Title: Secure Processor Hardware

Abstract:
Recent security attacks have highlighted the inadequacy of modern processor hardware. Speculative execution and shared caches provide easy opportunities for side and covert channel attacks. Unless these vulnerabilities are promptly and effectively fixed, future processors will either be hard to use or deliver crippled performance.

In this talk, I present novel ways to address these vulnerabilities. First, I present a novel strategy to defend against hardware speculation attacks in multiprocessors. I call it “Invisible Speculation” (InvisiSpec) [MICRO’18]. The idea is to make speculative accesses invisible in the cache hierarchy. InvisiSpec can defend not only against existing branch speculation attacks, but also against attacks resulting from any speculative memory access. Second, I present a new way to eliminate cache-based side-channel attacks due to cache evictions in inclusive cache hierarchies. The idea, called SHARP [ISCA’17], involves changing the cache replacement policy in the last-level cache to deliberately avoid creating inclusion victims in private caches. Finally, I show that providing non-inclusive cache hierarchies is insufficient to stave off cache-based side channel attacks. I subvert such cache hierarchies with a novel attack that targets the directory structures, rather than the caches [S&P’19]. I also outline a solution to thwart directory attacks.

Bio:
Mengjia Yan is a PhD student at the University of Illinois at Urbana-Champaign (UIUC), advised by Professor Josep Torrellas. Her research interest lies in the areas of computer architecture and hardware security, with a focus on cache-based side channel attacks and defenses. She is a UIUC College of Engineering Mavis Future Faculty Fellow in 2018, and has received the UIUC Computer Science W.J. Poppelbaum Memorial Award for academic performance and design creativity. She interned at NVIDIA Research during the summer of 2018. Prior to UIUC, she earned her MS degree from UIUC in 2016 and her BS degree from Zhejiang University in 2013.
Title: Broad-Based Side-Channel Defenses for Modern Processor Architectures

Abstract:
Even in the face of strong encryption, side-channel attacks can steal confidential or private information by observing some aspects of the computer system's behavior. Numerous side channels exist, and these side channels are pervasive across applications including databases, machine-learning applications, and browsers. Unfortunately, although many defenses exist, they either close only a limited number of side channels, or they only protect applications that execute a static sequence of instructions and memory accesses. Combining such point solutions may not only increase implementation complexity while also drastically limiting the programs that can be protected, but it also may not guarantee security since point solutions may negate each others' defenses.

In this talk, I will describe the key ideas behind our solutions for closing a broad class of side channels for a variety of programs including machine-learning libraries, graph kernels, font-renderers, floating-point libraries, and cryptographic kernels. Specifically, I will describe (1) techniques to discover a broad class of side channels in the ISA, the microarchitecture, and the hardware implementation, (2) compiler-based techniques to identify portions of the code that may leak information through the previously-discovered side channels and techniques to fix the vulnerable code portions, and (3) a technique to validate whether a solution closes the relevant side channels.

Bio: Ashay Rane is a PhD student at the University of Texas at Austin. He works with Professor Calvin Lin and Professor Mohit Tiwari on side-channel defenses. His research interests are systems security, compilers and runtime systems, computer architecture, and program verification.
Title: Extreme-Efficiency Computing

Abstract:
From datacenters to wearables, computers are operating in environments where resources like storage, energy and compute capability are increasingly constrained. Under these constraints, computers must deliver the best Quality of Service (QoS) and carefully regulate critical measures. One important challenge is that these computers are structured in multiple modular layers (e.g., hardware, OS and network), each having many heterogeneous and distributed components designed independently. This mandates resource controllers to be modular in managing each component and coordinate to achieve system-wide goals. The desire for both modularity and coordination causes a tension that requires novel control design.

In my talk, I will present a novel resource management framework based on Robust control theory for modular coordinated control of computers. The key idea is that the lack of full system knowledge is represented as uncertainty, and we use formal techniques to guarantee that each component’s controller is safe and optimal under this uncertainty. Controllers in large machines are connected in a network to coordinate system-wide goals, and this network is assembled in a modular manner. I prototyped my ideas on heterogeneous platforms, demonstrating substantial advancement in resource efficiency over the state of the art. I will also present our ongoing work on combining Machine Learning with Robust control theory.

Bio:
Raghavendra (Raghav) is a PhD candidate at the University of Illinois, where he is advised by Prof. Josep Torrellas. His research is on building extremely resource efficient computers. His work spans the areas of Computer Architecture, Control Theory, Machine Learning, OS, and Distributed System Design. He is a winner of the W. J. Poppelbaum Award by the Dept. of CS at Illinois for architecture design creativity, a Mavis Future Faculty Fellowship at Illinois, and an ACM SRC Competition at PACT'17. He interned at AMD Research and his modular control design is being considered for upcoming heterogeneous systems. He received his Masters in CS from Illinois in 2014 and worked at Nvidia before graduate school. He has a Bachelors in Engineering (Hons.) in Electrical and Electronics Engineering from the Birla Institute of Technology & Science (BITS) Pilani, India where he received the University Gold medal on graduation.
Title: Expressive Memory: Rethinking the Hardware-Software Contract with Rich Cross-Layer Abstractions

Abstract:
Recent years have seen the rapid evolution and advancements at all levels of the computing stack, from application to hardware. Key abstractions and interfaces among the levels, however, have largely stayed the same: hardware and software, for instance, still primarily interact with traditional abstractions (e.g., virtual memory, instruction set architecture (ISA)). These interfaces are narrow, as hardware is unaware of key program semantics and programmer intent; and rigid, in terms of the fixed roles played by hardware and software. This fundamentally constrains the performance, programmability, and portability we can attain.

