Metrics for Sustainability in Data Centers

ANSHUL GANDHI, DONGYOON LEE, ZHENHUA LIU, SHUAI MU, and EREZ ZADOK, Stony Brook University, USA
KANAD GHOSE, KARTIK GOPALAN, and YU DAVID LIU, Binghamton University, USA
SYED RAFIUL HUSSAIN, Pennsylvania State University, USA
PATRICK MCDANIEL, University of Wisconsin, Madison, USA

Despite several calls from the community for improving the sustainability of computing, sufficient progress is yet to be made on one of the key prerequisites of sustainable computing—the ability to define and measure computing sustainability holistically. This position paper proposes metrics that aim to measure the end-to-end sustainability footprint in data centers. To enable useful sustainable computing efforts, these metrics can track the sustainability footprint at various granularities—from a single request to an entire data center. The proposed metrics can also broadly influence sustainable computing practices by incentivizing end-users and developers to participate in sustainable computing efforts in data centers.

CCS Concepts: • **Social and professional topics** → **Sustainability**; • **Hardware** → **Energy metering**; **Enterprise level and data centers power issues**.

Additional Key Words and Phrases: sustainability, data centers, metrics, carbon footprint, operational carbon, embodied carbon.

1 INTRODUCTION

Sustainability is a societal challenge that must be addressed on all fronts. Data centers already contribute significantly to the global carbon footprint [48, 56]. Worse, the rise in popularity of resource-intensive Big Data, AI, cryptocurrency, and Machine Learning workloads is poised to make data center operations unsustainable [6, 35, 38, 43, 47, 61]. While existing hardware and virtualization technologies have helped regulate the energy consumed by data center servers, they ignore other important contributors to unsustainable computing, such as manufacturing and disposal costs. Achieving true "sustainable" data center computing will require taking into account all factors that contribute to the sustainability footprint of computing, including the manufacture, repair, and disposal/recycling of IT equipment.

A necessary first step for any sustainable computing approach is the ability to define and measure comprehensive sustainability metrics or cost functions. As business management thinker Peter Drucker rightly said "if you can't measure it, you can't improve it." The key question that we consider in this position paper is ***how do we define comprehensive metrics and cost functions that capture the sustainability footprint of data centers?***

Ideally, the metrics should allow sustainability accounting at various *granularities*—from a single request to an entire data center. This is necessary as system-scale metrics are inadequate in quantifying the overall sustainability impact of individual contributors, such as programming techniques used by developers, management practices used within virtualized systems operated by a renter in a colocated installation, etc. The metrics should also track sustainability costs across computing devices, including shared (*e.g.*, virtualized) computing environments. Finally, the metrics should be accurate, easy to measure, reproducible, and useful (*e.g.*, to encourage sustainable computing).

The eventual goal of this work is to develop metrics that lead to market-based strategies to incentivize sustainable computing. To that end, this position paper presents *new candidate metrics for sustainability in data centers*. We start by describing the various factors, beyond IT equipment energy (or "operational energy"), which contribute to the sustainability footprint of data centers. For instance, one significant factor that is often ignored is the energy spent in the entire lifecycle—production, delivery, and disposal—of IT equipment (referred to as "embodied energy"). Next, we discuss the various challenges involved in measuring and obtaining the end-to-end sustainability costs of these factors and potential solutions to address the challenges. We then present our proposed sustainability metrics that (i) take into account various sustainability costs incurred throughout the entire lifecycle of IT equipment, including wear-and-tear, and (ii) assign appropriate sustainability costs to individual jobs.

A key consideration in designing sustainability metrics is the *units* of reporting. While carbon-based units have largely been employed in practice [44, 63, 65], we argue that not all sustainability footprint factors (*e.g.*, disposal of IT equipment, water usage) can be easily quantified in such units. Further, it is not entirely clear whether such carbon-based reporting units are sufficient to promote sustainable computing among end-users of computing services. For example, would a cloud user be sufficiently incentivized to participate in sustainable computing efforts if they were informed that their workload contributed to, say, 2.5 grams of CO_2 emissions?

We posit that there is room for additional, complementary units of sustainability reporting that can address the gaps in carbon-based reporting units. For example, translating all sustainability measurements into ***monetary units*** can provide a complementary view of sustainability costs, in addition to carbon-based reporting. While such translations of carbon-to-dollars have been used in other fields (*e.g.*, carbon tax [68, 70] or regulatory credits [37, 58]), we argue that employing similar translations can facilitate sustainable (data center) computing as well.

Finally, we discuss unique opportunities that can be enabled by employing our proposed sustainability metrics. Specifically, we describe how access to per-job sustainability costs can allow end-users and software developers to be involved in data center sustainability efforts.

Authors' addresses: Anshul Gandhi; Dongyoon Lee; Zhenhua Liu; Shuai Mu; Erez Zadok, Stony Brook University, Stony Brook, New York, USA; Kanad Ghose; Kartik Gopalan; Yu David Liu, Binghamton University, USA; Syed Rafiul Hussain, Pennsylvania State University, USA; Patrick McDaniel, University of Wisconsin, Madison, USA.

2 DEFINING SUSTAINABILITY METRICS

Recent studies and efforts in sustainable computing have established the *need for new metrics*, beyond traditional power/energy metrics, to define and measure sustainability holistically [11, 26, 27, 60]. To facilitate sustainable computing efforts, such metrics should also be granular enough to identify sustainability bottlenecks, for example, at the level of a single request.

