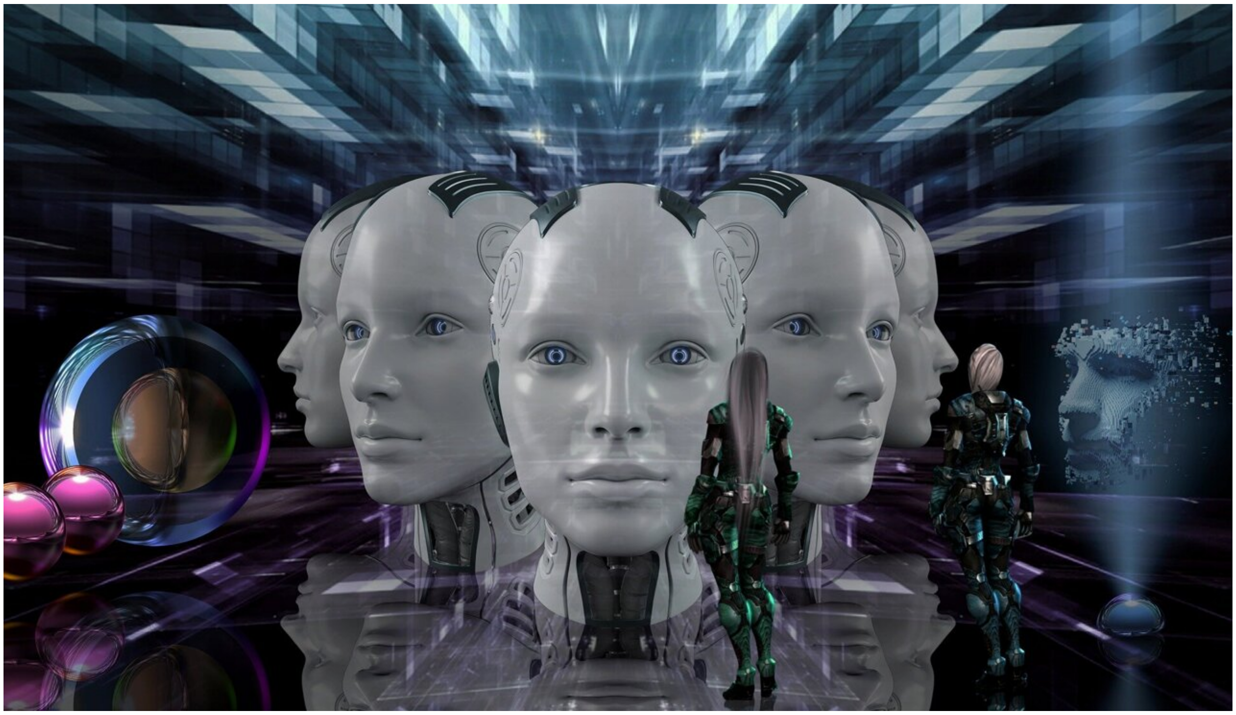
TechXplore

Supercomputing center dataset aims to accelerate AI research into optimizing high-performance computing systems

August 24 2022, by Kylie Foy



Credit: Pixabay/CC0 Public Domain

When the MIT Lincoln Laboratory Supercomputing Center (LLSC) unveiled its TX-GAIA supercomputer in 2019, it provided the MIT community a powerful new resource for applying artificial intelligence to their research. Anyone at MIT can submit a job to the system, which churns through trillions of operations per second to train models for diverse applications, such as spotting tumors in medical images, discovering new drugs, or modeling climate effects. But with this great power comes the great responsibility of managing and operating it in a sustainable manner—and the team is looking for ways to improve.

"We have these powerful computational tools that let researchers build intricate models to solve problems, but they can essentially be used as black boxes. What gets lost in there is whether we are actually using the hardware as effectively as we can," says Siddharth Samsi, a research scientist in the LLSC.

To gain insight into this challenge, the LLSC has been collecting detailed data on TX-GAIA usage over the past year. More than a million user jobs later, the team has released the dataset open source to the computing community.

Their goal is to empower computer scientists and data center operators to better understand avenues for data center optimization—an important task as processing needs continue to grow. They also see potential for leveraging AI in the data center itself, by using the data to develop models for predicting failure points, optimizing job scheduling, and improving energy efficiency. While cloud providers are actively working on optimizing their data centers, they do not often make their data or models available for the broader high-performance computing (HPC) community to leverage. The release of this dataset and associated code seeks to fill this space.

"Data centers are changing. We have an explosion of hardware platforms, the types of workloads are evolving, and the types of people who are using data centers is changing," says Vijay Gadepally, a senior researcher at the LLSC. "Until now, there hasn't been a great way to analyze the impact to data centers. We see this research and dataset as a big step toward coming up with a principled approach to understanding how these variables interact with each other and then applying AI for insights and improvements."

Papers describing the dataset and potential applications have been accepted to a number of venues, including the IEEE International Symposium on High-Performance Computer Architecture, the IEEE International Parallel and Distributed Processing Symposium, the Annual Conference of the North American Chapter of the Association for Computational Linguistics, the IEEE High-Performance and Embedded Computing Conference, and International Conference for High Performance Computing, Networking, Storage and Analysis.

Workload classification

Among the world's TOP500 supercomputers, TX-GAIA combines traditional computing hardware (central processing units, or CPUs) with nearly 900 graphics processing unit (GPU) accelerators. These NVIDIA GPUs are specialized for deep learning, the class of AI that has given rise to speech recognition and computer vision.

