Telecommunications Policy 48 (2024) 102649

Contents lists available at ScienceDirect

Telecommunications Policy

journal homepage: www.elsevier.com/locate/telpol

21st century progress in computing☆

Diane Coyle *, Lucy Hampton

Bennett Institute, University of Cambridge Alison Richard Building, 7 West Road, Cambridge, CB3 9DT UK

ARTICLE INFO

Keywords:
Computation
Prices
Productivity
AI
Cloud computing

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 2010, but not earlier. As firms use computational power along with other inputs including relevant human and organisational capital, to the extent that the productivity slowdown is linked to technology use the explanation is likely to lie in these other elements of the input bundle.

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.[1] The great majority of people use digital technologies frequently and value them highly (Brynjolfsson et al. 2020; Coyle & 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 & Mei, 2023; Goodridge & 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

☆ 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 are very grateful to Professor Erik Bohlin and one anonymous reviewer for their support and insight that has enabled us to strengthen our paper through the review process. 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.

* Corresponding author.
E-mail addresses: dc700@cam.ac.uk (D. Coyle), ljh92@cam.ac.uk (L. Hampton).

[1] 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.

https://doi.org/10.1016/j.telpol.2023.102649
Received 16 June 2023; Received in revised form 12 September 2023; Accepted 14 September 2023
Available online 4 October 2023
0308-5961/

fastest productivity-growth sector in the economy.

There is an extensive literature exploring the productivity puzzle, dividing 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 (Bloom et al. 2020; Gordon, 2017) – 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 (Bessen, 2022; Brynjolfsson et al. 2021).

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 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 of interest is what impact the 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 (Flamm, 2017; Ramu, 2018; Shalf, 2020; Theis & Wong, 2017). Or alternatively are there measurement challenges (similar to those identified in telecommunications) that mean price declines and corresponding output or productivity increases have been under-stated in current statistics? The answer matters not only for the narrow purpose of productivity measurement but also because – like all General Purpose Technologies – 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?

We address this using an engineering-based[2] 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 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, 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 computational power 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, 2023). On the other hand, the latest generation of AI applications require large amounts of computation (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 2010 (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 1995, and that there was reasonable evidence of a firm-level association between ICT use and productivity. A more recent literature review (Biagi, 2013, p. 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 U.S. 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 had 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 should 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, p. 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 percent of U.S.

[2] 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.

Table 1
Post-2005 major 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 eg 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

investment, 60 percent of U.S. consumption, and some 80 percent of U.S. 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. Hedonic regression, used by statistical agencies for some technology goods, he rejects 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: standardized operations per second. This begins with a standard measure, millions of instructions per second, taking account of the nature of the instruction, and standardizing by converting to 32 bit units. This calculation gives a physical, hardware performance measure, millions of standardized 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 2000 h 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 (Johnsson & Netzer, 2016; Lotfi-Kamran & Sarbazi-Azad, 2018), 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.[3] 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 in 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 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

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

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[4] 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). The model specification is as follows:

$$ln(perf_t) = \beta_0 + \beta_1 Year_t + \beta_2 SPLINE2005_t + \beta_3 SPLINE\_K_t + \beta_4 SPLINE2020_t + \varepsilon_t \quad (1)$$

where $perf_t$ is an index of computing performance, $Year_t$ is the year of hardware introduction, ε_t is the error term, and $SPLINE[YEAR]_t$ is a truncated power basis function defined as follows[5]:

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

The resulting model is piecewise (log-)linear, with the joins referred to as ‘knots’. 2005 is chosen as the first knot to allow a change in slope following the end of Dennard scaling and to test whether there is a slowdown in the decline of the cost of computing coinciding with the broader productivity slowdown, generally dated to the mid-2000s. 2020 is included to allow for a pandemic-related slowdown. Given that there is an apparent slowdown in the data in the 2010–2015 period, a third knot (K) is placed in whatever year within this range minimises the Akaike’s information criterion (AIC), and is allowed to vary across benchmarks.

The coefficients on the SPLINE [YEAR] variables capture the effect of increasing ‘year’ by 1 on the natural 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,[6] TechPowerUp,[7] and CPUWorld.[8] As Passmark does not report prices at introduction, we used the Wayback Machine[9] 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.[10] 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 GDP deflator (rebased to 2021) to obtain constant prices.

To obtain decadal price decline rates, we use a similar specification to the one used above for the performance data:

$$ln(price\_perf_t) = \beta_0 + \beta_1 Year_t + \beta_2 SPLINE2005_t + \beta_3 SPLINE\_K_t + \beta_4 SPLINE2020_t + \varepsilon_t \quad (3)$$

where all regressors are defined as before. As in the performance regressions, the value of K is allowed to vary across benchmarks and is determined using the AIC.

2.3. CPU results

Fig. 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. Fig. 1 demonstrates that hardware performance as measured by SPEC scores has improved by over 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 appear to improve at a faster rate until slowing down in the 2010–2015 period. Improvements also appear to slow down in 2020, though this is not true of the INT SPEED benchmark. 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

[4] 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 et al. (2000) and Weber (2010).

[5] t is the number of days between the date of hardware introduction and the 1st 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.

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

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

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

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

[10] 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.

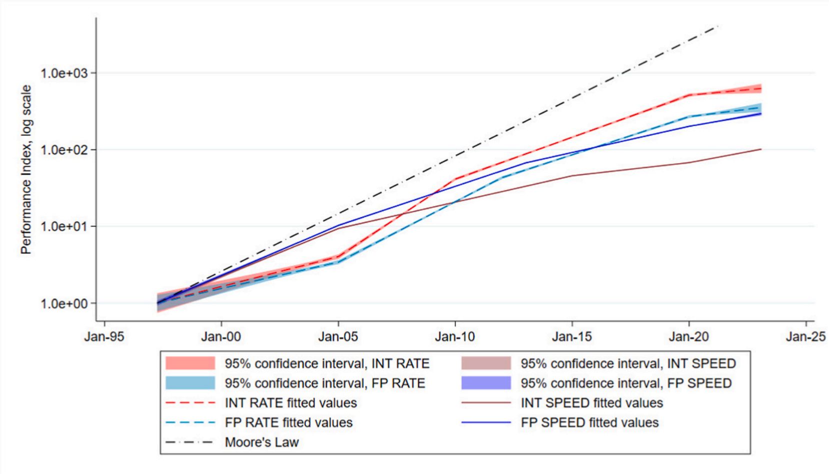

Fig. 1. Fitted values of the performance index of computations per second (averaged by CPU), by benchmark.
Notes: This figure shows fitted values of the relative performance index of computations per second at the CPU level over the date of hardware availability, with confidence intervals. Moore's Law (stating that the number of transistors in an integrated circuit doubles every 2 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. Fitted values are obtained using equation (1). Fitted values and 95% confidence intervals are shifted such that all the fitted values are equal to 1 in April 1997. 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.

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 Producer Price Index (PPI) for semiconductors (Semiconductors and Related Device Manufacturing: Other Semiconductor Devices, Including Parts Such as Chips, Wafers, and Heat Sinks).[11] Fig. 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 Fig. 2) as CPU and computer system price trends may diverge.

Placing the second knot in 2010 produced the best fit for each benchmark. In each case there was a statistically significant slowdown in the pace of decline in the cost of computation, although there were differences between benchmarks in when this occurred. The slowdown for the INT SPEED benchmark started around 2005, while the other slowdowns started around 2010. The cost of computation as measured by one, FP SPEED scores, is estimated to have increased since 2010. As well as the slowdown in performance improvements shown in Fig. 1, and so in price declines shown in Fig. 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 are slowdowns for each benchmark since 2020 (though only statistically significant in two cases), presumably related to the specific adverse supply shocks due to the pandemic.

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 Producer Price Index for the duration of the sample, the same finding as in Nordhaus (2001, 2007) for earlier years. This difference is also statistically significant ($p < 0.0001$), and even larger when the FP SPEED benchmark is omitted. 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.

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 contributing to slower progress in computation, there are new types of chip – originally designed for specialized 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

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

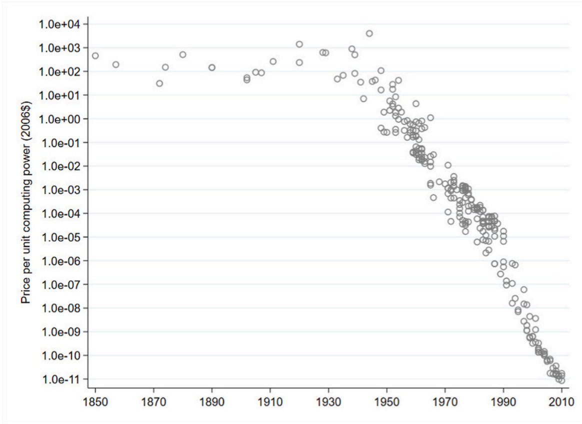

Fig. 2. Nordhaus cost of computation series (2006$)
Notes: This figure depicts Nordhaus' cost of computation series. Y axis is computations per second per 2006 USD, and x axis is date of observation or hardware availability. Nordhaus' computations per second measure is defined such that the speed of manual computation is equal to 1.
Source: Nordhaus (2007) appendix.

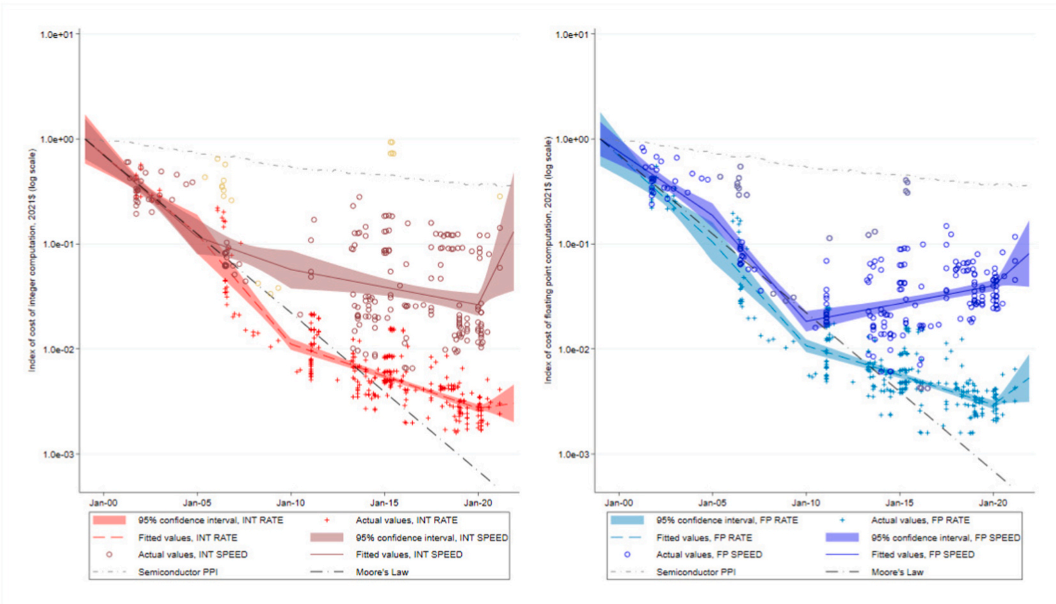

Fig. 3. Cost of CPU computation in 2021 $ (2000–2022), by benchmark
Notes: This figure shows the CPU cost of computation in 2021$ by benchmark, with INT (integer computation) scores shown on the left panel and FP (floating point computation) scores shown on the right. Lines are the fitted values from 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 Jan 1, 2000. Outliers are highlighted in sand and grey respectively. 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).