Existing solutions often focus on a single, possibly incomplete metric. For example, a popular metric used in data centers is the Power Usage Effectiveness (PUE) [7], defined as the ratio of total data center energy consumption to the energy consumed by IT equipment (servers, switches, etc.). However, PUE ignores the energy source (clean-energy vs. grid energy) and the wear-and-tear of IT equipment. Other metrics such as Carbon- [9] or Water-Usage Effectiveness [8] or Green PUE [5] are too coarse grained and cannot be mapped to individual requests, necessary for users to assess their impact on sustainability. The SCI rate metric [25] does provide carbon emissions per functional unit (*e.g.*, per user or per device), but it does not account for recycling potential or system overheads.

This section discusses the design of our proposed sustainability metrics and the various elements that contribute to our metrics. To combine different elements into a unified sustainability footprint, we start by employing the unit of "carbon dioxide equivalent" or **CO2e**, expressed in grams of CO_2 (or gCO2e) [44, 63, 65]. We discuss alternative units for reporting sustainability that may better incentivize general users in Section 3.

2.1 Design Philosophy

Our key objective in this position paper when designing our candidate metrics is to incorporate significant contributors to data center sustainability footprint—beyond just IT equipment energy—as discussed below. While we briefly comment on other objectives, such as metric accuracy and collection effort, we plan to investigate these in detail as part of future work.

(1) *Operational energy* is the energy consumed by all IT equipment in the data center (servers, routers, monitors, cooling systems, etc.), including the energy consumed by idle equipment and power delivery losses.
(2) *Energy source cleanliness* directly contributes to the sustainability footprint. For example, 1kWh of electricity generated from coal produces 820 gCO2e (grams of CO2-equivalent), whereas the carbon intensity for solar electricity is only about 48 gCO2e/kWh [39].
(3) *Embodied costs* refer to the sustainability costs incurred during the entire lifecycle of data center equipment, beyond the operating costs.
(4) *Device wear.* Different workloads can impact device health in different ways. For example, (write-heavy) I/O workloads wear out storage systems much faster than compute workloads, whereas the latter may wear out processors at a faster rate. As a result, the affected devices require additional maintenance or repair, thus incurring sustainability costs involved in manually attending to the devices and possibly replacing them.
(5) *Recycling costs.* Some equipment, such as hard disks or monitors, may be recycled instead of being disposed entirely. In such cases, there is a sustainability discount that applies when reusing the equipment.
(6) *Material consumption and human costs.* Data centers consume material for elements apart from IT equipment, such as the construction material required for data center building(s), water or glycol use for cooling and operating the data center, etc. These materials, and their disposal, contribute to the sustainability costs of data centers. Similarly, the personnel involved in data center operations and their actions (*e.g.*, driving to work) also contribute to sustainability costs.

2.2 Operational Sustainability Costs

It is critical to track the sustainability costs at a fine granularity (*e.g.*, a request or job) to drive meaningful sustainable computing efforts, such as resource management and scheduling/migration of workloads. This requires careful accounting of a job's power/energy usage at all hosts where the job executes. A job's end-to-end operational sustainability cost is then computed, for example, by summing up the job's share of all host-level and device-level (*e.g.*, switches, routers) power consumption and then converting them to CO2e values [65, 68, 70]. Note that the above computation captures the sustainability costs of distributed jobs (or applications) as it is a summation over all hosts and/or devices that the job executes on. Combining various, diverse energy costs may require more complex aggregation functions; the summation we employ is a simple first step. Note that operational sustainability costs are location- and time-dependent since the CO2e conversion values depend on the energy source mix in use at that time at the data center utility provider [55]. This trend can be exploited to schedule/migrate jobs at more sustainable sites and times of day [12, 40].

To fairly measure and charge a job for operational power use, one can leverage modern hardware and sensors exposed via the BIOS's ACPI interface. For non-shared environments, a job's power use can be obtained from direct power measurements via device-specific tools such as Intel's RAPL [21], NVIDIA's `nvidia-smi` utility [52] (for GPUs), ACPI [32], etc. The per-host power use can also be directly obtained via smart PDUs.

For shared environments, each job must be "charged" a fraction of the host's or device's power consumption, commensurate with the job's utilization of that host/device. Utilization details for each shared component (CPU, GPU, memory banks, etc.) can be obtained via monitoring available within modern hosts (*e.g.*, Intel's RDT and XTU [17, 18], NVIDIA's `nvprof` utility [53]), and complemented with readily available OS-level monitoring tools (*e.g.*, Linux's `/proc/PID/stat`). Note that we do *not* equally divide the power consumption of a host/device among the jobs executing on it, and instead charge jobs based on their utilization of the host/device.

To account for per-VM or per-container power use, a similar strategy can be employed by obtaining per-VM or per-container resource usage via the hypervisor or the orchestrator (*e.g.*, Kubernetes). Host-level virtualization overheads, such as VM exits, IPIs, world switches, and cache contention, can also be measured using

existing profiling/instrumentation mechanisms (*e.g.*, kprobes [45], perf [4]) and attributed to specific users, where possible.

