use is growing rapidly.

In this section we introduce these two extensions of available computing technology to our cost of computing calculation in the following way. First, we take the performance (measured in floating points per second or FLOPS) of GPUs since 2006 based on data

[15] In this paper, we measure GPU performance in single-precision FLOPS (rather than half-precision) as trends in price-performance are similar across precision formats. See https://epochai.org/blog/trends-in-gpu-price-performance for details.
[16] https://goingdigital.oecd.org/indicator/25

from Epoch AI,[17] and invert their FLOPS per $ series to get a cost per FLOPS. We introduce this to our earlier CPU price-performance series starting in 1999 to produce a combined index. But as we lack quantity data for use of the 'new good' of GPUs we need to make an assumption about the speed of adoption and the relative weights to assign to the cost of CPU series and the cost of GPUs. To illustrate the possible range, we construct three aggregate indices: one 'low-growth' index where we assume that the proportion of GPU use was 1% in 2006 and has grown by 10% a year, one 'high-growth' index where we assume that the proportion of GPU use has grown by 30% a year, and a 'Top500' index where we use a linear trend of GPU and CPU performance shares in the Top500 index for supercomputers as weights.[18] At the end of the sample, the GPU weight reaches 5% in the first case, 50% in the second, and 45% in the Top500 index. Secondly, we repeat this using the price per performance of cloud computing services, as for most users access to machine learning systems will occur through the use of cloud services. We extend the series constructed by Coyle and Nguyen (2018), who found a steep decline in this price index when Microsoft entered the market in 2014, and a levelling off thereafter.[19]

Our empirical framework for price-performance is similar to earlier sections. As the series starts in 2006, the specification for GPUs is as follows:

$$log(price_perf_t) = \beta_0 + \beta_1 Year_t + \beta_2 DUM2015_t + \varepsilon_t \quad (5)$$

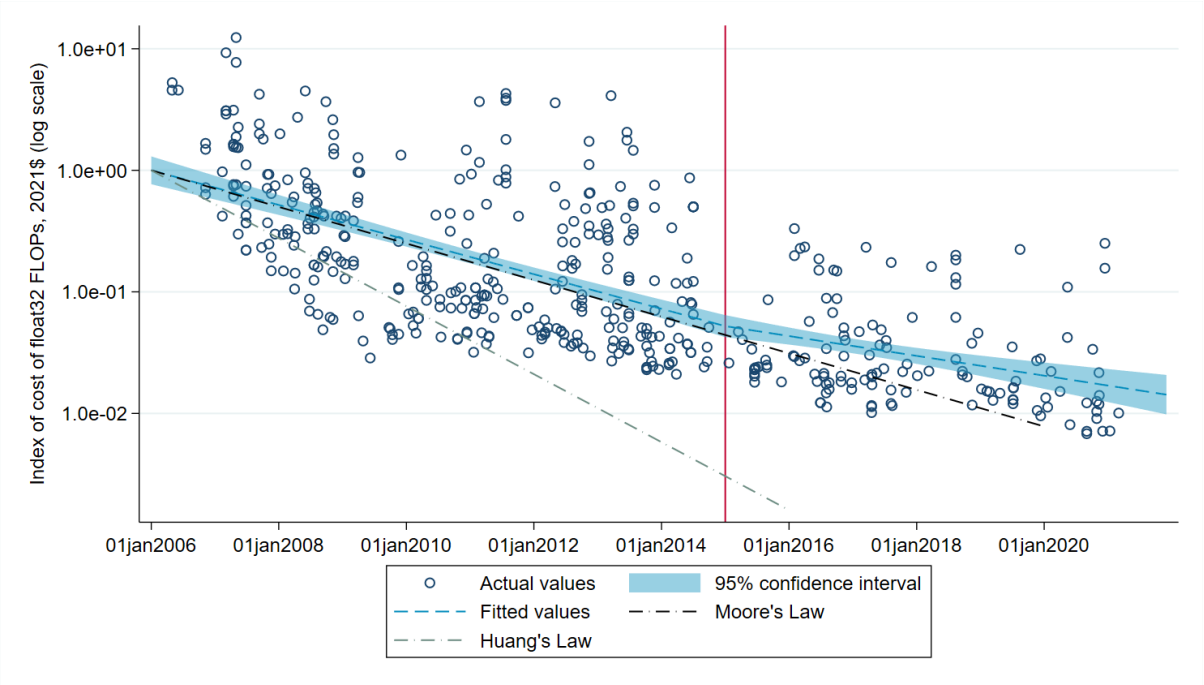
where $DUM2015_t$ is equal to 0 up to 2015 and increments by 1 each year after 2015. For the combined chip index, the specification is the same as in (3), where $price_perf_t$ is defined as the weighted average of CPU and GPU price-performance.

3.2 Results: GPUs and the Cloud

Table 4 shows the regression results for GPUs, where performance is measured in single-precision floating-point format (FP32) and prices are in 2021$. Figure 4 then plots the results. Although neither directly measure price-performance, Moore's Law and Huang's Law (GPU performance doubling approximately every 1.08 years) are shown for comparison. As in the CPU data, there is a statistically significant slowdown in the decline in cost of computation post-2015. However, the logarithmic rate of decline still averages around 8% per year for GPUs in this period, while for CPUs it is around 1.4%.

[17] https://epochai.org/blog/trends-in-gpu-price-performance. The dataset used by EpochAI in this blog post originates from Median Group (2018) and Sun *et al.* (2019).
[18] https://www.top500.org/statistics/overtime/
[19] In the UK, regulators are launching market inquiries on competition grounds into both cloud computing and AI foundation models (https://www.ofcom.org.uk/consultations-and-statements/category-2/cloud-services-market-study; https://www.gov.uk/cma-cases/ai-foundation-models-initial-review)

Figure 4: Cost of GPU computation in 2021$ (2006-2022)



Notes: This figure shows GPU log cost of computation in 2021$. The dashed lines are fitted values from equation (5) and shaded areas are 95% confidence intervals. Moore's Law (doubling every two years) and Huang's Law (performance doubling every 1.08 years) are shown in the black and teal dot-dashed lines respectively. All values are shifted such that the fitted value is 1 in January 2006. 1 Jan 2015 (the position of the knot) is shown in red.
Source: GPU price-performance data comes from Epoch AI (https://epochai.org/blog/trends-in-gpu-price-performance), which uses data from Median Group (2018) and Sun *et al.* (2019).

Table 4: Regression results for GPU data, and corresponding decline rates

	(1)	Decline rate
VARIABLES	GPU Index	(% per year)
Year	-0.142***	14.246***
	(0.010)	(16.183 - 12.309)
DUM2015	0.061***	8.147***
	(0.023)	(11.300 - 4.994)
Constant	5.894***	
	(0.508)	
Observations	470	
R-squared	0.474	
Breusch Pagan statistic	3.718	
p-value	0.0538	

Robust standard errors in parentheses in left column, 95% confidence intervals in right column
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table shows spline regression results for the GPU cost of computation (measured in FP32 units) on Year and Time variables based on equation (5). *Year* is calendar year, *DUM2015* is equal to 0 up to 2015

and increments by 1 after 2015. The second column shows the average annual decline for each period, calculated as the cumulative sum of the regression coefficients up to the date.
Source: GPU price-performance data comes from Epoch AI (https://epochai.org/blog/trends-in-gpu-price-performance), which uses data from Median Group (2018) and Sun *et al.* (2019).

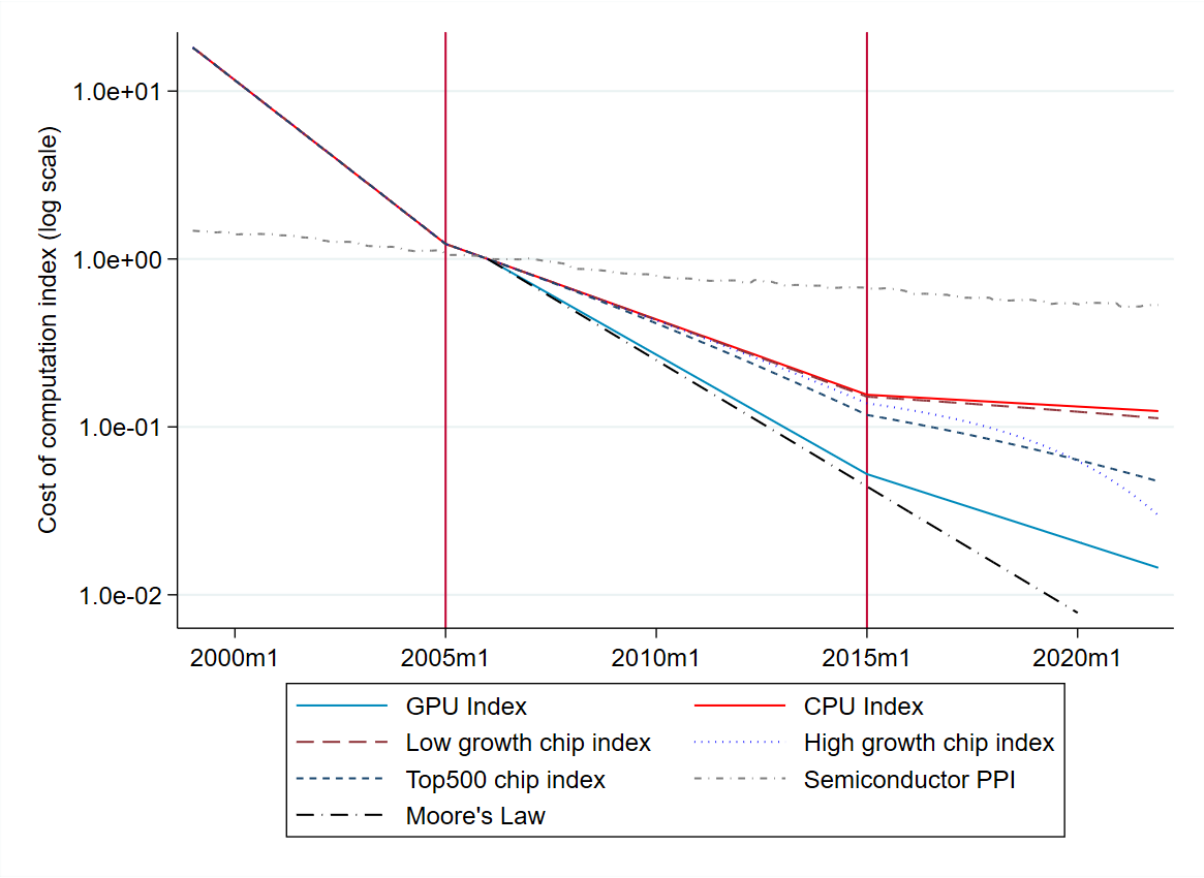
Table 5: Average annual decline rates for combined chip indices

PERIOD	CPU	GPU	Low-growth chip index	High-growth chip index	Top500 chip index	PPI
2000-2005	19.494	-	19.468	19.548	19.087	2.615
2005-2015	8.986	14.234	9.086	9.046	10.115	2.220
2015-2022	1.404	8.069	1.853	8.290	5.904	1.358

Notes: This table shows average annual decline rates from a regression based on equation 3 for the CPU, GPU and combined cost of computation indices. The combined indices were created by taking a weighted average of the GPU and CPU indices. 'Low-growth' denotes an index where we assume that the proportion of GPU use was 1% in 2006 and increased by 10% each year, 'high-growth' where we assume that the proportion of GPU was 1% in 2006 and increased 30% per year, 'Top500' where we use a linear trend of computation shares in the Top500 list as weights. Decline rates are calculated by taking the cumulative sums of the regression coefficients up to and including the period. These may differ from previous regressions since these data are collapsed such that there is only one observation for each index in a month. The R^2 and standard errors are not shown because fitted values from earlier regressions are used to impute missing values in this regression, hence they are misleading indicators of goodness of fit.
Source: as in Table 2 and Table 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

The magnitude of the difference between the CPU index and combined chip index with low-growth assumptions is fairly small: both decline at a logarithmic rate of approximately 9 per year, 2005-2015, while in the period 2015-2022 the CPU index declines by 1.4 per year and the combined index by 1.9. Assuming a higher, 30% annual growth rate produces an annual logarithmic decline in the cost of computation of 8.3 in the period 2015-2022, approximately six times as large as the decline rate in the CPU index and official BEA semiconductor index. In between these is the Top500 index, which declines at a rate of 5.9 per annum in the 2015-2022 period, or approximately four times the rate of the CPU and BEA indices.

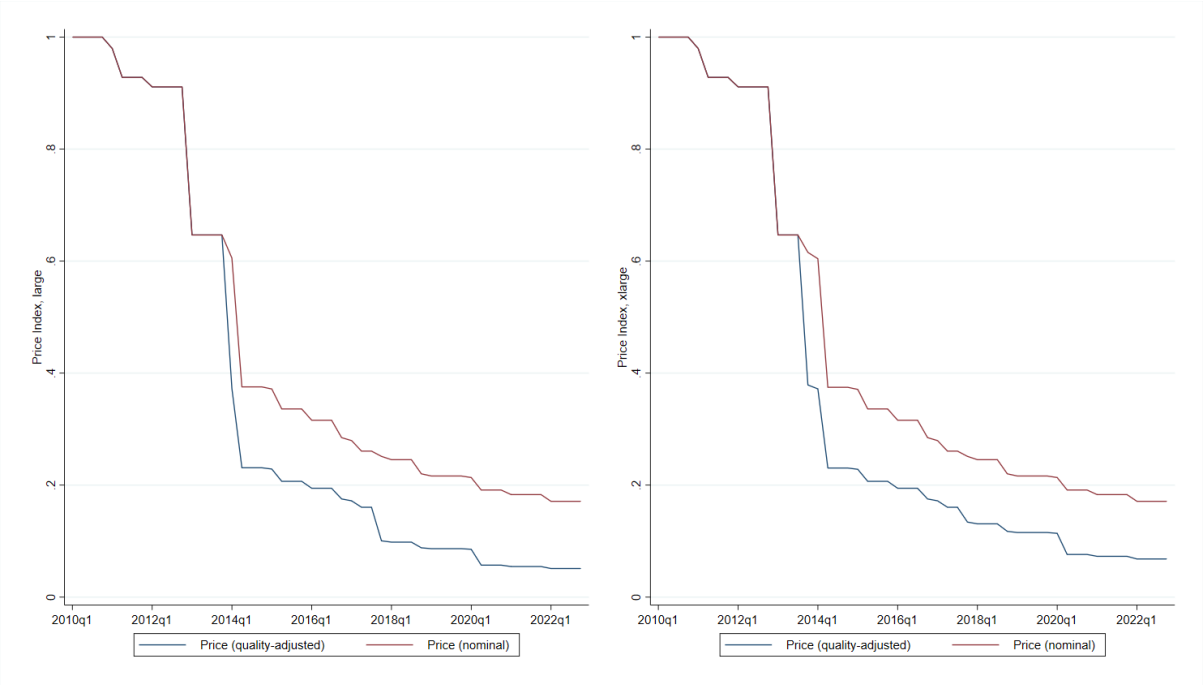
Figure 5: Combined cost of computation Chip Index in 2021$ (1999-2022)



Notes: This figure shows fitted values for the regressions of CPU (red), GPU (blue) and combined cost of computation indices, as well as the semiconductor PPI (grey). The combined indices were created by converting the fitted values of the CPU and GPU cost of computation indices to the same scale and taking a weighted average of each. For the 'Low-growth' (maroon) index, we assume that the proportion of GPU use was 1% in 2006 and increased by 10% each year. 'High-growth' (blue) denotes an index where we assume that the proportion of GPU was 1% in 2006 and increased 30% per year. In the 'Top500' (navy) index we use smoothed shares of computation in the Top500 lists as weights. Moore's Law is shown in black and the semiconductor PPI in grey. 1 Jan 2005 and 1 Jan 2015 (the positions of the knots) are denoted by the red vertical lines. The value of each index has been set to 1 in 1 Jan 2006.
Source: As in Table 2 and Table 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

We also update Coyle and Nguyen (2018) figures for prices of cloud use in the UK. Figure 6 shows the cloud price indices for Amazon 'large' and 'xlarge' instances – two popular classes of cloud service. The nominal price in each case is adjusted for the performance of the underlying hardware. Both figures show inflation-adjusted declines in cloud prices, most rapid around late 2013, when the UK market saw new entry, as well as a divergence in the unadjusted and quality-adjusted indices around the same time. The decline has continued but slowed since the mid-2010s.

Figure 6: Cloud price index, Linux (Q1-2010 – Q4-2022)



Notes: This figure shows nominal and quality-adjusted prices of AWS EC2 large and xlarge instances for Linux. Prices are hourly on-demand rates deflated by the aggregate price index. In blue, prices are also quality-adjusted for performance improvements.
Source: AWS API price lists (https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/price-changes.html) for prices, AWS press release for performance improvements, and pre-2019 data from Coyle and Nguyen (2018).

4. Discussion: implications for productivity

In this paper we have shown that taking account of continuing developments in computing technologies mean that the decline in price and hence growth in output and productivity in the computer sector have likely been understated in official figures, although we also find that the pace of decline in the quality-adjusted price of computation has slowed since the mid-2010s.

The dramatic engineering-enabled declines in the price of computation have been accompanied by substantial increases in usage. As with all general purpose technologies, this occurs first in the relevant sector itself, and then diffuses throughout the wider economy. At present there is scant data on AI (direct or via the cloud) usage or its purpose. Furthermore, both the network technology and software (for example in the form of machine learning applications such as generative models) are advancing extremely rapidly. According to the AI Index Report 2023,[20] AI capabilities are most likely to have

[20] https://aiindex.stanford.edu/report/

been previously adopted in robotic process automation, computer vision, virtual agents and natural language applications, but applications are growing rapidly in areas such as customer analytics and new AI-based products (Maslej *et al.* 2023). One early projection (Eloundou *et al.* 2023) suggests that four-fifths of the US workforce will have at least 10% of their tasks affected by new AI-based applications, and about one fifth will see half their tasks affected.

The fact of technological progress does not in itself ensure adoption and the consequent economic impacts, however. For example, Frey & Osborne (2017) estimated that 47% of US jobs (not tasks) were at high risk of automation "over some unspecified number of years," but evidence to date suggests that it is a minority of firms that are able to adopt digital technologies and use them to improve productivity (Brynjolfsson *et al.* 2021, Cathles *et al.* 2020, Coyle *et al.* 2022). The use of digital technologies to create economic value involves a bundle of products and capabilities, including computational hardware but also software, likely through cloud services (themselves bundling data centres, compute, communications networks and software including algorithms), and also specific human and organisational capital. Furthermore, although the capabilities of new AI systems are striking, the compute requirement for their development and training of has been increasing, and at an accelerating pace for the recent foundation models (Sevilla *et al.* 2022). On the other hand, algorithmic efficiency has been improving for a decade (Hernandez and Brown 2020).

In this paper we have addressed one potential element of the explanation. However, we found a dramatic cost decline in computation and cloud services, measured in several ways, which did not decelerate until a decade after the broader productivity growth slowdown. Other work (Fleming 2022) has also pointed to faster price declines than currently shown by official software price indices, by taking better account of changes in the production and business use of software leads. Fleming's software price index incorporates the price of cloud use by US businesses, which has been broadly flat in recent years, and also the growing substitution to free open source software. The pace of decline in his software price index accelerates post-2010 taking into account substitution to free software. It is therefore possible that official producer price indices for computer software and services understate the declines in prices firms actually pay, and similarly understate the sector's productivity growth. How different measures would ultimately impact estimates of productivity growth elsewhere depends on supply linkages.

The measured productivity slowdown in advanced economies since the mid-2000s on the face of it suggests that the diffusion of this technology bundle through the economy

has been limited in its impact. Indeed, there are broader issues to consider in understanding productivity outcomes.

One is what appears at first sight to be a further measurement issue. In their work on telecommunications services, Abdirahman *et al.* (2020,2022) show that the choice of weights makes a substantial difference to the producer output price index. The closer the weights on the service components are to volume weights, the faster the decline in the index; in the limit a unit value index shows a decline of about 90% in the price per byte transmitted in the UK in the decade from 2010. However, the choice of weights is non-trivial, as it raises a fundamental question about where to attribute the value being created in the economy by the greatly expanded use of ICTs. Volume weights (in the limit a unit value index) seem inappropriate because bytes of data do not all have equal value; consumers pay different amounts per byte depending on the services, albeit substituting over time to lower price-per-byte services. However, revenue weights do not seem the best choice either, partly because these reflect accounting choices by telecoms companies, but also because their (improved) cables and masts are not the main source of economic value creation. None of the measured elements of communications services – bytes of data, network infrastructure, access charges – reflects much of the economic value users gain from usage. Rather, this value stems from the derived demand for downstream content services or more broadly the things people can do with the technology bundle, which is a form of infrastructure. There is a value assignment problem just as with any other form of infrastructure. This is equally true of access to computation, albeit further complicated by the absence of data on usage rates.

What is clear is that there have been very substantial declines for many decades in the price of computation. Although it has slowed, the timing of the slowdown (mid-2010s) does not tally well with the timing of the broader productivity slowdown (mid-2000s). This suggests revisiting official price indices to better understand the productivity performance of the computer sector.

More broadly, the widespread evidence of growing use of computation through the economy – as would be expected for an activity whose price was falling so fast – raises profound questions about value creation and capture, beyond the scope of this paper. This is a networked, multi-device economy with a multitude of apps and platforms, and in which many advances are provided to users at zero price. Computational power is not the binding constraint on growth; rather (along with natural resources), the limitation, and solution to the productivity puzzle, is likely to lie in organisational constraints and the need for co-invention.

References

Abdirahman, M., Coyle, D., Heys, R., & Stewart, W. (2020). A Comparison of Deflators for Telecommunications Services Output. *Economie et Statistique / Economics and Statistics*, *517-518–519*, 103–122. https://doi.org/10.24187/ecostat.2020.517t.2017

Abdirahman, M., Coyle, D., Heys, R., & Stewart, W. (2022). Telecoms Deflators: A Story of Volume and Revenue Weights. *Economie et Statistique / Economics and Statistics*, *530–31*, 43–59. https://doi.org/10.24187/ecostat.2022.530.2063

Bessen, J. (2022). *The New Goliaths*. Yale University Press.

Biagi, F. (2013). ICT and Productivity: A Review of the Literature. *JRC Working Papers on Digital Economy*, Article 2013–09. https://ideas.repec.org//p/ipt/decwpa/2013-09.html

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, *34*(3), 279–298. https://doi.org/10.1016/S0167-9473(99)00101-2

Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are Ideas Getting Harder to Find? *American Economic Review*, *110*(4), 1104–1144. https://doi.org/10.1257/aer.20180338

Brynjolfsson, E., Collis, A., Diewert, W. E., Eggers, F., & Fox, K. J. (2020). Measuring the Impact of Free Goods on Real Household Consumption. *AEA Papers and Proceedings*, *110*, 25–30. https://doi.org/10.1257/pandp.20201054

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *American Economic Journal: Macroeconomics*, *13*(1), 333–372. https://doi.org/10.1257/mac.20180386

Byrne, D. M. (2022). *The Digital Economy and Productivity* (SSRN Scholarly Paper No. 4158722). https://doi.org/10.17016/FEDS.2022.038

Byrne, D. M., Corrado, C. A., & Sichel, D. (2018). *The Rise of Cloud Computing: Minding Your P'S, Q's and K's*. NBER. https://ssrn.com/abstract=3274414

Byrne, D. M., Hamins-Puertolas, A., & Harnish, M. M. (2023). *Transistors all the Way Down: Viability of Direct Volume Measurement (and Price Indexes) for Semiconductors*.

Cardona, M., Kretschmer, T., & Strobel, T. (2013). ICT and productivity: Conclusions from the empirical literature. *Information Economics and Policy*, *25*(3), 109–125. https://doi.org/10.1016/j.infoecopol.2012.12.002

Cathles, A., Nayyar, G., & Rückert, D. (2020). *Digital technologies and firm performance: Evidence from Europe*. http://dx.doi.org/10.2867/3688

Coyle, D., Lind, K., Nguyen, D., & Tong, M. (2022). Are digital-using UK firms more productive? *Economic Statistics Centre of Excellence (ESCoE) Discussion Papers*. ESCoE DP-2022-06. https://escoe-website.s3.amazonaws.com/wp-content/uploads/2022/03/10100111/ESCoE-DP-2022-06.pdf

Coyle, D., & Mei, J.-C. (2023). Diagnosing the UK productivity slowdown: Which sectors matter and why? *Economica*, 1–38. https://doi.org/10.1111/ecca.12459

Coyle, D., & Nguyen, D. (2023). *Free Digital Products and Aggregate Economic Measurement*. Economie et Statistique / Economics and Statistics.

Coyle, D., & Nguyen, D. (2018). Cloud Computing and National Accounting. *Economic Statistics Centre of Excellence (ESCoE) Discussion Papers*. ESCoE DP-2018-19. https://escoe-website.s3.amazonaws.com/wp-content/uploads/2020/07/13163356/ESCoE-DP-2018-19.pdf

Draca, M., Sadun, R., & Van Reenen, J. (2007). Productivity and ICTs: A review of the evidence. In R. Mansell, C. Avgerou, D. Quah, & R. Silverstone (Eds.), *The Oxford Handbook of Information and Communication Technologies* (pp. 100–147). Oxford University Press. http://www.oup.com/uk/catalogue/?ci=9780199266234

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

Flamm, K. (2017). Has Moore's Law Been Repealed? An Economist's Perspective. *Computing in Science & Engineering*, *19*(2), 29–40. https://doi.org/10.1109/MCSE.2017.30

Fleming, M. (2023). *Enterprise Information and Communications Technology Software Pricing and Developer Productivity Measurement*.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1), 1–67.

Goodridge, P., & Haskel, J. (2023). Accounting for the slowdown in UK innovation and productivity. *Economica*, 1–33. https://doi.org/10.1111/ecca.12468

Gordon, R. (2017). *The Rise and Fall of American Growth*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691175805/the-rise-and-fall-of-american-growth

Hernandez, D., & Brown, T. B. (2020). *Measuring the Algorithmic Efficiency of Neural Networks* (arXiv:2005.04305). arXiv. https://doi.org/10.48550/arXiv.2005.04305

[dataset] Hobbhahn, M., & Tamay, B. (2022, June 27). *Trends in GPU price-performance*. Epoch. https://epochai.org/blog/trends-in-gpu-price-performance

Johnsson, L., & Netzer, G. (2016). The impact of Moore's Law and loss of Dennard scaling: Are DSP SoCs an energy efficient alternative to x86 SoCs? *Journal of Physics: Conference Series*, *762*(1), 012022. https://doi.org/10.1088/1742-6596/762/1/012022

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, *74*(2), 132–157.

Leiserson, C. E., Thompson, N. C., Emer, J. S., Kuszmaul, B. C., Lampson, B. W., Sanchez, D., & Schardl, T. B. (2020). There's plenty of room at the Top: What will drive computer performance after Moore's law? *Science*, *368*(6495), eaam9744. https://doi.org/10.1126/science.aam9744

Lotfi-Kamran, P., & Sarbazi-Azad, H. (2018). Chapter One—Dark Silicon and the History of Computing. In *Dark Silicon and Future On-chip Systems* (Vol. 110). Elsevier. https://doi.org/10.1016/bs.adcom.2018.03.001

[dataset] Maltinsky, B., Gallagher, J., & Taylor, J. (2018). *Feasibility of Training an AGI using Deep RL: A Very Rough Estimate*. Median Group. http://mediangroup.org/rough_agi.html

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). *The AI Index 2023 Annual Report* (AI Index Annual Report). AI Index Steering Committee, Institute for Human-Centered AI. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

Munafo, R. P. (2022, April 2). *The SPEC Benchmarks at MROB*. The Published Data of Robert Munafo. https://mrob.com/pub/comp/benchmarks/spec.html

Nordhaus, W. D. (2001). *The Progress of Computing* (SSRN Scholarly Paper No. 285168). https://papers.ssrn.com/abstract=285168

Nordhaus, W. D. (2007). Two Centuries of Productivity Growth in Computing. *The Journal of Economic History*, *67*(1), 128–159. https://doi.org/10.1017/S0022050707000058

[dataset] PassMark® Software Pty Ltd. (2023). *PassMark Software—CPU Benchmarks*. https://www.cpubenchmark.net

Poquiz, J. L. (2023). *How much is your video call worth? Measuring the value of free digital goods*.

Ramu, A. (2018, August 14). The End of Moore's Law. And The Coming Computing Renaissance. *A.Capital Ventures*. https://acapital.com/2018/08/the-end-of-moores-law-and-the-coming-computing-renaissance

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning* (arXiv:2202.05924). arXiv. http://arxiv.org/abs/2202.05924

Shalf, J. (2020). The future of computing beyond Moore's Law. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *378*(2166), 20190061. https://doi.org/10.1098/rsta.2019.0061

[dataset] Shvets, G. (2022). *CPU-World: Microprocessor news, benchmarks, information and pictures*. CPUWorld. https://www.cpu-world.com/

Solow, R. (1987). We'd better watch out. *New York Times Book Review*. https://www.semanticscholar.org/paper/We%E2%80%99d-better-watch-out-Solow/cef149b3dbdaa85f74b114c2c7832982f23bcbf0

Sun, Y., Agostini, N. B., Dong, S., & Kaeli, D. (2020). *Summarizing CPU and GPU Design Trends with Product Data* (arXiv:1911.11313). arXiv. https://doi.org/10.48550/arXiv.1911.11313

[dataset] TechPowerUp. (2023, June 14). *TechPowerUp*. TechPowerUp. https://www.techpowerup.com/

Theis, T. N., & Wong, H.-S. P. (2017). The End of Moore's Law: A New Beginning for Information Technology. *Computing in Science & Engineering*, *19*(2), 41–50. https://doi.org/10.1109/MCSE.2017.29

Van Ark, B. (2016). The Productivity Paradox of the New Digital Economy. *International Productivity Monitor*, *31*, 3–18.

