Proposal for Dynamic DRAM Bank Allocation with Translate Lookahead Buffer Integration

Harshit Mawandia – 2020CS10348

Executive Summary

This proposal aims to enhance the Scale-Sim Accelerator Simulator's efficiency by introducing a multi-bank DRAM model that leverages dynamic DRAM bank allocation and integrates a Translate Lookahead Buffer (TLB). This approach not only adapts to real-time memory access patterns but also anticipates future requests through prefetching, thus optimizing bandwidth utilization and minimizing latency across various neural network accelerator configurations.

Introduction

Neural network accelerators require efficient data management to improve performance. Static mapping techniques for DRAM bank allocation do not account for the variability in access patterns, leading to suboptimal memory usage. Our proposal introduces a dynamic allocation method enhanced by a TLB to address this challenge, providing a flexible and efficient solution for managing DRAM banks.

Dynamic DRAM Bank Allocation

Our dynamic allocation method analyzes DRAM access patterns in real-time, adjusting the distribution of IFMAP, OFMAP, and FILTER data across DRAM banks to minimize cross-bank access and leverage spatial locality. This strategy ensures optimal data placement based on current workloads, significantly reducing latency and enhancing bandwidth efficiency.

- **Monitoring System**: Tracks access patterns and buffer states. Analyzes patterns and reallocates DRAM bank resources dynamically.
- **Performance Evaluation**: Assesses the impact of dynamic allocation on system performance through simulation.

Translate Lookahead Buffer Integration

The integration of a TLB represents a further enhancement to our dynamic allocation strategy. By storing addresses of recently accessed data, the TLB can prefetch data into active buffers based on historical access patterns, substantially reducing DRAM latency.

- **Cache of Memory Addresses**: Maintains a record of frequently or recently used addresses.
- **Intelligent Prefetching**: Uses predictive analysis to prefetch data, ensuring it's readily available in optimal DRAM banks.
- **Efficiency Assessment**: Evaluates the TLB's effectiveness in improving overall system performance.

Implementation Plan

The implementation involves phases, including the development of the monitoring system, creation of the adaptive allocation algorithm, integration of the TLB, and comprehensive system testing to evaluate performance enhancements.

Conclusion

By adopting a dynamic DRAM bank allocation method complemented by a TLB, this proposal offers a groundbreaking approach to optimizing neural network accelerator performance. This strategy not only adapts to changing data access patterns but also proactively manages data placement and prefetching, promising a significant leap forward in accelerator efficiency.