SCAIGATE: Science Gateway for Scientific Computing with Artificial Intelligence and Reconfigurable Architectures

David Ojika[1 2], Herman Lam[1], Bhavesh Patel[2], Ann Gordon-Ross[1]
[1]University of Florida [2]DELL EMC

***Abstract*— SCAIGATE is an ambitious project to design the first AI-centric science gateway based on field-programmable gate arrays (FPGAs). The goal is to democratize access to FPGAs and AI in scientific computing and related applications. When completed, the project will enable the large-scale deployment and use of machine learning models on AI-centric FPGA platforms, allowing increased performance-efficiency, reduced development effort, and customization at unprecedented scale, all while simplifying ease-of-use in science domains which were previously AI-lagging. SCAIGATE was an incubation project at the Science Gateway Community Institute (SGCI) bootcamp held in Austin, Texas in 2018.**

I. INTRODUCTION

Science gateways provide community-based access to shared, distributed, advanced technologies and resources that support science and engineering research and education [1]. These resources, in the form of data, software, high-performance computing, instrumentation and collaboration tools, enable the formation of scientific communities, accelerating the discovery process, and engaging citizens in the scientific process [2]. In particular, SCAIGATE – *scientific computing with artificial intelligence gateway* – is a science gateway which integrates field programmable gate-arrays (FPGAs) and artificial intelligence (AI) to facilitate machine learning through data preprocessing, training and inferencing. SCAIGATE will help computational scientists and researchers accelerate their data analyses workflows at a fraction of the time and effort compared to existing systems. FPGAs are attractive to AI because of their real-time processing capability, energy efficiency, and reconfigurability to the rapidly evolving AI innovations.

In recent times, AI and deep learning have witnessed explosive growth in almost every subject involving data. Complex data analyses problems that took prolonged periods, or required laborious, manual effort are now being tackled through AI and deep learning techniques at unprecedented accuracy [3]. Given the massive computing demands of these techniques, accelerator platforms – graphics processing units (GPUs) in particular – have been widely adopted to achieve speedup [4], even when such platforms (as compared to FPGA-based platforms) are costly, energy-inefficient and not well suited to real-time processing of streaming data in mission-critical applications (e.g., nanoscience imagery in electron microscopy, satellite image analysis in environmental monitoring, to name a few). For these reasons, SCAIGATE will be based on an FPGA pool, along with open-source software and science gateway interface that together support AI-centric science as

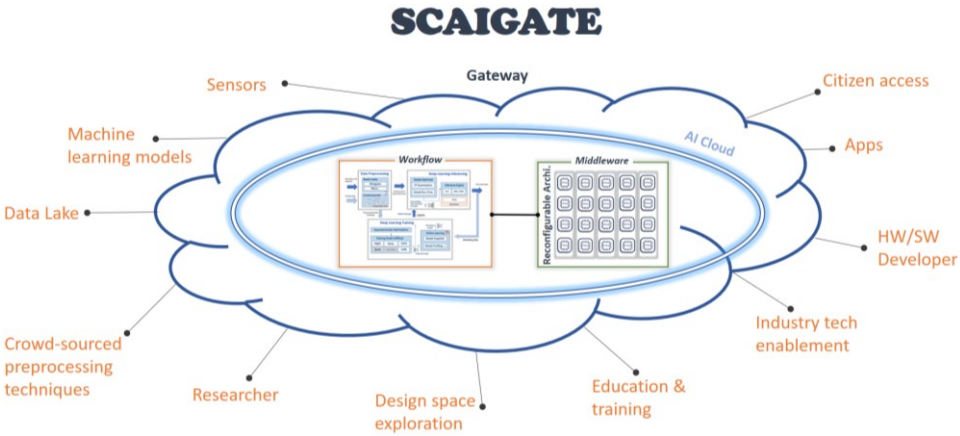

Fig 1: SCAIGATE ecosystem. Workflow and Middleware provide an "*AI Cloud*", exposing through the Gateway a unified science-as-a-service plaform backed by high-performance reconfigurable architectures (i.e., FPGAs).

a service (Fig. 1).

However, in contrast to mainstream processors, including CPUs and GPUs, FPGAs are more difficult to program, deploy and manage at large-scales, limiting their usage in scientific computing. Therefore, the vision of SCAIGATE is to provide a platform for deploying and experimenting with FPGA-enabled AI at production scale, advancing the capabilities of scientific computing, opening opportunities for AI-driven data analyses in a variety of science and engineering fields. The mission of SCAIGATE, and by extension, SCAIGATE ecosystem, is to greatly simplify the combination of FPGAs and AI techniques for scientific computing.

The SCAIGATE ecosystem comprises three fundamental components (see Fig. 1): (1) reconfigurable hardware, enabling deep learning acceleration with FPGAs; (2) system software, consisting of an FPGA middleware and a novel workflow management framework to support integration with scientific workflows; and (3) gateway, to simplify ease-of-use, expose AI services, and extend access to community through portals and application programming interfaces (APIs). The focus of this paper is on the workflow management framework – Cloudorch, aimed at orchestrating scientific workflows while seamlessly leveraging FPGA-accelerated cloud-based services.

II. CLOUDORCH

A major shortcoming of many scientific workflows is their limited interoperability, lack of component reusability, and curbed portability to new, advanced hardware (e.g., FPGAs). By leveraging community support and open-source software technology, Cloudorch will provide an FPGA-accelerated scientific workflow through platform as a service (PaaS), allowing scientists more focus on the hypothesis-test cycle instead of programming and maintaining toolchains, reducing re-invention, and accelerating discovery process. Cloudorch also aims to support the sharing of data preprocessing techniques, a crucial drawback in migrating non-AI based workflows to accelerated AI platforms. By abstracting key data processing workflows (data preprocessing, deep learning training and deep learning inference) as illustrated in Fig 2, Cloudorch will provide a scalable, end-to-end workflow, allowing users to go quickly from experiments to results. Because each Cloudorch component is a set of microservices that are loosely-coupled, users can compose customized workflows, train or import models, and deploy models effortlessly while leveraging FPGA hardware-acceleration.

Our previous work proposed a workload-intuitive framework, SWIF [5], and *FPGA as microservices* (FaaM) [6] to streamline the deployment of FPGAs in datacenters and the cloud, achieving 3x speedup and 40% memory-footprint savings in Apache Spark [7]. We have also benchmarked a variety of FPGA-based platforms, including the integrated Xeon-FPGA platform [8] and an Arria-10 accelerator platform [9]. As test-cases, we ported AlexNet deep learning model on three representative computing environments (university, cloud, and enterprise): University of Florida's NOVO-G#; Amazon AWS F1 compute instances; and Intel Programmable Accelerator Card (PAC) cluster at Dell, respectively, the results which will appear in the extended version of this paper.

III. CONCLUSIONS

Given the unique benefits of FPGAs (low-latency, energy-efficiency and reconfigurability), we have researched on ways of combining FPGAs with AI for scientific computing. While the initial results are promising, future work will be much more impactful through collaboration. To extend community access and foster scientific and engineering collaborations, we proposed SCAIGATE science gateway with the goal of advancing the capabilities of scientific computing with respect to AI and FPGAs. In particular, we proposed Cloudorch, a community-driven effort and framework to support scientific workflows with FPGA-based deep learning inferencing, while enabling end-to-end composability across entire deep learning stack. In the next coming months, we anticipate more collaboration with academic and industry partners, working closely on applications, tools and novel architectures to establish and show-case scientific and engineering use-cases of FPGA-accelerated AI at record performance, productivity and efficiency.

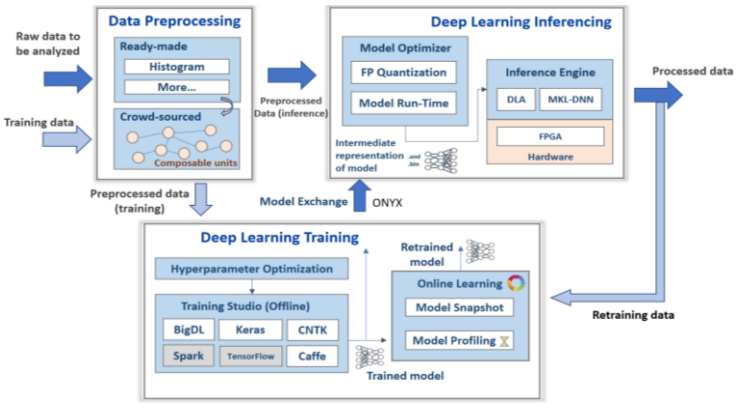

Fig 2: Cloudorch. Components are decoupled, allowing users to compose accelerated scientific workflows. Users can import/deploy models, train new models, or improve models on-the-fly by learning the models incrementally.

IV. ACKNOWLEDGEMENTS

The current work progress has been made possible through collaboration with the University of Florida's SHREC center and ApaLab, Dell, Intel, NERSC and the Science Gateway Community Institute (SGCI). We acknowledge the generous research awards from Microsoft Azure and Amazon AWS.

Project Contributor	Role
Univ. of Florida	FPGA expertise (research lead)
NERSC	ML expertise
CERN OpenLab	Data source
Intel	AI hardware/tools
Dell	Solutions provider
SGCI	Science gateway consultancy

REFERENCES

[1] Lawrence, Katherine A., Michael G. Zentner, Nancy Wilkins-Diehr, Julie Wernert, Marlon E. Pierce, Suresh Marru and Scott Michael. "Science gateways today and tomorrow: positive perspectives of nearly 5000 members of the research community." Concurrency and Computation: Practice and Experience 27 (2015): 4252-4268.

[2] Sandra Gesing Katherine Lawrence Nancy Wilkins-Diehr Maytal Dahan Michael Zentner Marlon E Pierce, "Science Gateways: Sustainability via On-Campus Teams"

[3] S. Uchida, S. Ide, B. K. Iwana and A. Zhu, "A Further Step to Perfect Accuracy by Training CNN with Larger Data," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, 2016

[4] https://www.nvidia.com/en-us/data-center/dgx-systems/

[5] D. Ojika, P. Majcher, W. Neubauer, S. Subhaschandra and D. Acosta, "SWiF: A Simplified Workload-Centric Framework for FPGA-Based Computing," 2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), Napa, CA, 2017

[6] David Ojika, Ann Gordon-Ross, Herman Lam, Bhavesh Patel, Gaurav Kaul, Jayson Strayer "Using FPGAs as Microservices: Technology, Challenges and Case Study", The Ninth Workshop on Big Data Benchmarks, Performance, Optimization and Emerging Hardware, BPOE 2018

[7] https://databricks.com/session/speeding-up-spark-with-data-compression-on-xeonfpga

[8]https://www.altera.com/solutions/acceleration-hub/overview.html

[9] https://www.altera.com/products/fpga/arria-series/arria-10/overview.html

[10] Y. Gil et al., "Examining the Challenges of Scientific Workflows," in Computer, vol. 40, no. 12, Dec. 2007