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

令和　　元 年　　６ 月 １４ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)（一般）
研究期間：2016～2018
課題番号：１６Ｈ０２８１６

研究課題名（和文）Approximate Computing ネットワークの研究

研究課題名（英文）A Study on Approximate-Computing Networks

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）　　13,000,000円

研究成果の概要（和文）：センサー観測データやソーシャルデータ処理に代表されるビッグデータ並列計算ではコンピュータとネットワークが提供する精度を落としても結果の大勢に影響せず,ユーザーが利用する上では十分なことが多い。そこで、本研究ではデータセンターなどの並列分散システムにおいて、情報の価値に応じた許容誤差でデータ転送することで、通信遅延、エネルギー性能比を向上させるApproximate Computing ネットワークを提案、探求した。

研究成果の学術的意義や社会的意義
1.近年、データ量が多い一方、計算精度を従来ほど厳しく要求しないビッグデータ処理の需要が劇的に増加しており、計算システムの在り方が質から量に変化しつつある。その変化に対応すべく、本研究では、Approximationに基づく新たな計算システム・ネットワークの設計法を提唱した点に社会的意義がある。
2.ネットワークが提供する信頼性の度合いを下げることで，新デバイス(たとえば光無線)の通信機器への導入障壁を緩和することが期待できる点に学術的意義がある。

研究成果の概要（英文）：Some big-data parallel applications, such as sensor observation data and social data processing, can accept multiple results caused by lowering the accuracy of computers and networks. In this study, we proposed approximate-computing networks to improve communication latency and communication-energy efficiency by allowing error-prone data transfers for parallel and distributed systems, such as data-centers.

研究分野：計算機システム・ネットワーク

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:数値 $0.1_{10進}$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。この点を逆に利用し、許容誤差を若干大きくすることで計算の精度を落とし、消費電力を削減し、スループットを向上させる Approximate Computing が注目されている。例えば、ディープラーニング系の計算を倍精度演算ではなく、半精度演算で行う試みが検討されている。一方、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は見られない。例えば、イーサネットの規格は 10^{-12} のビット誤り率を定めている。さらに InfiniBand ではより低いビット誤り率での安定したネットワーク運用が報告されている。

今後も、継続的に広帯域化していくために、光通信チャネルの変調フォーマットとして、スペクトル効率の高い直角位相振幅変調（QAM）などの高度なフォーマットの使用が見込まれる。しかし、この場合、ビット距離が近くなるため信号対雑音比耐性が低下し、CRC(巡回冗長検査)ではなく FEC(Forward Error Correction)によるエラー検出訂正を導入せざるを得なくなる。その FEC 計算のために、ケーブル毎に 100 ナノ秒、5 ワット程度のオーバヘッドが見込まれる。これは例えば 10 万本のケーブルを用いたスパコンの場合、5 ワット/本×100、000 本＝50 万ワット増加することとなる。通信遅延については、例えば 10 台のスイッチを経由してパケット転送する場合、11 本のケーブルを経由するため 1.1μ秒の通信遅延が増加し、並列計算システム全体として許容できるものではない。同様に、計算ノードのネットワークインタフェースや、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(200-400Gbps 級)が見込まれるが、現状と同じ通信遅延、消費電力のデータ転送すら実現できなくなる。

この問題を解決する 1 つの方法は、Approximate Computing の考え方に基づき、通信路においてビット誤りを許容する、すなわち、エラー検出訂正を省略、あるいは簡略化することで、飛躍的な通信性能の向上を実現することである。

ただし、ビット誤りは並列計算アプリケーションの挙動に影響を与えるために、並列計算アプリケーションの設計を慎重に行う必要がある。並列計算アプリケーションの開発環境として MapReduce の利用が挙げられる。MapReduce は大量データを計算機クラスタ上で並列処理するためのプログラミングモデルである。Map フェーズおよび Reduce フェーズをそれぞれ複数の計算機を用いて並列実行する。図 1 に MapReduce の概要を示す。MapReduce のような並列処理では、通信のビット誤りにより応答が不適切、あるいは応答が著しく遅い計算ノード（straggler と呼ぶ）がアプリケーション性能のボトルネックになってしまう。例えば、図 1 の MAP2 ノードの計算が著しく遅いとき、Map フェーズの結果が揃うまで Reduce フェーズにおいて待ち時間が発生してしまう。この待ち時間がそのまま実行時間の遅延に直結する。