The dataset covers CPU, GPU, and memory usage by job; scheduling logs; and physical monitoring data. Compared to similar datasets, such as those from Google and Microsoft, the LLSC dataset offers "labeled data, a variety of known AI workloads, and more detailed time series data compared with prior datasets. To our knowledge, it's one of the most comprehensive and fine-grained datasets available," Gadepally says.

Notably, the team collected time-series data at an unprecedented level of detail: 100-millisecond intervals on every GPU and 10-second intervals on every CPU, as the machines processed more than 3,000 known deep-learning jobs. One of the first goals is to use this labeled dataset to characterize the workloads that different types of deep-learning jobs place on the system. This process would extract features that reveal differences in how the hardware processes natural language models versus image classification or materials design models, for example.

The team has now launched the MIT Datacenter Challenge to mobilize this research. The challenge invites researchers to use AI techniques to identify with 95 percent accuracy the type of job that was run, using their labeled time-series data as ground truth.

Such insights could enable data centers to better match a user's job request with the hardware best suited for it, potentially conserving energy and improving system performance. Classifying workloads could also allow operators to quickly notice discrepancies resulting from hardware failures, inefficient data access patterns, or unauthorized usage.

Too many choices

Today, the LLSC offers tools that let users submit their job and select the processors they want to use, "but it's a lot of guesswork on the part of users," Samsi says. "Somebody might want to use the latest GPU, but maybe their computation doesn't actually need it and they could get just as impressive results on CPUs, or lower-powered machines."

Professor Devesh Tiwari at Northeastern University is working with the LLSC team to develop techniques that can help users match their workloads to appropriate hardware. Tiwari explains that the emergence of different types of AI accelerators, GPUs, and CPUs has left users suffering from too many choices. Without the right tools to take advantage of this heterogeneity, they are missing out on the benefits: better performance, lower costs, and greater productivity.

"We are fixing this very capability gap—making users more productive and helping users do science better and faster without worrying about managing heterogeneous hardware," says Tiwari. "My Ph.D. student, Baolin Li, is building new capabilities and tools to help HPC users leverage heterogeneity near-optimally without user intervention, using techniques grounded in Bayesian optimization and other learning-based optimization methods. But, this is just the beginning. We are looking into ways to introduce heterogeneity in our data centers in a principled approach to help our users achieve the maximum advantage of heterogeneity autonomously and cost-effectively."

Workload classification is the first of many problems to be posed through the Datacenter Challenge. Others include developing AI techniques to predict job failures, conserve energy, or create job scheduling approaches that improve data center cooling efficiencies.

Energy conservation

To mobilize research into greener computing, the team is also planning to release an environmental dataset of TX-GAIA operations, containing rack temperature, power consumption, and other relevant data.

According to the researchers, huge opportunities exist to improve the power efficiency of HPC systems being used for AI processing. As one example, recent work in the LLSC determined that simple hardware tuning, such as limiting the amount of power an individual GPU can draw, could reduce the energy cost of training an AI model by 20 percent, with only modest increases in computing time. "This reduction translates to approximately an entire week's worth of household energy for a mere three-hour time increase," Gadepally says.

They have also been developing techniques to predict model accuracy, so that users can quickly terminate experiments that are unlikely to yield meaningful results, saving energy. The Datacenter Challenge will share relevant data to enable researchers to explore other opportunities to conserve energy.

The team expects that lessons learned from this research can be applied to the thousands of data centers operated by the U.S. Department of Defense.

Other collaborators include researchers at MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). Professor Charles Leiserson's Supertech Research Group is investigating performance-enhancing techniques for parallel computing, and research scientist Neil Thompson is designing studies on ways to nudge data center users toward climate-friendly behavior.

Samsi presented this work at the inaugural AI for Datacenter Optimization (ADOPT'22) workshop last spring as part of the IEEE International Parallel and Distributed Processing Symposium. The workshop officially introduced their Datacenter Challenge to the HPC community.

"We hope this research will allow us and others who run supercomputing centers to be more responsive to user needs while also reducing the energy consumption at the center level," Samsi says.

More information: Baolin Li et al, AI-Enabling Workloads on Large-Scale GPU-Accelerated System: Characterization, Opportunities, and Implications, *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (2022). DOI: 10.1109/HPCA53966.2022.00093

Nathan C. Frey et al, Energy-aware neural architecture selection and hyperparameter optimization, *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (2022). DOI:

10.1109/IPDPSW55747.2022.00125

Dan Zhao et al, A Green(er) World for A.I., *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (2022). DOI: 10.1109/IPDPSW55747.2022.00126

Dan Zhao et al, Loss Curve Approximations for Fast Neural Architecture Ranking & Training Elasticity Estimation, *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (2022). DOI: 10.1109/IPDPSW55747.2022.00123

The MIT Supercloud Dataset

The MIT Supercloud Workload Classification Challenge

Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models

Baolin Li et al, RIBBON, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2021). DOI: 10.1145/3458817.3476168

*This story is republished courtesy of MIT News (*web.mit.edu/newsoffice/*), a popular site that covers news about MIT research, innovation and teaching.*

Citation: Supercomputing center dataset aims to accelerate AI research into optimizing high-performance computing systems (2022, August 24) retrieved 9 May 2024 from https://techxplore.com/news/2022-08-supercomputing-center-dataset-aims-ai.html