2022

Sustainable Computing - Without the Hot Air

Noman Bashir
University of Massachusetts Amherst

David irwin
University of Massachusetts Amherst

Prashant Shenoy
University of Massachusetts Amherst

Abel Souza

Bashir, Noman; irwin, David; Shenoy, Prashant; and Souza, Abel, "Sustainable Computing - Without the Hot Air" (2022). *Proceedings of the First Workshop on Sustainable Computer Systems Design and Implementation (HotCarbon)*. 4.
https://doi.org/10.48550/arXiv.2207.00081

Sustainable Computing – Without the Hot Air*

Noman Bashir, David Irwin, Prashant Shenoy, Abel Souza
University of Massachusetts Amherst

Abstract

The demand for computing is continuing to grow exponentially. This growth will translate to exponential growth in computing's energy consumption unless improvements in its energy-efficiency can outpace increases in its demand. Yet, after decades of research, further improving energy-efficiency is becoming increasingly challenging, as it is already highly optimized. As a result, at some point, increases in computing demand are likely to outpace increases in its energy-efficiency, potentially by a wide margin. Such exponential growth, if left unchecked, will position computing as a substantial contributor to global carbon emissions. While prominent technology companies have recognized the problem and sought to reduce their carbon emissions, they understandably focus on their successes, which has the potential to inadvertently convey the false impression that this is now, or will soon be, a solved problem. Such false impressions can be counterproductive if they serve to discourage further research in this area, since, as we discuss, eliminating computing's, and more generally society's, carbon emissions is far from a solved problem. To better understand the problem's scope, this paper distills the fundamental trends that determine computing's carbon footprint and their implications for achieving sustainable computing.

1 Introduction

The demand for computing is growing exponentially, and has been for some time [20], mostly because society continues to find useful applications for it. These applications have transformed modern society over the past fifty years, and now largely serve as the foundation of its information-based economy. Since computation is simply a processed form of energy, and energy use incurs both a monetary and environmental cost, there has long been a concern that exponentially growing computing demand would translate into exponentially growing energy demand, which could stifle technological innovation and damage the environment. For example, as early as 2007, the U.S. Environmental Protection Agency (EPA) sent a report to Congress projecting a doubling in aggregate data center energy consumption if historical trends continued [1]. While follow-up analyses in 2011 [31] and 2016 [41] suggested the 2007 forecast was inaccurate (with data center energy consumption increasing by only 24% from 2005-2010 and 4% from 2010-2014), more recent studies have come to widely different estimates of the growth in data center energy consumption [16–18, 24, 29]. The most optimistic of these analyses estimates only a 6% increase in data center energy consumption from 2010-2018, or roughly an average 0.65% increase per year, despite a 6× increase in capacity [36].

Ultimately, computing's aggregate energy consumption E (in kWh) is simply a function (shown below) of its demand d (in cycles) versus its energy-efficiency e (in cycles/kWh).

$$E = \frac{d}{e} \tag{1}$$

Likewise, the growth rate r_E in computing's energy consumption is simply a function of the growth rate in its demand (r_d) versus its energy-efficiency (r_e): if demand increases faster than energy-efficiency, i.e., $r_d > r_e$, then consumption will grow exponentially, and otherwise, it will shrink.

The optimistic analysis above attributes the small increase in computing's energy consumption to the incredible increase in computing's energy-efficiency over the past 15 years. This trend largely derives from computing's ongoing transition from smaller traditional data centers to "hyperscale" cloud platforms, which have a strong financial incentive to optimize their energy-efficiency to reduce operational costs. Indeed, the reported Power Usage Effectiveness (PUE) for Google's data centers—the ratio of their total energy to the energy of IT equipment—is now ∼1.1 [10], which is near the optimal value of 1 and also nearly 30% lower than the industry average of ∼1.57 [11]. However, while the trends above have been broadly characterized as a tremendous success story for industry, which they undoubtedly are, they also belie a significant problem. Specifically, *despite incredible improvements in computing's energy-efficiency over the past 15 years, by even the most optimistic analysis, its aggregate energy consumption still increased!* That is, $r_d > r_e$ over this period.