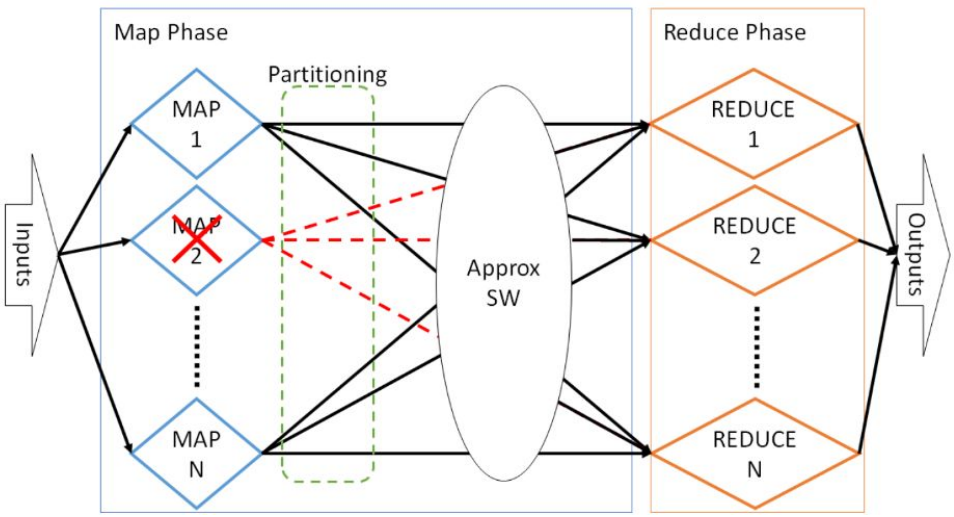

図 1 MapReduce の全体像と straggler(MAP2)の例

２．研究の目的

センサー観測データやソーシャルデータ処理に代表されるビッグデータ計算は、従来の厳密さが要求される計算と比べて、コンピュータとネットワークが保証する精度を大幅に緩和（誤差を大きく）しても結果の大勢に影響せず十分なことが多い。

そこで、本研究ではビッグデータ処理を行うデータセンター、スパコンを対象とし、情報の価値に応じた許容誤差でデータ転送するネットワークにより、通信遅延とエネルギー性能比を向上させる Approximate Computing ネットワークを提案、追求する。具体的には、Approximate Computing ネットワークに適用する光通信方式の同定を行う。次に、ネットワークのスイッチ

におけるボトルネック解消に向け、光スイッチを用いた光サーキットスイッチ技術の Approximate Computing ネットワークへの適用について検討を行う。さらに、典型的な並列プログラミングモデルとして、MapReduce において上述の straggler が引き起こす処理待ち時間を軽減する技術を開発する。

３．研究の方法

本研究では、(1) 情報の価値に応じた通信データのエラー耐性可変符号化、(2) ランダムショートカットと光通信技術の融合、(3) 並列アプリケーションの Approximation 化を探求し、統合することで Approximate Computing ネットワークの構成と有効性を示す。

(1) 情報の価値に応じたエラー耐性可変符号化

パケットのヘッダ情報とメッセージボディの各々にビット誤りが生じた場合に、計算ノード内の受信プロセスレベルでどのような影響が生じるかを解析する。例えば、パケットヘッダにビット誤りが生じた場合、受信プロセスはそもそもデータを受信できない。一方で、データの値が若干異なるのみという場合もありうる。つまり、エラー箇所ごとに受信プロセスに対する影響の度合いを調査する。そして、情報の価値に応じた エラー耐性を与えるための符号化について検討する。次に、エラー率を向上させる各エラー訂正・検出法(特に FEC と CRC)を通信遅延と消費電力の面で評価する。なお、過度なデータ圧縮やセキュリティ保護を行った通信については、デコード/エンコード処理遅延オーバヘッドが大きいことから現状の並列アプリケーション内の通信で用いていない。そのためこれらの通信に対する処理は本研究の対象外とする。

(2) ランダムショートカットと光通信技術の融合

ネットワークレイヤにおける Approximation による低消費電力・低遅延化にむけて、平均ホップ数・ネットワーク直径の小さなトポロジを開発する。そのために、研究代表者の鯉渕が提案したランダムトポロジをベースラインとし、2015 年より研究代表者の鯉渕が主催している「GraphGolf コンペ（http://research.nii.ac.jp/graphgolf)」での成果、すなわち、直径と次数の最適解を求めるオープングラフ理論問題の解を活用する。なお、Approximation によるパケットの低遅延到達性については研究代表者の鯉渕が保有するサイクルレベルネットワークシミュレータを拡張することで評価する。