Finally, a job must be charged with part of the total *unaccounted* power use of the host proportionally to the job's utilization, similar to a sales tax. Such unaccounted power use includes idle power of individual hardware components, power consumption of components that cannot be directly measured, and power consumed by the OS for maintenance functions such as garbage collection. Some data center components are closed "black boxes" that cannot be easily instrumented to measure power use, such as commercial storage servers/SANs or network switches. Power use of such units can be obtained via SNMP or via the rack's smart Power Management Unit (PMU) sockets where various components are plugged in. Per-job charges can then be applied proportionally based on the job's OS-reported network and storage traffic. By allocating unaccounted power across jobs, one can reconstruct the costs at various granularities (*e.g.*, host or cluster) by combining the costs of all jobs (executed and/or executing at that host or cluster).

To enhance our measurement accuracy, we will periodically compare the sum of power estimates of all jobs at a host with the aggregate per-host power measurements reported by PDUs. The difference in values will guide our end-to-end sustainability cost aggregation functions.

2.3 Embodied Sustainability Costs

These are the sustainability costs associated with the production, manufacturing, and disposal (including waste treatment) of all data center equipment. Embodied costs are often indirect costs with respect to data center computing (sometimes referred to as *scope-3 emissions* [67]) but are known to cause significant environmental harm [15, 16, 30, 71]. By accounting for embodied costs during decision making, data center operators can more broadly impact sustainability by incentivizing the purchase of "sustainably sourced" IT equipment.

However, we note that such costs are typically proprietary (verified through multiple industrial collaborators) and not released publicly. For example, estimating the CO2e usage incurred in building a server will require CO2e usage details from all (supply chain) vendors involved in manufacturing, shipping, and assembling the components of that server. Nonetheless, estimates of *some* of these costs (*e.g.*, CO2e costs of manufacturing specific IT equipment [15] or specific server models [62]) can be obtained. Whether or not such sparse, publicly available data can be used to extrapolate estimates to *all* data center equipment remains an important open question for sustainable computing.

The embodied costs will be amortized over the equipment lifetime; for example, servers are often replaced after 3–5 years [28]. A job will then be assigned a share of the per-equipment embodied costs in proportion to its total (possibly shared) usage time of that equipment, divided by the usable lifetime of the equipment.

2.4 Other Sustainability Costs

To account for the impact of job execution on device health, we propose to levy CO2e "taxes" on the job in proportion to the job's contribution in reducing the usable lifetime of the equipment it runs on. For example, if a job produces heavy I/O activity on a hard-disk drive or triggers overclocking on a processor, the associated wear-and-tear impact on the device, estimated via existing reliability studies and models [20, 24, 29, 69], can be added to the job's sustainability cost after appropriate amortization. However, it is possible that a job's execution can trigger overclocking or garbage collection, thereby impacting the sustainability costs and performance of all colocated jobs. Such unintentional costs will require further analysis for fair charging, an effort we defer to future work. Note that device health is also impacted by ambient factors such as heat and humidity [42].

To account for equipment recycling, we increase the lifetime estimate of the equipment by its predicted additional usage time. This results in jobs using that equipment to be taxed at a lower rate. For example, if job J runs for 1 day on an equipment with usable lifetime of 500 days and embodied costs of 100 gCO2e, then J will be taxed $\frac{100}{500} = 0.2$ gCO2e. If the equipment can be recycled, resulting in an additional 200 days of predicted usage, then the tax on J reduces to $\frac{100}{700} \approx 0.14$ gCO2e. Note that recycling may incur additional (one-time) costs which will have to be charged back to jobs; *e.g.*, RAM modules can be easily extracted and reused but HDDs may require careful scrubbing before reuse [33, 34]. In some cases, discarding and purchasing a more sustainable piece of equipment may be the better choice.

Finally, to account for other consumption, we rely on existing studies to translate the consumption to CO2e units. For example, one gallon of (unheated) water usage represents about 0.08 gCO2e [50, 57]. Of course, the CO2e costs of water usage also depend on the source of water (*e.g.*, desalination plant, non-replenishing aquifer, or a local river system). Note that it is possible that not all consumption (*e.g.*, activities of work-from-home staff) can be easily translated to CO2e values.

2.5 Our Proposed Sustainability Metrics

We now present our holistic, fine-grained metrics that consider the costs discussed in previous subsections:

- **Job Sustainability Costs** (JSC) is the operational CO2e spent while running the job, focusing on the factors discussed in Section 2.2. A "job" can be as small as a single network packet or as large as a giant DNN model that takes days to train [31]. Consider a job that consumes 1kJ energy executing on a host and an additional 0.08kJ due to power losses and cooling. If the energy source is 80% coal and 20% solar, then, using the carbon-intensity values stated in Section 2.1 and converting kJ to kWh, we have JSC $= 1.08 \times (0.8 \times 820 + 0.2 \times 48) \div 3600 \approx 0.2$gCO2e.
- **Amortized Sustainability Costs** (ASC) is the sum of JSC and the job's share (or tax) of embodied and other costs. Consider a job J that ran for 5 hours and incurred 40 gCO2e during its execution on equipment with a lifetime of 3 years, and the embodied costs (including wear-and-tear impact and recycling potential) of the equipment it exclusively ran on was 10,000 gCO2e. Then, $ASC = 40 + 10,000 \times \frac{5}{3 \times 365 \times 24} = 41.9$, meaning this job was taxed 1.90 gCO2e. If the equipment

was shared with other jobs, then the taxes are spread over jobs in proportion to their equipment usage time.
- **Sustainability Cost Rate** (SCR) is the rate of sustainability costs incurred per unit time for a job. SCR can be obtained by averaging JSC or ASC over short time periods, even instantaneously, or over the entire job lifetime. If the aforementioned job with JSC of 0.2 gCO2e completed in 20s, then its SCR is $\frac{0.2}{20}$ = 0.01 gCO2e/s.