In this talk, I will make a case for rethinking the semantic contract between hardware and software and discuss how designing richer hardware-software abstractions can fundamentally change how we optimize for performance today. I will introduce two of our recent works in ISCA 2018 that explore the design and benefits of such cross-layer abstractions in two different contexts. I will first introduce Expressive Memory (XMem), a new cross-layer interface that communicates higher-level program semantics from the application to the underlying OS and hardware architecture. XMem thus enables the OS/architecture to identify the program's data structures and be aware of each data structure's access semantics, data types, etc. We demonstrate that this key, otherwise unavailable, information enables intelligent and much more powerful optimizations in operating systems and hardware architecture that significantly improves overall performance, programmability, and portability.

I will also briefly introduce the Locality Descriptor, a cross-layer abstraction to express and exploit data locality in throughput-oriented architectures, such as modern GPUs. I will discuss how a challenging aspect of programming GPUs can be made much simpler with a rich cross-layer programming abstraction that also significantly enhances performance and portability.

Bio: Nandita Vijaykumar is a Ph.D. candidate at Carnegie Mellon University, advised by Prof. Onur Mutlu and Prof. Phil Gibbons. Her research focuses on the interaction between programming models, system software, and hardware architecture, and explores how richer cross-layer abstractions can enhance performance, programmability, and portability. She is excited about rethinking the roles played by different levels of the stack in the modern era of rapidly evolving, specialized, and data-centric computing landscapes. During my Ph.D., I have been fortunate to intern at Microsoft Research, Nvidia Research, and Intel Labs. She is currently a visiting student at ETH Zurich.
Title: Software, Architecture, and VLSI Co-Design for Task-Based Parallel Runtimes

Abstract: Fast-paced advances in software are placing heavy performance and energy demands on hardware ranging from mobile systems running edge compute to warehouse-scale servers operating on big data. However, as technology scaling slows and process-related gains steadily decrease, a great burden and an even greater opportunity has fallen to computer architects who design at the software-hardware boundary to reimagine the computing stack for a new era of post-scaling compute. Luckily, architects have tremendous potential to build far more efficient computing systems through aggressive software/hardware co-design. For example, task-based parallel programming frameworks are one of the most popular ways to exploit increasing thread counts in CMPs (e.g., Intel Cilk Plus, Intel TBB). These task-distribution frameworks can be co-designed with hardware to differentiate threads doing useful work from threads waiting for work, providing seemingly obvious but otherwise missing critical information. In parallel, new enabling techniques for circuit control are continually becoming feasible thanks to research in the VLSI and circuit communities. For example, research on integrated voltage regulation is making per-core dynamic voltage and frequency scaling (DVFS) gradually more realistic. With pressure and opportunities coming from both ends of the computing stack, can architects co-design across the computing stack to create future systems that efficiently specialize for the most critical software domains? In this talk, I will describe a series of work that co-designs architecture with the software runtime and also with underlying circuits based on integrated voltage regulation to enable highly efficient, flexible, yet lightly specialized systems for task-based parallel runtimes.

Bio: Christopher Torng is a Ph.D. candidate in the Computer Systems Laboratory at Cornell University. His research interests are in the areas of computer architecture and VLSI, and he is also particularly interested in the challenge of productive hardware design for computer architecture researchers. His recent research spans a range of topics including work-stealing runtimes, heterogeneous big.LITTLE systems, SIMT-like architectures, and integrated voltage regulation. In the past, he was the university student lead for the DARPA-funded Celerity SoC in TSMC 16nm, he was the project lead for two chips in TSMC 28nm and in IBM 130nm designed using a Python-based hardware modeling language, and he has also supported the design of two chips in TSMC 65nm and TSMC 180nm with circuits targeting challenges in IoT synchronization and on-chip voltage regulation.
Title: On the design and optimization of hardware accelerators with applications in deep learning

Abstract:
In this talk I will cover two projects I worked on during my PhD to improve deep learning inference efficiency through co-designing optimizations with the unique properties of DNNs. Minerva is an optimization framework developed to reduce inference power by 8x over an accelerator baseline without compromising model accuracy. The key contribution is three aggressive, unsafe optimizations: datatype quantization, activity pruning, and SRAM voltage scaling with proposed fault mitigation techniques. Each optimization provides an efficiency-accuracy tradeoff and Minerva tightly bounds the optimizations to maximize efficiency while preserving model accuracy. Next, I will cover my work on weight compression to lower the cost of distributing weights from the datacenter to edge devices. Weightless is a weight encoding technique based on the Bloomier filter, a probabilistic data structure that tests set membership and can return an associated value. Experimentally, I found that the DNN’s implicit error resiliency enabled models to tolerate large amounts of incorrect weights resulting from false-positives (a side effect of the data structure). However, with a retraining phase the data structures can be much smaller, in the best case compressing a layer of weights by 496x.

Bio:
I am currently a research scientist at Facebook working with the AI Infrastructure team to improve the efficiency and performance of machine learning at both the mobile and datacenter scale. I received my PhD from Harvard this May with a focus on low-power specialized hardware design for Deep Learning workloads and accelerator centric architectures. I am particularly interested in aggressive “unsafe” optimizations that alter a DNN’s computation to improve efficiency while simultaneously preserving high model accuracy. My hope is that with these types of optimizations, ubiquitous, real-time inference execution will be possible (e.g., on devices with power budgets of <10mW). I have worked on many projects along these lines and proposed optimizations including data-type reduction, cheap, fault-prone weight memories, as well as sparse execution and weight compression.