Efficient Utilization of GPU Cluster Resource for Stencil Computation

Guanghao Jin
Tokyo Institute of Technology
JST-CREST
Tokyo, Japan
jin.g.ab@m.titech.ac.jp

Toshio Endo
Tokyo Institute of Technology
JST-CREST
Tokyo, Japan
endo@is.titech.ac.jp

***Abstract*—In common way case, the domain size of the stencil computation is limited by the memory capacity GPUs in GPU cluster. To efficiently use the resource of GPU cluster, this paper proposes and evaluates parallel optimization method for stencil computation to utilize GPU memory, CPU memory and SSD of the multiple nodes while maintaining high performance. Furthermore, our new method utilizes multiple GPUs in each node to achieve higher performance. Also, it uses the CPU memory and SSD to enable bigger domain computation in each node. Then, it proposes new decomposition method among the nodes to achieve scalability. Evaluation of stencil simulation on 3D domains show that our new method for 7-point achieves good scalability while achieving 2.14 times higher performance than other methods on average.**

Keywords—stencil computation; GPU cluster; memory capacity; parallel optimization method, GPU memory, CPU memory, SSD

I. INTRODUCTION

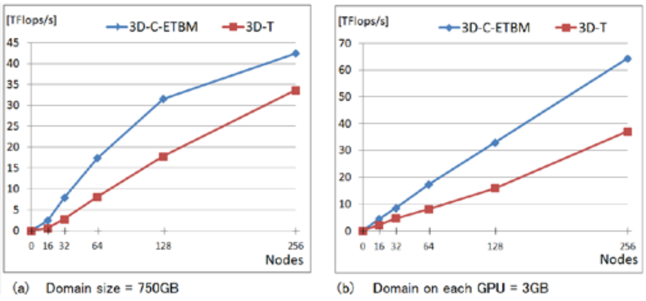
Recently, Graphics Processing Units for general-purpose computation (GPGPU) is proved to be a high-performance computing device which has been remarkably successful in *stencil computation*. In many GPU based supercomputer systems, the peak performance of the GPU is faster than the CPU while the available memory is smaller than that of CPU. For example, the device memory is 6GB and the host memory is 54GB on some nodes of TSUBAME2.5 supercomputer at Tokyo institute of technology. This property limits the domain sizes in stencil computation by the typical stencil computation that is executed on GPUs. By using multiple GPUs with proper domain decomposition, the limitation is mitigated; however, the total domain size is still limited by the aggregated memory capacity of used GPUs. We would be able to compute even larger domains if we used the larger capacity of host memory and local storage (like SSD). However, this is a great challenge to support larger domains while keeping high performance.

II. RELATED WORK

In order to enable the bigger domain computation in stencil case, *temporal blocking method* [1] has been proposed. In temporal blocking method, it focuses on a smaller part of the domain, and processes its computation for several time steps at once. On the other hand, it is known that this approach causes redundancy problem. Our previous paper proposed optimization techniques to solve redundancy problem while enabling bigger domain computation by the utilization of CPU memory, GPU memory and SSD in single node case [2].

III. PROPOSED METHOD

Firstly, our optimization method utilizes multiple GPUs in each node to achieve higher performance. Then, it utilizes CPU memory and SSD to enable bigger domain computation while solving redundancy problem of temporal blocking method. Among the nodes, it proposes new decomposition method to achieve scalability while reducing communication cost between nodes. We evaluate our method on TSUBAME2.5 peta-scale CPU-GPU hybrid supercomputer. We use 3 GPUs (K20X, 6GB memory) in each node. The results show that the performance of our method for 7-point achieves good scalability while achieving high performance, which is 2.3 times higher in strong scaling case and 1.97 times higher in weak scaling case than temporal blocking method.



(a) Domain size = 750GB

(b) Domain on each GPU = 3GB

Fig.1. (a) Strong scaling (b) Weak scaling

[1] Leonardo Mattes and Sergio Kofuji, “Overcoming the GPU memory limitation on FDTD through the use of overlapping subgrids,” ICMMT, pp.1536 – 1539, 2010.

[2] Tianqi Xu, Guanghao Jin, Toshio Endo and Satoshi Matsuoka. Efficient Utilization of Multi-level Memory System for Stencil Computation. IPSJ SIG Technical Report, 2014-HPC-147, 8pages, Japan, March 2014.