公開

研 究 終 了 報 告 書

「耐故障並列計算と高速ロシー結合網の協調」

研究期間： 2019 年 10 月～2023 年 3 月
研 究 者： 鯉渕 道紘

1. 研究のねらい

将来のクラウドハイエンドデータセンターの設計は、フォトニクスや短距離無線などの新通信デバイス統合による相互結合網の複雑化により制限を受けることが予想される。その制限のもと、現実的なコストで従来通りのデータロスが生じない、完璧な通信ハードウェアを設計することは困難を極める。そこで、本研究では、データの欠損を多少許容することで、高速処理を実現するロシー通信技術を探求する。ビッグデータ、機械学習、NP 困難問題の近似解を求める並列計算アプリケーションなどを対象に、ロシー通信への十分な耐故障性を有するように Algorithm Based Fault Tolerance (ABFT)技術を導入するフレームワークを実現する。ロシー通信技術と耐故障性を有するアプリケーションフレームワークは相互に支えあう技術であり、同時に成立することが重要である。そこで、両者を統合することで、高速並列コンピューティングの協調設計技術を確立する。

最終的に、クラウドハイエンドデータセンターにおいて通信エラーが生じても、各並列計算アプリケーションの高速実行に成功する並列コンピューティング基盤を明らかにする。そして、実機における動作検証と定量的な性能評価を通して、本設計技術の有効性を示す。

なお、本研究で対象とするエラーとは、恒久的なハードエラーではなく、ビット化けなどの一時的なエラーであるソフトエラーを指す。従来は「通信エラーが生じない」前提で、プログラマが並列計算アプリケーションを設計してきた。しかし、本研究によって検討されたように、ロシー通信環境下では、プログラマは通信エラーの発生を想定し、要所において簡易な検算をするなどして、耐故障性を有するように並列計算アプリケーションを開発することとなる。この点により、アプリケーション開発のコーディングコストは増大することになる。一方、通信ハードウェアは高速化に特化した開発を行うことで、コスト・半導体の微細化・ネットワークの複雑化の縛りから開放され、高性能通信技術の飛躍的な発展が可能になる。

2. 研究成果

(1) 概要

本研究では、耐故障性に関する機能を最小限に留める高速相互結合網と、アルゴリズムレベルで高い耐故障性を有する並列計算アプリケーションという二つの開発を進めた。そして、両者を協調させる設計技術を用いた高性能クラウド並列計算基盤の実現を目指した。

相互結合網に関して通信データの重要度に応じて非可逆データ圧縮技術を導入し、転送効率と計算結果の品質のトレードオフを探究した[2,3]。結合網の通信遅延要求は 1 マイクロ秒以内と極めて小さいため、圧縮と解凍の処理時間を百ナノ秒オーダーに抑える設計が必要となる。そこで、特定のビット列のパターンに一致する場合にのみ圧縮する FPC (Frequent Pattern Compression)アルゴリズムを応用する方式と、1 次元配列データの規則性を利用する

方式、浮動小数点数の仮数部の一部を切り捨てる方式を併用して通信データを低遅延で圧縮する技術を開発した[2]。さらに、出発地と目的地において、データ圧縮および解凍処理をパケット転送のパイプライン処理と並列実行する方式を開発した[3]。シミュレーションによる評価結果より、提案方式は、典型的な並列計算ベンチマークの解の要求精度を満たした上で、実行時間の大幅な向上を達成する場合があることを示した。

並列計算アプリケーションに対しては、Algorithm-Based Fault Tolerance (ABFT)を導入する技術を開発した。すべての通信時に発生するソフトエラーを ABFT により訂正することは、コストの面で現実的ではない。そこで、2 箇所までのエラーを ABFT 計算により訂正し、それ以上の箇所のエラーが同時に生じた場合、都度 end-to-end にて再送することで、並列行列計算の実行時間が大幅に短縮できる場合があることを示した。ABFT 支援については、通信データの非可逆圧縮の改良を行い、再送頻度を抑制するようにパケット全体の圧縮率を調整する技術に発展させた。

