Near-Memory Compute for AI Inferencing

Matt Bromage

Abstract:
The demand for AI inferencing continues to grow, creating novel problems in the design and cost of building out these types of data centers. To address this issue, we explore the use of low-cost, remote memory connected by low-latency interconnects. By identifying and offloading appropriate inferencing tasks to smaller cores located close to remote memory, here referred to as Near-Memory Compute (NMC), we show through simulation that the overall latency of inferencing execution can be reduced. In addition, through the use of cheaper pools of remote memory, the Total Cost of Ownership (TCO) of these inferencing data centers can also be reduced. This presentation offers a forward-thinking approach to data center design, with a focus on sustainability and efficiency.