CECS
CENTER FOR EMBEDDED & CYBER-PHYSICAL SYSTEMS
UNIVERSITY OF CALIFORNIA · IRVINE

CECS Seminar



"Unleashing Multi-GPU Computing to the Next-Level"

Dr. Xulong Tang

University of Pittsburgh

Wednesday, October 15th
2:00-3:00 p.m.
Location: EH 2430

Abstract:

Modern computing platforms increasingly rely on multi-GPU systems to deliver the performance, bandwidth, and memory capacity required by data-intensive applications spanning cloud and edge services to large language models (LLMs) and quantum workloads. However, scaling GPUs beyond single-device encounters fundamental system-level challenges, including address translation overheads, complex memory management, and inefficient multi-tenancy in shared environments. In this talk, I will present our recent efforts on hardware-software co-design to overcome these bottlenecks and unleash the full potential of multi-GPU computing. I will discuss novel techniques such as lightweight invalidations for page translation, short-circuiting page table walks, and multi-instance GPU (MIG) reconfiguration strategies that collectively deliver substantial performance and efficiency gains. Looking forward, I will highlight emerging research directions in scalable heterogeneous GPU architectures and memory management in GPU virtualization. An especially exciting avenue is applying these insights to LLMs, where efficient GPU memory use and multi-GPU coordination are critical for KV-cache reuse, reduced serving latency, and scalable throughput. Together, these innovations pave the way for the next generation of GPU systems that can power both large-scale scientific computing and next-generation intelligent applications.

Biography:

Dr. Xulong Tang is an Associate Professor in the Department of Computer Science at the University of Pittsburgh. His research spans computer architecture and systems, with a focus on multi-GPU architectures, scalable memory management, and heterogeneous computing across high-performance, edge, and quantum systems. His group has pioneered advances in GPU memory management, temporal graph neural network training, and quantum compilation. Dr. Tang has published extensively in premier venues such as MICRO, HPCA, ISCA, ASPLOS, PLDI, ICLR, and NeurIPS, and his work has received recognition including an Intel Research Award, an ICLR Spotlight Award, and a MICRO Best Paper Nomination. He is also active in the community, serving as General Co-Chair of ASPLOS 2026 and Vice Chair of the IEEE Micro Top Picks 2025 Selection Committee.

Hosted By: Prof. Sitao Huang