様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２２　日現在

機関番号：　１２６１２
研究種目：　若手研究(B)
研究期間：　2015～2016
課題番号：　１５Ｋ１５９５９

研究課題名（和文）密結合型専用ハードウェアを用いた大規模データ分散処理システムの開発

研究課題名（英文）Research for a Distributed System with Tightly-Coupled Specialized Hardware for Bigdata appications

研究代表者

吉見　真聡（Yoshimi, Masato）

電気通信大学・大学院情報理工学研究科・助教

研究者番号：００５４８０００

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究は，集積された大量のデータを質の高い情報に変換するビッグデータ解析において，FPGAを用いて構成するアクセラレータを導入することで，現実的な計算時間と可能な限り低いコストで実現する計算機システムの開発と性能の実証を行った．多数のデータ走査を要するビッグデータ解析の一部を，計算のボトルネックとなる通信と同時に行うことにより，大きな性能向上が可能となる．データベース処理と計算生物学のアプリケーションを対象にした研究を通して，計算速度を数倍から数十倍高速化できることを確認した．

研究成果の概要（英文）：An objective of this research aims to develop an energy-efficient and high-performance computer system for bigdata applications which convert massive amount of data to high-quality information. An FPGA-based accelerator couples storage, network and the main memory accelerates such applications to overwrap computation and transmission. We confirmed that the PC-cluster with accelerators overcomes software-based implementation over dozens of time by basic database operations and biological computation.

研究分野：　コンピュータシステム

キーワード：　コンピュータシステム　ビッグデータ　FPGA　リコンフィギャラブルシステム　データベース　計算生物学　コンピュータアーキテクチャ　アクセラレータ

１．研究開始当初の背景

近年，コンピュータ，携帯端末，センサーなどあらゆる情報機器がネットワークに接続され，それらからもたらされるビッグデータの活用が進んでいる．集積され続けるデータにもとづくビジネスインテリジェンスは，意思決定支援に重要な役割を担うようになっている．

蓄積された大量のデータを，多数の計算機を用いて解析するソフトウェアフレームワークが実用化されてきており，低い開発コストでアプリケーション開発が可能になりつつあるが，多次元データに対する複雑かつ分析的な問い合わせへの対応は未だ充分でない状況にある．オンライン分析処理(OLAP)と呼ばれるこのような処理を分散システムで実行する場合，長い計算時間を要する．データの読み込みと通信の頻度と速度の問題を改善する様々な高速化の研究が取り組まれてきたが，OLAP で重要となるストレージ，ネットワーク，専用ハードウェアの協調による高速化は現在でも実用化されていない．

ビッグデータ処理に要する分散システムは，Hive，Impalra などのクエリエンジンや，MapReduce，Spark などのソフトウェア技術に関する研究開発が盛んである．これらはアプリケーションの開発コストを低減させる一方で，計算機システムの大規模化によるエネルギー効率の低下を招いてもいる．特に，分散データの結合演算においては，通信オーバヘッドにより計算時間が大きく増加する問題がある．

ビッグデータのデータ蓄積速度は計算機の性能向上速度よりも速く，従来型の計算機を並列に動作させる計算機システムの場合，計算機の規模が増大を続けてしまっている．この状況を打開し，高効率な大規模データ処理基盤に関する取り組みとして，専用ハードウェアを活用する方法が挙げられる．GPU やメニーコアプロセッサのような，CPU バウンドな計算を並列プロセッサによって高速化する従来型のハードウェアとは異なり，データ入出力に注目した専用ハードウェアの活用は比較的新しい概念である．

現在までに，ストレージに演算機能を付加する Active Disk 構造や，既存の汎用的な分散処理アルゴリズムに合わせて開発するアクセラレータが使用されてきた [1,2]．I/O バウンドな処理は，計算結果を得るために複数回データスキャンを要するワークロードであるストレージの間に計算用のハードウェアを接続した Active Disk 構造により，読み出しデータから必要なフィールドのみを抽出する．一方，CPU バウンドな処理は文字列検索などデータ単位あたりの演算量が多いワークロードであるため，[1]のようにメニーコアアクセラレータへオフロードが有効である．

