# AI for Performance Engineering and Performance Engineering for AI

Lizy K. John
ljohn@ece.utexas.edu
The University of Texas at Austin
Austin, Texas, USA

## Abstract

Artificial Intelligence (AI) and Machine Learning (ML) techniques are revolutionizing various domains, including performance engineering. Performance engineering, which involves the evaluation, modeling, and optimization of system performance, has traditionally relied on established methodologies that have proven effective over the years. However, the growing complexity and heterogeneity of modern computing systems, particularly with the emergence of AI accelerators, has resulted in a shift in approach. AI and ML techniques are now being leveraged to achieve unprecedented levels of efficiency and scalability in performance engineering. Similarly, performance engineering is modifying metrics and methodologies for ML benchmarking. This talk will describe some opportunities and challenges **when AI meets Performance Engineering.**

AI/ML can address challenges in performance engineering by learning complex system behaviors from vast amounts of data, enabling adaptive and predictive performance models. One of the key advantages of using AI/ML in performance engineering is its ability to identify performance bottlenecks and predict system behavior under varying workloads. Machine learning models can analyze performance metrics in real time, allowing for automated tuning and optimization. This capability is particularly useful in cloud computing environments, where dynamic resource allocation is crucial for maintaining efficiency and cost-effectiveness. Moreover, AI-driven approaches can facilitate workload characterization and anomaly detection. By training models on historical data, AI systems can detect deviations from normal performance patterns, identifying potential issues before they impact system stability. This proactive approach to performance engineering reduces downtime and enhances overall system reliability.

AI and ML can also be used to create performance models, at levels ranging from circuit design to system level performance. An example from our prior work [5] used constrained lasso regression to estimate cross-platform microprocessor performance and power from performance counters. Essentially benchmark programs are run on one machine and processor performance monitoring counter outputs gathered to predict performance and energy on a different processor. Another effort [4] utilizes training on one platform followed by fine-tuning to predict cross-platform energy/power on Field Programmable Gate Arrays (FPGAs).


*ICPE '25, May 5–9, 2025, Toronto, ON, Canada*

ACM ISBN 979-8-4007-1073-5/2025/05
https://doi.org/10.1145/3676151.3720528

AI/ML is helping performance engineering; similarly performance engineering has to help design of AI/ML platforms. Performance evaluation and benchmarking of ML systems [2] is challenging, especially due to the complexity and heterogeneity of modern AI workloads and accelerators. AI workloads running on specialized hardware, such as Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), Neural Processing Units (NPUs) and Field-Programmable Gate Arrays (FPGAs) with dozens of software layers in the thick software stack bring many challenges to effective evaluation and benchmarking.

Performance metrics and methodologies must be adapted to AI-driven workloads and platforms. Metrics for benchmarking ML training will be different from inference benchmarking. Training could use metrics such as **Time To Accuracy (TTA)** which considers the time to train a model to a predetermined accuracy [1, 3]. Metrics such as time (latency) for inference in batch size of one may be most appropriate for edge devices, whereas average latency for a larger batch may be appropriate for cloud inference engines. With the proliferation of Generative AI and language models, some say tokens per second is the relevant metric.

Many choices exist in the software stack about what ML frameworks (eg: PyTorch, TensorFlow, etc.), what graph formats (ONNX, NNEF), what graph compilers (Glow, TVM, XLA etc.), what libraries (CuBLAS, MKL DNN, etc.), operating systems (Linux, RTOS, MacOS, Android, etc.), hardware targets (CPUs, GPUs, TPUs, NPUs, FPGAs, accelerators, etc.) should be used. The diversity of options at every level of the stack make benchmarking inference systems very challenging. The sensitivity of performance to libraries, formats, frameworks, etc. is very high, and needs to be studied. It is questionable whether an AI/ML model must be the cutting-edge model or a simpler model that many platforms can run. It is questionable whether very high accuracy thresholds should be chosen in the TTA metrics or a moderate accuracy achievable by many platforms and vendors must be chosen. ML datasets need to be predetermined for fair apple to apple comparison, and test sets identified.

The MLPerf benchmarks [2] are separated into categories of training and inference. Furthermore there are four categories of inference: cloud, edge, mobile, and tiny, as requirements and constraints are very different in cloud and edge inferencing. Results may be submitted in the closed division where the specified model must be used or in the open division, where the model can be changed.

