CASS Conference Highlights

Rajiv Joshi, *Fellow, IEEE*, Matthew Ziegler, *Senior Member, IEEE*, Jin-Ping Han, and Kaoutar El Maghraoui

6th IBM IEEE CAS/EDS AI Compute Symposium (AICS'23)

The 6th IBM IEEE CAS/EDS AI Compute Symposium was held hybrid at the T. J. Watson Research Center on 28 November 2023. The event was extremely successful and well attended by over 2000 folks from all over the world (in-person and virtual). The symposium featured 8 distinguished speakers (7 from industry and 1 from academia), over 30 student in-person posters, best poster awards, and a panel discussion. The registration list spanned citizens of 53 countries. The theme of the symposium, "From Chips to Chiplets," turned out to be an opportune and important topic for the current semiconductor industry direction. The symposium served as an educational as well as a brainstorming session for industry/academia/students across the world. The symposium covered a range of topics from emerging device technology, innovative circuits, chip and chiplet architecture, advanced packaging technologies, such as 2D to 3D packaging elements, and how these topics drive the rapid growth of AI and generative AI. Dr. Rajiv Joshi, General Chair and IEEE Life Fellow opened the symposium with welcoming remarks along with the goals and accomplishments of this symposium under the auspices of CAS and IBM.

Huiming Bu, VP of Global Semiconductors R&D and Albany Operations, IBM gave the first keynote talk—**"Path to 1 Trillion Transistors in the Era of AI"** (Fig. 1, left). He discussed how artificial intelligence is transforming our world and the demand for computing power is increasing at an unprecedented pace. Then he showed how semiconductor technology innovations in materials, transistors, interconnects, chip architectures, and advanced packaging are being pursued to meet this demand. He emphasized technology development in the semiconductor ecosystem is the key ingredient to making this happen, He shared the roadmap, challenges, and enablers of IBM research's Artificial Intelligence Unit (AIU) Chiplet and advanced packaging technologies are vital to enable the next generation of AIU.

Rob Aitken, Distinguished Architect, Synopsys gave a vivid talk on **"Impedance Matching AI, EDA, Chips, and Chiplets"** (Fig. 1, right). Approximately 35 high-school student participants from two local high schools applauded loudly and commented that they understood most of it and conveyed the following messages. As Moore's law has slowed and Dennard scaling has all but vanished, the technical world is exploring a wide variety of approaches to improve the performance, power, and cost of digital systems. Four key methods are explored in the talk: expanded use of AI, especially generative AI, improvements in design automation, novel chip architectures, and the use of multi-die systems, informally "chiplets." Each of these approaches has already demonstrated the ability to recapture some amount of classical scaling, yet while they are not orthogonal to one another, neither are they always aligned. The talk focused on some of the challenges of the approaches and the resulting opportunities for them to work together, which can be thought of metaphorically as a kind of impedance matching designed to preserve and amplify the improvements available from each source.

Tulika Mitra, Vice-Provost (Academic Affairs) and Provost's Chair Professor of Computer Science at the National University of Singapore (NUS), talked about **"Next-Generation Accelerator Design: Specialize or Generalize?"** (Fig. 2, left). In the AI era, where parallelism is essential for modern workloads, several domain-specific AI hardware accelerators rooted in dataflow computing models have emerged to cater to the demanding performance and energy-efficiency needs of such workloads. Yet, this specialization shift challenges the programmability of generalized solutions cherished by software developers. The central question arises: Can the advantages of both paradigms be harmonized? Similar to how versatile foundation models in generative AI can be fine-tuned to a wide range of specialized tasks, can we find a way to blend the adaptability of general-purpose processors with the strength of specialized accelerators? Further, software-defined hardware accelerators, an innovative, domain-agnostic embodiment

The authors are with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

Digital Object Identifier 10.1109/MCAS.2024.3395580
Date of current version: 15 August 2024

Fig. 1. Dr. Huiming Bu (Left) delivering a keynote talk **"Path to 1 Trillion Transistors in the Era of AI"** and Dr. Robert Aitken (right) giving a presentation related **"Impedance Matching AI, EDA, Chips, and Chiplets."**

Fig. 2. Prof. Tulika Mitra (Left) presenting **"Next-Generation Accelerator Design: Specialize or Generalize?"** and Dr. Dae-Woo Kim (right) describing **"3D package solution, the new boundary of Si and package technology."**

of the dataflow computing model are presented. This approach envisages a synergistic hardware-software co-design strategy, allowing the same silicon chip to be morphed and instantiated to support various dataflows via software. Consequently, the system offers the efficiency of specialized accelerators while preserving the flexibility needed to accommodate diverse applications through generalization. The challenges posed by this novel paradigm were described along with a spotlight on the substantial opportunities it presents in the realm of accelerator design.

