学際大規模情報基盤共同利用・共同研究拠点 公募型共同研究 2025年度採択課題
jh250079
17th Symposium
Joint Usage / Research Center for Interdisciplinary Large-scale Information Infrastructures

Toshihiro Hanawa (The University of Tokyo)

Study on Efficient FileIO for GPU

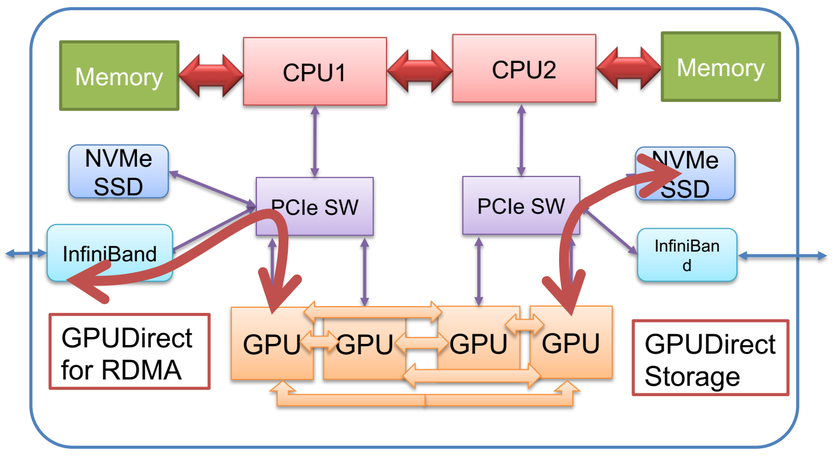
Introduction

- Using GPUs in practical applications often faces file IO bottlenecks due to the need to handle data in GPU memory separately from the host.
- While file IO from GPUs has conventionally been performed via the host CPU, NVIDIA offers GPUDirect Storage (GDS), an extension of GPUDirect for RDMA for file IO.
- This study aims to optimize the handling of direct file IO on GPUs and enhance processing efficiency in various applications, based on the results of **jh240081**, etc.

Purpose

- We analyze different architectures to optimize file IO for GPUs and will provide libraries and functions for use **across all JHPCN GPU systems**.
  ➔ Commonly, NVIDIA H100 is used, but different CPUs are used !!

University	System	CPU	GPU
Hokkaido U.	Grand Chariot 2 (GPU Nodes)	Intel Emerald Rapids	NVIDIA H100
U. Tokyo	Miyabi-G (JCAHPC)	NVIDIA Grace (GH200)	NVIDIA H100 (GH200)
Science Tokyo	TSUBAME 4.0	AMD Genoa	NVIDIA H100
Kyushu U.	Genkai (Node Group B)	Intel Sapphire Rapids	NVIDIA H100

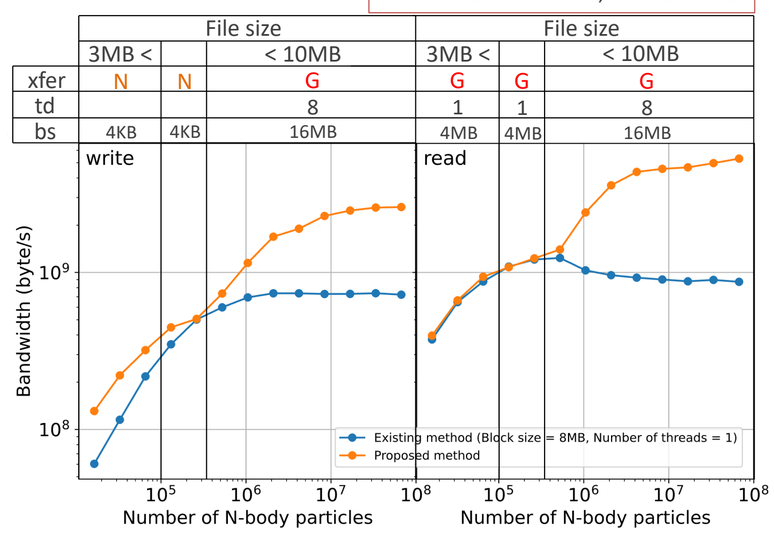


Overview of GPU node

Theme

Applications (Astrophysics:GOTHIC, Climate/Weather: Scale-RM)