Unfortunately, after decades of research, further improving energy-efficiency is becoming increasingly challenging, as it is already highly optimized. Thus, moving forward, increases in computing's energy-efficiency are likely to slow, especially once its transition to the cloud is complete. At the same time, demand is likely to continue increasing, if not accelerate, as new useful applications are developed. For example, recent progress in AI has the potential to enable a wide range of novel applications that are also computationally-intensive [2]. Importantly, these trends will not only have profound implications on computing's cost, but also its carbon emissions. Exponentially growing demand that is not offset by energy-

*Title inspired by the book Sustainable Energy - Without the Hot Air [35].

efficiency improvements, would quickly position computing as a substantial contributor to global carbon emissions. Yet, at the same time, there is now a broad consensus that society must rapidly reduce, and ultimately eliminate, its carbon emissions to halt climate change, which represents an existential threat to the earth's ecosystem and humanity.

Similar to the relationship above, computing's carbon footprint C (in g·CO_2) is simply a function (shown below) of its aggregate energy consumption E from Equation 1 versus its energy's carbon-efficiency c (in kWh/g·CO_2).

$$C = \frac{E}{c} = \frac{d}{c \times e} \tag{2}$$

Likewise, the growth rate r_C in computing's carbon footprint is simply a function of the growth rate in its aggregate energy consumption (r_E) versus its energy's carbon-efficiency (r_c): if consumption increases faster than carbon-efficiency, i.e., $r_E > r_c$, then it will grow exponentially, and otherwise, it will shrink. We can also use Equation 1 to substitute d/e for E in Equation 2. Here, $c \times e$ represents computing's carbon-efficiency (in cycles/g·CO_2), and highlights that computing's energy-efficiency and energy's carbon-efficiency are equally important in determining computing's carbon footprint.

The most recent estimates suggest electricity's carbon-intensity (in g·CO_2/kWh), which is the inverse of carbon-efficiency, in the U.S. decreased 30% between 2001 and 2017, largely due to the replacement of coal-fired power plants with natural gas and wind generation [30, 40]. This is equivalent to a 45.6% increase in energy's carbon-efficiency over the same period, or equivalently a ∼2.33% increase per year. Thus, the most optimistic assessments based on the reported averages above—a 0.65% per year increase in energy consumption [36] and a 2.33% per year increase in energy's carbon-efficiency [40]—suggest that computing's carbon footprint decreased slightly (by ∼1.64% per year on average[1]) from 2010-2017, and that this decrease was entirely due to improvements in energy's carbon-efficiency. In contrast, if the most optimistic assessments are inaccurate, then, based on the same reasoning, computing's carbon footprint likely increased.

Of course, the macro longitudinal analyses cited above are necessarily simplistic, coarse, and imprecise, and should be taken with a grain of salt, i.e., viewed skeptically. For example, our analysis does not take into account that electricity's carbon-efficiency varies widely across days, seasons, and regions, and thus it is also a function of when and where energy is consumed. That said, macro analyses can be useful in distilling the fundamental trends that matter. In particular, irrespective of the specific numbers, Equation 2 shows that the growth of computing's carbon footprint is based on the relative growth in its demand, energy-efficiency, and energy's carbon-efficiency. Thus, better understanding these relative growth rates can provide some insight into how computing's carbon footprint is changing, and also how to eliminate it.

Our simple analysis also paints a slightly different, and more nuanced, picture of computing's carbon footprint than recent industry announcements [3, 23, 37, 42]. While prominent technology companies have recognized the trends above and sought to reduce their carbon emissions, they understandably focus on their successes, which has the potential to inadvertently convey the false impression that this is now, or will soon be, a solved problem. This paper's title is a reference to a well-known book that made a similar observation about the energy industry [35]. For example, many technology companies have eliminated their net carbon emissions [3, 23, 37, 42], which they often refer to as running on "100% renewable energy." However, eliminating net carbon emissions is both different and much easier than eliminating direct carbon emissions. Unfortunately, such false impressions can be counterproductive if they unintentionally discourage further research, since, as we discuss, eliminating computing's, and more generally society's, real carbon emissions is far from a solved problem. To better understand the problem's scope, we examine relative trends in the growth of computing's demand and energy-efficiency, as well as its energy's carbon-efficiency, and their implications for achieving sustainable computing.

