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　 ７ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2016～2018
課題番号：１６Ｈ０２７９３

研究課題名（和文）超広帯域I/0を想定したアーキテクチャの検討

研究課題名（英文）Study of an architecture which leverages ultra-wide bandwidth I/0

研究代表者
工藤　知宏（KUDO, Tomohiro）

東京大学・情報基盤センター・教授

研究者番号：００２３４４５１

交付決定額（研究期間全体）：（直接経費）　　13,100,000円

研究成果の概要（和文）：広帯域通信が計算機システム内で利用できるようになることを想定し、計算機システムの構成法を検討した。将来、光通信技術の発展によりどの程度の帯域の通信をどの程度の消費電力で利用できるようになるかを検討し、この想定される性能をもとに、構成方式を検討し、アクセラレータ間を直接結合するアーキテクチャが有効であることを示した。また、広帯域通信を活用する資源管理技術やI/0方式の検討を行った。

研究成果の学術的意義や社会的意義
ムーアの法則に従った半導体集積度の向上は限界を迎えつつあり、これに伴うシステムの性能向上も望めなくなりつつある。このような状況下でシステムの性能向上を図るために、広域網の技術を適用した超広帯域光通信技術の計算機システムへの導入の可能性を検討し、広帯域通信を生かした新しい通信アーキテクチャの導入が計算機システムの性能向上に有効であることを示した。

研究成果の概要（英文）：Considering deployment of wide bandwidth optical communication, we have investigated communication architecture of future computing systems. First, we have examined the achievable bandwidth and power consumption of optical communication in computing systems. Then, according to the estimated performance of communication, we have investigated the organization of future computing system architecture, and showed the effectiveness of the direct coupled inter-accelerator communication architecture. Also we have investigated resource management scheme and I/0 architecture of the proposed computing system.

研究分野：計算機システム

キーワード：高性能計算　ポストムーア　通信方式　インターコネクト　データセンター

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

高性能な計算機システムでは、計算ノード間の通信帯域を向上させるために光通信技術が用いられるようになりつつあった。しかし、ファイバあたりの通信帯域は最大でも 100Gbps 程度であり、ノード内でのプロセッサの処理スループット(例えば倍精度 100GFLOPS の二項演算は、100G x 64bit x 3 = 19.2Tbps の入出力スループットを持つ）や、メモリの帯域と比べて小さい。一方、広域の通信では、波長多重(DWDM: Dense Wavelength Division Multiplexing)などの多重化技術によりファイバあたり 100Tbps 超の通信が実現されている[1]。しかし、広域向けの光通信技術は、消費電力やサイズ、コストなどの点で、計算機システム内部で用いることができなかった。

２．研究の目的

広帯域通信が計算機システム内で利用できるようになることを想定し、計算機システムの構成法を検討する。

３．研究の方法

高い通信性能を生かす計算機システムの構成法について方式の検討と評価を行った。また、今後実用化が期待できる計算機システム向け光通信技術についての検討を行うとともに、特に消費電力の観点から、システムの要件を検討した。FPGA 上に、プロセッサや専用計算回路を配置しその速度を変更することで、想定する将来の計算システムよりも遅いが通信性能と計算・メモリアクセス性能のバランスが同じ環境を構築し、実アプリケーションのコア部分を動作させて性能を評価した。

４．研究成果

(1) 全体アーキテクチャと通信方式の検討

これまで計算機性能の向上ペースは通信帯域の向上ペースを上回っていた。このため、計算機の利用においては通信量を減らしなるべくデータを動かさずに計算を行う Near-Data Processing 手法の開発が行われてきた。しかし、広域網で使われている広帯域通信技術を用いれば、データセンター内の通信帯域を現在の 100 倍にあたる 10Tbps 程度に飛躍的に向上させることができる。これは、通信と計算の性能のバランスが劇的に変化することを意味し、従来と異なる新しいコンピューティング方式、通信方式が求められる。

インターコネクションの帯域が DRAM のアクセス帯域より大きく、データを専用ハードウェア間で動かす電力が DRAM のアクセス電力に比べて小さければ、このようなデータを動かす方式に優位性が出てくる。現在最も帯域が大きい DRAM 規格である HBM2 で、帯域 2Tbps 程度であり、その消費電力は 5~6pJ/bit 程度とされている。一方、光通信技術の改良により、帯域 10Tbps、消費電力 1~5pJ/bit 程度のノード間通信の実現が見込めることが明らかになった。

このような広帯域通信を活かしたシステムの実現には、光インターコネクション、光デバイス技術だけでなく、メモリを用いずに広帯域通信を効率よく利用できる通信方式の開発も必要になる。

