Optimizing Machine Learning Benchmarking: A FAIR Approach to Energy Efficiency and Data Transparency

Christine R. Kirkpatrick,[1] Gregor von Laszewski,[2] Juri Papay,[3] Jeyan Thiyagalingam,[3] Gregg Barrett,[4] Wesley Brewer,[5]Julie Christopher,[1] Inês Dutra,[7] Murali Emani,[8] Piotr Luszczek,[9] Mallikarjun (Arjun) Shankar,[5] and Geoffrey C. Fox [2,11]

[1] San Diego Supercomputer Center, UC San Diego, CA, 92093, USA
[2] Biocomplexity Institute, Charlottesville, Virginia, 22911, USA
[3] Rutherford Appleton Laboratory, Science and Technology Facilities Council, United Kingdom
[4] Cirrus AI, Johannesburg, South Africa
[5] Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37831, USA
[7] Department of Computer Science, Faculty of Sciences of University of Porto, Porto, Portugal
[8] Argonne National Laboratory, Lemont, Illinois, 60439, USA
[9] University of Tennessee, Knoxville, Tennessee, 37996, USA
[11] Computer Science Department,, Charlottesville, Virginia, 22911, USA

Abstract

Machine learning (ML) training and inference are resource-intensive, leading to high energy consumption and increased environmental impact. This study applies FAIR (Findability, Accessibility, Interoperability, and Reusability) principles to MLCommons benchmarking results to enhance benchmarking transparency, power efficiency analysis, and data quality. By restructuring metadata and the results, standardizing power consumption reporting, and refining data accessibility, we improve the comparability and reproducibility of ML performance results across diverse hardware architectures. The introduction of Power Consumption Fingerprints provides a structured method for detecting workload-hardware mismatches, optimizing energy efficiency, and reducing computational overhead. Our findings demonstrate that FAIR-compliant benchmarking leads to more actionable insights, better logging practices, and improved interpretability, ultimately enabling the development of more energy-efficient AI systems.

1. Introduction

DOI:

AI benchmarking activities provide an opportunity to compare and contrast the runtime and task performance of different machine learning (ML) models across various hardware systems. These AI benchmarks are used in a conventional sense to assess how quickly and accurately ML models can be trained and/or how fast the inference can be on given datasets. However, raw performance metrics, such as training time or inference time, alone are not sufficient to fully describe the problems or systems in terms of their efficiency, for instance energy efficiency. This has encouraged the community to establish benchmarking efforts for quantifying energy efficiency or resource efficiency [Citations Needed]. Initiatives, such as MLPerf Power Benchmarking, focus on quantifying, for instance, energy efficiency of systems, using a suite of AI/ML benchmarks: A set of ML-specific tasks operating on a given set of datasets. Benchmarks contained in such efforts often produce a large volume of very detailed sets of results, which are often overlooked.

While such efforts have paved a way for quantifying energy efficiency and thus advancing ML performance, less attention has been given to the quality, structure, and usability of the benchmarking results themselves. While the focus on ML models and benchmark datasets (used for training or inference) is important for understanding energy efficiency of AI models, we firmly believe that significant, if not equal, emphasis must be placed on the benchmarking results themselves. Benchmarking results contain valuable information about how ML workloads interact with different hardware and software configurations, offering insights into power consumption, compute efficiency, and potential system bottlenecks. However, with current benchmark reporting practices lacking any form of standardization, extracting meaningful information around energy efficiency of ML models
or tasks become inconceivably difficult. Inconsistent metadata, missing power consumption data, and unstructured logging formats are some key examples that prevent benchmarking results from being effectively mined for gaining useful insights or for making useful decisions.

In this paper, we use a set of MLCommons benchmarks as the driving example to demonstrate the view that high-quality benchmarking results are crucial as high-quality ML models and (training or inferencing) datasets. More specifically, we show that by improving how benchmarking results are structured, stored, and shared, the practical limitations of AI benchmarking for deriving energy-efficient computing strategies can be mitigated. In doing that, we perform two interconnected operations, namely, (a) we perform a rigorous analysis of the results for highlighting the issues that could limit their utility, and (b) we assess the benchmark results using the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles for improving the usability of the results. With these two in place, we then propose a concrete set of enhancements that can make these results more useful for optimizing ML models for energy efficiency or performance improvements. One of the enhancements we propose is the concept of “power consumption fingerprints” in the benchmark results. These fingerprints, when adhered, can aid in detecting energy consumption mismatches between ML workloads and underlying hardware.

We believe that this approach to AI benchmarking can lead to several benefits, including, better resource allocation, improved energy-aware ML training and inference operations, ability to develop hardware-software co-designed algorithms for failure prediction and fault

 DOI:

diagnosis. and improved interpretation, validation and visualization of results, and hence, leading to better dissemination of the findings. All these can collectively pave a way for sustainable AI computing practices.

(Boehme et al. 2016)
(Petkov 2014)

The rest of this paper is organized as follows. In Section 1, we provide detailed background approaches to benchmarking for energy efficiency, benchmarks for the comparison of power consumption across different computational systems, and outlining known limitations of measurement. This is then followed by Section 2 where we describe our findings about energy consumption using ML analyses using data and tools made available by MLCommons Science.. Section 3 presents a detailed insight into the results through data analysis, including limitations of the current benchmark results and suggestions for improving the data through extended tools and research data management techniques. Section 4 concludes the paper with future directions.

2. Background

2.1 Benchmarking in High Performance Computing and AI/ML

Benchmarking has long been a cornerstone of evaluating computational systems, particularly in the High Performance Computing (HPC) community. The TOP500 list, ranks the world's most powerful supercomputers based on their performance in the High Performance LINPACK (HPL) benchmark (Meuer et al. 2021), which measures floating-point operations per second (FLOPS) using double-precision matrix-matrix multiplication (DGEMM). While HPL has been instrumental in assessing raw computational power, its reliance on dense linear algebra operations makes it less relevant for modern machine learning (ML) workloads, which often prioritize lower-precision arithmetic and sparse operations [HPL citation from JD]. Recognizing this limitation, the HPC community has introduced metrics such as HPL-MxP (formerly HPL-AI), which evaluates mixed-precision FLOPS. However, these benchmarks still focus on dense matrix operations, which do not fully capture the computational patterns of ML workflows [Citation needed].

To address the growing demand for ML-specific benchmarking, initiatives such as MLCommons have emerged. These efforts focus on evaluating ML models and systems using workloads that better reflect real-world AI applications, such as image classification, natural language processing, and scientific cases, like space group identification in electron diffraction data //Junqi's Paper//. However, while these benchmarks provide insights into runtime performance and accuracy, they often lack comprehensive energy efficiency metrics.

This gap highlights the need for benchmarking practices that not only measure computational performance but also quantify energy consumption and resource utilization.

2.2 Energy Consumption of Computing and AI/ML Workloads

The growing energy demands of computing systems have become a significant concern particularly with data centers now accounting for approximately 1-1.5% of global electricity consumption. This challenge is further compounded by the rapid expansion of artificial intelligence (AI and machine learning (ML) workloads, which are projected to increase data centre energy demand by up to 160% by 2030 (Prisco et al. 2020). It is not an overstatement to say that AI-specific systems, which are extensively used for AI/ML applications, represent a significant share of these energy demands.

////.

(AI is poised to drive 160% increase i...)

A single rack of modern AI-specific systems can easily consume between 25kW and 140kW, while next-generation AI-optimized systems are expected to push energy requirements even further, reaching 160–200 kW per rack //Tate, B., 2022, 2024//. The rising power demands of AI workloads have driven significant advancements in energy efficiency and cooling technologies. Notably, direct-to-liquid cooling (DLC) has emerged as a promising approach to mitigating thermal constraints in HPC and AI systems. Furthermore, initiatives such as the ARPA-E COOLERCHIPS program aim to reduce cooling-related energy expenditures to less than 5% of a data center's total IT load, enhancing overall energy efficiency //CITATION//.

Despite these innovations, accurately measuring and optimizing the energy consumption of AI/ML workloads remains a complex challenge. Power usage varies significantly due to heterogeneous hardware architectures, software configurations, and dynamic workload characteristics. For instance, HPC systems exhibit power fluctuations of up to 2 MW within seconds, depending on the computational phase (Prisco et al. 2020). These variations demonstrate the need for robust, standardized methodologies to monitor, model, and optimize power efficiency in AI-driven computing environments.

(Prisco et al. 2020)

DOI:

2.3 Power Consumption Metrics and Challenges

Quantifying the energy efficiency of computing systems is further complicated by the lack of standardized power consumption metrics. Processor manufacturers often provide Thermal Design Power (TDP) values, which indicate the maximum heat a cooling system must dissipate under sustained workloads. However, TDP is not designed to measure energy efficiency or operational power consumption, and it often underestimates peak power demands (Ganapathy & Warner, 2008). For example, AMD's Average CPU Power (ACP) metric has been criticized for providing overly optimistic estimates of power consumption . (De Gelas) ([untitled item]) //huck//. These limitations make it difficult to compare the energy efficiency of different systems or to optimize hardware-software configurations for energy-aware computing.

To address these challenges, the Energy Efficiency High Performance Computing Working Group //EE HPC WG// has developed methodologies and tools, such as the Power API, for measuring and analyzing energy consumption in HPC systems //EE HPC WG, 2018//. These efforts have laid the groundwork for more comprehensive energy efficiency benchmarking, but their application to AI/ML workloads remains limited. As AI/ML systems continue to grow in scale and complexity, there is an urgent need for standardized energy efficiency metrics and benchmarking practices that can provide reliable and comparable data for optimizing both hardware and software.

(De Gelas)
Hueck needs manual add ([untitled item])
(Ganapathy and Warner 2008)

2.4 Challenges in Understanding AI/ML Benchmarking Results

AI/ML benchmarking results are typically recorded in logs generated by custom programs, user-defined scripts, or specialized tools. These logs capture details on runtime performance, resource utilization, and sometimes energy consumption, but lack standardized formats and reporting practices, making them difficult to compare and analyze. Unlike benchmarks such as TOP500, which adhere to strict reporting standards, AI/ML benchmark results are often unstructured and inconsistent, as they are tailored to specific hardware configurations, software stacks, or research objectives. Power consumption data, for instance, may be recorded using different sampling rates, units, or levels of granularity, while metadata such as hardware specifications and software versions is often incomplete or inconsistent. These variations hinder the aggregation and comparison of benchmarking results, limiting their utility in optimizing energy efficiency across different ML systems.

Furthermore, reliance on ad-hoc logging practices and custom scripts introduces errors, missing values, and inconsistencies in data collection and processing, particularly in energy efficiency analyses, where even minor discrepancies in power measurements can significantly affect conclusions. While initiatives such as MLPerf provide structured performance benchmarks, their power measurement results often lack the necessary granularity and consistency for detailed energy assessments. This lack of standardization prevents the identification of broad trends or patterns, complicating the development of energy-aware optimization strategies. Addressing these challenges requires establishing consistent reporting formats, improving data quality, and ensuring interoperability across benchmarking initiatives to make benchmarking results more actionable for advancing energy-efficient AI computing.

3. Evaluation

3.1 Overview of MLCommons Benchmarks

MLCommons is a consortium dedicated to improving the accessibility, standardization, and performance of machine learning (ML) benchmarks /Cite MLCommons web page/, organized as multiple working groups, each focussing on different aspects of ML benchmarking, such as, performance, power efficiency, and domain-specific applications. MLPerf Training and Inference Benchmarks assess ML models across computational platforms, focusing on both training and inference tasks and covering areas such as image classification (ResNet), natural language processing (BERT), and scientific computing /Reference needed/. The MLCommons Science Benchmarks evaluate ML models in scientific applications, such as Earthquake forecasting and CloudMask detection /Thiyagalingam et al., 2022/, while the MLPerf Power Benchmarking initiative focuses on energy efficiency by measuring power consumption across ML workloads / Reference needed/. The MLCommons Algorithm Working Group evaluates training algorithms and benchmarks for the given models and hardwares / arxiv.org/pdf/2306.07179/. Additionally, the MLCommons Data Working Group enhances dataset accessibility, format standardization, and metadata quality to facilitate structured benchmarking / Reference needed/.

MLCommons working groups and members therein contribute benchmark results through systematic submissions. Participating member organizations, including leading technology companies, research labs, and academic institutions, run standardized ML workloads on their hardware and software configurations to evaluate performance, energy efficiency, and scalability. These submissions, reviewed and aggregated by relevant working groups, provide insights into the effectiveness of various architectures and drive optimizations for future ML models and hardware platforms.

 DOI:

3.2 Selection of Benchmark Results

To evaluate the impact of structured and FAIR-compliant benchmarking, we focus on a limited set of benchmark results from MLCommons, namely, ResNet, BERT and Earthquake Forecasting, that provide a representative cross-section of ML workloads in diverse application areas. We first provide a rationale for the selection of these benchmarks is as follows:

- **Broad Applicability:** The chosen benchmarks represent three major machine learning domains, namely, computer vision, natural language processing (NLP), and scientific computing. These areas are among the most widely studied and deployed in both research and industry applications.

- **Diversity of Computational Demands:** The selected benchmarks differ significantly in their computational profiles, covering both dense and sparse computation, varying memory requirements, and different levels of parallelism. More specifically, ResNet primarily evaluates deep convolutional networks, BERT represents transformer-based NLP workloads, and Earthquake Forecasting showcases scientific AI applications.

- **Hardware Comparability:** These benchmarks have been widely adopted across multiple hardware architectures, including GPUs, TPUs, and CPUs. This allows for meaningful comparisons of power efficiency and performance scaling across different processing environments.

- **Energy Efficiency Insights:** One of the key objectives of this study is to analyze power consumption and energy efficiency in ML training workloads. ResNet, BERT, and Earthquake Forecasting have been frequently used in MLPerf Power Benchmarking efforts, making them ideal candidates for studying energy efficiency under FAIR principles.

- **Benchmarking Standardization:** Each of these benchmarks is part of MLCommons' standardized benchmarking suite, ensuring consistency in performance measurements, dataset usage, and evaluation criteria. The standardized nature of these benchmarks reduces confounding factors that could arise from using non-standardized ML workloads.

- **Availability of Results:** The selected benchmarks have well-documented historical results from MLPerf Training benchmarks (v1.0, v1.1, v3.0) and MLCommons Science submissions, enabling comparative analysis before and after applying FAIR principles.

