Report from Dagstuhl Seminar 19152

Emerging Hardware Techniques and EDA Methodologies for Neuromorphic Computing

Edited by

Krishnendu Chakrabarty[1], Tsung-Yi Ho[2], Hai Li[3], and Ulf Schlichtmann[4]

1 **Duke University – Durham, US,** krish@duke.edu
2 **National Tsing Hua University – Hsinchu, TW,** tyho@cs.nthu.edu.tw
3 **Duke University – Durham, US,** hai.li@duke.edu
4 **TU München, DE,** ulf.schlichtmann@tum.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19152 "Emerging Hardware Techniques and EDA Methodologies for Neuromorphic Computing," which was held during April 7–10, 2019 in Schloss Dagstuhl – Leibniz Center for Informatics. Though interdisciplinary considerations of issues from computer science in the domain of machine learning and large scale computing have already successfully been covered in a series of Dagstuhl seminars, this was the first time that *Neuromorphic Computing* was brought out as the focus.

During the seminar, many of the participants presented their current research on the traditional and emerging hardware techniques, design methodologies, electronic design automation techniques, and application of neuromorphic computing, including ongoing work and open problems. This report documents the abstracts or extended abstracts of the talks presented during the seminar, as well as summaries of the discussion sessions.

Seminar April 7–10, 2019 – http://www.dagstuhl.de/19152
2012 ACM Subject Classification Computer systems organization → Neural networks, Hardware → Biology-related information processing, Hardware → Hardware-software codesign
Keywords and phrases Neuromorphic computing; nanotechnology; hardware design; electronic design automation; reliability and robustness
Digital Object Identifier 10.4230/DagRep.9.4.43

1 Executive Summary

Hai Li (Duke University – Durham, US)

The explosion of *big data* applications imposes severe challenges of data processing speed and scalability on traditional computer systems. However, the performance of von Neumann architecture is greatly hindered by the increasing performance gap between CPU and memory, motivating active research on new or alternative computing architectures. Neuromorphic computing systems, that refer to the computing architecture inspired by the working mechanism of human brains, have gained considerable attention. The human neocortex system naturally possesses a massively parallel architecture with closely coupled memory and computing as well as unique analog domain operations. By imitating this structure, neuromorphic computing systems are anticipated to be superior to conventional

computer systems across various application areas. In the past few years, extensive research studies have been performed on developing large-scale neuromorphic systems. Examples include IBM's TrueNorth chip, the SpiNNaker machine of the EU Human Brain Project, the BrainScaleS neuromorphic system developed at the University of Heidelberg, Intel's Loihi etc. These attempts still fall short of our expectation on energy-efficient neuromorphic computing systems with online, real-time learning and inference capability. The bottlenecks of computation requirements, memory latency, and communication overhead continue to be showstoppers. Moreover, there is a lack of support in design automation of neuromorphic systems, including functionality verification, robustness evaluation and chip testing and debugging. Hardware innovation and electronic design automation (EDA) tools are required to enable energy-efficient and reliable hardware implementation for machine intelligence on cloud servers for extremely high performance as well as edge devices with severe power and area constraints.

The goal of the seminar was to bring together experts from different areas in order to present and to develop new ideas and concepts for emerging hardware techniques and EDA methodologies for neuromorphic computing. Topics that were discussed included:

- Neuroscience basics
- Physical fundamentals
- New devices and device modeling
- Circuit design and logic synthesis
- Architectural innovations
- Neurosynaptic processor and system integration
- Design automation techniques
- Simulation and emulation of neuromorphic systems
- Reliability and robustness
- Efficiency and scalability
- Hardware/software co-design
- Applications

The seminar facilitated greater interdisciplinary interactions between physicists, chip designers, architects, system engineers, and computer scientists. High-quality presentations and lively discussions were ensured by inviting carefully selected experts who participated in the seminar. All of them have established stellar reputations in the respective domains. As a result, we developed a better understanding of the respective areas, generated impetus for new research directions, and ideas for areas that will heavily influence research in the domain of neuromorphic design over the next years.

At the end of the seminar, we identified the following four areas as being among the most important topics for future research: *computing-in-memory*, *brain-inspired design and architecture*, *new technologies and devices*, and *reliability and robustness*. These research topics are certainly not restricted to and cannot be solved within one single domain. It is therefore imperative to foster interactions and collaborations across different areas.