To also account for job "performance", our metrics can be augmented with Quality of Service (QoS), a performance measure that users wish to maximize, such as throughput, model accuracy, inverse of tail latency, or a generic utility function [41, 66]. With QoS and sustainability in mind, we propose our final metric:

- **Job Quality per Cost Rate** (JQCR) combines a job's QoS and JSC (or ASC) as $\frac{QoS}{Elapsed\ time \times JSC}$, where elapsed time is the job's running time or the monitoring duration of interest. JQCR is a value that we aim to maximize (higher QoS, lower sustainability costs, and shorter time). JQCR is a rate metric, hence the *Elapsed time* in the denominator. The intuition behind JQCR is that it represents the rate at which one unit of sustainability cost (JSC or ASC) improves QoS; maximizing JQCR thus maximizes the rate at which QoS increases for each unit of sustainability cost expensed. JQCR can cover any useful scope: a single job, VM, host, etc. Consider an e-commerce service comprising three VMs: web server, database server, and credit card charging server. If, over the course of one day, the three VMs processed 15,000 online purchases and the sum of all JSCs was 250 gCO2e, then the hourly JQCR of the e-commerce service "job" is $JQCR = \frac{15{,}000}{24 \times 250}$ = 2.5 purchases/hr/gCO2e. As another example, suppose a large ML job took 1 week to train a single model and obtained a classification accuracy of 90% and incurred a JSC of 70 gCO2e. Then, its JQCR for accuracy as the QoS is $\frac{90}{7 \times 70} \approx 0.18$ % accuracy/day/gCO2e. Note that JQCR can be computed for any job granularity and in any QoS and time units. In this example, if the training data consisted of 1 million images, then the image-level JQCR, in time units of hours and QoS of number of images processed, is $\frac{1{,}000{,}000}{7 \times 24 \times 70} \approx 85$ images/hr/gCO2e.

The JSC, ASC, and SCR metrics can be used to optimize sustainability costs for a given QoS target (or vice-versa) or for determining Pareto-optimal curves of sustainability costs vs. QoS, whereas JQCR can jointly maximize QoS and minimize sustainability costs. Our metrics are customizable: *e.g.*, a QoS-sensitive provider can use QoS^2 in the numerator of JQCR.

The above metrics only represent a first step. We will continue to improve the accuracy and coverage of our metrics, and to make our metric collection more scalable.

3 COMPLEMENTARY REPORTING UNITS

Carbon-based units have often been employed as the de-facto units for sustainability reporting by various entities, including the EPA [44, 63, 65]. However, as we discuss below, carbon-based units do have certain limitations, and this represents an opportunity for alternative units to complement carbon-based ones.

3.1 Limitations of Carbon-Based Units

The carbon-based units (*e.g.*, CO2e) used today for sustainability assessments are intended to capture the environmental impact of an activity. However, carbon-based values are not *easily* available for the manufacture, distribution, installation, end-of-life dismantling, and recycling of various data center equipment (*e.g.*, air handlers, cables, UPSs, batteries). Further, CO2e values do not exist for every equipment used by a data center, for example, the water used in heat exchangers (cooling towers or radiator units) and evaporative coolers. Similarly, chemicals and energy used for conditioning water prior to use also have an environmental impact that is not captured by carbon-based units.

Finally, and importantly, one of the biggest problems that carbon-based units have is that many users cannot relate to carbon, because it is literally invisible. Users begin to relate to carbon, and sustainability issues in general, when, sadly, the impact is much more direct *and* visible, especially health or financial—*e.g.*, a fire, flood, or hurricane destroying one's home. We thus propose that *an alternative unit of measurement of sustainability is necessary to bridge the gap left by carbon-based units.*

3.2 Monetary Units: A Complementary View

Monetary units, such as dollars, for better or worse, serve as a common denominator that can be understood quite easily by all computing users. Indeed, pocketbook issues have a greater impact on sustainability than altruism: *e.g.*, demand for energy-efficient cars skyrocketed in 2022 largely due to rising gasoline prices [23, 51]. Further, monetary units can certainly quantify, at least approximately, the (sustainability) cost of manufacturing, distribution, and recycling as well as the consumption of resources that are not amenable to carbon-based measures. Monetary values for the hard-to-quantify environmental impact of data centers, combined with already quantifiable CO2e values of data center operations, can provide a way to look at the sustainability implications of data centers through their larger lifecycle. We therefore posit that ***monetary-based sustainability metrics are vital and complementary to carbon-based metrics.***

The environmental impact of carbon and other GHGs (*e.g.*, methane) can be converted to dollars, even accounting for long-term impact. For example, insurance companies have studied risk analysis and can produce reasonably accurate home-insurance quotes for those who live even in high-risk areas (*e.g.*, near flood-prone shores or in "tornado alley"). Likewise, health experts are projecting healthcare costs associated with global warming. Converting carbon-based values to monetary ones is already happening. Government Cap-and-Trade Programs have existed for years [14] as well as trading platforms such as the Chicago Climate Exchange [59]. For example, carbon taxes are already used [68, 70] and companies like Tesla regularly sell carbon credits [37, 58].