2 Computing's Demand

By all indications, the demand for computing—the total number of cycles executed—has been growing exponentially for some time, likely since the dawn of computing [20]. The optimistic analysis above estimated a 6× increase in data center capacity from 2010-2018 (or ∼22% per year) [36]. Another recent study estimated that the capacity for the most efficient hyperscale data centers had doubled over the past five years [5]. While some of this growth surely represents existing demand transitioning from smaller traditional data centers to cheaper cloud platforms, much of it also likely represents new demand from cloud-native applications. For example, a recent report estimates that 75% of companies are now focusing on developing cloud-native applications [12]. A variety of other anecdotal evidence suggests computing demand may be accelerating. For example, the cycles devoted to cryptomining [8] and training state-of-the-art machine learning (ML) models [39] is growing much faster than Moore's Law. Computing is also continually displacing other activities, such as videoconferencing in lieu of traveling for meetings. While such displacement may improve energy-efficiency, which we discuss below, it undoubtedly increases computing's demand.

The only way to reduce computing demand (aside from not computing) is to improve algorithmic efficiency by enabling computation to do more (or the same) work using fewer cycles. To be sure, there are numerous and substantial remaining opportunities to improve algorithmic efficiency. For example, broad adoption of proof-of-stake consensus for cryptocurrencies would effectively eliminate soaring demand

[1] Carbon footprint's growth rate is (M-N)/(1+N), where M and N are the growth rates in energy's consumption and carbon-efficiency, respectively.

from cryptomining. Likewise, reducing the demand to train large-scale ML models has been a focus of recent research, and yielded some notable improvements [21,43]. Importantly, though, computing's demand is not only a function of each applications' efficiency, but also the total number of applications executed. That is, improving the efficiency of training ML models by 10× will not decrease demand if the number of models trained increases by 10×. Absent resource constraints, computing's potential applications still seem limitless, or at least only limited by people's imaginations. Thus, improvements to algorithmic efficiency may be hard-pressed to offset the growth in the sheer number of applications executed.

Finally, while industry has a strong incentive to increase algorithmic efficiency to reduce their operational cost, it is bounded by each problem's computational complexity. Obviously, we cannot solve computational problems without some minimal amount of computation. Further, industry's primary incentive is to increase its potential profit, which is effectively unbounded and generally correlates with increasing demand, regardless of efficiency. That is, while improving efficiency may increase profit, it is not always necessary or possible.

Key Point. *Computing demand is increasing, and possibly accelerating, as more useful applications are developed. Improvements to algorithmic efficiency are bounded and thus unlikely to staunch this growth over the long-term.*

3 Computing's Energy-Efficiency

Computing's energy-efficiency has also been increasing at an exponential rate for some time, a trend that has been referred to as Koomey's Law [32]. Koomey estimated that computing's energy-efficiency at peak capacity has been approximately doubling every 1.57 years from the 1950s up through 2010 (roughly in-line with Moore's law) [32], although a revised analysis suggested the pace slowed to every 2.6 years starting in 2000 (due, in part, to the end of Dennard scaling) [33]. Since computing platforms are often idle, the same report also estimated that average energy-efficiency, which considers idle periods, had continued to double every ∼1.5 years due to increases in average utilization and energy-proportionality. This latter point captures some of the energy-efficiency improvements from the transition to cloud platforms, which leverage statistical multiplexing at massive scales to increase average server utilization, as servers are more energy-efficient at higher utilization. As noted earlier, hyperscale cloud data centers have also improved their facilities' energy-efficiency by driving down their PUEs to within 10% of optimal [10].

