ヘテロジニアスな In-database データ解析・機械学習基盤
—GPU を活用した高速 PostgreSQL の実現と事業化—

1. 背景

近年、半導体の微細化技術の進展が鈍化しつつあり、これまでのように高度化するソフトウェアがハードウェア性能の進歩に“ただ乗り”できる状況は終わりを迎えつつあり、その対策として、計算機をスケールさせる並列分散処理や、GPU/FPGA のように特性の異なるハードウェアを組み合わせる“ヘテロジニアス・コンピューティング”が用いられるようになっている。

特にソフトウェアの高度化が著しく、計算能力に対する需要が旺盛であるのは、大量データの収集・蓄積と、分析・モデル化・予測といった一連のビッグデータ処理、統計解析処理や機械学習といった分野である。

本 PJ の代表である海外は、2012 年より GPU をデータベースのアクセラレータとして使用する技術(PG-Strom)を開発し、使いやすく、かつ安価で高速な大量データ処理基盤の研究開発に取り組んでいた。

2. 目的

本未踏 AD プロジェクトの目標は以下の通りである。

技術面:

- ✓ 中核技術の PG-Strom をエンタープライズ利用に耐えうる完成度へと引き上げ、競合製品の代替となりえる処理速度を達成する。
- ✓ 機械学習ライブラリや周辺ソリューション等、周辺ソフトウェアの開発
- ✓ 更なる性能向上・機能強化に向けた基本技術開発と検証

事業面:

- ✓ 事業体の立ち上げと事業体制の構築
- ✓ アーリーアダプタとなるエンドユーザを開拓。PoC の実施。
- ✓ 顧客窓口となる SIer/販売代理店の開拓

3. 製品・サービスの内容

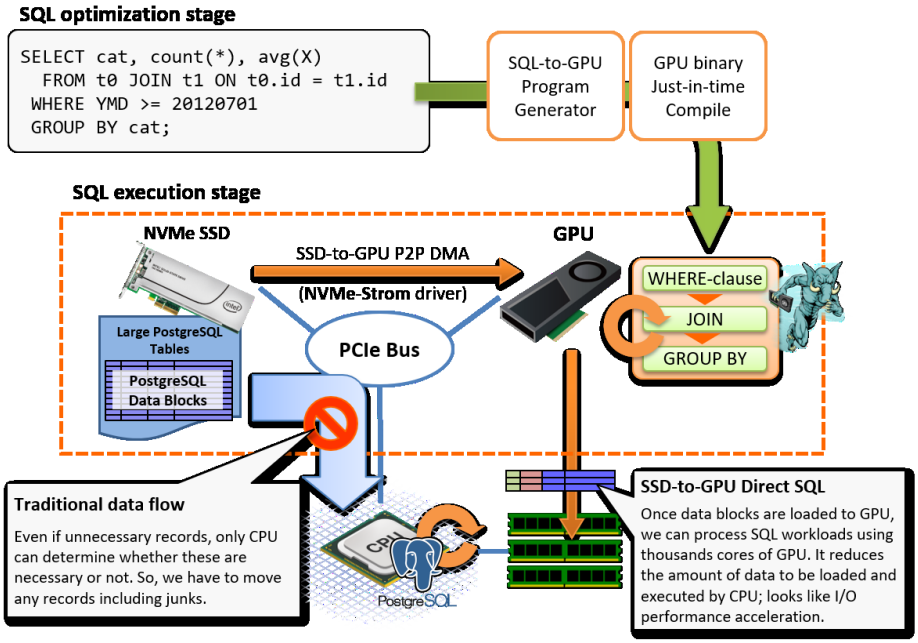
PG-Strom: SSD-to-GPU ダイレクト SQL 実行機能

通常、ストレージ上に格納されたＤＢデータブロックはメインメモリへと読み出され、CPU や（場合によっては）GPU によって各レコードの要不要が判断され、次いで JOIN/GROUP BY といった集計系の演算を行う事となる。

一方で、SQL ワークロードの特性上、これら読み出したデータを全て使うわけではなく、そのかなりの部分がまず条件句によってフィルタリングされ、また GROUP BY 句や集約演算によって集計された後は、オリジナルデータはそもそも必要ではない。

