6.111 Final Project Abstract: FPGA-based Accelerator for Vector Search

Lasya Balachandran, Sanjay Seshan
lasyab@mit.edu, seshan@mit.edu

Saturday 4th November, 2023

Abstract

With the application of graphs in large-scale modeling, such as social network analysis and image and video segmentation, among other applications, graphs are increasingly being used to encode and find complex relationships between data for machine learning models, leading to an increased need for optimization of these models. As a result, in order to better support model-specific algorithm efficiency, there has been work to create specialized hardware accelerators focusing on aspects such as memory accessing, latency, and resource allocation. However, current accelerators for graph problems are not scalable and can only be optimized for a single algorithm, such as graph random walks or matrix multiplication.

The goal of this project is to implement an accelerator for a general-use search algorithm. One of the main problems when working with large graphs is the large amount of computations, which is a reason why current accelerators have focused on optimizing only a single algorithm. One way we can reduce these computations is by running the calculations on only a subset of the graph. For example, the novel graph-based vector search algorithm iQAN uses a priority queue of certain length to approximate the most similar points and avoid brute-force-checking all of the points [1]. Graph sampling, where we select a random subset of vertices representative of the entire graph, can also be used to reduce a graph.

Due to the reprogrammability of FPGAs and following from previous work on hardware and software co-design for an inverted file index (IVF)-product quantization (PQ)-based vector search algorithm on FPGAs, we propose an FPGA-based accelerator that builds off of the iQAN and graph sampling framework to provide an interface to efficiency search graphs [2]. Some applications of this system include graph-based vector search (e.g. iQAN) for machine learning applications and graph pattern mining on the sampled graphs.

This project will build on the work of iQAN (graph-based vector search) and NextDoor (efficient graph sampling on GPUs) [1, 3]. It is also part of our UROP with Arvind and Xuhao Chen in the Computation Structures Group at CSAIL.

Design Goals

To produce a minimal viable product, we will first implement iQAN on an FPGA. We will then create an accelerator for iQAN on simple graphs that can fit in BRAM. After that, we will extend the minimal viable product to make a modular, variable-sized accelerator that can read large graphs from DRAM and SD cards.

We currently see our project requiring modules to perform the following:

1. Graph fetcher and cache
2. Distance calculation for n-dimensional vectors
3. Graph updater
4. Graph vector search (iQAN) and graph pattern mining
5. Random graph point sampler to determine starting point (not necessary for minimal viable product)

We expect these modules to change as we work more on the design phase of our accelerator. However, these modules, at a high level, are needed to build our system.

Plan

Our project will be divided as follows:

1. Create CPU implementations of iQAN and graph sampling algorithms (already in progress)
2. Implement random sampler and distance calculation (Lasya)
3. Implement graph fetcher and cache and graph updater (Sanjay)
4. Convert CPU implementation of iQAN to Verilog/SystemVerilog
5. Program sample use cases for testing performance and create reference implementations on CPU to evaluate performance on FPGA compared to CPU
6. Create a modular implementation of our system

We have specified how we tentatively expect to split components 2 and 3. We expect to have a better idea on how to split the work, including components 4, 5 and 6, in the next week or two after creating a block diagram of the modules and starting to build the system.

Relevant Papers

[1] iQAN: Fast and Accurate Vector Search with Efficient Intra-Query Parallelism on Multi-Core Architectures, 2023. https://johnpzh.github.io/assets/papers/PPoPP-2023_iQAN_Zhen.CameraReady.pdf

[2] Co-design Hardware and Algorithm for Vector Search, 2023. https://arxiv.org/pdf/2306.11182.pdf

[3] NextDoor: Accelerating Graph Sampling for Graph Machine Learning Using GPUs, 2021. https://github.com/chenxuhao/ReadingList/blob/master/sampling/NextDoor.pdf