Available on-line at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Optimizing GPAW on GPUs

Martti Louhivuori[a,*], Jussi Enkovaara[a,b]

[a]CSC - IT Center for Science Ltd., PO Box 405, 02101 Espoo, Finland
[b]Aalto University, Department of Applied Physics, PO Box 11100, 00076 Aalto, Finland

Abstract

In recent years, graphical processing units (GPUs) have generated a lot of excitement in computational sciences by promising a significant increase in computational power compared to conventional processors. While this is true in many cases for small-scale computational problems that can be solved using the processing power of a single computing unit, the efficient usage of multiple GPUs in parallel over multiple inter-connected computing units has been problematic. Increasingly the real-life problems tackled by computational scientist require large-scale parallel computing and thus it is crucial that GPU-enabled software reach good parallel scalability to reap the benefits of GPU acceleration. This is exactly what has been achieved for GPAW, a popular quantum chemistry program, by Hakala *et al.* in their recent work [1].

Comment

Calculation of the dynamic electronic structure of materials is a corner-stone of computational studies of material properties and accounts for a large portion of the computing resources consumed in high-performance computing. Density-functional theory (DFT) is a widely used approach to electronic structure calculations due to its versatile nature and lower computational costs compared e.g. to Hartree-Fock based methods. One such program that uses DFT for electronic structure calculations is GPAW, a highly scalable and extensively used program that is well suited for massively parallel computing.

So far, the main bottle-neck of GPUs has been the separation of CPU and GPU memory and the low memory bandwidth between the localised memories. A typical parallel algorithm splits a large data set or simulation system into smaller pieces and assign one such sub-system to each computing unit. Each of the computing units then performs the calculations on their part of the problem and the changes or results are then propagated to the other computing units as needed. Since communication between computing units is handled by the CPU, the latency of data transfer between the CPU and GPU memories means that one can not do frequent propagation of data among the parallel computing units if one uses the GPUs for the local calculations. To overcome this one needs to either limit the amount of communication, to optimise the communication patterns e.g. by interleaving computation and communication, or to devise a way to achieve a better load-balancing at algorithmic level.

In the article "Parallel Electronic Structure Calculations Using Multiple Graphics Processing Units (GPUs)" Hakala *et al.*[1] describe the methods they have used to achieve efficient parallellisation of GPAW using GPUs and the resulting speed-up on multiple machine architectures. In collaboration with PRACE experts, they have off-loaded the most computational intensive part of the calculations to be fully on the GPU-side, limited the amount of slow memory copies between the GPU and the CPU, and optimised the communication that uses the Message Passing Interface. By using the computing resources granted by the PRACE Research Infrastructure at the CURIE supercomputer (GENCI), they have shown speed-ups up to 15 times compared to using only CPUs. For more details and full analysis of the results, see the original article [1].

The significance of this work is easy to underestimate. By demonstrating that efficient parallelisation is feasible for GPUs, and more specifically, that even large-scale DFT calculations can be accelerated by GPUs, this work has paved way for a wider utilisation of GPUs in computational quantum chemistry. Despite the fact that this wider shift from conventional processors to more efficient multi-core architectures is an on-going effort and, in many ways, just beginning in earnest, it is exactly these kind of works that give tangible evidence that accelerators may indeed deliver in practice the kind of performance boost they promise in theory.

*Corresponding author.
tel. +358-9-457-2308 fax. +358-9-457-2302 e-mail. martti.louhivuori@csc.fi

Acknowledgements

This work was financially supported by the PRACE project funded in part by the EU's 7th Framework Programme (FP7/2007-2013) under grant agreement no. RI-283493. The work was achieved using the PRACE Research Infrastructure resources at the CURIE supercomputer (GENCI).

References

1. S. Hakala, V. Havu, J. Enkovaara, and R. Nieminen, Parallel Electronic Structure Calculations Using Multiple Graphics Processing Units (GPUs), In P. Manninen and P. Öster (eds.), PARA 2012, LNCS, vol. 7782, pp. 63–76. Springer, Heidelberg (2013)