本研究課題開始時点では，これらの IO バウンドな処理に着目した専用ハードウェアの活用は計算機ごとに搭載されるものが多く，計算機間の通信にも活用する仕組みは実用化されていなかった．

２．研究の目的

本研究の主な目的として，以下の 3 つが挙げられる．

① 大規模データの高速，低電力な専用ハードウェアの開発を通して，低消費エネルギー社会を実現する技術に貢献する

② 専用ハードウェアを搭載する複数の計算機を協調動作させるソフトウェアを開発し，ビッグデータ処理アプリケーションの開発コストを低減する

③ 上記 1,2,を用いて分散データベース管理システム(DBMS)を構築し，オンライン分析処理(OLAP)の実アプリケーションの実行性能，消費電力等を実験的に明らかにし，有効性を示す

本研究では，ビッグデータ処理を加速する新たな計算機システムを提案し，実証実験により性能を評価する．

高速化の要点は，計算機間，および，計算機内のストレージと主記憶の間を，FPGA を用いて構成する専用ハードウェアで密に結合した構造である．ビッグデータ処理の計算時間の多くが，大量のデータの入出力に占められることに着目すると，主記憶を介さずにデータを通信する機構と，データを読み出す間に同時に計算を行う機構の両方を実現する専用ハードウェアが，高速化に資すると考えられる．そこで，図 1 に示すように，通信とデータ入出力を担う専用ハードウェアを実装し，ビッグデータ処理に計算速度，消費エネルギーの観点から実証実験を通して性能の向上と効率性を明らかにする．

[1] T. Honjo, *et.al.*: Hardware acceleration of Hadoop MapReduce, BigData Conference, 2013.

[2] G. Davidson, et.al.: Data-centric computing with the netezza architecture, SANDIA REPORT, pp.1–24, 2006.

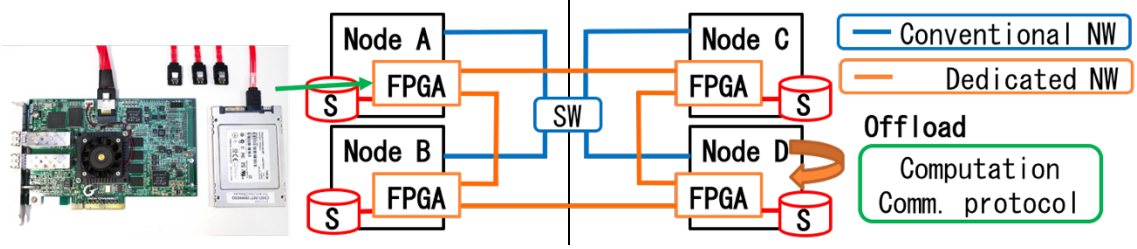

図 1　密結合型ハードウェアを用いた分散処理システム

３．研究の方法

下記の２つの研究に並行して取り組み，２．の目的を実証する．

1. 密結合型ハードウェアの開発
2. アプリケーションの実装，評価

まず第１に，FPGA を介してストレージと通信を接続する専用ハードウェアを実装する．第２に，ビッグデータ解析には様々なアプリケーションのうち，基本的なものをいくつか選択し，開発を続ける専用ハードウェアを対象とした実装と評価を行う．

計算機の規模の増大とアプリケーションの多様化を両立し，有用性を実証する．

４．研究成果

（１）密結合型 FPGA ボードの開発

株式会社アバールデータが開発した FPGA ボード APX7142 をベースに，SSD を接続可能なインターフェイスを増設し，自研究室で開発した SATA コアを接続できるように改造した（図１左側の FPGA ボード，図 2）．SATA2 速度で 4 台の SSD を接続することができ，1GB/s を超える入出力性能が得られることを確認した（図 3）．この FPGA ボードを 16 枚作成し，16 台の PC クラスタに搭載して，光のリングネットワークで相互接続した．この仕組みを，後述するアプリケーションの実装と評価で利用した（図２）．

