Towards Carbon Footprint Management in Hybrid Multicloud

Rohan Arora[1], Umamaheswari Devi[2], Tamar Eilam[1], Aanchal Goyal[2], Chandra Narayanaswami[1], Pritish Parida[1]
[1] IBM Research, Yorktown Heights, USA
[2] IBM Research, India

ABSTRACT

Enterprises today aspire to optimize the operating costs and carbon footprint (CFP) of their IT operations jointly without compromising their business imperatives. This has given rise to a hybrid approach in which enterprises retain the dynamic choice to leverage private data centers and one or more public clouds in conjunction. While cloud service providers (CSPs) have long provided APIs for estimating, reconciling, and optimizing operating costs, they have only recently started exposing APIs related to CFP.

Indeed, this is a step in the right direction. Nevertheless, our analyses of these APIs reveals many gaps that need to be addressed to facilitate sizing and placement decisions that can factor in carbon. First, there is a lack of standardized, transparent methodology for CFP quantification across different CSPs. Second, the coarse granularity of the CFP data provided today can help with post-facto reporting but is not suitable for proactive fine-grained optimization. Last, enterprises themselves are unable to independently compute the current CFP or estimate potential CFP savings since CSPs do not share the required power usage data.

To address these gaps, enterprises have started developing their own carbon assessment methodologies and tools to estimate the CFP of workloads running on public clouds using the available user-facing APIs. These systems hold the promise for *an independent and unbiased evaluation* and *estimation of relative savings* between different deployment options by cloud users. We describe and analyze the details of CSP-native carbon-reporting tools and their quantification methodology, and the "outside-of-the-cloud" estimation approaches. Finally, we present opportunities for future research in the direction of trustworthy, fine-grained, public cloud workload CFP estimation, which is a prerequisite for meaningful realization of carbon optimization.

CCS CONCEPTS

• **Hardware** → **Power estimation and optimization**; • **General and reference** → *Metrics*; *Performance*; *Evaluation*; • **Applied computing** → *Multi-criterion optimization and decision-making*.

KEYWORDS

Sustainable Computing, GHG emissions, Carbon Footprint, Cloud, Data Centers, GHG Accounting, Carbon-Aware Optimization

HotCarbon '23, July 9, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0242-6/23/07.
https://doi.org/10.1145/3604930.3605721

1 INTRODUCTION

Multiple inflection points are changing the way enterprise CIOs are thinking about software that runs in their enterprise. The rising energy footprint of data centers, which already accounts for 200 TWh per year (around 1% of total global electricity demand) [2], is something they need to contend with. This is driven by the proliferation of AI [52], including the recent explosive spurt in generative AI [42, 55], profusion of data created by sensors at the edge and via customer interactions [43, 44], cloud-based gaming, cryptocurrency mania [32], etc., coupled with the flattening of Moore's law (see e.g., [14], [5]), over-provisioning of compute resources [23], and the focus on speed of product delivery over efficiency during operations. The concomitant increase in enterprise carbon footprint (CFP)[1] has brought a sense of urgency within enterprises to first measure and then reduce their carbon emissions associated with the direct or indirect operations of their businesses [51]. This is a notable departure from the erstwhile sole focus on cost and performance and is in keeping with the sustainability grand challenge in computing [8]. A second trend is the evolution of the IT footprint for modern enterprises. Modern enterprise customers typically have their workloads deployed to both private data centers and more than one public cloud [11]. This is driven by a variety of reasons, including resilience and responsiveness, concentration risk, data sovereignty, and the need to co-locate data and compute for best performance, private data center capacity, availability of resources on specific clouds in various geographies, cost for cloud services, etc. [31]. The combination of these two important and growing trends is leading enterprises to ask the following questions.

(1) How can the energy and CFP of workloads running in private data centers and public clouds be quantified, optimized, and reported using a consistent approach?
(2) How can intelligent decisions be made regarding workload migration from private data center to cloud or from one CSP to another based on their CFP, availability of renewable energy, and other key factors while achieving SLOs?
(3) What are the estimated energy and CFP savings if they modernize the workload and/or the underlying IT hardware leveraging GPUs, TPUs, AIUs [22]?

To draw an analogy, cloud service providers (CSPs) and third-party vendors have long provided application programming interfaces (APIs) that allow their customers to systematically determine the services that are available in various geographies and their associated costs. Application performance management (APM) tools like Dynatrace [12, 48], Instana [28], and Prometheus [54] have been

[1] *Carbon footprint* refers to the amount of greenhouse gases (GHG) emitted due an activity. It is expressed in units of mass of carbon-dioxide equivalent gases (CO_2-eq) and indicates the sum total of all GHG emissions and not just CO_2.

Feature	Google Cloud Platform (GCP)	Microsoft Azure	Amazon Web Services(AWS)
Service coverage	Services list [19]	Azure and Microsoft 365 [35]	S3, EC2, and rest as one category
Geographic coverage	Worldwide	Worldwide	Worldwide
Aggregation level	Region, service, and subscription level	Region, service, and subscription level	Limited aggregation to geographies like AMER and EMEA and services for all accounts
Time granularity	Monthly	Monthly	Monthly
Emissions scope	Scope1+Scope2+Scope3, location based [24]	Scope1+Scope2+Scope3, location based	Scope1+Scope2, market based [24]
Power usage effectiveness (PUE) [21]	Quarterly and trailing twelve-month (TTM) PUE [18] for data center	Design and operational PUE for data center [38, 40]	Precise PUE not shared [3]
Carbon intensity (CI) [20]	Electricity maps [33] or IEA [1] in case of limited data	Limited [40]	Limited [3]
History	Starting Feb 15, 2021	Azure: last 5 years of enrollment; Microsoft 365: last 12 months usage	Starting Jan 1, 2020

Table 1: Comparison of client CFP services provided by the major CSPs and some data underpinning (PUE and CI) the services.

developed to collect service-level metrics and application topologies. These can be coupled with tools such as Grafana [6] and Kibana [49] to create dashboards. This suite of tools allows one to continuously monitor performance and also estimate the costs and performance of running a workload in a different hardware configuration/location.

