戦略的創造研究推進事業　CREST
研究領域「ビッグデータ統合利活用のための
次世代基盤技術の創出・体系化」
研究課題「EBD：次世代の年ヨッタバイト処理に
向けたエクストリームビッグデータの基盤技術」

研究終了報告書

研究期間　2013年　10月～2019年　3月

研究代表者：松岡　聡
（東京工業大学情報理工学院
特任教授）

§1 研究実施の概要

(1)実施概要

将来 Zeta(1021)Byte/日(あるいは Yotta(1024)Byte/年)という、今の Google/Amazon の個々の IDC に代表される 10 万ノード級のクラウドのデータ処理能力の、最大で 10 万倍に至る処理能力を達成するための EBD(Extreme Big Data)システム基礎技術の確立を達成することを目標とし、そのためにスーパーコンピューティング技術、特にメニーコア超並列処理と広帯域低遅延ネットワーク技術・不揮発性メモリ技術・及び高性能データベース技術を融合し、単なる「ビッグデータ」から「EBD」への相転移的な技術革新をはかることを目的に研究を実施した。

システムソフトウェアに関しては、次世代のストレージ装置として期待されるフラッシュデバイスや不揮発性メモリを想定した高速な並列データ・アクセスを実現するためのローカルオブジェクトストアの設計の最適化や、EBD アプリケーションとのコデザインによるネットワーク構成方法の設計、数千台の GPU を搭載したビッグデータマシンを対象とした超高速大規模分散ソートや数千・数万台の計算ノードまでスケールするグラフ処理カーネル、機械学習のための並列化学習カーネルなどの研究開発を進め、世界トップの成果を得るとともに、作成したソフトウェアを広く一般に公開した。

EBD インターコネクトについては、従来のスパコンのインターコネクトとは異なり、アクセスパターンが(隣接通信などに)事前に最適化されていない非定型なデータ流に対して広帯域低遅延化が要求されるものであるが、この点でランダム性を利用したネットワークトポロジの設計技術を開発し、その結果、同じ次数、スイッチ数の既存のデータセンター、スパコンネットワークと比べて最長通信遅延、平均通信遅延を大幅に削減した。さらに配線遅延を削減するマシンルームへの配線法などを開発した。また Approximate Computing の考え方に基づく EBD アプリケーションとネットワークのコデザインを行うことで広帯域通信を実現できることを示した。

また、これらを基盤に、大規模ゲノム相関、社会シミュレーション、超高速センサ・エクサスケール気象データ同化、などを代表的な「EBD アプリケーション」のインスタンスとして類型し、EBD アーキテクチャ上での実行に向けた設計、及び、評価を進めた。大規模ゲノム相関に関しては、類似配列検索の高速化とマウスの腸内細菌叢やヒトの歯周病巣細菌叢への応用に関し EBD 対 EBD の相同性計算を実現するソフトウェア GHOSTZ-GPU などを開発し、シンガポールの National Supercomputing Center Singapore が所有する Aspire-1 スーパーコンピュータ上、および同国 NTU 大学の SCELSE 研究所サーバに移植し、現地の研究者に開放した。社会シミュレーションに関しては、大規模エージェントベースシミュレーション基盤へ性能最適化のためのアーキテクチャと機構の提案と実装をし、気象データ同化に関しては、4D-LETKF (4-Dimensional Local Ensemble Transform Kalman Filter)ワークフローを対象に観測データサーチアルゴリズムの改良を行い、性能分析とその向上を行った。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1. Graph500 ベンチマークにおいて世界一を 8 回連続で達成

概要:

CREST「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」(研究総括:佐藤 三久 理研計算科学研究機構)における研究課題「ポストペタスケールシステムにおける超大規模グラフ最適化基盤」(研究代表者:藤澤 克樹、研究参加者:上野 晃司等)や理研計算科学研究機構(AICS)、富士通などと共同研究で、スパコンにおけるグラフの幅優先探索を行うビッグデータベンチマーク Graph500 で世界一を京コンピュータ上で 8 回取得し、統合アーキテクチャにおけるビッグデータ処理のアルゴリズムを大幅に進化させた。
http://www.graph500.org.

