Said Hamdioui

Delft University of Technology, The Netherlands

S.Hamdioui@tudelft.nl

Biography

Said Hamdioui is currently Chair Professor on Dependable and Emerging Computer Technologies, Head of the Quantum and Computer Engineering department, and also serving as Head of the Computer Engineering Laboratory (CE-Lab) of the Delft University of Technology, the Netherlands. He is also the co-founder and CEO of Cognitive-IC, a start-up focusing on hardware dependability solutions and consultancy. Hamdioui received MSEE and PhD degrees (both with honors) from TUDelft. Prior to joining TUDelft as a professor, Hamdioui worked for Intel Corporation (Califorina, USA), Philips Semiconductors R&D (Crolles, France) and Philips/ NXP Semiconductors (Nijmegen, The Netherlands). His research focuses on two domains: dependable CMOS nano-computing (including testability, reliability, and hardware security) and emerging technologies and computing paradigms (including memristors for logic and storage, in-memory-computing for big-data applications).

Hamdioui owns two patents, has published one book and contributed to other two, and has co-authored over 200 conference and journal papers. He has consulted for many semiconductor companies in the area of memory testing and has collaborated with many industry/research partners in the field of dependable nano-computing and emerging technologies. Hamdioui is a Senior member of the IEEE, was Associate Editor of *IEEE Transactions on VLSI Systems* (TVLSI), and he serves on the editorial board of *IEEE Design & Test*, *Elsevier Microelectronic Reliability Journal*, the *Journal of Electronic Testing: Theory and Applications* (JETTA), and the *ACM Journal on Emerging Technologies in Computing Systems* (JETC). He is also member of AENEAS/ENIAC Scientific Committee Council (AENEAS - Association for European NanoElectronics Activities). Hamdioui has been the recipient of many international/national awards. E.g., he was the recipient of European Design Automation Association Outstanding Dissertation Award 2001; and many Best Paper Awards (DATE, ISVLSI, ICCD, etc)

Lecture 1: Device Aware Test: The Means to Win the War Against Unmodeled Faults

Testing defects in logic and memory chips underwent a long evolution process. Testing went from early functional test methods to structural testing driven by well-defined fault models. The development of such models assumes that physical defects in devices can be modeled as linear resistors. Although it can be convincing for modeling opens and shorts in interconnects, this assumption has never been validated for devices. In addition, it is well known that scaling below 10nm is giving rise to many device failure mechanisms that cannot be modeled by linear resistors; not to mention emerging devices which are by nature non-linear such as RRAMs, PCMs and STT-MRAMs.

This tutorial introduces Device-Aware Test (DAT) as a mean to close the gap between manufacturing defects and the way they can be modeled. DAT does not assume that a defect in a device (or a cell) can be modeled electrically as a linear resistor (as the state-of-the art approach

suggests), but it rather incorporates the impact of the physical defect in the technology parameters of the device and thereafter in its electrical parameters. Once the defective electrical model is defined, a systematic fault analysis is performed to derive appropriate fault models and subsequently test solutions. The tutorial demonstrates the application/use and the superiority of DAT approach based on two industrial memory designs: STT-MRAMs and RRAMs. For the considered memories, the industrial results will be presented to show that DAT sensitizes realistic faults as well as new unique defects and faults that can never be caught with the traditional approach. In addition, the results will show how powerful is DAT for fast diagnosis and yield learning.

Lecture 2: Computation-in-Memory Hardware Architectures for Edge-AI

Emerging IoT-edge applications are extremely demanding in terms of storage, computing power and energy efficiency in order to enable the deployment of AI, hence generate the “information” locally rather than communication the data to e.g., the cloud. On the other hand, both today’s computer architectures and device technologies are facing major challenges making them incapable to deliver the required functionalities and features at economical affordable cost. In order for computing systems to continue deliver sustainable benefits for the foreseeable future society, alternative computing architectures and notions have to be explored in the light of emerging new device technologies.