The optimistic analysis from §1 estimated that the energy-efficiency improvements above have nearly kept computing's energy consumption constant, despite its exploding demand [36]. Indeed, the transition from highly inefficient small traditional data centers to highly efficient hyperscale data centers has yielded dramatic increases in energy-efficiency. Moreover, this transition is not yet complete with recent estimates suggesting nearly 20% of data center energy consumption still derives from traditional smaller, and less efficient, facilities, so there is still room for further improvement [14].

Yet, continuing to increase computing's energy-efficiency to keep pace with increases in demand may prove challenging for many reasons. Most importantly, the shift to hyperscale cloud data centers, which has yielded much of the improvement above, is a one-time event. Once the shift is complete, it is unclear where significant improvements will come from. One possibility is increasing the use of specialized hardware, which is more energy-efficient than general-purpose platforms. For example, cryptomining and ML have employed hardware specifically tailored to their function to dramatically increase their energy-efficiency (and performance). However, much of computing's demand remains general-purpose, with specialized tasks still constituting only a small fraction of it. For example, a recent paper estimates that only 15% of Google's energy consumption is due to ML [38]. More generally, improving computing's energy-efficiency has been a significant focus of research for at least three decades. Thus, there are likely few remaining substantial optimization opportunities using traditional methods, which may be one reason for the reported slowing of Koomey's law [33].

As with algorithmic efficiency, there is also a well-known physical limit to the energy-efficiency of our current form of computing, which is defined by Landaur's principle [34]. Current estimates are that if computing's energy-efficiency were to continue to double every ∼1.5 years, then it would reach this physical limit by 2050 [32], although it is not yet known how close CMOS circuits can, in practice, come to this limit. While, in theory, adopting reversible computing techniques can overcome Landaur's limit by performing computation without consuming any energy, it is a nascent, and largely theoretical, area that is far from any practical application [25].

Finally, even if computing's energy-efficiency were to continue doubling every 1.5 years, there is no guarantee it would cause computing's energy consumption to decrease. As noted earlier, even by the most optimistic estimates, computing's incredible energy-efficiency improvements have not reduced its energy consumption thus far. Interestingly, whether increases in energy-efficiency actually decrease energy's consumption is, in part, a function of economics. Specifically, as computing's energy-efficiency improves, its energy cost generally decreases, which in-turn affects its demand. The magnitude of this effect is a function of computing's price elasticity of demand, which dictates how much demand changes when prices change. Jevons Paradox, which is well-known in energy economics, occurs when demand elasticity is high enough that the increases in energy consumption from higher demand (caused by lower costs) is greater than the decrease in consumption from improved energy-efficiency [15, 44]. Thus, under Jevon's Paradox, improved energy-efficiency actually, and paradoxically, can lead to increased energy consumption. Even if Jevons Paradox does not occur, assessing the effect of

increases in computing's energy-efficiency on its energy consumption is largely an economic, and not technical, question.

Of course, improving computing's energy-efficiency is always beneficial, as it increases productivity and economic output, i.e., enables more to be done with less energy at lower cost. Thus, as with improving algorithmic efficiency, industry has a strong financial incentive to improve energy-efficiency. This incentive has likely driven the incredible energy-efficiency improvements over the past fifty years. However, improvements in computing's energy-efficiency do not necessarily, and have not historically, led to reductions in its energy consumption. In fact, if Jevons Paradox occurs, improving computing's energy-efficiency may contribute to increasing energy consumption.

Key Point. *Computing's energy-efficiency is continuing to increase, although its rate may be slowing. Improvements to computing's energy-efficiency are bounded, and do not necessarily, and have not historically, led to reductions in computing's energy consumption, due to faster growth in demand both from new applications and lower costs.*

4 Energy's Carbon-Efficiency

Unlike algorithmic- and energy-efficiency, there is no fundamental limit to energy's carbon-efficiency, since it is possible to use zero-carbon energy sources, such as solar, wind, geothermal, hydroelectric, nuclear, etc. The cost for solar and wind renewable energy sources, in particular, have also been decreasing exponentially for some time. Swanson's law, which captures this trend for solar energy, refers to the observation that solar photovoltaic PV module prices have tended to drop 20% for every doubling in production volume [45]. As a result, solar energy's cost (in $/watt) has dropped ∼10% each year on average over the past fifty years [19]. Renewable energy sources also have massive energy potential that could fuel exponential growth for the foreseeable future. For example, the amount of solar energy the earth receives each hour is more than global annual energy consumption [28].