つまり言い換えれば、データベース処理において少なくない割合を占める I/O ワークロードは、その多くをジャンクを転送するために費やしている。

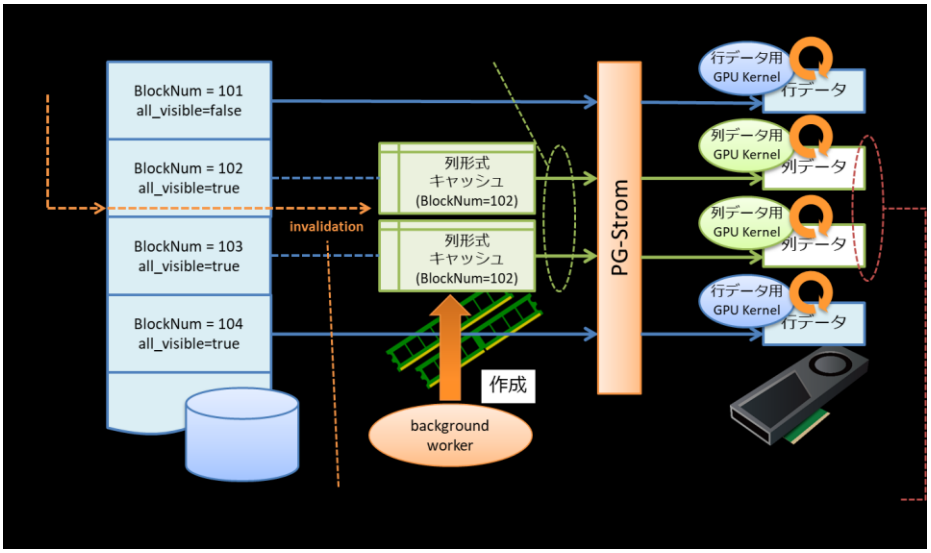
SSD-to-GPU ダイレクト SQL 実行機構は、SSD 上のデータブロックをまず GPU へと転送する。これは P2P DMA を用いており、CPU/RAM を経由しない。



データが GPU に転送されさえすれば、GPU での SQL ワークロード並列実行は元々保有していた機能である。SCAN/JOIN/GROUP BY をデータ読出しの途中で処理する事でデータ量を削減し、CPU が処理すべきレコード数を削減。結果として I/O を高速化する。

PG-Strom: In-memory 列指向ストア

SSD-to-GPU ダイレクト機能は非常に強力なストレージ機能強化であるが、H/W 構成を選んでしまう。そのため、現時点では NVMe-SSD と GPU を共存するタイプのインスタンスが提供されていないクラウド環境では適用が難しいという課題がある。

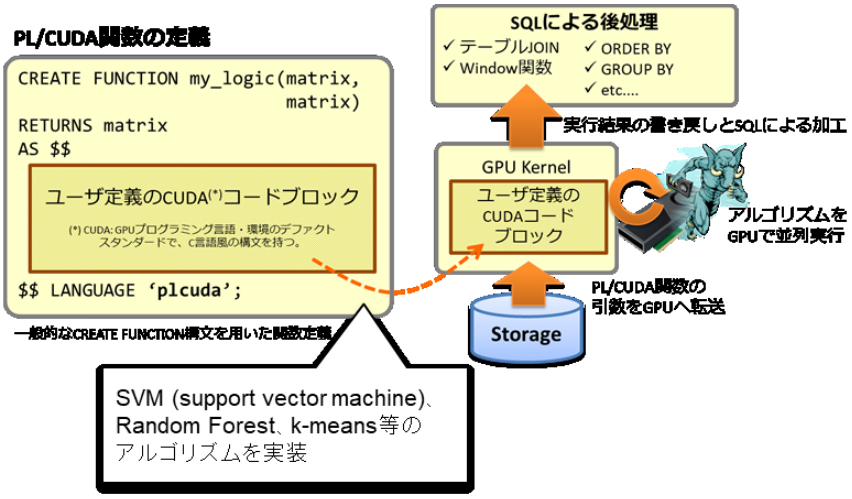


そのため、容量では NVMe-SSD を下回るものの、RAM に格納可能なサイズのデータを対象に、I/O 効率・GPU 処理効率を最適化するようデータ形式を列形式に変換して保持するキャッシュ機構を実装した。

これにより、100GB 未満の比較的ローエンド層に対してはクラウドサービス上での展開も可能となった。今後、AWS、Microsoft、Google 等主要ベンダー環境への展開や、契約面も含めたサービス提供準備の作業が残っている。

PG-Strom: 機械学習ライブラリ

PG-Strom の PL/CUDA 機能は複雑な統計解析アルゴリズムを SQL 関数として記述し、GPU ネイティブの実行速度で In-database 実行を可能にする機能である。上手く問題に適合する場合は、CPU 実装の 100 倍以上の高速化が可能となる。

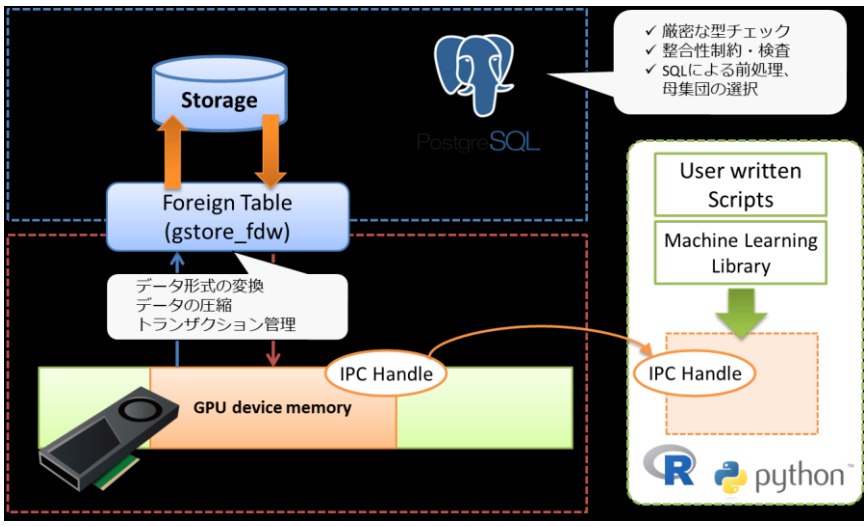


一方で、PL/CUDA 関数でアルゴリズムを記述する場合、高度な数学・統計・GPU プログラミングの技術が必要となるため、必ずしも使いよいものとは言えなかった。

本 PJ では、機械学習アルゴリズムのうち、ユーザヒアリングを踏まえ利用頻度の高い SVM(Support Vector Machine)、Random Forest、k-means clustering をライブラリとして実装し、製薬企業の研究部門の協力を得て Random Forest アルゴリズムの実証実験を行った。

PG-Strom: GPU メモリストア(gstore_fdw)

当初計画では PostgreSQL の Python スクリプト連携機能を通して深層学習フレームワークとの連携を図る予定であったが、ユーザヒアリングを通して、SQL を経由する事でユーザ獲得の敷居が上がる事と、PostgreSQL の可変長データ形式に起因する制限を回避できない事から、別アプローチを採用した。



GPU メモリストア機能は、GPU 上に一時的でないメモリ領域を確保し、PostgreSQL の FDW インターフェースを通して当該領域へ SQL を用いてデータの読み書きを行うための機能である。

GPU 上のメモリ領域は IPC Handle を通して他のプロセスと共有する事ができるため、いわば共有メモリとして扱う事ができ、外部の Python や R スクリプトから参照する事ができる。

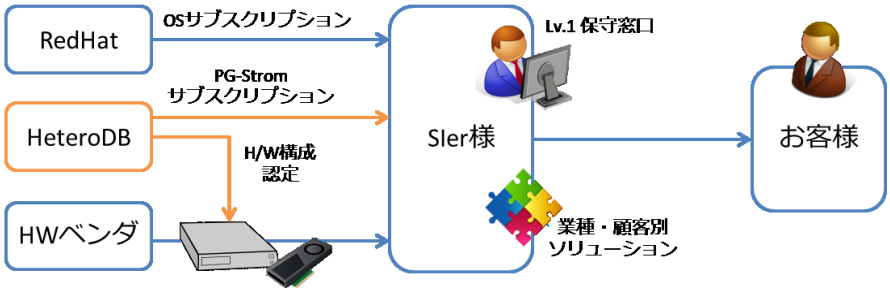
GPU 上の内部データ形式は任意に設計する事ができるので、外部のライブラリで利用されているデータ形式に適合させる事で、データ管理は RDBMS で、機械学習は Python でという得意分野に基づいた住み分けに繋がる可能性がある。
現時点では、PG-Strom 固有のデータ形式であるが、機械学習フレームワークである Chainer との連携を目指し、cuPy のデータ形式に対応させるべく作業を進めている。

