Weaver: Streamlining LLM Inference with Spatial Accelerators

March 12, 2024

Abstract

Inference for large language models (LLMs) heavily relies on General Matrix Multiply (GEMM) and General Matrix-Vector Multiply (GEMV) operations, consuming over 80% of the total computational effort. To tackle these computational challenges, spatial accelerators have been developed. These accelerators have massive Processing Elements (PEs) designed to expedite MM and MV operations, and the PEs are connected through mesh-like network-on-chips, providing massive on-chip memory bandwidth. Notable examples include Cerebras, Dojo, TPUv5, and Tenstorrent.

Despite their potential, spatial accelerators often struggle to fully deliver on their promise for LLM inference due to the high communication and memory demands of existing GEMM and GEMV algorithms. A key issue is the uneven computation pipeline lengths within the mesh structures, leading to pipeline stragglers and bottlenecks. As a result, conventional algorithms like Cannon and SUMMA exhibit insufficient efficiency, with Cannon experiencing $O(N^2)$ communication complexity and SUMMA incurring $O(2*N^2)$ memory costs, where N^2 represents the number of PEs in an accelerator.

In this talk, we will introduce Weaver, a novel LLM inference system specifically designed to capitalize on the capabilities of spatial accelerators. At its core, Weaver employs a new scalable matrix computing approach called MeshFold. By smartly folding matrix entries and interleaving them, the MeshFold approach allows GEMM and GEMV to minimize the length of the longest pipeline spread across the PEs and achieves load balancing, with proven $O(N)$ total communication cost and $O(N^2)$ total memory cost.

We've developed a prototype of Weaver on an advanced spatial accelerator Cerebras, and are in the process of adapting it for Tenstorrent. Early experiments have shown that Weaver can improve LLM inference latency by 3 times and increase throughput by 1.8 times compared to existing systems using Cannon and SUMMA.