The detailed description of the benchmarks can be found in MLCommons documentations / CITATION NEEDED/, but for the reasons of convenience, we provide a short description below:

- **ResNet Training Benchmark Results (**From MLPerf (or MLCommons?) Training Working Group): This benchmark is used for image classification tasks on the ImageNet dataset (Deng et al., 2009). The benchmark evaluates training efficiency across GPUs, TPUs, and CPUs. For this study, we included the results data from multiple Training benchmark versions (v1.0, v1.1, v3.0). When the results are perfect, they should help assess hardware acceleration benefits and power consumption efficiency across different systems.
- **BERT Training Benchmark Results** (From MLPerf (or MLCommons?) Training Working Group): Evaluates natural language processing (NLP) tasks, specifically masked language modeling /? / , and measures performance and scalability across different hardware configurations. The benchmark is intended to provide insights into the compute-intensive nature of transformer models and their energy consumption. As such, it is very useful for comparing efficiency gains from newer accelerator architectures.
- **Earthquake Forecasting Benchmark Results** (From MLCommons Science Working Group): Focuses on predicting seismic events using deep learning models / Main paper that Geoffrey cited /. The benchmark results incorporate power consumption tracing to assess the energy efficiency of the application, and the benchmark compares GPU and CPU performance (?) / Cite Geoffrey's Paper/.

We summarize these in Table 1 below.

Table 1: Summary of Benchmark Results Used for Our Studies

Benchmark	Source (MLCommons Working Group)	Task	Hardware
ResNet	Training	Image classification	GPU, CPU, TPU
BERT	Training	NLP - Masked LM	GPU, TPU
Earthquake	Science	Time Series Forecasting	GPU

3.3 FAIR Compliance & Enhancements

To improve the usability of the benchmark results, we applied FAIR (Findability, Accessibility, Interoperability, and Reusability) principles. Below, we describe the specific steps we have taken.

Findability Enhancements

DOI:

- Unique Identifiers: We assigned globally unique, persistent identifiers (GUPRIs) to submissions, ensuring consistent referencing (Wilkinson et al., 2016).
- Metadata Standardization: We enriched metadata with structured naming conventions for hardware, software versions, and configurations to eliminate ambiguities (e.g., clarifying “AMD ROME” into precise SKUs).

Accessibility Improvements

- Reformatted Data Storage: Power consumption data, execution times, and system configurations were extracted from raw logs and structured into a machine-readable format (CSV/JSON) instead of proprietary HTML tables.
- Standardized Directory Structure: We enforced a structured directory layout for benchmark results, ensuring that each run’s metadata and performance logs were easy to access and compare (Mons et al., 2017).

Interoperability Enhancements

- Adoption of Controlled Vocabularies: We mapped hardware descriptions to existing ontologies (e.g., processor models, accelerator types) to facilitate comparisons across datasets (Jacobsen et al., 2020).
- Cross-System Comparability: Benchmark results were transformed into a uniform schema, allowing comparisons across CPUs, GPUs, and TPUs, irrespective of manufacturer-specific variations.

Reusability Improvements

- Power Consumption Data Validation: We cleaned and validated energy consumption metrics to ensure consistency in reporting power usage across different hardware configurations.
- Benchmark Log Cleanup: We identified and removed inconsistencies (e.g., incorrect numbering of benchmark runs, missing software versions) to improve reproducibility.
- Integration with Cloudmesh: We developed an automated logging mechanism to extract and store structured energy traces, enabling long-term benchmarking usability.

3.4 Addressing Limitations in Power Metrics: TDP and Other Considerations

One of the major challenges in evaluating ML benchmarks is the unreliability of TDP (Thermal Design Power) as a performance and power metric. TDP is frequently used as a proxy for estimating power consumption, but it has several limitations. First of all, TDP represents theoretical maximum power, and not actual consumption. As such, the actual power usage varies significantly based on workload characteristics and system optimizations opposed to expected maximum thermal dissipation of the component under sustained load. Secondly,

different vendors define TDP differently. For example, AMD, Intel, and NVIDIA report TDP using different methodologies, making direct comparisons misleading. Finally, TDP does not capture workload-specific variations. In our case, ML workloads have fluctuating power demands due to different computational phases, such as data preprocessing, model training, and inference. To mitigate these issues, we incorporated a number of steps, including:

- We measured actual power usage wherever possible or available instead of relying solely on TDP. For this, we utilized real-time power consumption logs from system monitoring tools.
- We standardized power reporting by enforcing the use of uniform power measurement methodologies across all hardware setups.
- We performed cross-validation with MLCommons power benchmarks, ensuring consistency with best practices in ML power efficiency research.
- We examined energy consumption profiles across different training phases, and identified inefficiencies and optimized power usage.

With these, we ensured that our benchmarking analysis provides realistic views into ML workload energy efficiency, instead of oversimplifying the results by solely relying on TDP.

3.5 Handling Outliers and Performance Variations

Outlier detection is an essential part of benchmarking to ensure data integrity and meaningful comparisons. Although outlier detection is not part of the FAIR compliance, we include a post-processing step to identify the extreme cases where certain configurations would exhibit significantly higher power consumption than expected, likely due to inefficient resource utilization or suboptimal workload scheduling.

As such, the post-processing step includes the application of statistical techniques, including standard deviation analysis and quartile-based detection, to remove extreme results that could skew comparisons. This step should offer more stable performance trends, allowing for clearer insights into energy efficiency trade-offs. Although extreme values are likely to be outliers, whenever there are multiple extreme values, the notion of outliers becomes arguable. For this purpose, in our case, we define outliers as submission values that demonstrate extreme energy spikes without corresponding performance improvements, which in normal circumstances would be deemed unreliable for comparative analysis.

By incorporating outlier detection as part of the FAIR methodology, we ensured that our benchmarking results remained accurate, consistent, and truly representative of hardware efficiency trends.

4. Results and Analysis

4.1 Benchmark Data Before & After FAIR Application

 DOI:

To evaluate the impact of FAIR principles on benchmarking data, we analyzed the benchmark results before and after applying structured methodologies. The results demonstrate significant improvements in data accessibility, comparability, and usability. We first summarise the impact of applying FAIR principles to the benchmarking results in Table 2.

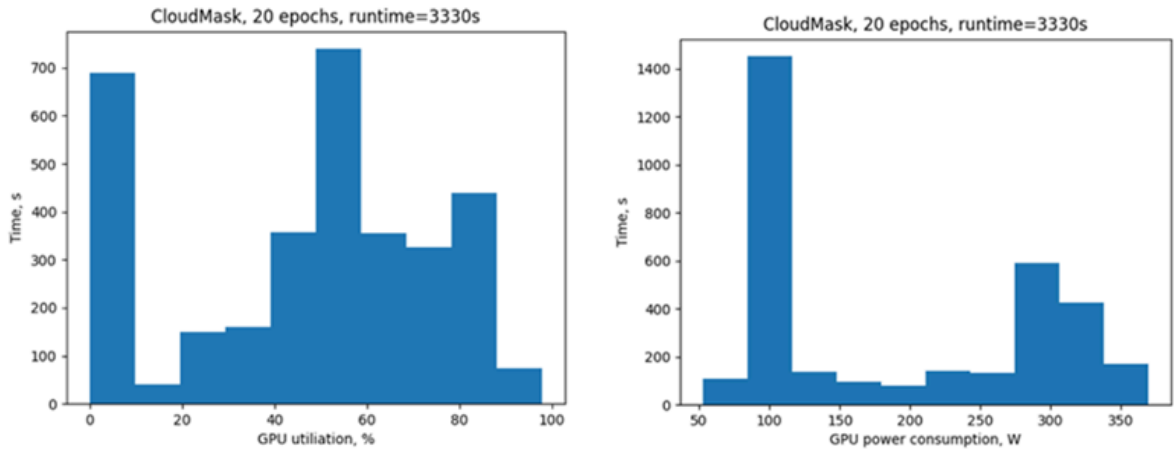

Table 2: Comparative Analysis of Benchmark Data Before and After FAIR Compliance

Metric	Before FAIR Compliance	After FAIR Compliance
Metadata Consistency	Missing software versions, vague hardware labels	Fully structured metadata with precise naming conventions
Data Accessibility	Logs stored in unstructured formats (e.g., raw text, proprietary HTML tables)	Standardized, machine-readable CSV/JSON format
Interoperability	Hardware configurations varied across submissions, making comparisons difficult	Controlled vocabularies enabled direct comparisons across CPUs, GPUs, TPUs
Power Consumption Data	Often missing, unstructured, or inconsistently logged	Standardized reporting, validated TDP and energy metrics