2 Table of Contents

3 Overview of Talks

3.1 Challenges in Circuit Designs for Computing-in-Memory and Nonvolatile Logics for Edge Computing

Meng-Fan Chang (National Tsing Hua University – Hsinchu, TW)

Joint work of Meng-Fan Chang, CX Xue

Memory has proven to be a major bottleneck in the development of energy-efficient chips for artificial intelligence (AI). Recent memory devices not only serve as memory macros, but also enable the development of computing-in-memory (CIM) for IoT and AI chips. In this talk, we will review recent trend of Intelligent IoT and AI (AIoT) chips. Then, we will examine some of the challenges, circuits-devices-interaction, and recent progress involved in the further development of SRAM, emerging memory (STT-MRAM, ReRAM and PCM), nvLogics and CIMs for AIoT chips.

3.2 System-Level Design Methodology for Compute-in-Memory DNN Architecture

Chia-Lin Yang (National Taiwan University – Taipei, TW)

Joint work of Chia-Lin Yang, Hsiang-Yun Cheng, Tzu-Shien Yang, Meng-Yao Lin
Main reference Meng-Yao Lin, Hsiang-Yun Cheng, Wei-Ting Lin, Tzu-Hsien Yang, I-Ching Tseng, Chia-Lin Yang, Han-Wen Hu, Hung-Sheng Chang, Hsiang-Pang Li, Meng-Fan Chang: "DL-RSIM: a simulation framework to enable reliable ReRAM-based accelerators for deep learning", in Proc. of the International Conference on Computer-Aided Design, ICCAD 2018, San Diego, CA, USA, November 05-08, 2018, p. 31, ACM, 2018.
URL http://dx.doi.org/10.1145/3240765.3240800

Compute-in-memory (CIM) provides a promising solution to improve the energy efficiency of neuromorphic computing systems. Memristor-based crossbar architecture has gained a lot of attention recently. A few studies have shown the successful tape out of CIM ReRAM macros. However, how to integrate these macros together to handle large-scale DNN models still remains a challenge in terms of reliability and performance. In this talk, I will cover two main issues of the system-level design for Compute-in-Memory DNN architecture. First, I will introduce a simulation framework, DL-RSIM. DL-RSIM simulates the error rates of every sum-of-products computation in the memristor-based accelerator and injects the errors in the targeted TensorFlow-based neural network model. A rich set of reliability impact factors are explored by DL-RSIM, and it can be incorporated with any deep learning neural network implemented by TensorFlow. This tool enables design space exploration considering inference accuracy, performance and power tradeoffs. Second, I will present a new methodology for exploiting sparsity on a crossbar architecture. Existing sparsity solutions assume an over-idealized ReRAM crossbar architecture assuming an entire 128x128 or 256x256 crossbar array could be activated in a single cycle. However, due to power constraints and reliability issues, vector-matrix multiplication needs to be operated in a smaller granularity, called operation unit (OU), in practice. This finer granularity of computation presents a new opportunity for exploiting sparsity.

3.3 Neuromorphic computing architectures for IoT applications

Federico Corradi (Stichting IMEC Nederland – Eindhoven, NL)

Joint work of Federico Corradi, Giacomo Indiveri, Francky Catthoor
Main reference Corradi F. et al., "ECG-based Heartbeat Classification in Neuromorphic Hardware", IEEE International Joint Conference on Neural Networks (IJCNN), 2019.

Neuromorphic computing is a promising approach for developing new generations of smarter and adaptive computing technologies that can drastically change the way in which we think of computers, and in which we interact with artificial systems. In this approach, brain-inspired models of neural computations, based on massively parallel networks of low-power silicon neurons and synapses are implemented in electronic microchips [1]. These novel devices represent the third generation of neural networks which not only allows us to investigate at the theoretical level the possibilities of using time as a resource for computation and communication but can also shed light in the way in which our own brain works. This talk will describes recent efforts in building microchips and architectures of spiking neurons that can go beyond standard von Neumann machines [3]. In particular, I will showcase an ultra-low power architecture that target always-on wearable monitoring and other internet of things applications [2, 4].

References

1 G. Indiveri, F. Corradi, and N. Qiao. *Neuromorphic architectures for spiking deep neural networks*. IEEE International Electron Devices Meeting (IEDM), 2015.
2 F. Corradi, S. Pande, J. Stuijt, N. Qiao, S. Schaafsma, G. Indiveri, and F. Catthoor. *ECG-based Heartbeat Classification in Neuromorphic Hardware*. IEEE International Joint Conference on Neural Networks (IJCNN), 2019.
3 N. Qiao, M. Hesham, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri. *A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses*. Frontiers in neuroscience, Vol 9, 2015.
4 A. Balaji, F. Corradi, A. Das, S. Pande, S. Schaafsma, and F. Catthoor. *Power-Accuracy Trade-Offs for Heartbeat Classification on Neural Networks Hardware*. Journal of Low Power Electronics, Vol.14, 2018

3.4 Logic Synthesis for Hybrid CMOS-ReRAM Sequential Circuits

Rolf Drechsler (Universität Bremen, DE)

Joint work of Saman Froehlich, Saeideh Shirinzadeh, Rolf Drechsler
Main reference Saman Froehlich, Saeideh Shirinzadeh, Rolf Drechsler: "Logic Synthesis for Hybrid CMOS-ReRAM Sequential Circuits," IEEE Computer Society Annual Symposium on VLSI Miami, Florida, U.S.A., July 15-17, 2019.