Weber, S. (2010). Bacon: An Effective way to Detect Outliers in Multivariate Data Using Stata (and Mata). *The Stata Journal*, *10*(3), 331–338. https://doi.org/10.1177/1536867X1001000302

Appendix

A. Glossary of terms

Central Processing Unit (CPU): A set of electronic circuitry that interprets, regulates and executes instructions. A CPU is typically contained on an integrated circuit chip or microprocessor.

Chip/Integrated Circuit/Microprocessor: A set of electronic circuits (including transistors) contained within a flat piece of semiconductor material (the chip).

Clock Speed/Rate: The frequency of pulses (in MHz) released by the clock generator of a processor, used to synchronise its components. Clock speed is indicative of the number of operations the CPU can execute per second.

Cores: A processing unit within a CPU. Multi-core processors contain multiple cores on the same chip.

Floating point computing: A type of computing involving floating point data, i.e., a data type representing real numbers approximately using significands and an integer exponent. Applications of floating-point computing are those that require higher precision than integer computing, such as financial and scientific modelling, weather forecasting, image manipulation and machine learning.

Integer computing: A type of computing involving integer data, i.e., a data type representing mathematical integers. Applications of integer computing include tasks that do not require decimal precision such as database management, route planning, video compression, word processing and file conversion.

Moore's Law: The empirical regularity that the number of transistors on a chip doubles every 18-24 months.

Throughput (RATE): Work done in a unit of time. In the SPEC benchmarks, RATE scores are throughput based, and measure the number of copies of a particular benchmark run by a computer relative to a reference machine. Higher scores mean a higher throughput.

Time (SPEED): In the SPEC benchmarks, the time taken to run a single copy of a particular benchmark relative to a reference machine. Higher scores mean less time is taken.

Parallelism: A type of computing where multiple computations (often sub-components of a larger task) are performed at the same time, usually through the use of multiple cores.

B. Detailed methodology

CPU performance series

For performance data, we use benchmark results from SPEC, a nonprofit consortium of 22 major computer vendors who aim "to provide the industry with a realistic yardstick to measure the performance of advanced computer systems"[21]. Data was obtained from the SPEC website for the SPEC 2000, SPEC 2006 and SPEC 2017 CPU benchmark versions. SPEC CPU benchmarks are developed from end-user applications, with the applications tested varying across versions to reflect the changing demand for tasks. A comparison of benchmarks in the 2006 and 2017 versions are presented in table B.1. Scores for each benchmark suite are geometric means of equally-weighted scores in each programme.

Table B.1: A comparison of SPEC benchmarks

CINT2006	CFP2006	CINT2017	CFP2017
PERL programming language	Fluid Dynamics	Perl interpreter	Explosion modelling
Compression	Quantum chemistry	GNU C compiler	Physics: relativity
C Compiler	Physics: quantum Chromodynamics	Route planning	Molecular dynamics
Artificial Intelligence: go	Physics/CFD	Discrete event simulation - computer network	Biomedical imaging: optical tomography with finite elements
Search gene sequence	Biochemistry/molecular dynamics	XML to HTML conversion via XSLT	Ray tracing
Artificial Intelligence: chess	Physics/general relativity	Video compression	Fluid dynamics
Physics: quantum computing	Fluid dynamics	Artificial intelligence: alpha-beta tree search (Chess)	Weather forecasting
Video compression	Biology/molecular dynamics	Artificial intelligence: Monte Carlo tree search (Go)	3D rendering and animation
Discrete event simulation	Finite element analysis	Artificial intelligence: recursive solution generator (Sudoku)	Atmosphere modelling

[21] http://jimgray.azurewebsites.net/benchmarkhandbook/chapter9.pdf

Path-finding Algorithms	Linear Programming, Optimisation	General data compression	Wide-scale ocean modelling (climate level)
XML Processing	Image Ray-tracing		Image manipulation
	Structural mechanics		Molecular dynamics
	Computational electromagnetics		Computational electromagnetics
	Quantum chemistry		Regional ocean modelling
	Fluid dynamics		
	Weather prediction		
	Speech recognition		

Source: SPEC

Each SPEC score is relative to a reference machine, which differ by version. A score of 100 in SPEC92, for example, means that the tested computer system performs 100 times faster than the Digital VAX 11/780. Hence SPEC scores from different versions are not on the same scale. To scale them appropriately, we use the conversion factors shown in table B.2 following the methodology outlined on the MROB website.

The conversion factors were obtained by identifying a computer system present in multiple versions of the benchmark suite. For example, the Cisco UCS C220 M5 was present in both the 2017 and 2006 versions of CPU INT SPEED, with a score of 9.86 in 2017 and 83 in 2006. Thus, to convert to 2006 scores, all 2017 INT SPEED scores are first divided by 9.86 to make all 2017 scores relative to the CISCO C220 M5, and then multiplied by 83 such that the CISCO C220 M5 is given the same score in 2017 as in 2006. Since the ratios of scores are unaffected, relative performance between computers is preserved in the conversion. For example, if a computer scores 19.72 in 2017, its relative performance is double that of CISCO C220 M5, and it will receive a score of 2*83 = 166 in 2006.

To depict speed and rate scores on the same scale, further conversions are needed. While speed and rate scores are fundamentally different measures of performance, there is a theoretical relationship between them. The formula for SPEC CPU1992 rate scores is as follows, where 056.ear is the name of the program to which scores are normalised:

$$SPECrate(program) = N * [T_{ref}(program)\ /T_{ref}(056.ear)]\ * [604800[T_{SUT}(program)]$$

where

N = number of copies run concurrently
$Tref(program)$ = time to run program on the reference machine, a VAX 11/780
$Tref(056.ear)$ = time to run 056.ear on the reference machine = 25500 seconds
604800 = number of seconds in a week
$TSUT(program)$ = time to finish last concurrent copy on system under test
SUT = system under test

This simplifies to:

$$SPECrateInt92 = 23.72 * P * SPECInt92$$
$$SPECrateFP92 = 23.72 * P * SPECFP92$$

where P is the number of cores. 23.72 is the rate score for the reference machine, the VAX 11/780. Note that for multicore systems, this conversion only gives theoretical maximum performance on the rate benchmarks, as it assumes that doubling the processors doubles rate performance, which is rarely achieved as processors typically share memory. For the same reason, it is likely to *underestimate* speed scores converted from rate scores.

We do not convert all rate scores to rate scores, as both measures of performance are informative, and as discussed above, such conversions are likely to be lossy for multi-core machines. Instead, we convert *single-core* rate scores to speed scores by dividing them by 23.72 so that single-core rate scores can be directly compared to speed scores, and then divide multi-core rate scores by 23.72 to preserve relative performance between single and multi-core rate scores. Note however that multi-core rate scores are not directly comparable with single-core speed scores, although (and most importantly for this analysis) rates of performance improvements can be compared. As scores are 1992 scores, they can be combined with Nordhaus' series, as shown in figure D.3.

It should be noted that these conversions are limited in several ways, and as such the 1992 equivalents to scores on more modern benchmarks are estimates[22]. Firstly, it assumes that relative performance is constant across benchmarks, which may not be the case due to changes in the benchmarks affecting memory usage and run times. This

[22] From the SPEC CPU2006 Q&A:
Q19: Is there a way to translate SPEC CPU2000 results to SPEC CPU2006 results or vice versa?

There is no formula for converting CPU2000 results to CPU2006 results and vice versa; they are different products. There probably will be some correlation between CPU2000 and CPU2006 results (i.e., machines with higher CPU2000 results often will have higher CPU2006 results), but there is no universal formula for all systems.

SPEC encourages SPEC licensees to publish CPU2006 numbers on older platforms to provide a historical perspective on performance.

problem worsens as the comparison period widens, as the performance gap between older computers and newer computers tested in new benchmarks is likely to shrink when converting backwards as computing tasks become less memory intensive. Secondly, perfect matches were not available in all years. The Sun Enterprise 3500/4500, for example, has different operating systems in different versions.

Table B.2: Conversion factors and computer systems for benchmark versions

	INT, SPEED	INT, RATE	FP, SPEED	FP, RATE
2017->2006	83/9.86	1520/159	153/116	1260/174
	Cisco UCS C220 M5	Cisco UCS C220 M5	Cisco UCS C220 M5	Cisco UCS C220 M5
2006->2000	116	162 / 67.2	147	104 / 53.3
	Sun Ultra Enterprise 2	Servidor Ituatec ZX440	Sun Ultra Enterprise 2	Servidor Ituatec ZX440
2000->1992	50.20/8.264	1030/18.200001 *51394/449	60.20/7.752	1708/19 *73044/493
	Sun Ultra 10 300 MHz	Sun Enterprise 3500/4500 and Alpha Server 8200 5/350	Sun Ultra 10 300 MHz	Sun Enterprise 3500/4500 and Alpha Server 8200 5/350

Source: The MROB website

SPEC scores are assigned to computer systems, which are combinations of processors, memory, compilers, operating systems and other attributes. To obtain chip-level performance, we collapse the data by CPU, number of cores, number of chips, clock speed, benchmark and date of hardware availability.

Price-performance series

Release price data was obtained from Passmark, Tech PowerUp, and CPUWorld. Data was obtained from Passmark by scraping old versions of the 'CPU Mega Page' in each year using WebArchive. Due to the occasional lack of availability of the archived version of the page, data could not be obtained at exact yearly intervals. The Passmark data was matched to the CPU performance data based on the processor name, number of chips and year of hardware availability, and prices for N-chip systems were calculated as N x the price of the single-chip system. TechPowerUp release price data was scraped directly from the website, while CPUWorld data was entered manually for a small number of CPUs as their terms and conditions prohibit scraping. TechPowerUp and CPUWorld data was matched to performance data based on the processor name and number of chips, with priced for multi-chip systems calculated as before. To obtain real prices, we use the US GDP aggregate price index and convert to 2021 values.

To obtain the capital cost of computation per second, we follow Nordhaus (2007) in using the standard user cost of capital formula with a 10% a year constant real interest rate, exponential depreciation at 10% per year, and 2000 hours per year of usage.

C. Robustness checks

C1 Varying the interest rate

In constructing the capital cost per second variable, we assume a constant real interest rate and depreciation rate of 10% per year. Varying these initial values and holding them constant does not affect the rates of decline. The results are sensitive, however, to changes in interest or depreciation rates: declining interest or depreciation rates would steepen the decline rate in cost of computation, all else equal. In this section we explore this by assuming an initial interest rate of 10% that declines by 5% per year to 3% in October 2021.

