様　式　Ｆ－１９－２

科学研究費助成事業　　研究成果報告書

平成 ３０ 年　 ６ 月 １５ 日現在

機関番号：６２６１５
研究種目：国際共同研究加速基金（国際共同研究強化）
研究期間：2017～2017
課題番号：１６ＫＫ０００９

研究課題名（和文）Approximateネットワークと並列アルゴリズムの協調（国際共同研究強化）

研究課題名（英文）Co-operation Between Approximate Networks and Parallel Algorithms(Fostering Joint International Research)

研究代表者

鯉渕　道紘（Koibuchi, Michihiro）

国立情報学研究所・アーキテクチャ科学研究系・准教授

研究者番号：４０４１３９２６

交付決定額（研究期間全体）：（直接経費）　　6,200,000円
渡航期間：　　７ヶ月

研究成果の概要（和文）：従来の大規模並列コンピュータは、古典的な科学シミュレーションが要求する高精度な計算をサポートしてきた。しかし、近年、データ量が多い一方、計算精度を従来ほど高く要求しない類のビッグデータ処理や並列処理計算が劇的に増加している。本研究では、これらの計算処理系がビット化けによる誤りを訂正せずに放置するApproximate Computing向けに設計されたネットワークにおいて正しく動作し、実行結果の大勢に影響しないようにするための並列アルゴリズムと、その支援技術を開発した。

研究成果の概要（英文）：Existing large-scale parallel computers have supported high accurate computation required by traditional scientific simulation. However, recently, big-data low-accuracy data processing and parallel computation would received attention as emerging applications. They do not require high-precise computation and communication. In this context, we extend approximate computing networks that leave bit errors. On approximate computing networks with special supports, we have developed parallel algorithms in order to obtain the acceptable quality of results.

研究分野：情報学

キーワード：相互結合網　Approximate Computing　計算機システム　フォトニックネットワーク　ハイパフォーマンスコンピューティング

１．研究開始当初の背景

コンピュータは数値を近似して表現(例:数 0.110 進は 0.0001100[1100]2 進で丸め)し、複数のプロセッサがハードウェアレベルで非決定的な順序により共有変数にアクセスするため、計算結果の潜在的誤差を完全に除去することが難しい。この点を逆手にとり、計算の許容誤差を若干大きくすることで計算の精度を落とし消費電力を削減、スループットを向上させる Approximate Computing に関して様々な回路設計、プロセッサ、プログラミング言語に関する先行研究がある。ディープラーニング系の処理を行う演算ビット数の削減などが著名だが、通信技術に関しては、研究代表者の鯉渕らの変調時に工夫をほどこす技術などのごく一部を除いて存在しない。これは、データセンター、スーパーコンピュータ(以後、スパコンと呼ぶ)のネットワークではソフトエラー(ビット化け)について高い信頼性を要求する標準規格があり、厳密に守られているため Approximate Computing の考え方に基づく関連研究が進んでいないためと考えられる。しかし、ソフトエラー対策であるビット化け検出、訂正の処理に係る電力、通信オーバヘッドはリンク帯域が大きくなるにつれ劇的に悪化する。例えば、400Gbps イーサネットに関する多くの規格では巨大なバッファリングを想定したリードソロモン符号による Forward Error Correction(FEC)の利用が見込まれている。この FEC 処理で約 50―100 ナノ秒の遅延が見込まれる。

同様に計算ノードのネットワークインタフェース、チップ内通信でも信頼性に関する処理が必要となる。つまり、今後、リンク帯域の向上(100-400Gbps 級)が見込まれるが、現状と同じ通信遅延、消費電力でのデータ転送すら実現することが難しくなる。

２．研究の目的

従来の大規模並列コンピュータは、古典的な科学シミュレーションが要求する高精度な計算をサポートしてきた。しかし、近年、データ量が多い一方、計算精度を従来ほど高く要求しない類のビッグデータ処理や並列処理計算が劇的に増加している。

本研究では、これらの計算処理系がビット化けによる誤りを訂正せずに放置する Approximate Computing 向けに設計されたネットワークにおいて正しく動作し、実行結果の大勢に影響しないようにするための並列アルゴリズムとネットワークの要素技術を開発する。

