Accelerating Data Processing at the Edge with Extreme Specialization

Data-management support for AI and complex workflows

ANTONINO TUMEO, MARCO MINUTOLI, VITO GIOVANNI CASTELLANA, ANKUR LIMAYE, CHENG TAN, ISMET DAGLI, NICOLAS BOHM AGOSTINI, SERENA CURZEL, VINAY AMATYA, and JOSEPH MANZANO, Pacific Northwest National Laboratory, USA

Challenges and Current Approaches. The US Department of Energy (DOE) operates the largest collection of experimental scientific instruments in the world for a variety of scientific disciplines. Each discipline's specific instruments and computing requirements introduce unique design challenges for addressing the data management and processing required by the application. The vision for the next generation of scientific instruments is that they will produce higher volumes of data at a higher velocity [1] and are expected to perform computing and learning tasks at the edge in order to automate experiments and/or assist domain scientists. Redesigning the computer hardware and software systems that are capable of supporting the data-management aspects required to build tightly integrated and intelligent scientific instruments will be fundamental to maintain DOE's leadership in scientific discovery.

The advent of the Internet-of-Things (IoT) has led the microelectronics research community to explore efficient, fast, scalable, reliable, and secure edge computing devices that are able to process data in real time at the source or near to the user. However, only the high-volume commercial edge computing devices can justify the high non-recurrent engineering costs for their development. These devices also have to collect and manage significantly lower amount of data, and are typically specialized for one or just a few specific use cases. Experimental instruments, such as electron microscopes, particle accelerators, mass spectrometers, and more, may need to set up a variety of specific experimental workflows to collect different types of data, at different times, and in some cases with limited opportunities to repeat the experiments. The sheer amount of data collected may not allow in situ processing, so methods to efficiently store, reduce, and move the data to a large-scale system for analysis and simulation are required. While some partial solutions may be achieved by combining existing high-performance embedded devices with reconfigurable hardware, they may not be able to address all the unique requirements of scientific instruments.

Opportunities. We believe the future of compute within DOE experimental facilities lies in a division of processing. At the experimental instrument or sensor, compute/data analysis will be present, providing high-speed online processing, conditioning, and extraction of features as the data is captured. Experimental steering will be possible through results learned during online processing leveraging custom accelerators and by providing feedback to sensors on where best to focus their data gathering. Such an approach opens new research opportunities – most notably, how best to partition the processing between edge-based and data center-based devices, how best to design and optimize each device for performance, power and cost, and finally, how to interconnect these elements to deliver the highest possible aggregate performance to instrument users while retaining flexibility.

To enable this vision, we believe that an opensource, modular, extensible, multilevel compiler toolchain to enable hardware/software codesign and agile end-to-end generation of domain specific system is required. Enabling domain scientists to move form novel algorithmic formulation to the implementation of a system with dedicated accelerators, exploiting either reconfigurable logic (e.g. Field Programmable Gate Arrays) or application specific integrated circuits (ASICs) without the assistance of a team of hardware designers, offers unique opportunities to accelerate the data processing and management pipelines from the scientific instruments to scientific discovery. A modular and multilevel compiler infrastructure [2] allows interfacing with the productivity tools adopted by domain scientists. They are

Corresponding Author: Antonino Tumeo; Authors' addresses: name.surname@pnnl.gov.

critical to initiate architecture independent optimizations and design space exploration of the generated systems as early as possible, to maximize the benefit of user-provided information. Dedicated hardware generation engines based on existing opensource compiler technologies can today leverage a richness of algorithmic solutions to generate highly optimized circuit designs, especially in the case of Finite State Machines with Datapath (FSMD). Interfacing with novel higher-level compiler infrastructures with their natural support for hierarchy and (task level, coarse grained) parallelism, opens further opportunities in generating and composing hierarchical hardware systems. The hardware generations process benefits from the availability of opensource or licensable hardware intellectural properties (IPs), which can beccome part of the resource library for such compiler-based toolchains, enabling algorithmic and hierarchical system-level design. This not only accounts for opensource instructions sets (such as RISC-V) but also templated accelerators [3] or even functional units. Compiler-based generators enable exploring the design space and setting parameters for these components (e.g., precision). Hence, they directly tie to the configurability of templated components. Additionally, they provide a path to supporting dynamic reconfigurability, for example, leveraging just-in-time compilation, where the intermediate representation (bit code) can be lowered to slightly different machine code depending on the overall system status and adapt dynamically to the experimental workflows.

Timeliness. DOE's scientific instruments present unique challenges. Some target very specific edge use cases, others are connected to large-scale instruments. In all cases, the ability to process multi-modal high-bandwidth data streams in real time, perform data compression and management, and identify key data points of actual value, is critical. As sensors evolve, the amount of collected data will explode. Only highly specialized, custom systems might satisfy these unique, contrasting requirements, but the complexity, and ultimately the costs, associated with their design, from software to hardware implementation, are too high. Despite the advances in tools and architectures, complexities and costs continue to rise. Industry alone cannot satisfy the needs of DOE's scientific instruments. Addressing these needs can only happen by leveraging community efforts to build adequate, end-to-end tools to enable the automated generation of specialized systems. The emergence of opensource compiler-based design automation tools, opensource hardware, chiplet based designs, establishes a unique context ripe for new research and investments to empower domain scientists with methods to support complex, data intensive experimental workflows.

Conclusion. We have identified needs for the co-design of efficient edge-computing and high-performance data processing facilities for DOE's intelligent scientific instruments. We have argued that to address these unique needs of these instruments we need new end-to-end design automation tools able to generate application-specific accelerators from high-level productive programming frameworks. In particular, we have highlighted the impact of these tools to generate hyper-specialized systems able to deal with the volume, velocity, variety, veracity, and value (big 5 Vs) of data provided by new sensors in scientific instruments. We believe that it will be possible to address these key data management needs only by leveraging a modular, end-to-end, compiler-based agile hardware design toolchain able to generate highly specialized data analytics and artificial intelligence accelerators.

REFERENCES

[1] E. Wes Bethel, Martin Greenwald, Kerstin Kleese van Dam, Manish Parashar, Stefan M. Wild, and H. Steven Wiley. 2016. Management, analysis, and visualization of experimental and observational data - The convergence of data and computing. In *12th IEEE International Conference on e-Science, e-Science 2016, Baltimore, MD, USA, October 23-27, 2016.* IEEE Computer Society, 213–222. https://doi.org/10.1109/eScience.2016.7870902

[2] Chris Lattner, Jacques A. Pienaar, Mehdi Amini, Uday Bondhugula, River Riddle, Albert Cohen, Tatiana Shpeisman, Andy Davis, Nicolas Vasilache, and Oleksandr Zinenko. 2020. MLIR: A Compiler Infrastructure for the End of Moore's Law. *CoRR* abs/2002.11054 (2020). arXiv:2002.11054 https://arxiv.org/abs/2002.11054

[3] Antonino Tumeo, Marco Minutoli, Vito Giovanni Castellana, Joseph B. Manzano, Vinay Amatya, David Brooks, and Gu-Yeon Wei. 2020. Invited: Software Defined Accelerators From Learning Tools Environment. In *57th ACM/IEEE Design Automation Conference, DAC 2020, San Francisco, CA, USA, July 20-24, 2020.* IEEE, 1–6. https://doi.org/10.1109/DAC18072.2020.9218489