The results for CPUs are displayed in table C1.1. The decline rates are steeper with declining interest rates and this difference is statistically significant ($p < 0.0001$ for each period). However, the difference is fairly small: e.g., the decline in the 2015-2022 period with constant interest rates is 1.4 per year, while with declining interest rates it is two per year. The finding that there is a slowdown post-2015 is also robust. Similarly, for the aggregate indices shown in table C1.2, each cost of computation index is higher in the post-2015 period than it is with a constant interest rate, but lower post-2015 than pre-2015.

Table C1.1: CPU cost of computation decline rates with declining interest rate

PERIOD	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED	Average
2000-2005	26.547***	12.103***	27.077***	16.544***	20.568***
	(31.880 - 21.214)	(17.459 - 6.746)	(32.812 - 21.342)	(21.839 - 11.249)	(23.284 - 17.851)
2005-2014	11.731***	7.044***	11.198***	9.130***	9.776***
	(13.355 - 10.107)	(9.125 - 4.962)	(12.859 - 9.538)	(11.064 - 7.195)	(10.693 - 8.858)
2015-2022	5.720***	3.010*	4.764***	-5.463***	2.008***
	(6.899 - 4.541)	(6.313 - - 0.293)	(6.029 - 3.499)	(-3.266 - - 7.659)	(3.090 - 0.926)

Notes: This table shows average annual logarithmic (base10) decline rates in the cost of computation for each CPU benchmark, with interest rates declining by 5% per year. The arithmetic mean and BEA chip PPI are displayed in the penultimate and last columns respectively. 95% confidence intervals are in parentheses. Outliers are removed. Decline rates in each period are the cumulative sums of the regression coefficients up to and including the period.

Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data. The PPI is the BEA PPI for Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks (https://fred.stlouisfed.org/series/PCU334413334413A).

Table C1.2: Aggregate cost of computation decline rates with declining interest rate

PERIOD	CPU	GPU	Low growth	High growth	Top500	PPI
2000-2005	20.573	-	20.548	20.628	20.168	2.615
2005-2015	9.777	15.019	9.878	9.838	10.907	2.220
2015-2022	2.008	8.662	2.457	8.884	6.500	1.358

Notes: This table shows average annual decline rates from a regression based on equation 3 for the CPU, GPU and combined cost of computation indices where the interest rate declines by 5% per year. The combined indices were created by taking a weighted average of the GPU and CPU indices. 'Low-growth' denotes an index where we assume that the proportion of GPU use was 1% in 2006 and increased by 10% each year, 'high-growth' where we assume that the proportion of GPU was 1% in 2006 and increased 30% per year, 'Top500' where we use computation shares in the Top500 list as weights. Decline rates in each period are the cumulative sums of the regression coefficients up to and including the period.
Source: as in Table 2 and Table 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

C2 Varying the knots

To account for the possibility of a slowdown following the end of Dennard scaling around 2004 and to create approximately equal spacing between knots, 2005 and 2015 were chosen as break points in the main analysis. To test the sensitivity of the results to the choice of knots, we repeat the analysis by selecting three knots based on percentiles of the CPU data. The new knots are in January 2006, 2013, and 2017. Regression results for CPUs are shown in table C2.1 and C2.2. In contrast to the official PPI, the average decline in cost of computation slows to approximately 0 between 2014 and 2017 before declining again post-2017. The difference between the cost of computation decline rate and the official PPI is statistically significant at the 1% level in each period. Similarly, the decline rate for each aggregate index slows in the 2013-2017 period before steepening again post-2017 (though less quickly than pre-2013), a pattern not shared by the official PPI.

Table C2.1: CPU cost of computation regression results

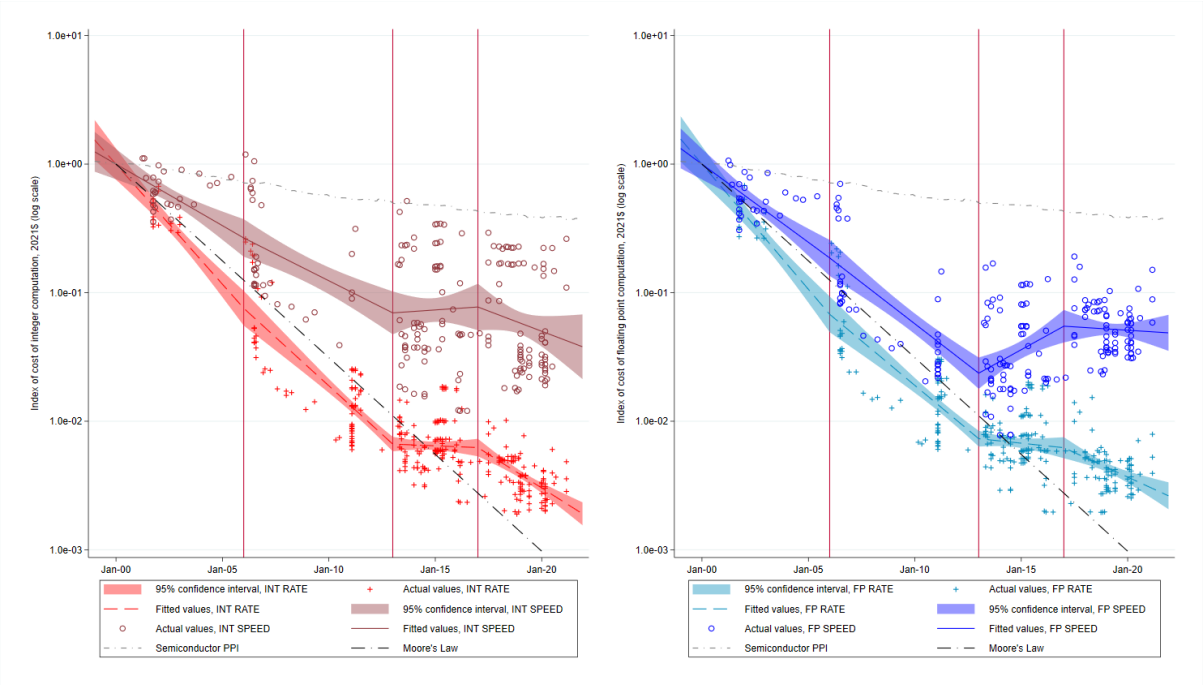
	(1)	(2)	(3)	(4)	(5)
VARIABLES	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED	PPI
Year	-0.187***	-0.096***	-0.195***	-0.122***	-0.025***
	(0.019)	(0.020)	(0.021)	(0.020)	(0.000)

DUM2006	0.036	0.012	0.057*	-0.006	0.002**
	(0.030)	(0.035)	(0.032)	(0.033)	(0.001)
DUM2013	0.144***	0.095*	0.121***	0.219***	0.009***
	(0.022)	(0.051)	(0.024)	(0.037)	(0.001)
DUM2017	-0.098***	-0.074	-0.059*	-0.102**	0.000
	(0.027)	(0.073)	(0.032)	(0.049)	(0.001)
Constant	8.422***	4.473***	8.775***	5.599***	3.159***
	(0.835)	(0.866)	(0.900)	(0.843)	(0.018)
Observations	299	204	281	208	276
R-squared	0.860	0.467	0.831	0.683	0.995
Breusch Pagan statistic	14.46	9.301	6.561	5.276	0.0502
p-value	0.000143	0.00229	0.0104	0.0216	0.823

Robust standard errors in parentheses for columns (1)-(4), non-robust errors in (5)
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table shows spline regression results for the logged (base 10) cost of computation index on Year and Time variable. YEAR is calendar year. DUM2006 is equal to 0 up to 2006 and increments by 1 after 2006. DUM2013 and DUM2017 are defined similarly. Outliers have been removed.
Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data. The PPI is the BEA PPI for Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks (https://fred.stlouisfed.org/series/PCU334413334413A).

Figure C2.1: CPU cost of computation index



Notes: This figure shows the CPU cost of computation in 2021$ by benchmark, with INT (integer computation) scores shown on the left panel and FP (floating point computation) scores shown on the right. Lines are the fitted values from equation (3) and shaded areas are 95% confidence intervals. Fitted values, confidence intervals and the index are shifted such that fitted values are equal to 1 in 1 Jan 2000. Knots are indicated with red lines. Moore's Law and the BEA semiconductor PPI are shown in black and grey respectively for comparison. *Source:* SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data. The Chip PPI is the BEA PPI for Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks (https://fred.stlouisfed.org/series/PCU334413334413A).

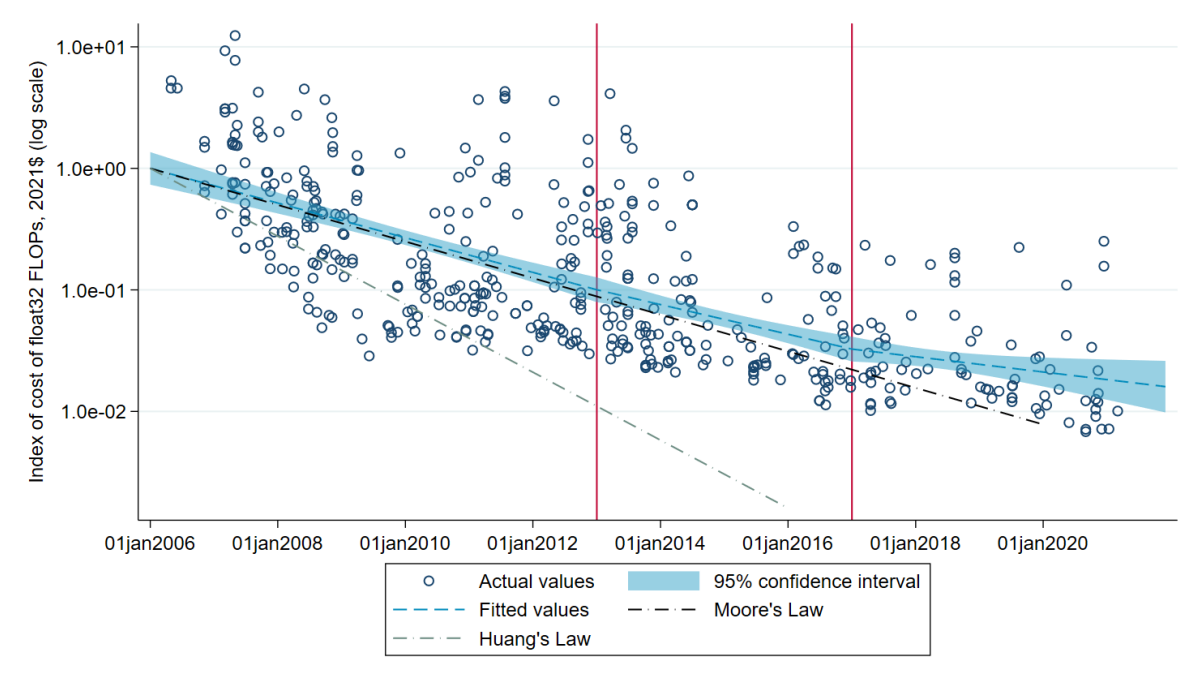
Table C2.2: CPU cost of computation decline rates

