Enhancing GPU Data Transfer Efficiency via Optimization of V100 Architecture and Markov-Based Predictive Caching With Quaning

Johnson, Austin (School: Gwinnett School of Mathematics, Science, and Technology)
Azmaien, Samiel (School: Gwinnett School of Mathematics, Science, and Technology)

High energy consumption in AI data centers can be attributed to several factors; however, the most notable issue is inefficiencies within their main processing units, the graphics processing units (GPUs). The main cause of these inefficiencies is data transfer bottlenecks, or when the GPU takes more time to sort data within its memory than it takes to process data in and out of the memory. This prevents GPUs from taking advantage of memory and processing capacities, thus decreasing overall efficiency. This project aims to determine the specific architectural configurations resulting in these inefficiencies, as well as optimize cache data management using Markov Chains. A Markov Chain is a mathematical system that can represent different states. By utilizing Markov Chains to sequence memory blocks into transition states, a neural network will model and optimize the Markov Chain, enhancing the process of prefetching cache data. Specifically, this project analyzes NVIDIA's V100 GPU, a modern GPU that is equipped with high bandwidth memory, which brings with it the potential for new inefficiencies. The results of this project showcase that at higher memory capacities, the V100 GPU shows peaks in bottlenecks with minimums in latencies, suggesting an inverse relationship between bottlenecks and latencies. Finally, the integration of Markov Chains and neural network optimization achieved an 85% L1 cache hit rate, 70% L2 cache hit rate, and 30% DDR memory hit rate, reducing memory overhead by 40% and improving data transfer efficiency by 25%, demonstrating a scalable solution for GPU-driven workloads.

Awards Won: