Dynamic Information Extraction with Edge AI

Audrey Corbeil-Therrien, Angelo Dragone, Ryan Herbst, Jana Thayer, Ryan Coffee

February 20, 2020

Problem Statement The Linac Coherent Light Source II (LCLS-II) holds great promise to reveal critical rare events such as the onset of battery failure, mutation events and atomic scale motion during DNA zipping and protein folding. There is also a growing portfolio of newly enabled experimental techniques such as femtosecond x-ray Fourier holography [2] and femtosecond dark field x-ray microscopy [3] that interrogate nano- and meso-scopic movies of the action of photochemistry and emergent materials properties. At the heart of all these objectives lies the need to identify both very weak and/or very rare events in overwhelmingly cluttered and noisy data. To target these science areas, SLAC is developing a suite of extreme data rate detectors including the ePix family of x-ray imaging detectors with an eventual multi-megapixel readout at one million frames per second within this decade. Such a raw data volume would be equivalent to producing over 100 years worth of 4K (Ultra-HD) video [4] every day, requiring nearly $1M[1] in storage [5]; conventional Datacenter hosted mining is not a viable option. Similar to the multi-threading paradigm shift that enabled continued exponential computing performance in spite of physical clock and power limitations of the mid-2000s (Fig. 1), the Datacenter HPC to Edge AI paradigm shift is upon us now. Although the scale of the data for LCLS-II and the upcoming Advanced Photon Source Upgrade (APS-U) pose truly unique challenges, the evolution of Industry 4.0 and 5G networked autonomous vehicles is revealing an critical need for data handling at the point of generation, at the sensor [6, 7].

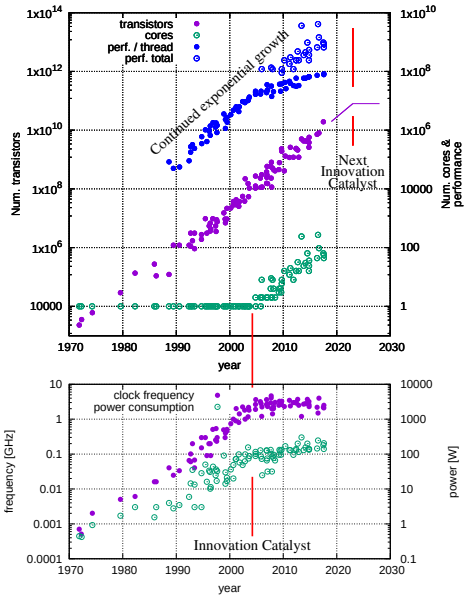

Figure 1: Adapted from Ref. [1]. Note that the limitations in the mid-2000s triggered the multi-threading paradigm.

We therefore propose streaming actionable information extraction that can flexibly handle a weekly or daily re-definition of actionable information. At our National Lab facilities, every week new users mount experiments with vastly different requirements and objectives and preclude a static data acquisition system. The ability to move CPU/GPU data analysis into firmware is an enabling step to high throughput but dynamically adaptive experiments that require ultra-low latency. We require real-time streaming inference that dynamically routes data through different Edge AI hardware throughout the data acquisition chain, be it FPGA, GPU, CPU, or other emerging hardware [8, 9, 10, 11, 12]. This raises the complementary requirement for data and model provenance tracking of the data flow through these chained Edge AI systems that will be actively manipulating the data in flight to storage. Active data and model provenance must therefore be baked directly into the adaptive Edge AI paradigm.

Proposed solution We take inspiration for our Edge AI from Nature's own distributed processing in the human nervous system. For example, a significant portion of sensory processing occurs in the sensory organs themselves well before the information reaches the deeper neural networks inside the brain, i.e. the unconscious rapid reactions stabilizing eye movements. We propose a similar function for our scientific

[1]This only considers hardware costs. Actual costs including power, space and personnel are much higher (estimated at $30 per GB per month) and accumulate over time.

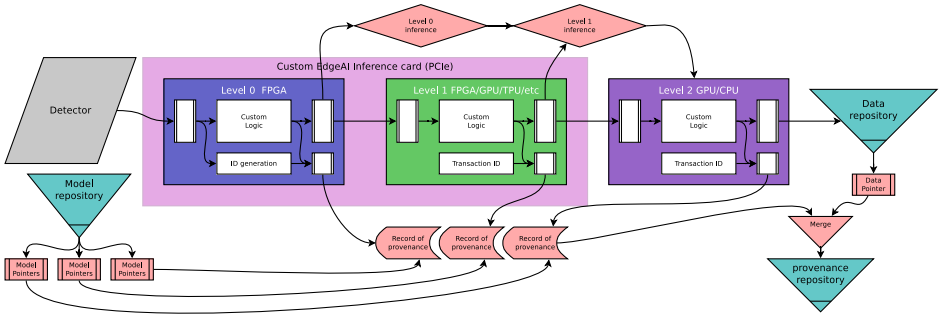

Figure 2: Schematic of information flow through heterogeneous hardware with ID generation and provenance tracking.

sensors; a processing unit at or near to the detector that can analyze incoming data in real-time and provide actionable information to the detector, the source, and the downstream (more complex) analysis networks. These inference engines will host dynamically adaptive algorithms and AI models that extract contextually relevant information before transformation and passage down the analysis chain. For example, the dynamic transformation chain will use early stage analysis to select the most appropriate subsequent compression algorithm. This Edge AI system will be hosted on Field Programmable Gate Arrays (FPGAs) as well as emerging flexible Edge AI hardware [8, 9, 10, 11, 12] that will be incorporated directly into the sensor, or nearly so, to minimize latency as well as alleviate the need for inappropriately large buffer memory. The flexibility of FPGAs and Edge AI devices is critical for user facilities like LCLS-II and APS-U where the firmware inference models will be completely replaced every weekly experimental cycle.