Resistive Random Access Memory (ReRAM) is an emerging nonvolatile technology with high scalability and zero standby power which allows to perform logic primitives. ReRAM crossbar arrays combined with a CMOS substrate provide a wide range of benefits in logic synthesis. However, the application of ReRAM to sequential circuits has not been studied. We present a general synthesis approach for hybrid CMOS-ReRAM sequential architectures, which aims to minimize the CMOS overhead.We apply this general approach to different design methodologies, such as BDD-based design and AIGs. In the experiments we show that ReRAM allows for a significant reduction in CMOS size.

3.5 Cognitive Computing-in-Memory: Circuit to Algorithm

Deliang Fan (University of Central Florida – Orlando, US)

Main reference Shaahin Angizi, Jiao Sun, Wei Zhang, Deliang Fan: "GraphS: A Graph Processing Accelerator Leveraging SOT-MRAM", in Proc. of the Design, Automation & Test in Europe Conference & Exhibition, DATE 2019, Florence, Italy, March 25-29, 2019, pp. 378–383, IEEE, 2019.
URL https://doi.org/10.23919/DATE.2019.8715270

In-memory computing architecture is becoming a promising solution to reduce massive power hungry data traffic between computing and memory units, leading to significant improvement of entire system performance and energy efficiency. Meanwhile, emerging post-CMOS non-volatile memory, like Magnetic Random Access Memory (MRAM) or Resistive RAM (ReRAM), has been explored as next-generation memory technology that could be used not only as high performance memory with non-volatility, zero leakage power in un-accessed bit-cell, high integration density, but also as an efficient neuromorphic computing-in-memory platform. In this talk, Dr. Deliang Fan, from University of Central Florida, Orlando, FL, presents his recent synergistic research in cognitive computing-in-memory, including two main topics: 1) deep neural network (DNN) acceleration-in-memory system with both in-memory-logic design and hardware driven neural network optimization algorithm; 2) end-to-end ReRAM crossbar based neuromorphic computing EDA tool powered by a robust neural network mapping framework that adapts to existing non-ideal effects of crossbar structure.

For the first DNN computing-in-memory topic, it involves both the NVM based in-memory logic design and hardware aware DNN weight ternarization algorithm. First, in order to efficiently implement fast, reconfigurable in-memory logic without energy consuming intermediate data write back, two different logic-sense amplifier designs are presented. They could implement reconfigurable complete Boolean logic and full adder in only one cycle by simultaneously sensing two or three resistive memory cells in the same bit-line. Then a fully parallel multi-bit adder is presented. The related works are published in ASPDAC 2019[1], DATE 2019[2] and DAC 2019[3]. To further make the DNN algorithm intrinsically match with the developed in-memory logic platform, a DNN weight ternarization algorithm is introduced to fully ternarize all weight parameters into two states (i.e. +1, 0, -1), which brings three main benefits: 1) model size is compressed by 16X; 2) multiplication and accumulation based convolution operations are converted into adder only computation due to fully ternary weights. Different ternarization algorithms are published in WACV 2019[4] and CVPR 2019[5], both showing less than 2% top-1 accuracy drop when fully ternarizing DNN including the first layer and last layer.

For the second ReRAM crossbar based neuromorphic computing topic, a comprehensive framework called PytorX is introduced. It performs end-to-end (i.e. algorithm to device) training, mapping and evaluation for crossbar-based deep neural network accelerator, considering various non-ideal effects (e.g. Stuck-At-Fault (SAF), IR-drop, thermal noise, shot noise and random telegraph noise) of ReRAM crossbar when employing it as a dot-product engine. In particular, to overcome IR drop effects, a Noise Injection Adaption (NIA) methodology is introduced by incorporating statistics of current shift caused by IR drop in each crossbar as stochastic noise to DNN training algorithm, which could efficiently regularize DNN model to make it intrinsically adaptive to non-ideal ReRAM crossbar. It is a one-time training method without need of retraining for every specific crossbar. The related paper is published in DAC 2019[6] and PytroX code will be released in Github.

References

1 Shaahin Angizi, Zhezhi He and Deliang Fan. *ParaPIM: a parallel processing-in-memory accelerator for binary-weight deep neural networks*. ACM Proceedings of the 24th Asia and South Pacific Design Automation Conference, Tokyo, Japan, 2019

2 Shaahin Angizi, Jiao Sun, Wei Zhang and Deliang Fan. *GraphS: A Graph Processing Accelerator Leveraging SOT-MRAM*. Design, Automation and Test in Europe, March 25-29, 2019, Florence, Italy

3 Shaahin Angizi, Jiao Sun, Wei Zhang and Deliang Fan. *AlignS: A Processing-In-Memory Accelerator for DNA Short Read Alignment Leveraging SOT-MRAM*. Design Automation Conference (DAC), June 2-6, 2019, Las Vegas, NV, USA

4 Zhezhi He, Boqing Gong, Deliang Fan. *Optimize Deep Convolutional Neural Network with Ternarized Weights and High Accuracy*. IEEE Winter Conference on Applications of Computer Vision, January 7-11, 2019, Hawaii, USA

5 Zhezhi He and Deliang Fan. *Simultaneously Optimizing Weight and Quantizer of Ternary Neural Network using Truncated Gaussian Approximation*. Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, CA, USA

6 Zhezhi He, Jie Lin, Rickard Ewetz, Jiann-Shiun Yuan and Deliang Fan. *Noise Injection Adaption: End-to-End ReRAM Crossbar Non-ideal Effect Adaption for Neural Network Mapping*. Design Automation Conference (DAC), June 2-6, 2019, Las Vegas, NV, USA

3.6 Scaling-up analog neural networks – practical design considerations

Alex Pappachen James (Nazarbayev University – Astana, KZ)

Main reference Olga Krestinskaya, Aidana Irmanova, Alex Pappachen James: "Memristive Non-Idealities: Is there any Practical Implications for Designing Neural Network Chips?", in Proc. of the IEEE International Symposium on Circuits and Systems, ISCAS 2019, Sapporo, Japan, May 26-29, 2019, pp. 1–5, IEEE, 2019.
URL https://doi.org/10.1109/ISCAS.2019.8702245

The implementation of analog neural networks in a memristive crossbar faces several challenges. The impact of non-idealistic behavior of devices and network components poses challenges to designing specifications for neural networks. The performance and reliability analysis of memristive neural networks involve a careful look into the device to device variability, signal integrity issues, signal noise issues, thermal issues, packaging issues, device aging, line resistors, device faults, and process variation. The scaling of analog neural networks are limited by such non-idealities [2], and to counter this the performance analysis needs to be mapped and benchmarked with the degradations from non-idealities. The analog networks are ideally suitable for near-sensor edge AI computing[3], where the signal degradation from the source is limited, and where the information processing can be done in a near-continuous mode of operation.

The memristor crossbar as a dot product engine for analog neural networks can be used for achiving low power and low area density for different types of networks such as hierarchical temporal memories, convolutional neural networks and binary neural networks [1]. In analog neural network implementations, the mapping of weights to specific resistance levels in memristor is challenging over a period of time due to physical stress on the devices over continuous read-write cycles. Even if it is programmed to a specific level, we notice that the number of levels within a memristor is an important factor in determining the robustness of the analog neural network towards performance indicators such as recognition

accuracy in a classification task [2]. It is important to also take into account the variability of conductance states between the devices within a crossbar along with non-idealities of interfacing circuits in the system level performance evalution and design of analog neural networks to ensure reliability of a given system application.

References

1 A. P. James. *Deep Learning Classifiers with Memristive Networks: Theory and Applications*. Springer, 2019.
2 O. Krestinskaya, A. Irmanova, and A. P. James. Memristive non-idealities: Is there any practical implications for designing neural network chips? In *IEEE International Symposium on Circuits and Systems*, 2019.
3 O. Krestinskaya, A. P. James, and L. O. Chua. Neuromemristive circuits for edge computing: A review. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–20, 2019.

3.7 Resource-aware machine learning and data mining

Jian-Jia Chen (TU Dortmund, DE)

Joint work of Sebastian Buschjäger, Kuan-Hsun Chen, Jian-Jia Chen, Katharina Morik
Main reference Sebastian Buschjäger, Kuan-Hsun Chen, Jian-Jia Chen, Katharina Morik: "Realization of Random Forest for Real-Time Evaluation through Tree Framing", in Proc. of the IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018, pp. 19–28, IEEE Computer Society, 2018.
URL http://dx.doi.org/10.1109/ICDM.2018.00017

The performance of machine learning and data mining algorithms are constrained by the underlying computing platforms. On one hand, the platform defines how fast a machine learning algorithm can be executed; on the other hand, the machine learning algorithm should be adaptive to be efficiently executed on different platforms. In this talk, I will summarize my perspectives of such resource-aware machine learning and data mining algorithms. Specifically, I will address by using simple decision trees or random forests, that are widely used in many applications. We show that effective configurations of cache layout can improve the performance of decision trees significantly. We also show that tuning hyper-parameters in machine learning algorithms can be constrained by the resources and demonstrate how to use model-based optimization (MBO) to handle such configurations in a resource-efficient manner.

3.8 Turing, or Non-Turing? That is the question

Johannes Schemmel (Universität Heidelberg, DE)

Neuromorphic computing, as a realization of Non-Turing, in-memory, event-based computing, will allow us to overcome the power wall our CPU-centric CMOS technology is facing. But that does not mean that the era of Turing-based computing will come to an end soon, or that Turing-based computing does not have its place in the neuromorphic world.

Our work in the area of neuromorphic computing is based on the premise that the biological brain is the ultimate cognitive system, especially considering its low power consumption of approximately 20 Watts. Each of the billions of neurons constituting the brain receives signals from thousands of others. Thereby forming a dense network of interconnects. This exchange of information happens at special sites of transmission, the synapses. The receiving neuron integrates its synaptic inputs continuously over space and time, thereby computing without a central clock or any global addressing schemes. Temporal and spatial correlations determine the dynamics of the system, and the synapse becomes the central point of learning. Any correlation depends on the interconnection schemes between the neurons, which is determined by the growth and development of the synaptic connectome.

Any theories we derive from experimental observations of this cellular interplay can be formulated in the language of mathematics. This leads to incredible complex systems of billions of coupled differential equations. Numerical simulations provide most of the insight we can get from these descriptions of biology. Numerical simulations including the temporal evolution of the synaptic connectome are currently only feasible for small networks. The power-wall our current CMOS technology is facing gives us not much hope for the future, as long as we rely on contemporary computing architectures.

Dedicated accelerators are one possible solution. They can improve the power-efficiency of these kind of simulations by several orders of magnitude. One special kind of accelerator are the physical model systems we presented in this talk: they solve the differential equations by emulating their dynamics in analog CMOS circuits instead of numerical calculations.

The state of the system is represented by physical quantities like voltage, current and charges instead of binary numbers. This approach was first published by Carver Mead in the 1980ies. The implementation we present is different from this initial work by biasing the transistors near or even within strong inversion instead of deep sub-threshold, utilizing transistors closer to their minimum size. This allows a higher component density and, due to the strongly reduced mismatch, a better calibratability and repeatability of the neuron and synapse parameters.

The natural time constants at these higher bias points lead to time constants several orders of magnitude shorter than in biology. This effectively accelerates the emulation of the underlying model, therefore the term "accelerated analog neuromorphic hardware" for our kind of neuromorphic systems.

The BrainScaleS-1 (BSS-1) system uses these principles to create highly connected systems from multiple microchips, each implementing 512 neurons and 114k synapses. The BSS-1 system is usually trained by hardware-in-the-loop methods, using an external compute cluster to calculate the synaptic connectome.

The analog emulation of a neural network, running at 10000 times the speed of its biological example within a BSS-1 system formed by more than 400 interconnected chips, realizes one of the very few implementations of true non-Turing computing worldwide. Alternatively, this program-free mode of computation is sometimes called non-Von-Neumann computing. Looking more closely, Turing-based computation is still utilized for a multiple of purposes in the BSS-1 system: training, system initialization, hardware calibration, runtime control and the handling of all input and output data is performed by a complex, multi-tiered software stack.

To create a self-sufficient, scalable neuromorphic system that has a minimized communication demand with any external hardware, all of this functions have to move inside of the system, similar to the self-sufficiency we observe in biology. In other words and to stay withing the analogy: we have to move everything besides the simple synapse and neuron models into

our microchips. Functions performed by the Astrocytes, the cell-nuclei, the intercellular, dendritic and axonal transport systems as well as the peripheral nerves connecting the brain with its body.

This will not be achievable by analog, dedicated hardware in the foreseeable future. Mostly, because we do not know enough details of these processes yet. Our hardware systems need a lot of flexibility regarding the implementation of theses features. With the second generation BrainScaleS architecture, BrainScaleS-2, we devised a solution to this issues, originally called "hybrid plasticity". The key aspect of which is the demotion of the analog neuromorphic part from the center of the system to the role of a co-processor.

The BSS-2 architecture is based on a high-bandwidth link between an SIMD microprocessor and an accelerated analog neuromorphic core, where the CPU takes over every role not covered by the neuromorphic core.

This Turing-non-Turing hybrid allows the realization of on-chip learning, calibration, initialization and environmental simulation. The accelerated analog core emulates the network in continuous time and performs all performance and energy intensive operations, like measuring spatial and temporal correlations or solving the differential equations that govern the membrane voltages. The network will always perform without any intervention from the CPU. But it will be a static system, where no parameters change anymore. The dynamic adjustment of all parameters, from synaptic weights, network topology to neuron parameters, is controlled by the CPU, allowing the full flexibility of software-defined algorithms for the implementation of learning algorithms. Since the CPU does not need to look at every individual spike, the learning speed is also accelerated, usually by the same factor as the network time constants. Therefore, learning processes need only seconds instead of tens of minutes.

This talk explains the Heidelberg BrainScaleS-2 accelerated analog neuromorphic architecture and demonstrates how it balances Turing and Non-Turing computing to combine power efficiency with the necessary flexibility and programmability.

References

1 Friedmann, Simon and Schemmel, Johannes and Grübl, Andreas and Hartel, Andreas and Hock, Matthias and Meier, Karlheinz. *Demonstrating hybrid learning in a flexible neuromorphic hardware system*. IEEE transactions on biomedical circuits and systems, 2017

2 Schemmel, Johannes and Kriener, Laura and Müller, Paul and Meier, Karlheinz. *An accelerated analog neuromorphic hardware system emulating NMDA-and calcium-based non-linear dendrites* 2017 International Joint Conference on Neural Networks (IJCNN)

3.9 Reliability / Robustness of Neuromorphic Computing Architectures – Has the time come yet?

Bing Li (TU München, DE) and Ulf Schlichtmann (TU München, DE)

Joint work of Shuhang Zhang, Grace Li Zhang, Bing Li, Hai (Helen) Li, Ulf Schlichtmann
Main reference Shuhang Zhang, Grace Li Zhang, Bing Li, Hai Helen Li, Ulf Schlichtmann: "Aging-aware Lifetime Enhancement for Memristor-based Neuromorphic Computing", in Proc. of the Design, Automation & Test in Europe Conference & Exhibition, DATE 2019, Florence, Italy, March 25-29, 2019, pp. 1751–1756, IEEE, 2019.
URL https://doi.org/10.23919/DATE.2019.8714954

Neuromorphic Computing is a rapidly emerging, very promising area of computing. Understandably, research focuses on determining the most efficient architectures, circuit design concepts etc. But as we have learned from traditional CMOS technology, reliability and robustness are very important areas of concern [2, 3, 4, 5]. For instance, memristors can only be programmed reliably for a given number of times. Afterwards, the working ranges of the memristors deviate from the fresh state. As a result, the weights of the corresponding neural networks cannot be implemented correctly and the classification accuracy drops significantly. To counter this effect, software training and hardware mapping can be combined to extend the lifetime of memristor crossbars up to 11 times, while the accuracy of classification is maintained [1].

Broadly speaking, neuromorphic computing schemes with CMOS-based devices and emerging technologies all face reliability and robustness challenges, specially when the feature size of the manufacturing technology is reduced further to achieve a higher integration density. In "classical", von-Neumann-based computing, analysis and optimization of reliability and robustness has already become as important as considering area, performance and power. How about neuromorphic computing? Since AI based on neuromorphic computing is often intended to be used significantly in safey-critical applications such as autonomous driving, has now already the moment come to pose the question – has the time come already now to worry about reliability and robustness of neuromorphic computing?

References

1 Shuhang Zhang, Grace Li Zhang, Bing Li, Hai (Helen) Li, Ulf Schlichtmann. *Aging-aware Lifetime Enhancement for Memristor-based Neuromorphic Computing*. Design, Automation and Test in Europe (DATE), March 2019
2 Dominik Lorenz, Martin Barke, Ulf Schlichtmann. *Efficiently analyzing the impact of aging effects on large integrated circuits*. Microelectronics Reliability 52 (8), 1546-1552, 2012
3 Dominik Lorenz, Martin Barke, Ulf Schlichtmann. *Aging analysis at gate and macro cell level*. International Conference on Computer-Aided Design (ICCAD), November 2010
4 Dominik Lorenz, Martin Barke, Ulf Schlichtmann. *Monitoring of aging in integrated circuits by identifying possible critical paths*. Microelectronics Reliability 54 (6-7), 1075-1082, 2014
5 Veit B Kleeberger, Christina Gimmler-Dumont, Christian Weis, Andreas Herkersdorf, Daniel Mueller-Gritschneder, Sani R Nassif, Ulf Schlichtmann, Norbert Wehn. *A cross-layer technology-based study of how memory errors impact system resilience*. IEEE Micro 33 (4), 46-55, 2013

3.10 Memory Device Modeling for the Simulation of Neuromorphic Systems

Darsen Lu (National Cheng Kung University – Tainan, TW)

With the recent advancement in the internet and mobility, which generated massive amounts of data, deep learning has become a powerful tool in creating domain-specific intelligence, and has found application in many areas. Hardware acceleration of deep learning becomes essential given the large computational power requirement, especially during training. Analog neuromorphic computation may be the ultimate (best) hardware for deep learning given its significantly lower power compared to the digital counterpart.