common in AI training runs.[12] 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

[12] 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.

investment and provision, and including access to the most up-to-date software and equipment (Byrne et al. 2018; Coyle & 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[13]; 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 from Epoch AI,[14] 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.[15] 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.[16]

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

$$ln(price\_perf_t) = \beta_0 + \beta_1 Year_t + \beta_2 SPLINE2010_t + \beta_3 SPLINE2020_t + \varepsilon_t \quad (5)$$

where $SPLINE2010_t$ is equal to 0 up to 2010 and increments by 1 each year after 2010, and so on for $SPLINE2020_t$. As in the previous section, the choice of 2010 was informed by minimisation of the AIC. 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 $. Fig. 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-2010. However, the logarithmic rate of decline still averages around 21% per year in the 2010–2020 period. Due to the sparsity of the data between 2020 and 2022, there is large uncertainty around the decline rate in this period (see Fig. 5).

Table 5 shows the decline rates for the combined GPU and CPU indices. The magnitude of the difference between the CPU index and combined index with low-growth assumptions is fairly small: both decline at a logarithmic rate of approximately 39 per year, 2005–2010, while in the period 2010–2020 the CPU index declines by approximately 6.7 per year and the combined index by 7.5. Assuming a higher, 30% annual growth rate produces an annual logarithmic decline in the cost of computation of 14.6 in the period 2010–2020, approximately twice as large as the decline rate in the CPU index and almost four times as large as that of the official BEA semiconductor index. The Top500 index is similar to the high-growth index during the 2010–2020 period but is much slower in 2020–2022 as the weight on the GPU index increases linearly, not exponentially.

We also update Coyle and Nguyen (2018) figures for prices of cloud use in the UK. Fig. 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.

4. Discussion

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 early 2010s.

[13] https://goingdigital.oecd.org/indicator/25.

[14] 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).

[15] https://www.top500.org/statistics/overtime/.

[16] 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).

Table 2
Regression results for CPU cost of computation series.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED	PPI
Year	−0.347*** (0.078)	−0.357*** (0.066)	−0.376*** (0.082)	−0.279*** (0.050)	−0.053*** (0.001)
SPLINE2005	−0.136 (0.123)	0.213* (0.129)	−0.080 (0.128)	−0.186** (0.085)	−0.011*** (0.002)
SPLINE2010	0.342*** (0.051)	0.067 (0.095)	0.327*** (0.053)	0.544*** (0.052)	0.027*** (0.001)
SPLINE2020	0.196 (0.123)	0.914** (0.377)	0.434*** (0.153)	0.288 (0.210)	0.023*** (0.004)
Constant	15.903*** (3.318)	16.018*** (2.785)	17.151*** (3.507)	12.830*** (2.114)	7.079*** (0.040)
Observations	299	192	281	196	276
R-squared	0.863	0.457	0.839	0.724	0.997
Breusch Pagan statistic	9.892	12.56	4.426	11.55	2.439
p-value	0.00166	0.000395	0.0354	0.000679	0.118

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
Notes: This table shows regression results for the logged (base e) cost of computation index based on equation (3). YEAR is calendar year. SPLINE2005 is equal to 0 up to 2005 and increments by 1 after 2005. SPLINE2010 is equal to 0 up to 2010 and increments by 1 after 2010. Outliers have been removed. Results with outliers are in table D3 in the appendix. The Breusch Pagan statistics test the null of homoskedasticity. Heteroskedasticity-robust standard errors are in parentheses for columns (1)–(4), but non-robust for (5), where the null is not rejected.
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	34.734*** (49.951–19.516)	35.696*** (48.457–22.935)	37.596*** (53.622–21.570)	27.912*** (37.504–18.321)	33.985*** (40.800–27.169)	5.341*** (5.546–5.135)
2005–2010	48.299*** (57.413–39.184)	14.372** (27.834–0.909)	45.569*** (54.959–36.180)	46.523*** (54.476–38.569)	38.690*** (43.788–33.593)	6.483*** (6.639–6.327)
2010–2020	14.064*** (15.848–12.279)	7.714*** (13.355–2.073)	12.917*** (14.905–10.928)	−7.875*** (−5.103 to −10.648)	6.705*** (8.412–4.997)	3.823*** (3.895–3.751)
2020–2022	−5.586 (17.506 to −28.678)	−83.689** (−12.708 to −154.670)	−30.466** (−1.532 to −59.400)	−36.705* (3.041 to −76.451)	−39.111*** (−16.767 to −61.456)	1.564*** (2.186–0.942)

Notes: This table shows cumulative sums of the estimated coefficients in Tables 2 and i.e., average annual logarithmic (base e) 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).

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,[17] AI capabilities are most likely to have 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 and 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

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

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

VARIABLES	(1) GPU Index	Decline rate (% per year)
Year	−0.642*** (0.067)	64.243*** (77.248–51.237)
SPLINE2010	0.431*** (0.080)	21.151*** (24.874–17.429)
SPLINE2020	−0.000 (0.375)	21.152 (92.897 to −50.594)
Constant	28.868*** (3.250)	
Observations	470	
R-squared	0.498	
Breusch Pagan statistic	0.122	
p-value	0.727	

***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, SPLINE2010 is equal to 0 up to 2010 and increments by 1 after 2010, and SPLINE2020 is defined similarly. The second column shows the average annual decline for each period, calculated as the cumulative sum of the regression coefficients up to the date. Robust standard errors are in parentheses in the left column, and 95% confidence intervals in the right column.
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).