光通信方式の同定に関しては、関連研究の動向調査をもとに、本課題の Approximate computing がターゲットとする 2020 年代後半以降に実用化される光通信技術について検討を行う。光サーキットスイッチ技術適用に関しては、関連研究の最新動向の調査を行うと共に、切り替え時間とスイッチポート数規模との間にトレードオフが存在する光スイッチ技術において、適切な光スイッチ技術の同定及びネットワークアーキテクチャの検討を行う。

(3) 並列アプリケーションの Approximation 化

計算ノード間の通信は、ネットワークスイッチを通過する点に着目し、ネットワークスイッチにて MapReduce における straggler の代理計算、および、代理応答を行う手法を探求する。ネットワークスイッチによる代理計算、代理応答は必ずしも正確なものではないため、本研究ではこれを ApproxSW と呼ぶ。

ApproxSW は straggler 以外の計算ノードの計算結果を受動的に収集することができる。そこで、straggler と傾向の近いノードからの計算結果を用いて straggler の代理応答を推測、近似することで生成する手法を提案、探求する。そして、ApproxSW を 10Gbit Ethernet（10GbE）インタフェースを 4 本有する FPGA ボード上に実現する。

４．研究成果

(1) 情報の価値に応じたエラー耐性可変符号化

ビット列から数値の符号化を工夫することで、ビット誤りがもたらす転送データ値の誤差を最小化する方式を開発した。対象とする並列アプリケーションにおいてプロセス間通信の転送データのフォーマットである IEEE754 浮動小数点数表現における符号部、指数部、仮数部の各ビット誤りがもたらす数値誤差は均一ではない。そこで、我々はビット誤りがもたらすアプリケーションの転送データの数値への影響を最小化するように符号部、指数部の上位ビットに対し直角位相振幅変調の一部の符号シンボルのみを用いることとして、精度と性能のトレードオフを探求した。さらに、その点を活用することができる最新のテクノロジを用いたネットワークアーキテクチャを示した。

転送データのフォーマットを通信のビット誤りによる誤差を抑えるように最適化する本成果は、チップ内、メモリ、ストレージなどの計算機システムのネットワークへの応用が見込める汎用性と将来性の高い技術であると考えられる。

(2) ランダムショートカットと光通信技術の融合

光トランシーバの小型化・低消費電力化に向けた技術開発が進んでいる。さらに、400Gbps 級の短距離向け通信規格では PAM-4（Pulse Amplitude Modulation）の多値変調方式が採用された。2020 年代後半にむけては 1 Tbps 超級の通信規格制定に向けたロードマップが議論され、高度な多値変調を可能とするコヒーレント検波方式の採用が検討されている。そのため、従来の高信頼通信性能を満たすために、従来と比べて高い誤り訂正処理能力を持つ技術の研究開発も同時に議論されている。誤り訂正技術について、必要な通信容量に応じて誤り訂正に必要と

なる DSP（Digital Signal Processing）回路を限ることで誤り訂正能力と電力をスケーリングさせる手法(引用文献①)、変調フォーマットの多値化、および、電力スケーリング誤り訂正技術という Approximate Computing ネットワークと親和性の高い技術動向が確認されている。そこで、それらの技術への Approximate Computing の応用技術を検討した。

データセンターのネットワークへの光スイッチを含む光通信技術の導入の研究開発は広く進められており、大まかに、ネットワークトポロジ再構成に光スイッチを適用する広帯域多ポート低速切替光スイッチ（～数 THｚ、100 ポート～、100ms～）を用いたアーキテクチャ、大粒度のフロー毎に切り替えを行う狭帯域多ポート高速切替光スイッチ（～100GHz、1000 ポート～、50us～）を用いたアーキテクチャ、小粒度のフローにも対応する少ポート高速切替光スイッチ（4 ポート～、100ns～）を用いたアーキテクチャに分類される。本研究では、広帯域（数 THｚ）で高速切替（数十 us）が可能なシリコンフォトニクス光スイッチの適用について検討を行った。シリコンフォトニクス光スイッチの適用には、多段構成によるポート数の拡張が重要となる。そこで、高次数のランダムネットワークトポロジ、あるいは GraphGolf コンペで公開されているすぐれた高次数のグラフをネットワークトポロジとして用いることで直径を小さくし、その結果、各パケットが目的地に到着するために、ビット誤りが生じる可能性が大幅に抑制できることを示した。