This tutorial addresses the potential, design, and test of Computation-in-memory architecture based on non-volatile (NV) devices such as ReRAM, PCM and STT-MRAM as alterative low power hardware architectures that could enable edge-AI. First, the talk briefly explains the limitation of both CMOS scaling and today’s computing architectures. Then it classifies the sate-of-the art computer architectures and highlight how the trends is going toward computation-in-memory (CIM) architectures in order to eliminated and/or significantly reduces the limitations of today’s technologies. The concept of CIM based on NV devices is discussed, and logic and arithmetic circuit designs using such devices and how they enable such architectures are covered; data measurements are shown to demonstrate the CIM concept in silicon. The strong dependency of application domains on the selection of appropriate CIM architecture and its building block, as well as the huge potential of CIM (in realizing order of magnitude improvement in terms of computing and energy efficiency) are illustrated based on some case studies. Finally, the research directions in different aspects of computation-in-memory are highlighted.

Yongpan Liu

Tsinghua University, China

ypliu@tsinghua.edu.cn

Biography

Yongpan Liu is currently a Professor with Tsinghua University, China, where he leads the Institute of Circuits and Systems. His research interests include nonvolatile processor, energy efficient machine learning accelerators, industry Internet and so on. He has published more than 60 journals (40 IEEE/ACM journals, 5 JSSC, 10 TCASI/II) and 100 conference papers (5 ISSCC, 3 VLSI, 4 A-SSCC, 11 DAC, 8 DATE, 2 ICCAD, 11 ASPDAC,) in the areas of VLSI Designs, EDA and Embedded Systems, focusing on nonvolatile computing, machine learning accelerator and embedded systems for internet of things. His work has received DAC under 40 Innovation Award 2017, Micro Top Pick 2016, Best Paper Award in ASPDAC 2017 and HPCA 2015, 2 Design Contest Awards in ISLPED 2012 and 2013, and 2 Best Paper Nominations in ASPDAC 2013 and 2016. Yongpan Liu has been an active volunteer in the design automation, VLSI, and embedded system conferences. He served as general secretary for ASPDAC 2020, technical program chair for NVMSA 2019, general chair for IWCR 2018, general chair for AWSSS 2016, exhibition chair for A-SSCC 2015, publicity chair for ICCD 2015-2016 and BioCAS 2016, and panel chair for AsiaHOST 2016. He also served as a program committee member for leading conferences, including ISSCC, ISCAS, DAC, ASP-DAC, ISLPED, ICCD, RTSS, A-SSCC. He is/was an Associate Editor for IEEE Transactions on CAS II, IEEE Transactions on CAD and IET Cyber-Physical Systems: Theory & Applications. He co-founded and serves as vice chair for ACM SIGDA North China Chapter and co-founded IEEE CEDA Beijing Chapter. He is an IEEE Senior Member and ACM Member.

Lecture 1: Energy Efficient Nonvolatile Intelligent IoT Processor and Its Applications

Internet of things are regarded as a very promising market in the next decade. However, batteries have become as a critical obstacle due to their limited operating time and frequent maintenance. Energy harvesting techniques are proposed to relieve those problems and self-powered sensor nodes are attracting more and more attentions. A typical self-powered sensor node consists of power supply and computation system, and it collects energy from ambient power sources, such as solar, vibration, temperature difference and RF energy. Several major design challenges exist in the present self-powered sensor nodes: 1) Limited output power: The typical generated power ranges from several mW to hundreds of uW, leading to a gap of several orders of magnitude between the harvested energy and the consumption of mainstream low power chips. 2) Frequent power failures: Lots of power failures occur frequently in self-powered systems, requiring efficient operations in an energy intermittent mode. 3) Hard to predict: The power profiles are determined by the ambient factors and hard to be predicted. This tutorial will provide several state-of-the-art techniques from circuit levels to system levels to handle above challenges, including nonvolatile processor design, architecture exploration, software and system optimization techniques for

energy harvesting sensor platform. Finally, we demonstrate a smart ultraviolet monitoring system using CNN-based pattern recognition on the platform.

Lecture 2: Energy Efficient Sparse Machine Learning Professor

Sparsity is widely existed in modern neural networks and how to support such sparsity in hardware is an important direction to enhance energy efficiency of machine learning chips. This talk will first begin with an introduction of various pruning algorithm techniques to achieve sparse neural network (i.e., unstructured and structured sparse networks). Furthermore, we review different up-to-date architectures and chips to make efficient inference and training of sparse neural network, covering both spatial domain as well as time domain sparsity. Finally, we discuss challenges and future directions to support sparsity in computing-in-memory artificial intelligent chips.

Ljiljana Trajkovic

Simon Fraser University, Canada

ljilja@cs.sfu.ca

Biography

Ljiljana Trajkovic received the Dipl. Ing. degree from University of Pristina, Yugoslavia, in 1974, M.Sc. degrees in electrical engineering and computer engineering from Syracuse University, Syracuse, NY, in 1979 and 1981, respectively, and a Ph.D. degree in electrical engineering from University of California at Los Angeles, in 1986.

She is currently a Professor in the School of Engineering Science at Simon Fraser University, Burnaby, British Columbia, Canada. From 1995 to 1997, she was a National Science Foundation (NSF) Visiting Professor in the Electrical Engineering and Computer Sciences Department, University of California, Berkeley. She was a Research Scientist at Bell Communications Research, Morristown, NJ, from 1990 to 1997, and a Member of the Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ, from 1988 to 1990. Her research interests include high-performance communication networks, control of communication systems, computer-aided circuit analysis and design, and theory of nonlinear circuits and dynamical systems.

Dr. Trajkovic served as IEEE Division X Delegate/Director (2019–2020) and IEEE Division X Delegate-Elect/Director-Elect (2018). She served as Senior Past President (2018–2019), Junior Past President (2016–2017), President (2014–2015), President-Elect (2013), Vice President Publications (2012–2013, 2010–2011), Vice President Long-Range Planning and Finance (2008–2009), and a Member at Large of the Board of Governors (2004–2006) of the IEEE Systems, Man, and Cybernetics Society. She served as 2007 President of the IEEE Circuits and Systems Society and a member of its Board of Governors (2004–2005, 2001–2003). She is Chair of the IEEE Circuits and Systems Society joint Chapter of the Vancouver/Victoria Sections. She was Chair of the IEEE Technical Committee on Nonlinear Circuits and Systems (1998). She was General Co-Chair of SMC 2020 and SMC 2020 Workshop on BMI Systems and served as General Co-Chair of SMC 2019 and SMC 2018 Workshops on BMI Systems, SMC 2016, and HPSR 2014, Special Sessions Co-Chair of SMC 2017, Technical Program Chair of SMC 2017 and SMC 2016 Workshops on BMI Systems, Technical Program Co-Chair of ISCAS 2005, and Technical Program Chair and Vice General Co-Chair of ISCAS 2004. She served as an Associate Editor of the *IEEE Transactions on Circuits and Systems (Part I)* (2004–2005, 1993–1995), the *IEEE Transactions on Circuits and Systems (Part II)* (2018, 2002–2003, 1999–2001), and the *IEEE Circuits and Systems Magazine* (2001–2003). She is a Distinguished Lecturer of the IEEE Systems, Man, and Cybernetics Society (2020–2021) and the IEEE Circuits and Systems Society (2020–2021, 2010–2011, 2002–2003). She is a Professional Member of IEEE-HKN and a Life Fellow of the IEEE.

Lecture 1: Complex Networks

The Internet, social networks, power grids, gene regulatory networks, neuronal systems, food webs, social systems, and networks emanating from augmented and virtual reality platforms are

all examples of complex networks. Collection and analysis of data from these networks is essential for their understanding. Traffic traces collected from various deployed communication networks and the Internet have been used to characterize and model network traffic, analyze network topologies, and classify network anomalies. Data mining and statistical analysis of network data have been employed to determine traffic loads, analyze patterns of users' behavior, and predict future network traffic while spectral graph theory has been applied to analyze network topologies and capture historical trends in their development. Machine learning techniques have proved valuable for predicting anomalous traffic behavior and for classifying anomalies and intrusions in communication networks. Applications of these tools help understand the underlying mechanisms that affect behavior, performance, and security of computer networks.

Lecture 2: Data Mining and Machine Learning for Analysis of Network Traffic

Collection and analysis of data from deployed networks is essential for understanding modern communication networks. Data mining and statistical analysis of network data are often employed to determine traffic loads, analyze patterns of users' behavior, and predict future network traffic while various machine learning techniques proved valuable for predicting anomalous traffic behavior. In described case studies, traffic traces collected from various deployed networks and the Internet are used to characterize and model network traffic, analyze Internet topologies, and classify network anomalies.

Shimeng Yu

Georgia Institute of Technology, USA

shimeng.yu@ece.gatech.edu

Biography

Shimeng Yu is an associate professor of electrical and computer engineering at the Georgia Institute of Technology. He received the B.S. degree in microelectronics from Peking University in 2009, and the M.S. degree and Ph.D. degree in electrical engineering from Stanford University in 2011 and 2013, respectively. From 2013 to 2018, he was an assistant professor at Arizona State University.

Prof. Yu's research interests are nanoelectronic devices and circuits for energy-efficient computing systems. His expertise is on the emerging non-volatile memories (e.g., RRAM, ferroelectrics) for different applications such as deep learning accelerator, neuromorphic computing, monolithic 3D integration, and hardware security.

Among Prof. Yu's honors, he was a recipient of the NSF Faculty Early CAREER Award in 2016, the IEEE Electron Devices Society (EDS) Early Career Award in 2017, the ACM Special Interests Group on Design Automation (SIGDA) Outstanding New Faculty Award in 2018, the Semiconductor Research Corporation (SRC) Young Faculty Award in 2019, and the ACM/IEEE Design Automation Conference (DAC) Under-40 Innovators Award in 2020, etc.

Prof. Yu served or is serving many premier conferences as technical program committee, including IEEE International Electron Devices Meeting (IEDM), IEEE Symposium on VLSI Technology, ACM/IEEE Design Automation Conference (DAC), ACM/IEEE Design, Automation & Test in Europe (DATE), ACM/IEEE International Conference on Computer-Aided-Design (ICCAD), IEEE International Symposium on Circuits and Systems (ISCAS), etc. He is a senior member of the IEEE.

Lecture 1: Circuit Design and Silicon Prototypes for Compute-in-Memory for Deep Learning Inference Engine

Compute-in-memory (CIM) is a new computing paradigm that addresses the memory-wall problem in the deep learning inference engine. SRAM and resistive random-access memory (RRAM) are identified as two promising embedded memories to store the weights of the deep neural network (DNN) models. In this seminar, first I will review the recent progresses of SRAM and RRAM-CIM macros that are integrated with peripheral analog-to-digital converter (ADC). The bit cell variants (e.g., 6T SRAM, 8T SRAM, 1T1R, 2T2R) and array architectures that allow parallel weighted sum are discussed. State-of-the-art silicon prototypes are surveyed with normalized metrics such as energy efficiency (TOPS/W). Second, we will discuss the array-level characterizations of non-ideal device characteristics of RRAM, e.g., the variability and reliability of multilevel states, which may negatively affect the inference accuracy. Third, I will discuss the general challenges in CIM chip design with regards to the imperfect device properties, ADC overhead, and chip to chip variations. Finally, I will discuss future research directions including

monolithic 3D integration of memory tier on top of the peripheral logic tier, as well enhancing the inference engine's security against DNN model leaking and reverse engineering.

Lecture 2: Tutorial - NeuroSim: A Benchmark Framework of Compute-in-Memory Hardware Accelerators from Devices/Circuits to Architectures/Algorithms

DNN+NeuroSim is an integrated framework to benchmark compute-in-memory (CIM) accelerators for deep neural network (DNN), with hierarchical design options from device-level, to circuit-level and up to algorithm-level. NeuroSim is a C++ based circuit-level macro model, which can achieve fast early-stage design exploration (compared to a full SPICE simulation). It takes design parameters including memory types (includes SRAM, RRAM/PCM and FeFET), non-ideal device parameters, transistor technology nodes (from 130 nm to 7nm), memory array size, training dataset and traces to estimate the area, latency, dynamic energy, leakage power. A python wrapper is developed to interface NeuroSim with deep learning platforms Pytorch/Tensorflow, to support flexible network topologies including VGG and ResNet for CIFAR/ImageNet. It supports weight/activation/gradient/error quantization in algorithm, and takes non-ideal properties of synaptic devices and peripheral circuits, in order to estimate training/inference accuracy. The framework is open-sourced and publicly available on GitHub. DNN+NeuroSim's user community is growing, including industry researchers from Intel, TSMC, Samsung and SK Hynix. Therefore, it is timely to conduct a tutorial for a broader education to the community, and help the researchers to use/modify the code more flexibly for their own research purposes.

