OpenCL implementations of principal component analysis for large scale data analysis and benchmarking of heterogeneous clusters.

Joshua C Bowden
CSIRO Advance Scientific Computing, Melbourne, Australia, josh.bowden@csiro.au

INTRODUCTION

Programming environments for General-Purpose computation on Graphics Processing Units (GPU) have improved rapidly in the past decade. They allow a programmer to tap into the potential of GPU based devices for non-graphics tasks. As widely adopted standard, OpenCL attempts to standardize the programming of the various devices constituting a heterogeneous computing system [1]. At present these devices typically include multi-core processors and GPUs from two main vendors. Unifying the programming of these devices under the OpenCL framework is an attractive option for programmers who desire access to the computational potential of diverse hardware resources.

To illustrate the use and potential of OpenCL the work produced various implementations of the Non-linear Iterative Partial Least Squares (NIPALS) algorithm for principal component analysis (PCA)[2]. The PCA method is used widely throughout scientific disciplines for exploratory data analysis and data reduction. It provides a starting point for a variety of further analyses such as regression and classification. The NIPALS algorithm allows PCA analysis through the sequential computation of principal components (p) of a data matrix (M), returning the scores (t) for each sample for each component found [3]. This algorithm can be time-demanding for large data sets owing to its iterative nature. Example input data is from a high throughput wood analysis system named SilviScan. Resulting outputs have been used for least squares estimation of materials properties, such as microfibril angle (MFA), crystal size and fibre pitch and roll. The technique is shown to highlight areas of the images which are correlated (positively and negatively) with the property of interest as they relate to the structural parameters of the complex textured materials.

A benefit of using widely adopted standards such as OpenCL is that it allows the comparison of performance of an algorithm on a variety of modern CPU architectures and GPU based system. The OpenCL implementations of NIPALS were used as a benchmarking program to test a range hardware for the core vector-matrix operations that are at the heart of the algorithm. These operations are limited in performance primarily by memory bandwidth due to their low ratio of computation to memory accesses. These results are compareded against optimised BLAS based library implementations as part of the same benchmarking suite. Modelling of timing results was shown to give accurate prediction of performance as systems scale to cluster sizes. Various steps in the code optimization are discussed, spanning the use of: a single GPU, multiple GPUs on a single node, and multiple GPUs on multiple nodes. Both single and double precision floating point operations can be tested. Results of benchmarking workstation, cluster and cloud based solutions are described. The measurement and modelling of these workloads results in a better understanding of the economies the different systems bring to research based computation.

REFERENCES

1. Munshi, A. "*OpenCL specification version 1.0."* The Khronos Group, Available from: www.khronos.org/registry/cl, 2008.
2. Bowden, J. C., "*Application of the OpenCL API for Implementation of the NIPALS Algorithm for Principal Component Analysis of Large Data Sets*," e-sciencew, pp.25-30, 2010 Sixth IEEE International Conference on e-Science Workshops, 2010.
3. Wold, H., *Research Papers in Statistics.* Eds. F. David. New York: Wiley, 1966, pp. 411-444.

ABOUT THE AUTHOR

Dr Joshua Bowden has an Honors degree in Chemistry from Flinders University and a PhD in Materials Science from the University of South Australia. He also has two graduate Certificates in Information Technology and Object Oriented Programming gained during the completion of his PhD. He undertook a post doctoral position at CSIRO in the area of informatics systems where he developed software solutions to the problems involving analysis of large data sets used to determine wood properties from a range of analytical equipment including X-ray diffraction and near infrared spectroscopy. While finishing his PhD he worked for three years at Queensland University of Technology in the bio-materials field investigating the microstructure and chemistry of cartilage. There he developed his interest in computational problems of multivariate methods. He has been working with the CSIRO Advance Scientific Computing group for the past year providing software support to a number of high performance computing projects.