On the other hand, CSPs have just begun to respond to the questions enterprises are posing around carbon metrics and estimation. The efforts are in preliminary stages and present several challenges around the granularity at which data is available, completeness of data, and consistency of different carbon estimates that need to be addressed. Given the complexity of the problems, akin to the APM domain, tools, services, and open-source projects to assess the energy and CFP from outside of the cloud are beginning to emerge to help answer the questions previously mentioned.

In this paper, we first compare CSP APIs and reports for carbon metrics and review emerging methods for calculating the energy and CFP. We then identify the gaps and discuss research opportunities for carbon performance measurement (CPM) tools. We suggest how current APM tools can be extended and explore new abstractions and tools to cover CPM. Our investigation leads us to conclude that (1) hardware manufacturers need to provide more granular and modular non-invasive power measurement hooks, and (2) CSPs and their users (enterprise customers) are jointly responsible for working together toward greater transparency that will allow for meaningful energy and CFP calculations for workloads.

2 CSP CARBON EMISSION REPORTS

In this section, we survey and compare the features of and methodology for calculating CFP data provided by the three major CSPs to the clients for use of their services, from the perspective of their suitability for CFP reporting and optimization. The three major CSPs—Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP)—have been providing dashboards and APIs for quantifying, analyzing, and optimizing operational costs for quite some time; however, CFP reports for a select set of services have been a fairly recent addition. All three CSPs follow the Greenhouse Gas Protocol (GHG) carbon reporting and accounting standards [25] to generate client CFP reports.

The GHG protocol defines three *scopes* (scope 1, scope 2, and scope 3) for GHG accounting and reporting purposes to help delineate direct and indirect emission sources and improve transparency [26]. Scope 1 covers a company's direct GHG emissions (e.g., emissions from owned or controlled vehicles, boilers, furnaces). Scope 2 accounts for GHG emissions from the generation of purchased electricity consumed by the company, while scope 3 is attributed to indirect emissions on account of the activities of the company that occur from sources not owned or controlled by the company. The three CSPs differ in the extent and the manner in which they report emissions under the three scopes. In this paper, we focus primarily on scope 2 emissions, the emissions due to the consumption of electricity in data centers.

Table 1 captures some of the salient features and similarities / differences across carbon emissions provided by leading CSPs and transparency in the underpinning data and methodology. We next briefly present and compare their CFP calculation methodologies.

GCP. Google's methodology is the best explained of the three CSPs [17]. For estimating CFP in GCP and proper apportioning of total machine energy usage to its internal services, dynamic power and idle power are separately identified. Hourly dynamic power is allocated based on relative internal service CPU usage, whereas the machine idle power is assigned based on each internal service's resource allocation (CPU, RAM, SDD, HDD). Also, the grid carbon emission intensity data is tracked hourly and multiplied with the hourly energy usage of each internal service to derive the internal service's location-based electricity CFP per hour and location. In short, following the GHG Protocol [25], compute and data center resource utilization is used as a weighting factor for apportioning CFP. It is augmented by accounting for location-based carbon emissions from electricity use and proportional allocations of emissions due to non-electrical sources.

Azure. Microsoft has released an Emissions Impact Dashboard consistent with GHG Protocol [25]. The methodology [37, 39] for scope 3 emissions calculates the energy and carbon impacts for each data center over time, using the information about most common cloud infrastructure parts (hard disks, FPGA, steel racks) and the manufacturing materials most commonly used in their data centers. Power usage for scopes 1 and 2 Azure emissions is categorized by

storage, compute, or network, and usage time is utilized for apportioning emissions. Unlike GCP, the official methodology document does not include much detail on how input data is collected or the time granularity at which input data is processed for computing CFP. According to their methodology as described in [34], power consumption and resource usage metrics are mostly estimated and not measured.

AWS. Amazon has recently released a sustainability dashboard. Unlike GCP and Azure, however, it does not have any attached API exposed yet. Also, across all three CSPs, one needs a billing account to be able to access emissions data. The carbon emissions model consists of five major areas: embodied emissions of both data center facilities and IT hardware, carbon intensity of the grid, and facilities and IT operational emissions [47]. Similar to Azure, AWS does not disclose details on data collection methodology or time granularity at which input data is processed for computing CFP.

The gaps and differences highlighted in Table 1 and evident from the brief descriptions above limit the customer in getting a holistic view of their enterprise CFP. Moreover, the lack of CFP numbers at the entity level (e.g.,host, VM, containers) is a bigger drawback in projecting and creating optimal resource-allocation strategies for tasks like migrating to cloud. The upcoming area of multicloud sustainability offers to customers the advantages of getting the best of all CSPs rather than being bound to one specific provider. Unless the carbon estimates provided by all third-party clouds are well understood, optimization strategies are difficult to create and execute in a reliable and robust manner across clouds and geographies. Though the CSP's CFP reports include emissions from scopes 1 and 3, in this paper, we restrict our focus to Scope 2 emissions. Also, we do not explicitly address data center overhead such as cooling, but only via the fidelity of PUE data [21].

