研　究　報　告　書

「多様な構造型ストレージ技術を統合可能な再構成可能データベース技術」

研究タイプ：通常型
研究期間：　平成 25 年 10 月～平成 29 年 3 月
研 究 者：　松谷　宏紀

1．　研究のねらい

科学技術分野において日本の国際競争力を高めるには大量データ（ビッグデータ）処理技術と機械学習ベースの AI（Artificial Intelligence）技術の普及と浸透が不可欠である。これらを活かすには本来莫大な量の計算リソースが必要であり、現状では、巨大なデータセンタを管理、運用する限られた企業がそのような計算プラットフォームを握っている。このような状況において、大量データおよび AI 利活用のための敷居を下げ、これらの技術革新の恩恵を最大限享受できるようにするには、計算プラットフォームの低消費電力化および低コスト化が必須である。そこで、本研究では、大量データ処理や機械学習に関するオープンソースプロダクトのフレームワークを維持しつつ、FPGA（Field-Programmable Gate Array）や GPU（Graphics Processing Unit）のような汎用アクセラレータを利活用することに着目した。実際、FPGA や GPU の利活用によって得られた性能向上の分だけ計算リソースを削減できるため、これらのアクセラレーションは大量データ処理や機械学習処理の低消費電力化および低コスト化に大きく寄与する。

アクセラレーション対象の 1 つ目として、大量データの蓄積と検索を司るデータベース技術に着目する。具体的には、キーバリューストア（KVS）型、カラム指向型、ドキュメント指向型、グラフ型データベースなど多様なデータベースに対し、FPGA や GPU のようなアクセラレータをどのように使うべきかを探求する。アクセラレーション対象の 2 つ目はストリーム処理であり、無限に生成される時刻順データに対する計算処理に着目する。具体的には、オンラインの外れ値検出などストリームデータに対する各種機械学習アルゴリズムを対象に FPGA を用いたアクセラレーションを行う。3 つ目はバッチ処理である。10GbE（10Gbit Ethernet）経由で接続された GPU クラスタを用いて既存のバッチ処理フレームワークを高速化する。当初はアクセラレーション対象をデータベース処理に絞っていたが、最終的には既存のラムダアーキテクチャの要素技術を網羅的にカバーできるようアクセラレーション対象を広げて行った。広範囲に渡るアクセラレーション事例を通して、大量データ利活用に向けたアクセラレーション戦略の指針を示すことが本研究のねらいである。

