PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mba-8luc-9vnj
Comments Due: May 28, 2025 **Submission Type:** API

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0345
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Fast Machine Learning Collaboration and AI-Accelerated Algorithms for Data-Driven Discovery Institute

General Comment

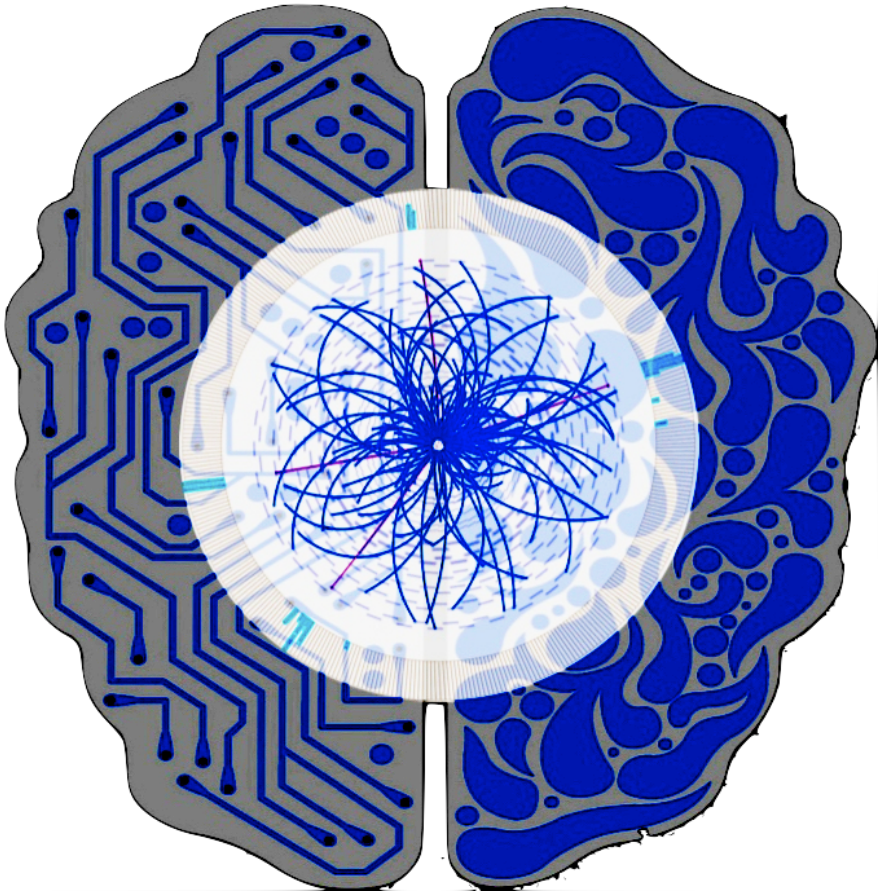
See attached file(s)

Attachments

FastML_AI_RFI

Artificial Intelligence Research and Development Strategic Plan for Fast Machine Learning

Fast Machine Learning Organization (https://fastmachinelearning.org/)
AI-Accelerated Algorithms for Data Driven Discovery(A3D3) Institute (https://a3d3.ai/)



Artificial Intelligence (AI) has had a transformative impact on society. Despite rapid progress, some areas of AI research progress more slowly due to difficulties in research and development. In particular, the development of low-latency, fast AI requires different goals and AI design to achieve the ultimate performance. AI for ultra-fast processing often requires a small batch coupled with extensive parallelism, power efficiency, and extremely compressed AI algorithms to ensure computational latency is minimal, while sustaining high throughput. This, ultimately, means that GPUs, CPUs, and other conventional processing computational paradigms need to be rethought, augmented, and sometimes replaced with bespoke AI algorithmic designs customized within the processor instruction set. However, once modified, algorithms can achieve processing times that are over 100 times faster than conventional GPU- and CPU-driven AI algorithms. This is a remarkable speedup that has the potential to significantly impact the AI community, leading to faster, more powerful AI.

However, at present, much of the development in this direction is limited to a few scientific domains. In fact, the most pressing fast AI problems have originated from attempts to solve specific scientific problems in different domains, with domain scientists driving the research. Despite that, we are quickly realizing that the lessons learned from the individual domains can be consolidated into a new scientific field that will have significant benefits for future AI technologies. In particular, Fast AI:

- extends the capability of what is possible with AI, leading to the adoption of AI in new systems that are not limited by latency restrictions,
- opens new paradigms for AI design and optimization strategies for AI deployment; this leads to new, efficient chip design and heterogeneous computing systems
- motivates the need for extensive use of AI-based algorithm compression strategies, allowing for an understanding of how to make efficient AI
- ultimately leads to the possibility of both faster and lower power usage of all AI algorithms [1, 2].

In light of the further advancement of these technologies, a community has emerged over the past six years entitled the Fast Machine Learning community, where scientists and industry members have come together to develop novel AI solutions. Although the drivers of this effort have started in scientific domains, an ecosystem has emerged that brings this technology from one domain to the global community.

The Fast Machine Learning Community was founded to address these complex problems and has built a community around the need for solutions in many scientific domains, covering high energy physics, astronomy materials science, plasma physics, neuroscience, quantum computing, satellite controls, and real-time systems. Out of this greater effort, several dedicated funded programs have emerged, including the NSF-funded Harnessing the Data Revolution Institute: AI Algorithms for Data-Driven Discovery (A3D3) aimed to develop fast AI algorithms for real-time high-energy physics, astrophysics, and neuroscience experiments.

The impact of science on data rate and latency requirements can be seen in Figure 1. A broad range of systems exists in many different scientific domains that demand high streaming data rates approaching the petabyte per second data rate. Additionally, the demands of these experiments often require algorithms to respond in fractions of a second, below one millisecond. The existing computational toolkit that serves commercial domains, such as Google Cloud, Netflix, or other extensive computing facilities, cannot address the combined significant data rates and timescales needed to run these scientific experiments. As a result, there are no commercial solutions to these problems, and novel AI solutions are required to address these scientifically driven challenges.

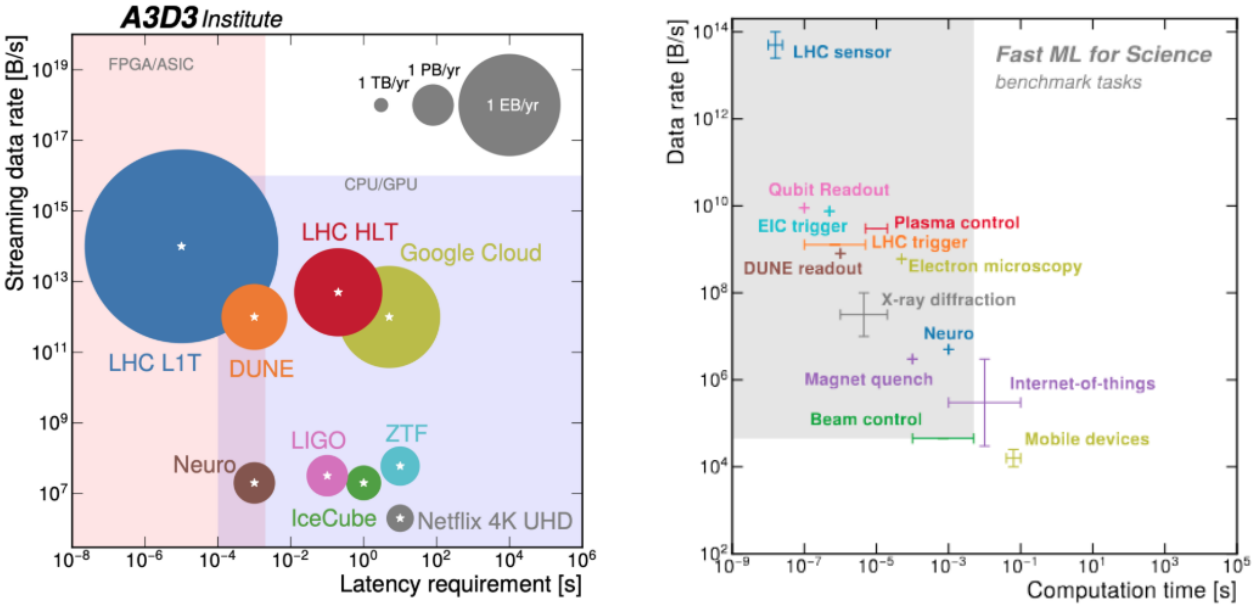


Figure 1: Left: Latency requirements and system-level data streaming rates for various experiments in the physics and neuroscience domain. Right: Latency requirements and task-level data rates for benchmark tasks across many domains of science.

We advocate for continued support of scientific research whose focus is on optimized AI strategies for new processing technology. While a large community is working to extend AI algorithm design, the community focusing on processor technology is smaller despite an equally significant need for development. Research support is critical to advance this field, as it will lead to the deployment of algorithms on customized hardware at scale.

To cultivate an ecosystem where we continue to drive optimized FastAI, we advocate for the continued support of cross-disciplinary efforts that aim to bring critical scientific problems from many different domains to the AI community. Building on the existing Fast Machine Learning/A3D3 effort for low-latency, efficient AI, we aim to grow the community while constructing cross-domain toolkits that allow for low-latency AI solutions. Already, we have had a significant amount of success through the development of the hls4ml software toolkit [3] and the SONIC software toolkit [4]. Continued support for this effort will allow us to extend Fast AI to more domains, finding science drivers where fast AI is needed. This direction differs from existing funded AI institutes, which tend to focus on just a few scientific domains. Here, we aim broadly to connect many domains with a common technological driver.

When dealing with large throughput systems, such as those present for big data experiments, including the Large Hadron Collider and the Vera C. Rubin Astronomical Observatory, there is a need to optimize computing resources to contend with the large data rates while being cost-effective. Current industry solutions are customized and are primarily driven by demands for large-scale deployment of large language models. Flexible approaches, which integrate with an existing broad series of workflows, are limited in scope. However, an inference toolkit that is flexible to many experiments and provides a rapid adoption of AI tools with optimized data ingestion is needed. Building on the chip-level ideas within the Fast Machine Learning Community, we also pursue strategies to optimize the use of heterogeneous systems (CPU/GPU and other processors) at a large scale. Here, our work targets a solution that builds on existing scientific workflows, allowing for the continual adoption of new AI algorithms as AI solutions emerge.

Connecting the research, fast machine learning work targets optimized use in a generic setting, allowing chip-level optimizations to be propagated to systems-level optimizations. This work leads to a holistic view of how we can adopt fast machine learning strategies both at the macro- and micro-levels. As a result, we aim to build a standard toolkit that captures all scales and can be applied to many scientific workflows. A survey of scientific applications, problems, and tools can be found in this A3D3 white paper [5].

Cultivating a cross-disciplinary effort that brings together the many scientific drivers is crucial to preserve the ecosystem and to advance developments in Artificial Intelligence algorithms and chip design. An institute that captures these many scientific domains and produces toolkits and a knowledge base capable of deploying fast AI algorithms would help preserve and grow the knowledge base. Scientific challenges put different demands on what is needed to ensure that algorithms can be run at high throughput and low latency. Their combined efforts are seeds for the creation of cross-domain toolkits capable of deploying Artificial Intelligence algorithms both within their domains and beyond. With existing toolkits, such as HLS4ML and SONIC [4, 6],the Fast Machine Learning Organization is already enabling a community to deploy artificial intelligence in many domains. While working as a single entity, this organization is funded through many different initiatives. The single most significant support is from the NSF A3D3 institute, whose mandate ends within 1.5 years, and which covers only 3 of the many science drivers. Continued support for a cross-disciplinary institute that broadly covers many domains would further advance this emerging field of artificial intelligence while consolidating the many different bespoke, domain-specific tools currently maintained.

In summary, the work within the Fast Machine Learning Organization goes beyond the conventional Artificial Intelligence toolkit and enables algorithms to be deployed within systems where computing throughput, latency, and integration demands were previously impossible. The knock-on impacts of this work are substantial and are leading to new devices that will have significant impacts beyond science. Ultimately, the fast machine learning community aims to bring the learned knowledge to the public to ensure everyone benefits from the technological advances that rapidly appear as artificial intelligence advances and demands for more and faster AI increase.

References

[1] A. M. Deiana *et al.*, "Applications and Techniques for Fast Machine Learning in Science," *Front. Big Data*, vol. 5, p. 787421, 2022. doi: 10.3389/fdata.2022.787421

[2] F. Fahim, B. Hawks, C. Herwig, J. Hirschauer, S. Jindariani, N. Tran, L. P. Carloni, G. Di Guglielmo, P. Harris, J. Krupa *et al.*, "hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices," *arXiv preprint arXiv:2103.05579*, 2021.

[3] J. Duarte, S. Han, P. Harris, S. Jindariani, E. Kreinar, B. Kreis, J. Ngadiuba, M. Pierini, R. Rivera, N. Tran *et al.*, "Fast inference of deep neural networks in fpgas for particle physics," *Journal of instrumentation*, vol. 13, no. 07, p. P07027, 2018.

[4] J. Krupa *et al.*, "GPU coprocessors as a service for deep learning inference in high energy physics," *Mach. Learn. Sci. Tech.*, vol. 2, no. 3, p. 035005, 2021. doi: 10.1088/2632-2153/abec21

[5] M. Agarwal *et al.*, "Applications of Deep Learning to physics workflows," 6 2023.

[6] Fast Machine Learning Collaboration, "fastmachinelearning/hls4ml," 2023. [Online]. Available: https://github.com/fastmachinelearning/hls4ml