Joshua Kim

Curriculum Vitae

Department of Computer Science
University of Texas at Austin

Research Interests

- Computer architecture and hardware-software co-design for datacenter efficiency.
- Acceleration of complex problems using heterogeneous architectures.
- Compiler tools that offload resource management and parallel code from programmers.

Education

2025–Present **Ph.D Computer Science**, *University of Texas at Austin*
Advisors: Jovan Stojkovic, Christopher J. Rossbach
Topic: Designing Processor Architectures for Datacenter Servers

2021–2025 **Bachelor of Science, Computer Science & Mathematical Sciences**, *Carnegie Mellon University*, *(Dual Degree)*
Thesis: Towards a Scalable Multi-GPU System for Encrypted AI **GPA:** 3.73/4.0

Research Experience

2025-Present **Research Assistant, UT Austin**, *Advisors: Jovan Stojkovic, Christopher J. Rossbach*, Designing Processor Architectures for Datacenter Servers

2024-2025 **Undergraduate RA, Carnegie Mellon University**, *Advisor: Dimitrios Skarlatos*
Towards a Scalable Multi-GPU System for Encrypted AI

2023-2024 **CS Academy, Carnegie Mellon University**, *Advisor: David Kosbie*
Clustering in Plagiarism Detection

Scholarships & Awards

2025 **Allen Newell Award for Excellence in Undergraduate Research**, Carnegie Mellon University

2024-2025 **CRA Outstanding Undergraduate Research Awards - Honorable Mention**, Computing Research Association

Technologies

Languages: C, C++, CUDA, Python, Java, OCaml, JavaScript/TypeScript, SML, x86/64 Assembly, SQL

Hardware: CPU, GPU, AWS Trainium

Benchmarking Technologies: DCPerf, Intel Perfspect, VTune, CMT-CAT, Nvidia-SMI, Nsight Systems

Selected Projects

1. Designing Processor Architectures for Datacenter Servers at Hyperscale

Modern datacenters run workloads with varying characteristics, and with the rise of the 'microservice' paradigm, we see differences in the components of a single application as well - resulting in sub-optimal performance and resource management. We propose a heterogeneous chiplet-based manycore architecture for datacenter servers, where each chiplet has different hardware configurations (e.g. cache sizes). When an application runs on the server, our system detects phase changes within an application and migrates the execution of the application to a chiplet that has the optimal hardware configuration for the current phase.

2. Scalable Multi-GPU System for Encrypted AI

Fully homomorphic encryption (FHE) is a cryptographic scheme that offers user data privacy by supporting computations directly on encrypted data - however, it suffers a severe performance overhead of four orders of magnitude compared to plaintext computations. To make FHE feasible in modern datacenters, we employ a scale-out approach on multi-GPU systems. We provide a GPU-targeting acceleration backend, along with a set of kernel-fusion, cross-device communication, and memory management techniques to address overheads unique to GPU implementations of FHE. *(Work is currently in process of submission to ISCA'26)*

Teaching Experience

Undergraduate Teaching Assistant, Carnegie Mellon University

Fall 2024 – Summer 2025	15-213/513: Introduction to Computer Systems *(Head TA)*
Spring 2025	15-418/618: Parallel Computer Architecture
Fall 2023 – Spring 2024	21-259: Calculus in Three Dimensions