As mentioned in §1, energy's carbon-efficiency has been steadily increasing for the past 20 years, mostly due to the adoption of natural gas and wind. This trend has been independent of any efforts by the computing industry to reduce its carbon footprint. However, isolating and capturing the trend the carbon-efficiency of computing's energy is more challenging, as it depends on the strictness of carbon accounting and attribution methods used. Carbon offsets are the loosest, and most widely used, method of carbon accounting: they enable "offsetting" the use of carbon-intensive grid energy with zero-carbon renewable energy generated at another location and time. Technology companies have led in the adoption of carbon offsets, and many have used them to eliminate their net carbon footprint, which is often referred to as running on 100% renewable energy [3, 23, 37]. However, while carbon offsets are beneficial in subsidizing renewable energy, they are only a temporary mechanism as society transitions to lower carbon energy, since near zero-carbon there will not be any carbon left to offset. In addition, the use of carbon offsets means that even net zero companies are still responsible for a significant amount of direct carbon emissions.

To reach zero-carbon, companies must progressively adopt stricter forms of carbon accounting. To this end, Google recently announced that it aims to be "carbon free" by 2030, in part, by piloting a stricter form of carbon offset, called Time-based Energy Attribute Certificates (TEACs), which have an hourly location-specific accounting regime [7]. However, TEACs are still carbon offsets, just at a higher temporal and spatial resolution than typical offsets, which are usually 1 year and the entire earth, respectively. That is, TEACs match consumption of grid energy within an hour to renewable generated that hour within the same grid. Thus, while TEACs are an improvement upon existing annualized location-agnostic carbon offsets, they, by definition, also cannot be used to reach zero-carbon. Of course, since the grid cannot physically isolate different energy sources, in reality, all loads that consume grid energy share in its carbon emissions. Thus, the strictest form of carbon accounting attributes the grid's carbon emissions to all its loads based on their energy use. As a result, reducing and ultimately eliminating computing's carbon emissions will require changing its operations to be responsive to variations in grid energy's carbon emissions and availability.

Thus far, we have focused on trends in operational carbon, i.e., carbon emissions from using grid energy. There has also been an increasing focus on accounting for and reducing "embodied carbon," which represents the carbon emissions from producing a product or service [22, 26, 27]. For example, computing's embodied carbon emissions are based on the carbon emissions from manufacturing the facilities and IT equipment that host it. Importantly, though, one company's embodied carbon is another company's operational carbon. For example, a cloud platform's embodied carbon is, in part, a chip manufacturer's operational carbon. The primary purpose in accounting for embodied carbon is to provide an incentive in the supply chain for companies to reduce their operational carbon. That is, if companies made purchasing decisions to reduce their embodied carbon, it would incentive upstream suppliers to in-turn reduce their operational carbon. Accounting for embodied carbon is akin to a value added tax (VAT), as carbon emissions, similar to a VAT, are associated with the value added at each production stage of a good or service.

Unfortunately, unlike with algorithmic- and energy-efficiency, there are not yet strong financial incentives for companies to reduce their operational or embodied carbon emissions, as energy prices do not yet incorporate the cost of carbon's negative externalities to the environment. As a result, while energy's carbon-efficiency has been improving, and is unbounded, its long-term trend is unclear.

Key Point. *Energy's carbon-efficiency is increasing, although its long-term trend is unclear due to the lack of financial incentive to improve it. Improvements to energy's carbon-efficiency*

are unbounded, as it is possible to only use zero-carbon energy. Since there will be no carbon offsets at zero-carbon, eliminating computing's carbon emissions will ultimately require eliminating the grid's carbon emissions.