3 THIRD-PARTY CFP ESTIMATION EFFORTS

Cloud users have been seeking to understand the environmental impact due to their share of cloud usage for quite some time now. Challenges in estimating this by tenants (users) themselves are described in detail in [41]. At the time that [41] was authored, CSP-based carbon accounting was available only for Azure [36]. Though the three leading CSPs and some other prominent players currently support carbon accounting for their services natively, as discussed in Sec. 2, they differ in many aspects and currently do not fully meet client needs. Hence, there has been quite some effort for estimating CFP on public clouds, either by cloud tenants themselves or on their behalf by third-party service providers and tools developed by open-source efforts. We will refer to this collective effort as third-party approaches (TPAs), the most prominent of which is by Cloud Carbon Footprint [16], an open-source *cloud carbon emissions measurement and analysis tool*, with climatiq [9], teads [15], and Nordcloud Klarity's GreenOps [29], among others.

One question that rightly arises is how reliable the TPA estimates could be given the lack of visibility to them on the internals of a public cloud deployment. To better answer this, we first present brief overviews of the services deployment architecture in cloud and the metering and bookkeeping required to enable fair and accurate attribution of CFP per workload. (Recall from Sec. 2 that not all CSPs are transparent about their bookkeeping.) We will then go over the assumptions and methodology used by the TPAs to identify the gaps and discuss the possibilities for bridging the same.

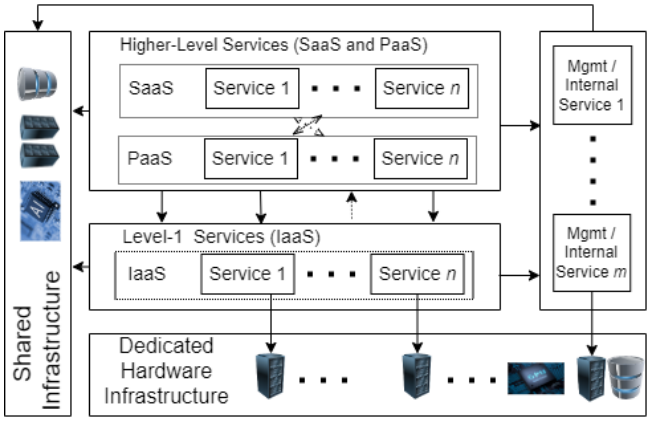

Figure 1: High-level deployment architecture of services in a cloud.

3.1 Cloud Deployment and Service Metering

Cloud deployment architecture. A high-level services deployment architecture for public clouds is shown in Fig. 1. Cloud services are heterogeneous, broadly classified into *infrastructure*, *platform*, and *software as a service* (i.e., IaaS, PaaS, and SaaS) classes. IaaS class is the closest to hardware and simplest in terms of energy attribution due to limited dependency on other services. Google's Compute Engine (GCE), Kubernetes Engine (GKE), and Amazon's Simple Storage Service (S3) are examples of IaaS services. PaaS and SaaS services are layered on top of IaaS services, with some dependency from SaaS to PaaS too. Google AppEngine and Google Docs fall under the PaaS and SaaS categories, respectively. Apart from external services, there can be some internal services, management, and control-plane services at any of the layers that are used by the external services or manage them (e.g., load balancers and control plane of container services and serverless functions).

As depicted in Fig. 1, each cloud service can be deployed on dedicated and/or shared compute, storage, and networking hardware. Lower-level services are more likely to have a direct hardware footprint, whereas the higher-level services will have an indirect footprint via their use of lower-level services. Distributing the energy consumed at the hardware level among the users of the diverse services in proportion to their actual usage at the hardware through the entire software stack in a fair and accurate manner will require extensive and rigorous metering and bookkeeping. A possible high-level attribution approach is depicted in Fig. 2.

Service metering in Clouds for CFP. In Fig. 2, the data in input boxes (1) through (3) is obtained by live metering of equipment- or rack-level electricity consumption and per-process hardware resource usage tracked by the operating system (OS). In the case of higher-level services, additional service-specific metering (3) will be required by the provider services. For example, a DBaaS service (as a provider service) can perhaps track the resource usage of each of its users (through system calls, loadable kernel modules, eBPF [13], etc.) so that its electricity consumption can be rightly apportioned. On the other hand, tracking direct use of resources at the user-level can be difficult for SaaS applications like streaming

services or Google Docs. Instead, these services generally only track service-specific usage units, commonly termed *functional units*, such as the number and size of videos streamed, and use those for energy attribution purposes in combination with aggregate functional units, resources, and electricity consumed by the service.