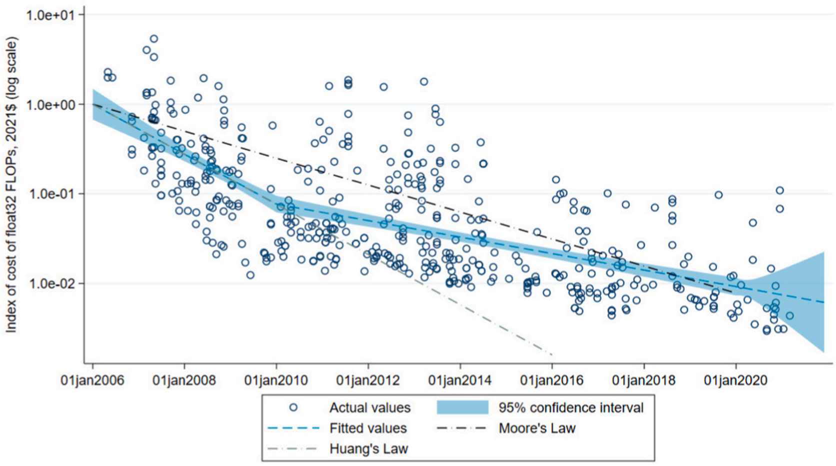

Fig. 4. Cost of GPU computation in 2021 $ (2006–2022)
Notes: This figure shows GPU log cost of computation in 2021$. The dashed lines are fitted values from equation (5) and shaded areas are 95% confidence intervals. Moore's Law (doubling every 2 years) and Huang's Law (performance doubling every 1.08 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.
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).

and organisational capital. Furthermore, although the capabilities of new AI systems are striking, the computational requirements 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 & Brown, 2020).

In this paper we have addressed one element suggested as a possible contributor to the productivity growth slowdown, namely that computational progress has slowed, perhaps due to the slowdown in Moore's Law (Azar, 2022. https://www.newyorkfed.org/research/staff_reports/sr970). However, we found a dramatic cost decline in computation and cloud services, measured in several ways, which did not decelerate until after the broader productivity growth slowdown. Other work (Fleming, 2023) 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 ICT sector's productivity growth. How different deflator measures would ultimately impact estimates of productivity growth across the whole economy depends on supply linkages, as software and services are for the most part intermediate inputs, and on the subsequent expansion of economic activity. For example, the revised telecoms services deflator constructed by Abdirahman et al. (2020,2022)

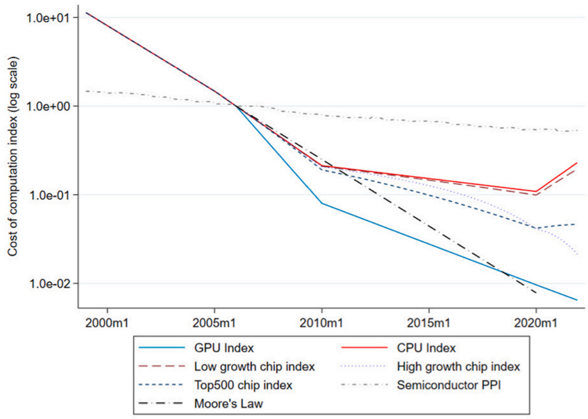

Fig. 5. Combined Cost of Computation Chip Index in 2021$ (1999–2022)
Notes: This figure shows fitted values for the regressions of CPU (red), GPU (blue) and combined cost of computation indices, as well as the semiconductor PPI (grey). The combined indices were created by converting the fitted values 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. The value of each index has been set to 1 in Jan 1, 2006.
Source: as in Tables 2 and 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

Table 5
Average annual decline rates for combined chip indices.

PERIOD	CPU	GPU	Low-growth chip index	High-growth chip index	Top500 index	PPI
2000–2005	33.981		33.978	34.607	33.874	5.341
2005–2010	38.700	63.144	38.867	36.499	39.843	6.483
2010–2020	6.708	21.183	7.459	14.558	15.104	3.823
2020–2022	−39.106	20.710	−34.789	48.906	−1.342	1.564

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 Tables 2 and 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

contributed to a non-trivial upward revision to volume-terms GDP growth in the UK over a decade.

However, there are broader issues to consider in understanding productivity outcomes. Abdirahman et al. (2020,2022) show that the choice of weights makes a substantial difference to the producer output price index for telecoms services. 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 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 appropriate 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. The measurement dilemma is that 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 similar value assignment with the use of computational resources.

However, we have shown that there have been very substantial declines for many decades in the price of computation. Although the pace of decline has slowed, the timing of the slowdown (early 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, as these substantially understate our measures.

