4th IBM IEEE CAS/EDS AI Compute Symposium (AICS’21)

The 4th IBM IEEE CAS/EDS AI Compute Symposium, known as (AICS’21), was held over two days (Oct 13–Oct 14, 2021). The event was very well attended and received great responses from the audience all over the world. The symposium was also an initiative supported by IBM Academy of Technology (https://www.ibm.com/blogs/academy-of-technology/). Dr. Joshi has been the main interface for CAS and EDS for organizing this successful event. This is the second time the event was organized as a virtual symposium. The audio/video presentation on this virtual platform went smoothly. **More than 2400 viewers over two days, participation from 50 countries, over 54 student posters, best paper poster awards, excellent panel discussions, 11 distinguished speakers from industry and academia were the salient features of this symposium. There were more than 5200 views on the LinkedIn post about the symposium. The theme of the symposium was “From Ground up to Cloud”. In short, the symposium covered a range of topics from device technology, to circuits, architecture, algorithms and sustainability—to make innovations for the cloud with an emphasis on green AI.**

Prof. Hoi-Jun Yoo (Professor of School of Electrical Engineering and the director of the System Design at KAIST, Korea) opened the symposium with his excellent presentation related to “Training on Chip—Next Wave of Mobile AI Accelerators”. Most mobile Deep Neural Network (DNN) accelerators target only inference of DNN models on edge devices, whereas on-device training was out of reach in mobile platforms due to its excessive computational requirements. Training-on-Chip (ToC) with user-specific data is becoming more important than ever because of privacy issues and communication latency of training on remote servers. He highlighted a number of approaches in realizing ToC. General purpose hardware and software co-optimization techniques aiming to maximize throughput and energy-efficiency of DNN training were brought out with examples, such as, sparsity exploitation and bit-precision optimization for training. In addition, application specific training accelerators for Deep Reinforcement Learning (DRL) and Generative Adversarial Network (GAN) were discussed, touching on issues regarding system implementation with the fabricated silicon.

Dr. Teo Laino (IBM Distinguished Research Staff Member) followed up with a very interesting talk about “A Cloud-based AI-driven Autonomous Lab”. One of the most significant outcomes of chemistry is the design and production of new molecules. The application of domain knowledge accumulated over decades of laboratory experience has been critical in the synthesis of many new molecular structures. Nonetheless, most synthetic success stories are accompanied by long hours of repetitive synthesis. Automation systems were created less than 20 years ago to assist chemists with repetitive laboratory tasks. While this has proven to be very effective in a few areas, such as, high-throughput chemistry, the use of automation for general-purpose tasks remains a tremendous challenge even today. Automation necessitates that chemistry operators write different software for different tasks, each of which codifies a specific and distinct type of chemistry. Meanwhile, in organic chemistry, Artificial Intelligence (AI) has emerged as a valuable complement to human knowledge for tasks such as predicting chemical reactions, retrosynthetic routes, and digitizing chemical literature. Dr. Laino’s talk highlighted the first cloud-based AI-driven autonomous laboratory implementation. The AI assists remote chemists with a variety of tasks, including designing retrosynthetic trees and recommending the correct sequence of operational actions (reaction conditions and procedures) or ingesting synthetic procedures from literature and converting them into an executable program. The AI self-programs the automation layer and makes decisions on synthesis execution using feedback loops from analytical chemistry instruments, with supervision from synthetic chemists. He presented the AI core technology and how it performs across different types of synthetic tasks.

Subsequently, Mr. Gunnar Hellekson (Vice President at Red Hat) gave an exciting talk about an open approach to AI and its integration into cloud. Business innovation is driven by big ideas, moving faster than ever before. Today, we can do things we could only dream of a few years ago. Massive global changes are shifting the

Digital Object Identifier 10.1109/MCAS.2022.3189894
Date of current version: 5 September 2022

way people live and work and require organizations to rethink their teams, processes, and technologies to stay competitive. Today, organizations across all geographies and industries can innovate, create more customer value and differentiation and compete on an equal playing field. This new reality demands that enterprises embrace digital transformation and pivot quickly or fail. AI (Artificial Intelligence) is a critical part of the digital transformation journey for many organizations. Smart cities, wearable health technologies, smart energy grids, autonomous vehicles, manufacturing, and agriculture are just some of the key markets being transformed by AI. He pointed out that how technology, open source communities and new ways of collaborating are driving business innovations like AI. AI investments across every industry are accelerating to develop differentiated services and gain competitive advantages. The difference between complex hardware and software for key elements like security, data is diminishing. Many businesses are aware of the benefits, but there are a number of challenges delaying their implementation plans. This is the natural home of open source, providing the components and foundation, and creating a space for innovators to come together and share their great ideas in a way that's self-sustainable.

Next Dr. Evgeni Gousev (Senior Director at Qualcomm) gave a wonderful overview of tinyML: enabling ultra-low power machine learning at the very edge. In his talk, he covered many aspects such as tinyML fundamentals, its markets and values, and gave many examples. He discussed recent developments from Qualcomm and provided information about the tinyML foundation, ecosystem, projects and events and educational activities. Dr. Gousev further defined tinyML as machine learning architectures, techniques, tools, and approaches capable of performing on-device analytics for variety of sensing modalities (vision, audio, motion, chemicals, etc.) at mW the power range or below, targeting predominantly battery-operated devices. The key tinyML growth drivers include more efficient hardware, energy efficient algorithms, more mature software infrastructure, tools, diverse ecosystems, growing the number of applications, corporate investment, VC investment, and increased start-ups. It is predicted that 1B tinyML devices would be shipped in 2024 and would approach 5.4B in 2026. The growth is in double digits. He described many useful applications of tinyML, such as, voice recognition, environmental sensors, predictive maintenance, gesture control, and augmented reality.

The final talk on the first day was given by Dr. Venkat Thanvantri (VP of Machine Learning R&D at Cadence). He presented advances in AI/ML for chip design. The emergence of machine learning (ML) has unlocked many new applications and transformed user experiences. Electronics Design Automation (EDA) is an application area that delivers value by providing automation and abstraction. ML technology is having a similarly transformative impact on EDA, accelerating execution of algorithms, improving quality of results, and now significantly improving the productivity of users. In particular, he detailed the use of Reinforcement Learning to automate and optimize results for the digital design and signoff flow.

On the second day of the symposium, Dr. Suk Hwan Lim (Executive VP at Samsung) addressed fine grained domain specific architectures for diverse workloads. He covered deep learning applications and workloads, domain specific neural processing units/accelerators (NPU), and future directions for NPUs. There are several deep learning (DL) applications – speech recognition, voice activation, text to speech, authentication, image classification, object detection, semantic segmentation and image processing, among others. Training is typically implemented in the cloud, while inference is deployed on the edge. The compute complexity varies by 5–7 orders of magnitude for these applications. Thus, he categorized fine grained diverse processors for diverse applications—micro NPU for audio/always-on applications, general NPUs for small spatial resolutions with deep network, and image processing NPUs for large special resolutions with shallow networks. Significant improvements in energy, area, efficiency and utilization are needed and can be achieved through algorithms/compiler/architecture/circuits.

Next, Prof. Song Han (Assistant Professor at MIT) described the role of tinyML and how greener AI can be achieved. TinyML and efficient deep learning make AI greener and easily deployable to IoT. AI applications can generate high power and have detrimental affects on the environment. There is a push for data compression, pruning, and other techniques to reduce computation and thereby power. New models can be developed to improve latency and accuracy. Manual design is challenging and automation is needed. Prof. Han has proposed and developed a hardware-aware neural network search called "Once-for-All". While computationally expensive, it requires training only once and then produces multiple models which can be used for inference. This approach reduces the data usage and results in less computation and hence a lower carbon footprint. Also, sparse attention and progressive quantization ideas are used to prune the tokens in Natural Language Processing (NLP). Activation is the main bottleneck and not the trainable parameter space. Activation minimization leads to significant memory reduction in the IoT. All the techniques described in his talk can help AI to be greener.