Table 3: Scaled System Performance for the MLCommons 2022 Training Submissions for ResNet and BERT Benchmark

ResNet	Etc	Etc

4.2 Energy Efficiency Analysis

After applying FAIR-compliant restructuring, we conducted an in-depth energy efficiency analysis. Ensuring data completeness and consistency enabled us to do a number of things, which have not been easy, if not impossible before. These are (a) Identify inefficient hardware configurations that consume excessive power for minimal performance gains, (b) compare power efficiency across architectures using uniform metrics instead of vendor-specific reporting, and (c) generate more reliable energy consumption fingerprints, allowing for detection of inefficiencies in workload execution.

4.3 Case Study: Power Consumption Fingerprints

To further illustrate the benefits of FAIR principles, we introduce Power Consumption Fingerprints as a structured approach to analyzing energy efficiency in ML workloads. Using Cloudmesh, we generated energy traces for different computational phases in the Earthquake Forecasting Benchmark. The results revealed patterns in power usage, highlighting inefficiencies in data loading and model execution.

DOI:

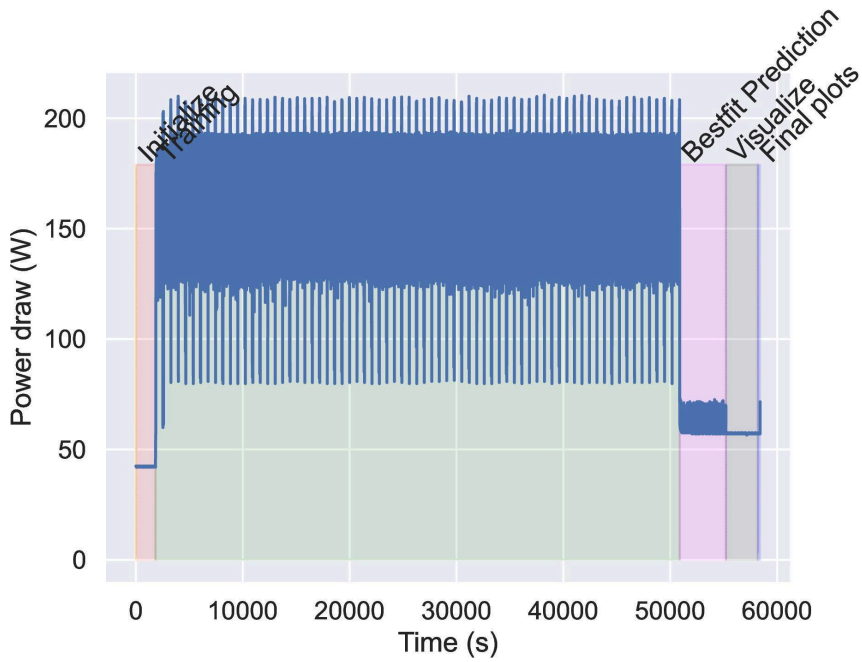

Figure 4: Power Consumption Trace of Cloudmesh Earthquake Forecasting Benchmark

Through this method, we detected inefficient power spikes in model training, prompting optimizations such as adaptive batch sizing and preloading strategies to reduce unnecessary energy consumption.

4.4 Additional Observations

One of the key benefits of FAIR compliance is the improved ability to systematically analyze and compare benchmarking results. Standardized logging and structured metadata enabled improved interpretability of results across different submissions, enhanced reproducibility, ensuring that benchmarking data remains useful for future comparisons, and easier integration with MLCommons power benchmarking efforts.

Furthermore, to understand the trade-offs between performance and energy consumption, we analyzed multiple benchmarking runs under different configurations. The results indicate that while some configurations achieve higher throughput, they often do so at the expense of power efficiency.

5. Conclusions and Future Work

This study demonstrated that applying FAIR principles to MLCommons benchmarking results significantly improves data accessibility, comparability, and usability. By restructuring the benchmarking results, ensuring consistent power reporting, and standardizing directory structures, we enhanced the reliability of benchmark results for both performance and energy efficiency analyses. The introduction of Power Consumption Fingerprints provided a structured approach to detecting and mitigating inefficient energy use, while improvements in logging and reporting allowed for greater interpretability and reproducibility. Addressing inconsistencies in power consumption metrics and eliminating unreliable TDP estimates further strengthened the benchmarking framework, ensuring that results are more actionable for optimizing machine learning workloads.

Future research should focus on expanding FAIR-compliant benchmark reporting to additional workloads, integrating real-time power monitoring for adaptive benchmarking, and incorporating energy efficiency considerations into AutoML-like frameworks. Exploring privacy-preserving benchmarking approaches, developing Green AI benchmarks for sustainability assessments, and extending FAIR principles to other domains such as biomedical AI and climate modeling would further enhance the impact of this work. As AI continues to underpin every aspect of computing, standardizing the way the results are reported will be critical in shaping benchmarks that optimize not just for performance, but also for transparency, energy efficiency, and ethical AI development.

Acknowledgements

The work of C.R. Kirkpatrick was supported in part by the National Science Foundation under Grants 1916481 and 2226453. The work from Gregor von Laszewski and Geoffrey Fox was supported in part by DE-SC0023452: FAIR Surrogate Benchmarks Supporting AI and Simulation Research, NSF 2346173 POSE: Phase II: MLCommons Research for Science: Enabling Open-Source Ecosystems for Scientific Foundation Models by Community Standards and Benchmarks, and NSF 2411009 Elements: A Sustainable, Resource-Efficient Cyberinfrastructure for Notebook Interactive ML Training Workloads. Juri Papay and Jeyan Thiyagalingam are supported by the Blueprinting AI for Science at Exascale (EP/????/????), and by the AI for Realistic Science (AIRS) funding from EPSRC, and DSIT/UKRI, respectively. We also would like to thank Ismael Kherroubi for his help with editing this paper.

DOI:

References

<<OLD SECTiONS — DO NOT DELETE YET >>

The MLCommons consortium began in 2018 with MLPerf benchmarks. The aim is to drive improvements in ML through sharing of performance data using established benchmarks (Thiyagalingam et al. 2022; MLCommons, n.d.a). The consortium has profiled systems such as commodity hardware readily available for purchase, and proprietary and specialized hardware in the commercial cloud, whose relative performance would not otherwise be known to the larger ML community. The benchmarks have expanded into eight

types (see Supplementary Material). The analysis discussed here focuses on the image classification benchmark in “Training” and one of the classification benchmarks for “Science.”

MLCommons benchmark results provide data that can be mined for insights and an opportunity to include power consumption data with results. As with the Green500, this would allow observations to be a balance of computational efficiency and power consumption. In recent benchmark rounds for MLPerf Inference (3.0) and Mobile (3.0), many submissions included accelerators’ TDP values.

MLCommons benchmark data includes a summary table of results shared on the MLCommons website, as well as corresponding log files with benchmark data stored as text files, and system profiles in JSON. Each entry includes up to five results per submission for a particular machine or cluster. The results are stored as files with an organized directory structure on the MLCommons GitHub (MLCommons; n.d.b).