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

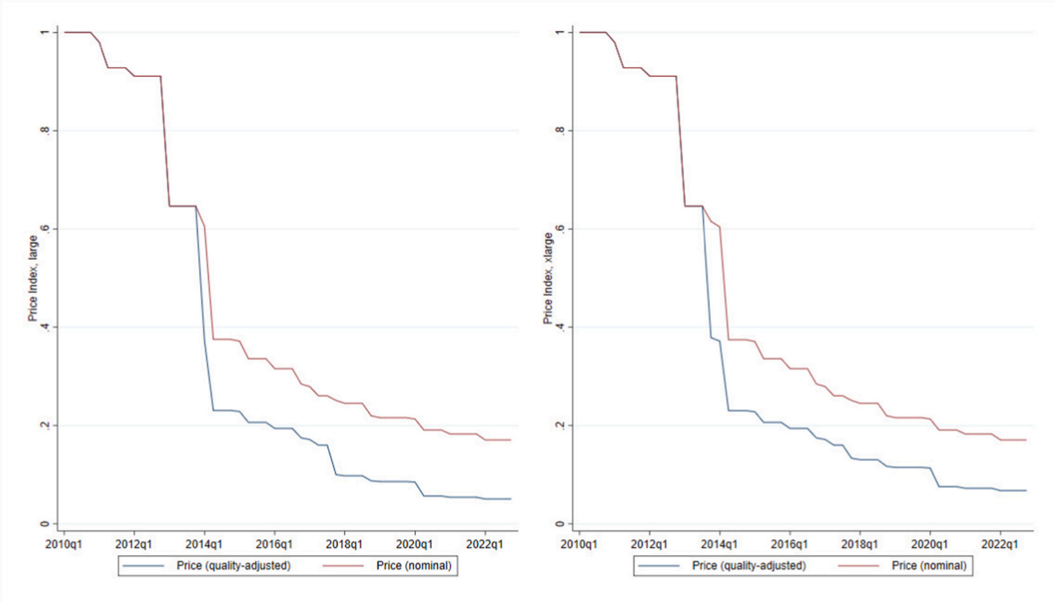

Fig. 6. Cloud price index, Linux (Q1-2010 – Q4-2022)
Notes: This figure shows nominal and quality-adjusted prices of AWS EC2 large and xlarge instances for Linux. Prices are hourly on-demand rates deflated by the aggregate price index. In blue, prices are also quality-adjusted for performance improvements.
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).

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.

Data availability

Data will be made available on request.

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 modeling, 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”.[18] Data was obtained from the SPEC website for the SPEC 2000, SPEC 2006; 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 B1. Scores for each benchmark suite are geometric means of equally-weighted scores in each program.

Table B.1
a comparison of SPEC benchmarks

CINT2006	CFP2006	CINT2017	CFP2017
PERL programming language	Fluid Dynamics	Perl interpreter	Explosion modeling
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 modeling
Path-finding Algorithms	Linear Programming, Optimization	General data compression	Wide-scale ocean modeling (climate level)
XML Processing	Image Ray-tracing		Image manipulation
	Structural Mechanics		Molecular dynamics
	Computational Electromagnetics		Computational Electromagnetics
	Quantum Chemistry		Regional ocean modeling
	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 B2 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 $2 \times 83 = 166$ 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:

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

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

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 = 25 500 s
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 D3.

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.[19] 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 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 *51 394/449	60.20/7.752	1708/19 *73 044/493
	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

[19] 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.*

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 h 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 in magnitude: e.g. the decline in the 2010–2020 period with constant interest rates is 6.7 per year, while with declining interest rates it is 8.2 per year. The finding that there is a slowdown post-2010 is also robust. For the aggregate indices shown in table C1.2, the growth rate is higher post-2010 than it is with a constant interest rate, though there is still a slowdown after 2010.

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	36.986***	37.933***	39.851***	30.173***	36.236***
	(52.200–21.772)	(50.683–25.184)	(55.873–23.828)	(39.760–20.585)	(43.049–29.423)
2005–2010	50.444***	16.590**	47.712***	48.693***	40.860***
	(59.557–41.331)	(30.047–3.133)	(57.101–38.324)	(56.647–40.738)	(45.956–35.763)
2010–2020	15.610***	9.215***	14.463***	−6.352***	8.234***
	(17.394–13.826)	(14.856–3.575)	(16.451–12.475)	(-3.577 to −9.127)	(9.942–6.527)
2020–2022	−4.807	−82.735**	−29.673**	−35.774*	−38.247***
	(18.248 to −27.862)	(-11.815 to −153.654)	(-0.768 to −58.578)	(3.931 to −75.479)	(-15.924 to −60.570)

Notes: This table shows average annual logarithmic (base e) decline rates in the cost of computation for each CPU benchmark, with interest rates declining by 5% per year. The arithmetic mean is displayed in the last column. 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	36.233		36.230	36.860	36.127	5.341
2005–2010	40.870	65.246	41.035	38.661	42.006	6.483
2010–2020	8.238	22.776	8.991	16.101	16.646	3.823
2020–2022	−38.241	21.271	−33.933	49.679	−0.597	1.564

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 Tables 2 and 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

C2 Varying the Knots

Motivating the choice of knots in the main analysis was the aim of testing for a slowdown in the cost of computation coinciding with that of the broader productivity growth slowdown, accounting for the apparent slowdown in the data between 2010 and 2015, and allowing a break in trend during the pandemic. To test the sensitivity of the results to the location of the knots, we repeat the exercise using the Multivariate Adaptive Regression Splines (MARS) algorithm (Friedman & Silverman, 1989) which automatically chooses the location and number of knots. The algorithm proceeds in two stages: a forward pass and a backward pass. In the forward pass, MARS adds truncated power basis functions in turn to achieve the maximum reduction in the sum of squared residual errors until a maximum number of terms is reached. This typically produces an overfitted model, so the backwards pass prunes the model of any terms that increase the Generalised Cross Validation (GCV) criterion, which both rewards fit and imposes a penalty on the number of terms.

MARS is applied to each CPU benchmark with outliers excluded to generate figures C2.1 and C2.2.[20] GRSQ (1 – GCV/GCV of the null), a measure of the predictive power of the model, is 0.862 for INT_RATE, 0.439 for INT_SPEED, 0.831 for FP_RATE and 0.718 for FP_SPEED. Similar to the main analysis, the algorithm placed knots in the 2010–2015 period for ¾ of the benchmarks. INT benchmarks additionally exhibited slowdowns pre-2010. Additionally, there is a rise in the cost of computation in 2020 for 2 of the benchmarks, although the data is sparse in this region and confidence intervals are consequently wide. In general, and particularly for the average cost of computation in figure C2.2, the results presented here exhibit similar trends to those in the main analysis.