従来の計算機間通信方式は、サーバ計算機間のメモリ間コピーを基本としていた。サーバのメインメモリにデータをコピーすることなく、アクセラレータの持つメモリ間で直接コピーを行う手法も開発されているが、通信帯域がメモリアクセス帯域よりも大きい場合の通信手法は検討されていなかった。また、通信においてはフローコントロールが必要で、広帯域の通信ではフローコントロールのオーバヘッドが顕在化してしまう。そこで我々は、従来一つの LSI 内のロジック間で用いられている通信手法を、アクセラレータ間の通信にも用いる手法を提案した。これは、図１のように、通信する複数の装置（チップ）に同一のクロック信号を与えて、チップ間通信をチップ内の通信と同様に考えるというものである。

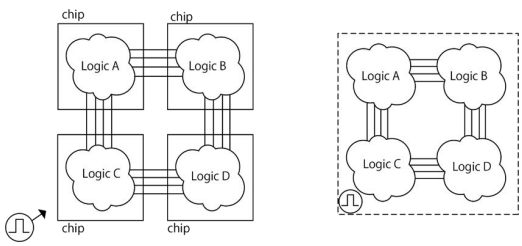

図１　大規模回路の分割実装モデルによる通信の概念図

チップ間の通信では、チップ内の通信のように多数の信号線を用いることはできず、高速シリアル信号を用いる必要がある。しかし、この高速シリアル信号は、チップ内の並列信号を多重化して通信しているものであり、端点側では並列信号として扱うことが可能である。例えば、10Tbps の高速シリアル通信は、1Gbps の信号を１万本束ねたものと考えることができる。一方、チップ間の通信には多重化(SERDES)と伝搬遅延等による大きな遅延が伴う。これは、チップ内での複数のレジスタを介した通信とモデル化することにより、図２のように全体を一つのチップと同様に扱うことが可能となる。

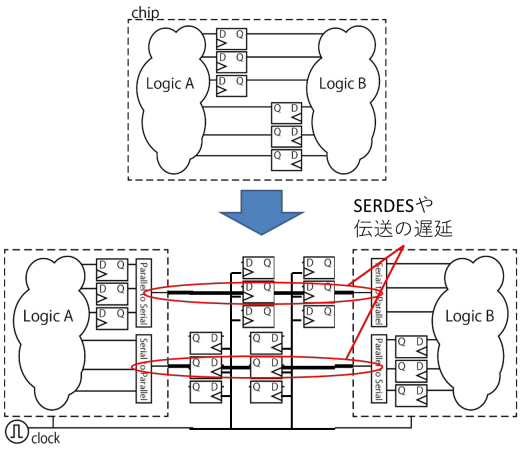

図２ 高速シリアル信号上の同期通信

この方式では、システム全体が同期して動作するため、非同期通信委起因するチップ間のフローコントロールが不要となる。実装する回路自体の動作タイミングを極力予想可能とすることにより、特に広帯域通信ではフローコントロールを排除し、オーバヘッドを削減することが可能である。一方、現在の高速シリアル信号の実装ではトレーニングサイクルが必要であり、試作した FPGA システムではトレーニングサイクルの発生時刻をコントロールすることができなかった。このような通信方式の実現には、高速シリアル信号の実装方式の改良も必要であることが明らかになった。

(2) I/O 性能を活用するシステムソフトウェアの検討

メモリに匹敵する I/O 性能を活用するためのシステムソフトウェアについて次のように検討を行った。並列計算機システムは、共有メモリ型システムと分散メモリ型システムにわけられる。HPE The Machine のようなラック規模を対象とするシステムでは前者に近いアプローチも取り得るが、スケーラビリティに課題がある。本研究では、個々の演算器がローカルメモリを持つ分散メモリ型システム上において、複数の演算器がパイプライン的に処理を実行することを前提とする。そして、我々が研究開発しているアクセラレータクラウドを実現するシステムソフトウェア FlowOS を題材に、本提案アーキテクチャに対するシステムソフトウェアの検討を行った。

FlowOS はアクセラレータの資源管理を行う FlowOS-RM とプログラミング環境を提供する FlowOS Job から構成される。FlowOS-RM では、アプリケーションを実行する演算器にタスクを割り当て、光サーキットスイッチを制御することで、演算器間を広帯域 I/O で接続する。タスク割当は、波長などのハードウェア資源量の制約を考慮して実行する必要がある。FlowOS Job では複数のタスクからなる有向グラフとしてアプリケーションを記述できる。なお、タスクのカーネルは C/C++やアクセラレータに特化したプログラミング言語(CUDA、HDL など)で実装し、FlowOS Job ではそれらを組み合わせることでプログラムを構成する。API レベルでは、タスク間の通信はバッファとそのコピーとして抽象化される。アーキテクチャ側が提供する RDMA を用いることで、ソフトウェアによるオーバヘッドを極力削減した通信を実現可能である。検討の結果、提案するシステムソフトウェアを用いることで、高効率かつスケーラブルな計算機システムが構築できる見込みを得た。

