Anshumali Shrivastava

Session 4

Title: Dynamic Contextual Sparsity: The Next Paradigm for Scaling GenAI

Abstract:
Despite stagnation in hardware and HPC infrastructure improvements, GenAI continues to scale. The cost of ChatGPT drops by 10x annually, and DeepSeek has achieved another order of magnitude in efficiency gains. Clearly, there is significant untapped potential in optimizing chips through smarter algorithmic paradigms. This talk will explore **contextual (dynamic) sparsity**, a novel paradigm already driving orders-of-magnitude improvements in scaling. We will delve into existing advancements and uncover surprising new possibilities.

Bio: Anshumali Shrivastava is an associate professor in the computer science department at Rice University. He is also founder of two startups ThirdAI Corp and xmad.ai. Both the companies build around Dr. Shrivastavas's research has successfully raised corporate VC capital and their mission is democratizing GenAI & AI Agent for everyone. Anshumali's broad research interests include probabilistic algorithms for resource-frugal deep learning. In 2018, Science news named him one of the Top-10 scientists under 40 to watch. He is a recipient of the National Science Foundation CAREER Award, a Young Investigator Award from the Air Force Office of Scientific Research, a machine learning research award from Amazon, and a Data Science Research Award from Adobe. He has won numerous paper awards, including Best Paper Award at NIPS 2014, MLSys 2022, and Most Reproducible Paper Award at SIGMOD 2019. His work on efficient machine learning technologies on CPUs has been covered by popular press including Wall Street Journal, New York Times, TechCrunch, NDTV, etc.