However, monetary units are not perfect either, at least for three reasons. (1) Devising fair and accurate cost models requires deep understanding of both computing and economics. As such, economists'

involvement will be integral in devising models and market structures to determine the sustainability taxes and budgets. (2) Rich countries, corporations, and individuals can afford to change their habits perhaps with little impact to their standards of living; but poorer and third-world countries may not be able to change so easily. Thus, economists would have to devise cost models and associated "exchange rates" that will work world-wide. (3) When money is involved, so grow the incentives to cheat and steal. There would have to be better policing and checks-and-balances to ensure that no entity can misreport their sustainability monetary impact, or steal sustainability related taxes.

4 ENABLING OPPORTUNITIES

Sustainability is a market differentiator. Our holistic sustainability metrics, applicable at a per-job basis, allow for various opportunities to improve computing sustainability. For example, data center providers can immediately employ our metrics to identify expensive (sustainability-wise) jobs or services, which can then be migrated or scheduled at times and locations where cleaner energy is available. Importantly, our metrics can enable broader participation from key stakeholders of sustainable computing who have thus far been ignored—the users.

4.1 Opportunities for Cloud Users

Data centers have an implicit (economic and environmental) incentive to make their operations sustainable; looming governmental regulations will only strengthen this incentive [10, 13, 22, 49]. We assert that mechanisms are needed to include *users* in the data center sustainability management effort. To incentivize users to participate in data center sustainability efforts, the first step, as also noted by industry, is to provide users with visibility into the sustainability costs of their actions [19, 46].

Our metrics (JSC, SCR, etc.) can provide this information readily to users in real-time, *e.g.*, via `/proc/pid` or visually via a dashboard. The dashboard can notify users about high sustainability costs, especially if the provider has mechanisms in place to throttle expensive jobs. Such notifications serve as feedback to incentivize users to improve the sustainability of their jobs. Our metrics can also be employed by users to plug in various values (*e.g.*, cloud data center location and time-of-use) and conduct what-if analysis for their sustainability use.

If using monetary units for sustainability metrics, then additional user incentivization can be achieved in public clouds by possibly discounting user's resource usage costs in proportion to their sustainability efforts (similar to a "cashback" strategy). For private clouds, a similar strategy could be employed but with tokens (*e.g.*, Service Units [SUs] in NSF Chameleon [2, 36]) which are either periodically replenished or allocated based on user needs [1, 64]. These tokens could then be expended in proportion to the sustainability cost metrics of user jobs, thus incentivizing users to sustainably execute their jobs.

4.2 Opportunities for Software Developers

Current programming models are largely oblivious to the sustainability implications of their design decisions. Developers are not entirely at fault here as they rarely have visibility into the sustainability metrics that are only available at the data center level. With our proposed metrics, the programming abstractions can be designed to be aware of—and adaptive to—JSC, ASC, SCR, and JQCR. For example, the hyper-parameters of CNN training can be dynamically adjusted based on the goals of sustainability. New programming models can be designed to capture these recurring programming idioms that provide application-specific approaches in balancing the trade-off between maximizing sustainability and minimizing the loss in the Quality of Service (QoS).

Furthermore, the proposed sustainability metrics will allow software developers to design and optimize software not only for QoS but also for sustainability. For example, software can be designed to take advantage of heterogeneous computing (*e.g.*, CPU, GPU, FPGA) and memory (*e.g.*, DRAM, NVM, HBM) resources with different sustainability implications. Software can also be designed to dynamically adapt to different execution environments (*e.g.*, the availability of renewable energy, the age and health of IT equipment) to trade off QoS and sustainability. For instance, fault-tolerant distributed software may adjust the checkpointing frequency, replication factor, or the backup storage medium to balance sustainability cost, QoS, and consistency guarantees.

5 CONCLUSION

To be truly end-to-end, metrics used for tracking and improving data center sustainability must look beyond traditional energy/power-based ones. These metrics must be comprehensive and encompass sustainability factors that are difficult, if not impossible, to track using energy/power-based formulations alone.

This position paper presented candidate metrics that account for some of these sustainability factors. A particular difficulty in quantifying the missing factors, such as heat loss and noise pollution, is the lack of any standards or conventions for the necessary instrumentation and measurement methodologies. While some efforts do exist in this space, such as the Greenhouse Gas Protocol (GHGP) [3] and the OCP Sustainability Initiative [54], more work is needed to standardize sustainability quantification. The lack of publicly available sustainability use data, especially embodied costs (see Section 2.3), is also an impediment. Finally, while we are dealing with sustainable computing, there is need to involve, at the very least, economists and public policy experts to make meaningful and lasting change. Mechanisms, such as those suggested in Sections 3 and 4, will be needed to implicitly incentivize computing users to participate in data center sustainability efforts before it is too late.

ACKNOWLEDGMENT

This work was supported in part by NSF grants 2214980, 1730128, 1750109, 2106434, 1918225, 1729939, 1900706, 2106263, 1910532, 2153747, and 1738793.