Following Prof. Han's talk, Prof. H.-S. Philip Wong (Willard R. and Inez Kerr Bell Professor, Stanford University) presented "N3XT-3D-MOSAIC: Domain-Specific Technology for AI Compute", noting that 21st century applications are going to be data-centric. Data analytics, machine learning, and AI applications are going to dominate, from data center to mobile and IoT, from collecting and processing, to curating the data, to deriving information. Many systems will need to learn and adapt on the fly. Three-dimensional integration is one of the major technology directions for integrated circuits. Prof. Wong gave an overview of the new materials and device technologies that may need to be developed to realize monolithic 3D integration with multiple logic transistor and memory device layers. He gave two examples of compute-in-memory chips that feature RRAM integration with CMOS logic as an illustration of how future 3D systems may be designed.

Dr. Steve Pawlowski (VP at Micron Technology) next talked about rethinking the memory-compute subsystem. The best machines struggle on workloads requiring higher memory and network performance. The challenge remains to improve overall systems design starting with memory. Dr. Pawlowski described evolving usage models driving compute demands that require vast amounts of data. He suggested current system bottlenecks will get worse unless something is done. New memory-compute architectures will have the greatest impact on systems. However, new architectures will take time to evolve. Market driven use cases can speed adoption. Integration of compute into memory to get greatest energy and performance is key. New devices must build from and evolve the dominant software ecosystem. Many future neural networks will require higher bandwidths, which increases energy consumption. Memory energy is interconnect dominated. Memory bandwidth is also pin and locality dominated. A closer coupling of memory and compute is the path forward.

Subsequently, Dr. Pradeep Dubey (Senior Fellow at Intel) gave a great talk about the "Era of Ubiquitous AI". Artificial intelligence (AI) is touching, if not transforming, every aspect of our lives. AI is impacting not just what computing can do for us, rather how computing gets done. Fast-evolving AI algorithms are driving demand for general-purpose computing that cannot be met by "business as usual" engineering. At the same time, programmers are often data scientists, not computer scientists; expecting programmers to figure out increasingly complex hardware on their own just doesn't work. Architects are therefore needed more than ever—chip architects to create new processors, systems architects to design new data centers, software architects to design new frameworks, and AI architects to churn out new models and new algorithms. Are we up to the task? Or do we need to augment human architects with AI to meet the challenge?

The symposium's concluding talk was given by Tamar Eilam (IBM Fellow). She presented IBM's initiative related to sustainable and responsible computing. IBM traditionally has focused on privacy of data, security, and ethics. Another key consideration is reducing carbon footprint (CF). For example, some AI jobs consume CF equivalent to that of the lifetime of 5 cars. It is predicted that electricity usage in the data center would increase by 8% by the year 2030. CF is driving AI to move to cloud. It is crucial to quantify Carbon Footprint Energy (CFE), which is defined as product of IT equipment energy, power usage effectiveness (overhead power conversion and cooling), and source of the energy (coal, nuclear, etc.). Organizations need to report CFE to discover hot spots and optimize. To make greener AI, the use of renewable energy, controlling the data utilization, and finding alternate ways to reduce power are essential. In order to do this, workloads need to be monitored, enabling a breakdown of CF by cloud tenant and application.

The symposium also featured a poster session, organized into 3 parallel tracks. Out of 54 posters, the top 3 best posters awarded from each track. The list of winners is given on the symposium website: https://www.zurich.ibm.com/thinklab/AIcomputesymposium.html

The symposium closed with a panel discussion on Responsible Computing, with four distinguished panelists including Tamar Eilam (IBM), Irena Risch (McGill), Evgeni Gousev (Qualcomm), and Bouchra Bouqata (Amazon). Panelists addressed questions spanning a range of controversial topics, including the environmental impact of data center workloads, building trust in AI, eliminating bias in AI, and ensuring data privacy.

Replays of entire two-day symposium are available on the symposium website: https://www.zurich.ibm.com/thinklab/AIcomputesymposium.html

Dr. Rajiv Joshi, IEEE Life Fellow
Dr. Arvind Kumar, IEEE Member
Dr. Matt Ziegler, IEEE Member
Affiliation—T. J. Watson Research Center,
Yorktown Heights, NY 10598
Executive Sponsor—Dr. Mukesh Khare
AoT Sponsor—Dr. John "Boz" Handy Bosma
Committee—Rajiv Joshi, Matt Ziegler, Arvind Kumar, Xin Zhang, Krishnan Kailas, Kaoutar El Maghraoui, Jin-Ping Han, Anna Topol, John Rozen