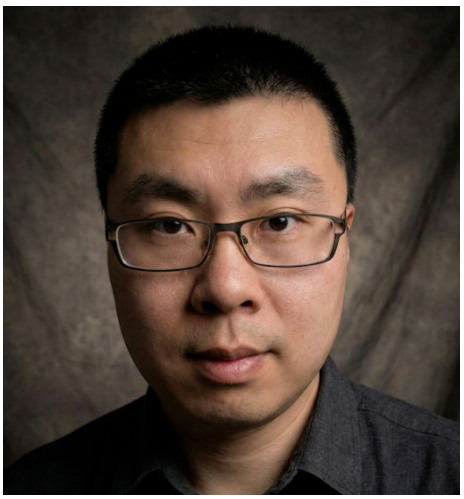
AI Hardware Faculty Candidate Seminar



Yue Cheng
Associate Professor
University of Virginia

Thursday, March 5th
10:30 A.M • MSEE 112

Sustaining AI Infrastructure through Model-Aware Compression

Abstract
Large-scale model hubs (e.g., Hugging Face, ModelScope) and numerous private repositories collectively host millions of pretrained and fine-tuned models, primarily large language models (LLMs). As the de facto infrastructure for AI development and model sharing, these platforms support a vast ecosystem of downstream applications across research and industry. However, their storage footprint has grown explosively---Hugging Face alone hosted over 77 PB of model artifacts by late 2025 and continues to expand exponentially, posing mounting sustainability challenges.

In this talk, I will present a new perspective on sustainable AI model storage that moves beyond generic data reduction toward model-aware compression. I will first show that fine-tuned models within a model family exhibit numerical similarity, revealing latent redundancy that traditional storage pipelines fail to exploit. Building on this insight, I will introduce ZipLLM, which redesigns storage reduction around model lineage and delta compression. I will then demonstrate why model-level assumptions break down in real-world repositories and show that storage redundancy emerges at the tensor granularity. To address this limitation, I will present TensorDex, a tensor-centric model compression system that achieves significant lossless storage reduction for large-scale model hubs. Finally, I will share a vision for a tensor-centric AI infrastructure and discuss future research directions.

Bio
Yue Cheng is an associate professor of Data Science and Computer Science at the University of Virginia. His research interests include systems for AI and AI systems, serverless computing, and data and storage systems. His group has built a number of techniques to improve the efficiency, scalability, and sustainability of cloud and AI platforms. Some of his works have led to large-scale deployments and adoptions in public clouds and power the AI applications used by millions every day. He is a recipient of several awards and honors, including an Amazon Research Award (2020), an NSF CAREER Award (2021), a Meta Research Award (2022), the 2022 IEEE CS TCHPC Early Career Researchers Award for Excellence in HPC, and a Samsung GRO Award (2023).

Host
Yi Ding ~ yiding@purdue.edu

Zoom-https://purdue-edu.zoom.us/j/95866124567 ~ Meeting ID- 958 6612 4567

PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering