2023 ACE/SCR Annual Review | OCTOBER 4-5,2023

ACE CENTER FOR EVOLVABLE COMPUTING

Semiconductor Research Corporation

I UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Cornell University

Georgia Tech

HARVARD UNIVERSITY

KU THE UNIVERSITY OF KANSAS

MIT Massachusetts Institute of Technology

UNIVERSITY OF MICHIGAN

THE OHIO STATE UNIVERSITY

PURDUE UNIVERSITY

Stanford University

TEXAS The University of Texas at Austin

UC San Diego

W UNIVERSITY of WASHINGTON

ACE Themes & Tasks

The goal of the ACE Center is to devise novel technologies for scalable distributed computing that will improve the performance and the energy efficiency of diverse applications by 100x over the expected computer systems of 2030.

Distributed computing in 2030 will be defined by the need to process vast swaths of data for insights in a timely manner. Minimizing data movement to curtail energy consumption in an energy-conscious Earth will be the overriding constraint. The compute infrastructure will be a seamless hierarchy of compute centers from edge to geo-distributed mega-datacenters. Each compute center will contain a large number of heterogeneous hardware accelerators, and tasks of unprecedented small granularity will seamlessly ship computation to where data is. To further minimize data movement, key data will be replicated or re-materialized on demand. The computational environment will be highly dynamic, with the constant introduction of new classes of accelerators for barely-emerging workloads, and of new applications/protocols that could benefit from yet-to-be-conceived accelerators.

THEME 1: HETEROGENEOUS COMPUTING PLATFORMS	
3134.001	Evolvable Distributed Accelerators
3134.002	Composable Distributed Acceleration
3134.003	Making Distributed Accelerator Ensembles Usable: Multilatency & Code Mapping
3134.004	Energy-efficiency Driven CPU-centric Nodes
THEME 2: DISTRIBUTED EVOLVABLE MEMORY & STORAGE	
3134.005	Scalable heterogeneous Memory Hierarchies
3134.006	Scalable Management of distributed Memory & Storage Assets
3134.007	Near- and In- memory/storage Acceleration
THEME 3: FINE-GRAINED COMMUNICATION AND COORDINATION	
3134.008	An Accelerator-Rich Datacenter Architecture and Beyond
3134.009	An Evolvable Network Stack
3134.010	A Self-balancing Planet-Scale Distributed Runtime
3134.011	In-network Computing
3134.012	Hardware-Supported Intelligent Distributed Data Stores
THEME 4: SECURITY, PRIVACY AND CORRECTNESS	
3134.013	Data-centric Security that Evolves with Threat Models and Systems
3134.014	Domain-specific TEE on Evolving heterogeneous Accelerators
3134.015	Security and Privacy Assurance
3134.016	Design for Verification of Evolvable Hardware Accelerators
THEME 5: DEMONSTRATORS	
3134.017	Demonstrator 1: A Reconfigurable Multi-Accelerator Compute Engine
3134.018	Demonstrator 2: A heterogeneous Large Cluster with Specialized Intelligence
3134.019	Demonstrator 3: Applications Benchmark

ACE PI Directory

NAME & AFFILIATION	ASSOCIATED TASKS	EMAIL
Josep Torrellas Center Director, Illinois	All	torrellas@cs.uiuc.edu
Tarek Abdelzaher Illinois	3134.019	zaher@illinois.edu
Mohammad Alian University of Kansas	3134.005, 3134.007, 3134.011	alian@ku.edu
Adam Belay MIT	3134.010, 3134.013, 3134.018	abelay@csail.mit.edu
Manya Ghobadi MIT	3134.008, 3134.009, 3134.011	ghobadi@csail.mit.edu
Rajesh K. Gupta UCSD	3134.002, 3134.004, 3134.016	rgupta@ucsd.edu
Christoforos Kozyrakis Stanford University	3134.003, 3134.008, 3134.009, 3134.014, 3134.018, 3134.019	christos@cs.stanford.edu
Tushar Krishna Georgia Tech	3134.001, 3134.002, 3134.008, 3134.001, 3134.017, 3134.018	tushar@ece.gatech.edu
Arvind Krishnamurthy University of Washington	3134.0006, 3134.009, 3134.011, 3134.012, 3134.018	arvind@cs.washington.edu
José F. Martínez Cornell	3134.004, 3134.005, 3134.006, 3134.007, 3134.017, 3134.018	martinez@cornell.edu
Charith Mendis Illinois	3134.002, 3134.003, 3134.009, 3134.017	charithm@illinois.edu
Subhasish Mitra Stanford	3134.015, 3134.016, 3134.017	subh@stanford.edu
Muhammad Shahbaz Purdue University	3134.003, 3134.004, 3134.009, 3134.011, 3134.012, 3134.017, 3134.018	mshahbaz@purdue.edu
Gookwon Edward Suh Cornell	3134.013, 3134.014, 3134.015, 3134.017, 3134.018	suh@ece.cornell.edu
Steven Swanson UCSD	3134.005, 3134.006, 3134.007, 3134.017, 3134.018	swanson@cs.ucsd.edu
Michael Taylor University of Washington	3134.001, 3134.004, 3134.017	profmbt@uw.edu
Mircea Radu Teodorescu Ohio State University	3134.004, 3134.013, 3134.014, 3134.015, 3134.016, 3134.017, 3134.018	teodorescu.1@osu.edu
Mohit Tiwari University of Texas at Austin	3134.013, 3134.014, 3134.015, 3134.017, 3134.018	tiwari@austin.utexas.edu

Minlan Yu Harvard University	3134.008, 3134.009, 3134.010, 3134.011, 3134.012, 3134.018	minlanyu@g.harvard.edu
Zhengya Zhang University of Michigan	3134.001, 3134.002, 3134.003, 3134.017, 3134.018, 3134.019	zhengya@eecs.umich.edu
Zhiru Zhang Cornell University	3134.001, 3134.002, 3134.003, 3134.004, 3134.007, 3134.014, 3134.016, 3134.017	zhiruz@cornell.edu

Center Administration

NAME	ROLE	EMAIL
Josep Torrellas	Center Director	torrella@illinois.edu
Minlan Yu	Assistant Director	minlanyu@g.harvard.edu
Mircea Radu Teodorescu	Director of Logistics	teodorescu.1@osu.edu
Jill Peckham	Executive Director	jpeckham@illinois.edu

Students Presenting Project Deep Dives

RESEARCH SCHOLAR	PRESENTATION TITLE & BIO
Jovan Stojkovic Illinois Email: jovans2@illinois.edu PI: Josep Torrellas Avail for Hire Date: Internship Summer 2024	**Theme 1 Deep Dive Presentation:** *Server Design in the Age of Microservices and Serverless Computing* **Bio:** Jovan Stojkovic is a fourth year PhD student at University of Illinois at Urbana-Champaign advised by Professor Josep Torrellas. Hist research focuses on the cloud computing data platforms and deployment paradigms, such as microservices and serverless computing. He explores ways to make systems fast, reliable, and efficient in a holistic manner: from the hardware up to the platform and application layers. Jovan's work has been published at top-tier computer architecture conferences, such as ISCA, ASPLOS and HPCA. He was awarded the Kenichi Miura Award for excellence in High-Performance Computing. Prior to joining UIUC, Jovan completed his undergraduate studies at the University of Belgrade and graduated as the best student of his class.
Xiyuan Zhang UCSD Email: xiz032@ucsd.edu PI: Rajesh Gupta Available for Hire Date: 5/2024	**Theme 1 Deep Dive Presentation:** *ML on the Edge: Physics-Informed Data Denoising for Real-Life* **Bio:** Xiyuan Zhang is a Ph.D. student at Computer Science and Engineering, University of California, San Diego. She is advised by Prof. Rajesh Gupta and Prof. Jingbo Shang. Prior to UCSD, she obtained her B.S. degree in Computer Science with honors from Zhejiang University in 2020. Her research interests are in robust and efficient machine learning for sensing systems. She has received the Qualcomm Innovation Fellowship and has been selected as CPS Rising Star. She has also held internships in AWS AI Labs, MIT and UC Davis

Suyash Mahar UCSD Email: smahar@ucsd.edu PI: Steven Swanson Available for Hire Date: NA	**Theme 2 Deep Dive Presentation:** *Telepathic Datacenters: Efficient and High-Performance RPCs using Shared CXL Memory* **Bio:** Suyash Mahar is a fourth-year Ph.D. student at UC San Diego interested in the datacenter's memory efficiency. He has worked with Google, Meta, and Intel on datacenter efficiency, studying their memory hierarchy and acceleration opportunities. Before starting his Ph.D. program, he worked on architecture and safety of persistent memories at the University of Virginia, CMU, and Technion. His works on memory systems have appeared in Eurosys, ASPLOS, PACT, and ICCD.
Mark Zhao Stanford Email: myzhao@stanford.edu PI: Christos Kozyrakis Avail for Hire Date: 6/2024	**Theme 3 Deep Dive Presentation:** *End to End Optimization of Large-scale ML Training* **Bio:** Mark is a Ph.D. student at Stanford, advised by Christos Kozyrakis. His research centers around improving the scalability, performance, and security of systems for datacenter-scale applications such as machine learning. He was recently a visiting researcher at Meta, where he worked on data infrastructure for ML training. Mark was selected as an MLCommons ML and Systems Rising Star, and he is supported by a Stanford Graduate Fellowship and a Meta Ph.D. Fellowship.

Shijia Wei Univ of Texas Austin Email: shijiawei@utexas.edu PI: Mohit Tiwari Available for Hire Date: 1/2024	**Theme 4 Deep Dive Presentation:** *Understanding Security Domains and their Implications for Architects* **Bio:** Shijia is a final-year PhD student working with Professor Mohit Tiwari at UT Austin. Shijia’s research interests lie in computer architecture and systems security with a goal of bridging the gap between application-layer security requirements and low-level hardware/system primitives. His thesis currently focuses on microarchitectural side channels. Before joining UT, Shijia obtained his Bachelor's degree from Zhejiang University.

Theme 1 & Application Benchmarks Poster Session

RESEARCH SCHOLAR	POSTER DETAILS
Yixiao Du Cornell Email: yd383@cornell.edu PI: Zhiru Zhang Avail for Hire Date: 5/2024	**Title:** *Building Evolvable Accelerators for Sparse Data Processing* **Abstract:** As general-purpose scaling yields diminishing benefits and modern applications become increasingly data intensive, there has been a surge of research focused on using specialized hardware to accelerate sparse workloads. This poster presents our recent research efforts on building efficient yet versatile sparse accelerators, which aim to strike a balance between domain specialization and adaptability to accommodate the rapidly evolving application requirements and technological capabilities. We will begin with GraphLily, an FPGA-based graph processing overlay leveraging the GraphBLAS abstraction to accelerate a rich set of graph processing algorithms. Next, we will demonstrate our latest efforts to develop a versatile sparse accelerator that supports a broader range of sparse linear algebra kernels and compute patterns. Additionally, we will outline our ongoing work on developing a unified abstraction to support a multitude of sparse formats that are customized for varying degrees and patterns of sparsity.

