Strengthening the Adoption of AI in Research and Cyberinfrastructure

Paola A. Buitrago
paola@psc.edu
Pittsburgh Supercomputing Center
Carnegie Mellon University
Pittsburgh, Pennsylvania

Nicholas A. Nystrom
nystrom@psc.edu
Pittsburgh Supercomputing Center
Carnegie Mellon University
Pittsburgh, Pennsylvania

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have vast potential to advance research, and their increasing effectiveness will both shape and enhance cyberinfrastructure (CI) to enable the breakthroughs of the future. The rapidly expanding array of algorithms and technologies is a transformation with potential exceeding that of other architectural advances over the past few decades -– for example, simulations are being accelerated by factors of up to two billion [3], vastly outpacing what has been possible through Dennard scaling – and it is also the greatest fundamental shift in how researchers engage in computational science and engineering.

We face tremendous opportunities and exciting challenges in helping the research community realize the benefits of AI and ML. Hardware resources must continue to integrate high performance computing (HPC) and scalable AI resources effectively into heterogeneous systems, including a range of special-purpose AI architectures that will be superior for specific workloads, and provide user-friendly software environments to make the resources accessible to domain specialists (i.e., not only traditional HPC users). User training of many kinds is, and will continue to be, absolutely essential. Equally exciting is the prospect of applying AI and ML to the operation of cyberinfrastructure, or *Smart CI*, which has begun yet has considerable untapped potential. The increasing complexity of heterogeneous CI and the extremely diverse workloads that are executed on it make Smart CI vitally important to maximize the resources' potential and users' productivity.

1 OPEN RESEARCH CHALLENGES AND PROMISING RESEARCH DIRECTIONS

Fundamental AI research is flourishing [6] and beyond the scope of this white paper. Instead, we focus on three specific, promising areas of research that are of specific relevance to NSF-supported cyberinfrastructure: 1) scalable AI for scientific data; 2) AI-accelerated simulation; and 3) AI for system operations, *AIOps*. For all three, there is an opportunity to identify use cases that go beyond commercial ones that will inform selection of appropriate hardware technologies for production research platforms.

Scalable AI for Scientific Data: Research is needed to provide more robust tools for data acquisition and preparation, to increase the scalability of deep learning training (which will directly benefit use of generative adversarial networks (GANs) and deep reinforcement learning, methods that are being used to address challenges involving limited or sensitive data), to work with multimodal data, and in many other areas. Applications are analyses of very large datasets from instruments or simulations, identifying rare events, combining different kinds of data to allow answering new questions, facilitating the discover and reuse of scientific data, and helping to find relevant insights from scientific literature.

AI-Accelerated Simulation: Research is needed to improve the development of surrogate models for accelerating and potentially improving the accuracy of simulations, including scalable online training (i.e., incorporating new data into a model on an ongoing basis) for use with ensembles of simulations and data assimilation. Published results already demonstrate up to nine orders of magnitude acceleration [3] with no loss of accuracy [7], and impact on other applications is probable given sufficient human resources and user training. Applications are widespread, including both nontraditional fields (genomics, protein signaling pathways, 3D image processing, etc.) and traditional HPC (weather, chemistry, astrophysics, engineering, etc.).

AIOps: There are many promising directions, limited primarily by human resources, to improve CI through AIOps. Applying AIOps to system instrumentation data can improve system operation, reliability, and performance by identifying degrading components for proactive repair, spotting issues that negatively impact system performance, and optimizing queue structure. AIOps can improve users' experience, particularly for nontraditional users who are unaccustomed to advanced CI environments, as is being addressed by PSC's internal project *Calima*, led by Buitrago, for large systems such as *Bridges* [5], *Bridges-AI* [2], and *Bridges-2*. AIOps can also help with resource requests by helping to match advanced, heterogeneous computational resources to project requirements, helping to match reviewers to resource proposals, and then assessing whether the expected resource utilization is achieved.

2 TESTBEDS

Innovative hardware designed for specific aspects of AI and ML is being developed at an unprecedented pace. For example, emerging architectures accelerate deep learning training by orders of magnitude, deliver a petaflop/s on a chip for streaming inference, and maximize power efficiency for IoT sensor arrays.

It is essential for the research community to have access to a testbed (or testbeds) of such technologies to gain timely experience and identify those technologies that would benefit science and engineering. That knowledge would then inform larger-scale deployments of the technologies that prove to be most valuable. In many cases, architectural and AI expertise is needed to fully exploit new hardware technologies: simply providing users with access to newly emerging hardware is not sufficient. This kind of testbed is exactly what PSC is doing through its NSF-supported *Open Compass* [1] project,

which is evaluating the performance of deep learning networks relevant to research (which differ significantly from industry-motivated benchmarks such as MLPerf and GLUE) on new AI technologies, leveraging a frequently-refreshed AI technology testbed at PSC. Sustaining and expanding such effort is vital for effectively bringing new AI technologies to science and engineering research.

3 INNOVATIONS

Wide-ranging innovations will allow CI for AI-enabled science to reach its full potential. These innovations range across all aspects of resource provisioning, operations, and user engagement. Some examples are as follows.

AIOps can improve system reliability, availability, and performance. Early work has already been done to applying machine learning to system instrumentation data, for example, to identify and proactively repair components that are degrading.

The research community needs specialized, advanced hardware that delivers great scalability for specific types of problems. (This is in addition to testbeds of frequently-refreshed hardware as described in §2.) For example, individual training runs can take hours to weeks even on today's most advanced GPUs, and fully training a model can take hundreds or thousands of runs across different network architectures and sets of hyperparameters. New kinds of hardware that are architected specifically to accelerate deep learning training can accelerate those runs by orders of magnitude, potentially transforming researchers' ability to produce higher-accuracy models. Other kinds of specialized hardware are being developed for inferencing, streaming data, and edge devices (which can stream data back to large CI resources).

