Energy Reduction on GPU-based Supercomputers by Utilizing Mixed Precision Arithmetic*

K. Rojek, R. Wyrzykowski

Czestochowa University of Technology

The energy consumption becomes critical due to the significant increase of operation costs in modern computer systems [2]. In consequence, reducing the energy consumption turns into the primary objective for scientific and industrial environments which are related to large-scale calculations. Given the rapidly climbing power bills, as well as the negative impact of energy production technologies on the environment, achieving power and energy efficiency of parallel systems and applications has become one of the most challenging issues.

One of the most common techniques used to reduce the energy consumption is DVFS (Dynamic Voltage and Frequency Scaling) [6]. However, its usage requires a special access to the machine and can be applied mostly to experimental clusters, where researchers can get *sudo* user account. It is problematic to apply this technique in supercomputing clusters, where the security is a very important aspect. For those reasons, we decided to focus on the mixed precision method [1] to reduce the energy consumption. This method can by easily adapted to applications that are executed by a large group of users, not only by a small group of researchers that execute their codes on some experimental clusters. Instead of the hardware reconfiguration such as frequency scaling, switching nodes off, powering down cores, the reduction of energy is achieved through modifications of the application code.

Using the mixed precision arithmetic is traditionally based on the static selection of precision with respect to high accuracy of results for all the possible tasks that can be performed by a given application. In our work, we focus on a more flexible alternative [3]. The proposed method assumes a dynamic selection of the data precision for each array processed by an algorithm. Our approach includes a short, self-adaptable stage that calibrates the precision of computation depending on the specific simulation. In this way, we can decrease the precision of calculation for more error-tolerant simulation and increasing it for more restrictive ones.

We verify our method for the Multidimensional Positive Definite Advection Transport Algorithm (MPDATA) [4,7] that contains a set of stencil-based computing kernels with heterogeneous patterns. Besides the CGR solver, MPDATA is the second major part of the dynamic core of the EULAG geophysical model, which was developed for simulating thermo-fluid flows across a wide range of scales and physical scenario, including numerical weather prediction, simulation of urban flows, turbulences, and ocean currents. The mixed precision arithmetic is especially useful in our research since, for example, weather forecasts simulate such quantities as temperature, pressure, velocity, etc. Their values are usually rather error-tolerant. Another advantage of this technique is that it allows us to reduce both the execution time and energy consumption of applications. It is especially useful for GPUs [5], where the ratio between performance of double and float computations is from $32:1$ (Maxwell-based GPUs) to $2:1$ (Pascal-based GPUs).

The efficient realization of the mixed precision arithmetic for a real scientific algorithm is not a trivial task. MPDATA is a complex algorithm and evaluating the right group of instructions which can be safely set to float requires to analyze a lot of possible scenarios. For this reason, we investigate machine learning methods and select the random forest algorithm as a candidate to find a high efficient configuration consisting of the precision setup (single or double) for each array of the MPDATA algorithm.

*This work was supported by the National Science Centre, Poland, under grant no. UMO-2015/17/D/ST6/04059, and by the grant from the Swiss National Supercomputing Centre (CSCS) under project ID d25. The authors are grateful to the Czestochowa University of Technology for granting access to NVIDIA Tesla K80 GPU providing by the MICLAB project No. POIG.02.03.00.24-093/13.

The energy efficiency of the proposed method is examined using two GPU-based clusters. The first of them is the Piz Daint supercomputer, currently ranked 3rd at the TOP500 list (Nov. 2017). It is equipped with NVIDIA Tesla P100 GPU accelerators based on the Pascal architecture. The second is the MICLAB cluster containing NVIDIA Tesla K80, based on the Kepler architecture. By comparing the mixed and double precision versions of MPDATA, we conclude that our method allows us to reduce the energy consumed by the application up to 30% on Piz Daint and up to 36% on MICLAB, considering the energy consumed by the graphic cards.

Changing the precision of arithmetic affects a few factors, including the computation speed, data size, and scalability of the application. It has improved the performance by 33% for Piz Daint and 36% for MICLAB, in relation to the double precision arithmetic. It also allows us to reduce the number of GPU resources to achieve the same performance as using the double precision arithmetic.

Furthermore, the proposed method allows users to decrease the energy consumption required by their simulations without any special access to machines such as the superuser account. Using the mixed precision arithmetic, our approach permits us to achieve the energy reduction at the similar level as the DVFS technique investigated in paper [6]. What is also important, unlike the DVFS technique, the proposed method reduces the execution time of MPDATA.

References

1. P. Igounet, E. Dufrechou, M. Pedemonte, P. Ezzatti, A Study on Mixed Precision Techniques for a GPU-based SIP Solver, 2012 Third Workshop on Applications for Multi-Core Architectures (WAMCA), 2012, pp. 7-12.

2. E. Meneses, O. Sarood, L.V. Kale, Energy profile of rollback-recovery strategies in high performance computing, Parallel Computing 40, 2014, pp. 536-547

3. R. Nathan, H. Naeimi, D.J. Sorin, X. Sun, Profile-Driven Automated Mixed Precision, https://arxiv.org/pdf/1606.00251

4. J.M. Prusa, P.K. Smolarkiewicz, A.A. Wyszogrodzki, EULAG, a computational model for multiscale flows, Computers & Fluids 37(9), 2008, pp. 1193-1207

5. K. Rojek, R. Wyrzykowski, Parallelization of 3D MPDATA algorithm using many graphics processors, Lect. Notes Comput. Sci. 9251, 2015, pp. 445-457

6. K. Rojek, A. Illic, R. Wyrzykowski, L. Sousa, Energy-aware Mechanism for Stencil-Based MPDATA Algorithm with Constraints, Concurrency and Computation: Practice and Experience, 29(8), 2017

7. B. Rosa, et al., Adaptation of multidimensional positive definite advection transport algorithm to modern high-performance computing platforms, Int. Journal of Modeling and Optimization 5(3), 2015, pp. 171-176