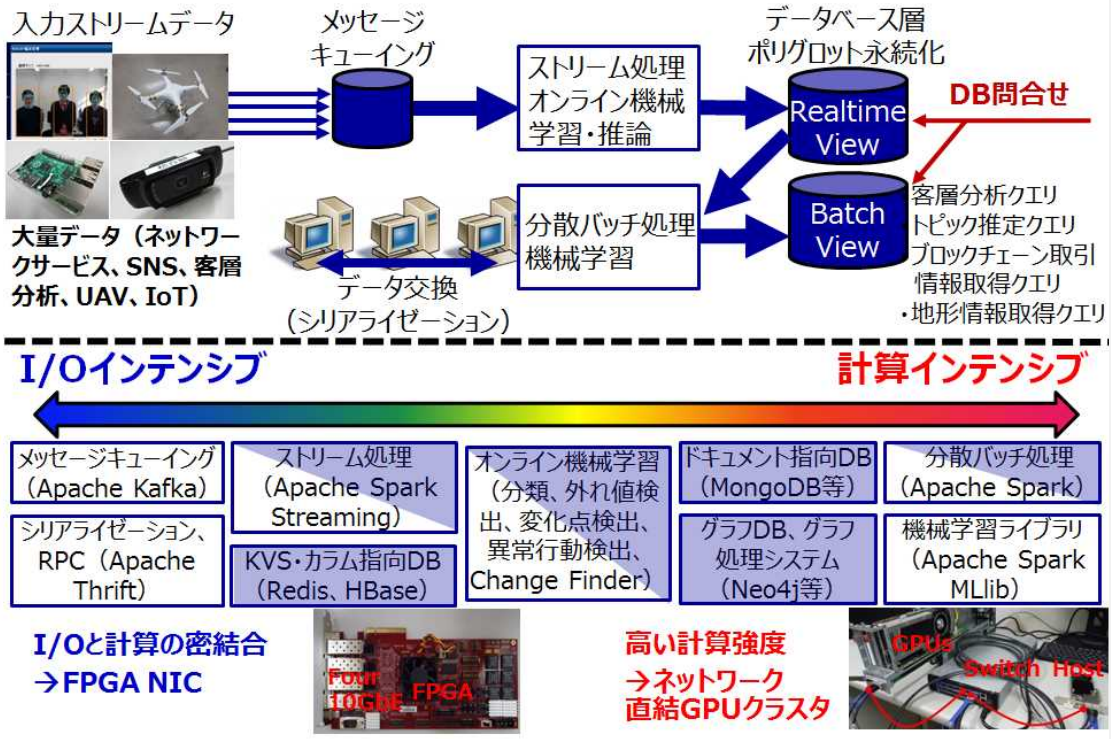
2．　研究成果

（1）概要

大量データ利活用のための統合システムの一例を次ページの図に示す。上図の左端から入力されたストリームデータはストリーム処理やバッチ処理によって加工もしくは集計され、データベースに蓄積されている。①データの収集では、ネットワークサービスや IoT（Internet of Things）デバイスによって生成された大量データを集め、データベース（Realtime View）を更新

する。データの受信と配送にはメッセージキューイングシステム、データベースの更新処理にはストリーム処理フレームワークが利用される。外れ値検出や変化点検出のようなオンライン機械学習も行われる。②データの集約では、1 日に数回など定期的に全データを解析し、集計結果をデータベース(Batch View)に格納する。このためにバッチ処理フレームワークおよび機械学習フレームワークなどが利用される。また、分散処理のために計算機間でデータを交信する際にはシリアライゼーション処理が必要になる。③データの蓄積と検索は、上述の Realtime View と Batch View を格納するデータベースに相当する。Realtime View には時々刻々と更新される直近のデータ、Batch View には定期的に集計される過去の集計済み全データが格納される。ユーザからの問い合わせに対し、Realtime View と Batch View の検索結果を結合したものを返答することで全データ性および即応性の両立を図る。

以上が本研究で想定する大量データ利活用に向けた統合システムである。このためのデータ検索、データ収集、データ集約処理など広範囲に渡るトピックを FPGA や GPU のような汎用アクセラレータを用いていかに高効率化するかという点を探求した。以下にその詳細を示す。



上図：大量データ処理向け統合システムの一例。下図：そのための要素技術と計算強度。

（2）詳細

研究テーマ A「データ検索のアクセラレーション」

KVS は、データをキーとバリューの組として扱う非常にシンプルなデータベースである。計算負荷が低いため、ネットワーク処理がボトルネックになりやすい。そこで、ネットワーク処理と計算を密結合できるデバイスとして FPGA ベースのネットワークインタフェース(FPGA NIC と呼ぶ)を用いた KVS のアクセラレーションを研究した。業績[1]では 10GbE インタフェースを 4 個有する NetFPGA-10G ボードを用いて KVS の一種である Redis のハードウェアキャッシュを実現した。

ただし、FPGA ボードに搭載できる DRAM 容量は大きくはないため、業績[1]では FPGA ベースのハードウェア KVS キャッシュに加え、Linux カーネル内にソフトウェアベースの KVS キャッシュを設けることを提案した。我々は前者を L1 NoSQL キャッシュ、後者を L2 NoSQL キャッシュと呼び、要求されたデータが L1 NoSQL キャッシュに存在しない場合は L2 NoSQL キャッシュをルックアップし、L2 NoSQL キャッシュにもヒットしない場合はアプリケーション層で動作する Redis に問い合わせが行くようにした。

カラム指向型では、各行データは複数個のカラムから構成され、行キーを基にソートされた状態で扱われる。このため HBase では startRow と stopRow を用いた範囲問い合わせも可能である。カラム指向型も KVS 同様、ネットワーク処理がボトルネックとなりやすいため、我々は HBase を対象に FPGA NIC を用いたアクセラレーション手法[23]や Linux カーネル内キャッシュを用いたアクセラレーション手法[9]を提案してきた。