As a specific example, consider per-account energy computation for the Virtual Machine (VM) service. If the VMs that comprise the service are isolated from the other services and are deployed on dedicated hardware, then the total energy consumption on that hardware will constitute the aggregate energy for the VM service, denoted $E(\mathrm{S_{VM}})$. In such a case, in Fig. 2, $E(\mathrm{S_{VM}})$ will be provided by (2). (If the VM service uses shared hardware exclusively or uses dedicated hardware in addition, then $E(\mathrm{S_{VM}})$ will be derived from (1) exclusively or both (1) and (2).) Let $E(\mathrm{S_{VM}}) = E_{\mathrm{idle}}(\mathrm{S_{VM}}) + E_{\mathrm{active}}(\mathrm{S_{VM}})$, where E_{idle} and E_{active} denote the idle and active parts of the total energy.[2] Let $VM_1 \cdots VM_n$ denote the n VMs that are part of the service, $vCPU_1 \cdots vCPU_n$ and $M_1 \cdots M_n$ the number of vCPUs and amount of memory provisioned for the VMs, respectively. Let $u_1 \cdots u_n$ denote the CPU utilization of the VMs. The GHG protocol suggests that the idle energy of a service be apportioned among its users by provisioned units, while active energy will be by actual usage. Thus, energy attribution for VM_i can be given by:

$$E(VM_i) = E_{\mathrm{idle}}(\mathrm{S_{VM}}) \times \left(\frac{w_c \cdot vCPU_i}{\sum_k vCPU_k} + \frac{w_m \cdot M_i}{\sum_k M_k} \right) + E_{\mathrm{active}}(\mathrm{S_{VM}}) \times \frac{u_i \cdot vCPU_i}{\sum_k u_k \cdot vCPU_k}, \quad (1)$$

where w_c and w_m denote the fractions of idle energy due to compute and memory, respectively. If the service has overheads due to common networking equipment or control plane nodes, the contribution from those should be correctly factored in.

The GHG protocol or other standards do not mandate the temporal granularity (per minute vs. month) or hardware granularity (per machine vs. rack) at which the resource consumption is metered. In fact, rationalized and well-documented estimates can be substitutes for metering. Similarly, guidelines are quite open in the functional units used by higher-level services. Though the standards mandate appropriately attributing resource consumption at the shared infrastructure and services, the exact methodology is left open. Thus, the approaches adopted by different cloud providers can have significant differences, leading to difficulty in comparing and managing CFP. Further, comprehensive carbon reporting spanning all the catalogued services is currently not known to be supported by any CSP.

To see how the differences in the aspects discussed here can impact the reported numbers and the actions they can guide, note that in Eq. (1), idle energy is split by CPU and memory provisioned. CSPs have the choice of including other resources such as storage and networking too, or even excluding memory. Also note that run-time memory usage is not factored into (1). Thus, depending on which resources are considered, energy and carbon attributions can differ. Attributions will also differ based on what overheads are considered, whether they are metered or estimated, how they are metered or estimated, and how they are apportioned.

[2] *Idle energy* is the energy consumed by or attributed to a component (hardware or software component) when it is in idle state, while *active energy* is the additional energy consumed or attributed to the component when it is used to perform work.

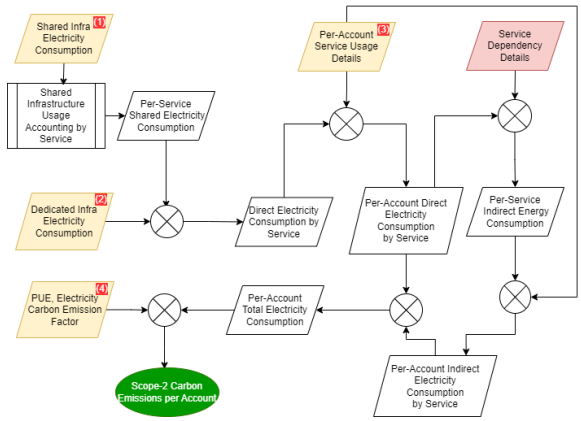

Figure 2: Per-account carbon quantification framework.

As a second example, consider two users using equal amount of resources for equal duration in a day but at different complementary time periods: one when the grid is powered by renewables, and the other when fossil fuels are used. If usages or CI or both are averaged over a day, then both the users will see identical CFP and cannot effectively use the metric. The magnitude of error could be glaring at the current CSP reporting granularity of one month. Differences in other aspects can also be similarly misleading or wanting. Appropriate levels of granularity and bookkeeping required can depend on the service, whose determination can be a challenge by itself.

Thus, to enable comparison of the carbon metrics reported by various CSPs and to use them for carbon-aware placement and optimization, it is essential that various aspects/features of the methodology be standardized. These include the following:

- Set of resource utilization metrics, service functional units, and overheads used in apportioning electricity consumption
- Granularity at which metrics are collected
- Averaging interval for the collected metrics
- Electricity consumption apportioning formulas at various levels, including overheads
- How CEF is determined (location-based vs. market-based) and how carbon offsets and renewable energy certificates are handled
- Granularity at which PUE and CEF are applied
- Details of emissions calculations for scopes 1 and 3

Until accounting becomes standardized, CSPs should be more transparent and share the details on the preceding aspects of their specific implementation.

3.2 Client and Third-Party Approaches

TPAs make up for the lack of visibility into the details of a cloud deployment and lack of information due to limited direct measurements by models, estimations, approximations, and assumptions. TPA approaches can be broadly classified into *online* or *dynamic* approaches and *offline* or *static* approaches.