5 Implications for Sustainable Computing

The trends above have important implications for sustainable computing moving forward. Specifically, given the fundamental limits to improving computing's algorithmic- and energy-efficiency, the only way to sustain exponential growth in its demand, while also eliminating its carbon footprint, is to improve its energy's carbon-efficiency. However, the trends in energy's carbon-efficiency are not yet clear. In particular, the terminology above around different forms of carbon accounting, e.g., "100% renewable energy," "carbon-neutral," "carbon-free," "zero-carbon," 'embodied carbon," etc., is complex and fully understanding it requires some non-trivial technical background on how society's energy system works. To anyone without such a background, which includes much of the general public as well as many computing researchers, the use of the terms above may inadvertently convey the false impression that computing's carbon emissions are already at zero, or soon will be. Such messaging is often pejoratively referred to as "greenwashing." False impressions of computing's carbon footprint are a significant issue, as they can diminish the perception of progress in decarbonizing computing, or even discourage further research altogether.

In the end, as we discuss, the various forms of carbon accounting and offsets are temporary measures that, by definition, will not be applicable at zero-carbon. To reach zero-carbon, computing, and more generally society, will have to significantly change how it operates to directly use renewable and low-carbon energy. Of course, the problem with renewable energy is that, while it is potentially plentiful, cheap, and clean, it is also highly unreliable. In particular, solar and wind vary widely and uncontrollably over time based on the earth's movement and weather. As a result, transitioning the grid to operate entirely on zero-carbon energy will require either i) significant over-provisioning within the energy system, e.g., of batteries, solar, wind, etc., which is likely cost-prohibitive, or ii) significant flexibility in the system's loads.

Fortunately, compared to other loads, computing is uniquely flexible with substantial performance, temporal, and spatial flexibility, enabling it to shift the intensity, time, and location of its execution to better align with when and where renewable and other low-carbon energy is available. To the best of our knowledge, computing is the only load with substantial spatial flexibility that is capable of migrating its energy consumption over long distances. In addition, computing can also leverage numerous software-based fault-tolerance techniques, e.g., checkpointing, replication, and recomputation, to continue execution despite unexpected renewable shortages, which may require throttling or shutting down servers. Thus, computing has the potential to leverage its multiple dimensions of flexibility to not only lower its direct carbon footprint, but offset variations in renewable energy's availability. As a result, computing is not just another grid load, as it can also act as an energy resource, akin to a battery, that the grid can deploy to balance demand with a variable supply [13]. In some sense, improving energy's carbon-efficiency is related to improving computing's energy-flexibility by enabling it to adapt to when and where low-carbon energy is available.

While many have recognized computing's unique dimensions of energy flexibility, there has been much less research on exercising them to optimize energy's carbon-efficiency compared to computing's energy-efficiency, even though, as Equation 2 shows, carbon-efficiency is just as important as energy-efficiency in determining computing's carbon footprint. One reason for the lack of research is likely that, unlike with algorithmic- and energy-efficiency, there is neither a direct nor strong financial incentive to improve energy's carbon-efficiency, although this may change as renewable energy prices drop. That said, there is a weak, but increasing, indirect incentive to track and improve energy's carbon-efficiency both to appeal to environmentally-conscious consumers (and employees), and as a hedge against future changes in the energy system, such as energy constraints due to geopolitical events, stricter carbon regulations imposed by governments, or further significant drops in renewable or battery prices.

Another reason for the lack of research may also be that optimizing carbon-efficiency requires deeper visibility into energy's carbon emissions, which has historically not been available. Recently, carbon information services, such as electricityMap [4] and WattTime [6], have emerged, and are beginning to address this issue by tracking grid energy's carbon emissions for different regions over time, and making them available online. The data shows that grid energy's carbon emissions vary significantly by region and over time. Cloud platforms have started adopting these services to enable their users to estimate the carbon emissions of their energy consumption, and adjust their operations to reduce emissions [9].