Gerasimos Gerogiannis
Illinois

Email: gg24@illinois.edu
PI: Josep Torrellas
Avail for Hire Date: 6/2026

Title: *SPADE: A Flexible and Scalable Accelerator for SpMM and SDDMM*

Abstract: The widespread use of Sparse Matrix Dense Matrix Multiplication (SpMM) and Sampled Dense Matrix Dense Matrix Multiplication (SDDMM) kernels makes them candidates for hardware acceleration. However, accelerator design for these kernels faces two main challenges: (1) the overhead of moving data between CPU and accelerator (often including an address space conversion from the CPU's virtual addresses) and (2) marginal flexibility to leverage the fact that different sparse input matrices benefit from different variations of the SpMM and SDDMM algorithms. To address these challenges, this paper proposes SPADE, a new SpMM and SDDMM hardware accelerator. SPADE avoids data transfers by tightly-coupling accelerator processing elements (PEs) with the cores of a multicore, as if the accelerator PEs were advanced functional units---allowing the accelerator to reuse the CPU memory system and its virtual addresses. SPADE attains flexibility and programmability by supporting a tile-based ISA---high level enough to eliminate the overhead of fetching and decoding fine-grained instructions. To prove the SPADE concept, we have taped-out a simplified SPADE chip. Further, simulations of a SPADE system with 224--1792 PEs show its high performance and scalability. A 224-PE SPADE system is on average 2.3x, 1.3x and 2.5x faster than a 56-core CPU, a server-class GPU, and an SpMM accelerator, respectively, without accounting for the host-accelerator data transfer overhead. If such overhead is taken into account, the 224-PE SPADE system is on average 43.4x and 52.4x faster than the GPU and the accelerator, respectively. Further, SPADE has a small area and power footprint.

CoAuthors: Gerasimos Gerogiannis, Serif Yesil, Damitha Lenadora, Dingyuan Cao, Charith Mendis and Josep Torrellas

Ahan Gupta
Illinois

Email: ag82@illinois.edu
PI: Charith Mendis
Avail for Hire Date: 5/15/2024

Title: FLuRKA: Fast fused Low-Rank & Kernel Attention

Abstract: Many efficient approximate self-attention techniques have become prevalent since the inception of the transformer architecture. Two popular classes of these techniques are low-rank and kernel methods. Each of these methods has its own strengths. We observe these strengths synergistically complement each other and exploit these synergies to fuse low-rank and kernel methods, producing a new class of transformers: FLuRKA (Fast Low-Rank and Kernel Attention). FLuRKA provide sizable performance gains over these approximate techniques and are of high quality. We theoretically and empirically evaluate both the runtime performance and quality of FLuRKA. Our runtime analysis posits a variety of parameter configurations where FLuRKA exhibit speedups and our accuracy analysis bounds the error of FLuRKA with respect to full-attention. We instantiate three FLuRKA variants which experience empirical speedups of up to 3.3x and 1.7x over low-rank and kernel methods respectively. This translates to speedups of up to 30x over models with full-attention. With respect to model quality, FLuRKA can match the accuracy of low-rank and kernel methods on GLUE after pre-training on wiki-text 103. When pre-training on a fixed time budget, FLuRKA yield better perplexity scores than models with full-attention.

Damitha Lenadora
Illinois

Email: damitha2@illinois.edu
PI: Charith Mendis
Avail for Hire Date: 5/2026

Title: *SENSEi: Input Sensitive primitive compositions for GNNs*

Abstract: Graph neural networks (GNN) have become an important class of neural networks that have gained popularity in domains such as social and financial network analysis. As a result, there have been many frameworks and optimization techniques proposed in the literature to accelerate GNNs. However, getting consistent high performance across many input graphs with different sparsity patterns and embedding sizes has remained difficult. In this paper, we observe that different algebraic reassociations of GNN computations lead to interesting input-sensitive performance characteristics. We use these observations to introduce novel dense and sparse matrix primitive compositions targeting convolution-based and attention-based GNNs and show how their profitability changes with the input graph, embedding size, and target hardware. We developed SENSEi, a system that uses a data-driven adaptive strategy to select the best composition given the input graph and embedding sizes. Our evaluations on a wide range of graphs and embedding sizes show that SENSEi achieves speedups on canonical convolution and attention-based GNNs, of up to 2.049× and 1.153× on graph convolutional networks, and up to 51.123× and 6.868× on graph attention networks, on CPUs and GPUs respectively, compared to the widely used Deep Graph Library. We also show that our technique generalizes and gives speedups to other convolution (SGC, TAGCN) and attention (GATv2, GaAN) based GNN variants, as well as the decisions made by SENSEi do not change across sampled graphs, enabling it to support sampled variants. Further, we show that the compositions yield notable synergistic performance improvements on top of other established sparse optimizations, such as sparse matrix tiling, by evaluating against a well-tuned baseline.

CoAuthors: Vimarsh Sathia, Gerasimos Gerogiannis, Serif Yesil, Josep Torrellas, Charith Mendis

Chunao Liu
Purdue

Email: liu2849@purdue.edu
PI: Muhammad Shahbaz
Avail for Hire Date: 5/2024

Title: *μManycore: A Cloud-Native CPU for Tail at Scale*

Abstract: Microservices are emerging as a popular cloud-computing para- digm. Microservice environments execute typically-short service requests that interact with one another via remote procedure calls (often across machines), and are subject to stringent tail-latency constraints. In contrast, current processors are designed for tradi- tional monolithic applications. They support global hardware cache coherence, provide large caches, incorporate microarchitecture for long-running, predictable applications (such as advanced prefetching), and are optimized to minimize average latency rather than tail latency. To address this imbalance, this paper proposes μManycore, an architecture optimized for cloud-native microservice environments. Based on a characterization of microservice applications, μManycore is designed to minimize unnecessary microarchitecture and miti- gate overheads to reduce tail latency. Indeed, rather than supporting manycore-wide hardware cache coherence, μManycore has multiple small hardware cache-coherent domains, called Villages. Clusters of villages are interconnected with an on-package leaf-spine net- work, which has many redundant, low-hop-count paths between clusters. To minimize latency overheads, μManycore schedules and queues service requests in hardware, and includes hardware sup- port to save and restore process state when doing a context-switch. Our simulation-based results show that μManycore delivers high performance. A cluster of 10 servers with a 1024-core μManycore in each server delivers 3.7× lower average latency, 15.5× higher throughput, and, importantly, 10.4× lower tail latency than a cluster with iso-power conventional server-class multicores. Similar good results are attained compared to a cluster with power-hungry iso-area conventional server-class multicores.

CoAuthors: Jovan Stojkovic, Muhammad Shahbaz, Josep Torrellas

Ashitabh Misra
Illinois

Email: misra8@illinois.edu
PI: Tarek Abdelzaher
Avail for Hire Date: 5/2026

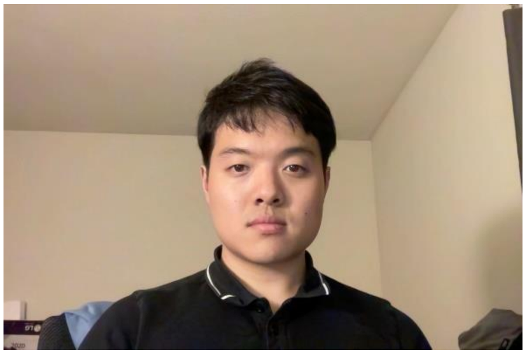
AND

Sakshi Tayal
Illinois

Email: stayal2@illinois.edu
PI: Tarek Abdelzaher
Avail for Hire Date: 8/2024

Title: *Adaptive Precision Inference for Audio Signal Classification*

Abstract: Quantization techniques have shown great promise in reducing inference times and memory footprint of Deep Neural Networks (DNNs), which are critical to real-time cyber-physical systems that run in resource-constrained environments. Some notable schemes include post-training quantization, pre-training quantization, and learnable dynamic precision quantization. Due to the dynamic nature of the operating environment of IoT devices, static fixed-point inference across the lifetime of the inference engine results in sub-optimal accuracy on out-of-distribution inputs. We propose a temporally dynamic precision inference engine for real-time audio signal classification that learns an efficient precision selection scheme that defers casting of each layer to runtime. The resulting precision is contingent on the theoretical properties of an initial fixed number of audio frames. The cost of precision selection is amortized over a predefined time period leading to an overall reduction in the number of arithmetic operations. Our framework achieves equivalent performance as static fixed-point precision quantization per inference and is robust to a wider range of input variations. The framework consists of an LSTM convolution encoder using spectrograms as the input followed by dense weight layers for classification. The model performs a casting of the initial 32-bit fixed-point layer following each predefined time period. This results in dynamic precision across layers and time. The input signal is downsampled to 1500 Hz to isolate relevant frequencies and improve performance. Initial experiments performed on ARM-Cortex M3 show that casting has a negligible contribution to the runtime. Our preliminary implementation achieves a classification accuracy of 74% for classification between two vehicles. Further hyperparameter tuning is expected to increase the accuracy.

Hyoungwook Nam Illinois  Email: hn5@illinois.edu PI: Josep Torrellas Avail for Hire Date: 5/2024	**Title:** *ML-Based Hierarchical Power Control of Serverlessoads for Sustainability* **Abstract:** In this work, we are leveraging machine learning method to perform hierarchical power control of datacenter running serverless applications. We use ML to model dynamic power and performance behavior of the system on-line. Using the ML models that are generated dynamically, we use their gradients to decide better power distribution for efficiency. Such an optimization method can scale to any hierarchy -- from sockets in a node to a multi-node cluster.
Bilal Saleem Purdue Email: bsaleem@purdue.edu PI: Muhamm Shahbaz Avail for Hire Date: 8/2025	**Title:** *Towards a Performant and Scalable Cloud-Native 5G Mobile Core Architecture* **Abstract:** To support the rapidly evolving mobile use cases (e.g., AR/VR, autonomous driving, and massive IoT), the 5G mobile core is being architected as a next-generation microservice-based workload running on edge clouds. Yet, the current proposals to improve its performance still revert to old methods and techniques used in traditional NFV-based core designs (e.g., consolidating functions on dedicated servers). In this paper, we conduct the first in-depth study of a 5G-compliant open-source mobile core (i.e., Aether) to characterize its various bottlenecks. Our measurements show that, unlike NFV-based designs, the volume of CPUs, memory, and bandwidth are not the primary bottlenecks in Aether. Instead, it is the execution time (e.g., encoding/decoding messages to/from base stations) and the contention for resources (such as Go scheduler, network IO, and synchronization) that arise due to the disaggregation of a mobile core into its smaller constituents, when serving multiple UEs. Based on these measurements, we propose a scale-out, cloud-native version of a 5G core, Aether+, which extends Aether by redesigning its stateful components (e.g., AMF and SMF) as stateless services. Doing so allows Aether+ to dynamically and independently scale these services as the UE number and traffic increase. **CoAuthors:** Jingqi Huang, Jiayi Meng, Iftekhar Alam, Ajay Thakur, Christian Maciocco, Muhammad Shahbaz and Y. Charlie Hu

Vimarsh Sathia
Illinois

Email: vsathia2@illinois.edu
PI: Charith Mendis
Date Avail for Hire: 5/2024

Title: *Exploring and Exposing Redundancy-Aware Optimizations for Temporal Graph Neural Networks*

Abstract: We address optimization challenges in the realm of dynamic graphs by focusing on Temporal Graph Attention Networks (TGATs). Despite their effectiveness in predictive tasks, existing optimization methods for Graph Neural Networks (GNNs) fall short when applied to TGATs and TGNNs. To bridge this gap, we detail optimization opportunities in TGOpt, which exploit redundancies in temporal node embedding computations. Our results led to inference speedups of up to 4.9× on CPU and 2.9× on GPU, with notable gains of 6.3× on the CPU for the Reddit Posts dataset.

