様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年５月３０日現在

機関番号：12608
研究種目：若手研究(B)
研究期間：2011～2012
課題番号：23700031
研究課題名（和文）　ヘテロ型クラスタ計算機上でのペタスケール大規模データ処理
研究課題名（英文）　Peta-scale Data Processing for Large-scale Heterogeneous Clusters

研究代表者
佐藤　仁　　(Sato　Hitoshi)
東京工業大学・学術国際情報センター・特任助教
研究者番号：00550633

研究成果の概要（和文）：
次世代のヘテロ型アーキテクチャ上での大規模データ処理の実現のための基盤技術の確立を目的として，GPU を搭載したヘテロ型クラスタ計算機向けの MapReduce 処理系の研究開発を進めた．TSUBAME2.0 スーパーコンピュータ上で 256 ノード，768GPU を用いて，MapReduce モデルを採用したグラフ処理アプリケーションを対象に実行し，大規模実行における性能のスケーラビリティ，ボトルネックを明らかにし，さらに次世代の超大規模データ処理基盤の実現のための礎を築いた．

研究成果の概要（英文）：
We developed a MapReduce framework for GPU-based heterogeneous clusters as an instance of a future large-scale new platform with heterogeneous many core processors for big data applications. We confirmed performance scalability and overhead of our MapReduce framework by running MapReduce-based graph processing applications using 256 nodes 768 GPUs of the TSUBAME2.0 supercomputer. Our results lead the foundation of software technology for next-generation extreme big data processing.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウエア
キーワード：並列処理・分散処理・大規模データ処理

１．研究開始当初の背景

近年，情報技術の発達により，人類が取り扱うデータ量が爆発的に増加している．例えば，Google 社では 1 日に 20 ペタバイト以上のデータに対してクラスタ計算機を用いた並列分散処理を行っていると報告されている．科学技術計算の分野においてもこの傾向は当てはまり，高エネルギー物理学，生物学，天文学などの分野においても大量のデータに対して並列分散処理を行う試みが広く行われている．

このような大規模データに対する並列分散処理を行うためのプログラミングモデルとして MapReduce が近年注目されている．MapReduce は，分散した key-value ペアデータに対して Map や Reduce などの統一的な操作を並列で適用し，データへのアクセスの局所性を考慮したスケーラブルなデータ処理を実現する技術である．典型的な例では，ウェブのデータ解析などの処理に適用され，数千・数万台規模の大規模なクラスタ計算機においてノード数に応じたスケーラビリティを得ている．しかし，このようにクラスタ計算機上での並列性で性能を得たとしても，それによる電力や設置スペースの増加，ひいてはコストの増加が問題となる．例えば，多く

の大規模なクラスタ計算機(東京工業大学(東工大)の TSUBAME スーパーコンピュータ(スパコン)や Google,Yahoo,Microsoft 社のデータセンターなど)では，消費電力量が数百 KW から数十 MW に達しており，また，それだけでなく，冷却，騒音，重量などの設置性が問題となる.つまり，同じ性能が得られるのであるならば，計算機としては省電力でかつ省スペースであることが望ましい.

一方で，近年，省電力，省スペースで高い性能を実現する技術が多数登場してきている．例えば，Intel,AMD などの汎用 CPU のプロセッサのマルチコア化や,GPU,Cell などのアクセラレータ，フラッシュメモリを記憶装置に利用した SSD などが挙げられる．また，近年の大規模計算機システムではこれらの技術が多く採用されつつある(東工大 TSUBAME1.2 やその次期スパコンシステムである TSUBAME2.0 をはじめ，米国のペタフロップス級のスパコンシステムであるロスアラモス国立研究所の Roadrunner や中国の国立スーパーコンピューティングセンター深センの Nebulae など)．しかし，このような環境で既存の MapReduce 処理を適用すると，ヘテロ型のアーキテクチャ上での汎用 CPU やアクセラレータの演算性能や，ホストマシンのメインメモリ，アクセラレータ上のメモリ,SSD や HDD など様々な階層の記憶媒体へのアクセス性能を考慮しなければならず，効率的な実行は困難である．このため，このような次世代のヘテロ型アーキテクチャを対象にした効率的な大規模データ処理を実現するための基盤技術の研究開発が必要である.

２．研究の目的

研究目的は以下の 4 点に細分化される.