ドキュメント指向型ではデータをドキュメントの集合、グラフ型データベースではデータをグラフとして扱う。前者では正規表現ベースの文字列探索、後者ではグラフ探索を伴う問い合わせが生じる。これらの処理は一般的に計算負荷が高いため、単一 GPU を用いたドキュメント指向型データベース MongoDB の高速化[8]、グラフ型データベース Neo4j の高速化[3]を提案してきた。ただし、データベースで扱うデータに比して GPU のデバイスメモリはあまりにも小さいため、GPU とデータベース間でデータ交信が頻発し、性能の新たなボトルネックとなっていた。そこで、GPU 内にキャッシュできるデータ容量をスケーラブルに増強するために、我々は PCI-Express over 10GbE 技術によってネットワーク接続された GPU クラスタを用いるアプローチを提案している[7]。業績[7]では多数の GPU のデバイスメモリを分散共有メモリとして扱い、分散ハッシュテーブルから着想を得たハッシュ技法によってデータを各 GPU のデバイスメモリに分散させている。

研究テーマ B「データ収集のアクセラレーション」

主要な要素技術はストリーム処理フレームワークとオンライン機械学習である。これらの処理は総じて計算負荷は低めであり、ネットワーク処理と計算を密結合できる FPGA NIC を用いたアプローチが有利であると考えている[2,5,6]。

オンライン機械学習処理の代表例として異常検出が挙げられる。異常検出はさらに外れ値検出、変化点検出、異常行動検出などに分類できる。例えば、ネットワークから流れてくるサンプルデータに対し、通常とは異なる値、傾向、振る舞いなどを検出するために利用できる。我々はこのうち外れ値検出を FPGA NIC で高スループット化する研究を行ってきた[2,6]。業績[2]では、マハラノビス距離を用いて外れ値を検出する手法を FPGA NIC 上に実現し、10GbE ラインレートの 95.8%のスループットを実現した。業績[6]では、k-Nearest Neighbor（kNN）や Local Outlier Factor（LOF）アルゴリズムを FPGA NIC 上に実現する方法を提案している。LOF による外れ値検出では、過去のサンプルデータと入力サンプルデータの比較が必要だが、当然、FPGA NIC の限られた DRAM に過去の全サンプルデータを保持することはできない。そこで、最近アクセスされた過去のサンプルデータクラスタのみを FPGA NIC にキャッシュしておき、一方で、FPGA NIC にキャッシュされている限られた情報だけでは外れ値かどうか判別できないような入力データについてはアプリケーション層にて完全な LOF 処理を行うアプローチを提案した。

ストリーム処理フレームワークに関しては、まず、One-at-a-time 方式と Micro-batch 方式に

大別できる。One-at-a-time 方式ではデータ要素１つ１つに対し決められた処理を適用するのに対し、Micro-batch 方式では短い時間間隔の間に到着したデータ要素をまとめて 1 つの Micro-batch とし、この Micro-batch 毎に決められた処理を適用する。前者の実例として Storm、後者の実例として SparkStreaming が挙げられる。通常のソフトウェア処理の場合、高機能な処理を One-at-a-time 方式で実現しようとすると計算負荷が高くなり過ぎる。一方、Micro-batch 方式では Micro-batch のサイズに応じて外れ値や変化点などのイベントを検出するまでの遅延が増大してしまう。そこで、我々は One-at-a-time 処理と Micro-batch 処理を組み合わせる 2 段階ストリーム処理を研究している。具体的には、Micro-batch 方式の SparkStreaming に対し、FPGA NIC 上に実現した One-at-a-time 処理を組み込むアプローチを提案した**[5]**。

研究テーマ C「データ集約のアクセラレーション」

バッチ処理フレームワークとして我々は Hadoop/MapReduce に加え Spark にも注目している。Spark では RDD（Resilient Distributed Dataset）と呼ばれる分散共有メモリ上にデータを保持し、RDD から別の RDD を生成する処理（Transformation）や RDD を集約する処理（Action）が行われる。

バッチ処理は計算負荷が高くなりやすく、GPU によるアクセラレーションが向くことが多い。業績**[4]**では Spark の RDD に対する Transformation 処理や Action 処理を GPU にオフロードしているが、「研究テーマ A」でも述べたとおり、これだけでは GPU と Spark の間のデータ転送が新たなボトルネックとなる可能性がある。そこで、業績**[4]**では業績**[7]**と同じアプローチを採用している。具体的には、PCI-Express over 10GbE 技術によってネットワーク接続された GPU クラスタを前提に、GPU のデバイスメモリに RDD をキャッシュしておく。GPU のネットワークトポロジによっては CPU から近い GPU、遠い GPU が生じてしまうが、業績**[4]**では RDD の系統グラフを基に頻繁にアクセスされる RDD は近い GPU、それ以外は遠い GPU にキャッシュするアイディアを提案している。