Online approaches. Online or dynamic approaches leverage resource consumption metrics that can be captured directly from the hardware at runtime by the users of the service. These are best suited to IaaS services (e.g., those providing dedicated bare-metal servers, VMs, or container clusters). For these services, clients

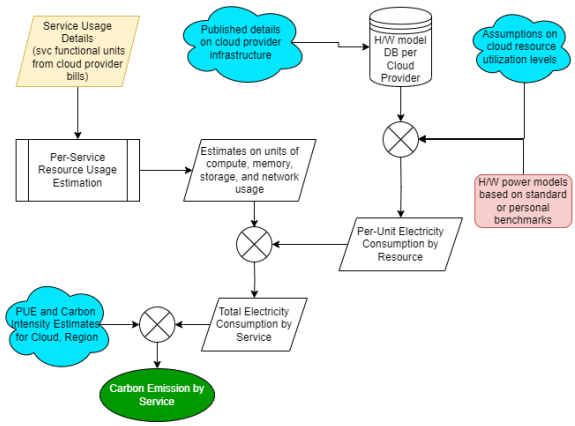

Figure 3: Offline third-party carbon estimation framework.

may be able to install custom agents that capture the resource and power metrics pertaining to their usage from the OS via system calls or utilities that rely on special device drivers. The RedHat project Kubernetes Efficient Power Level ExporteR (Kepler) is one such effort [46], which uses eBPF [13] to gather system-level fine-grained metrics to compute pod-level energy. It is unclear whether superuser privileges required for collecting detailed metrics would be readily available in all cases. Further, determining overhead usage incurred on behalf of the service on nodes (that host management/control/internal services) that the agent does not have access to will be difficult to gather.

Offline approaches. An overview of the more common offline (static) approach used by many TPAs is shown in Fig. 3. TPAs rely on service consumption costs and usage reports provided by the CSPs as the starting point, from which coarse estimates on the usage of different resources—namely CPU, memory, storage, and network—are derived. The approach also relies on publicly available sources for assumptions on hardware configuration details, data-center-wide resource utilization levels [30], and benchmarks, both from standards bodies such as from SPEC [10] and ad hoc ones, for estimating electricity per unit of resource usage. In other words, a single coefficient is estimated per unit of usage for each of the resources over a CSP's data centers in a region, regardless of hardware heterogeneity or the time of usage or the service for which the resource is consumed. As can be seen, the estimates can have wide discrepancies in comparison to actual measurements. Furthermore, TPAs seem to completely ignore overheads incurred at supporting systems and services and are oblivious to service dependencies. These limitations are, in general, acknowledged by the TPA providers [16].

3.3 Relevance and Validation of TPAs

Although CSP reports are available now with limited scope (which is expected to expand), anecdotal evidence suggests that the demand for third-party products and services for cloud CFP estimation is growing and that the third-party providers (TPPs) have plans of expanding the portfolio of their services. This interest is likely to continue not only until CSP reporting is sufficiently mature, but even beyond, due to the following reasons. First, despite the imprecise nature of their estimation, which can have non-trivial errors in the absolute terms, TPAs seem to be sufficient for reasonably gauging the relative carbon merits of different environments and leveraging it for holistic CFP management in hybrid multicloud environments. By design, due to the need to optimize by choosing among the services of multiple providers, TPPs will likely have a prominent role to play in carbon optimization in hybrid multicloud environments just as in the APM space. Second, the TPAs can be expected to evolve and enable a virtuous cycle of continuous improvement to reporting by CSPs. We envision that the discrepancy in reporting that is very likely between TPPs and CSPs will compel the latter to open up about their hidden overhead costs, infrastructure, service dependency, and accounting methodology, which will in turn enable TPPs to refine their methodology and serve as an independent evaluator as well as provider of solutions for carbon-aware cross-platform multicloud deployments. Finally, the cloud itself is still evolving, with new types of hardware and more novel and agile service offerings. Hopefully, the new services will be built with carbon and energy as first principles and integral facilities for CFP accounting, which would enable CFP reporting from the get-go. In the absence of such features, TPAs will have a significant part to play—even in the estimation space—in the interim.

Evaluating third-party estimates. TPAs currently provide only point estimates, which neither include confidence intervals nor have been validated (to the best of our knowledge) due to the lack of ground truth for their assumptions. With CSP reports trickling in (which can serve as *pseudo* ground-truth), TPPs can start providing accuracy metrics for their estimates. TPPs, however, may have to come up with creative methods for gathering a sufficient number of usage reports to provide statistically relevant accuracy measures.

4 RESEARCH OPPORTUNITIES

Carbon performance management (CPM) for hybrid multicloud is nascent and is, as such, ripe for innovation on various fronts, especially to keep up with and include within its ambit the emerging workloads and services, and progress in specialized hardware, such as GPUs, TPUs, and AIUs [22] catering to them. Similarly, CPM should cover non-x86 based systems such as the Z Mainframes [27], sought after by enterprises for their energy efficiency. Research is needed on the following: (1) The design of the right interfaces between different layers in the system stack spanning chip and system hardware / firmware, OS, and virtualization layers, and an open framework to unify the heterogeneous space for seamless energy and other input monitoring, fair energy apportioning, and CFP quantification and estimation. This will require the participation of many players, including hardware vendors, CSPs and their users, and the open source community. (2) Subtle aspects such as the impacts and amortizations from tenant history and multi-tenancy to CFP, which are not well understood and require attention, since co-tenancy is a given on shared public cloud. (3) New techniques that are evolving for estimating CFP for "to-be" workload placement scenarios, design of workload shifting, migration, and dispatching algorithms to increase the use of renewables and drive the cloud to true zero carbon [7]. The immediate need, however, is to improve CFP estimates of TPAs, under different data availability and transparency conditions, for the rest of the research to be evaluated objectively and be impactful. Here we discuss some thoughts.