また、保有する光伝送実験システムを用いて、44Gbaud DP-16QAM 信号の現実的な伝送性能を測定したところ、ビット誤り率 10^{-3} を達成するには OSNR（Optical Signal to Noise Ratio）が 24dB 以上、10^{-5} を達成するには OSNR30dB 以上が必要であるとの結果が得られた。現在、実現されているシリコンフォトニクス光スイッチの性能をもとに、十分に大規模な光スイッチを構成するための多段構成を検討したところ、伝送後の光信号の OSNR は 25dB 程度と試算された。そのため、重要な情報に対しては簡易な誤り訂正処理を行う、あるいは、光スイッチの多段数が少ない範囲で通信を行う必要があるとの検討結果を得た。この点からもランダムネットワークトポロジおよび GraphGolf コンペで公開されているグラフで表されるネットワークトポロジが Approximate Computing ネットワークにおいて有望であるという結論に達した。

(3) 並列アプリケーションの Approximation 化

MapReduce の性能ボトルネックである遅延タスク(straggler)の問題に対処すべく、ApproxSW における近似計算によって、遅延タスクの代理応答を実現した。これは、Map フェーズの遅延タスクを検出し、その結果をネットワークスイッチ上で近似的に計算する。シミュレーション評価の結果、この近似計算による精度の低下は 7%に留まることがわかった。これほどの高い精度が得られることは、本研究開始時点では予期しておらず、新たな知見といえる。

また、ApproxSW を 10Gbit Ethernet インタフェースを有する FPGA を用いて実装することで、10Gbps までの通信量に対応することに成功した。NetFPGA-10G ボードを用いた実機評価の結果、10Gbps の 96%の通信量を保って機能実装することができた。さらに、Xilinx 社の Virtex-5 FPGA を搭載した NetFPGA-10G ボード上に ApproxSW を実装し、実機でスループット測定を行った結果、最大 33.51Gbps の性能が得られた。以上より、ApproxSW により、Approximate Computing ネットワークを用いた並列計算システムでの効果的な並列アプリケーションの実行が可能となることが分かった。

CPU と同様に通信機器もムーア則による性能向上、デナードのスケーリング則によるトランジスタの低消費電力化が 2020 年代後半には終焉する。そのため、ムーア則、デナードのスケーリング則に頼らない技術開発が求められている。本研究成果は、データの価値と伝送の確実性を関連づけることにより、大幅な省電力化と低遅延化を実現した点に特長があり、ムーア則によらずに通信性能の向上をするネットワーク基盤技術としての活用が期待される。

＜引用文献＞

① 亀谷総一朗、久保和夫、石井健二、土肥慶亮、杉原隆嗣、電力スケーリング誤り訂正による光通信網の構成、信学技報、vol. 117, no. 88、PN2017-6、pp. 1-6、2017

５．主な発表論文等

〔雑誌論文〕（計 9 件）

① Koya Mitsuzuka、 Michihiro Koibuchi 、 Hideharu Amano、 Hiroki Matsutani、 Proxy Responses by FPGA-based Switch for MapReduce Stragglers、 IEICE Transactions on Information and Systems、 査読有、 E101-D、 2018、 pp. 2258-2268、 10.1587/transinf.2017EDP7287

② Yao Hu、 Michihiro Koibuchi、 The Impact of Job Mapping on Random Network Topology、 CANDAR Workshops(CSA)、 査読有、 2018、 pp. 79-85、 10.1109/CANDARW.2018.00024

③ Ke Cui、 Michihiro Koibuchi、 Performance Evaluation of Collective Communication on Random Network Topology、 CANDAR Workshops(CSA)、 査読有、 2018、 pp. 159-162、 10.1109/CANDARW.2018.00036

④ Truong Thao Nguyen、 Hiroki Matsutani、Michihiro Koibuchi、Low-Reliable Low-Latency Networks Optimized for HPC Parallel Applications、 Proc. of the 17th IEEE

International Symposium on Network Computing and Applications (NCA)、 査読有、 2018、 pp. 1-10、 10.1109/NCA.2018.8548063

⑤ Thao-Nguyen Nguyen、Michihiro Koibuchi、Cable-geometric error-prone approach for low-latency interconnection networks、 17th IEEE/ACM International Symposium on Cluster、 Cloud and Grid Computing (CCGRID)、 査読有、 vol. 1、 2017、 pp. 699-702、 10.1109/CCGRID.2017.46

⑥ Koya Mitsuzuka、Ami Hayashi、Michihiro Koibuchi、Hideharu Amano、Hiroki Matsutani、In-Switch Approximate Processing: Delayed Tasks Management for MapReduce Applications、 Proc. of the 27th International Conference on Field-Programmable Logic and Applications (FPL)、査読有、vol. 1、2017、pp. 1-4、10.23919/FPL.2017.8056802

⑦ 平澤 将一、Truong Thao Nguyen、鯉渕 道紘、広帯域低レイテンシの Approximate ネットワークに対する 自動チューニング手法、 16 回情報科学技術フォーラム(FIT2017)、 査読無、 1 巻、 2017、 CC005 (計 8 項)

⑧ Nguyen T. Truong、Ikki Fujiwara、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、査読有、vol. 4、2017、pp. 989-1001、10.1109/TPDS.2016.2613043

⑨ 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、 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、 査読無(招待論文)、 2 巻、 2017、pp. 34-37

〔学会発表〕（計 12 件）

① Michihiro Koibuchi、Low-Latency Error-Prone Optical Networks for Fast Approximate Computation on High-End Datacenters、 OECC/PSC2019 (24th Optoelectronics and Communications Conference / International Conference on Photonics in Switching and Computing 2019)（招待講演）(国際学会)、2019 (to appear)

② Michihiro Koibuchi、Approximate HPC Networks for Imperfect Computing、 CREST International Symposium on Bi g Data Application（招待講演）(国際学会)、2019

③ 鯉渕 道紘、 光無線によるビッグデータ処理向け相互結合網の研究開発、 総務省関東総合通信局、戦略的情報通信研究開発セミナー2019（招待講演)、 2018

④ Kiyo Ishii、Shu Namiki、Highly available optical physical layer for future telecom and datacom networks、 CANDAR workshop (CANREXI)（招待講演）(国際学会)、2018

⑤ 石井 紀代、並木 周、Photonics in Switching and Computing PSC2018 報告ネットワークシステムアーキテクチャ関連、 電子情報通信学会 EXAT 研究会（招待講演)、2018

⑥ 鯉渕 道紘、Approximate Computing と関連する通信技術、 光ネットワーク産業・技術研究会、 第３回討論会公開ワークショップ（招待講演)、2018

⑦ 鯉渕 道紘、スーパーコンピュータのための光速相互結合網、 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)（招待講演)、2018

⑧ 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)（国際学会)、2017

⑨ Nguyen T. Truong、Henri Casanova、鯉渕 道紘、 Discussion on Approximate Interconnection Networks、 組込み技術とネットワークに関するワークショップ (ETNET)、2017

⑩ 鯉渕 道紘、IoT/ビッグデータ専用計算システムのための光通信技術、組込みシステムシンポジウム(ESS)2016（招待講演)、2016

⑪ Akihiko Hamada、Hiroki Matsutani、Design and Implementation of Hardware Cache Mechanism and NIC for Column Oriented Databases、Proc. of the 11th International Conference on ReConFigurabl e Computing and FPGAs (ReConFig)（国際学会)、2016

⑫ Kiyo Ishii、 Shu Namiki、 Toward Exa-scale Photonic Switch System for the Future Datacenter、IEEE/ACM International Symposium on Networks-on-Chip (NOCS)、(国際学会)、2016

〔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〕

ホームページ等

Approximate ネットワークに関する研究(国立情報学研究所鯉渕研究室)
http://research.nii.ac.jp/~koibuchi/research08.html

６．研究組織

(1)研究分担者

研究分担者氏名：松谷 宏紀

ローマ字氏名：（MATSUTANI, Hiroki）

所属研究機関名：慶應義塾大学

部局名：理工学部（矢上）

職名：准教授

研究者番号（8 桁）：70611135

研究分担者氏名：石井 紀代

ローマ字氏名：（ISHII, Kiyo）

所属研究機関名：国立研究開発法人産業技術総合研究所

部局名：エレクトロニクス・製造領域

職名：主任研究員

研究者番号（8 桁）：90612177