3. 今後の展開

短期的な計画として、既存のラムダアーキテクチャの要素技術のうち本研究によってアクセラレーションできていない処理のアクセラレーション指針をまずは確立したい。具体的には、2 ページ目の図のメッセージキューイングミドルウェアや RPC（Remote Procedure Call）、シリアライゼーションに関する部分である。アクセラレーションのための指針はすでにあり、研究も開始している。これらのアクセラレーションに関する既存研究はほとんど存在しないため、インパクトのある成果になると期待している。また、本研究では 10GbE インタフェースを有する FPGA NIC を用いて研究を行ったが、100GbE 版の FPGA NIC が入手でき次第、100GbE に移行する予定でもある。

本研究では、多数のセンサや IoT デバイス、世界規模のネットワークサービスによって生成される大量データが絶え間なく通過するネットワークに着目し、FPGA NIC 上に各種機械学習アルゴリズムを高性能ハードウェアとして実現した。このアプローチはネットワークを流れるデータからパターンを学習し、知識を得るという用途一般に応用できる。本研究では主として外れ値検出アルゴリズムに焦点を当てたが**[2,6]**、現在では、いくつかの変化点検出アルゴリズムを対象に研究を開始するなど対象範囲を広げている。NIC やスイッチという特殊、かつ、厳し

いリソース制約のもと、実用的な機械学習アルゴリズムをいかに専用ハードウェアとして実現するかはさらに深く研究すべき課題である。実際、NIC やスイッチへの機械学習機能のオフローディング、さらに、そのような NIC やスイッチによる分散協調学習は、エッジヘビーコンピューティングを実現するうえでのキーテクノロジーに成り得ると考えている。

4. 評価

（1）自己評価

当初計画では、アクセラレーションは対象をデータベースのみで、使用するアクセラレータは FPGA のみとしていた。これでは 2 ページ目の図のごく一部分しかカバーできない。その後、研究を遂行していく過程でアクセラレーション対象はラムダアーキテクチャの要素技術の大半をカバーするまでに広がり、FPGA NIC や FPGA スイッチ[1,2,5,6,9,21,22,23]、GPU [3,4,7,8,12]、PCI-Express over 10GbE [4,7]、Linux カーネル内キャッシュ[9]など様々なアクセラレーション戦略を適材適所で使い分けるようになった。このため、研究期間の途中から「大量データ利活用のための各種要素技術を網羅的にカバーすべくアクセラレーション事例を蓄積し、大量データ利活用に向けたアクセラレーション戦略の指針を示す」ことを目指すべく、研究目標を緩やかに変更していった。上述の「今後の課題」で言及したとおり、まだまだ研究しなければならないアクセラレーション課題は残っているものの、3 年半という研究期間を鑑みるに最善を尽くすことができたものと考えている。今後は、NIC やスイッチへの機械学習機能のオフローディング、さらに、そのような NIC やスイッチによる分散協調学習などエッジヘビーコンピューティングに向けた要素技術群をハードウェア屋の観点から探求して行きたい。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータの統合利活用のためには，強力なコンピュータやネットワークが不可欠である．一方，ハードウェアを増強する際の制約要因の中で，消費電力の占める割合が，近年，次第に増加しつつある．もちろん，ハードウェア自体のコストも大きな制約条件になる．

本研究は，電力効率とコストパフォーマンスの高いビッグデータ統合利活用を可能とするシステム・アーキテクチャを探求し，その有効性を実証的に示すものである．ビッグデータ取得時に利用されるストリーム処理，蓄積と検索に利用される NoSQL の処理，解析に利用される Map-Reduce などのバッチ処理について，典型的なものを対象に，消費電力削減と性能向上の両立を目指している．そのために，FPGA, GPU などの既存のアクセラレータを適所適材で組み合わせ，さらに NIC やアクセラレータ内のものを含むメモリ階層の潜在能力を引き出す方式を提案している．そして，これらの分野において，国際会議での最優秀論文賞の受賞や多数の論文の公表など，優れた学術成果をあげている．

本研究の成果である要素技術群と研究の過程で得られた知見が，今後，多様な実アプリケーションのワークロードや要件を反映した統合アーキテクチャの提案として結実することを期待したい．