We then introduce TGLite, a lightweight framework to enable the efficient construction of TGNN models on Continuous-Time Dynamic Graphs(CTDGs). To capture message flow dependencies and accommodate temporal attributes, we introduce the *TBlock* abstraction. TBlocks serve as a central representation on which many different operators can be defined, such as temporal neighborhood sampling, scatter/segmented computations, as well as optimizations tailored to CTDGs. On 4 existing TGNN models, TGLITE is able to accelerate runtime performance of training (1.06 – 3.43×) and inference (1.09 – 4.65×) across different experimental settings when compared against TGL framework.

Jianming Tong
Georgia Tech

Email: jianming.tong@gatech.edu
PI: Tushar Krishna
Avail for Hire Date: 1/2024

Title: *SUSHI: Model-System-Accelerator Co-Design for Real-Time Latency/Accuracy Navigation in Edge Applications*

Abstract: A growing number of applications depend on Machine Learning (ML) functionality and benefits from both higher quality ML predictions and better timeliness (latency) at the same time. A growing body of research in computer architecture, ML, and systems software literature focuses on reaching better latency/accuracy tradeoffs for ML models. Efforts include compression, quantization, pruning, early-exit models, mixed DNN precision, as well as ML inference accelerator designs that minimize latency and energy, while preserving delivered accuracy. All of them, however, yield improvements for a single static point in the latency/accuracy tradeoff space. We make a case for applications that operate in dynamically changing deployment scenarios, where no single static point is optimal. We draw on a recently proposed weight-shared SuperNet mechanism to enable serving a stream of queries that uses (activates) different SubNets within this weight-shared construct. This creates an opportunity to exploit the inherent temporal locality with our proposed SubGraph Stationary (SGS) optimization. We take a hardware-software co-design approach with a real implementation of SGS in SushiAccel and the implementation of a software scheduler SushiSched controlling which SubNets to serve and what to cache in real-time. Combined, they are vertically integrated into SUSHI---an inference serving stack. For the stream of queries SUSHI yields up to 25% improvement in latency, 0.98% increase in served accuracy. SUSHI can achieve up to 78.7% off-chip energy savings.

CoAuthors: Athinagoras Skiadopoulos, Zhiqiang Xie, Mark Zhao, Saksham Agarwal, Johann Hauswald, Jacob Adelmann, David Ahern, Carlo Contavalli, Michael Goldflam, Raghu Raja, Daniel Walton, Rachit Agarwal, Shrijeet Mukherjee, Christos Kozyrakis

Tianshi Wang
Illinois

Email: tianshi3@illinois.edu
PI: Tarek Abdelzaher
Avail for Hire Date: TBD

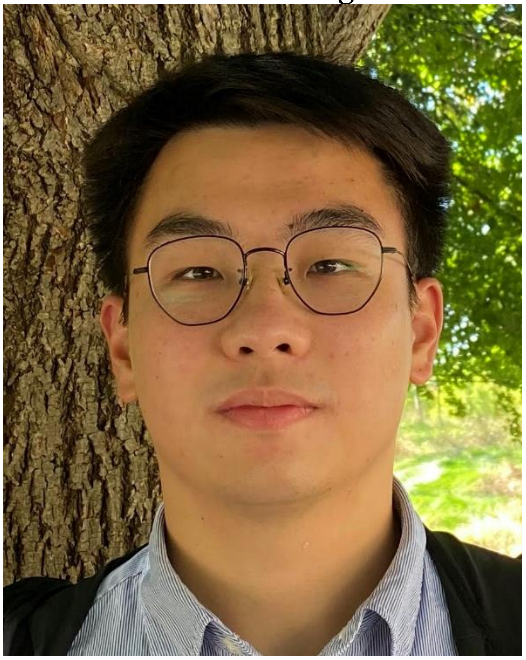
Title: *SudokuSens: Enhancing Deep Learning Robustness for IoT Sensing Applications using a Generative Approach*

Abstract:

Abstract: This poster introduces SudokuSens, a generative framework for automated generation of training data in machine-learning-based Internet-of-Things (IoT) applications, such that the generated synthetic data mimic experimental configurations not encountered during actual sensor data collection. The framework improves the robustness of resulting deep learning models, and is intended for IoT applications where data collection is expensive. The work is motivated by the fact that IoT time-series data entangle the signatures of observed objects with the confounding intrinsic properties of the surrounding environment and the dynamic environmental disturbances experienced. To incorporate sufficient diversity into the IoT training data, one therefore needs to consider a combinatorial explosion of training cases that are multiplicative in the number of objects considered and the possible environmental conditions in which such objects may be encountered.

Our framework substantially reduces these multiplicative training needs. To decouple object signatures from environmental conditions, we employ a Conditional Variational Autoencoder (CVAE) that allows us to reduce data collection needs from multiplicative to (nearly) linear, while synthetically generating (data for) the missing conditions. To obtain robustness with respect to dynamic disturbances, a session-aware temporal contrastive learning approach is taken. Integrating the aforementioned two approaches, SudokuSens significantly boosts the robustness of deep learning for IoT applications. We show that SudokuSensis general enough to benefit a variety of downstream neural network architectures and improve the performance of multiple temporal activity classification tasks.

CoAuthors: Tarek Abdelzaher

Tianyu Wei Univ. of Michigan  Email: billywty@umich.edu PI: Zhengya Zhang Avail for Hire Date: 4/2027	**Title:** *A high-bandwidth, energy-efficient chiplet interface for composable acceleration platform* **Abstract:** Integrating heterogeneous chiplets within a single package emerges as a promising and cost-effective strategy for constructing new compute platforms capable of a wide spectrum of workloads. Designing energy-efficient chiplet interfaces that satisfy the high bandwidth demands of various applications is an intricate task. In this study, we present a high-performance interface design by an automated design flow. We target the UCIe standard as our initial target. The interface design features auto-calibration and built-in self-testing to enable seamless adaptation. The automated design flow is streamlined through a series of automated steps from I/O cell synthesis, automatic place and route (APR), to the generation of bump maps and distribution of clock signals. The automations will contribute to the subsequent development of an I/O interface generator to expedite chiplet design cycle. **CoAuthors:** Wei Tang, Zhengya Zhang

Yao Yao
Illinois

Email: yaoy4@illinois.edu
PI: Josep Torrellas
Avail for Hire Date: Summer 2024

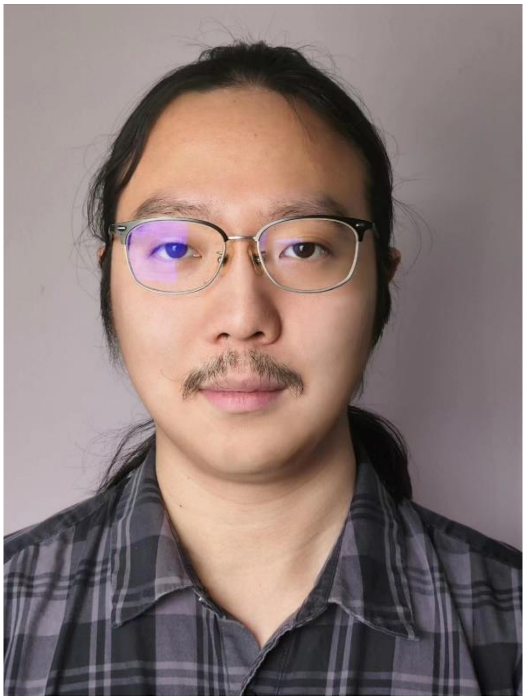
Title: *Optimize Graph Attention Network Training and Inference on CPUs*

Abstract: Traditional Deep Neural networks (DNNs) such as Convolutional Neural Networks are only applicable to Euclidean data, such as a grid of pixels in an image, but lack the power to process non-Euclidean data, such as graphs. Graph Neural Network (GNN) is a type of DNNs that specializes in processing graph structured data. It is becoming popular and has wide application domains such as Recommender Systems, Social Networks, and Knowledge Graphs. However, the performance of running these heavily memory-bound GNNs on CPUs can be limited due to the stress on memory. Typically, a GNN layer, such as in GraphSAGE and Graph Convolutional Network (GCN), is composed of a memory-intensive aggregation phase, where each vertex collects information from its neighbors, and a compute-intensive update phase, where a deep learning operator such as a fully-connected layer processes the collected information. Graph Attention Network (GAT) is a special type of GNNs that incorporates attentions on its edges to learn the importance of the neighbors for each vertex. It gives substantial performance improvement at the cost of increasing computational complexity. However, this also potentially introduces rooms for optimizations using layer-fusion techniques, where we can accelerate its execution on CPU by fusing the phase for attention calculation and the phase for aggregation such that the memory accesses can be overlapped with the computation and DRAM traffic can then be significantly reduced. Therefore, in this project, we are interested in exploring different possible ways, including layer fusion, to optimize full-batch GAT training and inference on CPUs.

