SAND2021-2665C

Achieving Extreme Heterogeneity: CoDesign using Neuromorphic Processors

Suma George Cardwell, Frances S. Chance, Craig M. Vineyard and James 'Brad' Aimone
Cognitive and Emerging Computing, Sandia National Laboratories
Email: sgcardw@sandia.gov

Topic: Architectures, Applications, Emerging technologies, Codesign methodologies

As conventional systems saturate in power efficiency, innovations in both architectures and algorithms are required to meet the computing needs of the future. With the slowing of Moore's law and the recent popularity of deep neural networks there has been renewed focus on emerging technologies, such as neuromorphic computing. Inspired by the brain, neuromorphic architectures leverage properties such as massive parallelism, sparse activity, and event-driven computing. Neuromorphic computing has the potential to be impactful for machine learning, scientific computing, modeling cognitive tasks as well as applications at the edge. Codesign tools are critical for the adoption of such novel technologies. If designed into a heterogeneous system with other accelerators and conventional computing platforms, this technology has the potential to augment the capabilities of High Performance Computing (HPC) platforms (1). This paper highlights the need for new heterogeneous tools and architectures through the lens of neuromorphic computing.

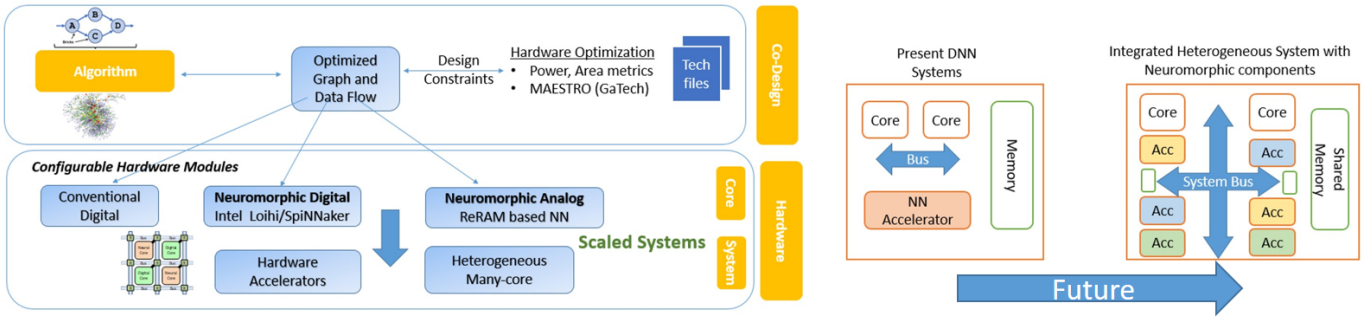

Figure 1: Developing Heterogeneous Architectures will require codesign tools that span algorithms and hardware. The future of computing will likely be extremely heterogeneous with different accelerator types. Image reproduced from (1)

Incorporating different classes of processors on single HPC node has been key to moving towards exascale computing. The scientific computing ecosystem is also changing with data collection outpacing theory in many fields like neuroscience, medicine, and, climatology. There has been explosion of different accelerator approaches in industry like TPUs, Cerebras (wafer-scale), Mythic (analog), each with a unique approach to overcoming performance bottlenecks. It is evident that future HPC approaches will be highly heterogeneous, where HPC system could include both conventional (CPUs, GPUs), and non-conventional approaches (neuromorphic hardware, Processing-In-Memory). Digital neuromorphic chips like Intel's Loihi have shown 100x efficiency gains compared to GPUs and CPUs and can be scaled to build larger systems (2). Analog neuromorphic architectures promise even further savings in energy efficiency, area, and latency than their digital counterparts (3; 4).

Current Challenges While a lot of progress has been made in codesign methodologies and their adoption, a major gap exists in the integration of novel computing paradigms like neuromorphic computing in heterogeneous computing. This is in part due to the lack of a cohesive codesign tool and also due the diverse nature of neuromorphic backends which range from digital, analog, mixed-signal to beyond-CMOS approaches.

Challenge#1: Codesign tools that support novel architectures. A lot of work is being currently done to incorporate GPUs, FPGAs to build heterogeneous systems. This is in part due the API and prototyping tools that are made available, as well as the ease of access to these devices for testing and validation. Such a framework is missing for neuromorphic processors– which are still evolving and have diverse approaches to the architecture and devices used.

Challenge#2: Developing applications for neuromorphic from HPC to the edge. Challenge 1 plays into challenge 2 in that, lack of codesign tools and ease of usability limits the different applications users can develop for these novel architectures. Thus, the barrier of entry is high, which hinders adoption.

Challenge#3: Exploration of next-generation heterogeneous neuromorphic architectures. We need to explore complex neurons and connectivity mechanisms that would make neuromorphic systems even more capable and apply to diverse set of problems. This will also include exploring novel devices, new integration techniques (3D architectures, photonics) and novel algorithms that exploit their characteristics.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Opportunity: The identified challenges present many opportunities through new tools and techniques, new technologies, and groups collaborating through open-source tools in the codesign process.