Andre Tost, Distinguished Engineer, IBM Watsonx Client Engineering, presented recent developments in the talk **"Generative AI in the Enterprise World with Watsonx"** (Fig. 3, left). The advent of generative AI technology has revolutionized the enterprise world, transforming business processes and user experiences alike. One key aspect of this shift is the integration of Large Language Models within IT landscapes, enabling increased automation and natural language interfaces. This talk delves into real-world use cases that leverage generative AI, showcasing concrete tasks assigned to Large Language Models across diverse industries and regions. These ideas address unique challenges that arise with the implementation of generative AI. Then strategies for addressing privacy, compliance, and trustworthiness

concerns, as well as approaches for managing the increased resource consumption associated with this technology are explored. Then examples of generative AI's impact on various industries, such as finance, manufacturing, and retail are highlighted. The usage of Large Language Models to streamline processes, enhance customer interactions, and unlock new business opportunities are showcased. Furthermore, the ethical considerations that accompany the use of generative AI, by examining the importance of transparency in AI decision-making, and the need for compliant and representative training data are stressed.

Dae-Woo Kim, Corporate VP, AVP, Samsung Electronics talked about **"3D package solution, the new boundary of Si and package technology"** (Fig. 2, right). Despite the high cost of the silicon fabrication process, chip sizes would like to increase beyond the reticle size limit by adding more and more functional blocks for high performance computing. In particular, with the continuous demand for high performance and high capacity in memory products, the amount of data created, processed, stored, and transferred is increasing tremendously. To overcome these challenges, advanced packages based on RDL (Re-Distribution Layer), flip chip bonding, and TSV (Through Silicon Via) have been actively used for heterogeneous integration in electronic packages for the past decade. Heterogeneous integration using advanced packaging technology (2.5D and 3D) and chiplets have been attracting a lot of attention as these approaches enable higher bandwidth with low power consumption at a reduced cost. Advanced packaging has been developed for optimum chip-to-chip interconnections, so more and more silicon fabrication processes are being adopted in package technology. The 2.5D silicon interposer architecture has been widely used for horizontal interconnection between logic to logic and logic to high bandwidth memory integration. The 3D stacking architecture is for vertical interconnections enabling small form factor, increasing signal speed, and reducing power consumption and power dissipation. To provide a bonding pitch solution less than 10 um for extreme I/O density, power and signal integrity, and thermal property, HCB (Hybrid Cu Bonding) must be considered for next generation bonding solutions. In this talk, recent advanced package technology and the roadmap for the Samsung AVP business unit was shared for memory, mobile, and HPC products.

Samuel Naffziger, SVP and Corporate Fellow, Advanced Micro Devices, talked about **"Technology and Architecture Requirements for Energy Efficient AI"** (Fig. 3, right). While the explosion in the capabilities of generative AI and the associated benefits has received a tremendous amount of attention, both in the technology world and in popular press, so has the unprecedented amount of compute and energy required to train and serve these extremely complex models. Many of these concerns are well founded given the hundreds of Mega-Watt-hours required to train large language models (LLMs). To frame the challenge, the trends in energy use for generative AI and where the power is consumed for modern training systems need to be understood. In this context, the talk proposes adopting a holistic approach to energy efficiency to address issues related to how to avoid overwhelming the world's power grid and energy generation capabilities. This approach involves the reduction of energy per computation through the use of lower precision math formats and associated algorithms, the adoption of advanced packaging and 3D architectures to reduce data movement power, and advanced optical interconnects. The most powerful lever

Fig. 3. Mr. Andre Toast (Left) presenting a talk related to **"Generative AI in the Enterprise World"** and Samuel Naffziger (right) describing **"Technology and Architecture Requirements for Energy Efficient AI."**

Fig. 4. Dr. Arif Khan (Left) talked about **"The More-than-Moore Juggernaut for Differentiation and Disaggregation for Processors and SoCs in the Generative AI World"** and Dr. Debendra Das Sharma (right) presented **"Universal Chiplet Interconnect Express™ (UCIe™): An Open Interconnect Standard for Innovation with Chiplets."**

ties these improvements together with algorithmic and hardware-software synergy for efficiently mapping AI problems onto the optimized system. If these approaches are adopted, then the history and recent developments in our industry point to an ability to deliver the benefits of LLMs while reining in energy use.

Arif Khan, Sr. Group Director, Cadence, presented a great talk **"The More-than-Moore Juggernaut for Differentiation and Disaggregation for Processors and SoCs in the Generative AI World"** (Fig. 4, left). This past year, ChatGPT was quite the phenomenon as generative AI hit the peak of the hype cycle and made AI part of our everyday lexicon. This presentation discussed the key market trends driving AI and the demand for newer processor/SoC, chip-to-chip, and module architectures that address the needs of this space. The unsustainable pace of AI die-size growth has come up against the reticle limit. The rise in cost per transistor (CPT) is outweighing scaling benefits from advances in generational process technology. The need for high numerical aperture EUV (High NA-EUV) at nodes beyond 3 nm reduces the reticle size by half. These diminishing silicon economies of scale have pushed foundries, EDA companies, and the manufacturing ecosystem to enable chiplet designs. New developments in packaging technology (through-silicon vias and stacking, interposers, bridging, bump-pitch scaling) and standardization of die-to-die interfaces are providing technology gains to offset the challenges of die sizes and CPT.

Standards bodies and IP implementers have risen to the challenge of providing solutions to interface bottlenecks. This talk also highlighted important standards in memory, such as the latest LPDDR and HBM versions, along with key interface standards such as 112G/224G, peripheral component interconnect express (PCIe), and Compute Express Link (CXL), and chiplet and die-to-die interfaces such as Universal Chiplet Interconnect Express **(**UCIe) that are critical to these new architectures for AI products. The juggernaut for differentiation and disaggregation in the *more-than-Moore* era continues to build momentum. These trends and recent technology advances, while exploring future directions—including questions such as "Can AI be used to design the next 3D-IC AI processors?" are explored through this talk.

Debendra Das Sharma, Intel Senior Fellow, co-GM Memory and I/O Technologies, Intel, presented **"Universal Chiplet Interconnect Express™ (UCIe™): An Open Interconnect Standard for Innovation with Chiplets"** (Fig. 4, right). High-performance workloads demand on-package integration of heterogeneous processing units, on-package memory, and communication infrastructure such as co-packaged optics to meet the demands of the computing landscape in the generative AI era. On-package interconnects are a critical component to deliver power-efficient performance with the right feature set in this evolving landscape.

Universal Chiplet Interconnect Express (UCIe), is an open industry standard with a fully specified stack that comprehends plug-and-play interoperability of chiplets on a package; like the seamless interoperability on board with well-established and successful off-package interconnect standards such as PCI Express® and Compute Express Link (CXL)®. The usages and key metrics associated with different technology choices in UCIe and how this open standard could potentially evolve to incorporate more compelling usage models in the future were described.

During the symposium, a poster session as well as a panel discussion was conducted (Fig. 5). The details of the program are listed in the following link. https://www.zurich.ibm.com/thinklab/AIcomputesymposium.html

Fig. 5. Dr. Rajiv Joshi (Left) gave **opening/closing remarks** and Dr. Arvind Kumar (right) moderated the **panel discussion**.

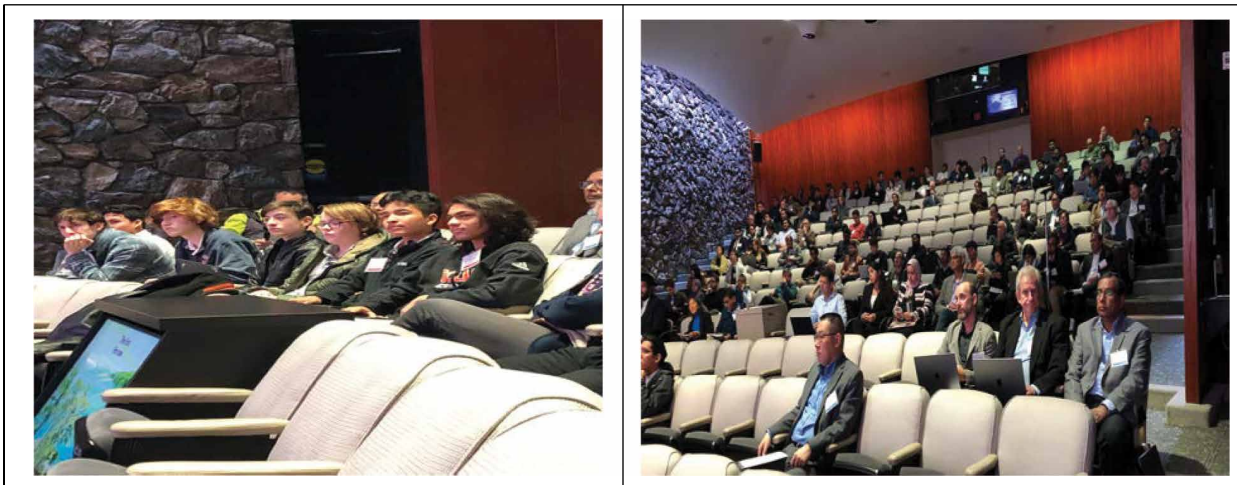

Fig. 6. Yorktown High School Students (left) and the audience (right) engrossed in talks.

This year, the AI Compute Symposium has expanded its participant base to include local Yorktown Heights high school students for the first time (Fig. 6). This initiative marks a significant step towards engaging and nurturing young talent in AI. It provides a unique platform for these students to immerse themselves in the latest trends in AI hardware and software. During the symposium, these bright young minds had the chance to contribute actively, particularly in the poster sessions. Their involvement went beyond mere participation—in a remarkable achievement, one high school student co-authored and presented a poster showcasing their research and insights alongside seasoned professionals.

Overall this symposium included industry veterans and exposed the young and the mature audience to future directions in technology and motivated all to pursue abundant opportunities in the field of chips, chiplets, and systems.

General Chairs
Rajiv Joshi, Matthew Ziegler, and Jin-Ping Han

Technical Program Committee
Anna Topol, Kaoutar El Maghraoui, Krishnan Kailas, Xin Zhang, Arvind Kumar, Linda Rudin, Cheng Chi, Atom Watanabe, and John Rozen