Ultimately, the primary implication for achieving sustainable computing from the trends above is that research should emphasize improvements to the carbon-efficiency of both computing's energy (by adapting to when and where low-carbon energy is available), as well as the grid's energy (by leveraging computing as an energy resource). The former is important for reducing computing's direct carbon emissions, while the latter is important for reducing society's carbon emissions, which are related and also affect embodied carbon. Given the lack of a strong financial incentive to improve carbon-efficiency, academic research in this area is especially important. Indeed, historically, an explicit purpose of academic research has been to focus on problems that industry does not address due to lack of a near-term financial incentive.

Acknowledgements. This research is supported by NSF grants 2105494, 2021693, 2020888, as well as VMware.

References

[1] EPA Report to Congress on Server and Data Center Energy Efficiency. Technical report, U.S. Environmental Protection Agency, August 2007.

[2] OpenAI Blog, AI and Compute. https://openai.com/blog/ai-and-compute/, March 16th 2018.

[3] Reuters, Amazon Vows to be Carbon Neutral by 2040, buying 100,000 Electric Vans. https://www.reuters.com/article/us-amazon-environment/amazon-vows-to-be-carbon-neutral-by-2040-/buying-100000-electric-vans-idUSKBN1W41ZV, September 19th 2019.

[4] Electricity Map. https://www.electricitymap.org/map, Accessed September 2020.

[5] Hyperscale Data Center Count Reaches 541 in Mid-2020; Another 176 in the Pipeline. Technical report, Synergy Research Group, 2020.

[6] WattTime. https://www.watttime.org/, Accessed September 2020.

[7] 24/7 by 2030: Realizing a Carbon-free Future. https://www.gstatic.com/gumdrop/sustainability/247-carbon-free-energy.pdf, Accessed May 2022.

[8] Digiconomist, Bitcoin Energy Consumption Index. https://digiconomist.net/bitcoin-energy-consumption, Accessed May 2022.

[9] Google Cloud Carbon Footprint Console. https://cloud.google.com/carbon-footprinthttps://cloud.google.com/carbon-footprint, Accessed June 2022.

[10] Google Data Centers: Efficiency. http://google.com/about/datacenters/efficiency/, Accessed May 2022.

[11] Uptime Institute Global Data Center Survey 2021: Growth Stretches an Evolving Sector. https://uptimeinstitute.com/resources/asset/2021-data-center-industry-survey, Accessed May 2022.

[12] VentureBeat, report: 75% of Companies are Focusing on Cloud-Native Apps. https://venturebeat.com/2022/05/04/report-75-of-companies-are-focusing-on-/cloud-native-apps/, May 4th 2022.

[13] A. Agarwal, J. Sun, S. Noghabi, S. Iyengar, A. Badam, R. Chandra, S. Seshan, and S. Kalyanaraman. Virtual battery: Redesigning cloud computing for renewable energy. In *HotNets*, November 2021.

[14] International Energy Agency. Global Data Centre Energy Demand by Data Centre Type, 2010-2022. https://www.iea.org/data-and-statistics/charts/global-data-centre-energy-demand-by-data-/centre-type-2010-2022, March 2021.

[15] B. Alcott. Jevons' Paradox. *Ecological Economics*, 54(1):9–21.

[16] Anders SG Andrae. Projecting the Chiaroscuro of the Electricity Use of Communication and Computing from 2018 to 2030. *Preprint*, 2019.

[17] Anders SG Andrae and Tomas Edler. On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges*, 2015.

[18] Lotfi Belkhir and Ahmed Elmeligi. Assessing ICT Global Emissions Footprint: Trends to 2040 & Recommendations. *Journal of Cleaner Production*, 2018.

[19] A. de la Tour, M. Glachant, and Y. Ménière. What Cost for Photovoltaic Modules in 2020? Lessons from Experience Curve Models. Technical report, Interdisciplinary Institute for Innovation, May 2013.

[20] Peter J. Denning and Ted G. Lewis. Exponential Laws of Computing Growth. *Communications of the ACM*, 60(1):54–65, January 2017.

[21] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathy Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts. Technical report, Google Inc., December 2021.

[22] Carole-Jean Wu et al. Sustainable AI: Environmental Implications, Challenges and Opportunities. In *MLSys*, August 2022.

[23] Darrell Etherington. TechCrunch, Google Claims Net Zero Carbon Footprint over its Entire Lifetime, Aims to only use Carbon-Free Energy by 2030. https://techcrunch.com/2020/09/14/google-claims-net-zero-carbon-footprint-/over-its-entire-lifetime-aims-to-only-use/-carbon-free-energy-by-2030/, September 14th 2020.

[24] Hugues Ferreboeuf. LEAN ICT- Towards Digital Sobriety. https://theshiftproject.org/en/article/lean-ict-our-new-report/, March 2019.

[25] Michael P. Frank. Foundations of Generalized Reversible Computing. In *Conference on Reversible Computation*, June 2017.

[26] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool. In *ISCA*, June 2022.

[27] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing Carbon: The Elusive Environmental Footprint of Computing. In *HPCA*, February 2021.

[28] R. Harrington. Business Insider, This incredible fact should get you psyched about solar power. https://www.businessinsider.com/this-is-the-potential-of-solar-power-2015-9, September 29th 2022.

[29] Ralph Hintemann. Efficiency Gains are Not Enough: Data Center Energy Consumption Continues to Rise Significantly. Technical report, Borderstep Institute for Innovation and Sustainability, 2018.

[30] Stephen P. Holland, Matthew J. Kotchen, Erin T. Mansur, and Andrew J. Yates. Why Marginal CO2 Emissions Are Not Decreasing for U.S. Electricity: Estimates and Implications for Climate Policy. *Proceedings of the National Academy of Sciences*, 119(8):e2116632119, 2022.

[31] J. Koomey. Growth in Data Center Electricity Use 2005 to 2010. https://www.koomey.com/research.html, 2011.

[32] Jonathan Koomey, Stephen Berard, Maria Sanchez, and Henry Wong. Implications of Historical Trends in the Electrical Efficiency of Computing. *IEEE Annals of the History of Computing*, 33(3):46–54, March 2010.

[33] Jonathan Koomey and Samuel Naffziger. Moore's Law Might be Slowing Down, but not Energy Efficiency. *IEEE spectrum*, 2015.

[34] R. Landauer. Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development*, 5(3):183–191, July 1961.

[35] David MacKay. *Sustainable Energy - Without the Hot Air*. UIT cambridge, 2008.

[36] Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. Recalibrating Global Data Center Energy-use Estimates. *Science*, 367(6481):984–986, February 2020.

[37] Kevin O'Sullivan. The Irish Times, Facebook Commits to Net-Zero Carbon Emissions by 2030. https://www.irishtimes.com/news/environment/facebook-commits-to-net-zero-carbon-/emissions-by-2030-1.4354701, September 15th 2020.

[38] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Technical report, Google Inc., April 2022.

[39] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. The AI Index 2019 Annual Report. *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*, 2019.

[40] Greg Schivley, Ines Azevedo, and Constantine Samaras. Assessing the Evolution of Power Sector Carbon Intensity in the United States. *Environmental Research Letters*, 13(064018), June 2018.

[41] Arman Shehabi, Sarah Josephine Smith, Dale A. Sartor, Richard E. Brown, Magnus Herrlin, Jonathan G. Koomey, Eric R. Masanet, Nathanial Horner, Ines Lima Azevedo, and William Linter. United States Data Center Energy Usage Report. Technical Report LBNL-1005775, Lawrence Berkeley National Lab (LBL), June 2016.

[42] Brad Smith. Official Microsoft Blog, Microsoft will be Carbon Negative by 2030. https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/, January 16th 2020.

[43] David R. So, Wojciech Manke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V. Le. Primer: Searching for Efficient Transformers for Language Modeling. In *NeurIPS*, December 2021.

[44] S. Sorrell. Jevons' Paradox Revisited: The Evidence for Backfire from Improved Energy Efficiency. 37(4):1456–1469, 2009.

[45] R. Swanson. A Vision for Crystalline Silicon Photovoltaics. *Progess in Photovoltaics: Research and Applications*, 14(5), August 2006.