Using GPU for accelerate dataset classification

{ Marcelo de Caux } Universidade Federal Fluminense Instituto de Computação

Objectives

The goal of this work is to add GPU power to improve performance of a simple dataset classifier intended to help another work, requirements are.

1. Be able to classify more than 10 million records
2. Perform this classification process in less than 30 seconds for 10 million records
3. Run in a GTX1060 Nvidia GPU
4. Use CUDA for engage the GPU

Materials & Methods

The following materials were required to create the classifier:

- The database with Bitcoin life-span quotes
- A computer with a GTX1060 GPU
- A classification rule
- CUDA 9 or greater installed and running

The following equation is used for classification:

```
class=(Open=Close ? 512:
Open > Close ? 256:1024)+
(Min < Open ? 1:4)+
(Min < Close ? 2:8)+
(Max > Open ? 16:64)+
(Max > Close ? 32:128)
```

Codes, data and documentation can be found at Github

Web https://github.com/marcelo-caux/Classification-using-GPU

Introduction

The need for build this classifier comes up when I decide to work in a prediction model for Bitcoin quotes, the idea behind classify available records is to identify types movement records available for the prediction work, records are composed by a timestamp representing one minute of a day, the opening value (second zero), the maximum and minimum values along this minute, the closure value (at second 59). Classify records according their behavior permits analyse and even clusterize periods of time in a day of trade.

Results 2

The experiment was centered in the execution time improvement of the classification process and got the following results (better ones).

Strategy	Total Time	Total Speedup
Sequential	9.85704	1.000
Unified Memory	9.35286	1.05391
2 Streams	9.34855	1.05439

Table 1: Strategies and times

For Parallel strategies I have tried different Thread numbers looking for better performance, but 256 threads got better results.

Strategy	Thrd	Blocks	TotTime	Clas.T.
UM	128	133042	9.36744	7.72242
UM	256	66521	9.35286	7.73000
UM	1024	16631	9.36721	7.77017
2S(20)	256	3327	9.34855	7.59817
2S(40)	256	1664	9.43475	7.61154
2S(100)	256	666	9.69975	8.14614

Table 2: UM=Unified Memory/2S=2 Streams

Results 1

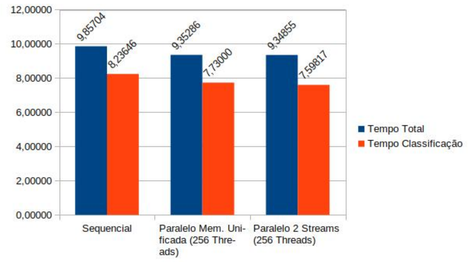

Figure 1: Execution Times for Strategies

Speed up measures, using wall clock standard, shows better numbers for the classification stage than for total process, this possibly happens because after perform the classification in GPU it's needed to copy back classification array from GPU to CPU and release more structures from memory than in the sequential approach. The maximum speedup achieved was 1.054 for total process and 1.084 for the classification step.

I have built two different parallel codes using GPU, first using Unified Memory and latter using 2 Streams, although results are better in parallel strategies, I've got only a marginal gain in processing time, compared to sequential code, for a 17 million records dataset.

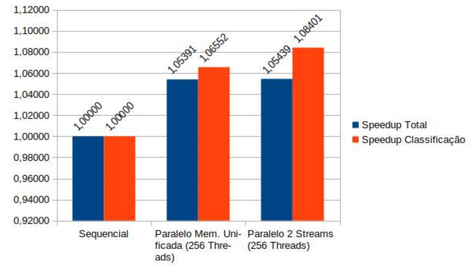

Figure 2: Achieved Speedup

Conclusion

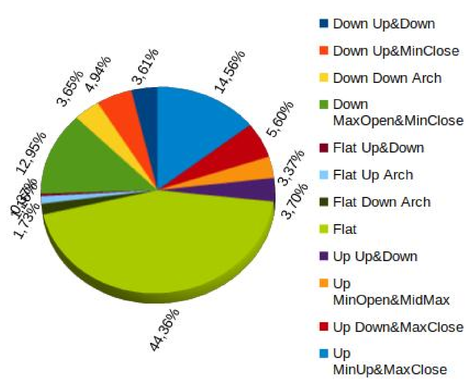

Figure 3: Final Classification

- The usage of GPU for accelerate the available dataset classification has put out poor results, some performance improvement was achieved but not at the expected level, which was reduce classification time in 50%, best results were only 5.4%.
- The Streams strategy shows best performance usage and deserves more research in using more than 2 streams, testing with 4 and 8 streams should results in better performance.
- The graph shows the data among it's classification, and will be used to better understand Bitcoin behavior on different periods of time.

References

[1] NVIDIA. Accelerated computing. https://developer.nvidia.com/accelerated-computing-training. Training page.

Future Research

Using 2 Streams shows a better performance than Unified Memory and sequential process, Try with more streams should make better execution times. Streams has another tricky part, how to divide the dataset, I've tried 20, 40 and 100 parts, will try other values can make better results.

Contact Information

Email marcelo.caux@gmail.com
Phone +55 (21) 99642-3915