NEC Scalable Technology File System for AI (ScaTeFS for AI)
GDS(GPU Direct Storage) 動作検証レポート

目次

GDS 動作検証について

1　ご利用にあたっての注意事項

本レポートは動作検証レポートであり、弊社が動作保証するものではありません。
動作確認情報は、各ページに掲載されている評価環境での検証結果に基づいたものです。

2　GDS の概要

GDS は、NVIDIA AI Enterprise における GPU ダイレクトテクノロジーの一つであり、GPU とストレージ間のデータ転送の高速化を実現する機能です。CPU 側のメモリへデータをコピーすることなく GPU から直接データ転送を行うことができます。ローカル及びリモートのストレージで動作が可能です。今回、弊社の分散・並列ファイルシステム製品である NEC Scalable Technology File System for AI(ScaTeFS for AI) を用いて、GPU と ScaTeFS for AI のファイルサーバ間とのデータ転送を GDS のアクセスとする検証を実施しました。

GDS の利用方法や詳細につきましては以下を参照ください。
https://docs.nvidia.com/gpudirect-storage/index.html

ScaTeFS for AI につきましては以下を参照ください。
https://jpn.nec.com/gpu/scatefs_ai/index.html

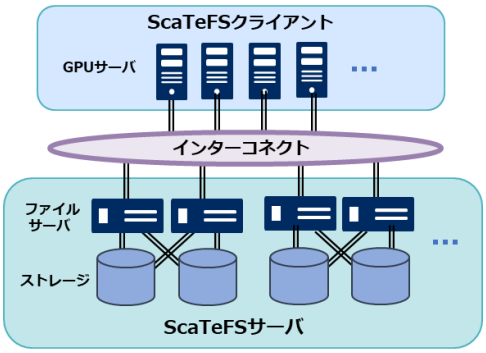
3　検証目的

映像・画像・音声・テキスト等の大規模データでの学習においては、GPU サーバからデータ取得（read 処理）を中心とした大量のアクセスが行われます。
今回の検証では、Express5800/R120j-2M を ScaTeFS for AI のファイルサーバとし、後述する環境下にて GDS が問題なく動作すること、およびデータ取得(read 処理)での GDS の効果を検証し、その結果を記載しています。

4　動作検証

4.1 動作検証システム構成

ScaTeFS for AI は、システムの大規模化、データの大容量化に対応できる分散・並列ファイルシステムです。以下は ScaTeFS for AI のシステム構成例です。ScaTeFS サーバは 2 台のファイルサーバと配下のストレージをセットとして、お客様の環境や要件などにあわせて構成することができます。



今回の検証で使用した ScaTeFS for AI は、ファイルサーバ 2 台(Express5800/R120j-2M)およびストレージ 2 台(iStorage V100)で構成し、ScaTeFS for AI のクライアント 1 台(GPU サーバ)からアクセスしています。ScaTeFS for AI のクライアントと ScaTeFS のファイルサーバ間は 100GbE のネットワークで接続しています。

4.2 ScaTeFS for AI のサーバ構成 (Express5800/R120j-2M)

本章では、ScaTeFS for AI のファイルサーバとして動作検証を実施した R120j-2M についての構成（1 台あたり）について説明します。

4.2.1　Express5800/R120j-2M サーバ手配構成

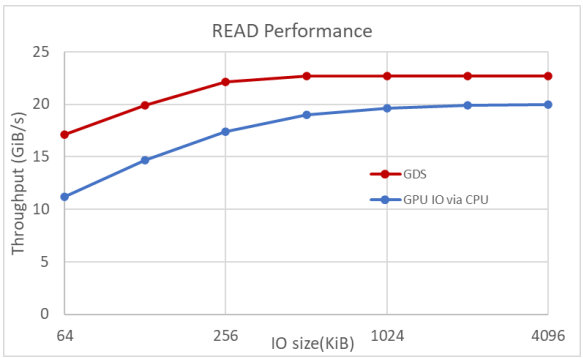
製品名	対象型名	数量	補足事項
Express5800/R120j-2M	N8100-2989Y	1	8x 2.5 型ドライブモデル
増設バッテリ用ケーブル	K410-513(00)	1	フラッシュバックアップユニット用ケーブル
OCP カード接続ケーブル	K410-525(00)	1	1st CPU 側
CPU ボード	N8101-1840	2	12C/2GHz/silver4410Y
2U 標準ヒートシンク	N8101-1856	2	
メモリダミーキット	N8102-746	1	冷却性能改善のために必要なメモリブランク
16GB 増設メモリボード	N8102-759	16	異なる型番のメモリとの混在搭載不可

フラッシュバックアップユニット	N8103-218	1	RAID コントローラ用
RAID コントローラ	N8103-243	1	
1000BASE-T 接続 LOM カード (4ch)	N8104-206	1	
10GBASE-T 接続ボード(2ch)	N8104-219	1	ファイルサーバ間の Interconnect 用
リモートマネジメント拡張ライセンス (Advanced)	N8115-33	1	リモートコンソールでインストール、保守行う場合に選択
トップカバーオープン検知キット	N8115-44	1	
2nd ライザカード(3xPCI + 1xGPU 搭載キット)	N8116-113	1	
増設用 2.5 型 480GB SATA RI SSD	N8150-1826	3	OS, ScaTeFS for AI データ退避用
電源ユニット(1600W)	N8181-162A	2	
2U 高性能ファン	N8181-209	1	
Fibre Channel コントローラ(2ch)	N8190-176	2	ストレージ接続用

上記構成に GDS アクセス用の NIC として NVIDIA Mellanox ConnectX-6(MT28908)x2 を搭載しています。また、ScaTeFS for AI のクライアント側では GPU として NVIDIA A100 40GB を使用しています。

5 検証結果

ScaTeFS for AI のファイルサーバとして Express5800/R120j-2M を用いた環境において、GDS の動作検証を行った結果、問題が発生しないことを確認しました。またデータ取得(read 処理)での GDS の効果として以下の通り性能向上が確認できました。



上記は、1 つの GPU から 32 並列でファイルデータを読み取るときのスループットを示しています。

全体的に通常の CPU 経由の読み取りによるスループットより GDS のスループットの方が高く、GPU から大量のデータを読み取るケースでの性能向上の効果が見て取れます。

動作検証時の主要ソフトウェアのバージョン・内容は下記になります。

- Linux : Red Hat Enterprise Linux 8.6 ,
  kernel 4.18.0-372.32.1.el8_6_x86_64
- ScaTeFS for AI Client : 1.0.0.0
- ScaTeFS for AI server : 1.0
- NVIDIA MOFED : 5.8-1.1.2.1

6 関連リンク

GPU ソリューション: サーバ・ストレージ・エッジコンピューティング (NEC)
PC サーバ Express5800 シリーズ (NEC)
NEC ストレージ「iStorage」(NEC)
分散・並列ファイルシステム ScaTeFS for AI（NEC）
NVIDIA GPUDirect Storage （NVIDIA 社）

7 お問い合わせ先

NEC インフラ・テクノロジーサービス事業部門 コンピュート統括部
ai-scatefs@support.jp.nec.com

8 改版履歴

版数	公開日時	変更内容
第 1 版	2024 年 8 月	第 1 版リリース