While AI brings numerous opportunities to enhance performance modeling and evaluation, it also presents significant challenges that must be addressed. One challenge lies in the interpretability of AI models. While AI can provide powerful insights into system performance, its decision-making processes are often

considered black-box in nature. Understanding why a model predicts certain performance trends requires explainable AI techniques, which are still an active area of research. Without transparency, it becomes difficult to trust and validate AI-driven optimizations. Additionally, the reliance on data-driven methods introduces concerns regarding data quality and bias. AI models are only as good as the data they are trained on, and inaccuracies or biases in training data can lead to misleading performance predictions. Ensuring the integrity and representativeness of training datasets is crucial for the reliability of AI-enhanced performance engineering solutions.

Creating reliable datasets for ML-based performance modeling is another challenge. The *Imagenets* for Performance Engineering are yet to be constructed and curated.

The intersection of AI/ML and performance engineering marks a transformative shift in how system performance is evaluated and optimized. AI-driven approaches offer new levels of efficiency, scalability, and adaptability, making them invaluable in modern computing environments. However, the challenges of heterogeneous hardware, model interpretability, and data quality must be carefully managed to fully harness the potential of AI in performance engineering. By addressing these challenges and exploring innovative solutions, the future of performance engineering can be significantly enhanced through AI and ML technologies. This talk will explore the many ways AI/ML can influence performance engineering, and vice versa, when AI meets performance engineering.

## CCS Concepts

• **General and reference** → **Evaluation**; **Performance**; **Measurement**; **Experimentation**; • **Computing methodologies** → *Machine learning*; *Natural language generation.*

## Keywords

Performance Modeling, ML Benchmarks, Time-To-Accuracy

**ACM Reference Format:**
Lizy K. John. 2025. AI for Performance Engineering and Performance Engineering for AI. In *Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering (ICPE '25), May 5–9, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3676151.3720528

## Biography

**Lizy Kurian John** is Professor and Truchard Foundation Chair in Engineering at the University of Texas at Austin. Her research interests include workload characterization, performance evaluation, and high performance architectures for emerging workloads. She is recipient of many awards including Joe J. King Professional Engineering Achievement Award (2023), and The Pennsylvania State University Outstanding Engineering Alumnus Award (2011). She has authored 3 books and has edited 4 books including a book on Computer Performance Evaluation and Benchmarking. She holds 18 US patents and is an IEEE Fellow (Class of 2009), ACM Fellow, AAAS Fellow and Fellow of the National Academy of Inventors (NAI).

## Acknowledgments

Lizy K. John's research is supported in part by the United States National Science Foundation Grants #2326894, #2425655, the Semiconductor Research Corporation (SRC) Task 3148.001, Texas Advanced Computing Center (TACC) and NVIDIA Applied Research Accelerator Program Grant.

## References

[1] C. Coleman, D. Kang, D. Narayanan, L. Nardi, T. Zhao, J. Zhang, P. Bailis, K. Olukotun, Chris Ré, and M. Zaharia. 2019. Analysis of DAWNBench, a Time-to-Accuracy Machine Learning Performance Benchmark. *SIGOPS Oper. Syst. Rev.* 53, 1 (July 2019), 14–25. https://doi.org/10.1145/3352020.3352024

[2] MLCommons Association. 2025. MLCommons Benchmarks. https://mlcommons.org/benchmarks/

[3] S. Verma, Q. Wu, B. Hanindhito, G. Jha, R. Radhakrishnan, and L. K. John. 2019. Metrics for Machine Learning Workload Benchmarking. *FastPath in conjunction with ISPASS* 2019, 1 (March 2019). https://web.archive.org/web/20210807053037/https://researcher.watson.ibm.com/researcher/files/us-ealtman/Snehil_Metrics_for_Machine_Learning_Workload_Benchmarking.pdf

[4] Zhigang Wei, Aman Arora, Emily Shriver, and Lizy Kurian John. 2024. Efficient FPGA-based power model adaption with Transfer-Learning and Meta-Learning. *COGARCH Workshop, in Conjunction with The 51st International Symposium on Computer Architecture (ISCA 2024)* 2024, 1 (2024).

[5] X. Zheng, L. K. John, and A. Gerstlauer. 2016. Accurate phase-level cross-platform power and performance estimation. In *DAC '16* (Austin, Texas) *(DAC '16)*. ACM, Article 4, 6 pages. https://doi.org/10.1145/2897937.2897977