3.1 Limitations of Measurement

Using TDP or other self-reported manufacturer specifications is limiting where those specifications have not been disclosed. In the case of MLCommons benchmarks and the various systems used by those submitting benchmarks, some utilize proprietary processors where the specifications are not available (e.g., the Graphcore IPU GC200’s TDP). Specifications are unverifiable in some cases. Google does not publish specifications for its TPUs, as compared to NVIDIA or AMD (Introduction to Cloud TPU)(Unprecedented Acceleration at Every S...) (Morgan 2024). Google’s self-reported analysis claims that the TPU v4 is 1.3-1.9 times more energy efficient than the NVIDIA A100 (Jouppi et al. 2023). Using the NVIDIA’s published TDP for the A100, one can estimate that the TPU’s TDP is 280W. The lower multiplier (1.3x) is consistent with the typical range of other accelerators in the v1.1 Training benchmark submissions range that fall between 140-280W.

3.2 Resource Utilization Example

Thanks to the progress of chip manufacturing, AI accelerators have achieved Tera and Petaflop scale. However, this is “theoretical peak performance,” which refers to the physical limit that a device can achieve, provided that the application is fine-tuned to match the architecture of the chip. Though power consumption between accelerators and systems is useful for comparison’s sake, consumption is not constant but comes in bursts based on the operations the chip is handling. Furthermore, applications rarely take the full advantage of the full performance potential due to the mismatch between the applications’ workload characteristics and the underlying hardware. On the level of processing elements and memory, the growing gap between the speed of processing and data access has become the key factor impacting performance. Consequently, the cost of delivering data along with power consumption have become the targets of chip design optimization.

In order to gain some idea about the interaction between the application and the hardware, profiling information must be collected. There are many profiling tools that collect large volumes of detailed information. However, the increasing amount of data limits the duration of the monitoring period, and making sense of the fine grain details can also be a challenge. To resolve this dilemma, a simple sampling technique which reads the status of the computing device in regular intervals and summarizes the data by histograms provides

 DOI:

an easy visualization of power consumption. The amount of collected data is minimal, allowing long-term monitoring. The histograms provide a high-level view of resource utilization, power consumption and temperature variation. These histograms can be considered as a “fingerprint” which provides a quantitative measure of how well an application matches the underlying hardware. They are generated using the Cloudmesh library and integrated them in the CloudMask benchmark. In the example below (Figure 1), the CloudMask benchmark that uses image classification to identify clouds in satellite imagery is profiled.

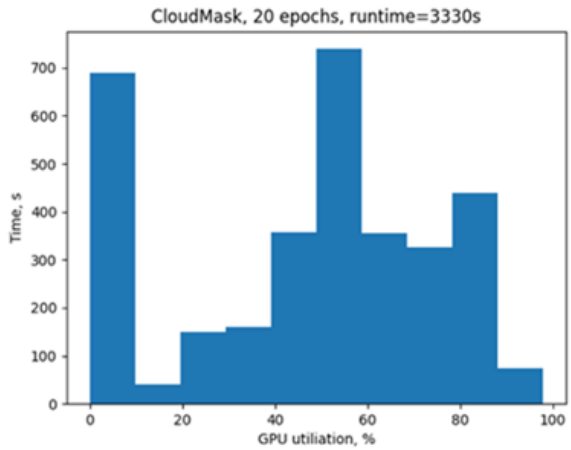

FIGURE 1. GPU utilization of CloudMask benchmark

Traditionally, runtime has been the main concern of performance studies, with power consumption historically being paid less attention. However, this attitude has changed recently, and it is no longer sufficient to run simulations or analyses quickly, but also with the smallest amount of power. The “power consumption fingerprint” of the CloudMask benchmark is illustrated in Figure 2.

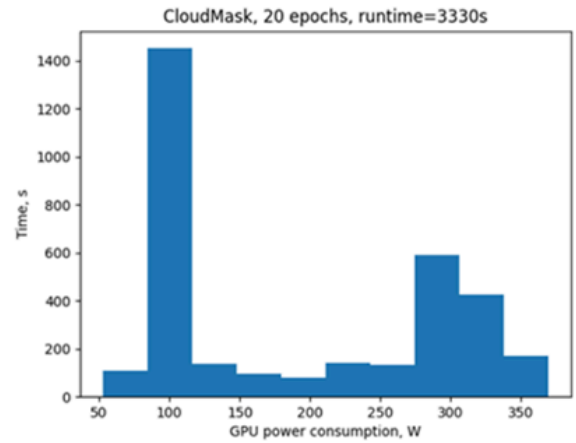

Figure 2. The power drawn by the CloudMask benchmark

Figure 2 illustrates differing utilization and power consumption of GPUs during the ML processing. Understanding GPU utilization of differing compute architectures and software stacks could also lead to more efficient use of GPUs.

4. Results and Analysis

With the above context in mind, a subset of MLCommons data was targeted for further analysis.

4.1 Accessing & Cleaning

The data was collected from multiple rounds of training benchmarks (v1.0, v1.1, and v3.0) with a focus on the ResNet benchmark because it had the most entries. ResNet refers to a residual neural network —a type of deep learning model— that was employed for image classification on the ImageNet dataset. The summary results table provided by MLCommons (MLCommons, n.d.c.) was amended with additional values and reviewed to match the benchmark results on GitHub. The metadata for the processor and accelerator TDP were added to the table where available (see Supplementary Materials).

4.2 Data Issues and Limitations

While the desire at the outset of this work was to use feature selection or similar ML techniques for further insight, the MLCommons ResNet benchmark results data is a relatively small dataset (36-80 submissions per round), and too small for ML inquiry. Within these small sets of records, the individual data quality is lacking and only a subset has the needed detail for further analysis.

4.3 Nonspecific Vocabulary & Quality

The MLCommons benchmark data's quality and specificity limits its use for further analysis. A Google submission lists the processor as "AMD ROME," of which there are two dozen types. In some cases, the processor can be deduced from other system attributes (e.g., the L2 cache). The system profiles contain typos and do not use a controlled vocabulary to constrain entries to valid software versions. Entries inaccurately document software versions. (e.g., "Dell 'Tensorflow 21.05'" which is a version of a specific NVIDIA container that includes Tensorflow) which limits reproducibility.

4.4 Data Validation and Provenance

Each submitted benchmark includes five runs, but uniform numbering is not enforced. Most submissions number these 1-5, but some label them 0-4. This discrepancy disrupts result comparisons between benchmark runs.

4.5 Too Much or Too Little Information

Different log verbosity setting choices can provide very different results and complicate analysis. For example, the submitted results ranged from text files with 100 lines to 30,000 lines, with 3,500 lines being most typical.

 DOI:

4.6 Opportunities for Improving MLCommons Benchmark Data Results

Several actions could be undertaken to improve the machine accessibility and ease analysis of the results (see Table 1).