PERIOD	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED	Average	PPI
2000-2007	18.679***	9.569***	19.466***	12.159***	14.968***	2.526***
	(22.468 - 14.889)	(13.501 - 5.636)	(23.528 - 15.404)	(15.981 - 8.337)	(16.920 - 13.016)	(2.611 - 2.441)
2007-2014	15.089***	8.325***	13.814***	12.804***	12.508***	2.353***
	(17.417 - 12.762)	(11.709 - 4.942)	(16.208 - 11.421)	(15.717 - 9.891)	(13.902 - 11.115)	(2.417 - 2.289)
2014-2017	0.682	-1.147	1.757	-9.137***	-1.961	1.491***
	(3.306 - -1.942)	(6.099 - -8.393)	(4.810 - -1.296)	(-4.065 - -14.209)	(0.468 - -4.391)	(1.608 - 1.374)
2017-2022	10.477***	6.294	7.611***	1.050	6.358***	1.469***
	(13.407 - 7.548)	(14.207 - -1.618)	(11.100 - 4.122)	(6.083 - -3.983)	(8.965 - 3.752)	(1.593 - 1.346)

95% confidence intervals in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table shows average annual logarithmic (base 10) decline rates in the cost of computation for each CPU benchmark. The arithmetic mean and BEA chip PPI are displayed in the penultimate and last columns respectively. 95% confidence intervals are in parentheses. Outliers have been removed. Decline rates in each period are the cumulative sums of the regression coefficients up to and including the period.
Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data. The PPI is the BEA PPI for Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks (https://fred.stlouisfed.org/series/PCU334413334413A).

Figure C2.2: GPU cost of computation index



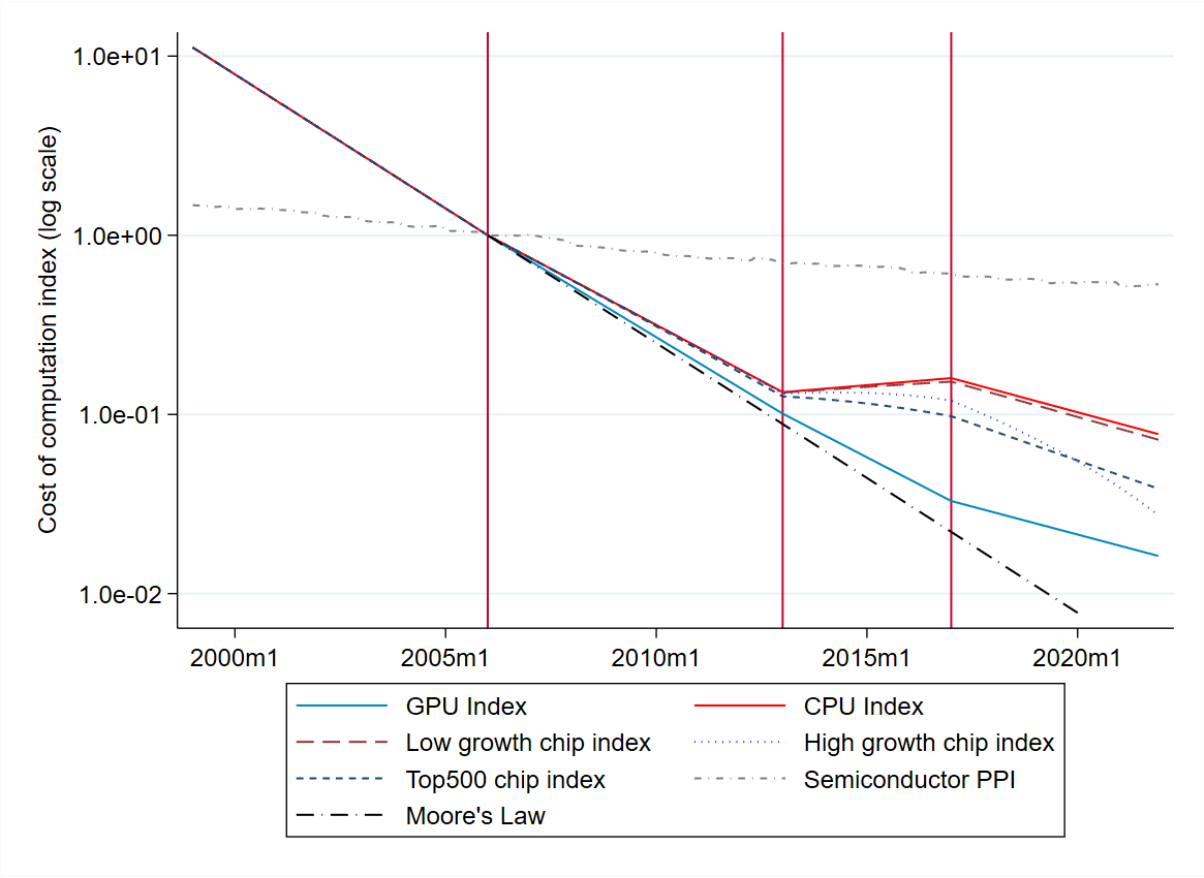
Notes: This figure shows GPU log cost of computation in 2021$. The dashed lines are fitted values from equation (5) and shaded areas are 95% confidence intervals. Moore's Law (doubling every 2 years) and Huang's Law (doubling every 1.08 years) are shown in the black and teal dot-dashed lines respectively. All values are shifted such that the fitted value for the cost of FP32 FLOPs is 1 in January 2006. Jan 2013 and Jan 2017 (the knots) are shown in red.
Source: GPU price-performance data comes from Epoch AI (https://epochai.org/blog/trends-in-gpu-price-performance), which uses data from Median Group (2018) and Sun *et al.* (2019).

Table C2.3: Combined cost of computation index decline rates

PERIOD	CPU	GPU	Low-growth chip index	High-growth chip index	Top500 chip index	PPI
2000-2006	14.968	-	14.968	14.972	14.946	2.526
2006-2013	12.512	14.221	12.538	12.548	12.751	2.353
2013-2017	-1.957	12.192	-1.528	0.471	2.819	1.491
2017-2022	6.353	6.237	6.600	12.732	8.411	1.469

Notes: This table shows average annual decline rates from a regression based on equation 3 for the CPU, GPU and combined cost of computation indices. The combined indices were created by taking a weighted average of the GPU and CPU indices. 'Low-growth' denotes an index where we assume that the proportion of GPU use was 1% in 2006 and increased by 10% each year, 'high-growth' where we assume that the proportion of GPU was 1% in 2006 and increased 30% per year, 'Top500' where we use a linear trend of the computation shares in the Top500 list as weights.
Source: as in Table 2 and Table 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

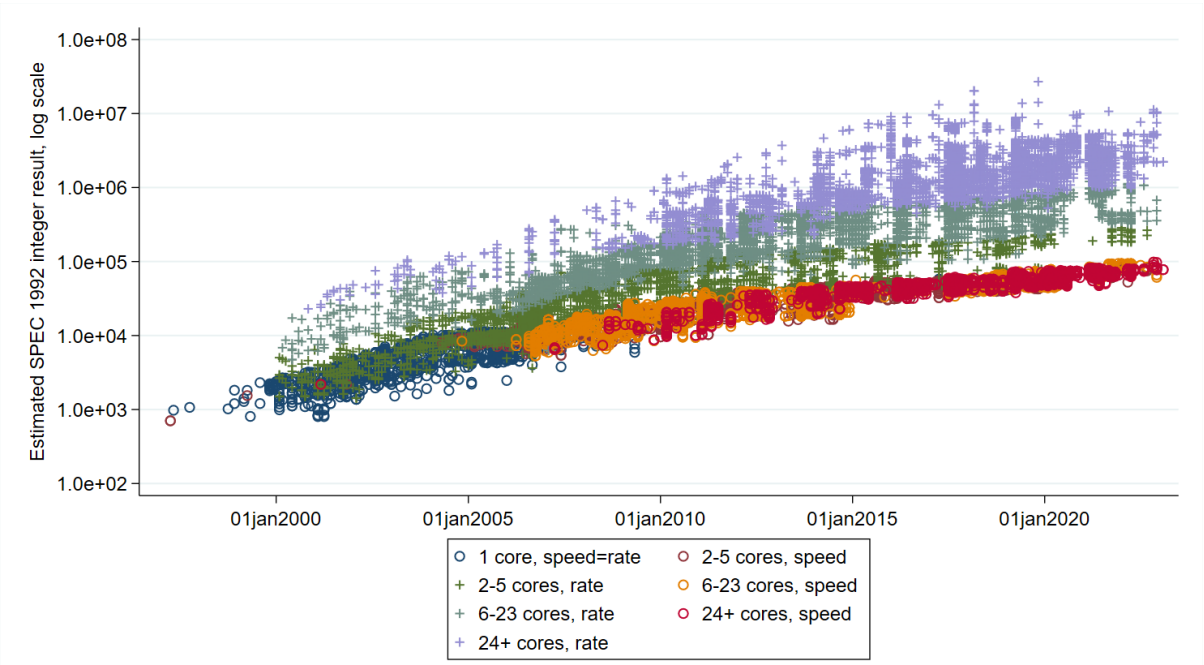
Figure C2.3: Combined cost of computation index