CoAuthors: Zhangxiaowen Gong, Christopher W. Fletcher, Christopher J. Hughes, Josep Torrellas

Canlin Zhang Georgia Tech Email: canlinz2@gatech.edu PI: Tushar Krishna Avail for Hire Date: 1/2026	**Title**: *3D Architecture for Accelerating Memory-Intensive AI Workloads* **Abstract:** There are three noticeable trends for state-of-the-art ML Models: 1. Model sizes are increasing as the number of parameters grows exponentially; 2. Operation density is decreasing due to the popularity of attention-based language models; 3. Higher sparsity that either occur naturally (e.g., GNNs) or from model pruning, which further reduces operation density. Those trends makes ML models highly memory-bound. To mitigate this problem, prior work typically tries to increase on-chip memory bandwidth or to enable data reuse by connecting the processing elements (PEs) with flexible on-chip interconnects. However, those two approaches are difficult or costly to implement on 2D IC with significant area, energy and timing overhead. This work looks into two HW/Technology co-design approaches to tackle the memory-bound problem: 1. We propose a novel, flexible on-chip interconnect design using logic-on-logic 3D, achieving lower area overhead and higher bandwidth than its 2D IC counterparts; 2. We propose two system-level design approaches, scale-up and scale-out, to improve design scalability and to enable more data reuse opportunities in 3D IC architectures. **CoAuthors:** Gauthaman Murali, Tushar Krishna, Sung Kyu Lim
Jiayun Zhang UCSD Email: jiz069@ucsd.edu PI: Rajesh K. Gupta Avail for Hire Date:6/2024	**Title:** *Federated Learning in Heterogeneous Edge Computing Environments* **Abstract:** The growing demand for data privacy has catalyzed the rise of federated learning as a privacy-preserving distributed learning paradigm that closely integrates with edge computing. The real-world deployment of federated learning needs to deal with diverse edge computing environments such as heterogeneous running capabilities across devices. Ideally, we need models that scale to fit devices with different capabilities and design effective model aggregation methods. Existing approaches follow the idea of identifying shared patterns (i.e., layers) in local models and aggregating the common part. These methods have strong assumptions of how models are scaled, constraining the applicability to different model architectures. We propose a novel federated learning framework to enable diverse model architectures in heterogeneous edge computing environments. We replace the conventional weight-averaging aggregation with a graph hypernetwork to generate model parameters. This hypernetwork-based approach is effective in generalizing model weights across diverse model architectures. Moreover, our framework supports personalization to manage the heterogeneity inherent in data distribution across devices.

Junkang Zhu
Univ. of Michigan

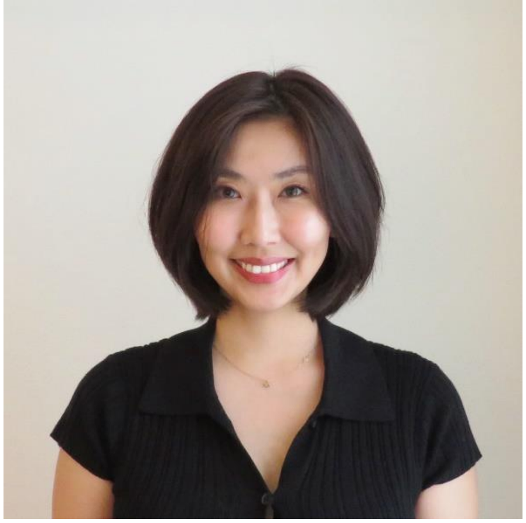


Email: jkzhu@umich.edu
PI: Zhengya Zhang
Avail for Hire Date: 10/2024

Title: *An Evolvable and Composable Chiplet Design for Future Heterogeneous Machine Learning and Big data Processing*

Abstract: Future machine learning and big data processing require both intensive computation power and support for extensive heterogeneous computation kernels. New hardware accelerators for such applications rely on evolvability and composability to fulfill these demands. Evolvability allows an accelerator to be reconfigured and reprogrammed to support a wide range of heterogeneous computation kernels. With composability, an accelerator can host multiple heterogeneous computation kernels, and multiple accelerators can communicate and coordinate with each other for extensible and scalable hyperscale computing. We present ECOM, an evolvable and composable chiplet design for future heterogeneous machine learning and big data processing. The chiplet design consists of CPU tiles and evolvable CGRA tiles. The CPU tiles provide programmability to schedule heterogeneous workloads and reconfigure the CGRA tiles. Each evolvable CGRA tile contains an array of programmable processing elements (PEs) connected through a reconfigurable interconnect network. An evolvable CGRA can be reconfigured and reprogrammed to effectively and efficiently perform computation and data movement in a variety of kernels. Different computation kernels can be mapped onto one or multiple evolvable CGRA tiles to compose larger computation tasks. The composable mapping is highly flexible and can be achieved across PEs in a CGRA tile and across CGRA tiles in a chiplet. Furtherly, multiple chiplets equipped with high-bandwidth standard interfaces can communicate and coordinate with each other for hyperscale machine learning and big data processing.

Theme 2 Poster Session

RESEARCH SCHOLAR	POSTER DETAILS
Narangerelt Batsoyol UCSD  Email: nbatsoyo@ucsd.edu PI: Steven Swanson Avail for Hire Date: NA	**Title:** *DPU-accelerated Near-Storage Data Filtering* **Abstract:** In the context of data-intensive applications, transferring large datasets over constrained network links (such as the Internet) often results in performance bottlenecks. To address this issue and improve overall system performance, we introduce a novel framework leveraging Data Processing Units (DPUs) for near-storage data filtering. DPUS are specialized system-on-chip solutions that integrate high-performance CPUs, network interfaces, and programmable acceleration engines. These units facilitate the computational offloading of data filtering tasks, allowing data to be processed closer to where it is stored. Our framework operates transparently, requiring no alterations to the existing storage infrastructure, thereby maintaining flexibility and security isolation. This approach is especially well-suited for applications dealing with rapidly growing data volumes, such as database queries on Parquet files stored in data lake-houses, scientific research analytics, and the preparation of machine learning training sets. By performing data filtering close to storage, we achieve substantial reductions in data transfer volume, thereby optimizing overall system performance.

Ehsan Hajyasini UCSD Email: ehajyasini@ucsd.edu PI: Steven Swanson Avail for Hire Date: NA	**Title:** *Telepathic Datacenters: Fast RPCs With Shared CXL Memory* **Abstract:** Compute Express Link (CXL) enables memory sharing between devices, presenting opportunities to rethink application-to-application communication within data centers. We propose utilizing CXL to optimize remote procedure calls (RPCs) in microservices. Current RPCs suffer from high overheads stemming from serialization, deserialization, and data copying, which consume up to 27% of CPU cycles. We aim to mitigate this "data center tax" by designing RPC frameworks that leverage CXL shared memory. Our approach relies on CXL shared memory for communication and falls back to RDMA/TCP for datacenter scale requests. This CXL/RDMA hierarchy can be used to create a unified virtual address space in the datacenter, enabling true zero-copy messaging via pointer passing. Realizing the potential of CXL shared memory poses challenges, including isolation, signaling, and orchestration. In this study, we implement communication channels and several sandboxing, isolation, and protection mechanisms, benchmark against microservice workloads, and prototype replacements for network services. Finally, we explore different failure models for shared memory in which the server and client can fail independently. **CoAuthors:** Suyash Mahar, Zifeng Zhang, Steven Swanson
Amin Mamandipoor Kansas Email: amin.mamandi@ku.edu PI: Mohammad Alian Avail for Hire Date: 5/2024	**Title:** *SmartDIMM: In-Memory Acceleration of Upper Layer I/O Protocols* **Abstract:** With high-throughput I/O devices deployed in data- center servers, DRAM is on the path of processing the layered, asynchronous I/O software stack. In this setting, the buffer devices of memory modules are an ideal place for inline acceleration of upper-layer I/O protocols (ULPs). In this work, we architect Smart- DIMM, a platform for near-memory acceleration of ULPs. We prototype SmartDIMM using Samsung AxDIMM and implement the end-to-end offload of Transport-Layer Security (TLS) and (de)compression, two key datacenter ULPs that are categorized under datacenter tax operations. We compare the performance of SmartDIMM with CPU, SmartNIC, and PCIe-based accelerator offload implementations. Our results show that TLS offload on SmartDIMM outperforms the CPU implementation as well as SmartNIC, and PCIe-based offload configurations. Compared to a server that executes (de)compression and (en)decryption on CPU, SmartDIMM delivers 21.0%-10.28× higher request per second and 36.3%-88.9% lower memory bandwidth utilization. **CoAuthors:** John Salihu, Mohammad Alian

Neel Patel Kansas Email: nmpatel@ku.edu PI: Mohammad Alian Avail for Hire Date: 5/2025	**Title:** *Datacenter Compression Design Space Exploration* **Abstract:** As the data footprint of hyperscaler applications increases, so does the need for increased memory and storage capacity. Low-latency datacenter applications which perform in-memory computations drive memory capacity demands ever-higher [1], while Data Warehouse services report daily data ingestions reaching ~300TB/day [2]. This has left datacenters seeking new methods to mitigate TCO increases incurred by these data-intensive applications. One widely deployed solution is to use lossless compression algorithms to shrink data memory footprints, network bandwidth consumption, and storage requirements. With compression finding its way into applications ranging from binary deployment to local-DRAM fast caches, a natural question is how to perform this "datacenter tax" as efficiently as possible, minimizing cost while meeting application SLOs. Unfortunately, there is no best-size fits all approach to compression when considering the various constraints imposed by datacenter services. Compression algorithms and their corresponding parameters (e.g., compression window size) must be chosen to best suit the calling application. For some services, this means favoring decompression speeds to meet strict latency requirements, Others seek higher compression ratios to maximize space savings [3]. Here, we evaluate the trade-offs in this complex design space to inform system designs which seek to perform (de)compression efficiently while best suiting the needs of the calling application
John Salihu Kansas Email: jsalihu@ku.edu PI: Mohammad Alian Avail for Hire Date: 6/2024	**Title:** *Opportunistic Cache Cleaning for Scalable Memory Hierarchies* **Abstract:** Prior works showed that DMA leaks are frequent in the age of high bandwidth networks and proposed hardware and software techniques mitigating DMA leaks. In this work, we look at the DMA leak problem from a different perspective and argue that leakage to DRAM is not always bad if we can intelligently control the leaks. Modern CPUs implement high bandwidth interfaces to the memory, such as HBM and CXL, and experience a high degree of fluctuation in memory bandwidth utilization when running network-intensive applications. Such fluctuations are due to the bursty nature of handling network packets in the network hardware and software stack. We leverage the low utilization periods of the high bandwidth DRAM interface in modern CPUs to intelligently clean the dirty cachelines in DDIO ways, eliminating the negative impact of DMA leaks on the application performance. **CoAuthors:** Amin Mamandipoor, Mohammad Alian

Akhil Shekar Virginia From JUMP 2.0 PRISM Center collaborating with ACE on project 005 Email: as8hu@virginia.edu PI: Kevin Skadron/ José F. Martínez	**Title:** *Membrane: A PIM-based Architecture to Accelerate Database OLAP Queries* **Abstract:** This work explores application of processing- in-memory (PIM) techniques to Online Analytical Processing (OLAP) database workloads. We explore how to map queries onto subarray-level PIM, which enables parallelism across sub- arrays and banks. We systematically explore mapping strategies and trade-offs between bit-serial/element-parallel and bit- parallel/element-serial designs adapted from the prior Sieve and Fulcrum architectures, respectively. We find that join operations do not map well to subarray-level PIM architectures, and thus we need to use a software pre-join/denormalization method to transform join operations to selection/filter operations. We also learn that certain operations, such as aggregation, remain better served using the CPU. Thus, we propose a cooperative approach for analytic query processing between CPU and PIM. We then explore several dimensions in the design space of PIM architectures, including different ways to perform filter operations, and a new way to return data to the CPU. We conclude that a traditional columnar-database layout with a scalar processing element in the PIM-enabled subarrays (Membrane-H) for the table scan, combined with a rank-level unit (RLU) for gathering the selected elements, is the best configuration. An evaluation of an end-to- end query processing on the popular analytic benchmark SSB at scale factor 100 (a 60GB database) yields a 45.39× geometric-mean speedup over a hand-optimized AVX-512 implementation of SSB. **CoAuthors:** Lingxi Wu, Kevin Gaffney, Martin Prammer, Helena Caminal, Yimin Gao, Ashish Venkat, José F. Martínez Jignesh Patel, Kevin Skadron