4. 新規性・優位性

本 PJ で開発した PG-Strom を搭載する DB アプライアンスサーバは、7.5GB/s という DWH 専用機に匹敵する処理速度はもとより、①PostgreSQL という最も広範に利用されている RDBMS と全く同様に利用する事ができ、②行データ形式を採用しているためトランザクション系処理とも容易に統合可能、それに伴い③システム全体をシンプル化する事で DB 管理者の負荷を軽減する事が可能となる。
一方で、GPU や NVMe-SSD といったコモディティ部品とオープンソースソフトウェアを利用している事から、競合製品よりも概ね1桁小さな TCO を実現している。
また、PL/CUDA や機械学習ライブラリは必ずしも透過的な高速化を実現する機能ではないが、GPU の得意分野である機械学習を In-database 実行し、“データのある場所で計算”を可能にするのは他に例を見ないユニークな機能である。

5. 事業普及（または活用）の見通し

事業体である HeteroDB 社は小規模なスタートアップであり、中核技術以外に強みを持たない。基本的な戦略は、SIer や販売代理店と提携する事で、営業網、保守体制、顧客・業種別ソリューションなどで弱点を補完する事である。



未踏 AD プロジェクト期間を通して、累計 xx 社の SIer/販売代理店を訪問。うち、大手 2 社との提携に向け、ソリューションの検討や商流・保守体制などの協議に入っている。また、ベンチャー2 社とも共同でソリューション開発に合意しており、3 月以降の実証実験で予定性能を発揮できれば、これも間接販売の窓口とできる可能性がある。
PG-Strom を搭載した DB アプライアンス、および PG-Strom サブスクリプション製品は 2018 年 4 月からの販売開始を目指しており、販売開始から初年度で 10 ユーザの獲得を目標にしている。

6. 期待される波及効果

PG-Strom の目指す市場は、既存製品のある DWH/Database のマーケットであり、その中でも情報系システムの代替を目指すものである。
我々のソフトウェアは必ずしもピーク性能を実現するための設計とはなっていないが、その代わり、少なくない数のエンジニアが使い慣れた PostgreSQL という基盤との差異をほとんど意識せず利用する事ができる構成となっているほか、データを移動することなくその場で計算するというコンセプトを頑なに守っている。

これは、本質的に希少リソースであるデータサイエンティストの時間を、雑用と言ってよい数々のデータ管理・システム管理から解き放つ事を目的としている。したがって、本製品による第一の波及効果として、データサイエンティストの生産性向上を挙げる事ができるであろう。

第二に、本領域は既存製品が数多く存在する領域であるが、市場シェアの多くを握る Oracle 社や Microsoft 社の製品をはじめ、その大半が海外製のソフトウェアである。海外製ソフトウェアへのライセンス費等の支払いは純然たる外貨の流出であり、乗数効果を加味すれば、国内企業から海外企業への決して低額ではないソフトウェアライセンス費の支払いは負の経済効果をもたらす。

性能面でこれら先行する競合製品に匹敵し、より安価な製品を提供する事で、国内IT投資が国内経済に再投資される事が期待される。

7. 未踏イノベータ名（所属）

海外 浩平（ヘテロDB株式会社 チーフアーキテクト 兼 代表取締役社長）
柏木 岳彦（ヘテロDB株式会社 チーフセールスエンジニア 兼 取締役副社長）
遠藤 克浩（慶應義塾大学 理工学部）

（参考）関連 URL

http://heterodb.com/ ヘテロDB株式会社
https://github.com/heterodb/pg-strom PG-Strom Project