Analog neuromorphic computation often uses non-volatile memory devices as basic element (synapse). We have developed, in collaboration with our colleagues at NCKU, compact models for ferroelectric memory, resistive memory, phase change memory, and flash and have used them towards neuromorphic circuit applications.

To translate compact-model-based device characteristics to estimated system-level performance such as recognition rate, power consumption, and circuit speed, a simulation platform is required. We have developed simNemo, a platform which predicts deep learning performance for given memory device characteristics, circuit architecture, and neural network model. So far we have focused on modeling resistive RAM devices in MLP neural networks, and have using the MNIST database for benchmark.

Through the process of building and using simNemo, we have learned the challenges of analog neuromorphic computing and its EDA methodologies. First, even though device-to-device variation can be compensated by weight adjustment during traning to some extent, there is a limit. If the variation is too large, it can limit the synaptic degree of freedom such as allowing only positive or negative weight. Second, cycle-to-cycle variation, which in many cases is the fundamental property of a memory device, cannot be compensated by training. However, it can be mitigated by special programming technique. Take the memristor/RRAM for example, we may program the device with large current so that it achieves a very low resistivity, and the impact of cycle-to-cycle variation due to a single trapping or de-trapping event becomes minimal. Third, linearity and symmetry are very important properties for memory devices used for neuromorphic applications. Symmetry is more important, as non-symmetric positive versus negative weight updates may even cause training algorithm to fail by introducing bias towards a certain direction, so that the algorithm never converges to the optimal point. On the other hand, closed-loop programming is one way to overcome this issue. Forth, endurance is a challenge for online training since the memory device needs to be updated many times. The use or larger batch size helps. Fifth, analog-to-digital conversion always introduce quantization error. One solution is digital-assisted training of analog neuromorphic systems by storing weight update residues in a digital form during training. Finally, analog summing of current poses the dynamic range challenge as each wire has certain noise floor. This can be solved by the integrate-and-fire circuitry at the output to effectively increase the dynamic range of current sum.

3.11 Processing Data Where It Makes Sense in Modern Computing Systems: Enabling In-Memory Computation

Onur Mutlu (ETH Zürich, CH)

Main reference Onur Mutlu, Saugata Ghose, Juan Gómez-Luna, Rachata Ausavarungnirun: "Processing Data Where It Makes Sense: Enabling In-Memory Computation", CoRR, Vol. abs/1903.03988, 2019.
URL https://arxiv.org/abs/1903.03988

Today's systems are overwhelmingly designed to move data to computation. This design choice goes directly against at least three key trends in systems that cause performance, scalability and energy bottlenecks: 1) data access from memory is already a key bottleneck as applications become more data-intensive and memory bandwidth and energy do not scale well, 2) energy consumption is a key constraint in especially mobile and server systems, 3) data movement is very expensive in terms of bandwidth, energy and latency, much more so than computation. These trends are especially severely-felt in the data-intensive server and energy-constrained mobile systems of today.

At the same time, conventional memory technology is facing many scaling challenges in terms of reliability, energy, and performance. As a result, memory system architects are open to organizing memory in different ways and making it more intelligent, at the expense of slightly higher cost. The emergence of 3D-stacked memory plus logic, the adoption of error correcting codes inside the latest DRAM chips, and intelligent memory controllers to solve the RowHammer problem are an evidence of this trend.

In this talk, I will discuss some recent research that aims to practically enable computation close to data. After motivating trends in applications as well as technology, we will discuss at least two promising directions: 1) performing massively-parallel bulk operations in memory by exploiting the analog operational properties of DRAM, with low-cost changes, 2) exploiting the logic layer in 3D-stacked memory technology in various ways to accelerate important data-intensive applications. In both approaches, we will discuss relevant cross-layer research, design, and adoption challenges in devices, architecture, systems, applications, and programming models. Our focus will be the development of in-memory processing designs that can be adopted in real computing platforms and real data-intensive applications, spanning machine learning, graph processing, data analytics, and genome analysis, at low cost. If time permits, we will also discuss and describe simulation and evaluation infrastructures that can enable exciting and forward-looking research in future memory systems, including Ramulator and SoftMC.

References

1 Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, Rachata Ausavarungnirun, *Processing data where it makes sense: Enabling in-memory computation*. Journal of Microprocessors and Microsystems, Volume 67, Pages 28-41, June 2019.

