CS 410 Project Proposal
Andrew Vamos (avamos2@illinois.edu)
Misha Ryabko (mryabk2@illinois.edu)
Benjamin De Vierno (bid2@illinois.edu)

Requirements

In your proposal, please answer the following questions:

1. What are the names and NetIDs of all your team members?
   Team Member:
   - Andree Vamos: avamos2@illinois.edu
   - Misha Ryabko: mryabk2@illinois.edu
   - Benjamin De Vierno: bid2@illinois.edu

   Who is the captain? The captain will have more administrative duties than team members.
   Misha will be acting as captain during the project.

2. What is your free topic?
   Accelerated text inversion using CUDA/Triton.

   Please give a detailed description:
   Heterogeneous inverted indexes are a popular data structure used in text retrieval in text information systems. The creation of these specialized systems pose challenges in expansive datasets as they are highly compute intensive. Our accelerated inverted index aims to offload parts of the inversion process onto the GPU. This allows the inversion to be optimized through the parallel processing power of the GPU. We plan to integrate the Triton platform https://openai.com/research/triton from Nvidia to manage the deployment of our GPU-accelerated indexing components.

   What is the task?
   Accelerate the text inversion process through the utilization of GPUs.

   Why is it important or interesting?
   Organizations collect an increasingly large number of documents and rely on quick retrieval of such documents to gather information and insights. The inversion process in inverted indexes is an important and compute intensive process in text search and retrieval. The utilization of GPUs make this process more efficient particularly on very large datasets.

   What tools, systems or datasets are involved?
   Triton, cuda, Nvidia command line toolsets (nvperf, nvcc, Nvidia-smi, etc). We are going to make a control inverted index process on the CPU, and for data sets we are going to randomly generate sample inverted index inputs.

3. Which programming language do you plan to use?

CS 410 Project Proposal
Andrew Vamos (avamos2@illinois.edu)
Misha Ryabko (mryabk2@illinois.edu)
Benjamin De Vierno (bid2@illinois.edu)

Triton/CUDA

4. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.
   Implement inverted index with memory offloading: 10 hours
   Implement parallel inverted index creation with offloading: 20 hours
   Perform benchmarking, testing: 10 hours
   Write up analysis/compare performance against prior art: 5 hours
   General meeting time for architecture, writing: 5 hours * 3 people = 15 hours

At the final stage of your project, you need to deliver the following:

- Your documented source code and main results.
- Self-evaluation. Have you completed what you have planned? Have you got the expected outcome? If not, discuss why.
- A demo that shows your code can actually run and generate the desired results. If there is a training process involved, you don't need to show that process during the demo. If your code takes too long to run, try to optimize it, or write some intermediate results (e.g. inverted index, trained model parameters, etc.) to disk beforehand.