REFERENCES

[1] 2020. TC-TBF - Linux man page. https://linux.die.net/man/8/tc-tbf.
[2] 2023. Chameleon Cloud - A configurable experimental environment for large-scale cloud research. https://www.chameleoncloud.org.
[3] 2023. Greenhouse Gas Protocal. https://ghgprotocol.org.
[4] 2023. perf: Linux profiling with performance counters. https://perf.wiki.kernel.org/.

[5] Fawaz AL-Hazemi, Alaelddin Fuad Yousif Mohammed, Lemi Isaac Yoseke Laku, and Rayan Alanazi. 2019. PUE or GPUE: A Carbon-Aware Metric for Data Centers. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*. PyeongChang, South Korea, 38–41.

[6] Esmail Asyabi, Azer Bestavros, Erfan Sharafzadeh, and Timothy Zhu. 2020. Peafowl: In-application CPU Scheduling to Reduce Power Consumption of In-memory Key-Value Stores. In *Proceedings of the 11th ACM Symposium on Cloud Computing (SOCC '20)*.

[7] Victor Avelar, Dan Azevedo, and Alan French. 2012. PUE: A comprehensive examination of the metric. *White Paper WP-49, The Green Grid* (2012).

[8] Dan Azevedo, Christian Belady, and J Pouchet. 2011. Water usage effectiveness (WUE): A green grid datacenter sustainability metric. *White Paper WP-35, The Green Grid* (2011).

[9] Dan Azevedo, M Patterson, J Pouchet, and R Tipley. 2010. Carbon usage effectiveness (CUE): A green grid data center sustainability metric. *White Paper WP-32, The Green Grid* (2010).

[10] Mark Ballard. 2020. Data Center Operators Vie for Leverage as Europe Eyes Efficiency Rules. https://www.datacenterknowledge.com/regulation/data-center-operators-vie-leverage-europe-eyes-efficiency-rules.

[11] Noman Bashir, Tian Guo, Mohammad Hajiesmaili, David Irwin, Prashant Shenoy, Ramesh Sitaraman, Abel Souza, and Adam Wierman. 2021. Enabling Sustainable Clouds: The Case for Virtualizing the Energy System. In *Proceedings of the ACM Symposium on Cloud Computing (SoCC '21)*. Seattle, WA, USA, 350–358.

[12] Noman Bashir, Tian Guo, Mohammad Hajiesmaili, David Irwin, Prashant Shenoy, Ramesh Sitaraman, Abel Souza, and Adam Wierman. 2021. Enabling Sustainable Clouds: The Case for Virtualizing the Energy System. In *Proceedings of the 12th ACM Symposium on Cloud Computing (SoCC '21)*. Seattle, WA, USA.

[13] Rabih Bashroush. 2019. EU regulation on servers and data storage products adopted. https://www.opencompute.org/blog/eu-regulation-on-servers-and-data-storage-products-adopted.

[14] The California Air Resources Board. 2022. Cap-and-Trade Program. https://ww2.arb.ca.gov/our-work/programs/cap-and-trade-program.

[15] Gary Cook and Elizabeth Jardim. 2017. Guide to Greener Electronics. https://www.greenpeace.org/usa/wp-content/uploads/2017/10/Guide-to-Greener-Electronics-2017.pdf.

[16] Gary Cook and Elizabeth Jardim. 2019. Clicking Clean Virginia: The Dirty Energy Powering Data Center Alley. https://www.greenpeace.org/usa/wp-content/uploads/2019/02/Greenpeace-Click-Clean-Virginia-2019.pdf.

[17] Intel Corporation. 2019. Intel Resource Director Technology (Intel RDT) on 2nd Generation Intel Xeon Scalable Processors Reference Manual. https://software.intel.com/content/www/us/en/develop/articles/intel-resource-director-technology-rdt-reference-manual.html.

[18] Intel Corporation. 2020. Intel Extreme Tuning Utility (Intel XTU). https://downloadcenter.intel.com/product/66427/Intel-Extreme-Tuning-Utility-Intel-XTU-.

[19] IBM Corporation. 2020. Reducing the carbon footprint of computing. https://www.ibm.com/community/z-and-cloud/use-case/sustainability.

[20] Ayse K. Coskun, Richard Strong, Dean M. Tullsen, and Tajana Simunic Rosing. 2009. Evaluating the Impact of Job Scheduling and Power Management on Processor Lifetime for Chip Multiprocessors. In *Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '09)*. Seattle, WA, USA, 169—180.

[21] Howard David, Eugene Gorbatov, Ulf R. Hanebutte, Rahul Khanna, and Christian Le. 2010. RAPL: Memory Power Estimation and Capping. In *Proceedings of the 16th ACM/IEEE International Symposium on Low Power Electronics and Design* (Austin, Texas, USA) *(ISLPED '10)*. ACM, New York, NY, USA, 189–194. https://doi.org/10.1145/1840845.1840883

[22] C.D. Ditlev-Simonsen. 2021. *A Guide to Sustainable Corporate Responsibility: From Theory to Action*. Springer International Publishing.

[23] Stephen Edelstein. 2022. Gas prices spur demand for high-mpg vehicles, amid short supplies. https://www.greencarreports.com/news/1135287_gas-prices-high-mpg-vehicles-supplies-are-short.