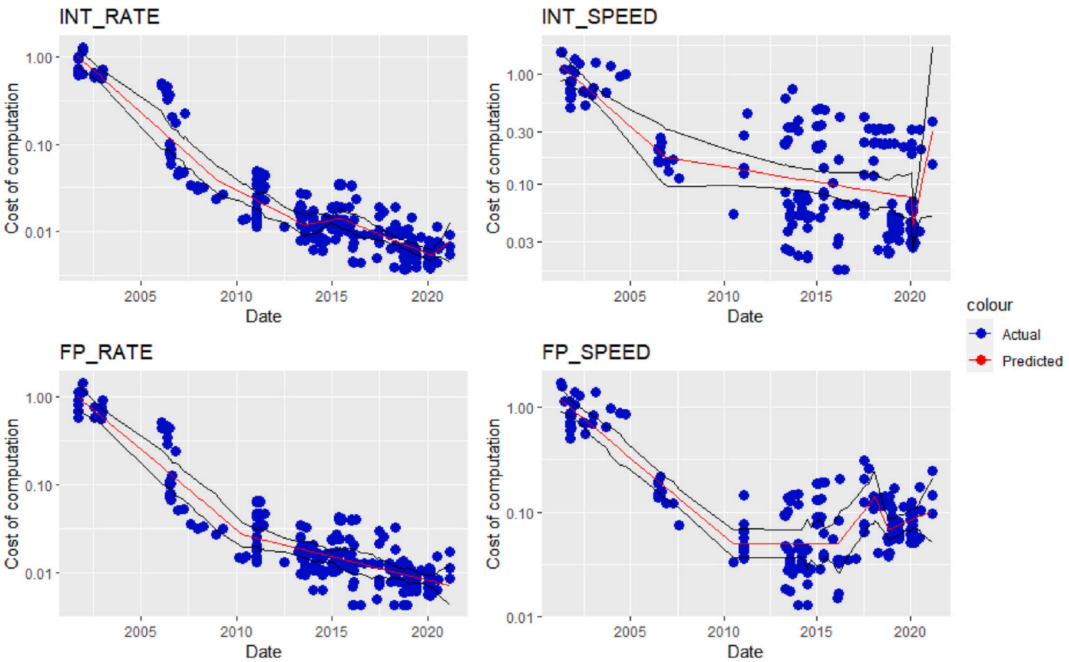

Fig. C2.1. Cost of computation over time using MARS. *Notes:* This figure shows the CPU cost of computation in 2021$ by benchmark using the MARS algorithm, with INT scores shown in the first row and RATE scores in the first column. 95% confidence intervals are obtained using a cross-validation procedure that accounts for the uncertainty in the selection of the knots: for more details see http://www.milbo.org/doc/earth-varmod.pdf. When generating INT_SPEED predicted values, the number of knots was reduced to avoid overfitting in the 2015–2020 period. Fitted values, confidence intervals and the index are shifted such that fitted values are equal to 1 in Jan 1, 2000. Outliers are removed from the sample. *Source:* as in Fig. 3. .

[20] Regression tables have not been included because standard errors, p values and confidence levels on the coefficients for MARS models are likely to be misleading: see section 13.8 of http://www.milbo.org/doc/earth-notes.pdf for details.

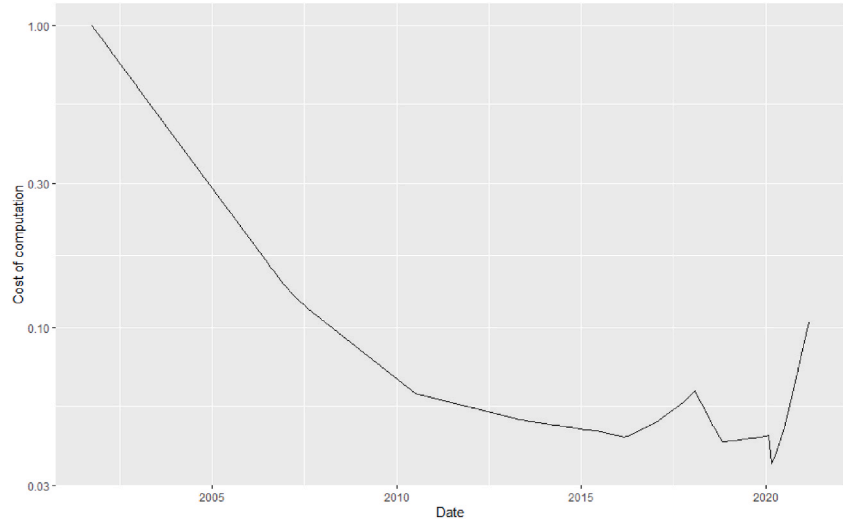

Fig. C2.2. Combined cost of computation over time using MARS. *Notes:* This figure shows the CPU cost of computation in 2021$ averaged over the four benchmarks. *Source:* as in Fig. 3. .

D Additional Figures

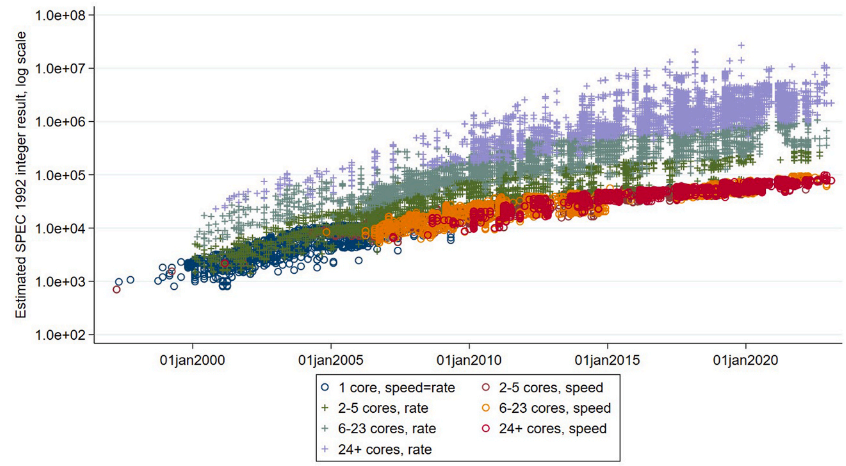

Fig. D.1. CPU Computing performance (integer), by cores
Notes: This figure plots estimated SPEC 1992 integer scores for SPEC CPU2000, SPEC CPU2006 and SPEC CPU2017 benchmarks. Different colours indicate the number of cores. RATE scores are indicated by plus markers. *Source:* SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations.