Notes: This figure shows fitted values for the regressions of CPU (red), GPU (blue) and combined cost of computation indices. The combined indices were created by converting the fitted values of the CPU and GPU cost of computation indices to the same scale and taking a weighted average of each. 'Low-growth' (maroon) denotes an index where we assume that the proportion of GPU use was 1% in 2006 and increased by 10% each year. 'High-growth' (blue) denote an index where we assume that the proportion of GPU was 1% in 2006 and increased 30% per year. 'Top500' (navy) is an index where we use a linear trend of the shares of computation in the Top500 lists as weights. Moore's Law is shown in black for comparison. 1 Jan 2005 and 1 Jan 2015 (the positions of the knots) are denoted by the red vertical lines. The value of each index has been set to 1 in 1 Jan 2006.
Source: as in Table 2 and Table 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

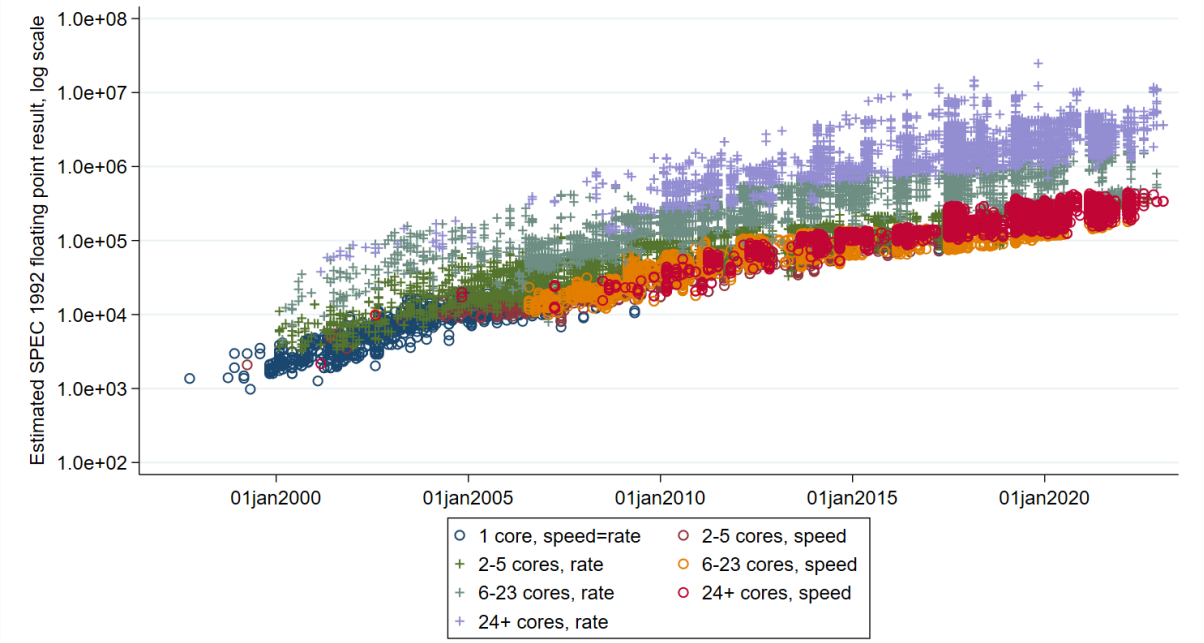
D. Additional figures

Figure D.1: CPU computing performance (integer), by cores



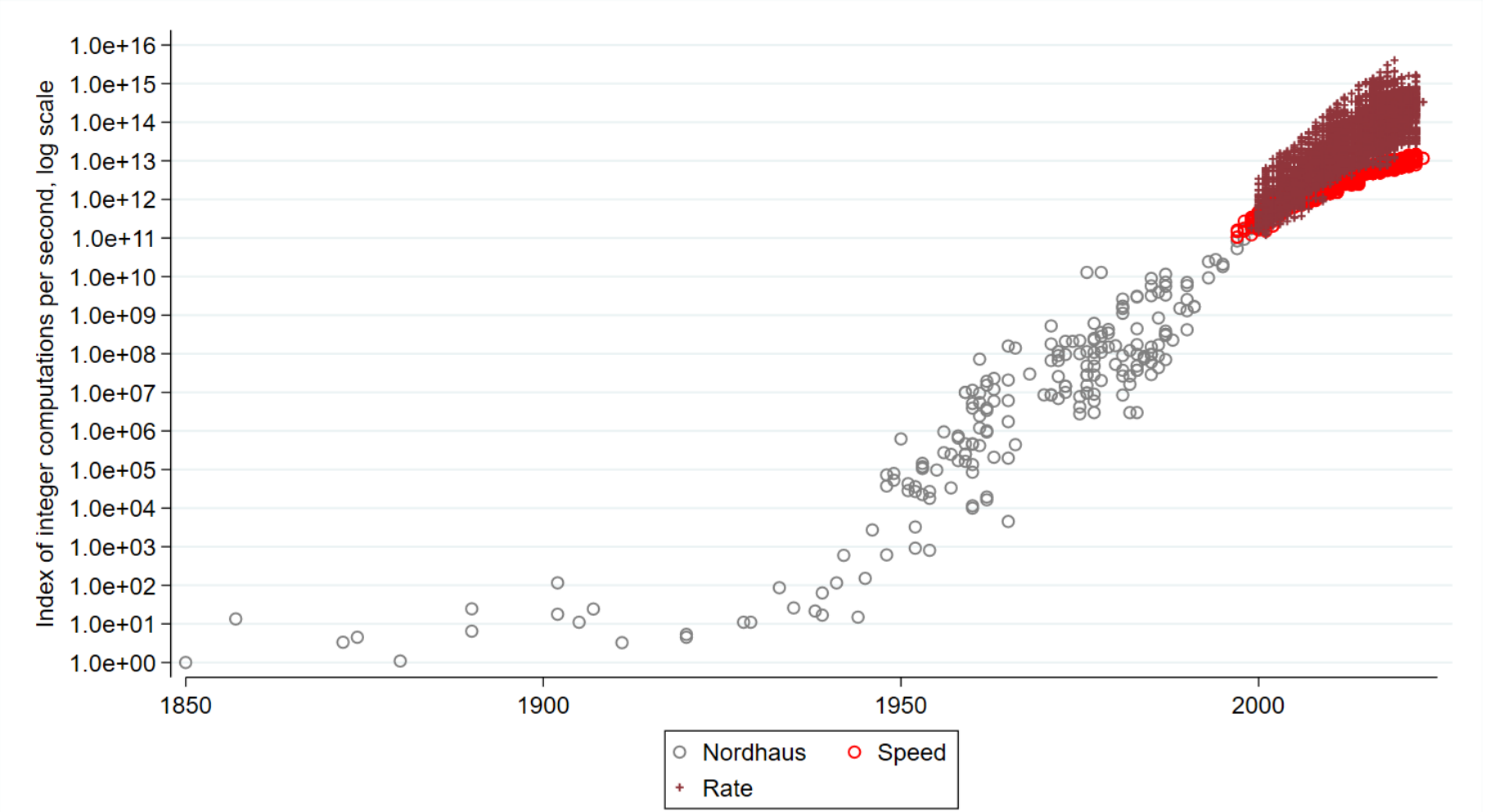
Notes: This figure plots estimated SPEC 1992 integer scores for SPEC CPU2000, SPEC CPU2006 and SPEC CPU2017 benchmarks. Different colours indicate the number of cores. RATE scores are indicated by plus markers.
Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations.

Figure D.2: CPU computing performance (floating point), by cores



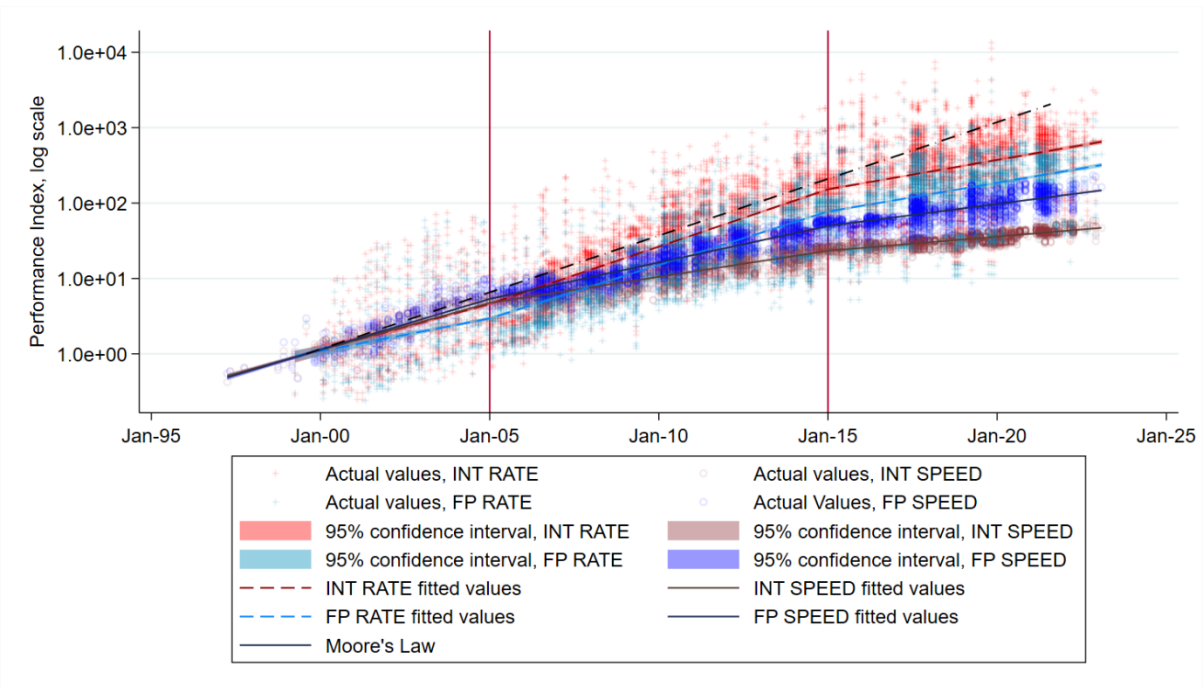
Notes: This figure plots estimated SPEC 1992 floating point scores for SPEC CPU2000, SPEC CPU2006 and SPEC CPU2017 benchmarks. Different colours indicate the number of cores. RATE scores are indicated by plus markers.
Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations.

Figure D.3: Long-run index of computer system integer computations per second, 1950-2023



Notes: This figure shows the long-run index of integer computations per second, anchored to human-level performance (1). Speed and rate scores are indicated separately.
Source: The pre-2007 series uses data from the appendix of Nordhaus (2007), while the post-2007 series uses performance scores from SPEC CPU2000, 2006 and 2017 benchmarks (https://www.spec.org/benchmarks.html#cpu) and author's calculations for the conversions.

Figure D.4: CPU performance index with actual values



Notes: This figure shows actual and fitted values of the relative performance index of computations per second at the CPU level over the date of hardware availability. The performance index is constructed using SPEC INT RATE (red), INT SPEED (brown), FP RATE (light blue) and FP SPEED (blue) scores converted to their 1992 equivalents. To convert from the computer system level to the CPU level, scores are averaged by processor, clock speed, number of chips, number of cores and date of hardware availability. Outliers have been removed from the sample. Scores are then divided by the score of the first observation in each group to obtain relative performance. Fitted values are obtained using equation (1). The actual and fitted values and 95% confidence intervals are shifted such that the value of each benchmark in August 1999 is 1. Moore's Law is shown in black.
Sources: SPEC (https://www.spec.org/benchmarks.html#cpu), authors' calculations.

Table D.1: CPU performance over time, outliers removed

VARIABLES	(1) INT, RATE	(2) INT, SPEED	(3) FP, RATE	(4) FP, SPEED
Year	0.122***	0.125***	0.087***	0.134***
	(0.008)	(0.002)	(0.008)	(0.002)
DUM2005	0.030***	-0.055***	0.055***	-0.038***
	(0.009)	(0.003)	(0.008)	(0.003)
DUM2015	-0.074***	-0.032***	-0.065***	-0.038***
	(0.004)	(0.001)	(0.004)	(0.002)
Constant	-4.132***	-4.706***	-2.555***	-5.077***
	(0.341)	(0.097)	(0.333)	(0.106)
Observations	7,921	5,250	7,452	4,902
R-squared	0.747	0.943	0.747	0.945
Breusch Pagan statistic	1.565	690.8	0.0511	1.368
p-value	0.211	0	0.821	0.242

Standard errors in parentheses, robust in col. 2

*** p<0.01, ** p<0.05, * p<0.1

Table D.2: CPU performance over time, outliers included

VARIABLES	(1) INT, RATE	(2) INT, SPEED	(3) FP, RATE	(4) FP, SPEED
Year	0.118***	0.127***	0.089***	0.136***
	(0.009)	(0.002)	(0.009)	(0.003)
DUM2005	0.033***	-0.059***	0.049***	-0.041***
	(0.010)	(0.003)	(0.010)	(0.003)
DUM2015	-0.074***	-0.030***	-0.062***	-0.055***
	(0.004)	(0.001)	(0.004)	(0.002)
Constant	-3.913***	-4.789***	-2.643***	-5.136***
	(0.400)	(0.103)	(0.421)	(0.116)
Observations	7,972	5,334	7,518	5,250
R-squared	0.730	0.934	0.718	0.854
Breusch Pagan statistic	10.48	436.5	47.04	686.5
p-value	0.00121	0	0	0

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes: These tables show the results for the regression of relative CPU performance, measured in log10 estimated SPEC 1992 scores divided by the first score in the sample, on year and time variables based on equation (1). The regression is conducted separately for SPEC INT RATE (column 1), INT SPEED (column 2), FP RATE (column 3) and FP SPEED (column 4) scores. To convert from the computer system to the CPU level, scores are averaged by processor, clock speed, number of chips, number of cores and date of hardware availability. In the first table, observations identified as outliers are removed. Year is the calendar year, DUM2005 is a variable equal to 0 if the year is less than 2005 and increments by 1 each year after 2005, and DUM2015 is defined similarly.
Sources: SPEC (https://www.spec.org/benchmarks.html#cpu), authors' calculations.

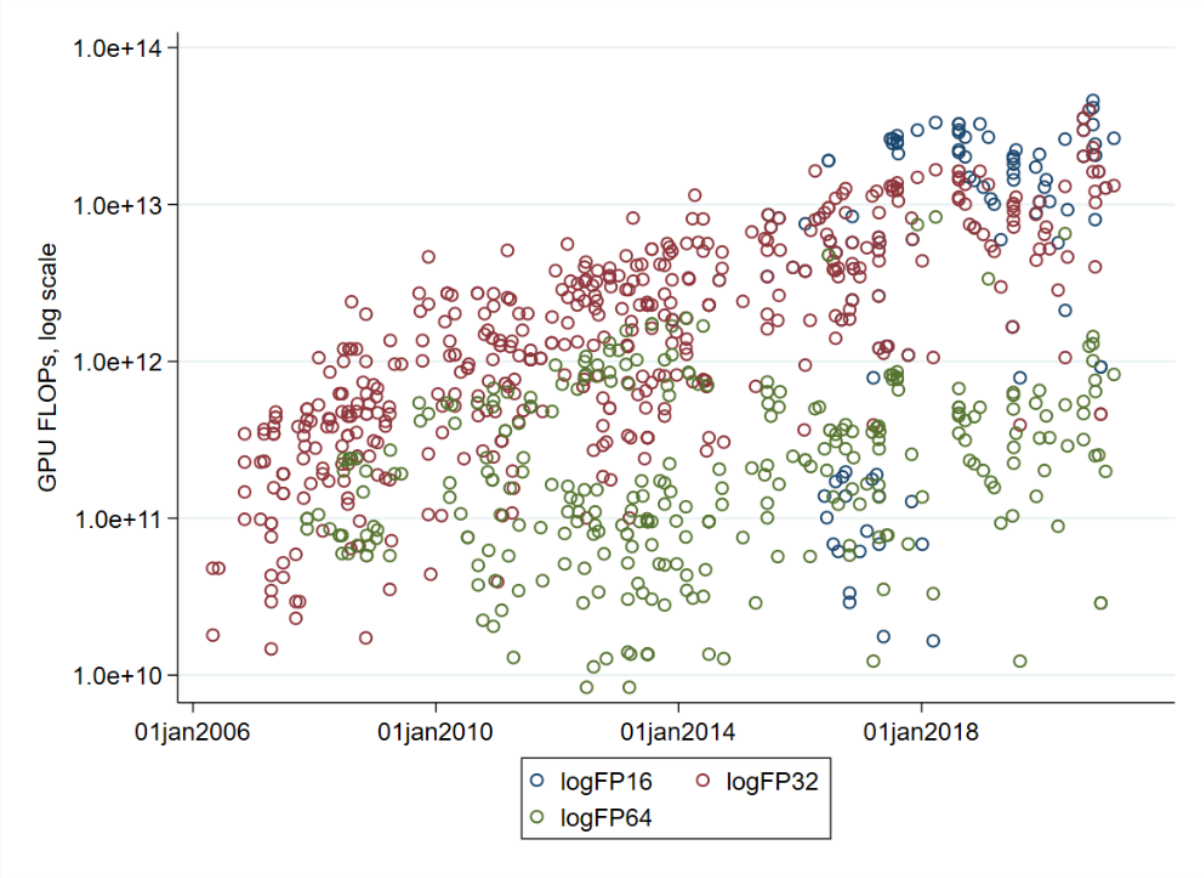
Table D.3: CPU price performance with outliers

VARIABLES	(1) INT, RATE	(2) INT, SPEED	(3) FP, RATE	(4) FP, SPEED
Year	-0.254***	-0.122***	-0.260***	-0.159***
	(0.027)	(0.028)	(0.029)	(0.028)
DUM2005	0.145***	0.072*	0.156***	0.081**
	(0.035)	(0.038)	(0.037)	(0.038)
DUM2015	0.058***	0.011	0.063***	0.131***
	(0.013)	(0.029)	(0.013)	(0.022)
Constant	11.270***	5.573***	11.517***	7.181***
	(1.167)	(1.198)	(1.261)	(1.184)
Observations	299	210	281	217
R-squared	0.839	0.366	0.815	0.546
Breusch Pagan statistic	14.63	9.894	7.535	14.70
p-value	0.000131	0.00166	0.00605	0.000126

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table shows spline regression results for the logged (base 10) cost of computation index on Year and Time variable. YEAR is calendar year. DUM2006 is equal to 0 up to 2005 and increments by 1 after 2005. DUM2015 is defined similarly. Outliers are included in these regressions.
Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations for performance data, Passmark (https://www.cpubenchmark.net/CPU_mega_page.html), WebArchive (https://web.archive.org/), TechPowerUp (https://www.techpowerup.com/cpu-specs/?sort=name) and CPU World (https://www.cpu-world.com/) for price data.

Figure D.5: GPU FLOPs trends



Source: EpochAI (https://epochai.org/blog/trends-in-gpu-price-performance)

Table D.4: Aggregate index with outliers

PERIOD	CPU	GPU	Low-growth chip index	High-growth chip index	Top500 chip index
2000-2005	19.903	-	19.874	19.935	19.456
2005-2015	8.565	14.234	8.675	8.673	9.799
2015-2022	1.998	8.069	2.436	8.859	6.384

Notes: This table shows average annual decline rates from a regression based on equation 3 for the CPU, GPU and combined cost of computation indices with outliers included. The combined indices were created by taking a weighted average of the GPU and CPU indices. 'Low-growth' denotes an index where we assume that the proportion of GPU use was 1% in 2006 and increased by 10% each year, 'high-growth' where we assume that the proportion of GPU was 1% in 2006 and increased 30% per year, 'Top500' where we use a linear trend of the computation shares in the Top500 list as weights. Decline rates in each period are the cumulative sums of the regression coefficients up to and including the period.
Source: As in Table 2 and Table 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

UNIVERSITY OF CAMBRIDGE

www.bennettinstitute.cam.ac.uk

Bennett Institute for Public Policy Cambridge