Analog/Hybrid Co-Design Flow Methodology

Benjamin Parpillon[1,2], Amit Trivedi[2], Haitong Li[3], Jennifer Hasler[4], Farah Fahim[1]
bparpill@fnal.gov, amitrt@uic.edu, haitongli@purdue.edu, jennifer.hasler@ece.gatech.edu, farah@fnal.gov
[1]Fermi National Accelerator Laboratory, [2]University of Illinois Chicago, [3] Purdue University, [4]Georgia Institute of Technology

1 Topic

Our position paper discusses hybrid analog and digital systems design methodology.

2 Challenge

The rapid growth of analog sensor data has outpaced intelligent processing capabilities in many domains, including scientific experiments, causing an *analog data deluge* that obscures valuable information. At the LHC, Petabytes per second of data are generated thus requiring very high-speed offline filtering to avoid data pileups. A 2021 report [1] depicted in Fig.1, illustrates the total amount of data produced by LHC collisions in one year exceeded the total size of files ever stored on Amazon cloud storage services by approximately two order of magnitude. More than 90% of collision data is generated by a single detector system the **silicon pixel detectors**.

The data bottleneck is by far the greatest challenge faced by the HEP community. Currently, more than 99.995% of collision data is filtered out during offline data analysis looking only for rare interactions hoping to discover new Physics. The Level 1 trigger is mostly responsible for large amount of data to be rejected on-sensor reducing the information transfer from PBps to TBps. The innermost layers of the CMS/ ATLAS detectors currently do not contribute to the Level 1 trigger.

3 Opportunity

The data bottleneck challenge creates a unique opportunity for the HEP community to develop new computing co-design paradigm and methodology.

Current digital systems are unable to handle data rates needed at the LHC due to conversion overheads, limited parallelism, and resource intensiveness [2]. *On-sensor analog deep learning* can eliminate data conversion and storage overheads by filtering non-essential signals at the acquisition stage. Similarly, at the

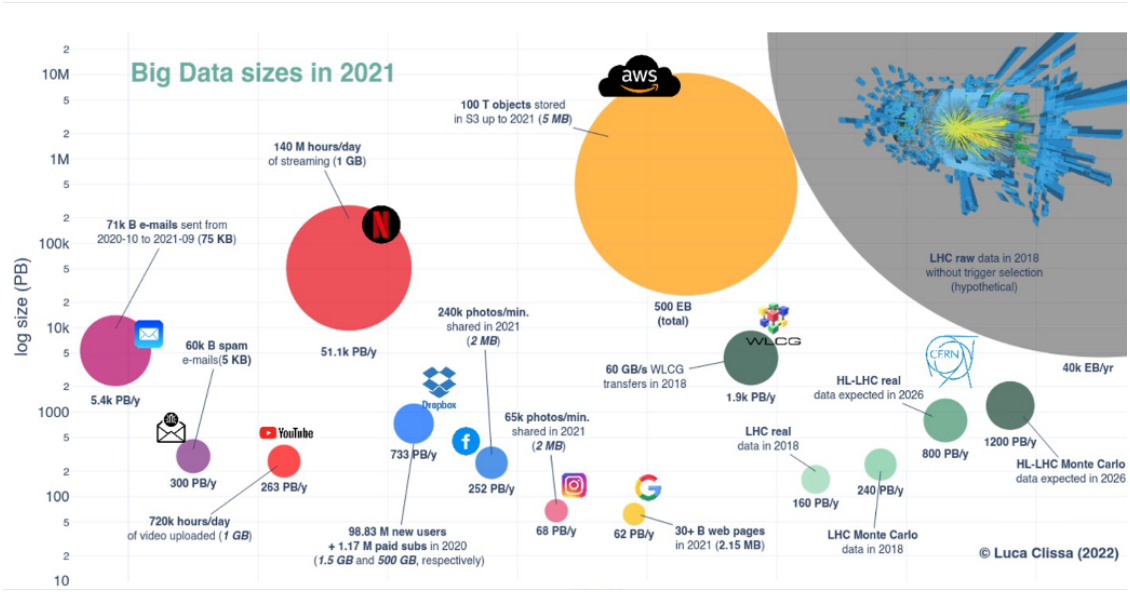

Figure 1: **Big Data in HEP:** Orders of magnitude involved in different data sources for several big data players. The area of each bubble represents the amount of data streamed, hosted or generated. The accompanying text annotations emphasize the crucial factors considered in the estimation process. Average per-unit sizes are indicated in parentheses, where italic denotes measures derived from reasonable assumptions due to the absence of available references