Mingyao Shen
UCSD

Email: mis015@ucsd.edu
PI: Steven Swanson
Avail for Hire Date: 9/2024

Title: *CXL-based SSD-autonomic scheduling system*

Abstract: NAND flash memory-based solid-state drives (SSDs) have been widely used in data centers due to their better performance compared with hard disk drives (HDDs). However, SSDs do not always provide low access latency, which can be attributed to their background jobs and uneven workload distribution. This results in unstable performance and adversely affects the quality of service (QoS) requirements. To address the issue of SSDs' long tail latency, we propose an SSD-autonomic distributed scheduling system based on the new cache-coherent memory access protocol, Compute Express Link (CXL). The system employs CXL.mem and CXL.cache to provide high-performance state communication, which allows SSDs to handle scheduling work. By offloading scheduling work from host CPUs to processors in SSDs, the computing capacity required for scheduling work naturally scales with the storage capacity, even when storage devices are disaggregated. Additionally, scheduling work does not interfere with the main workloads processed on host CPUs. Since SSDs' processors manage the scheduling work, scheduling decisions can be made instantly based on SSDs' internal states, which are not visible to host CPUs for most commodity market SSDs. CXL also enables low-overhead request redirection. By carefully designing the backup method and choosing concurrent data structures, while the original SSD is busy processing normal requests or background jobs, requests can be redirected to other SSDs to mitigate the effects on latency.

Cecilio Tamarit
Cornell

Email: ct652@cornell.edu
PI: José F. Martínez
Avail for Hire Date: 2027

Title: *Increasing the Efficiency of Associative Processors via CMOS-Compatible Hybridization*

Abstract: Associative processors (AP) have recently re-emerged as an appealing architecture that provides vast amounts of data-level parallelism. Internally, APs carry out arithmetic and logic operations on very long vectors (tens of thousands of elements or more) via sequences of bulk search and update operations, without the need for ALU circuitry. Emerging memory technologies (EMTs) could further enhance APs through gains in density and energy efficiency. However, EMTs often suffer from slower write speeds, higher write energy costs or lower endurance. High write latencies and wearout levels, in particular, can be lethal to APs' performance and endurance, as most arithmetic and logic operations involve multiple bulk updates. In this work, for the first time, we propose a hybrid CMOS-EMT AP solution that reaps the energy and area advantages of EMTs in addition to the performance and endurance benefits of CMOS. A small fraction of the total AP vector register storage is implemented in CMOS, which the microarchitecture engages selectively to take advantage of CMOS' faster and more resilient writes. At the same time, a FeFET-based organization serves as the primary storage of the vector registers, resulting in significant area-delay-power (ADP) improvement over a full-CMOS implementation. All of this is transparent to the programmer, as it requires no changes to the ISA or the program. We evaluate our proposed mechanism using a sophisticated cycle-approximate execution-driven simulation infrastructure. Results show that our hybrid AP design is hardly 1\% slower than a full CMOS implementation while at the same time achieving a 2.29x ADP^{-1} improvement over a pure FeFET design (1.11x ADP^{-1} improvement over pure CMOS) and essentially eliminating the FeFET design's endurance disadvantages.

CoAuthors: Socrates Wong, Dayane Reis, Xiaobo Sharon Hu, Michael Niemier, José Martínez

Johnson Umeike
Kansas

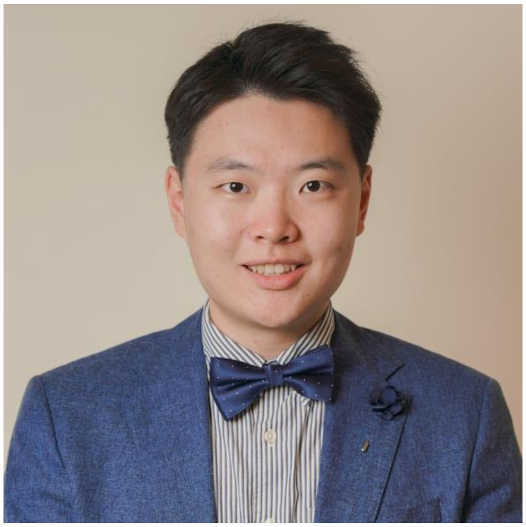
Email: johnson.chinedu@ku.edu
PI: Mohammad Alian
Avail for Hire Date: 5/2024

Title: *Userspace Networking in gem5*

Abstract: Full-system simulation of computer systems is critical to capture the complex interplay between various hard- ware and software components in future systems. Modeling the network subsystem is indispensable to the fidelity of the full- system simulation due to the increasing importance of scale- out systems. The network software stack has undergone major changes over the last decade, and kernel-bypass networking stacks and data-plane networks are rapidly replacing the conventional kernel network stack. Nevertheless, the current state-of- the-art architectural simulator, gem5, still uses kernel networking which precludes realistic network application scenarios. In this work, we first show the limitation of gem5's current network stack in achieving a high network bandwidth. Then we enable kernel bypass networking stack on gem5. We extend gem5's NIC hardware model and device driver to enable the support for userspace device drivers to run the DPDK framework. We also implement a network load generator hardware model in gem5 to generate various traffic patterns and perform per-packet timestamp and latency measurements without introducing packet loss. We develop a suite of five networking micro-benchmarks for stress testing the host network stack. These applications can run on both gem5 and a real system with a fast turnaround for gem5. Our experimental results show that enabling userspace networking improves gem5's network bandwidth by 5.4× compared with the current Linux software stack. We characterize the performance differences when running the DPDK network stack on a real system and gem5 and evaluate the sensitivity of DPDK performance to various system and microarchitecture parameters. This work is the first step in refactoring the networking subsystem in gem5.

CoAuthors: Siddharth Agarwal, Derrick Quinn, Nikita Lazarev, Mohammad Alian

Kailin Yang
Cornell



Email: ky362@cornell.edu
PI: José Martínez
Avail for Hire Date: 5/2024

Title: *VersaTile: Flexible Tiled Architectures via Processing-Using-Memory Cores*

Abstract: As modern applications demand more data, processing-in-memory (PIM) and processing-using-memory (PUM) architectures have emerged to address the challenges of data movement and parallelism. In this paper, we propose VersaTile, a heterogeneous, fully CMOS-based tiled architecture that combines conventional out-of-order (OoO) superscalar CPUs and processing-using-memory (PUM) cores, both leveraging the RISC-V ISA and its standard vector extensions for vector-SIMD execution. VersaTile fosters collaboration between multiple low-latency CPUs and high-throughput PUM cores by sharing the same software stack and adopting a CPU programming and compilation frontend. Moreover, we introduce PUM Fusion, a mechanism enabling the aggregation of multiple PUM cores' memory arrays into a single vector super-unit with modest hardware support and no programming effort, to pursue optimal performance across a wide range of applications. We provide a detailed case study including a scalable floorplan example, as well as a comprehensive evaluation over various design points. Our experiments show that when only using PUM cores, VersaTile can achieve, on average across the Phoenix benchmark suite and 3D convolution, a 5.7× speedup with respect to area-equivalent OoO CPU cores with SIMD ALUs (up to 23×), and 4.6× with respect to an equivalent-sized monolithic PUM baseline (up to 29×). For the apps with both DLP (vector) and ILP (scalar) regions, VersaTile can use PUM and OoO cores collaboratively to achieve better performance than solely using either one of them, up to 4.4×.

Theme 3 & Application Benchmarks Poster Session

RESEARCH SCHOLAR	POSTER DETAILS
Charles Block Illinois Email: coblock2@illinois.edu PI: Josep Torrellas Avail for Hire Date: 2027	**Title:** *Two-Face: Combining Collective and One-Sided Communication for Efficient Distributed SpMM* **Abstract:** Sparse matrix times dense matrix multiplication (SpMM) is commonly used in applications ranging from scientific processing to graph neural networks. Often, when this operation is performed in a distributed system, the communication costs dominate due to poor data reuse. Prior work has investigated algorithms that execute data transfers in a sparsity-unaware manner or in a sparsity-aware manner. In the former category, techniques such as collectives or shifting algorithms are employed to transfer data in a coarse-grained manner without considering the input sparsity pattern. In the latter category, the locations of input sparse matrix nonzeros determine asynchronous, fine-grained accesses. Although both can be effective, each of these approaches contains pitfalls. On the one hand, sparsity-unaware transfers can lead to unnecessary data transfers. On the other hand, sparsity-aware transfers typically carry a high software overhead and require more network round-trips. We claim that a combination of the two communication flavors can produce a more efficient distributed SpMM kernel. Towards this goal, we utilize MPI collectives for larger, contiguous data transfers, and finer-grained asynchronous one-sided communications for residual data. We propose and implement an algorithm, Two-Face, which partitions the input into a collective portion and a one-sided portion. We describe how this algorithm can be calibrated, and detail its implementation using MPI and OpenMP. We evaluate Two-Face against several baselines using large real-world sparse matrices and show that Two-Face displays an average speedup of 1.99x over the next-best baseline. Additionally, we compare Two-Face's scaling behavior to our best performing baseline and show that Two-Face scales well with the number of nodes in the system. **CoAuthors:** Gerasimos Gerogiannis, Charith Mendis, Ariful Azad, Josep Torrellas

Ajay Brahmakshatriya
MIT

Email: ajaybr@mit.edu
PI: Manya Ghobadi
Avail for Hire Date: 8/2025

Title: *LAKEPLACID: Compiling Datacenter Applications to the Microsecond Latency Regime*

Abstract: We present LAKEPLACID, a compiler-based framework that enables data center applications with legacy TCP/UDP sockets to achieve µs-scale latency with minimal programmer effort. LAKEPLACID leverages an important but perhaps overlooked observation: depending on the workload, only a small fraction of the code impacts the overall performance of some networking applications. We refer to this small fraction of the application as the EliteCode, and use a custom-designed compiler to automatically identify parts of the application logic that belong to the EliteCode. LAKEPLACID automatically transforms the EliteCode to run inside the kernel using an optimized TCP/UDP network stack. To enable data center operators to fine-tune EliteCode's behavior, LAKEPLACID's approach is parameterized, reconfigurable, and automatic. We implement three µs-scale applications using LAKEPLACID: Memcached, NGINX, and an echo server. Our evaluations demonstrate that LAKEPLACID achieves 2.55µs median round-trip latency, which is on par with the performance of eRPC and Demikernel. LAKEPLACID realizes this low latency while requiring the developer to change only ≈0.5% of the code, leaving the rest of code optimizations to its compilers

CoAuthors: Manya Ghobadi and Saman Amarasinghe

Gohar Irfan Chaudhry
MIT

Email: girfan@mit.edu
PI: Adam Belay
Avail for Hire Date: NA

AND

Zain Ruan
MIT

Email: zainruan@mit.edu
PI: Adam Belay
Avail for Hire Date: NA

Title: *Towards Self-Balancing Cloud Storage*

Abstract: In today's cloud, the best available option for high performance storage is to dedicate a locally attached flash device to a specific workload. This is necessary because flash performs poorly when it is shared across tenants. Unfortunately, this leaves the flash device mostly idle because it is unlikely enough demand will be generated to saturate it. We propose a new self-balancing approach to cloud storage that can efficiently share flash resources among many tenants over the network. Our goal is to drive up utilization while delivering performance that is equivalent or better to locally attached flash. However, to realize this vision, we must overcome two challenges. First, sharing flash often leads to hotspots, which can cause long delays in accessing disk blocks. Second, when mixing reads and writes on the same device, flash suffers from a collapse in throughput and higher tail latency. To resolve these problems, we propose fine-grained request steering and adaptive block replication/migration to prevent hotspots and segregate reads and writes onto specific flash devices. Our preliminary analysis suggests our solution has the potential to improve utilization by 300% without impacting performance.