FAIR Principle	Opportunities to implement FAIR
F1 (meta)data are assigned a globally unique and persistent identifier	Assign unique identifiers to results
A1 (meta)data are retrievable by their identifier using a standardized communications protocol **A1.1** the protocol is open, free, and universally implementable	Results accessible via open API
I2 (meta)data use vocabularies that follow FAIR principles	(FAIR) vocabularies used
R1.2 (meta)data are associated with their provenance	Provenance included in results and not only derived from directory structure

Table 1. Summary of MLCommons Data Improvements by FAIR Principle

- **Format results in a more reusable format.** The power and energy results could be made available in an extensible format, CVS or Google sheet, rather than an HTML table.
- **Extend Cloudmesh capabilities.** Analyzing the HTML summary results requires scraping the data, an easy way to introduce errors. The Cloudmesh tool allows for some of this capability and could be extended by MLCommons . The Cloudmesh stopwatch writes in parallel to mllog and can write out in any format needed (e.g., html, yaml, csv, json, txt, latex and mllog).
- **Use unique identifiers.** Assigning globally unique, persistent and resolvable identifiers (GUPRIs) to submissions, submitters, systems, and data used would aid in the accuracy of tracing results and analysis.
- **Output directory standard.** I
- **Create or extend controlled vocabularies and map to existing ontologies.** For example, identify or create a controlled vocabulary with a standardized listing of existing processors. An existing MLCommons working group could maintain the list and amend as new processors are released or used in benchmark submissions.
- **Include power consumption data.** This could take the form of the kW per node and the Power Usage Effectiveness (PUE) of the host site's data center.
- **Enforce the submission directory structure or replace with metadata.** The submitter's directory structure substitutes for explicit metadata (e.g. the specific run tied to a specific submission). Enforcing directory structures would make scripted analysis easier. An additional improvement would be to add the implied Information from the directory structure as actual metadata within the submitted files.

4.7 MLCommons Benchmark Results Data

Two areas became the focus for analysis: relative efficiency of the ResNet and BERT benchmarks, and comparison of results in minutes as compared to power consumption.

ResNet is a machine learning model for image classification. BERT is a machine learning model for natural language processing.

4.8 Relative Power Efficiency

The MLCommons Training and HPC-submitted results provide useful information on the compute performance as a function of system (accelerator) type and size.

ResNet	4	4	8	64	1024	3456	4096
A100	0.92	1	0.95	0.76	0.39		0.16
TPUv4						0.28	0.28
H100			1.86	1.52			

BERT on	4	4 SXM	8	64	1024	3456	4096
A100	0.72	1	0.99	0.84	0.31		0.16
TPUv4						0.17	0.18
H100			2.62	2.32			

Table 2. Scaled System Performance for the MLCommons 2022 Training Submissions for ResNet and BERT Benchmark

Table 2 presents submitted performance numbers as speedup efficiency defined as follows:

$$\frac{(Execution\ time\ 1)*(Number\ of\ nodes\ 1)}{(Execution\ time\ 2)*(Number\ of\ nodes\ 2)},$$

where reference System 1 is a 4-node A100 with SXM (a high bandwidth socket connecting the GPUs to the system) and System 2 runs over other submitted systems. The column labels the system characteristics, and the rows are different accelerators. In some cases, the number of nodes is different from the column value and then it is listed in parentheses in the cell.

The listed efficiency decreases as system size increases due to decreasing parallel efficiency (usually communication overhead). Numbers increase for H100 and TPUv4 accelerators as these are faster than the A100. The rapid efficiency decrease as the scale of the job and nodes used increases translates into greater power inefficiency. Further, given the size of these two datasets, one would likely run with larger system sizes - not realizing that the decreased execution time comes at a greater power cost. Results with a larger choice of system sizes could allow a more informed choice of maximum useful size with a criterion such as the efficiency >0.5. Larger models such as ChatGPT or Bard would run much better on large systems

 DOI:

4.9 Training v1.0 and v3.0 Analysis for ResNet Benchmarks.

The goal was to illustrate the relative tradeoffs between different benchmarking results and the electricity consumed. Submission results were plotted by time and energy consumed in kilowatts. Where TDP was reported as a range, the lowest TDP was used. While the actual energy consumption of the results varies widely and is unknown, these hypothetical consumption estimates are built from the most energy intensive hardware component consumption specifications and allow one to contemplate the consumption tradeoffs. Another way to put the consumption into perspective is to calculate the power cost. Electricity rates and price tiers vary by region and utility provider, and can be very complex. Even though the monetary cost of a benchmark proved compelling in some cases, it includes too many additional variables and sensitivities, especially for advanced computing centers that rely on their institutions to pay for power via overhead (indirect cost recovery) collected from contracts and grants. For these reasons, the focus of the analysis is on kWs consumed.

4.10 Outliers

The experiments that did not use accelerators nor other outliers made it difficult to see patterns in the majority of the data, but were also a source of important inquiry. Of the complete benchmarks where costs could be discerned, less than 10% of them consumed more than 20 kWs. For example, an Intel submission that used 8 Intel Xenon Platinum CPUs and no GPUs took ~16 hours and consumed 314 kW.

The highest consumption benchmark was the third fastest (NVIDIA: 629 kW for .4 minutes using 620 AMD EPYC 7742 and 2,400 NVIDIA A100s). By contrast, the second fastest time used 40% less energy: Google at 369 kW for .28 minutes using 1,024 AMD Rome CPUs and 2,048 TPUs. Using the same CPUs and TPUs, Google had the fastest benchmark (619 kW for .23 minutes using 1,728 AMD Rome CPUs and 3,456 TPUs). Reviewing Google's two best benchmarks illustrates that adding more computational power does not always scale linearly. Further, unless the use case requires speed at any cost, occupying the additional 1,408 TPUs (or 1.45 times more accelerators) for not much performance gain (.05 seconds or less than 20% faster) is not the best use of finite resources.

For this reason, we reran the analysis after excluding the outliers. This allowed for closer examination of the results with relative power consumption below 8kW. Plotting benchmarks in minutes-by-kilowatts-of-energy-consumed helps illustrate the inefficiency in time and electricity consumption for running ML image classification processes without GPUs or other accelerators. Our experiments uncovered that benchmarks without GPUs take three times as long to run and consume and are at the high end for electricity consumption across all submissions. In other words, ML computation without accelerators occupies nodes better utilized for other tasks. The slowest GPU configuration from Fujitsu was almost four times as fast as the quickest CPU-only benchmark result from Intel.

The same analysis was conducted with the Training v3.0 benchmarks — once again, excluding outliers. The two fastest benchmarks were from Google, with the fastest (at .23 minutes) consuming 1.67 times the electricity of the runner-up (at .28 minutes. Another fast NVIDIA benchmark at 4.91 minutes and 18.95 kW using NVIDIA A100 GPUs outperformed the Dell submission using NVIDIA A100s with half the cache at 15.5 minutes and only 6.6 kW.

Our analysis suggested a relative lower electricity consumption of the NVIDIA A100-PCIE-40GB accelerator in concert with PyTorch and Tensorflow optimized for this particular GPU than compared with other hardware and software combinations.

Information about energy consumption over speed can be valuable for different AI use cases and contexts. In healthcare, air traffic control, or military applications, speed may be the priority at any cost. However, many research applications may have no perceived benefit to quicker processing. Where consumers bear the costs, they may opt for the cheapest utility costs.

Some of the lowest energy consumption benchmark results don't use novel hardware configurations. For example, in Training v3.0 on ResNet, NVIDIA has one of the fastest results, but for the highest power consumption: 0.369 mins for 632 kW. When a modest amount of GPUs is used (8 vs. 768 H100s) and CPUS (2 vs. 192 Xeon Platinum 8480C), it is both fast and energy efficient: 13.601 minutes for 8.4 kW. For ~96 times the GPUs, the NVIDIA result is only ~37 times as fast. In other words, the benchmark improvement is not linear given the additional resources used. This can be due to a variety of factors including how many floating point operations are invoked, data copy times and the type of interconnect used.

4.11 Energy Sample Tracing with the Earthquake Science Application Using Cloudmesh

Comparing computation time with energy consumption is useful for the reasons illuminated above, but it does not give the complete picture of electricity usage. During computation, energy fluctuates, sometimes significantly. To understand this, we examined one of the key MLCommons Science benchmarks: Cloudmesh.