signal processing stage, analog deep learning can leverage *non-von Neumann architectures* to minimize data movements and employ *physics-based computing*, such as using Kirchoff's law for summation by representing operands in the charge or current domain, to maximize energy efficiency [3]. Thereby processing signals closer to the source in the analog domain offers significant benefits in performance, speed, area, and power, and can enable novel signal processing paradigms such as asynchronous real-time waveform analysis and direct signal-to-inference capabilities.

The flow need to address three sections illustrated in Fig. 2 :

- A first **pre-silicon** stage algorithmic flow to convert bit representations of software learning models to analog representations (such as charge, time, or current) suited for silicon placement. The algorithm also needs to solve optimal mapping of deep learning layers to a system of analog crossbars while maximizing throughput by concurrent processing, minimizing data transmission lengths and rates, maintaining load balancing, minimizing model load cycles from off-chip memories, *etc.*
- A second **on-silicon** stage that leverages novel mixed-signal processing architectures as well CMOS based and CMOS+X (FeFET, Floating Gates, PRAM, ReRAM, MRAM...) devices to fully exploit analog processing capabilities while addressing challenges such as design complexity and process variability.
- Finally, a third **post-silicon** stage is needed to address the need for continual monitoring and correction of the computing substrate against chip-to-chip, across-chip, and time-varying degradation such as process variability, aging, and temperature-induced variations.

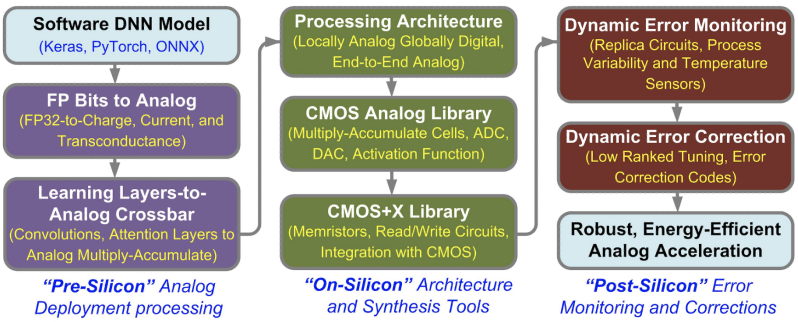

Figure 2: Co-designed analog acceleration flow consisting of pre-, on-, and post-silicon stages for reliable software-to-analog conversion, scalable synthesis and design, and runtime error monitoring and corrections.

4 Timeline

The AI/ML community for electronics design of HEP applications has now reached critical mass. The demand and need for increasingly more complex AI/ML models require new computing paradigm and design methodologies that supports analog/hybrid implementations. Recently, we introduced the support of Siemens Catapult HLS [4] as a backend of `hls4ml` to target specifically the ASIC flow. The significance for the industry of our framework has brought us to collaborate with Siemens EDA to release *Catapult AI NN* [5].

References

[1] Luca Clissa, Mario Lassnig, Lorenzo Rinaldi. How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry. *Frontiers Big Data*.

[2] Igor L Markov. Limits on fundamental limits to computation. *Nature*, 512(7513):147–154, 2014.

[3] Jennifer Hasler. Opportunities in physical computing driven by analog realization. In *2016 IEEE international conference on rebooting computing (ICRC)*, pages 1–8. IEEE, 2016.

[4] Siemens. Catapult HLS. `https://eda.sw.siemens.com/en-US/ic/ic-design/high-level-synthesis-and-verification-platform`.

[5] Siemens Digital Industries Software. Catapult ai nn simplifies development of ai accelerators, May 2024.