CoAuthors: Zhenyuan Ruan, Adam Belay

Vic Feng
Harvard

Email: wfeng@g.harvard.edu
PI: Minlan Yu
Avail for Hire Date: 6/2026

Title: *F3: Fast and Flexible FPGA-based Network Telemetry*

Abstract: Traffic monitoring in the dataplane is vital for reacting to network events such as microbursts, incast, and attacks. However, current solutions are constrained by the limited resources available on modern ASICs and may not provide the flexibility required to identify repeating patterns, such as applications whose flows communicate with a server at regular intervals. While such flexibility can be achieved using a co-processing CPU, it is generally too slow to provide insights quickly enough. In this paper, we show how an FPGA co-processor placed alongside the switching pipeline enables flexible traffic mon- itoring at data plane rates. While FPGAs have large mem- ory and expressive processing, their throughput is signifi- cantly lower than switch ASICs. To bridge the throughput gap, we split query execution between the switch and FPGA and present methods that prevents processing all packets in FPGA. Further, our design leverages the FPGA's partial recon- figuration capabilities to allow the addition of new queries without the downtime which is required by solutions that reprogram the switch. As a result, our system misses up to 5.0x fewer DDoS attack vectors than ACC-Turbo, the state of the art on-switch solution and up to 24% fewer microburst-contributing flows for the same precision rate

Yigong Hu
Illinois

Email: yigongh2@illinois.edu
PI: Tarek Abdelzaher
Avail for Hire Date: TBD

Title: *Towards Foundation Models for Internet of Things Applications*

Abstract: Modern machine learning models have shown promising capabilities for Internet of Things (IoT) sensing applications. However, data collection and labeling are costly, and the amount of labeled training data constrains the performance and robustness of such models. Although labeled data is expensive, unlabeled data is easily obtainable when a large number of IoT sensors continuously measure the environment and stream the result. We propose to utilize this large amount of unlabeled IoT time-series data to pretrain foundation models that can be later adapted to different downstream tasks to improve the performance and robustness of IoT sensing applications. Unlike text or image, IoT data is not always meaningful because the physical phenomenon of interest may not be measured all the time. We design algorithms for data selection and model pretraining and present the early results with an example of sensing based on seismic signals.

Jinning Li
Illinois

Email: jinning4@illinois.edu
PI: Tarek Abdelzaher
Avail for Hire Date: TBD

Title: *Information-Theoretic Variational Graph Auto-Encoders for Unsupervised Belief Representation Learning and Ideology Detection*

Abstract: This project proposes a novel unsupervised algorithm for belief representation learning in social networks that jointly embeds users and content items into an underlying belief space, facilitating a number of downstream tasks, such as stance detection, stance prediction, and ideology mapping. We propose the Information-Theoretic Variational Graph Auto-Encoder (InfoVGAE) that learns to project both users and content items (e.g., posts that represent user views) into an appropriate disentangled latent space. To better disentangle latent variables in that space, we develop a total correlation regularization module, a Proportional-Integral (PI) control module, and adopt rectified Gaussian distribution to ensure the orthogonality. The latent representation of users and content can then be used to quantify their ideological leaning and predict their stances on issues. We evaluate the performance of the proposed InfoVGAE on three real-world datasets, of which two are collected from Twitter and one from the U.S. Congress voting database. The evaluation results show that our model outperforms state-of-the-art unsupervised models by reducing 10.5% user clustering errors and achieving 12.1% higher F1 scores for ideological separation of content items. In addition, we discuss the scalability bottleneck of the proposed InfoVGAE algorithm and potential improvements to speed up the proposed belief representation learning algorithm.

CoAuthors: Huajie Shao, Dachun Sun, Xinyi Liu, Ruijie Wang, Yuchen Yan, Jinyang Li, Shengzhong Liu, Hanghang Tong, and Tarek Abdelzaher

Jaehong Min Univ. of Washington Email: jaehongm@cs.washington.edu PI: Arvind Krishnamurthy Avail for Hire Date: 6/2025	**Title:** *eZNS: An Elastic Zoned Namespace for Commodity ZNS SSDs* **Abstract:** Emerging Zoned Namespace (ZNS) SSDs, providing the coarse-grained zone abstraction, hold the potential to significantly enhance the cost-efficiency of future storage infrastructure and mitigate performance unpredictability. However, existing ZNS SSDs have a static zoned interface, making them in-adaptable to workload runtime behavior, unscalable to underlying hardware capabilities, and interfering with co-located zones. Applications either under-provision the zone resources yielding unsatisfied throughput, create over-provisioned zones and incur costs, or experience unexpected I/O latencies. We propose eZNS, an elastic-zoned namespace interface that exposes an adaptive zone with predictable characteristics. eZNS comprises two major components: a zone arbiter that manages zone allocation and active resources on the control plane, a hierarchical I/O scheduler with read congestion control and write admission control on the data plane. Together, eZNS enables the transparent use of a ZNS SSD and closes the gap between application requirements and zone interface properties. Our evaluations over RocksDB demonstrate that eZNS outperforms a static zoned interface by 17.7% and 80.3% in throughput and tail latency, respectively, at most. **CoAuthors:** Chenxingyu Zhao, Ming Liu, and Arvind Krishnamurthy

Murrayyim Parvez Purdue Email: parvezm@purdue.edu PI: Muhammad Shahbaz Avail for Hire Date: 12/2025	**Title:** *NetEye: Extending the capabilities of a Programmable Switch using Time-Shifted Streams* **Abstract:** Managing and securing networks requires collecting and analyzing network traffic in real time. To this end, network operators often rely on telemetry systems and machine learning models to monitor the state of their network. These systems rely on programmable data plane targets to scale query execution. They offer high packet-processing speeds, but their limited computing and memory resources necessitate employing approximation techniques (e.g., sampling, sketches, and iterative refinement) that affect accuracy. In this paper, we explore a different way to increase the computational capacity of a programmable switch to increase the accuracy of a given system. We augment the recirculation path of a packet by leveraging the additional computational and storage capabilities of a modern near-switch. Packet recirculation helps us in resolving queries and classifying packets with a minimal hit on accuracy while incurring an acceptable delay. We introduce a buffer-based packet-header collection and storage architecture, named NetEye, that allows us to store packets worth of data streams in an efficient manner. On the near-switch device, we employ compression mechanisms, which reduces storage overhead by 92% and network bandwidth by 6.4%, allowing for dynamic resource usage on the switch. Consequently, our system supports more than fifteen multiple simultaneous queries without compromising accuracy to scale their execution. **CoAuthors:** Enkeleda Bardhi
Antonis Psistakis Illinois Email: psistaki@illinois.edu PI: Josep Torrellas Avail for Hire Date: Summer 2024	**Title:** *Efficient Recovery from Faults in Leaderless Distributed Systems* **Abstract:** In the high-performance realm of modern distributed systems, resilience against failures such as crashes and network partitions poses a significant challenge. The solution lies partly in data distribution across nodes and their durable mediums, particularly with the rising prevalence of low-latency persistent memories. The complexity increases in leaderless distributed systems that permit client requests to be served by multiple nodes. This work introduces a novel system, IASO, designed for efficient recovery in leaderless distributed systems equipped with persistent memory. IASO allows systems to harness the high performance typically associated with leaderless configurations, while also providing resilience against failures under Linearizable consistency and various persistency models. and various persistency models. **Coauthors:** Burak Ocalan, Fabien Chaix, Ramnatthan Alagappan, Josep Torrellas

Sudarsanan Rajasekaran MIT Email: rsudhir@mit.edu PI: Manya Ghobadi Avail for Hire Date:5/2025	**Title:** *CASSINI: Network-Aware Job Scheduling in Machine Learning Clusters* **Abstract:** We present CASSINI, a network-aware job scheduler for machine learning (ML) clusters. CASSINI introduces a novel geometric abstraction to consider the communication pattern of different jobs while placing them on network links. To do so, CASSINI uses an affinity graph that finds a series of time-shift values to adjust the communication phases of a subset of jobs, such that the communication patterns of jobs sharing the same network link are interleaved with each other. Experiments with 13 common ML models on a 24-server testbed demonstrate that compared to the state-of-the-art ML schedulers, CASSINI improves the average and tail completion time of jobs by up to 1.6x and 2.5x, respectively. Moreover, we show that CASSINI reduces the number of ECN marked packets in the cluster by up to 33x. **Coauthors:** Manya Ghobadi and Aditya Akella
Athinagoras Skiadopoulos Stanford Email: askiad@stanford.edu PI: Christos Kozyrakis Avail for Hire Date: 2026	**Title:** *High-throughput and Flexible Host Networking via Control and Data Path Physical Separation* **Abstract:** End-host network stacks can offer high performance or protocol flexibility, but not their combination. This limitation can largely be attributed to the tight integration of the network data and control path in current solutions. We argue that physical separation of the data and control path enables a performant and flexible host network stack. We present a co-designed hardware NIC and software stack that can execute arbitrary transport protocols anywhere (e.g., in kernel in a CPU, in user space in a CPU, or even in specialized packet processing accelerators), while asserting control over a zero-copy data path directly between the NIC and the memory of arbitrary devices (e.g., CPUs, GPUs, or other storage/compute components). **CoAuthors:** Zhiqiang Xie, Mark Zhao, Saksham Agarwal, Johann Hauswald, Jacob Adelmann, David Ahern, Carlo Contavalli, Michael Goldflam, Raghu Raja, Daniel Walton, Rachit Agarwal, Shrijeet Mukherjee, Christos Kozyrakis