Approaches for improving TPA. Consider a VM service whose plans are characterized by the number of vCPUs (q_c) and the amount of memory in GB (q_m) provisioned. TPAs currently assume a single or a small number of distinct coefficients each for the vCPUs and memory to arrive at a carbon estimate. Usage report from the CSP will indicate the total number of hours (h) the service was used for, the cost of the service, and, optionally, the associated CFP. The TPA's CFP estimate for a usage report of this service would roughly be ($ cost kept aside for simplicity):

$$\mathsf{CFP} = (q_c \times h \times w_c + q_m \times h \times w_m) \times \mathsf{PUE} \times \mathsf{CI}\,, \tag{2}$$

where w_c and w_m are the resource usage coefficients—that is, estimates for average energy consumed in kWh by a vCPU and a GB of memory, respectively, per hour in the region of the CSP where the VM is provisioned. Similarly, PUE and CI are TPA estimates.

In (2), w_c and w_m are estimates averaged (possibly with some weights) over the entire spectrum of CPU and server configurations and memory types, whereas PUE and CI would be averages over the time of interest. Fine-grained PUE and CI values, even if available, cannot be leveraged with coarse-grained usage reports.

Usage reports from the TPP's client pool can be used, to begin with, to assess the extent of discrepancy between their estimates and CSP-provided CFP values. Clustering techniques can be used to gain insights into the nature of discrepancy by classes such as the resource type and size, usage duration, and time of the year; CSP reports can be combined with self-reported data from clients (such as the time of usage, % usage of resource) to understand if they play a part. Finally, the collective data can be used to construct models that can help in dis-aggregating the CSP report into various components, thus helping to discover the extent of discrepancy in each of the input quantities assumed by the TPA or their models or both and possibly finetune them.

If, indeed, the CSP's CFP computation follows the model in (2) but for some overhead terms, then their CFP can be expressed as:

$$\mathsf{CFP}^{\mathsf{CSP}} = (q_c \times h \times w_c^{\mathsf{CSP}} + q_m \times h \times w_m^{\mathsf{CSP}} + O_h \times h + O) \times \mathsf{PUE}^{\mathsf{CSP}} \times \mathsf{CI}^{\mathsf{CSP}}, \tag{3}$$

where O_h and O indicate the duration-of-use dependent and independent overheads associated with the service in kWh. (Note that resource-volume-specific overhead, if any, will be absorbed by the resource energy coefficients.) The superscript CSP indicates that the associated quantity pertains to CSP. Letting $F = \mathsf{PUE}^{\mathsf{CSP}} \times CI^{\mathsf{CSP}}$, (3) can be expressed as

$$\begin{aligned} \mathsf{CFP}^{\mathsf{CSP}} &= q_c \cdot h \cdot w_c^{\mathsf{CSP}} \cdot F + q_m \cdot h \cdot w_m^{\mathsf{CSP}} \cdot F + O_h \cdot h \cdot F + O \cdot F \\ &= Q_c \cdot R_c^{\mathsf{CSP}} + Q_m \cdot R_m^{\mathsf{CSP}} + R_h \cdot h + R, \end{aligned} \tag{4}$$

where R_c^{CSP} and R_m^{CSP} can be thought of as carbon coefficients per unit of CPU and memory, respectively, and R_h and R as the per-hour and constant overhead carbon coefficients, respectively. Treating $\mathsf{CFP}^{\mathsf{CSP}}$ as the target variable y and $[Q_c, Q_m, h]$ as the feature vector $\mathbf{x}$, ML-based models can be used to determine the carbon coefficients. If good models can be fit with small errors, then the assumption of a single or narrow-ranged coefficient per dimension would be validated. Alternatively, clusters of records that fit a model instance can be analyzed to determine the underlying cause for the differences in model parameters of the different clusters. Variations may also be analyzed using data gleaned from the clients, such as time of usage, estimated utilization levels, and type of workload executed. Such analysis can be used by the TPAs to improve their CFP estimates. The approach sketched herein is just an initial step and can be extended to other resources and services.

The Way Forward. The opportunity and possible role for TPAs in the space of CPM for multicloud would depend to a large extent on how energy and carbon emission transparency of cloud and hybrid-cloud eco-systems evolve. Reverse engineering to determine cloud operational parameters, such as the one described above, can continue until there is sufficient transparency. Some ideas on enabling visibility into cloud energy system that call for virtualizing the entire energy sub-system to extend control to individual applications for optimizing their emissions have been explored in [4] and [50], Some other specific calls to action to various stakeholders in the CPM space that can help carbon-aware deployments in multicloud are as follows: (1) Development of an open framework to enable CFP computations and subsequent CFP reduction decisions. Currently APM tools like Instana collect resource utilization metrics and compute service-level metrics at very fine granularity. Similarly, resource optimization products like IBM Turbonomic [45, 53] collect resource utilization metrics and have recently been extended to tap into power utilization metrics within virtualization platforms like VMWare VSphere. Such efforts should be extended to a broader range of hardware devices, platforms, and services, and preferably, made open source. (2) A common benchmark suite for a variety of benchmark workloads, and their corresponding power and CFP estimates, can be established for different hardware configurations. These hardware configurations could vary in CPUs, GPUs, etc., and storage units. Such a benchmarking mechanism can draw from SPEC (which includes suites for IaaS Cloud and Power [10], in addition to the well-known CPU suite); this continues to drive hardware vendors to build better, efficient, and performant hardware. There is a need to establish a similar benchmarking system for carbon-aware computing or incorporate carbon awareness into SPEC that takes into account both the embodied and operational emissions to enable carbon vs. compute performance of various CSPs. (3) Development of mechanisms to seamlessly update the energy and carbon metering and attribution approaches to maintain currency and validate the approaches for correctness and consistency.