2. Keita Iwabuchi, Hitoshi Sato, Yuichiro Yasui, Katsuki Fujisawa and Satoshi Matsuoka, “NVM-based Hybrid BFS with memory efficient data structure”, The IEEE International Conference on Big Data 2014 (IEEE BigData 2014) pp.529-538, 2014. (DOI: 10.1109/BigData.2014.7004270)

概要:

次世代の大規模計算機環境の実現において価格面のコストと消費電力の上昇が大きな制約条件となっている。そこで、従来のメモリ(DRAM)の他に、速度などの性能は劣るものの消費電力や容量あたりの単価が低い不揮発性メモリを利用し高い電力効率を達成した。大規模なグラフ処理における電力効率性のベンチマークである Green Graph500 ランキングの Big Data カテゴリにおいて 2014 年の 6 月に世界 3 位となる成果を収めた。

3. Ikki Fujiwara, Michihiro Koibuchi. Tomoya Ozaki, Hiroki Matsutani, Henri Casanova, “Augmenting Low-latency HPC Network with Free-space Optical Links”, The 21st IEEE International Symposium on High Performance Computer Architecture (HPCA 2015), pp.390-401, Feb. 2015 (DOI: 10.1109/HPCA.2015.7056049)、要約が IEEE Spectrum (Giving supercomputers a second wind, Boyd, J., June 2015, p.20) で紹介された。

概要:

スーパーコンピュータ、データセンター、EBD マシンなどの大規模計算機システムでは、導入時に設定したネットワークトポロジを更新することが通常困難である。一方で、EBD アプリケーションの通信パターンには多様性があり、そのプロセスとネットワークトポロジのマッピングが難しい場合が生じる。本研究では、これらのアプリケーションを EBD インターコネクトが効率良くサポートするために、光無線技術によるネットワークトポロジの再構成技術を提案、探求した。

＜科学技術イノベーションに大きく寄与する成果＞

1. Kento Sato, Kathryn Mohror, Adam Moody, Todd Gamblin, Bronis R de. Supinski, Naoya Maruyama and Satoshi Matsuoka, ″A User-level InfiniBand-based File System and Checkpoint Strategy for Burst Buffers”, 2014 14th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (CCGrid), Chicago, USA, May 2014, pp.21-30 (DOI: 10.1109/CCGrid.2014.24)

概要:

エクストリームビッグデータ処理のための、ストレージ基盤として有力候補である、EBD IO の I/O 性能最適化や信頼性を定量的に評価した論文で、 採択率 19.1%と大変選別の厳しい国際会議 CCGrid2014 に論文が採択され、さらに Best Paper に選ばれた。

2. 超高速メタゲノム解析技術の開発

概要:

ヒト体内や環境内の微生物を網羅的に解析するメタゲノム解析では、未発見の微生物が含有される率が高いために、大量のデータに対してきわめて高精度の相同性検索を実施する必要がある。開発した GHOST-MP は、既存の標準的な並列ソフトウェア mpiBLAST よりも、同ノード数で 80 倍以上高速であり、数万ノードまでスケールする。当研究内で、東京歯科大学との共同で、ヒト歯周病の臨床研究をすでに実施中である。

3. Nguyen T. Truong, Ikki Fujiwara, ¥underline{Michihiro Koibuchi}, Khanh-Van Nguyen, Distributed Shortcut Networks: Low-latency Low-degree Non-random Topologies Targeting the Diameter & Cable Length Trade-off, IEEE Transactions on Parallel and Distributed Systems, pp.989-1001, Vol. 28, Issue. 4, 2017, 10.1109/TPDS.2016.2613043

概要:

EBD インターコネクト向き低遅延ネットワークトポロジを提案した。スモールワールド性を用いた

ショートカットリンクを巧みに用いることが特徴である。マシンルームでの配線法、高い耐故障性、高い拡張性を有することから、実用性が高いことを示した。

＜代表的な論文＞

- Suzuki S, Kakuta M, Ishida T, Akiyama Y. Faster sequence homology searches by clustering subsequences, PLoS ONE, 31(8): 1183-1190, 2015.