The Edge AI system will also include an accountability system that stitches the data ID and the shot-to-shot state of the algorithm chain into the data header, thus logging the precise actions taken on the particular data event into a unique data provenance record or ledger (see Fig. 2). The provenance ledger continues to track and auto-increment a use metric according to the derived scientific value of both data and algorithm. This "value aware" ledger can then be used for an automated dynamic data retention policy whereby lifetime in archive scales proportional to the evolving scientific value of a particular data set and/or algorithm. In other words, the more that data individuals are used for publications or even training AI inference models and the higher the impact of resulting projects and products, the longer the data individual remains active and discoverable in a data sharing marketplace. Aggregation would reveal the integrated value for scientific facilities, experimental techniques, and sensor technologies based on quantifiable impact.

The immediacy of our need for ultra-low latency and dynamic data compression at SLAC coupled with our long history of detector development from sensor to readout electronics to complex data pipeline design makes SLAC a unique environment with the necessary infrastructure to develop this Edge AI platform. The LCLS-II Data System has both offline systems and a data reduction layer that is tightly coupled to the detector to do near real-time processing hosted in FPGA hardware as a co-processor within the real-time processing layer or as stand-alone analysis accelerators. Finally, SLAC has experts in scientific instrumentation, data analysis, FPGA development, and machine learning all on-site to create, implement, and deploy the Edge AI system.

Deliverables The primary objective of this project is deployment of an extremely low latency, real-time Edge AI system that maximizes data reduction while also maximizing domain-specific information preservation. Ultra-low latency streaming analysis provides actionable information for autonomous feedback control of both the light source and the detector, thus enabling an fully adaptive instrument, source, and analysis pipeline. Since the transformation algorithms and ML models will adapt autonomously to the data in-flight, the system will encode a state-identifying fingerprint directly into the data header that serves as a provenance ledger. This data and algorithm provenance ledger can in turn be used to imbue results, respective algorithms, experimental methods, and facility beamlines with quantitative metrics that perpetually track their evolving derived scientific value.

The project will initially target the ePix family of detectors, developing bespoke raw data interpretation—extracting relevant information rather than raw data—via the Edge AI framework. The framework will also target commercial waveform digitizers and image capture cards for Edge AI which in turn will help formulate new interface standards for on- or near-detector FPGAs and emerging AI accelerating microchips.

References

[1] Karl Rupp. Microprocessor trend data – git repository. Available at: https://github.com/karlrupp/microprocessor-trend-data.

[2] Tais Gorkhover, Anatoli Ulmer, Ken Ferguson, Max Bucher, Filipe R. N. C. Maia, Johan Bielecki, Tomas Ekeberg, Max F. Hantke, Benedikt J. Daurer, Carl Nettelblad, Jakob Andreasson, Anton Barty, Petr Bruza, Sebastian Carron, Dirk Hasse, Jacek Krzywinski, Daniel S. D. Larsson, Andrew Morgan, Kerstin Mühlig, Maria Müller, Kenta Okamoto, Alberto Pietrini, Daniela Rupp, Mario Sauppe, Gijs van der Schot, Marvin Seibert, Jonas A. Sellberg, Martin Svenda, Michelle Swiggers, Nicusor Timneanu, Daniel Westphal, Garth Williams, Alessandro Zani, Henry N. Chapman, Gyula Faigel, Thomas Möller, Janos Hajdu, and Christoph Bostedt. Femtosecond x-ray fourier holography imaging of free-flying nanoparticles. *Nature Photonics*, 12(3):150–153, 2018.

[3] H. Simons, A. King, W. Ludwig, C. Detlefs, W. Pantleon, S. Schmidt, F. Stöhr, I. Snigireva, A. Snigirev, and H. F. Poulsen. Dark-field x-ray microscopy for multiscale structural characterization. *Nature Communications*, 6(1):6098, 2015.

[4] filecast blog. How to move large video files, 2016. Available at: https://filecatalyst.com/how-to-move-large-video-files/.

[5] Hilbert Hagedoorn. Decline price per gb for hdds comes to an end, 2017. Available at: https://www.guru3d.com/news-story/decline-price-per-gb-for-hdds-comes-to-an-end.html.

[6] Rob van der Meulen. What edge computing means for infrastructure and operations leaders, October 2018. Available at: https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders/.

[7] Ann Taylor. Edge computing is in most industries' future, April 2019. Available at: https://www.networkworld.com/article/3391016/edge-computing-is-in-most-industries-future.html.

[8] Google. Edge tpu performance benchmarks, 2019. Avaialable at: https://coral.ai/docs/edgetpu/benchmarks/.

[9] Google. Edge tpu, 2019. Available at: https://cloud.google.com/edge-tpu/.

[10] Chris Nicol. A coarse grain reconfigurable array (cgra) for statically scheduled data flow computing. 'Avaialable at: https://wavecomp.ai/wp-content/uploads/2018/12/WP_CGRA.pdf'.

[11] M. Wijtvliet, L. Waeijen, and H. Corporaal. Coarse grained reconfigurable architectures in the past 25 years: Overview and classification. In *2016 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS)*, pages 235–244, July 2016.

[12] Unique new streaming processor architecture, 2019. Available at: https://cornami.com/technology-products/.