Ruije Wang
Illinois

Email: ruijiew2@illinois.edu
PI: Tarek Abdelzaher
Avail for Hire Date: 5/2024

Title: *Online Inference Acceleration by Learning to Sample and Refresh on Streaming Temporal Graphs*

Abstract: This paper studies online link prediction on streaming temporal graphs, aiming to efficiently update deployed models on freshly acquired temporal data to ensure sustained long-term performance. State-of-the-art methods fall short in retaining and adapting informative knowledge distilled from existing data onto freshly gathered data for online updates, as they either cater exclusively to offline scenarios where all training data is available upfront or lack sufficient modeling of temporal information and temporal graph structures during online updates. We propose a temporal meta-training framework, namely OnlineSAFE, that extracts enduringly valuable knowledge across data collection periods during the offline phase and efficiently fine-tunes the model to encode newly emerging patterns during the online phase. To this end, we design a bi-level optimization to meta-learn the model parameters that ensure sustained long-term performance and adaptability to new data, where outer/inner loops are nested to optimize the global model parameters and the fine-tuning procedure, respectively. Considering the potentially distinct distribution exhibited in the new data, we analyze and derive an empirical bound based on the PAC-Bayes theory to enhance the stability and generalizability of the online updating process. Furthermore, we investigate a simple but effective sample reduction heuristic that accelerates online updates by bypassing edge samples that lack additional information. Extensive experiments on four real-world streaming graphs demonstrate the effectiveness and efficiency of OnlineSAFE, compared with 17 state-of-the-art baselines.

CoAuthors: Tarek Abdelzaher, Charith Mendis

Ertza Warraich
Purdue

Email: ewarraic@purdue.edu
PI: Muhammad Shahbaz
Avail for Hire Date: 6/2025

Title: Ultima: Robust and Tail-Optimal All-Reduce for Distributed Deep Learning

Abstract: Distributed Deep Learning (DDL) is the de-facto standard for training large-scale models (comprising billions of parameters) that form the backbone of numerous mainstream enterprise applications. Central to DDL's efficiency is the synchronization process, where model gradients are exchanged among workers of the distributed cluster. However, this synchronization is often hampered by stragglers --— workers that lag behind --- leading to system-wide delays. To overcome this, we introduce Ultima, a DDL framework that capitalizes on deep-learning models' inherent resiliency against some degree of gradient loss. Ultima introduces a novel approach by embracing a time-bounded, unreliable transport mechanism for DDL communication as a way to address the stragglers. Ultima pairs this transport with a novel Transpose Allreduce collective algorithm which curbs the propagation of gradient loss when using the unreliable time-bounded transport. Additional design choices in Ultima further disperse the occurred losses, contributing to the system's overall resilience. Our evaluations show that Ultima is able to achieve speed-ups of up to 60% in straggler-prone environments over state-of-the-art DDL frameworks and preserves comparable performance with these frameworks in optimal lossless environments.

CoAuthors: Omer Shabtai, Shay Vargaftik, Lalith Suresh, Matty Kadosh, Muhammad Shahbaz

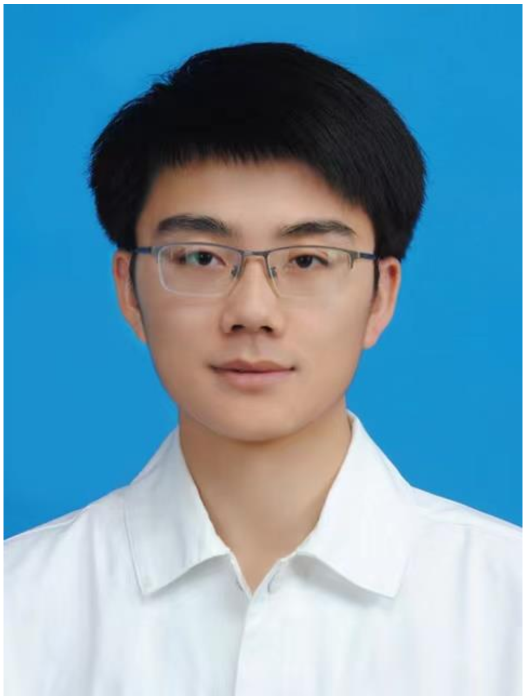
William Won
Georgia Tech

Email: william.won@gatech.edu
PI: Tushar Krishna
Avail for Hire Date: 1/2025

Title: *ASTRA-sim2.0: Modeling Hierarchical Networks and Disaggregated Systems for Large-model Training at Scale*

Abstract: As deep learning models and input data are scaling at an unprecedented rate, it is inevitable to move towards distributed training platforms to fit the model and increase training throughput. State-of-the-art approaches and techniques, such as wafer-scale nodes, multi-dimensional network topologies, disaggregated memory systems, and parallelization strategies, have been actively adopted by emerging distributed training systems. This results in a complex SW/HW co-design stack of distributed training, necessitating a modeling/simulation infrastructure for design-space exploration. In this paper, we extend the open-source ASTRA-sim infrastructure and endow it with the capabilities to model state-of-the-art and emerging distributed training models and platforms. More specifically, (i) we enable ASTRA-sim to support arbitrary model parallelization strategies via a graph-based training-loop implementation, (ii) we implement a parameterizable multi-dimensional heterogeneous topology generation infrastructure with analytical performance estimates enabling simulating target systems at scale, and (iii) we enhance the memory system modeling to support accurate modeling of in-network collective communication and disaggregated memory systems. With such capabilities, we run comprehensive case studies targeting emerging distributed models and platforms. This infrastructure lets system designers swiftly traverse the complex co-design stack and give meaningful insights when designing and deploying distributed training platforms at scale.

CoAuthors: Taekyung Heo, Saeed Rashidi, Srinivas Sridharan, Sudarshan Srinivasan, Tushar Krishna

Chenxingyu Zhao Univ of Washington  Email: cxyzhao@cs.washington.edu PI: Arvind Krishnamurthy Avail for Hire Date: 9/2025	**Title:** *Efficient Offloading Channel for DPU* **Abstract:** In this poster, we first identify four cross-PCIe Host-DPU communication primitives by analyzing the architectural peculiarities of a DPU SoC and systematically characterize their capabilities and limitations. We then design and implement a Host-DPU offloading channel by carefully synthesizing these underlying primitives and tailoring them to our requirements. Essentially, the channel operates as an adapter interface, a collection of elastic communication abstractions and an execution framework.

Annus Zulfiqar
Purdue

Email: zulfiqaa@purdue.edu
PI: Muhammad Shahbaz
Avail for Hire Date: 6/2026

Title: *Gigaflow - An Accelerator for the Slow Path at the End Host*

Abstract: Packet-processing data planes at the end-hosts have been enhanced in performance over the last decade to the point that, nowadays, they are increasingly implemented in hardware (e.g., in SmartNICs and programmable switches). However, little attention is given to the slow path residing between the data plane and the control plane, as it is not typically considered performance-critical. Recent research indicates that due to the growth in physical network bandwidth and topological complexity of modern networks, the slow path is set to become a new key bottleneck in Software-Defined Networks (SDN). We present the design and implementation of a new Domain Specific Accelerator (DSA) for the slow path at the end-host that sits between the hardware-offloaded data plane and the logically-centralized control plane. Our accelerator aims to capture most of the CPU-bound slow path traffic on virtual switches (flow cache misses from user traffic), thus reducing the load on end-host CPUs. We implement our slow path accelerator as a new caching layer in the Open vSwitch and implement its hardware-offload using NetFPGA on Xilinx Alveo data center accelerators.

CoAuthors: Venkat Kunaparaju, Ben Pfaff, Gianni Antichi, Muhammad Shahbaz

Theme 4 Poster Session

RESEARCH SCHOLAR	POSTER DETAILS
Dingyuan Cao Illinois Email: dc29@illinois.edu PI: Josep Torrellas Avail for Hire Date: 5/2026	**Title:** *ElaCache: Fine-Grain Dynamic Partitioning of Coherence Directories in Multiprocessors* **Abstract:** Cache side channel attacks pose severe security challenges in the multi-tenant cloud environment. To mitigate this vulnerability, several cache partitioning schemes have been proposed to provide isolation between mutually untrusted domains. However, none of the existing cache partition schemes incorporates the coherence directory into its partition, thus leaving this attack surface unprotected. This can lead to side-channel leakage across security domains. In this work, we propose ElaCache, which is a novel partitioning scheme that partitions both extended directory(ED) and traditional directory(TD) in order to provide a strong isolation in the cache hierarchy. To provide such isolation without impacting the application's performance, we utilize an indirection structure to provide fine-grained partitioning, while leveraging incoming memory traffic to profile applications' resource needs and adjust allocation sizes accordingly. Experiments show that ElaCache can achieve good performance while providing strong isolation compared to an unprotected cache hierarchy.

Saranyu Chattopadhyay
Stanford

Email: saranyuc@stanford.edu
PI: Subhasish Mitra
Avail for Hire Date: 2025

Title: *Pre-silicon G-QED Verification*

Abstract: G-QED -- Generalized Quick Error Detection -- is a highly thorough pre-silicon verification technique that significantly boosts design productivity. G-QED can be applied to any digital design that satisfies the following conditions: (1) actions, architectural states and idling, similar to instructions, software-visible states and idling in processors, can be defined; and (2) the content of each architectural state element can be read by an action to produce corresponding design outputs. G-QED is provably sound and complete, i.e., it detects all logic bugs without any false fails, within the capabilities of existing Bounded Model Checking (BMC) tools. Results on a wide range of processor and hardware accelerator designs demonstrate the effectiveness and practicality of G-QED. For an industrial case study using production-ready AI engines, G-QED detected 9 new critical bugs (in addition to all bugs detected by the industrial verification flow) with a drastic productivity boost -- 3 person weeks of verification effort using G-QED vs. 1 person-year using the industrial verification flow.

CoAuthors: Keerthikumara Devarajegowda, Bihan Zhao, Florian Lonsing, Brandon A. D’Agostino, Ioanna Vavelidou, Vijay D. Bhatt, Sebastian Prebeck, Wolfgang Ecker, Caroline Trippel, Clark Barrett, Subhasish Mitra

Moein Ghaniyoun
Ohio

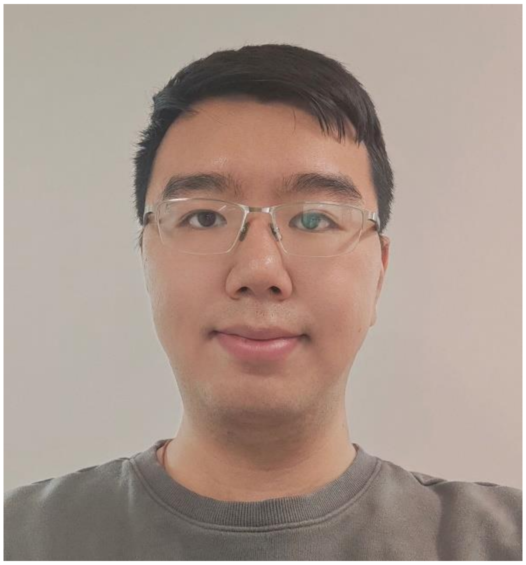
Email: ghaniyoun.1@osu.edu
PI: Radu Teodorescu
Avail for Hire Date: 5/2024

Title: *TEESec: Pre-Silicon Vulnerability Discovery for Trusted Execution Environments*

Abstract: Trusted execution environments (TEE) are CPU hardware extensions that provide security guarantees for applications running on untrusted operating systems. The security of TEEs is threatened by a variety of microarchitectural vulnerabilities, which have led to a large number of demonstrated attacks. While various solutions for verifying the correctness and security of TEE designs have been proposed, they generally do not extend to jointly verifying the security of the underlying microarchitecture. We present TEESec, the first pre-silicon framework for discovering microarchitectural vulnerabilities in the context of trusted execution environments. TEESec is designed to jointly and systematically test the TEE and underlying microarchitecture against data and metadata leakage across isolation boundaries. We implement TEESec in the Chipyard framework and evaluate it on two open-source RISC-V out-of-order processors running the Keystone TEE. Using TEESec we uncover 10 distinct vulnerabilities in these processors that violate TEE security principles and could lead to leakage of enclave secrets.

CoAuthors: Kristin Barber, Yuan Xiao, Yinqian Zhang, Radu Teodorescu

Pengzhi Huang
Cornell



Email: ph448@cornell.edu
PI: Edward Suh
Avail for Hire Date: 6/2025

Title: *STAMP: Efficient Privacy-Preserving Machine Learning With Lightweight Trusted Hardware*

Abstract: In this poster, we present a new secure machine learning inference platform assisted by a small dedicated security processor, which will be easier to protect and deploy compared to today's TEEs integrated into high-performance processors. (i) We achieve significant performance improvements compared to state-of-the-art distributed Privacy-Preserving Machine Learning (PPML) protocols, with only a small security processor that is comparable to a discrete security chip such as the Trusted Platform Module (TPM) or on-chip security subsystems in SoCs similar to the Apple enclave processor. (ii) Our platform guarantees security with abort against malicious adversaries under honest majority assumption. (iii) Our technique is not limited by the size of secure memory in a TEE and can support high-capacity modern neural networks like ResNet18 and Transformer. While previous work investigated the use of high-performance TEEs in PPML, this work represents the first to show that even tiny secure hardware with really limited performance can be leveraged to significantly speed-up distributed PPML protocols if the protocol can be carefully designed for lightweight trusted hardware.

CoAuthors: Thang Hoang, Yueying Li, Elaine Shi, G. Edward Suh

Saikat Majumdar
Ohio State

Email: majumdar.42@osu.edu
PI: Radu Teodorescu
Avail for Hire Date:

Title: *Voltage Noise-Based Adversarial Attacks on Machine Learning Inference in Multi-Tenant FPGA Accelerators*

Abstract: Deep neural network (DNN) classifiers are known to be vulnerable to adversarial attacks, in which a model is induced to misclassify an input into the wrong class. These attacks affect virtually all state-of-the-art models. While most adversarial attacks work by altering the classifier input, recent variants have also targeted the model parameters. This work proposes a new attack vector on DNN models that leverage computation errors, rather than memory errors, deliberately introduced during DNN inference to induce misclassification. In particular, it examines errors introduced by voltage noise into FPGA-based accelerators as the attack mechanism. We present an approach for precisely characterizing the distribution of faults under noise of individual input devices, by examining classification errors in select inputs. We show how, by fine-tuning the parameters of the attack (noise levels and target DNN layers) the attacker can produce the desired misclassification class, without altering the input. We demonstrate the attack on an FPGA device and show the attack success rate ranges between 80% and 99.5% depending on the DNN model.

CoAuthors: Radu Teodorescu

Mohammad Rahmani Fadiheh
Stanford

Email: fadiheh@stanford.edu
PI: Subhasish Mitra
Avail for Hire Date: 2024

Title: *A Scalable Solution for End to End Formal Verification of Millions Gate Designs*

Abstract: Scalability is the biggest hurdle in functional verification with more bug escapes as design size increases. This happens after spending a major chunk of the design project time in just verifying the design. We present a novel provably complete and scalable verification approach that can handle very large designs (over a million gates) that would otherwise not fit into a commercial formal tool. Instead of creating separate hand-crafted abstractions for each verified sub-component in a large design, Our approach uses a generic abstract model to reduce the overall complexity thereby saving time and effort. The proposed approach 1. does not need an understanding of the gory implementation details of the sub-component to be abstracted thereby drastically reducing the verification time and effort. 2. guarantees complete scalable verification for over million-gate designs. We believe that with some design discipline, false counterexamples can also be avoided in this approach. Preliminary results have shown that our approach can handle academic and industrial designs that are too large for loading into any off-the-shelf formal verification tool, including NVIDIA's 16M gate AI accelerator. Our novel abstraction technique reduces the design size (number of gates) by 10-20X. Our technique enabled the detection of new as well as previously detected bugs in these designs.

CoAuthors: Saranyu Chattopadhyay, Caroline Trippel, Clark Barrett, Subhasish Mitra

Muhammad Umar
Cornell

Email: mu94@cornell.edu
PI: Edward Suh
Avail for Hire Date: 7/2025

Title: *Efficient Memory Protection for Secure Machine Learning*

Abstract: Machine learning, especially deep learning, is a data-intensive application that can potentially consume private or sensitive data, which demands a strong security protection. A promising approach to provide strong confidentiality and integrity guarantees even under untrusted system software and potential physical tampering is to rely on trusted hardware to create a trusted execution environment (TEE). One important facility provided by TEEs is to protect sensitive data values and access patterns in the untrusted off-chip memory (DRAM). However, current techniques to protect memory incur a high overhead. In this poster, we describe our proposed techniques to lower the off-chip memory protection overhead. Firstly, to protect confidentiality and integrity of data in DRAM, TEEs use memory encryption and integrity verification, which incurs high performance overhead as it requires additional memory accesses for protection metadata such as version numbers (VNs) and MACs. To mitigate this, we exploit the simple access patterns of machine learning algorithms to generate the VNs on-chip, and optionally have MACs to protect chunks of larger granularity. As such, we propose MGX and SoftVN memory protection schemes for accelerator and processor TEEs respectively and show significant overhead reduction across a variety of deep learning benchmarks. Secondly, we study the confidentiality leakage via the memory access pattern side-channel in deep learning recommender systems, specifically via the indices of the categorical embedding tables. Typically, TEEs employ ORAM schemes to obfuscate the memory access pattern, which incurs a huge overhead especially for large table sizes. In this work, we propose to use an alternative technique to embedding tables, Deep Hash Embedding (DHE), to eliminate the input-dependent memory access pattern, as this technique has a deterministic access pattern with similar accuracy. We show some preliminary results on the overhead reduction due to using a combination of DHE and ORAM schemes for different table sizes.

CoAuthors: Weizhe Hua, Akhilesh Parag Marathe, Wenjie Xiong, Zhiru Zhang, and G. Edward Suh

Jingren Wei
Ohio

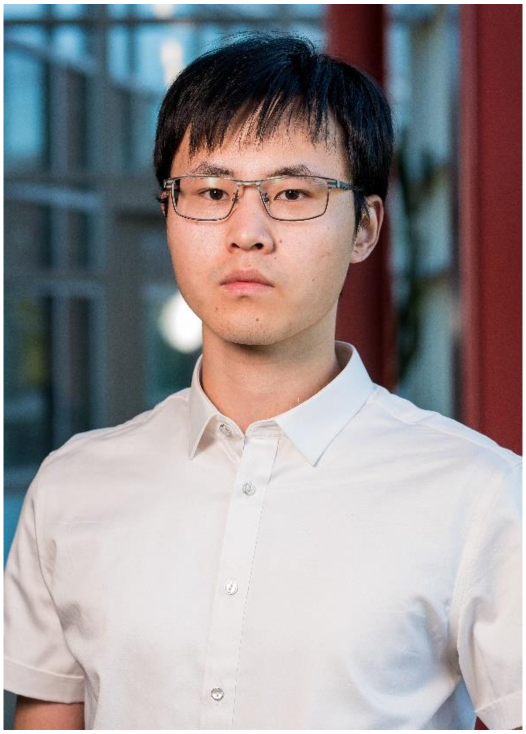
Email: wei.1276@osu.edu
PI: Radu Teodorescu
Avail for Hire Date: 2027

Title: *Adversarial Attacks on Machine Learning-Based Hardware Prefetchers*

Abstract: Machine Learning-based data prefetchers have emerged as a promising solution for capturing irregular memory access patterns more effectively than traditional rule or table-based prefetchers. However, ML models are known to be vulnerable to so-called adversarial attacks, in which inputs are manipulated to induce models to produce outputs that are beneficial to an adversary. Moreover, in order to accommodate irregular memory prefetch requests, most machine learning-based prefetchers have implemented cross-page prediction. This enables attackers to construct an adversarial memory access sequences that deceives a victim prefetcher model into making a prefetch request for a page that should be inaccessible to the attacker. We present the first comprehensive study of adversarial attacks on ML prefetchers. Evaluation on five different state-of-the-art ML-based prefetchers shows that adversarial attacks can be constructed with high success rates.

CoAuthors: Moein Ghaniyoun, Radu Teodorescu

Zirui Neil Zhao
Illinois



Email: ziruiz6@illinois.edu
PI: Josep Torrellas
Avail for Hire Date: 4/2024

Title: Everywhere All at Once: Co-Location Attacks on Public Cloud FaaS

Abstract: Microarchitectural side-channel attacks exploit shared hardware resources and pose severe threats to modern cloud environments. Achieving physical host co-location with a victim, a crucial step in these attacks, is challenging due to the widespread adoption of the virtual private cloud (VPC) and the ever-growing size of data centers. Moreover, cloud computing is increasingly moving towards Function-as-a-Service (FaaS) environments, characterized by highly-dynamic function instance placements and limited control for attackers. In this paper, we present the first comprehensive study of risks and techniques for co-location attacks in public FaaS environments. We develop two novel physical host fingerprinting techniques and propose a new, inexpensive methodology for large-scale instance co-location verification. Utilizing these techniques, we conduct an extensive study on Google Cloud Run, uncovering exploitable instance placement behaviors. Leveraging our findings, we devise a highly effective strategy for function instance launching that achieves 100\% co-location probability and covers 59\%--100\% of victim instances in three major Cloud Run data centers.

CoAuthors: Adam Morrison, Christopher W. Fletcher, Josep Torrellas