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. Yuta Tokusashi, Hiroki Matsutani, "A Multilevel NOSQL Cache Design Combining In-NIC and In-Kernel Caches", Proc. of the 24th IEEE International Symposium on High Performance Interconnects (Hot Interconnects 24), pp.60-67, Aug 2016.
2. Ami Hayashi, Yuta Tokusashi, Hiroki Matsutani, "A Line Rate Outlier Filtering FPGA NIC using 10GbE Interface", ACM SIGARCH Computer Architecture News (CAN), Vol.43, No.4, pp.22-27, Sep 2015.
3. Shin Morishima, Hiroki Matsutani, "Performance Evaluations of Graph Database using CUDA and OpenMP-Compatible Libraries", ACM SIGARCH Computer Architecture News (CAN), Vol.42, No.4, pp.75-80, Sep 2014.
4. Yasuhiro Ohno, Shin Morishima, Hiroki Matsutani, "Accelerating Spark RDD Operations with Local and Remote GPU Devices", Proc. of the 22nd IEEE International Conference on Parallel and Distributed Systems (ICPADS'16), pp.791-799, Dec 2016.
5. Kohei Nakamura, Ami Hayashi, Hiroki Matsutani, "An FPGA-Based Low-Latency Network Processing for Spark Streaming", Proc. of the 4th IEEE International Conference on Big Data (BigData'16) Workshops, pp.2410-2415, Dec 2016.
6. Ami Hayashi, Hiroki Matsutani, "An FPGA-Based In-NIC Cache Approach for Lazy Learning Outlier Filtering", Proc. of the 25th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP'17), 8 pages, Mar 2017.
7. Shin Morishima, Hiroki Matsutani, "Distributed In-GPU Data Cache for Document-Oriented Data Store via PCIe over 10Gbit Ethernet", Proc. of the 22nd International European Conference on Parallel and Distributed Computing (Euro-Par'16) Workshops, 12 pages, Aug 2016.
8. Shin Morishima, Hiroki Matsutani, "Performance Evaluations of Document-Oriented Databases using GPU and Cache Structure", Proc. of the 13th IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA'15), pp.108-115, Aug 2015.
9. Korechika Tamura, Hiroki Matsutani, "An In-Kernel NOSQL Cache for Range Queries Using FPGA NIC", Proc. of the 1st International Conference on FPGA Reconfiguration for General-Purpose Computing (FPGA4GPC'16), pp.13-18, May 2016.

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

招待講演等

10. 松谷 宏紀, "ビックデータ利活用のための計算基盤", 電子情報通信学会 コンピュータシステム（CPSY）研究会, 招待講演, Dec 2016.
11. 鯉渕 道紘, 松谷 宏紀, 藤原 一毅, "大規模コンピュータ・ネットワークの建築学", 国

立情報学研究所 H28 年度第 2 回 産官学連携塾, Oct 2016.

12. Hiroki Matsutani, "Accelerator Design for Various NOSQL Databases", The 16th International Forum on MPSoC for Software-defined Hardware (MPSoC'16), Invited Talk, Jul 2016.
13. Hiroki Matsutani, "Accelerator Design for Various NOSQL Databases", Big Data French-Japanese Workshop, The Embassy of France in Japan, Invited Talk, Nov 2014.
14. 松谷 宏紀, "ビッグデータ向け計算機アーキテクチャの研究動向と研究事例", インターネットコンファレンス 2014 (IC'14), 招待講演, Nov 2014.
15. 松谷 宏紀, "ポリグロット永続化のための NoSQL アクセラレータ", 情報処理学会 データベースシステム(DBS)研究会, 招待講演, Aug 2014.
16. 松谷 宏紀, "多様な構造型ストレージ(NOSQL)のためのアクセラレータ設計", 日本電気株式会社, 招待講演, Jul 2014.

受賞等

17. 情報処理学会 特選論文 (2016)
    (受賞論文:FPGA NIC 向けノンパラメトリックオンライン外れ値検出機構)
18. Best Paper Award, The 6th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART'15)
    (受賞論文:A Line Rate Outlier Filtering FPGA NIC using 10GbE Interface)
19. 電子情報通信学会 コンピュータシステム研究会 優秀若手講演賞 (2014)
    (受賞論文:カラム指向型データベース向けハードウェアキャッシュ機構の検討)