[24] J. G. Elerath. 2004. Server Class Disk Drives: How Reliable Are They. In *IEEE Reliability and Maintainability Symposium*. 151–156.

[25] Green Software Foundation. 2022. Software Carbon Intensity (SCI) Specification. https://greensoftware.foundation/projects/software-carbon-intensity-sci-specification.

[26] National Science Foundation. 2021. NSF/VMware Partnership on the Next Generation of Sustainable Digital Infrastructure (NGSDI). https://www.nsf.gov/pubs/2020/nsf20594/nsf20594.htm.

[27] National Science Foundation. 2022. Dear Colleague Letter: Design for Sustainability in Computing. https://www.nsf.gov/pubs/2022/nsf22060/nsf22060.jsp.

[28] Gartner, Inc. 2013. *Desktop Total Cost of Ownership: 2013 Update*. Technical Report. Gartner Group/Dataquest. www.gartner.com.

[29] K. M. Greenan, J. S. Plank, and J. J. Wylie. 2010. Mean Time to Meaningless: MTTDL, Markov Models, and Storage System Reliability. In *HotStorage '10: Proceedings of the 2nd USENIX Workshop on Hot Topics in Storage*.

[30] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien Hsin S. Lee, Gu Yeon Wei, David Brooks, and Carole Jean Wu. 2021. Chasing Carbon: The Elusive Environmental Footprint of Computing. In *Proceedings of the 27th IEEE International Symposium on High Performance Computer Architecture (HPCA '21)*. IEEE Computer Society, 854–867.

[31] Ubaid Ullah Hafeez, Xiao Sun, Anshul Gandhi, and Zhenhua Liu. 2021. Towards Optimal Placement and Scheduling of DNN Operations with Pesto. In *Proceedings of the 22nd International Middleware Conference (Middleware '21)*. Virtual Event, 39–51.

[32] Intel Corporation. 2016. Advanced Configuration and Power Interface (ACPI) Introduction and Overview. https://acpica.org/sites/acpica/files/ACPI-Introduction.pdf.

[33] N. Joukov, H. Papaxenopoulos, and E. Zadok. 2006. Secure Deletion Myths, Issues, and Solutions. In *Proceedings of the Second ACM Workshop on Storage Security and Survivability (StorageSS 2006)*. ACM, Alexandria, VA, 61–66.

[34] N. Joukov and E. Zadok. 2005. Adding Secure Deletion to Your Favorite File System. In *Proceedings of the third international IEEE Security In Storage Workshop (SISW 2005)*. IEEE Computer Society, San Francisco, CA, 63–70.

[35] Kostis Kaffes, Dragos Sbirlea, Yiyan Lin, David Lo, and Christos Kozyrakis. 2020. Leveraging Application Classes to Save Power in Highly-Utilized Data Centers. In *Proceedings of the 11th ACM Symposium on Cloud Computing (SOCC '20)*.

[36] Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons Learned from the Chameleon Testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

[37] Arjun Kharpal. 2021. What 'regulatory credits' are — and why they're so important to Tesla. https://www.cnbc.com/2021/05/18/tesla-electric-vehicle-regulatory-credits-explained.html.

[38] Keith Kirkpatrick. 2023. The Carbon Footprint of Artificial Intelligence. In *Communications of the ACM, Vol. 66, No. 8, August 2023*. ACM, 17–19.

[39] V Krey, Masera O, Blanford G, Bruckner T, Cooke R, Fisher-Vanden K, Haberl H, Hertwich E, Kriegler E, Mueller D, Paltsev S, Price L, Schloemer S, Uerge-Vorsatz D, Van Vuuren D, Zwickel T, Blok K, De La Rue Du Can S, Janssens-Maenhout G, Van Der Mensbrugghe D, Radebach A, and Steckel J. 2014. *Annex II: Metrics & Methodology*. Cambridge University Press, Cambridge and New York (UK and USA). 1281–1328 pages.

[40] Russell Lee, Jessica Maghakian, Mohammad Hajiesmaili, Jian Li, Ramesh Sitaraman, and Zhenhua Liu. 2021. Online Peak-Aware Energy Scheduling with Untrusted Advice. *ACM E-energy* (2021).

[41] Z. Li, A. Mukker, and E. Zadok. 2014. On the Importance of Evaluating Storage Systems' $Costs. In *Proceedings of the 6th USENIX Conference on Hot Topics in Storage and File Systems* (Philadelphia, PA) *(HotStorage'14)*.

[42] Ioannis Manousakis, Sriram Sankar, Gregg McKnight, Thu D. Nguyen, and Ricardo Bianchini. 2016. Environmental Conditions and Disk Reliability in Free-cooled Datacenters. In *14th USENIX Conference on File and Storage Technologies (FAST 16)*. Santa Clara, CA, 53–65.

[43] Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. 2020. Recalibrating global data center energy-use estimates. *Science* 367, 6481 (2020), 984–986.

[44] V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, M.I. Gomis X. Zhou, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield. 2018. IPCC, 2018: Annex I: Glossary. *Global Warming of 1.5 C. An IPCC Special Report on the impacts of global warming of 1.5 C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* (2018).

[45] A. Mavinakayanahalli, P. Panchamukhi, J. Keniston, A. Keshavamurthy, and M. Hiramatsu. 2006. Probing the Guts of Kprobes. In *Proceedings of the 2006 Linux Symposium*, Vol. 2. Ottawa, Canada, 109–124.