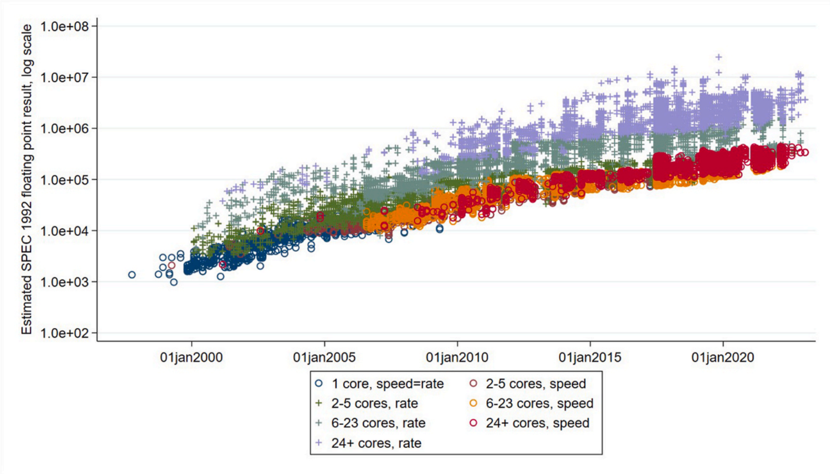

Fig. D.2. CPU Computing performance (floating point), by cores
Notes: This figure plots estimated SPEC 1992 floating point scores for SPEC CPU2000, SPEC CPU2006 and SPEC CPU2017 benchmarks. Different colours indicate the number of cores. RATE scores are indicated by plus markers.
Source: SPEC (https://www.spec.org/benchmarks.html#cpu) and authors' calculations.

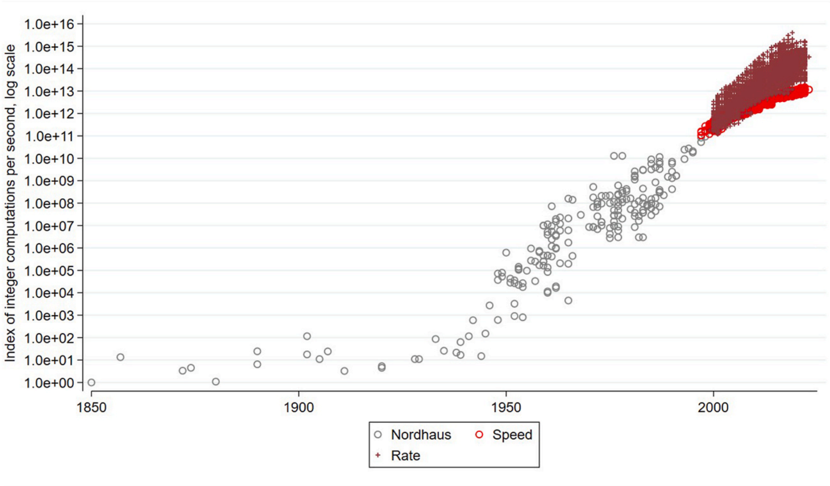

Fig. D.3. Long-run index of computer system integer computations per second, 1950-2023. *Notes:* This figure shows the long-run index of integer computations per second, anchored to human-level performance (1). Speed and rate scores are indicated separately. *Source:* The pre-2007 series uses data from 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.

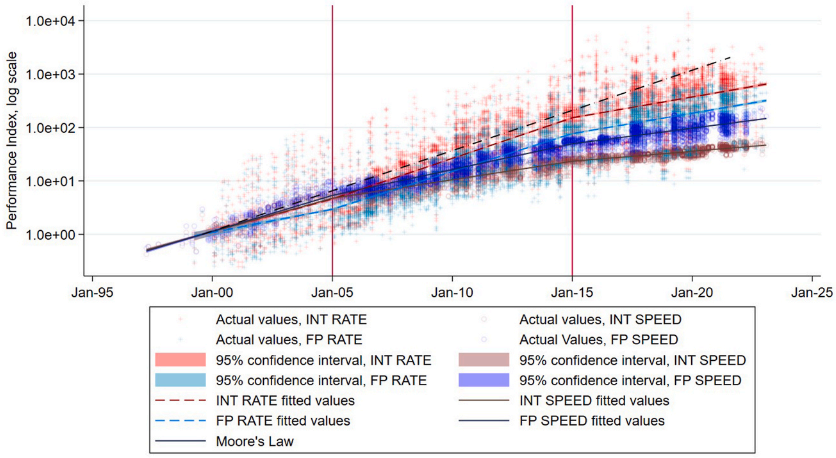

Fig. D.4. CPU performance index with actual values. *Notes:* This figure shows actual and fitted values of the relative performance index of computations per second at the CPU level over the date of hardware availability. The performance index is constructed using SPEC INT RATE (red), INT SPEED (brown), 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. 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 April 1997 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