3.12 Low Power Embedded Machine Intelligence on a Neurosynaptic Processor

Qinru Qiu (Syracuse University, US)

Spiking neural networks are rapidly gaining popularity for their brain-inspired architecture and operation. It has the potential to achieve low cost, high noise resilience, and high energy efficiency due to the distributed nature of neural computation and the use of low energy spikes for information exchange. This talk introduces our work on applying the neurosynaptic processor, such as IBM TrueNorth, to implement different neural networks. How distributed learning can potentially be achieved in a stochastic spiking neural network will also be discussed. It is the first work to implement not only feedforward networks, but also recurrent networks on the spiking based neurosynaptic processor. It has the potential to enable energy efficient implementation of a wider range of applications and very low cost, in hardware learning.

3.13 Hardware and Software for Spike-based Memristive Neuromorphic Computing Systems

Garrett S. Rose (University of Tennessee, US)

Main reference James S. Plank, Catherine D. Schuman, Grant Bruer, Mark E. Dean, and Garrett S. Rose, "The TENNLab Exploratory Neuromorphic Computing Framework," IEEE Letters of the Computer Society, vol. 1, pp. 17–20, July-Dec. 2018.
URL https://doi.ieeecomputersociety.org/10.1109/LOCS.2018.2885976

Given the slow down in Moore's Law scaling and limits in classic von Neumann computer architectures, spike-based neuromorphic computing has gained increasing interest for a range of potential applications. Here we consider spiky neuromorphic systems where neurons are implemented using analog/mixed-signal CMOS and synapses are built from memristors (or "memory resistors"). The specific neuromorphic framework presented also allows for recurrent pathways in the networks constructed, a feature that is particularly useful for stateful neuromorphic processing. One such spiky recurrent neuromorphic system is mrDANNA (Memristive Dynamic Adaptive Neural Network Array), a hybrid CMOS/memristor implementation from the University of Tennessee. A full mrDANNA system offers the potential to construct small neural network applications operating on a fraction of the power consumed by analogous deep learning systems.

3.14 DNN on FPGA and RRAM

Yu Wang (Tsinghua University – Beijing, CN)

URL https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

Artificial neural networks, which dominate artificial intelligence applications such as object recognition and speech recognition, are in evolution. To apply neural networks to wider applications, customized hardware are necessary since CPU and GPU are not efficient enough. FPGA can be an ideal platform for neural network acceleration (inference part) since it is programmable and can achieve much higher energy efficiency compared with general-purpose processors. However, the long development period and insufficient performance of traditional FPGA acceleration prevent it from wide utilization.

We propose a complete design flow to achieve both fast deployment and high energy efficiency for accelerating neural networks on FPGA [FPGA 16/17]. Deep compression and data quantization are employed to exploit the redundancy in algorithm and reduce both computational and memory complexity. Two architecture designs for CNN and DNN/RNN are proposed together with compilation environment. Evaluated on Xilinx Zynq 7000 and Kintex Ultrascale series FPGA with real-world neural networks, up to 15 times higher energy efficiency can be achieved compared with mobile GPU and desktop GPU.

We talk about why we start Deephi (DL is a uniform respresentation of $y = f(x)$). We also show the collective figure about all the current inference accelerator, and point out the 1-10TOPS/W limitation by the current technology. Meanwhile it is worth to think about where the computing is happening: cloud + edge + end devices, and the application domains defines the AI chips.

Another important perspective is the software toolchain/compiler is very important AI chips. We talk about the recent compiler design for CPU/GPU/FPGA, and try to think about the right tool for neurotrophic computing. We will discuss the possibilities and trends of adopting emerging NVM technology for efficient learning systems to further improve the energy efficiency. We would like to call for some kind of collaborative effort on the open software design for Neuromorphic computing or the computing in memory research.

Participants

- Krishnendu Chakrabarty
Duke University – Durham, US
- Meng-Fan Chang
National Tsing Hua University – Hsinchu, TW
- Jian-Jia Chen
TU Dortmund, DE
- Yiran Chen
Duke University – Durham, US
- Federico Corradi
Stichting IMEC Nederland – Eindhoven, NL
- Rolf Drechsler
Universität Bremen, DE
- Deliang Fan
University of Central Florida – Orlando, US
- Tsung-Yi Ho
National Tsing Hua University – Hsinchu, TW
- Alex Pappachen James
Nazarbayev University – Astana, KZ
- Bing Li
TU München, DE
- Hai Li
Duke University – Durham, US
- Darsen Lu
National Cheng Kung University – Tainan, TW
- Christoph Maier
University of California – San Diego, US
- Onur Mutlu
ETH Zürich, CH
- Qinru Qiu
Syracuse University, US
- Garrett S. Rose
University of Tennessee, US
- Yulia Sandamirskaya
Universität Zürich, CH
- Johannes Schemmel
Universität Heidelberg, DE
- Ulf Schlichtmann
TU München, DE
- Yu Wang
Tsinghua University – Beijing, CN
- Chia-Lin Yang
National Taiwan University – Taipei, TW