[46] Microsoft Azure. 2023. Emissions Impact Dashboard. https://www.microsoft.com/en-us/sustainability/emissions-impact-dashboard.

[47] Rich Miller. 2021. The Bitcoin Energy Debate: Lessons from the Data Center Industry. https://datacenterfrontier.com/the-bitcoin-energy-debate-lessons-from-the-data-center-industry.

[48] F. F. Moghaddam, M. Cheriet, and K. K. Nguyen. 2011. Low Carbon Virtual Private Clouds. In *Proceedings of the 2011 IEEE International Conference on Cloud Computing*. Washington, D.C., USA, 259–266.

[49] Sebastian Moss. 2020. The EU wants data centers to be carbon neutral by 2030. https://www.datacenterdynamics.com/en/news/eu-wants-data-centers-be-carbon-neutral-2030.

[50] River Network. 2010. Water Energy Toolkit: Understanding the Carbon Footprint of Your Water Use. https://www.rivernetwork.org/wp-content/uploads/2015/10/Toolkit_Emissions2-8-12.pdf.

[51] npr.org. 2022. Gas prices got you wanting an electric or hybrid car? Well, good luck finding one. https://www.npr.org/2022/03/25/1088287767/fuel-efficient-cars-hybrid-electric-gas-prices-surge.

[52] NVIDIA Corporation. 2023. NVIDIA System Management Interface. https://developer.nvidia.com/nvidia-system-management-interface.

[53] NVIDIA Corporation. 2023. NVIDIA Visual Profiler. https://developer.nvidia.com/nvidia-visual-profiler.

[54] Open Compute Project. 2022. OCP Sustainability Initiative. https://www.opencompute.org/wiki/OCP_Sustainability_Initiative.

[55] Katayoun Rahbar, Jie Xu, and Rui Zhang. 2014. Real-time energy storage management for renewable integration in microgrid: An off-line optimization approach. *IEEE Transactions on Smart Grid* 6, 1 (2014), 124–134.

[56] Chuangang Ren, Di Wang, Bhuvan Urgaonkar, and Anand Sivasubramaniam. 2012. Carbon-Aware Energy Capacity Planning for Datacenters. In *Proceedings of the 20th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS '12)*. Arlington, VA, USA, 391–400.

[57] Christine Ro. 2020. The hidden impact of your daily water use. https://www.bbc.com/future/article/20200326-the-hidden-impact-of-your-daily-water-use.

[58] Gustavo Henrique Ruffo. 2020. Tesla Earned $428 Million With Carbon Credits In Q2 2020: Why That's Bad. https://insideevs.com/news/438345/tesla-428-million-carbon-credits-q2-2020.

[59] Richard L. Sandor. 2003. Chicago Climate Exchange. https://en.wikipedia.org/wiki/Chicago_Climate_Exchange.

[60] P. Shenoy and T. Wenisch. 2015. NSF Workshop on Sustainable Data Centers: Final Report.

[61] Sam Steers. 2021. How crypto mining affects data centre sustainability. https://datacentremagazine.com/data-centres/how-crypto-mining-affects-data-centre-sustainability.

[62] M. Stutz, S. O'Connell, and J. Pflueger. 2012. Carbon footprint of a dell rack server. *2012 Electronics Goes Green 2012+* (2012), 1–5.

[63] The White House. 2021. Federal Sustainability Plan: Catalyzing America's Clean Energy Industries and Jobs. https://www.sustainability.gov/pdfs/federal-sustainability-plan.pdf.

[64] J. Turner. 1986. New directions in communications (or which way to the information age?). *IEEE Communications Magazine* 24, 10 (1986), 8–15.

[65] U.S. Environmental Protection Agency. 2016. Definition | CO2e. https://www3.epa.gov/carbon-footprint-calculator/tool/definitions/co2e.html.

[66] Muhammad Wajahat, Bharath Balasubramanian, Anshul Gandhi, Gueyoung Jung, and Shankar Narayanan. 2020. MERIT: Model-driven Rehoming for VNF Chains. In *Proceedings of the 1st IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS '20)*. Washington, DC, USA.

[67] Noelle Walsh. 2021. Supporting our customers on the path to net zero: The Microsoft cloud and decarbonization. https://blogs.microsoft.com/blog/2021/10/27/supporting-our-customers-on-the-path-to-net-zero-the-microsoft-cloud-and-decarbonization/.

[68] Akio Yamazaki. 2017. Jobs and climate policy: Evidence from British Columbia's revenue-neutral carbon tax. *Journal of Environmental Economics and Management* 83 (2017), 197 – 216.

[69] J Yang and F. Sun. 1999. A Comprehensive Review of Hard-Disk Drive Reliability. In *Annual Reliability and Maintainability Symposium*.

[70] Kun Zhang, Qian Wang, Qiao-Mei Liang, and Hao Chen. 2016. A bibliometric analysis of research on carbon tax from 1989 to 2014. *Renewable and Sustainable Energy Reviews* 58 (2016), 297 – 310.

[71] Ferdinand Zotz and Maximilian Kling. 2017. Support to selected Member States in improving hazardous waste management based on assessment of Member States' performance. https://ec.europa.eu/environment/waste/studies/pdf/20180227_Haz_Waste_Final_RepV5_clear.pdf.