- Takemasa Miyoshi, Guo-Yuan Lien, Shinsuke Satoh, Tomoo Ushio, Kotaro Bessho, Hirofumi Tomita, Seiya Nishizawa, Ryuji Yoshida, Sachiho A. Adachi, Jianwei Liao, Balazs Gerofi, Yutaka Ishikawa, Masaru Kunii, Juan Ruiz, Yasumitsu Maejima, Shigenori Otsuka, Michiko Otsuka, Kozo Okamoto, and Hiromu Seko, “Big Data Assimilation” Toward Post-Petascale Severe Weather Prediction: An Overview and Progress,″ in Proceedings of the IEEE, vol. 104, no. 11, pp. 2155-2179, 2016

- Kakuta M, Suzuki S, Izawa K, Ishida T, Akiyama Y. A massively parallel sequence similarity search for metagenomic sequencing data, International Journal of Molecular Sciences, 18(10): 2124, 2017.

§２ 研究実施体制

(1)研究チームの体制について

① 「松岡」グループ
研究代表者:松岡 聡(東京工業大学情報理工学院 特任教授)
研究項目
・EBD システムアーキテクチャの設計及びシステムソフトウェアの研究開発
・EBD システムアーキテクチャの基本設計
・EBD システムソフトウェアの基本設計
・EBD 基本ソフトウェアの OSS 化
・アプリケーションによる評価・高度化
・社会データの取得
・分析プログラムの開発
・モデルによる解析
・予測プログラムの開発
・プログラムの性能最適化
・深層学習の高速化

② 「建部」グループ
主たる共同研究者:建部 修見(筑波大学計算科学研究センター 教授)
研究項目
・EBD 分散オブジェクトストアの研究
・分散オブジェクトストアの設計
・プロトタイプ実装・性能評価
・アプリケーションによる評価・高度化

③ 「鯉渕」グループ
主たる共同研究者:鯉渕 道紘(情報・システム研究機構国立情報学研究所アーキテクチャ科学研

究系　准教授）
研究項目
・EBD インターコネクトの研究開発
・低遅延トポロジとルーティング
・ストレージへの直接通信機構

④ 「秋山」グループ
主たる共同研究者：秋山　泰（東京工業大学情報理工学院　教授）
研究項目
・EBD データ処理 API の開発
・大規模ゲノム解析等での実応用評価

⑤ 「三好」グループ
主たる共同研究者：三好　建正（理化学研究所計算科学研究センターデータ同化研究チーム　チームリーダー）
研究項目
・ゲリラ豪雨予測を可能にする次世代ビッグデータ同化アプリケーションの EBD コデザイン
・フェールセーフの EBD ワークフローの開発
・Geographical Search アルゴリズムの最適化
・EBD のプラットフォームの設計・開発のコデザイン

