

00491: The Graphics Processing Unit Enhanced Computer for Large-Scale Text Mining

Xiaohui Cui and Thomas E. Potok

This is a stage report of the Seed money project #491 started in July 2008. This report covers the work performed during the interval from July 2008 through the end of September 2008. In this research, we are conducting a proof-of-principle research of analyzing large scale datasets by using GPU enhanced computer.

Project Description:

We are quickly reaching an age in which a capability is needed for text mining (TM) of terabyte-scale unstructured text corpora for prompt decision making. Analyzing large-scale text collections requires high-performance computing and various algorithmic changes to current TM approaches. The graphics processing unit (GPU) can solve some highly parallel problems much faster than the traditional sequential processor (CPU). Thus, a deployable system using a GPU to speedup large-scale TM processes would be a much more effective choice (in terms of cost/performance ratio) than using a computer cluster. However, due to the GPU's application-specific architecture, harnessing the GPU's computational prowess for TM is a great challenge. The objective of this proposal is to prove the feasibility of utilizing the computational capabilities of GPUs to speedup TM on large scale datasets.

Mission Relevance:

In addition to the applications outlined in this proposal, our research result could be applied in intensive computer modeling, for supporting research such as climate change data processing, advanced computing architecture, risk analysis for national energy infrastructure, and confined plasma and high-energy particle beams problems that are currently of concern to DOE. The approach will provide general massive volume data analysis methods and advanced computing architecture that will be beneficial in many of the other research areas of interest. A successful result will enable us to demonstrate the use of much cheaper GPU-enhanced systems to provide massive document processing capacities. This capability will benefit to agencies, such as DHS, DARPA, and the intelligence communities, who have armies of analysts searching text on a daily basis in pursuit of the proverbial needle in a haystack.

Results and Accomplishments

In this three months period, we developed and presented a parallel Latent Semantic analysis (LSA) implementation on the GPU, using NVIDIA® Compute Unified Device Architecture and Compute Unified Basic Linear Algebra Subprograms. LSA aims to reduce the dimensions of large Term-Document datasets using Singular Value Decomposition. Implementing LSA on GPU is an important part for text mining the large scale dataset on GPU. With the ever expanding size of data sets, current implementations are not fast enough to quickly and easily compute the results on a standard PC. The performance of this implementation is compared to traditional LSA implementation on CPU using an optimized Basic Linear Algebra Subprograms library. The GPU version of

the algorithm is five to six times faster than the CPU version for large matrices (1000x1000 and above) that had dimensions divisible by 16. By using this dimension reduction technology, it is possible to fit 100k document vectors into a 1 GB graphic card for TM. Part of this research is presented in an intern student research paper, "Massively Parallel Latent Semantic Analysis Using a Graphics Processing Unit". The paper has been selected for publication in the Department of Energy's Journal of Undergraduate Research, Vol. IX. The intern student has been invited by DOE office of science to present this research at the annual meeting of the American Association for the Advancement of Science (AAAS) on February 12-16, 2009 in Chicago, Illinois. An extended discussion of this research is being prepared and will be submitted to IPDPS (23rd IEEE International Parallel and Distributed Processing Symposium). We have presented the results of our research to DHS, Navy, and FBI and we are pursuing additional funding opportunities.