３．研究の方法

研究代表者の鯉渕が Approximate ネットワーク設計を担当し、海外共同研究者の Henri Casanova が計算機上で実行する並列アルゴリズムの開発を担当するコデザインを遂行する。

本研究は基課題[1](概要を図 1 に示す)を発展させることで遂行する。基課題では、Approximate Computing ネットワークの実現性を追求することに焦点を置いているため、並列アルゴリズム、耐故障技術の専門家が不在であった。そのため本国際共同研究では、Casanova がその上で完動する汎用の並列アルゴリズムを開発し、研究代表者がその支援を行うネットワーク設計を行う。

本ネットワーク設計において、基課題で開発している Approximate Computing ネットワークを拡張し、ランタイムでリンク帯域を可変とするための仕組みを導入する。この仕組みにより多くの並列アルゴリズムを効率的に実装可能とする。

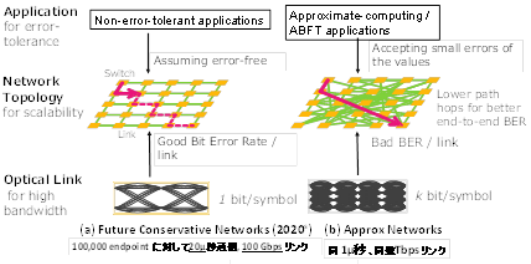

図1: 本研究の概要

(1) Approximate ネットワーク(研究代表者の鯉渕担当)

パケットがリンクを経由する毎にビット化けが生じる恐れがある。そのため、パケットが目的地に到達するまでに経由するホップ数を最小化することが重要である。この点でネットワークトポロジの構成について探求する。また、海外共同研究者の Casanova が開発する並列アプリケーションのトラヒックパターンを用いた場合にパケットの平均ホップ数が最小となるようにネットワークトポロジとジョブマッピングについて検討を行う。評価方法として、グラフ解析によるホップ数の計算および、フリットレベルネットワークシミュレーションによるスループットとレイテンシの解析により行う。ビットエラー率が与える数値誤差については数学的解析により検討を行う。

(2) Approximate 並列アプリケーション(海外共同研究者の Casanova 担当)

アプリケーションデータの価値と伝送の確実性を比例させるスキームを実現する。フーリエ変換などの実装アルゴリズムに関して、ビット化けの許容度について理論的に提示し、Quality of Results(QoR)の要求が厳しいアプリケーションに対しては、ランタイムで計算ノードが検算を行うことでエラーから回復する技術の適用を検討する。最終的には Approximate ネットワークに適したアプリケーションの特性についての知見を提示する。

連携の詳細は以下である。研究代表者の鯉渕は、全体の研究総括と Approximate ネットワークの設計を担当する。そして設計技術の

検証のため、海外共同研究者の Casanova が開発しているイベントドリブンでプログラムの挙動を解析できるスパコン・シミュレータ SimGrid を用いて評価を行う。そして、Casanova が設計するビットエラーを許容する並列アプリケーションとアルゴリズムが完動できるように、ネットワークの可変ビットエラーレートでの符号最適化を検討する。

４．研究成果

(1)Approximate ネットワーク

Approximate ネットワークの考え型と基本設計は先行研究[2][3]で行われており本研究では、これらの技術の利用を前提とする。

エラー検出訂正を行わない（原則再送を行わない）ネットワークでは、ケーブルを単方向リンクで構成する通信方式が有利であることが分かった。具体的には、単方向リンクネットワークは、従来手法である双方向リンクでネットワークを構成した場合と比べて、経由するスイッチ数を削減、すなわち、ホップ数を削減できるため、エラーの発生確率が小さくなり、有益なことがわかった。