（2）国内外の研究者や産業界等との連携によるネットワーク形成の状況について

松岡グループは、TSUBAME2 の後継のスーパーコンピュータであり、世界初の EBD 向けのスーパーコンピュータとなる TSUBAME3 のプロトタイプとして、日本電気株式会社、米国 NVIDIA 社、米国 Green Revolution Cooling 社、米国 Supermicro 社，米国 Intel 社、Mellanox 社と連携して、TSUBAME-KFC の開発を進めた。また、TSUBAME3 に関しても、前記各社に加え米国 Hewlett-Packard Enterprise 社などとも連携してシステム開発を行った。例えば、Data Direct Networks(DDN)社とは、現在デファクトで利用されているオープンソースの並列ファイルシステムである Lustre の次世代バージョンやスーパーコンピュータの I/O を高速化するバーストバッファに関して、性能評価を進めた。Amazon AWS ともスパコン-クラウド間のデータ共有・連携に関して、Cloud Burst Buffer の開発を進めている。また、九州大学藤澤克樹教授らの CREST プロジェクトのグループ及び、理化学研究所計算科学研究機構と Graph500 に関する共同研究、ローレンスリバモア国立研究所と不揮発性メモリを考慮したグラフ処理、東京工業大学篠田浩一教授らの CREST プロジェクトのグループやデンソー社、富士通研究所と深層学習に関する学習の高速化の研究をそれぞれ進めている。加えて、英国ダラム大学や米国オークリッジ国立研究所と、並列分散エージェントシミュレーションの高速化に関して、共同研究を実施している。

建部グループは Argonne National Laboratory の Rob Ross 氏らと共同でオブジェクトストアのシュミレーションスタディを進めた。また、富士通研究所と大量データ管理のためのストレージシステムソフトウェアについての共同研究を進めた。

鯉渕グループは Henri Casanova ハワイ大教授と密に連携して EBD インターコネクトの研究を行っている。彼はマイクロソフトアカデミックサーチの Citations が 4,236 など突出したインパクトをスケジューリング、並列科学計算分野に残している。
(http://academic.research.microsoft.com/Author/779650/)
本異分野連携により、不規則性を用いたネットワーク設計に関する様々な成果が生じており、今後もこの枠組みを継続する予定である。

秋山グループでは、当事業で開発した GHOST ソフトウェアの応用として、基礎医学分野では東京大学医科学研究所の宮野悟教授、植松智特任教授らとマウス腸内細菌叢のメタゲノム解析の共同研究を実施、臨床医歯学分野では東京歯科大学の石原和幸教授らと歯周病のメタゲノム解析で共同研究を実施。さらに、農研機構と微生物電池の分野で、さらに国立医薬品食品衛生研究所と住宅環境におけるハウスダスト中の微生物解析の分野でメタゲノム解析の共同研究を実施。またミトコンドリア病と呼ばれる難病に関する NPO 法人からも疾病モデルマウスの腸内解析に関する委託を受けて共同研究を実施している。

三好グループでは、本研究と密接に連携しながら並行して進めている CREST 研究課題「「ビッグデータ同化」の技術革新の創出によるゲリラ豪雨予測の実証」（研究代表者：三好建正）ではにおいて、フェーズドアレイ気象レーダーを開発、設置、運用している情報通信研究機構及び大阪大学、気象衛星ひまわり 8 号を打ち上げ、運用している気象衛星センター、気象庁のメソ数値天気予報モデル NHM を使ったデータ同化研究に取り組んでいる気象研究所、数値天気予報モデル SCALE を開発している理化学研究所計算科学研究機構複合系気候科学研究チーム、ジョブ間通信機構の最適化に取り組んでいる理化学研究所計算科学研究機構システムソフトウェア研究開発チームと共同して研究を進めている。また、宇宙航空研究開発機構降水観測ミッションによる降水観測データの利用に関して、宇宙航空研究開発機構、東京大学大気海洋研究所と共同して研究を進めている。このほか、ポスト「京」重点課題 では、上記の連携に加えて、海洋研究開発機構、東京工業大学、京都大学、東北大学、名古屋大学、気象庁、琉球大学、神戸大学と連携、協力しており、さらに連携の幅を広げており、全国に渡ったネットワーク形成が進んでいる。海外に関しては、データ同化アルゴリズムに関してメリーランド大学と MOU を締結して密接に連携して取り組んでいるほか、アルゼンチンのブエノスアイレス大学、アルゼンチン気象局とは共同研究が進んでおりいる。また、ドイツ気象局、ミュンヘン大学とは国際会合を共催するし、国際的なネットワークの形成にも力を入れている。