Codesign tools for Heterogeneous Neuromorphic Architectures: Accelerator tools have democratized the ability to test and validate different dataflow architectures. The neuromorphic field needs such open-access codesign tools available to the larger community that supports varied backends. This could entail analytical tools, cycle-accurate tools, as well as tools that enable exploring integration of neuromorphic accelerators with conventional processors. We can leverage deep learning accelerator modeling tools like Timeloop(5), MAESTRO (6), and NVDLA to explore heterogeneous architectures by extending them with different analog and digital neuromorphic kernels. Analytical tools work through computing energy related operations (number of memory read/write, number of MAC, NOC communications) given a certain technology node. These tools are beneficial for rapid testing and architecture prototyping. Such tools can feed into cycle-accurate explorations, e.g. Sandia's Structural Simulation Toolkit (SST) that can yield cycle-accurate simulations. Other tools that account for the performance of emerging device technologies are CrossSim and PUMA (3; 7). High level tools like Sandia's Fugu, are also needed to enable designing spiking neural networks while being hardware agnostic. A modular approach to codesign tools is also important, especially tools that enable integration of new architectures and device characteristics. This requires not just looking at non-Von Neumann architectures but also novel non-CMOS devices. While the software space is constantly evolving, building tools that can be re-usable, open, adaptable will be crucial to adoption.

Impact diverse set of applications: Many applications have been demonstrated for neuromorphic systems. Examples include solving PDEs using random walks on a neuromorphic platform (8), and a spiking implementation of Locally-competitive algorithms (LCA) (9) that implicitly solves the LASSO optimization problem with improved energy costs compared to conventional solvers (10). Our hypothesis is that multi-precision networks using neuromorphic processors will perform better than conventional computing approaches for scientific computing and machine learning algorithms. Recent programs like DARPA FENCE, focus on neuromorphic event sensor and processors that will bring low SWaP advantages to the edge. Thus, diverse applications will facilitate the development of new architectures that support a diverse set of algorithms and create an eco-system where users inform neuromorphic hardware developers. There is a tradeoff in codesigning applications and hardware, but applications are no longer immune to the hardware they run on, to gain performance benefits. This requires a strategic investment in codesign tools and approaches. Perhaps, the development of useful mini-apps for such heterogeneous architectures will be a good first step in this direction. Collaboration between industry, academia, and research laboratories is also important. Efforts like Intel's Neuromorphic Research Community and outreach by IBM (TrueNorth) for academic and research partners are good examples of this.

Next-generation extremely heterogeneous architectures: While we need novel neuromorphic devices to accelerate computation, we also need novel algorithms and architectures. A lot of current neuromorphic hardware uses simplified neuron models that can be scaled to billions of neurons. However, we hypothesize that designing complex neurons will augment the capabilities these systems currently offer. For example, introducing dendritic processing will introduce nonlinear summation, spatio-temporal processing, and increased connectivity. Techniques to do brain-inspired local learning is another area of active research that could impact the use of neuromorphic processors as not just inference but training engines. Next-generation neuromorphic circuits and systems based upon nonlinear dendritic processing and local learning will balance the trade-off between scalability and the biological complexity. Novel approaches in fabrication like three-dimensional architectures and wafer-scale technology as well as in-memory computing devices could further alleviate current communication and connectivity bottlenecks. This would require synergistic collaboration across devices, architectures, software, and algorithms.

Timeliness or Maturity Neuromorphic architectures have the potential to have an impact in the next 5-10 years as implementations in silicon exist today (Loihi, TrueNorth, SpiNNaker, offerings from BrainChip, GrAI Matter Labs). Non-CMOS approaches are promising and industry trends (Imec/Global Foundries) show that these architectures will be available for mass production soon. Neuromorphic accelerators can impact the efficiency of machine learning, scientific computing, and edge applications with two-three orders of magnitude improvement in energy and speed. Codesign tools will enable algorithm and hardware designers to account for novel accelerators in their design flows better. Hence, developing open-source codesign tools that include a wide variety of novel backends like neuromorphic processors is imperative to achieve extreme heterogeneity in the future.

Acknowledgments Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

[1] S. G. Cardwell, C. Vineyard, W. Severa, F. Chance, F. Rothganger, F. Wang, S. Musuvathy, C. Teeter, and J. B. Aimone, "Truly heterogeneous HPC: Co-design to achieve what science needs from HPC," *Smoky Mountain Computational Sciences and Engineering Conference*, Accepted for Talk, Publication 2020.

[2] P. Blouw, X. Choo, E. Hunsberger, and C. Eliasmith, "Benchmarking keyword spotting efficiency on neuromorphic hardware," in *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop*, 2019, pp. 1–8.

[3] S. Agarwal, A. Hsia, R. Jacobs-Gedrim, D. R. Hughart, S. J. Plimpton, C. D. James, and M. J. Marinella, "Designing an analog crossbar based neuromorphic accelerator," in *2017 Fifth Berkeley Symposium on Energy Efficient Electronic Systems & Steep Transistors Workshop (E3S)*. IEEE, 2017, pp. 1–3.

[4] J. Hasler and H. B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Frontiers in neuroscience*, vol. 7, p. 118, 2013.

[5] A. Parashar, P. Raina, Y. S. Shao, Y.-H. Chen, V. A. Ying, A. Mukkara, R. Venkatesan, B. Khailany, S. W. Keckler, and J. Emer, "Timeloop: A systematic approach to dnn accelerator evaluation," in *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2019, pp. 304–315.

[6] H. Kwon, M. Pellauer, and T. Krishna, "Maestro: an open-source infrastructure for modeling dataflows within deep learning accelerators," *arXiv preprint arXiv:1805.02566*, 2018.

[7] A. Ankit, I. E. Hajj, S. R. Chalamalasetti, G. Ndu, M. Foltin, R. S. Williams, P. Faraboschi, W.-m. W. Hwu, J. P. Strachan, K. Roy *et al.*, "Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 715–731.

[8] W. Severa, R. Lehoucq, O. Parekh, and J. B. Aimone, "Spiking neural algorithms for markov process random walk," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[9] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, "Sparse coding via thresholding and local competition in neural circuits," *Neural computation*, vol. 20, no. 10, pp. 2526–2563, 2008.

[10] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.