(1)ヘテロ型クラスタ計算機を対象にした大規模データ処理実行基盤の実現

汎用 CPU コアと GPU アクセラレータが混在したヘテロ型クラスタ計算機上での大規模データ処理に適した MapReduce システムミドルウェアの設計と実装を行う．アクセラレータを対象にした MapReduce 処理系として，Mars や MapReduce on Cell などが既に存在する．これらの研究は，現状では単一計算機ノード上でのアクセラレータの利用に留まっているのに対し，本研究では，汎用 CPU コアと GPU アクセラレータが混在した大規模ヘテロ型クラスタ型計算機を対象としている点が大きく異なる．特に，GPU アクセラレータを備えたスパコンである東工大 TSUBAME2.0 と連携して研究を進めることにより，GPU アクセラレータを多数搭載したヘテロ型大規模並列計算機システムでの効率的な MapReduce 処理の実現，及び，大規模データ処理の基盤技術の確立を目指す.

(2)ヘテロ型クラスタ計算機を対象にしたタスクスケジューリング手法の確立

ヘテロ型クラスタ型計算機上で効率的な MapReduce 処理を実現するためには，Map・Reduce タスク処理の特性や,汎用 CPU コアや GPU アクセラレータなどの実行環境，また，処理対象となるデータの配置を考慮したタスクスケジューリングが必要となるが，既存の MapReduce 処理系ではこれらは考慮されていない．そこで，我々は，ヘテロ型クラスタ計算機を対象にしたタスクスケジューリングアルゴリズムの複合的数理モデルを構築し，TSUBAME2.0 において評価・検証を行う.

(3)大規模データ処理に特化したアクセラレータへの効率的なデータ転送手法の実現

大規模データ処理をヘテロ型クラスタ計算機に適用するためには，並列ファイルシステム上のデータをホストマシンのメモリを介して GPU アクセラレータ上のメモリへストリームとして転送する必要があるが，このようなデータ転送の最適化手法の確立は未だ発展途上であり，大規模データ処理への適用事例は少ない．そこで，我々は，ペタバイト規模のデータ処理を目指し，並列ファイルシステムから複数の GPU アクセラレータへの効率的なデータ転送手法を構築し，TSUBAME2.0 において評価・検証を行う.

(4) *タスク実行時のモニタリング手法の確立*

上記の(2)ヘテロ型クラスタ計算機を対象にしたタスクスケジューリング手法や(3)GPU アクセラレータへの効率的なデータ転送手法を実現するためには，実行されるタスクの精緻なモニタリング(具体的には，タスク処理時の演算性能，メモリ使用量，I/O など)を行い，タスク処理時の挙動やデータへのアクセスパターンの把握が必要となる．我々は，MapReduce 実行処理系と GPU アクセラレータを含む計算機ノード上の性能プロファイリング手法とを連携して，タスク実行時の網羅的なモニタリング手法を確立し，モニタリングツールとして開発を行う.

本研究は，実際の科学技術計算アプリケーションへの適用を主眼としており，開発したシステムは，オープンソースソフトウェアとして公開することを視野に入れている．単に学術的知見のみならず，論文を超えたインパクトのある成果を狙う．また，大学の所有するスパコンシステム上の科学技術計算に留まらず，データセンターでのウェブデータ解析処理など，実社会への応用範囲が広い基礎研究である点も特色である．また，本研究は，一義的には，MapReduce 処理系を対象として

いるが，個々の要素技術は，その他の大規模データ処理手法(例えば，DAGMan， Pegasus，DryadLinQ などのデータ処理用ワークフローシステム)にも適用可能であると確信している．

３．研究の方法

２ヵ年計画でヘテロ型クラスタ計算機を対象にした大規模データ処理基盤の確立を行う．初年度は，プロトタイプとして，1〜4 台規模の GPU アクセラレータを搭載したヘテロ型クラスタを対象に，既存の MapReduce 処理系に対し，タスクのモニタリング，スケジューリング，GPU アクセラレータへのデータ転送の効率化手法などの要素技術の研究開発を行う．次年度では，プロトタイプを数百から数千台規模に適用することを目指し，CPU，GPU を含めたコア数や，MapReduce ジョブが対象とするデータサイズに応じたスケーラビリティを TSUBAME2.0 で検証する．研究成果は，単に学術会議での報告にとどまらず，オープンソースソフトウェアとして公開することを目指す．

４．研究成果

初年度は，汎用 CPU コアと GPU アクセラレータが混在したヘテロ型クラスタ計算機を対象にした MapReduce システムの基本的な設計と実装を行った．まず，実アプリケーション GPU 上の MapReduce 処理の性能特性を明らかにするために，1 台の計算ノード上の GPU を対象 にした既存の MapReduce 実装である Mars に対して，Generalized Iterative Matrix-Vector multiplication(GIM-V)モデルによるグラフ処理，具体的には，PageRank，Random Walk with Restart，Connected Components を実装し，Hadoop による GIM-V 処理実装である PEGASUS との比較を行った．その結果，1 反復あたり，PageRank で 2.17〜9.53 倍，Random Walk with Restart で 2.18〜5.47 倍、Connected Components で 2.41〜8.46 倍の高速化を確認し，GPU による MapReduce 処理の有効性を示した．更に，Mars に対して複数ノード上の GPU を使用した MapReduce の実行を可能にする 拡張を行った．この拡張した MapReduce に対して，GIM-V モデルによる PageRank を実装し，TSUBAME2.0 スーパーコンピュータ上の 64 ノード，64GPU を使用して実行したところ、Map 処理が 7.17 倍の高速化を示すことを確認した．一方で，Sort，Reduce 処理では高速化を示さず，性能改善の余地があることを確認した．これは，GPU 毎の負荷やデータ転送量が不均衡になることが要因であることを確認しており，更に大規模なヘテロ型クラスタ計算機上で，効率的な MapReduce 処理を実現するためには、GPU 毎のタスクスケジューリング，データ割り当てなどの動的な自動チューニング必要であるという指針を得た．

次年度は，前年度までの成果に基づき，MapReduce システムの大規模環境での実行に注力した．具体的には，単一 GPU 向けの既存の MapReduce 処理系である Mars を，大規模複数 GPU 向けの拡張をさらに進め，GPU アクセラレータを搭載した大規模ヘテロ型スーパーコンピュータ上で効率的に実行するための最適化を行った．更に，拡張した Mars の実装上に，MapReduce プログラミングモデルに基づいたペタバイト級の大規模グラフ処理手法である GIM-V と呼ばれる行列ベクトル積の処理モデルを適用し，ページランクを移植した．また，GIM-V によるページランクアルゴリズムを，TSUBAME2.0 上の 256 ノード，768GPU を用いて実行したところ，2^30(10 億 7 千万)頂点，2^34(172 億)枝のグラフに対して 87.04ME/s の性能を達成することを確認し，また，2^29 頂点，2^33 枝を超えるグラフに対して，CPU のみを用いる場合と比較して 1.52 倍の高速化を確認した．一方で，shuffle 処理，reduce 処理における性能低下やオーバーヘッドを確認した．また ，これらの経験を元に，更なる大規模計算環境へ適用や性能最適化，最新デバイス機能の適用を目的に，数千〜数万のアクセラレータ を搭載したスパコン上でのデータ並列処理を目指したソフトウェア基盤として Hamar(Highly Accelerated MapReduce)の開発を進め，さらに次世代の超大規模データ処理基盤の実現のための礎を築いた．

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 0 件）

〔学会発表〕（計 9 件）

(1) Koichi Shirahata, Hitoshi Sato, Toyotaro Suzumura, and Satoshi Matsuoka. “A Scalable Implementation of a MapReduce-based Graph Processing Algorithm for Large-scale Heterogeneous Supercomputers” In Proceedings of the 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid (CCGrid2013), Delft, The Netherlands, 15 May 2013.

(2) Koichi Shirahata, Hitoshi Sato, Toyotaro Suzumura, Satoshi Matsuoka, “”A Scalable Implementation of a MapReduce-based Graph Algorithm for Large-scale Heterogeneous Supercomputers”, GPU Technology

Conference 2013 (GTC2013), San Jose, USA, 18 Mar 2013. (Poster)

(3) Koichi Shirahata, Hitoshi Sato, Toyotaro Suzumura, and Satoshi Matsuoka. “A GPU Implementation of Generalized Graph Processing Algorithm GIM-V” In Proceedings of the 3rd International Workshop on Parallel Algorithm and Parallel Software (IWPAPS 2012), pp.207-212, Beijing, China, 28 Sep 2012.

(4) Hitoshi Sato and Satoshi Matsuoka. “Hadoop on the Tsubame2.0 Supercomputer” In Proceedings of the 6th International Conference on Ubiquitous Information and Technologies & Applications, Seoul, Korea, 15 December 2011. (Invited Paper)

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況 (計 0 件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況 (計 0 件)

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
Hadoop GPU 拡張 公開ページ
https://github.com/koichi626/hadoop-gpu.git

６．研究組織
(1)研究代表者
佐藤 仁　(Sato Hitoshi)
東京工業大学・学術国際情報センター・特任助教
研究者番号：00550633

(2)研究分担者 ()
研究者番号：

(3)連携研究者 ()
研究者番号：