	(1)	(2)	(3)	(4)
VARIABLES	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED
YEAR	0.180***	0.289***	0.158***	0.301***
	(0.021)	(0.005)	(0.018)	(0.006)
SPLINE2005	0.286***	−0.130***	0.204***	−0.066***
	(0.027)	(0.006)	(0.022)	(0.007)
SPLINE_K	−0.214***	−0.080***	−0.134***	−0.077***
	(0.012)	(0.003)	(0.010)	(0.004)
SPLINE2020	−0.187***	0.052***	−0.139***	−0.034***
	(0.029)	(0.006)	(0.028)	(0.012)
Constant	−5.259***	−10.902***	−4.162***	−11.322***
	(0.930)	(0.225)	(0.794)	(0.243)
Observations	7,900	5,302	7,430	4,836
R-squared	0.756	0.939	0.752	0.949
Breusch Pagan statistic	6.606	453.1	2.389	0.701
p-value	0.0102	0	0.122	0.402

Standard errors in parentheses, robust in cols 1 and 2.
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table D.2
CPU performance over time, outliers included

	(1)	(2)	(3)	(4)
VARIABLES	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED
YEAR	0.193***	0.289***	0.183***	0.295***
	(0.023)	(0.005)	(0.023)	(0.006)
SPLINE2005	0.255***	−0.130***	0.162***	−0.053***
	(0.029)	(0.006)	(0.026)	(0.008)
SPLINE_K	−0.198***	−0.081***	−0.117***	−0.129***
	(0.013)	(0.003)	(0.010)	(0.005)
SPLINE2020	−0.185***	0.055***	−0.141***	0.028
	(0.029)	(0.006)	(0.029)	(0.026)
Constant	−5.770***	−10.937***	−5.114***	−11.079***
	(1.001)	(0.239)	(1.002)	(0.273)
Observations	7,972	5,334	7,518	5,250
R-squared	0.733	0.935	0.718	0.854
Breusch Pagan statistic	13.86	455.9	50.52	660.4
p-value	0.000197	0	0	0

Robust standard errors in parentheses.
***p < 0.01, **p < 0.05, *p < 0.1.
Notes: K is 2010, 2015, 2012 and 2013 in columns 1–4 respectively. These tables show the results for the regression of relative CPU performance, measured in the natural logarithm of 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, SPLINE2005 is a variable equal to 0 if the year is less than 2005 and increments by 1 each year after 2005, and SPLINE_K and SPLINE2020 are defined similarly.
Sources: SPEC (https://www.spec.org/benchmarks.html#cpu), authors' calculations.

Table D.3
CPU price performance with outliers

	(1)	(2)	(3)	(4)
VARIABLES	INT, RATE	INT, SPEED	FP, RATE	FP, SPEED
YEAR	−0.347***	−0.178**	−0.376***	−0.120*
	(0.078)	(0.077)	(0.082)	(0.068)
SPLINE2005	−0.136	−0.043	−0.080	−0.405***
	(0.123)	(0.141)	(0.128)	(0.112)
SPLINE2010	0.342***	0.135	0.327***	0.583***
	(0.051)	(0.096)	(0.053)	(0.060)
SPLINE2020	0.196	1.154***	0.434***	0.334
	(0.123)	(0.378)	(0.153)	(0.213)
Constant	15.903***	8.500***	17.151***	6.127**
	(3.318)	(3.243)	(3.507)	(2.871)
Observations	299	210	281	217
R-squared	0.863	0.395	0.839	0.603
Breusch Pagan statistic	9.892	11.37	4.426	12.13
p-value	0.00166	0.000747	0.0354	0.000496

***p < 0.01, **p < 0.05, *p < 0.1.
Notes: This table shows spline regression results for the logged (base e) cost of computation index on Year and Time variable. YEAR is calendar year. SPLINE2005 is equal to 0 up to 2005 and increments by 1 after 2005. SPLINE2010 and SPLINE2020 are defined similarly. Outliers are included in these regressions. Heteroskedasticity-robust standard errors are in parentheses.
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.

Fig. 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	33.981		33.978	34.607	33.874
2005–2010	38.700	63.144	38.867	36.499	39.843
2010–2020	6.708	21.183	7.459	14.558	15.104
2020–2022	−39.106	20.710	−34.789	48.906	−1.342

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 Tables 2 and 4, and https://www.top500.org/statistics/overtime/ for Top500 shares.

References

Abdirahman, M., Coyle, D., Heys, R., & Stewart, W. (2020). A comparison of deflators for telecommunications services output. *Économie et Statistique/Economics and Statistics, 517–518*, 103–122. https://doi.org/10.24187/ecostat.2020.517t.2017. *–519*.

Abdirahman, M., Coyle, D., Heys, R., & Stewart, W. (2022). Telecoms deflators: A story of volume and revenue weights. *Économie et Statistique/Economics and Statistics, 530–31*, 43–59. https://doi.org/10.24187/ecostat.2022.530.2063

Azar, P. D. (2022). *Moore's Law and economic growth (federal reserve bank of New York staff reports, no. 970)*. Federal Reserve Bank of New York. https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr970.pdf?sc_lang=en.

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? *The 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*. https://doi.org/10.17016/FEDS.2022.038. SSRN Scholarly Paper No. 4158722).

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*. https://doi.org/10.2867/3688

Coyle, D., Lind, K., Nguyen, D., & Tong, M. (2022). Are digital-using UK firms more productive?. In *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. (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.

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

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*. https://doi.org/10.48550/arXiv.2303.10130 (arXiv:2303.10130). arXiv.

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., & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics, 31*(1), 3–21. https://doi.org/10.2307/1270359

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*. https://doi.org/10.48550/arXiv.2005.04305. arXiv:2005.04305). arXiv.

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), Article 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), Article 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

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). *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

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

Ramu, A. (2018). *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), Article 20190061. https://doi.org/10.1098/rsta.2019.0061

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.

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). *STATA Journal, 10*(3), 331–338. https://doi.org/10.1177/1536867X1001000302