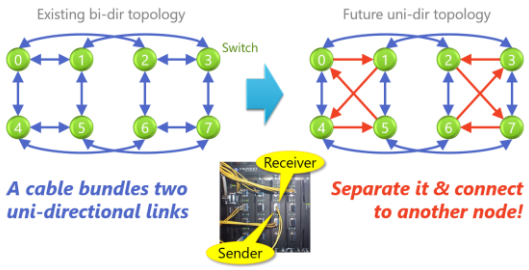

図 2:単方向リンクネットワークの例

幸い、現在、商用の光ケーブルの多くは、1 つのケーブルが 2 つのリンク(sender/receiver)で構成されているため、図 2 のようにバンドルを外すことで利用することができる。

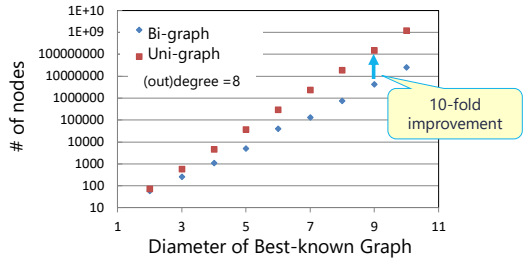

図 3:次数 8 の最大グラフの頂点数

図 3 は、与えられた次数と直径に対して構成できる既知のグラフの最大頂点数を示している。これは Degree Diameter 問題として知られたグラフ理論分野の著名な問題の現状の最良解をプロットしたものといえる。縦軸の数値の大きい方が良いことを示している。単方向グラフ(uni-graph)の場合はほぼ理論最適な構成法が任意の次数、直径に対してすでに確立されている。一方、無方向グラフ(Bi-graph)の場合は理論最適なグラフがあまり発見されていない。そのため、図 3 に示した通り、同一の次数、直径に対して単方向グラフの方がホップ数を削減できる。この点で Approximate ネットワークにおいて単方向リンクで構成されたネットワークを用いるべきという結論にいたった。

次に、リンクレベルのフロー制御について検討を行った。並列コンピュータの相互結合網ではパケットロスは原則遅延削減のために許容されない。そのためロスレスネットワークを達成するために、Stop/Go シグナルを受信ノードから送信ノードへ折り返すことで経由スイッチでのバッファオーバフローを抑えている。しかし、単純な単方向リンクでネットワークを構成した場合、このシグナル伝送ができない。

Approximate ネットワークでは、ビット化けによる再送は行わないので、NACK/ACK メッセージの送受信は不要であるが、バッファオーバフローの問題には対応する必要がある。

そこで、我々は単方向リンクで構成された Approximate ネットワークにおいて HotPotato ルーティングの利用を提唱した。HotPotato ルーティングでは、各スイッチにおいて入力ポート数と出力ポート数が同一であることを前提とする。実際に現状のほぼすべての商用スイッチはこの条件を満たしている。そして、すべての入力ポートにあるパケットは他のパケットにブロックされることなく、出力ポートに転送される。ただし、そのために、異なる入力ポートにあるパケットが同一の出力ポートを選択した場合、1 つのパケット以外は、最短経路でなくとも他の出力ポートを選択する必要が生じる可能性がある。よって、HotPotato ルーティングはネットワークの負荷が大きくなるにしたがって、パケットの平均ホップ数が大きくなるデメリットがある。本研究では、その影響を検証するためにフリットレベルシミュレーションにより評価を行った。その結果、64 スイッチと 256 スイッチの単方向ネットワークにおける平均ホップ数はネットワーク負荷が高くなっても微増するに留まることが分かった。

さらに、トラヒックパターンに応じて最適なネットワーク構成を取る方法については、空間的、さらには時間的な局所性を用いて設計する既知の方法[4]をそのまま利用できることが分かった。その結果、NAS 並列ベンチマークの多くのアプリケーションではネットワーク内のパケットの衝突の可能性がほとんどないような設計を小さい次数のスイッチで構成できることが分かった。

(2) Approximate 並列アプリケーション

一部の科学技術計算のカーネルプログラムではプログラム実行中に 1 ビットのエラー発生のみが許容できるなど、エラー耐性を多少なりとも持つことが分かった。そのため、

アプリケーションが要求する Quality of Results (QoR) を満たすようにプログラムの改良を行った。具体的には QoR の要求が厳しいアプリケーションに対しては、ランタイムで計算ノードが検算を行うことでエラーから回復する Algorithm-Based Fault Tolerance (ABFT) 技術を適用し、Approximate ネットワークで十分に完動できるように拡張を進めた。

本研究開始前には、ビッグデータ処理では入力データから大まかな傾向が理解できればよいため、Approximate ネットワークに適していると考えられたが、ポインタやリストを用いたアクセスが頻発する処理、グラフのサーチ処理の一部などでは適用可能な部分が限定され苦戦した。一方、厳密な計算が要求される、すなわち、高い精度が要求されると予想された高性能計算(High Performance Computing)では、反復計算を用いる例を含めてビット化け耐性を保持しているため適用可能な場合がある。このような知見は当初の予想と異なるものであり、興味深い結論が得られた。

今後は、SimGrid イベントドリブンシミュレーションにおいて用いた本実装を Approximate メモリ処理にも応用できるよう拡張させていく予定である。不揮発性メモリでは値の保存期間、信頼性について課題を抱えている。そのため、本技術の有望な応用先と考えられる。

以上の成果より、ムーアの法則による性能向上が成立しなくなると予想される 2025 年以降も Approximate ネットワークによりスパコン、データセンターのネットワークの性能を帯域と遅延面で向上する設計に関するパラダイムシフトを提唱できるよう引き続き検討を行う予定である。これは、2000 年頃、計算機システムの設計指標が、単純な性能に加えて消費電力が加わったことと同じ位のインパクトのある転換点である。すなわち、ポストムーア時代には、性能、消費電力、精度の 3 軸のトレードオフを最適化する技術を確立すべきであると考える。

<引用文献>

[1] 平成 28 年度-30 年度:科学研究費助成事業(基盤研究 B)Approximate Computing ネットワークの研究(16H02816)(代表:鯉渕 道紘)

[2] Daichi Fujiki, Kiyo Ishii, Ikki Fujiwara, Hiroki Matsutani, Hideharu Amano, Henri Casanova, Michihiro Koibuchi, High-Bandwidth Low-Latency Approximate Interconnection Networks, The International Symposium on High-Performance Computer Architecture (HPCA), 査読有, 2017, pp. 469 - 480

[3] 鯉渕 道紘、「不完璧」なデータセンターとスーパーコンピュータを目指そう、査読無、一般財団法人日本 ITU(国際電気通信連合) 協会、ITU ジャーナル誌、2017、Vol. 47 No. 2

[4] Wai Hong Ho, Timothy Mark Pinkston, A Design Methodology for Efficient Application-Specific On-Chip Interconnects, IEEE Transactions on Parallel Distributed Systems, 17(2), 2006, pp. 174-190

５．主な発表論文等
（研究代表者は下線）

〔雑誌論文〕（計 1 件）

① Michihiro Koibuchi, Tomohiro Totoki, Hiroki Matsutani, Hideharu Amano, Fabien Chaix, Ikki Fujiwara, Henri Casanova, A Case for Uni-Directional Network Topologies in Large-Scale Clusters, Proc. of the 19th IEEE International Conference on Cluster Computing (Cluster'17), 2017, pp.178-187
アクセス先 DOI:10.1109/CLUSTER.2017.33

〔学会発表〕（計 1 件）

①鯉渕 道紘, スーパーコンピュータの光速相互結合網, 電子情報通信学会総合大会(C1-3 次世代コンピューティングと光技術), 2018 年

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ
「Approximate ネットワークに関する研究」
http://research.nii.ac.jp/~koibuchi/research08.html

６．研究組織
(1)研究代表者
鯉渕 道紘 (KOIBUCHI, Michihiro)
国立情報学研究所・アーキテクチャ科学研究系・准教授
研究者番号： ４０４１３９２６

(2)研究協力者
〔主たる渡航先の主たる海外共同研究者〕
ヘンリ カサノバ(CASANOVA, Henri)
ハワイ大学マノア校・Information and Computer Science Department・教授

〔その他の研究協力者〕
松谷 宏紀 (MATSUTANI, Hiroki)
慶應義塾大学・理工学部・准教授