図 2　開発した FPGA ボードと実験環境

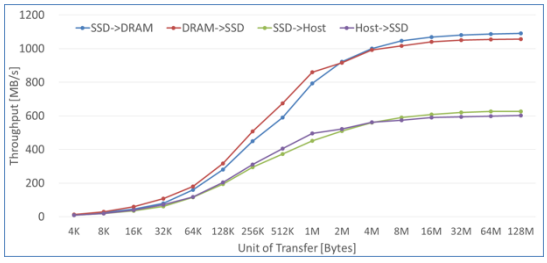

図 3 SSD4 台のデータアクセス性能

（２）大規模テーブル結合処理の高速化

数十 GB のテーブルの結合処理を，密結合型 FPGA ボード間で行われる通信を利用して高速に行うアルゴリズムを開発（図 4）し，実証実験を行った．テーブルは FPGA に接続された SSD に格納され，ホスト PC 群の主記憶を介さずに，FPGA ボード間の直接通信でデータを交換する．このことによりメモリウォール問題を軽減し，ホスト PC の CPU リソースを本来の計算のみに振り分けることができるようになる．

16 ノードのシステムで，128GB テーブル同士の結合処理において，直接通信による効果は 20%程度あり，アルゴリズム上の性能向上と合わせて，チューニングされたソフトウェア実行と比べ，計算速度が約 1.5 倍から 5 倍程度向上することを確認した（図 5）．

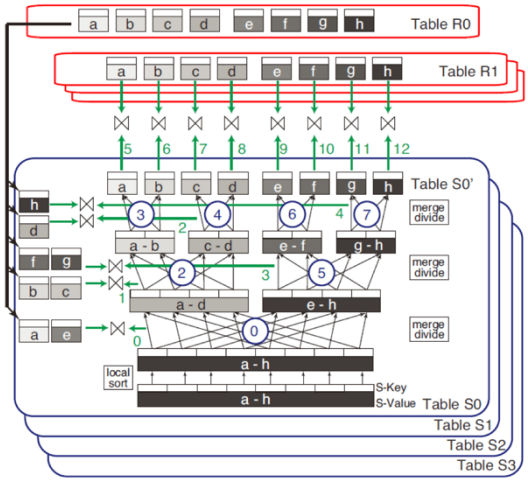

図 4 並列結合アルゴリズム PPJoin

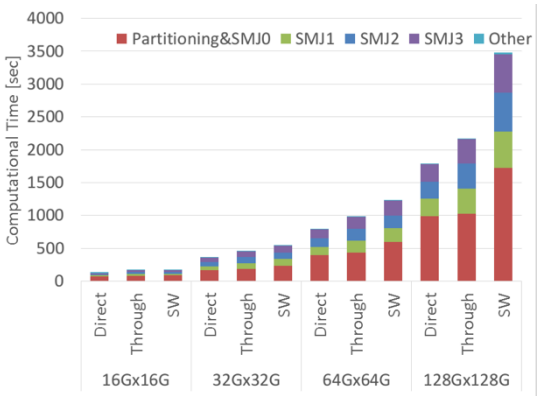

図 5 データベース結合処理の性能向上

（３）生物相同性検索システム BLAST

より科学計算向けのアプリケーションとして，DNA やアミノ酸配列の類似性を検索する生物相同性検索システム BLAST を対象に，高速実行する計算機システムの開発に取り組んだ．FPGA ボードを介してデータベースを高速に分割，分配するアルゴリズムと，SSD に格納

されたデータベースを読み出す際に,BLAST の計算の前半部を FPGA 内で行う専用ハードウェアを実装した.

データベースの高速配布は，配列の先頭に分割用のタグを付与し，受信側でタグを判定してデータを取り込む仕組みであり，従来のソフトウェア(mpiBLAST)でのデータ配布と比べ,10 倍以上の性能向上が確認された(図 6).

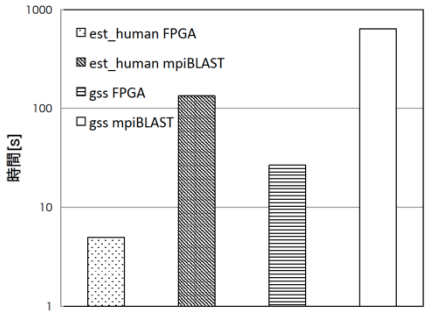

図 6　BLAST データベースの高速

BLAST の計算の前半部は,文字列に対する，単純だが膨大な演算を繰り返す必要があり，計算時間全体の 70～80%を占める．この演算は，単純な文字比較であるため，論理回路として FPGA 上に構成することができる(図 7).

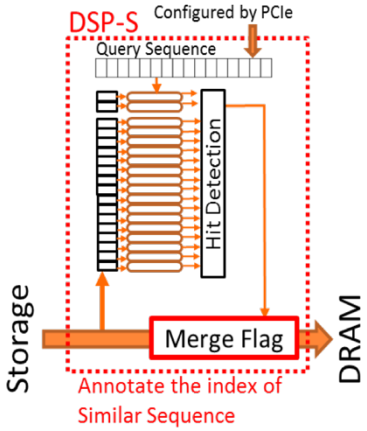

図 7 BLAST 前処理部のハードウェア実装

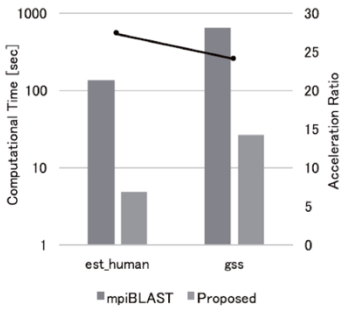

図 8 BLAST の性能比較

2 種類の比較的大きい配列を対象に DB 検索による実験を行ったところ，前処理の演算がデータの読み出し時間とオーバラップされ，計算速度が 20 倍以上向上したことが確認された.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　３件）

1. Masato Yoshimi, Yasin Oge, Tsutomu Yoshinga: Pipelined Parallel Join and Its FPGA-based Acceleration, Transactions on Reconfigurable Technology and Systems, ACM, (査読有り，採録決定).
2. Masato Yoshimi, Celimuge Wu, Tsutomu Yoshinaga: Accelerating BLAST Computation on an FPGA-enhanced Cluster, in Proceedings of The Fourth International Symposium on Computing and Networking, pp.67-76, 2016(査読有).
3. Takuma Nakajima, Masato Yoshimi, Celimuge Wu, Tsutomu Yoshinaga: A Light-weight Content Distribution Scheme for Cooperative Caching in Telco-CDNs, in Proceedings of The Fourth International Symposium on Computing and Networking, pp.126-132, 2016(査読有).

〔学会発表〕（計　５件）

1. 野島幸大，城間隆行，中島拓真，吉見真聡，策力木格，吉永努：ネットワーク内キャッシュによるISPネットワーク通信電力の削減．信学技報，pp. 223-228，電子情報通信学会，2016年8月10日，キッセイ文化ホール(長野県松本市).
2. 溝田敦也，城間隆行，中島拓真，吉見真聡，策力木格，吉永努：インタークラウドを活用した自動災害復旧システム．信学技報，pp. 229-234，電子情報通信学会 2016年8月10日，キッセイ文化ホール(長野県松本市).
3. Masato Yoshimi, Yasin Oge, Celimuge Wu, Tsutomu Yoshinaga: Design Evaluation of Low-Latency Handshake Join on FPGA. 信学技報，pp. 253-258，電子情報通信学会 2016年3月24日，福江文化会館(長崎県福江市).
4. 中島拓真，城間隆行，吉見真聡，策力木格，吉永努：動画の人気変動に追従する異種キャッシュ混在ネットワークの検討．信学技報，pp. 247-252，電子情報通信学会 2016年3月24日，福江文化会館(長崎県福江市).
5. 小川芳光，オゲ ヤースィン，吉見 真聡，策力木格，吉永 努：データストリーム集約演算 HW の並列化．信学技報，pp.79-84，2016年1月19日慶應義塾大学日吉キャンパス(神奈川県横浜市).

〔その他〕
ホームページ等
http://comp.is.uec.ac.jp/

６．研究組織
(1)研究代表者
吉見　真聡（Masato Yoshimi）
電気通信大学・大学院情報理工学研究科・助教
研究者番号：00548000

(2)研究分担者
無
(3)連携研究者
無
(4)研究協力者
無