Efficiency and Effectiveness Analysis of a Scratchpad Memory on FPGA and GPU for Diffuse Radiation Transfer Simulation

FURUKAWA, Kazuki YAMAGUCHI, Yoshiki YOSHIKAWA, Kohji KOBAYASHI, Ryohei
FUJITA, Norihisa BOKU, Taisuke UMEMURA, Masayuki
University of Tsukuba
1-1-1 Tennôdai, Tsukuba, Ibaraki, 305-8573, Japan
furukawa.kazuki.gr@lila.cs.tsukuba.ac.jp

1 INTRODUCTION

Radiation hydrodynamics is a fundamental scientific concept to unveil the cosmic physics process in astrophysics. The radiative transfer (RT) simulation, ARGOT (Accelerated Radiative transfer on Grids using Oct-Tree) [2], is well-optimized for the RT simulation. However, it still requires an accelerator to simulate within a reasonable time. Thus, this project focuses on the acceleration of the ART (Authentic Radiation Transfer) [3] scheme, one of the composition of the ARGOT, by data reuse optimization.

In the ART scheme, each ray goes straight into the simulation space, and the rays progress in parallel. It means this is not a matrix calculation but ray tracing. The available bandwidth of off-chip memory, including HBM, has been the most critical factor in the ART acceleration because of complicated and enormous memory access. Thus, PRISM (PRefetchable and Instantly accessible Scratch-pad Memory) are proposed for the ART acceleration available in FPGA and GPU to achieve sufficient acceleration. The efficiency and effectiveness are also discussed.

2 PRISM MECHANISM

As an ART-specific RTL buffering scheme, we have proposed the PRefetchable and Instantly accessible Scratch-pad Memory (PRISM) [1] for FPGAs. In this study, we extended it to GPU implementation.

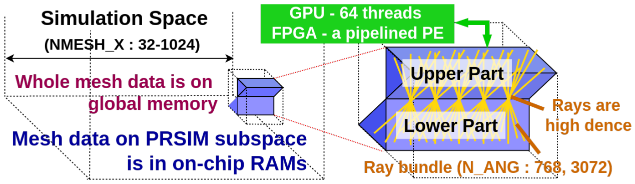

Figure 1: Memory access locality of the PRISM mechanism

The form of the PRISM subspace looks like the two long triangular prisms like Figure 1, which is cut out from the simulation space. The key concept is data reuse by local storage with high throughput to reduce the memory accesses. Its data-reuse efficiency is derived from the nature that the Rays diffuse in all directions. Namely, it is profitable since multiple Rays pass through one mesh, especially many at the centre, where Rays are highly dense.

The long prism form also enhances the memory access locality and enables the overlapping feature of arithmetic operations and memory accesses. Note that the PRISM usages UltraRAM in Xilinx FPGA and Shared memory with NVIDIA GPUs as local storage.

3 EXPERIMENTAL RESULT

Figure 2 shows the implementation result, whose vertical axis presents the number of RT equations that can be calculated in one second, i.e. computing performance. We can see that the ART with PRISM on FPGA is better when the simulation space is small, whereas one on GPU is better when large. One of the reasons can be the overhead such as GPU's device initialisation. Comparing the result of the original ART with the others, the PRISM is more effective in large scale simulations as the number of global accesses is significantly reduced on every implementation.

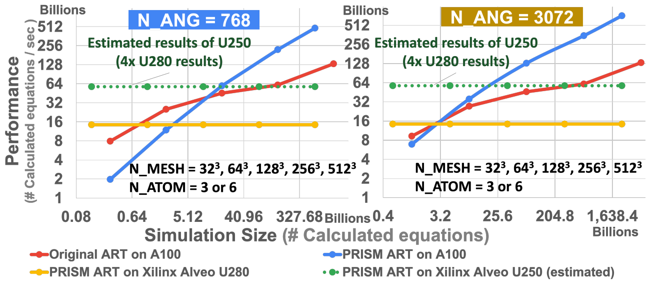

Figure 2: Performance comparisons of each simulation size

4 CONCLUSION

Using our proposed method, we conclude that the original ART can be accelerated by both FPGAs and GPUs. An FPGA yields better when the simulation space is small, while a GPU is better when large because of GPU's high parallelism. We also prove that the PRISM reduces the memory access bottleneck and contributes to a significant increase in the utilisation of the arithmetic circuits.

ACKNOWLEDGMENTS

This work was supported in part by the MEXT Next Generation High-Performance Computing Infrastructures and Applications Research and Development Program, entitled "Development of Computing-Communication Unified Supercomputer in Next Generation", and in part by the JSPS KAKENHI #21H04869. We would also acknowledge the support of Xilinx Inc. through the Xilinx University Program.

REFERENCES

[1] K. Furukawa et al. 2021. An efficient RTL buffering scheme for an FPGA-accelerated simulation of diffuse radiative transfer. In *2021 International Conference on Field-Programmable Technology (ICFPT)*. IEEE, Auckland, New Zealand, 1–9.
[2] T. Okamoto et al. 2012. ARGOT: accelerated radiative transfer on grids using oct-tree. *Monthly Notices of the Royal Astronomical Society* 419, 4 (Feb. 2012), 2855–2866.
[3] T. Satoshi et al. 2015. A new ray-tracing scheme for 3D diffuse radiation transfer on highly parallel architectures. *Publications of the Astronomical Society of Japan* 67, 4 (May 2015). 62.