CATEGORY: BIG DATA ANALYTICS - BD07
POSTER P5305
CONTACT NAME Mason Saucier: jmsaucier@cs.ua.edu

GPU TECHNOLOGY CONFERENCE

THE UNIVERSITY OF ALABAMA ENGINEERING

Fuzzy String Matching of Vehicle Identification Numbers in a Highly Parallel Environment

Mason Saucier, Matthew Hudnall, Brandon Dixon

CAPS CENTER for ADVANCED PUBLIC SAFETY

Goals

The goal of this project is to apply a fuzzy string matching algorithm to a vehicle identification number (VIN). The Levenshtein Distance (LD) is memory intensive, and can easily max out a CPU. By moving this to the GPU, we aim to speed up

- Speed
  - Reduce work for each thread
  - Simplify global memory usage by abusing shared memory locations.
- Simplicity
  - Only need to check a single VIN at a time.
  - Currently using csv files for data loading.

Results

Computer Specs

- CPU
  - Intel Core i7-3820 – Quad Core
  - 3.60 GHz
  - 32GB RAM
- GPU
  - NVIDIA GeForce GTX Titan
  - 6GB GDDR5

CPU vs GPU

- CPU allowed for easier memory management of dataset (12.8 million VINs)
- The GPU implementation allowed for a 13x – 15x speed increase, but could lead to memory constraints if the size of the data grew to larger sizes.
- A match of a single VIN against the dataset average around 1 second.

Next Steps

- Future optimizations could lead to a 5x speed increase on the GPU by using properties that are only applicable to VIN strings.
- All VINs after 1981 are 17 digits in length.
- Every VIN has a quick computation, called a “Check Digit,” which will verify if it is a valid VIN.
- Scalability will require more GPU resources in order to supply the users with real-time results.

Algorithm Details

Load File Contents
- Characters stored in 1-D array.
- Lengths of each entry stored in a separate array.

Levenshtein Distance
- Each thread computes the LD of a string.
- This computation is only done on a single string because of memory constraints.

Find Minimum Distance
- Each thread then stores the local minimum in shared memory.
- The blocks minimum is then stored in global memory.

Final Steps
- CPU computes the minimum LD across all blocks.
- The resulting VIN is then returned along with the distance that was computed.

Future Work

- Reduce the initial problem set to improve speed.
- Integrate with a web service for remote calling.
- Refine the algorithm to remove naïve pieces.
- Increase hardware to overcome memory constraints.