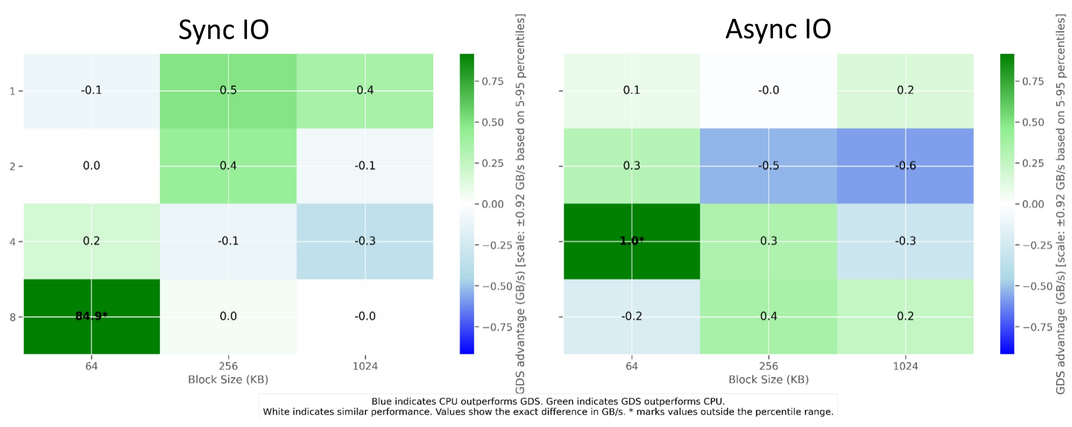
- GOTHIC: a gravitational octree code optimized for GPU [2]
  - Most routines, including tree construction and traversal, are performed on the GPU, while file I/O for reading initial conditions and writing snapshots or restart files relies on HDF5 functions via the host CPU.
  - Evaluate using the GH200 system ➔ Miyabi-G system
- Scale-RM: a climate/weather simulation code
  - Such as simulations of tropical cyclones, in order to observe the detailed time evolution of high-precision data on large-scale phenomena, it is necessary to devise ways to reduce IO overhead.
    - In previous project, we investigated to City-LES under similar motivation.



FileIO performance improvement on HDF5 [3,4]

System Software for GPU FileIO (Led by BITS Pilani, India)

- To optimize performance when choosing between CPU memory copy and GPU Direct storage, consider factors such as I/O access patterns, parallel file system efficiency, and data transfer rates
- We will develop a generic framework to transparently assist GPU-based applications in selecting the optimal transfer method and parameters for various file formats (NetCDF, VisIt, h5py, ParaView, and Binary) by leveraging advanced data-driven models, adaptive algorithms, and efficient I/O access patterns.



GDS vs CPU performance comparison with 10 GB dataset by HDF5
(RTX A2000 12 GB, IO threads: 1,2,4,8, Block size: 64KB, 256KB, 1MB)

Advanced GPU Direct IO: assist GPU memory by NVMe SSD with High IOPS

- GDS (GPUDirect Storage): enable direct FileIO between GPU and NVMe SSD / cluster file system, such as Lustre FS. However, in the case of small IO transaction, large overhead is observed.
- BaM (Big accelerator Memory): enable access to the storage based on NVMe protocol directly from GPU. However, it cannot treat file structure [5].

- SCADA: new programming model for high-throughput, fine-grained, GPU-initiated access, mainly based on object storage [6]
  - Apply novel NVMe SSD technology and framework to not only ML but also HPC applications with extremely large-scale data

Reference

[1] Y. Miki, M. Mori, T. Kawaguchi, and Y. Saito: ``Hunting a Wandering Supermassive Black Hole in the M31 Halo Hermitage'', The Astrophysical Journal, 783, 87 (2014)
[2] Y. Miki, M. Umemura, “GOTHIC: Gravitational oct-tree code accelerated by hierarchical time step controlling”, New Astronomy, vol. 52, pp. 65-81 (2017)
[3] M. Tominaga, T. Hanawa, Y. Miki, “Toward Optimizing File IO on GPU Clusters”, GTC 2024 poster
[4] 富永瑞己, 塙敏博, 三木洋平,「GPU 直接IO を用いたファイルIO の高速化」、xSIG 2024, 2024 年8 月(Best Master’s Student Award)
[5] Z. Qureshi et al., “GPU-Initiated On-Demand High-Throughput Storage Access in the BaM System Architecture,” ASPLOS 2023
[6] C.J. Newburn et al., “Speed-of-Light Data Movement Between Storage and the GPU”, GTC 2025