- Cloudmesh was developed with the purpose in mind to ease the capturing of energy traces via sampling from applications obtained with various hyperparameters.. Cloudmesh in this example utilizes the undelaying monitoring capabilities of NVIDIA GPUs. However it would be easily possible to expand it to also be used on AMD GPUs as well as the integration of various other libraries such as PAPI (PAPI)and the framework could be expanded to use other energy parameters. However, for this paper we focused on the development of Cloudmesh the capabilities for the purpose and in support of MLCommons science. It includesThe ability to augment a code with stopwatch timers and events that are available in human readable format and in mllog log files.
- A simple-to-use logging program that can be run parallel to the execution of the main program to be benchmarked, creating logging events at predefined times. The logging program utilizes the built-in NVIDIA logging capabilities. We augmented the program to generate visualizations from the logs, including log traces for earthquake prediction (see Figure 3) and histograms (see Figures 1-2).
- A framework to create workflows that allows for them to be applied to various HPC systems with a variety of GPUs. This enables comparisons between said workflows while guaranteeing the reproducibility of the experiments by considering variations in hardware configurations.

DOI:

- The storage of energy traces that are in a separate file in various formats and can be submitted as part of the benchmark submission process. If needed, an mllog based format could also be created, but the current Cloudmesh common format is useful as each energy event is much smaller than events created by mllog. This allows easy parsing as well as the support of human readable formats.

Using these capabilities, we uncovered that for the Earthquake application the power draws based on its algorithmic phases a wide variety of energy. While data preparation and post analysis although time consuming only need between 50-70W including the usage of GPUs, the vacillates of the core Deep Learning between 70W and 200W (see Figure 3). With such augmentation, we also discovered significant file system performance issues in the used HPC center that ultimately were fixed by the replacement of the data storage facility addressing the performance issues (von Laszewski et al. 2023).

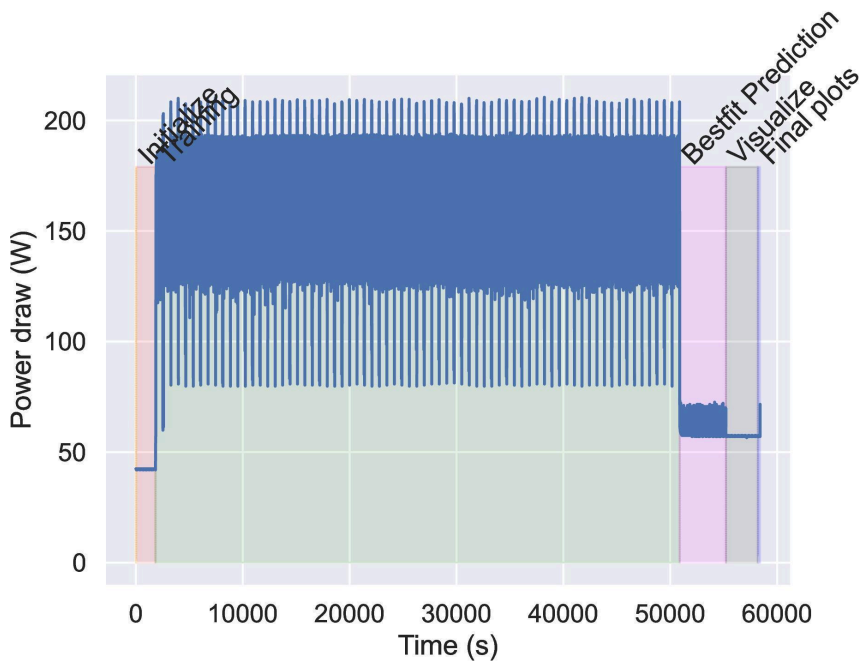

Figure 3. An energy trace for the MLCommons Earthquake forecasting application. Energy traces can be helpful in identifying different phases of the algorithms. Here, the phases to conduct initialization are distinguished: training, validation (e.g., best fit prediction), and visualization of the results. Such phases can then be further analyzed in order to optimize access to filesystem or network resources as they may project a bottleneck in the available hardware and result in not optimized performance.

5. Conclusions

The MLCommons benchmark results could be enhanced for machine actionability and made more analysis-ready through the application of the FAIR Principles. Future work may include ways to constrain input of the system profiles, with potential use of defined ontologies for interoperability of log data. Further work is needed to improve data management practices of benchmark data —such as through the FAIR Principles— and to understand practical ways to incorporate power measurements into benchmark results. Tools

—such as the capabilities developed for Cloudmesh— help to quickly analyze energy demands. Through system profiling, one can uncover issues of scaling due to interconnect or programing constraints. Even with the current limitations of the benchmark results data, they provide insight for choosing hardware, software, and configurations, and lead to informed decisions about speed and power consumption tradeoffs.

The results of ML benchmarks are a trove of information for optimizing resource utilization and costs. This is especially true in academic research and any sector with limited resources or access. Even where funds are awarded for procurement, electricity and water consumption may be constrained in many instances. Cloud and HPC ML loads alike can be optimized through the application and analysis of HPC-style benchmarks. This can contribute to resource optimization to maximize science production.

Acknowledgement

The work of C.R. Kirkpatrick was supported in part by the National Science Foundation under Grants 1916481 and 2226453.

The work from Gregor von Laszewski and Geoffrey Fox was supported in part by DE-SC0023452: FAIR Surrogate Benchmarks Supporting AI and Simulation Research, NSF 2346173 POSE: Phase II: MLCommons Research for Science: Enabling Open-Source Ecosystems for Scientific Foundation Models by Community Standards and Benchmarks, and NSF 2411009 Elements: A Sustainable, Resource-Efficient Cyberinfrastructure for Notebook Interactive ML Training Workloads.

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "AI for Science" theme within that grant, by the Alan Turing Institute, and by the Benchmarking for AI for Science at Exascale (BASE) project under the EPSRC Grant EP/V001310/1.

We like to thank Ismael Kherroubi Garcia for help with editing this paper.

References

** DONE **
Boehme D., Gamblin, T., Beckingsale, D., Bremer, P.T., Gimenez, A., LeGendre, M., Pearce, O., and Schulz, M., "SC'16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis," in *2016 IEEE Supercomputing Conference* in Caliper: Performance Introspection for HPC Software Stacks, Salt Lake City, UT, USA, 2016, pp. 550-560, online: https://dl.acm.org/doi/10.5555/3014904.3014967 [accessed December 08, 2024]

** DONE ** De Gelas, J., "Testing the latest x86 rack servers and low power server CPUs," *AnandTech*, July 22 2009, online: https://www.anandtech.com/show/2807/ [accessed December 12, 2024]

** DONE ** Ganapathy D., and Warner, E.J., "Defining thermal design power based on real-world usage models," in *2008 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, Orlando, FL, USA, 2008, pp. 1242-1246, doi: 10.1109/ITHERM.2008.4544402.

** DONE ** Goldman Sachs, “AI is poised to drive 160% increase in data center power demand,” *Goldman Sachs Insights*, May 14 2024, online: https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand [accessed December 09, 2024]

** DONE ** Google Cloud, “ Introduction to Cloud TPU,” *Google Cloud*, Last updated 2024-12-04 UTC, online: https://cloud.google.com/tpu/docs/intro-to-tpu [accessed December 12, 2024]

** ADD TO PAPER PILE BY HAND ** Huck, S., “White Paper Measuring Processor Power: TDP vs. ACP,” *Intel Corporation*, April 2011, online: https://www.intel.com/content/dam/doc/white-paper/resources-xeon-measuring-processor-power-paper.pdf [accessed December 12, 2024]

** DONE ** Jouppi, N.P., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B., and Young, C., “Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings,” *Cornell Library*, Apr. 2023. doi: 10.48550/arXiv.2304.01433.