最後に、大規模並列計算システムを想定したシミュレーション評価と、小規模 PC クラスタにおける実機評価を行った。具体的には、最先端のオンボード・シリコンフォトニクストランシーバーを搭載したカスタム FPGA ボード間通信において、光アンプの出力を抑えることで生じるビット誤り率を測定した。その結果、物理的にはランダムエラーが発生しているにも関わらず、64B67B ラインコードにより、アプリケーション層では一部 64 ビットのバーストエラーとして処理されていると考えられる事象が確認された。そのため、ラインコード内にバーストエラー防止用の簡便なエラー検出訂正符号を持つ方式を開発し、近似処理の統計的手法が応用できるように検討した。また、独自通信命令の追加により、FPGA クラスタ上でシリコンフォトニクストランシーバー経由のロシー通信を用いる並列計算を効率的に開発する環境を実現した。

(2) **詳細**

課題 A:ロシー結合網の設計

並列計算におけるプロセス間通信データを非可逆圧縮することで、相互結合網のスループットが劇的に向上できる場合があることを示した。相互結合網の通信遅延要求は 1μ 秒以内と極めて小さいため、圧縮と解凍の処理時間を百ナノ秒オーダーに抑えることが必要となる。そのため、画像処理やストレージの圧縮で用いられている高効率ではあるが、処理遅延が大きい複雑な圧縮アルゴリズムを相互結合網に応用することは現実的でない。そこで、特定のビット列のパターンに一致する場合にのみ圧縮する FPC (Frequent Pattern Compression)アルゴリズムを応用する方式、1 次元配列データの規則性を利用する方式、浮動小数点数の仮数部の一部を切り捨てる方式を併用して通信データを低遅延で圧縮する技術を開発した[2]。さらに、出発地と目的地において、データ圧縮および解凍処理をパケット転送のパイプライン処理と並列実行する方式を開発した[3]。

非可逆データ圧縮の鍵となる点は、浮動小数点数の扱いである。反復並列計算や NP 困難問題の近似解並列計算では、プロセス間で浮動小数点数の配列データの交換を頻繁に行う。本研究では、これらが近似計算であることを利用して、この浮動小数点数を Quality-of-Results (QoR)を一定に保ちつつ数値を丸める近似圧縮方式を開発した[2]。

まず、解析結果から、本研究で開発した非可逆圧縮を行うことで、最大 2 倍近い圧縮率を達

成できた。これは gzip などの既知の圧縮技術と比べると若干悪いが、浮動小数点数の圧縮としては十分に高いと考えられる。

次に、64 台の計算ノード間においてランダムな通信パターンが生じた場合の相互結合網のシミュレーション結果から、通信データを2倍に圧縮できた場合、計算ノードにおいて圧縮解凍遅延が生じるが、実効メッセージスループットが 1.5 倍という大きな効果を得ることができた。その他の知見としては、通信データを10倍に圧縮できた場合でも、スループットは10倍にはならず、3.5 倍程度にとどまる。これは、各スイッチにおけるパケット処理の一部(例:ルーティング計算)が、フリット単位ではなく、パケット単位で行われるためである。

この他に、結合網のビット誤り率に応じて誤り検出訂正の強度を変更し、圧縮技術と協調動作する技術についても開発を行った。

最後に FPGA 間の高速ロシー結合網の設計を行った。最先端のオンボード・シリコンフォトニクストランシーバーを搭載した FPGA ボード間通信に関してハードウェア量を抑えるために、イーサネットや InfiniBand などの標準的なネットワークスタックに頼らずに、生データをそのまま転送する方式(図 1(d))を提案した[1]。直接 FPGA 間を相互接続するためには、現状では図 1(b)と図 1(c)の 2 つの方式がある。図 1(b)は、イーサネットや InfiniBand のプロトコルに従って、パケットを生成するが、ルーティング情報やフロー情報がビット化けした場合に通信が停止する、あるいは大きな遅延が生じる点で本研究課題において問題である。図 1(c)は、独自にネットワークスイッチを FPGA 内に内蔵することで、高速通信を達成可能であるが、ビット化けに対する耐性は一般的なパケット通信と同様である。そこで、我々は図 1(d)に示した通り、FPGA 間完全結合方式を提案し、ルーティング情報とデータを分けて管理し、転送することでビット化けが生じた場合でも、データ転送を実現した。

