GPU のための効率的な IO に関する研究

Study on Efficient IO for GPU

塙　敏博
東京大学

１．研究目的

HPC システムにおいて GPU がアクセラレータとして広く利用されているが，実アプリケーションの GPU 移植では，計算が高速化されてもファイル IO がボトルネックとなって全体の効果が得られないことが多い．本研究は，NVIDIA が提案する次世代 GPU 直接 IO 設計思想 Scaled Accelerated Data Access（SCADA）の HPC 向け応用に向けたベンチマーク設計を目的とし，NVIDIA およびキオクシア株式会社との共同研究として進めている．SCADA は現在構想・開発段階のため，現時点ではプロトタイプである Big accelerator Memory（BaM）を用いて予備実験を行っている．BaM は GPU スレッドが直接 NVMe ストレージにアクセスし，CPU 主導の IO 処理によるオーバーヘッドを大きく低減する技術である．本研究は東京大学を中核拠点とする JHPCN 課題“Study on Efficient File IO for GPU”と連携して実施しており，本学際共同利用プログラムで割り当てられた Pegasus の NVIDIA H100 環境は，後述する階層型行列 H-matrix を用いた実アプリケーション応用ベンチマークでの活用を計画している．

２．研究成果の内容

(1) BaM の基本性能評価：手元にあるテスト環境（milan1: AMD EPYC 7713, NVIDIA A100 80GB，KIOXIA CM6-V NVMe SSD）において BaM の基本性能評価を実施した．バンド幅比較では IO 粒度を 4KB～256KB に変化させ，4KB の細粒度では BaM が GDS 比約 10.7 倍のバンド幅を示す一方，128KB 以上の大粒度では IO リクエスト発行回数の減少と CPU のクロック優位により GDS が優位となることを確認した．また，単精度浮動小数点行列積 SGEMM では，行列をタイルに分割し SSD-GPU 間の IO を伴いながら計算する実装により，N=131,072、T=2,048 において BaM は CPU 経由比で約 4.34 倍高速であり，特に小タイル（ランダム性の高いアクセス）で優位となることを確認した．本成果は情報処理学会 HPC 研究会（SWoPP 2025）にて発表した．

(2) H-Matrix アプリケーションへの適用：境界要素法等で扱う密行列を階層的にクラスタリングし低ランク近似で圧縮する $\mathcal{H}$-Matrix（積分方程式法シミュレーション向け高速計算ライブラリ $\mathcal{H}$ACApK を Fortran から C++に移植したものを使用）に対して，BaM による GPU 直接 IO を組み込んだアプリケーションを設計した．GPU 上

で構築した $\mathcal{H}$-Matrix を密葉/低ランク葉のバッチに分割して SSD へ書き戻し，BiCGSTAB 法ソルバの反復毎にバッチを読み出して計算する実装である．Stanford Dragon を入力とした静電ポテンシャル問題（$\mathcal{H}$-Matrix 総容量 42.16GB）の評価において，読み出したバッチを GPU メモリ上にピン留めする工夫を加えることで，BaM は CPU 経由（O_DIRECT）と比較して全実行時間で最大約 1.31 倍，読み出し時間で最大約 1.71 倍高速となることを確認した．本成果は情報処理学会 HPC 研究会（2026 年 3 月）にて発表した．

３．学際共同利用プログラムが果たした役割と意義

本研究の対象である GPU 直接 IO 技術は，演算加速器を備える次世代 HPC システムでの大規模データ処理における必須の要素技術である．本年度は SCADA の予備実験段階としてテスト環境（A100 + 単一 NVMe SSD）での BaM の基本性能評価および $\mathcal{H}$-Matrix アプリケーションへの組込みを中心に実施したが，本学際共同利用プログラムで割り当てられた Pegasus の NVIDIA H100 環境は，今回得られた知見を実機上でさらに大規模データ・複数 SSD 構成へ拡張・検証するための重要な計算資源であり，後述する次年度以降の評価において活用する予定である．

４．今後の展望

今後は，より大規模なデータセットおよび複数台の NVMe SSD を用いた構成への拡張により，PCIe ピーク帯域までのスケーリング評価と BaM・SCADA の性能比較を進める．$\mathcal{H}$-Matrix アプリケーションについては，BiCGSTAB 法に代わる，1 バッチに対し複数反復をまとめて計算可能なソルバや複数ベクトル同時の Matvec 実装により GPU キャッシュ活用を高めるとともに，バッチを KB 単位以下に細分化して IO と計算をオーバラップさせる手法に取り組む．加えて，2026 年サンプル出荷予定のキオクシア Super High IOPS SSD（1000 万 IOPS 以上）を用いた細粒度ランダムアクセス性能評価，および SCADA が利用可能になった際のオブジェクトストレージベース環境での評価へ拡張する．本学際共同利用プログラムで割り当てられた Pegasus の H100 環境は，これら拡張実験を実施する重要な計算資源として活用し，演算加速器を搭載するポスト富岳をはじめとする次世代システムへ知見を還元することを目指す．

５．成果発表

(1) 学術論文

(2) 学会発表

・竹島 颯, 塙 敏博, 三木 洋平,「超大規模データを指向した GPU 直接 IO の性能評価」，情報処理学会研究報告 ハイパフォーマンスコンピューティング, Vol.2025-HPC-200, No.35, SWoPP 2025, 2025 年 8 月，高松市.

・竹島 颯, 塙 敏博, 三木 洋平, 伊田 明弘,「大規模データを指向した GPU 直接 IO に基づく $\mathcal{H}$-Matrix アプリケーションの設計と性能評価」，情報処理学会研究報告 ハイパフォーマンスコンピューティング, Vol.2026-HPC-203, No.64, 2026 年 3 月，札幌市.

(3) その他

使用計算機	使用計算機に○	配分リソース※		
		当初配分	移行*	一般利用による追加
Pegasus	○	3,080		
Miyabi-G				
Miyabi-C				