Tutorial for VLSI Design 2024 - 500 word Summary

Title: Agile Software-Hardware Co-Design of AI-Centric Heterogeneous SoCs

Intelligent edge systems constitute a key growth segment within the cloud-backed cognitive IoT marketplace. The EPOCHS ("Efficient Programmability of Cognitive Heterogeneous Systems") research project at IBM (with collaborative university partners) is driven by the specific edge application domain of connected autonomous vehicles. Our team has developed leading edge methodologies for agile software-hardware co-design of heterogeneous SoCs to support the target application domain (along with adjacent domains like extended reality-driven use cases). As part of this project, we have successfully taped out two functional EPOCHS chips in 12 nm technology, and have demonstrated full-stack (software-hardware co-designed) solutions using both FPGA and ASIC versions of the SoCs. Specific use cases that have been illustrated in the context of our active technology transition phase are: (a) collaborative perception, involving a pair of communicating autonomous vehicles; (b) detecting hazards while scanning for objects and vehicles during autonomous navigation; (c) sentiment analysis of human-expressed conversations or commands (using an NLP algorithm). In this tutorial, we summarize the key innovations and open-source tools-driven agile methodology derived from this 5-year DARPA-sponsored project (2018-2023) that is now transitioning into commercially deployable solutions in partnership with clients. In particular, we cover the following aspects of the overall topic area in considerable detail: (a) the fundamentals of Columbia's ESP-driven agile SoC methodology, with demonstrated proof points in 10-100X improvement in designer productivity; (b) domain-specific hardware accelerators for AI/NLP – architecture and design methodology that address customer requirements in real-time performance, energy efficiency and security; (c) intelligent task scheduling and compiler support for domain-specific SoCs targeted for the important application space of connected autonomous vehicles.

- **Lead Speaker Bio-Sketch:**

Pradip Bose is a Distinguished Research Scientist and Manager of the Efficient & Resilient Systems department at IBM T. J. Watson Research Center. He has 40+ years of experience at IBM, where he was a member of the pioneering research project on RISC super scalar processors (at IBM Research) that led to the POWER-series processors and systems at IBM in the early 1990s. He was the lead performance engineer in POWER-3, and he has been involved in the pre-silicon power-performance modeling and power management of virtually all the POWER processors since then. He has also actively led the migration of power management techniques from the POWER platform to IBM's mainframe platform (z Systems) since the z13 product. Since 2012 he has served as the Principal Investigator of several successive DARPA-sponsored projects that have focused on highly energy-efficient edge processor design methodologies, with a more recent (2018-2023) emphasis on domain-specific SoCs that are easily programmable. These projects were driven by specific edge application use-cases of interest to the U.S. Department of Defense as well as commercial AI-centric enterprises within IBM and other cloud computing vendors. Pradip holds a B.Tech. (Hons.) degree in Electronics and Electrical Communication Engineering from I.I.T. Kharagpur (1977); M.S. and Ph.D degrees in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign (UIUC, 1981 & 1983). He holds well over 150 U.S. patents and the title of IBM Master Inventor. He also has over 180 peer-reviewed publications and several edited books to his credit. He is a Fellow of the IEEE and is a member of IBM's elite

Academy of Technology (recently renamed as the Open Innovation Community). He has given numerous tutorials and organized workshops at leading international conferences (e.g., ISCA, MICRO, HPCA in the computer architecture domain as well as DSN, ITC, ICCD and VLSI Design in the design community). He served as the co-General Chair of VLSI Design (with Prof. Susmita Sur-Kolay), held in Kolkata in 2016.

The tutorial will also include embedded video segments from

(a) Prof. Luca Carloni and one of his students, e.g., Joseph Zuckerman (Columbia University) covering the fundamentals of the ESP agile SoC methodology. (https://www.esp.cs.columbia.edu). Prof. Luca Carloni's bio-sketch is available at: http://www.cs.columbia.edu/~luca// and repeated below for convenience: Luca Carloni is Professor and Chair of Computer Science at Columbia University in the City of New York, where he leads the System-Level Design Group. He holds a Laurea Degree *Summa cum Laude* in Electronics Engineering from the University of Bologna, Italy, a Master of Science in Engineering from the University of California at Berkeley, and a Ph.D. in Electrical Engineering and Computer Sciences from the University of California at Berkeley. At Berkeley Luca was the 2002 recipient of the Demetri Angelakos Memorial Achievement Award in recognition of altruistic attitude towards fellow graduate students. Luca received the Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2006, was selected as an Alfred P. Sloan Research Fellow in 2008, received the ONR Young Investigator Award in 2010 and the IEEE CEDA Early Career Award in 2012. Luca is an IEEE Fellow (class of 2017). His research interests include methodologies and tools for system-on-chip platforms with emphasis on heterogeneous computing, intellectual property reuse, design of networks-on-chip, embedded software and distributed embedded systems. The System-Level Design Group has made several research contributions available in the public domain, including in particular ESP, the open-source heterogeneous system-on-chip platform. Luca coauthored one hundred and sixty refereed papers and is the holder of two patents. He received the best paper award at DATE'12 for the paper "Compositional System-Level Design Exploration with Planning of High-Level Synthesis" and at CloudCom'12 for the paper "A Broadband Embedded Computing System for MapReduce Utilizing Hadoop." His ICCAD'99 paper "A Methodology for Correct-by-Construction Latency-Insensitive Design was selected for The Best of ICCAD, a collection of the best papers published at the IEEE International Conference on Computer-Aided Design from 1982 to 2002. Luca is currently an associate editor of the IEEE Transactions on Computers, the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems and the ACM Transactions in Embedded Computing Systems. He has served on the technical program committee of several conferences including DAC, DATE, ICCAD, and EMSOFT. In 2010 he served as technical program co-chair of the International Conference on Embedded Software (EMSOFT), the International Symposium on Networks-on-Chip (NOCS), and the International Conference on Formal Methods and Models for Codesign (MEMOCODE). He served as vice general chair (in 2012) and general chair (in 2013) of Embedded Systems Week (ESWeek), the premier event covering all aspects of embedded systems and software. Luca was the co-leader of the Platform Architectures theme in the Gigascale Systems Research Center (GSRC) and participated in the Center for Future Architectures Research (C-FAR).

(b) A couple of key research colleagues of Dr. Pradip Bose at IBM Research.