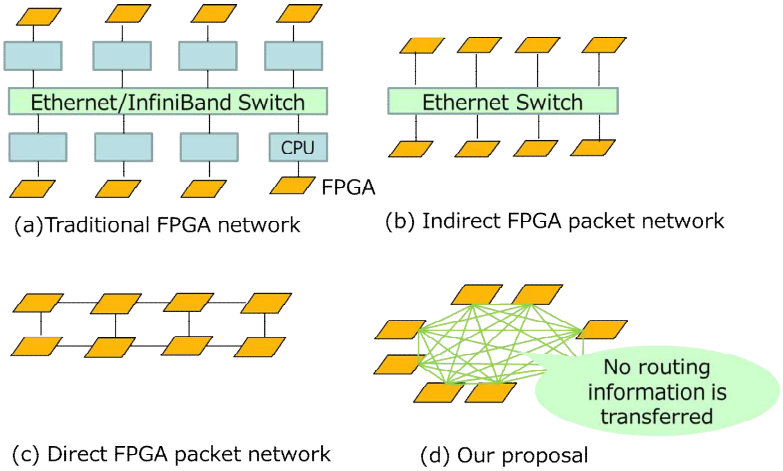

図 1:提案 FPGA 間通信方式(d)

具体的には、Start of Packet (SoP)や End of Packet (EoP)などのビット化けが許容されない制御情報を最小限のデータ量に抑えることで、通信のビット化けがシステムに与える影響を最小限にし、並列アプリケーションが異常終了する可能性を低減した。この方式で、8 台の FPGA 間通信に本ロシー結合網を実働することに成功した。ビット誤り率が測定可能である範囲(多くの場合 10^{-6} 以下)において、アプリケーションレベルでのビット誤り訂正が軽量であるため、ロシー通信でも FPGA 間の通信は性能劣化しないことが期待できる。

本ロシー結合網では、通信データにルーティング情報を含まない。そのため、ホップ毎に通信セットアップ遅延が生じるため、直径が小さいネットワークトポロジを採用することが性能面で有利である。その点で、Graph Golf コンペティション(https://research.nii.ac.jp/graphgolf/ 研究代表者の鯉渕は主催のひとり)で報告された直径の小さいグラフを FPGA 間ネットワークのトポロジに採用し、2 ホップにて本 FPGA ボード 256 台を相互接続できることを示した。

以上の通り、課題 A では、非可逆圧縮、ビット誤り率に応じてエラー検出訂正処理を最適化する技術を提案し、アプリケーション毎の有効性を示し、実機におけるロシー結合網の設計を行った。

課題 B:ネットワークシミュレータの開発

SimGrid イベントドリブンネットワークシミュレータを ABFT 並列計算の実行およびロシー結合網の挙動が再現できるように拡張した。SimGrid では、並列計算システムにおいてメッセージパッシングインタフェース(MPI)で記述された並列プログラムの実行時間を高精度で見積ることができる。本開発では、MPI 通信で生じるデータの任意の位置にビット化けを発生させる機能を用いて、本研究において提案しているロシー結合網の評価を行うことができるようになった。さらに、SimGrid のソースルーティング機能を使うことで、ロシー結合網上でのメッセージコンバイン技術や、集合通信のスケジューリング法の評価が可能となった[1]。また、実機での検証、通信イベントのトレース取得のため、MPI 通信を二重に実行し、ビットエラーを記録する関数を定義した。その評価を行った結果、SimGrid イベントドリブンネットワークシミュレータにおいて並列計算アプリケーション上でデータ圧縮した場合と、送信元ノードのネットワークインタフェース上でハードウェアにより高速にデータ圧縮した場合の処理遅延の差が、実機で動作させた場合と比べて相対的に小さく表示される傾向にあることが分かった。

以上の通り、課題 B の開発により、任意の並列計算システムにおける本ロシー結合網と耐故障性を有する並列計算アプリケーションの協調技術について定量的な評価を行うことができるようになった[1,2]。

課題 C: Algorithm Based Fault Tolerance(ABFT)技術

ABFT 支援について,通信データの非可逆圧縮方式の改良を行い、再送頻度を抑制するようにパケット全体の圧縮率を調整する技術に発展させた[2]。この技術により、アプリケーション実行結果の要求される質(Quality of Results:QoR)とロシー結合網のビット誤り率に応じて、アプリケーションの実行速度を向上させることができるようになった。

さらに、独自のロシーデータ転送用の命令を追加することにより、MPI プログラムの開発を容易に行うことが可能となった。具体的には、8 台の計算ノードで構成された PC クラスタにおいて、正確な通信を行う MPI 関数と、通信データのビット化けを許容する MPI 関数を point-to-point 通信、集合通信について提供した。そして、ビット化けを許容する場合のみ、実際のオンボードのシリコンフォトニクスのトランシーバーを経由して通信を行う。そして、この PC クラスタ上において、巡回セールスマン問題の近似解を求める並列計算において動作を確認した。

以上、課題 C では、並列アプリケーションからの ABFT 利用支援技術を開発した。

3. 今後の展開

(1)長期: 本領域アドバイザーの成瀬 誠先生が研究代表者である文部科学省・科研費学術

変革 A「光の極限性能を用いたフォトニックコンピューティングの創成」(22H05197、FY2022-FY2026)において、計画班長として、耐故障性を有する並列計算と光通信技術の協調という新しいアプローチを発展させることで、光コンピューティングの基礎研究に貢献する予定である。

(2)短期: 2030 年頃のスーパーコンピュータ(スパコン)のフィーシビリティスタディへの適用を検討している。文部科学省・次世代計算基盤に係る調査研究(研究代表者:牧野 淳一郎,2022-2023)において、研究分担者として参画し、スパコン内アクセラレータボード間の相互結合網のアーキテクチャへ本研究成果の応用を検討している。

以上、長期、短期のビジョンを明確にした上で、本成果を展開する予定である。

4. 自己評価

研究目的の達成状況

本研究では、通信ハードウェアの耐故障性に関する設計を最小限に留め、耐故障性を有する並列計算アプリケーションとの協調というコデザインにより、新通信デバイスの異種混合を実現する要となる高性能ロシー通信アーキテクチャを創出することを目指した。具体的には、「プログラムの耐故障性（計算科学）と、新通信デバイスの高性能化（計算機科学）の両者の協調設計技術」を重要視した。このビジョンに基づいて、課題 A,B,C に沿った研究を行うことができ、充実感のある活動を行うことができた。特に、技術研究組合光電子融合基盤技術研究所(PETRA)が開発した 800Gbps FPGA ノード帯域を有するオンボードシリコンフォトニクストランシーバーを対象として、本コンピューティング技術を展開する貴重な経験を得ることができた。さらに、並列分散システム分野のトップ国際会議の 1 つである IPDPS2022(36th IEEE International Parallel &Distributed Processing Symposium)の regular paper や、そのワークショップ APDCM2021(Advances in Parallel and Distributed Computational Models)の基調講演において本成果を発表する機会に恵まれた。

研究の進め方（研究実施体制及び研究費執行状況）

FPGA 間通信を本研究の応用対象に含める課題を 2021 年 11 月に追加した。PETRA が開発したオンボードシリコンフォトニクストランシーバーを対象にして、本コンピューティング技術を展開することができた。その評価結果から、スケーラビリティに関するシステムアーキテクチャの課題が見え、さらに、良いコンピューティング技術の開発[1]という、光デバイスとシステム研究の進化の好循環を体感することができた。

研究成果の科学技術及び社会・経済への波及効果

本研究で提示する革新的コンピューティングでは、人間が主体的にプログラムを通してコンピューティングの振舞である計算をオンラインでチェックする側面を持つ。Society 5.0 では無数の edge mobile computing (EMC)と、運用主体が見えづらいクラウドコンピューティング基盤に囲まれる世界となる。そのため、プログラム自体がコンピュータ、特にハードウェアの挙動を過信せずに主体性をもって検証しながら進めることが安定、安心をもたらす。つまり、本研究が提唱する ABFT の積極的な利用は、Society 5.0 高度 IT 社会のコンピューティングのあり方の 1 つを示すものである。

5. **主な研究成果リスト**

(1)代表的な論文(原著論文)発表

研究期間累積件数:5件

1. Kien Trung Pham, Truong Thao Nguyen, Hiroshi Yamaguchi, Yutaka Urino, Michihiro Koibuchi, ``Scalable Low-Latency Inter-FPGA Networks'', The 28th IEEE International Parallel & Distributed Processing Symposium (IPDPS) , pp.234-245, 2022 (査読有)
概要 PC クラスタのネットワークトポロジ、フロー制御、間接ルーティング、集合通信、WDM ケーブリング法を提案した。提案方式は Dragonfly ネットワークトポロジを採用した典型的なネットワーク構成と比べて集合通信が 7 倍速くなることが分かった。さらに、提案方式では転送データにビット化けが生じた場合でも目的地まで配送され、適切にアプリケーションにデータが受け渡しできる。
2. Yao Hu, Michihiro Koibuchi, "Accelerating MPI Communication using Floating-point Compression on Lossy Interconnection Networks", The 46th IEEE Conference on Local Computer Networks (LCN) pp.355-358, 2021(査読有) *概要* 非可逆データ圧縮により、仮想的に通信帯域を大きく見せることで、アプリケーションの実行時間の短縮を達成した。さらに、エラー検出符号をアプリケーションにおいて実装することで、ロシー結合網を用いて正しい並列計算の結果を得るフレームワークを提案した。イベントドリブンネットワークシミュレーションの結果により挙動を確認した。
3. Naoya Niwa, Yoshiya Shikama, Hideharu Amano, Michihiro Koibuchi, "A Case for Low-Latency Network-on-Chip using Compression Routers", 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), pp. 134-142, 2021 (査読有) *概要* 送信元ルータにおいて、パケット転送のパイプライン処理と、転送データの非可逆圧縮処理のパイプライン処理を同時に行うことで、圧縮処理遅延を隠蔽しつつ、転送データ量を削減する方式を提案した。巡回セールスマン問題、共役勾配法などの並列計算において、プロセス間通信時間の短縮により高速化を達成し、かつ、十分な解の計算精度を得ることができた。

(2)特許出願

該当無し

(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)

1 鯉渕 道紘、日本学士院, 学術奨励賞、.2021 年 2 月,

「並列計算機システムの相互結合網へのランダム性導入に関する先駆的研究」

2. 鯉渕 道紘、日本学術振興会, 日本学術振興会賞、.2021 年 2 月,

「並列計算機システムの相互結合網へのランダム性導入に関する先駆的研究」

3. 鯉渕 道紘, ``Approximate Computing と関連する通信技術" 電子情報技術産業協会(JEITA)、「非ノイマン型情報処理へ向けたデバイス技術分科会」, 2020 年 9 月 9 日(オンライン)、招待講演

4. 鯉渕 道紘,ネットワーク視点からの取り組み「不完璧なスーパーコンピュータ」, 825 回マルチメディア推進フォーラム「ポスト・ムーアの切り札： Approximate Computing」, 2020 年 7 月 15 日(オンライン), 招待講演

5. Michihiro Koibuchi, “Graph Golf Competition Seeking for Small-Diameter Graphs”, 23th Workshop on Advances in Parallel and Distributed Computational Models (APDCM), May 17, 2021, 基調講演