Solving the Memory Wall in the AI Era

Abstract: As artificial intelligence (AI) continues to evolve, the exponential growth in model complexity and data volume is placing unprecedented demands on memory bandwidth and capacity. Overcoming the challenges posed by the memory wall is therefore essential for designing computing platforms that can sustain the rapid advancement of AI. Addressing this bottleneck requires efforts spanning algorithmic optimizations, novel architectural designs, and breakthroughs in memory technologies. In this talk, I am going to share our recent work on solutions to tackle the memory wall challenge, including compute-in-memory architecture, computational storage, and model training strategy designed to exceed the capacity limitations of GPU memory.

Bio: Dr. Chia-Lin Yang is a Distinguished Professor in the Department of Computer Science and Information Engineering at National Taiwan University (NTU). She previously held positions as the Executive Secretary of the Office of Science and Technology, leading the national technology roadmap planning in Taiwan. Dr. Yang's research focuses on computing platform design, with a particular emphasis on innovative AI hardware/software system, low-power system, and memory hierarchy/SSD storage. She served as the Editor-in-Chief of IEEE Computer Architecture Letters and the Program Chair for DAC 2025. Her research has been recognized with several best-paper awards, and was selected for inclusion in the ISCA@50 25-Year Retrospective, which highlights significant contributions to the field over the past quarter-century.