The research community also needs more flexible ways to build customized software environments. Tools such as Anaconda work for some software stacks, but not for all. Containers are an increasingly popular alternative, and container technologies such as Singularity [4] are widely adopted in HPC. However, many users lack local resources to build their own containers, and for some purposes (e.g., to incorporate system-specific libraries) it would be extremely helpful to build containers on the target platform. For that, it would be helpful to deploy "sandbox" environments which provide isolated, secure spaces in which users can have the necessary privilege levels to build or refine the containers they need.

The complexity of the AI landscape also introduces challenges in applying for, and granting, research requests. Innovations are needed in the process, guidelines, and tools to properly match project and compute needs to the right resources. Improved methods are needed for handling research proposals to accommodate the different, non-traditional nature of AI/ML and data science research. There is also an opportunity to use AIOps to improve the proposal review process, specifically by applying natural language understanding to infer the actual emphasis of each proposal, to assist in recruiting and assigning the most appropriate reviewers.

4 TRAINING

User training in scalable AI for NSF-supported CI is greatly needed and in very high demand. PSC leads the XSEDE HPC Monthly Workshop series [8], which rotates through topics of MPI, OpenMP, OpenACC, and Big Data & AI. To date, these workshops have hosted 11,921 participants. The most popular workshops, which are now offered every other month, are on Big Data & AI, for which 7,158 individuals from 83 different institutions participated in 19 events. Up to 26 institutions and 612 individuals have participated per event, using PSC's Wide Area Classroom format, and each event features extensive hands-on exercises.

The XSEDE workshops have proven to be extremely successful and valuable. A valuable next step would be to develop more advanced modules focusing on AI to foster workforce development and more effective use of advanced CI. Specifically, users need guidance on the following:

- *Data acquisition and preparation*, including tools for orchestrating training (e.g., hyperparameter optimization) and measurement of bias and correcting for it.
- *Scalable AI*, or using many AI accelerators together, including distributed across many nodes and the use of advanced hardware, to reduce the time needed to train models, thereby facilitating higher accuracy and testing more ideas.
- *Measuring and improving performance*, including profiling, I/O optimization, and maximizing the efficiency of task and data parallelism.
- *Choosing optimal resources and estimating requirements*, including how to determine which hardware technologies and software frameworks would best fit specific tasks and how to estimate computational requirements.
- *Advanced topics*, including explainability, developing and validating surrogate models, multimodal data, reinforcement learning, and domain-specific networks and data types.

5 TOOLS AND TECHNIQUES

Usability and effective tools are vital for CI- and AI-enabled science and engineering. Users increasingly come from laptop and cloud backgrounds where providers have prioritized usability to gain market share. CI providers must place the user first and deliver the following:

- *Highly usable systems* on par with cloud providers.
- *Intuitive, interactive monitoring tools* for both service providers and end users.
- *Tools to assist with data acquisition and preparation*, such as cleaning and labeling.
- *Flexible data handling tools* to tackle the complex workflows that occur when using AI for in a new field or new domain challenge.

SUMMARY

The resurgence of innovation in computer architecture and the high scientific impact of artificial intelligence and machine learning present unique opportunities for deploying transformative cyberinfrastructure platforms, enhancing their operation through use of AIOps, engaging in research on AI & ML issues specific to computational science, training the computational science community, and strengthening the software ecosystem with emphases on usability and data. For each of these areas, the Pittsburgh Supercomputing Center has initiatives already underway, yet significantly more effort is needed.

REFERENCES

[1] Paola A. Buitrago and Nicholas A. Nystrom. 2019. Open Compass: Accelerating the Adoption of AI in Open Research. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)* (Chicago, IL, USA) *(PEARC '19)*. Association for Computing Machinery, New York, NY, USA, Article Article 72, 9 pages. https://doi.org/10.1145/3332186.3332253

[2] Paola A. Buitrago, Nicholas A. Nystrom, Rajarsi Gupta, and Joel Saltz. 2020. Delivering Scalable Deep Learning to Research with Bridges-AI. In *High Performance Computing: 6th Latin American Conference, CARLA 2019*, J. L. Crespo-Mariño and E. Meneses-Rojas (Eds.). Communications in Computer and Information Science, Vol. 1087. Springer, Basel, Switzerland.

[3] M. F. Kasim, D. Watson-Parris, L. Deaconu, S. Oliver, P. Hatfield, D. H. Froula, G. Gregori, M. Jarvis, S. Khatiwala, J. Korenaga, J. Topp-Mugglestone, E. Viezzer, and S. M. Vinko. 2020. Up to two billion times acceleration of scientific simulations with deep neural architecture search. arXiv:stat.ML/2001.08055

[4] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. 2017. Singularity: Scientific containers for mobility of compute. *PLoS ONE* 12, 5 (2017), e0177459. https://doi.org/10.1371/journal.pone.0177459

[5] Nicholas A Nystrom, Paola A Buitrago, and Philip D Blood. 2019. Bridges: Converging HPC, AI, and Big Data for Enabling Discovery. In *Contemporary High Performance Computing: From Petascale toward Exascale, Volume Three*, Jeffrey S. Vetter (Ed.). CRC Press, Boca Raton, FL.

[6] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. 2019. *The AI Index 2019 Annual Report*. Technical Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

[7] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. 2019. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications* 10, 1 (2019), 2903. https://doi.org/10.1038/s41467-019-10827-4

[8] John Urbanic and Thomas Maiden. 2018. Evaluating the Wide Area Classroom after 10,500 HPC Students. In *2018 IEEE/ACM Workshop on Education for High-Performance Computing (EduHPC)*. IEEE, New York, NY, 51–60.