MLCommons. “An Open AI Engineering Consortium.” *MLCommons*, n.d.a, online: https://mlcommons.org/about-us/ [accessed December 09, 2024]

** DONE ** MLCommons. “README.md,” *GitHub*, n.d.b, online: https://github.com/mlcommons/ [accessed December 09, 2024]

MLCommons, “MLPerf Training benchmark suite results,” *MLCommons*, n.d.c, online: https://mlcommons.org/benchmarks/training/ [accessed December 08 2024]"

** DONE** NVIDIA, “NVIDIA A100 TENSOR CORE GPU,” *NVIDIA Corporation*, June 2021, online: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf [accessed December 12, 2024]

** DONE ** PAPI, Web Page, https://icl.utk.edu/papi/, 2024.

** DONE ** Petkov, V., “Automatic performance engineering workflows for high performance computing,” Ph.D. dissertation, Technische Universität München, München, 2014, online: https://mediatum.ub.tum.de/doc/1174068/1174068.pdf [accessed December 08, 2024]

** DONE** Prickett Morgan, T., “Lots of questions on Google’s ‘Trillium’ TPU v6, a few answers,” *The Next Platform*, June 10 2024, online: https://www.nextplatform.com/2024/06/10/lots-of-questions-on-googles-trillium-tpu-v6-a-few-answers/ [accessed December 12, 2024]

** DONE ** Prisco, J., Stewart, G., Huber, H., Rannow, R., Hick, J., Martinez, D., Hong, B., and Deshpande, A.M., “2020 IEEE International Conference on Cluster Computing (CLUSTER),” in *2020 IEEE Cluster Conference* in Investigative Report on Electrical Commissioning in HPC Data Centers, Sept. 2020, pp. 519-522, doi: 10.1109/CLUSTER49012.2020.00074.

Rozite, V., Bertoli, E, and Reidenbach, B., “Data Centres and Data Transmission Networks.” *International Energy Agency*, 11 July 2023, online: www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks [accessed December 09, 2024]

Tate, B., Based on interviews with Thomas Tate, Brian Balderston San Diego Supercomputer Center, 2022, 2024.

Thiyagalingam, J. von Laszewski, G., Yin, J., Emani, M., Papay, J., Barrett, G., Luszczek, P., Tsaris, A. Kirkpatrick, C., Wang, F., Gibbs, T., Viswananath, V., Shankar, M., Fox, G. C., and Hey, T. “AI Benchmarking for Science: Efforts from the MLCommons Science Working Group,” in *ISC High Performance 2022 International Workshops*, Hamburg, Germany, 2022, pp. 47-64, doi: 10.1007/978-3-031-23220-6_4.

*** DONE **** TOP500. “GREEN500.” *TOP500*, June 2024, online: top500.org/lists/green500/2024/06/ [accessed December 09, 2024]

** DONE ** von Laszewski, G., Fleischer, J.P., and Fox, G. C., “Hybrid Reusable Computational Analytics Workflow Management with Cloudmesh,” arXiv, cs.DC, rCornell Library, Oct. 2022. doi: 10.48550/arXiv.2210.16941, https://arxiv.org/abs/2210.16941.

** DONE ** von Laszewski G. , Fleischer J. P. , Knuuti Robert , Fox Geoffrey C. , Kolessar Jake , Butler Thomas S. , Fox Judy, Opportunities for enhancing MLCommons efforts while leveraging insights from educational MLCommons earthquake benchmarks efforts, Frontiers in High Performance Computing, Volume 1, 2023, https://www.frontiersin.org/journals/high-performance-computing/articles/10.3389/fhpcp.2023.1233877, DOI 10.3389/fhpcp.2023.1233877, ISSN 2813-7337.

** DONE ** von Laszewski, G., Fleischer, J. P., Fox, G. C. , Papay, J., Jackson, S., and Thiyagalingam, J., "Templated Hybrid Reusable Computational Analytics Workflow Management with Cloudmesh, Applied to the Deep Learning MLCommons Cloudmask Application," 2023 IEEE 19th International Conference on e-Science (e-Science), Limassol, Cyprus, 2023, pp. 1-6,

doi:10.1109/e-Science58273.2023.10254942,https://ieeexplore.ieee.org/document/10254942

Author Bios

Christine R. Kirkpatrick is at the San Diego Supercomputer Center, UC San Diego where she heads the Research Data Services division. Kirkpatrick is a Ph.D. Candidate at the University of Porto's Computer Science Department and a member of the IEEE Computer Society. Contact her at christine@sdsc.edu.

Gregg Barrett is the CEO of Cirrus AI, Johannesburg, South Africa. Contact him at gregg.barrett@cirrusai.net.

Wesley Brewer is a senior research scientist in HPC and AI at Oak Ridge National Laboratory. His current research interests include topics in AI for Science and Digital Twins. He obtained a Ph.D. in Computational Engineering from Mississippi State University. He is a member of ACM SIGHPC. Contact him at brewerwh@ornl.gov.

Julie Christopher is a technical project manager at the San Diego Supercomputer Center at the University of California San Diego. Contact her at jchristopher@sdsc.edu.

Inês Dutra is a lecturer in the Department of Computer Science, School of Sciences University of Porto, Portugal. She obtained a M.Sc. degree in Systems Engineering and Computer Science from Federal University of Rio de Janeiro and a Ph.D. in Computer Science from Bristol University, UK. Contact her at ines.dutra@gmail.com

Murali Emani is a computer scientist at Argonne National Laboratory. His current research interests are in scalable machine learning, emerging HPC and AI hardware, and benchmarking. Murali received the Ph.D. degree in Informatics from The University of Edinburgh, UK. He is a member of the IEEE Computer Society and ACM. Contact him at memani@anl.gov.

Piotr Luszczek is a Research Associate Professor at the Electrical Engineering and Computer Science Department of the University of Tennessee in Knoxville. He received his Ph.D. for novel research into numerical methods for achieving portable performance to address the needs of both dense and sparse computations with direct and iterative approaches in mind. He is a member of ACM, IEEE Computer Society, and SIAM. Contact him at luszczek@icl.utk.edu.

Juri Papay is a Senior Research Engineer with the IT Innovation Centre, University of Southampton, UK. He received his Ph.D. in computer science from the University of Warwick.. Contact him at juri.papay@stfc.ac.uk.

Mallikarjun (Arjun) Shankar is the acting Division Director of the National Center for Computational Sciences at the Oak Ridge National Laboratory. He received his Ph.D. in Computer Science from the University of Illinois, Urbana-Champaign. He is a joint appointee at the University of Tennessee's Bredesen Center, a senior member of the IEEE and a senior member of the ACM.

Gregor von Laszewski is a Research Professor at the Biocomplexity Institute and Initiative at the University of Virginia. He obtained a Ph.D. in Computer Science at Syracuse University. He has long-standing experiences in High-Performance Computing, Cloud Computing, and Grid Computing. Contact him at laszewski@gmail.com.

Geoffrey Fox obtained a Ph.D. in Theoretical Physics from Cambridge University, where he was Senior Wrangler. He is now a Professor at the Biocomplexity Institute & Initiative and Computer Science Department at the University of Virginia. He is a Fellow of APS and ACM and works on the interdisciplinary interface between computing and applications; currently, AI for science. Contact him at gcfexchange@gmail.com.

DOI: