Hamed Tabkhi

360 Huntington Ave, 409 DA, Boston, MA 02115, USA
Phone: +1 (617) 763-4429, Email: *tabkhi@coe.neu.edu*
Homepage: *http://www.northeastern.edu/esl/users/Hamed-Tabkhi*

Research Statement

Research Objective

My research interest lies in novel architecture and design methodology for embedded computer systems to address societal challenges in health, safety and assistance. To this end, my primary focus is on emerging data analytic and signal-processing applications such as real-time embedded vision, cyber-physical systems, wireless baseband, biomedical and robotics. I design, architect and implement novel heterogeneous many-core platforms that consume little power (few watts) while offering very high performance to run complex algorithms for low cost, widespread deployment. My aim is to bridge the gap between growing complexity of algorithms and limited embedded architecture capabilities. For this purpose, my research roadmap is raising abstraction and architecture awareness through algorithm/architecture co-design. Overall, the result of my research opens a path for embedded deployment of advanced streaming applications to support us in many aspects of our daily life.

Current Research

My current research can be categorized into three major areas:

1) Real-Time Embedded Vision

The aim of this research is to integrate real-time vision capabilities into the physical environment for the smart understanding and analyzing of scenes without human intervention. The primarily targeted applications are video surveillance, advanced driver assistance, and patient-monitoring systems. The embedded realization of advanced vision applications is notoriously difficult. These applications often demand adaptive vision algorithms with complex machine learning computation as well as immerse data access to keep and continuously update models of the scene. Algorithm examples are Mixture-of-Gaussians (MoGs) background subtraction, Kanade Lucas optical flow, and Deep Convolutional Neural Networks (D-CNNs) object classification. To tackle the architecture complexity of these algorithms, I collaborate with vision algorithm experts for innovative cross layer optimizations between the algorithm and architecture (tuning algorithms to fit the architecture, customizing architecture to match algorithms). Some outcomes of this research have been data access separation between streaming pixels and algorithm-intrinsic (scene model), compression on streaming data access, and reconfigurable architecture enabling pyramid-based 2D vision processing [1, 2]. I am leading a team of graduate and undergraduate students to architect and design our insight in real-time embedded vision computing for a diverse set of applications. We have already demonstrated our design principles by prototyping an object tracking vision flow targeted for video surveillance on Xilinx Zynq platform. Our design operates at Full-HD resolution 30 frame per second execution 40GOPs at only 1.4Watts [3]. My research in embedded vision computing has been supported and awarded by Analog Devices Inc. (ADI). Overall, the results of my research demonstrate tremendous opportunity when studying and optimizing algorithm/architecture together.

2) Flexible Function-Level Acceleration for Domain-Specific Computing

The aim of this research is to reconcile the execution efficiency (performance and power) and programmability (flexibility) of processor architectures. This research, which started out as a joint collaboration with Analog Devices Inc. (ADI), proposes Function-level Processor (FLP) [4]. FLP is a novel architecture class which provides flexibility to execute many applications within a market while maintaining the efficiency comparable to custom hardware accelerators. The key insight in FLP is to match the architecture execution granularity with the programming abstraction by raising architecture programmability to the function-level granularity. FLP is a shift from optimizing individual applications (super specialization) to optimizing common functions and functions compositions presented across many applications (domain specialization). Instead of monolithic dedicated accelerators for applications, an FLP offers a set of composable function primitives to construct many applications. A version of FLP with the industrial name of Pipeline-Vision Processor (PVP) targeted for embedded vision market is successfully running on ADI BF60x vision processors. It efficiently runs 10 vision applications on the same platform [5]. Overall, FLP is not a general-purpose processor, but it is sufficiently programmable to efficiently execute many signal processing applications within the same market. Application examples are embedded vision, software defined radio, and cyber-physical systems. One branch of my research is to utilize FLP principles to integrate many custom hardware accelerates to the system [6, 7]. My aim is to shift from the current processor-centric view to a more equal, peer view between ACCs and the host processor.

3) Enhancing GPUs and FPGAs for Data-Intensive Applications

The aim of this research is to enhance Graphic Processing Units (GPUs) and Field Programmable Gate Arrays (FPGAs) to accelerate data-intensive applications with irregular execution patterns (e.g. big data analysis and classification, machine learning and machine vision). Both GPUs and FPGAs offer extensive parallel execution. However, the challenge is to match the parallelism intrinsic to the algorithms (or expressed in the languages) with the parallelism semantic of GPUs and FPGAs. In a joint collaboration with Northeastern University Computer Architecture Research (NUCAR) directed by Prof. David Kaeli, we study the interaction between the parallel algorithms and parallel architectures. One direction of our research is to enhance/revisit GPU architecture to match the execution semantic of irregular but massively parallel applications. As an example, we defined a new metric, Warps Progression Similarity (WPS), to capture and analyze thread-level execution similarity on GPUs [8]. We also demonstrated that WPS-aware scheduling on GPUs results in a higher throughput for irregular applications. Another major direction is to study the interaction between Open Computing Language (OpenCL) programming layer and the generated architecture on FPGAs [9]. OpenCL for FPGAs can provide a very attractive solution for high-performance computing by offering a customized data-path while abstracting away many implementation details. Overall, our aim is to bridge the gap between parallel algorithms and parallel architectures to streamline parallel algorithm development across heterogeneous nodes (FPGAs and GPUs).

Future Research Plan

For the short term, my aim is to follow and expand my current research areas. I will expand my collaboration with vision algorithm experts. In particular, we will further investigate into the architecture challenges of deep-learning vision algorithms as well as networking concepts across many embedded vision nodes. Our aim is to deploy real-time embedded vision processing to enhance research in the other fields (e.g. anthropology and primatology) for unsupervised pattern detection and behavioral analysis. I also see tremendous opportunities for architecting and programming at function-level granularity. My team is working on new FLP instances for other challenging markets (e.g. software-defined radio and radar). To streamline the process, we are working on automation tools to identify commonly used functions and compositions across many applications in the targeted markets. Furthermore, I am expanding my research collaboration with NUCAR research lab in parallel computing for data-intensive application with irregular execution patterns (e.g. big data classification and machine learning). We are launching a new research plan to design and architect a novel reconfigurable processor that combines the benefits of GPUs and FPGAs.

For the long term, my plan is to unify the process of algorithm design and architecture exploration. My aim is to make the embedded deployment of advanced signal processing applications cheap and affordable to be widespread used for health, safety, and assistance. To this end, one major goal of my research is to capture my insight in algorithm/architecture co-design as a novel design methodology. Ultimately, my research aims to create an Architecture Compiler (AC) that jointly explores, optimizes and refines architecture and algorithm. AC receives functional/technical requirements and produces a solution (a package of algorithm and architecture).

Funding Opportunities

I consider funding potentials both from government and industry. Currently, I am working on two NSF proposals. (1) A real-time embedded vision solution for primate behavioral analysis. This is an interdisciplinary collaboration with primatology scientists as well as vision algorithm experts to streamline research in monkeys' social behavior analysis. (2) A flexible function-level architecture for

streaming applications. The aim of this proposal is to expand the FLP principles for spatial/temporal multi-streaming as well as designing tools to automate FLP architecting and programming.
My plan is to actively collaborate with different companies. In general, my research leads to product-based solutions which are highly attractive and suitable for industry collaboration. In particular, I do see an enormous interest in real-time vision processing for autonomous driving, video surveillance, robotics and patient monitoring systems. My Ph.D. dissertation in embedded vision supported by ADI and my plan is to start an independent collaboration with ADI on vision as well as radar processing. I also know people from Nvidia and Qualcomm Inc. My aim is to utilize my insight in architectures for real-time vision processing to impact the future products of these companies. I also see significant potential in FLP context. As an example, I am in touch with MathWorks to study the interaction between Simulink (as a programming model), and FLP (as the execution model). Our research aim is to utilize FLP as a machine to efficiently execute algorithms developed in Simulink.

Sincerely,

Boston, Massachusetts, USA, March 2016
Hamed Tabkhi

Selected Publications:

** indicates students under my supervision*

[1] H. Tabkhi, R. Bushey and G. Schirner, "Conceptual Abstraction Levels (CALs) for Managing Design Complexity of Market-Oriented MPSoCs", Elsevier Journal of Microprocessors and Microsystems, vol.39, no.8, pp. 704-719, Nov. 2015.

[2] H. Tabkhi, M. Sabbagh and G. Schirner, "A Power-Efficient Real-Time Solution for Adaptive Vision Algorithms", *IET Computers & Digital Techniques*, vol.9, no.1, pp.16-26, Jan. 2015.

[3] H. Tabkhi, M. Sabbagh and G. Schirner, "An Efficient Architecture Solution for Low-Power Real-Time Background Subtraction", *IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAPs),* Toronto, Canada, Jul. 2015.

[4] H. Tabkhi, R. Bushey and G. Schirner, "Function-Level Processor (FLP): A Novel Processor Class for Efficient Processing of Streaming Applications", S*pringer Journal of Signal Processing and Systems*, in Press (accepted in 09/21/2015).

[5] H. Tabkhi, R. Bushey, G. Schirner, "Function-Level Processor (FLP): A High Performance, Minimal Bandwidth, Low Power Architecture for Market-Oriented MPSoCs", *IEEE Embedded Systems Letters*, vol.6, no.4, pp.65-68, Dec. 2014.

[6] N. Teimouri*, H. Tabkhi, G. Schirner, "Improving Scalability of CMPs with Dense ACCs Coverage", *IEEE Design Automation and Test in Europe (DATE),* Dresden, Germany, Mar. 2016.

[7] N. Teimouri*, H. Tabkhi, G. Schirner, "Revisiting Accelerator-Rich CMPs: Challenges and Solutions", *ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Francisco (CA), USA, Jun. 2015.

[8] C. Zhang*, H. Tabkhi, and G. Schirner, "Studying Inter-Warp Divergence Aware Execution on GPUs", IEEE *Computer Architecture Letters,* in Press (accepted in 09/03/2015).

[9] A. Momeni*, H. Tabkhi, G. Schirner and D. Kaeli, "Exploring the Efficiency of the OpenCL Pipe Semantic on an FPGA", *International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART),* Boston (MA), USA, Jun. 2015.