5 CONCLUSION

Enterprise computing is witnessing several important changes, including growing demand for resources, new hardware architectures, software abstractions, and workload deployment models, along with the use of renewable/stored energy sources. This conflation is leading to both a need for and an opportunity to jointly optimize for CFP via resource optimization, workload placement, and proactive energy modulation. As discussed, the methodology employed by CSPs to compute the CFP across services is not transparent. This is in addition to the coarse-grained nature in time and aggregation over a large number of services. Therefore, carbon performance management (CPM) tools and methods are needed at multiple levels to measure and apportion energy consumption, show energy footprint, translate energy to carbon, and present actionable recommendations to make informed choices and trade-offs. We strongly believe that CPM should become a first-class discipline that coexists with APM, ARM, and SLM. Manufacturers of hardware (GPUs, TPUs, AIUs, memory and storage units, network interconnect and switches), CSPs, and developers of APM and ARM tools need to work together with a sense of urgency to develop open, transparent frameworks to achieve both CFP estimation and reduction.

REFERENCES

[1] International Energy Agency. [n.d.]. Data and Statistics. https://www.iea.org/data-and-statistics/data-sets

[2] International Energy Agency. [n.d.]. Global data centre energy demand by data centre type, 2010-2022. Web page. https://www.iea.org/data-and-statistics/charts/global-data-centre-energy-demand-by-data-centre-type-2010-2022

[3] Jeff Barr. [n.d.]. Cloud Computing, Server Utilization, the Environment. https://aws.amazon.com/blogs/aws/cloud-computing-server-utilization-the-environment/

[4] 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)*. Association for Computing Machinery, New York, NY, USA, 350–358. https://doi.org/10.1145/3472883.3487009

[5] Mark Bohr. 2007. A 30 Year Retrospective on Dennard's MOSFET Scaling Paper. *IEEE Solid-State Circuits Society Newsletter* 12, 1 (2007), 11–13.

[6] Mainak Chakraborty and Ajit Pratap Kundan. 2021. *Grafana*. Apress, 187–240. https://doi.org/10.1007/978-1-4842-6888-9_6

[7] Andrew A. Chien. 2021. Driving the Cloud to True Zero Carbon. *Commun. ACM* 64, 2 (jan 2021), 5.

[8] Andrew A. Chien. 2022. Computing's Grand Challenge for Sustainability. *Commun. ACM* 65, 10 (Oct 2022), 5.

[9] climatiq. [n.d.]. Cloud Computing Carbon Emissions. REST API. https://www.climatiq.io/cloud-computing-carbon-emissions

[10] Standard Performance Evaluation Corporation. [n.d.]. SPEC. https://www.spec.org/

[11] Cem Dilmegani. 2023. Multi-Cloud vs. Hybrid Cloud: A Comprehensive Guide in 2023. Web article of AI Multiple. https://research.aimultiple.com/multi-cloud-vs-hybrid-cloud

[12] Dynatrace. [n.d.]. Application Monitoring; Cloud Monitoring. Website. http://dynatrace.com/solutions/application-monitoring;http://dynatrace.com/solutions/cloud-monitoring

[13] eBPF. [n.d.]. Dynamically program the kernel for efficient networking, observability, tracing, and security. https://ebpf.io/

[14] Interesting Engineering. [n.d.]. No more transistors: The end of Moore's Law. Web article. https://interestingengineering.com/innovation/transistors-moores-law

[15] Teads Engineering. [n.d.]. Carbon footprint estimator for AWS instances. Web App. https://engineering.teads.com/sustainability/carbon-footprint-estimator-for-aws-instances/

[16] Cloud Carbon Footprint. [n.d.]. Cloud Carbon Footprint. Open Source Project. https://github.com/cloud-carbon-footprint/cloud-carbon-footprint

[17] Google. [n.d.]. Carbon Footprint reporting methodology. web article. https://cloud.google.com/carbon-footprint/docs/methodology

[18] Google. [n.d.]. Google Data Center PUE performance. https://www.google.com/about/datacenters/efficiency/

[19] Google. [n.d.]. Services covered by Google Carbon Footprint. https://cloud.google.com/carbon-footprint/docs/covered-services

[20] National Grid. [n.d.]. What is carbon intensity? https://www.nationalgrid.com/stories/energy-explained/what-is-carbon-intensity

[21] The Green Grid. [n.d.]. PUE: A Comprehensive Examination of the Metric. https://datacenters.lbl.gov/sites/default/files/WP49-PUE%20A%20Comprehensive%20Examination%20of%20the%20Metric_v6.pdf

[22] IBM. [n.d.]. IBM AIU—A System On A Chip Designed For AI. https://research.ibm.com/blog/ibm-artificial-intelligence-unit-aiu

[23] IBM Inc. [n.d.]. Are Your Data Centers Keeping You From Sustainability? Blog. https://www.ibm.com/cloud/blog/are-your-data-centers-keeping-you-from-sustainability

[24] World Resources Institute. [n.d.]. GHG Protocol Scope 2 Guidance. https://ghgprotocol.org/sites/default/files/Scope2_ExecSum_Final.pdf Page 4.

[25] World Resources Institute. [n.d.]. Greenhouse Gas Protocol carbon reporting and accounting standards. https://ghgprotocol.org/standards

[26] World Resources Institute. [n.d.]. The Greenhouse Gas Protocol. https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

[27] Christian Jacobi and Charles Webb. 2020. History of IBM Z Mainframe Processors. *IEEE Micro* 40, 6 (2020), 50–58.

[28] Andrea Janes, Xiaozhou Li, and Valentina Lenarduzzi. 2022. Open Tracing Tools: Overview and Critical Comparison.

[29] Nordcloud Klarity. [n.d.]. Identify and reduce your cloud carbon footprint. https://klarity.nordcloud.com/blog/join-greenops-revolution-with-klarity/

[30] Lawrence Berkeley National Laboratory. [n.d.]. United States Data Center Energy Usage Report. Web. https://eta.lbl.gov/publications/united-states-data-center-energy

[31] Gartner Laurence Goasduff. 2019. Why Organizations Choose a Multicloud Strategy. Web article of Gartner. https://www.gartner.com/smarterwithgartner/why-organizations-choose-a-multicloud-strategy

[32] Jingming Li, Nianping Li, Jinqing Peng, Haijiao Cui, and Zhibin Wu. 2019. Energy consumption of cryptocurrency mining: A study of electricity consumption in mining cryptocurrencies. *Energy* 168 (2019), 160–168. https://www.sciencedirect.com/science/article/pii/S0360544218322503

[33] Electricity Maps. [n.d.]. Reduce carbon emissions with actionable electricity data. Web Service. https://app.electricitymaps.com/ accessed: 20 May 2023.

[34] Microsoft. [n.d.]. The carbon benefits of cloud computing. Web. https://download.microsoft.com/download/7/3/9/739BC4AD-A855-436E-961D-9C95EB51DAF9/Microsoft_Cloud_Carbon_Study_2018.pdf

[35] Microsoft. [n.d.]. Cloud for Sustainability API (Preview) overview. https://learn.microsoft.com/en-us/industry/sustainability/api-overview

[36] Microsoft. [n.d.]. Connect the Microsoft Sustainability Calculator. https://docs.microsoft.com/en-gb/power-bi/service-connect-to-microsoft-sustainability-calculator

[37] Microsoft. [n.d.]. Design methodology for sustainable workloads on Azure. https://learn.microsoft.com/en-us/azure/well-architected/sustainability/sustainability-design-methodology

[38] Microsoft. [n.d.]. How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability. https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability/

[39] Microsoft. [n.d.]. Microsoft Cloud for Sustainability API calculation methodology. https://learn.microsoft.com/en-us/industry/sustainability/api-calculation-method

[40] Microsoft. [n.d.]. Microsoft datacenter sustainability fact sheets. https://datacenters.microsoft.com/globe/fact-sheets

[41] David Mytton. 2020. Assessing the suitability of the Greenhouse Gas Protocol for calculation of emissions from public cloud computing workloads. *Journal of Cloud Computing: Advances, Systems and Application* (2020), 11 pages.

[42] OpenAI. 2023. GPT-4 Technical Report. arXiv:cs.CL/2303.08774

[43] Priceonomics. [n.d.]. The IoT Data Explosion: How Big Is the IoT Data Market? Web article. https://priceonomics.com/the-iot-data-explosion-how-big-is-the-iot-data/

[44] Priceonomics. [n.d.]. Sensors Explosion and the Rise of IoT. Blog. https://www.diamandis.com/blog/sensors-and-iot

[45] Pethuru Raj and Anupama Raman. 2018. *Multi-cloud Management: Technologies, Tools, and Techniques*. Springer International Publishing, Cham, 219–240. https://link.springer.com/chapter/10.1007/978-3-319-78637-7_10

[46] RedHat. [n.d.]. Kepler (Kubernetes Efficient Power Level Exporter). https://github.com/sustainable-computing-io/kepler Open Source Project.

[47] 451 Research. [n.d.]. The Carbon Reduction Opportunity of Moving to Amazon Web Services. https://d39w7f4ix9f5s9.cloudfront.net/e3/79/42bf75c94c279c67d777f002051f/carbon-reduction-opportunity-of-moving-to-aws.pdf

[48] Navin Sabharwal and Piyush Pandey. 2020. *Container Application Monitoring Using Dynatrace*. Apress, 183–233. https://doi.org/10.1007/978-1-4842-6216-0_7

[49] Vishal Sharma. 2016. *Getting Started with Kibana*. Apress, 29–44. https://doi.org/10.1007/978-1-4842-1694-1_3

[50] Abel Souza, Noman Bashir, Jorge Murillo, Walid Hanafy, Qianlin Liang, David Irwin, and Prashant Shenoy. 2023. Ecovisor: A Virtual Energy System for Carbon-Efficient Applications *(ASPLOS 2023)*. 252–265.

[51] Holly Stower. [n.d.]. The Future Faster: Corporate Sustainability Monitoring. Web article by Cleantech Group. https://www.cleantech.com/the-future-faster-corporate-sustainability-monitoring/

[52] The New York Times. [n.d.]. A.I. Here, There, Everywhere. Web article. https://www.nytimes.com/2021/02/23/technology/ai-innovation-privacy-seniors-education.html

[53] Turbonomic. [n.d.]. IBM Turbonomic. https://www.ibm.com/products/turbonomic

[54] James Turnbull. 2018. *Monitoring with Prometheus*. Turnbull Press.

[55] Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Donguk kim, Sung-Ho Bae, Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon, and Choong Seon Hong. 2023. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? arXiv:cs.AI/2303.11717