Xiao-Ping (Steven) Zhang

Ryerson University, Canada

xzhang@ee.ryerson.ca

Biography

Xiao-Ping (Steven) Zhang received B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, both in Electronic Engineering. He holds an MBA in Finance, Economics and Entrepreneurship with Honors from the University of Chicago Booth School of Business, Chicago, IL.

Since Fall 2000, he has been with the Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON, Canada, where he is currently a Professor and the Director of the Communication and Signal Processing Applications Laboratory. He has served as the Program Director of Graduate Studies. He is cross-appointed to the Finance Department at the Ted Rogers School of Management, Ryerson University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015 and 2017. He is a frequent consultant for biotech companies and investment firms. He is the Co-Founder and CEO for EidoSearch, an Ontario-based company offering a content-based search and analysis engine for financial big data. His research interests include statistical signal processing, machine learning, image and multimedia content analysis, sensor networks and IoT, and applications in big data, finance, and marketing.

Dr. Zhang is Fellow of the Canadian Academy of Engineering, Fellow of the Engineering Institute of Canada, Fellow of the IEEE, a registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society. He is the general Co-Chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2021. He is the general co-chair for 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business, and the general co-chair for 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He is an elected Member of the ICME steering committee. He is the General Chair for the IEEE International Workshop on Multimedia Signal Processing, 2015. He is the Publicity Chair for the International Conference on Multimedia and Expo 2006, and the Program Chair for International Conference on Intelligent Computing in 2005 and 2010. He served as a Guest Editor for Multimedia Tools and Applications and the International Journal of Semantic Computing. He was a tutorial speaker at the 2011 ACM International Conference on Multimedia, the 2013 IEEE International Symposium on Circuits and Systems, the 2013 IEEE International Conference on Image Processing, the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, the 2017 International Joint Conference on Neural Networks and the 2019 IEEE International Symposium on Circuits and Systems. He is Senior Area Editor for the *IEEE Transactions on Signal Processing* and the *IEEE Transactions on Image Processing*. He was Associate Editor for the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Circuits and Systems for Video Technology*, the *IEEE Transactions on Signal Processing*, and the *IEEE Signal Processing Letters*. He received 2020 Sarwan Sahota Ryerson Distinguished Scholar Award –

the Ryerson University highest honor for scholarly, research and creative achievements. He is selected as IEEE Distinguished Lecturer by the IEEE Signal Processing Society for the term 2020 to 2021, and by the IEEE Circuits and Systems Society for the term 2021 to 2022.

Lecture 1: Localization and Tracking for Internet of Things

With the emerging deployment of 5G wireless systems and the rapid growing number of smart sensors and deploying sensors on or around physical objects, the Internet of Things (IoT) seamlessly integrates a world of networked smart objects, makes their information be shared on a global scale, and provides an ability of intelligent computing and information processing, such as reporting status, position, and surrounding condition of each sensor node. Localization and tracking is a key problem, which has been already studied in various fields, including sonar, radar, seismic, mobile communications, wireless sensor networks. However, many solutions may not directly suit an IoT scenario where large quantities of sensor nodes that perform distributed sensing and collaborative information processing tasks are interconnected together over a wireless channel. It is almost impossible to collect full network sampling data for accurate localization since any inter-sensor communication requires a large burden on sensor batteries. Typical metrics are measured at the local sensors including sample covariance matrices (SCM), time of arrival (TOA), gain ratios of arrival (GROA), angles of arrival (AOA) and frequency differences of arrival (FDOA). Estimating the source position as accurate as possible by utilizing the above-mentioned metrics is full of challenges. In this talk, we introduce the fundamentals of typical source localization and target tracking methods, including least-squares, maximum likelihood, convex relax optimization. We discuss state-of-the-art localization and tracking approaches using global navigation satellite systems (GNSSs), array sensor networks (ASN) as well as for indoor positioning for IoT applications, for example, subband information fusion, auxiliary variables-based algorithms, localization penalized maximum likelihood, and weighted least-squares using AOA-GROA-TOA, and other state-of-the-arts in localization and tracking systems for IoT applications.