(3) 大規模不規則ネットワーク用ルーチング手法の検討

将来のデータセンターの結合網は、光を利用した巨大な不規則ネットワークになる可能性が高く、これに対応するルーチング手法を検討した。不規則ネットワークのスループットを向上させるためには、複数経路を利用してトラフィックを分散する必要がある。従来手法である Equal-cost multi-path routing (ECMP)は、最短距離経路のみを利用するため、経路数が不足し、短いほうから k 本の複数経路を利用する k-shortest path routing は、スループットが必ずしも最大化されない問題点があった。そこで、k 本の経路を線形計画法による最適化の結果から探索する k-optimized-path routing を提案した。この手法では、すべてのトラフィックに対するワーストケースを最適化する方法と、既知のトラフィックに対するスループットを最適化する方法を場合に応じて使い分ける。この手法により、トポロジ非依存(図３左)は最大 35%、依存（図３右）は最大 98%スループットを改善した。

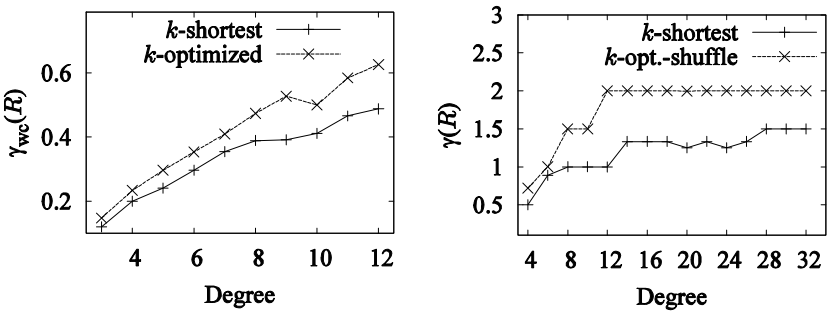

図３：スループットの改善

(4) FPGA NIC を用いた広帯域 I/O アーキテクチャ向けアプリケーションの研究

広帯域 I/O を想定したアーキテクチャ検討の一部として、FPGA を用いたネットワークサービスの高速化を研究した。具体的には、FPGA を用いた 10Gbit Ethernet（10GbE）ネットワークインタフェース（以下、FPGA NIC と呼ぶ）上で、ネットワークサービスの一部を高効率に実現することを研究した。FPGA NIC を用いたこのような取り組みとして、Key-Value Store（KVS）と呼ばれるデータストアの高効率化がよく研究されている。これは、KVS に対するデータ取得要求の応答結果を FPGA NIC 上にキャッシュしておき、再度、同じデータが要求されたときには FPGA NIC 上にキャッシュされている応答結果を返信するものである。

(5) 広帯域 I/O 接続方式の検討

広帯域 I/O 接続方式について次のように検討を行った。現時点において、高バンド幅メモリアクセスを実現する技術として、2.5 次元積層に基づく HBM(High Bandwidth Memory)技術と Hybrid Memory Cube(HMC)に代表されるメモリキューブが挙げられる。しかし、HBM は実装面積の面で容量が限られる。そのため、メモリの大容量化のためには、プロセッサチップと数十 cm 離れたボード上に設置できるメモリキューブが有力である。メモリキューブは HBM と比べて通信距離が圧倒的に長いため、その消費電力が問題になっている。具体的には、3 次元積層した HMC 自体の消費電力よりも通信路の消費電力の方が大きくなることがたびたび起こる。これは電気パケット通信を用いることが原因の一つである。
そこで、光サーキットスイッチ通信をメモリ通信に用いる研究はこれまでほとんどなかったが、本研究で開発した光サーキットスイッチ通信のメモリ通信への応用を検討した。つまり、光サーキットスイッチを用いることで、この通信路の消費電力は大幅に削減されることが期待される。メモリキューブのスイッチは低次数(HMC の場合は 4)のリンクで相互接続する。そして、ローカルプロセッサが各メモリキューブを管理するため、メモリキューブへのアクセスアパターンはローカルプロセッサ間である程度限定される。このスイッチの低次数とアクセスパターンの局所性を利用することで必要となるサーキット数を抑える現実的な実装が可能となる。よって、光サーキットスイッチ通信の応用先として、ボード上の数十 cm の短距離通信に用いることが有望であると導くことができる。

(6) データセンター/HPC 向け光技術の将来技術予測

文献検索や国際会議への出席などによりデータセンター/HPC 内ネットワークの技術動向調査を進め、2030 年頃までの同ネットワーク総容量の増加傾向を推定するとともに、簡易なモデルを基にデータセンター全体の消費電力予測を行った。その結果、現状の外挿では 2028 年頃までにトランシーバのラインレートは 10Tbps、データセンター内のネットワーク総容量は 1Ebps を超え、これを既存技術の延長で実現したとすると、楽観的にもネットワークだけで 500MW 以上の電力量を要することが分かった（図４）。

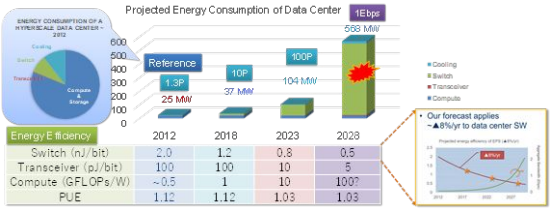

図４　データセンターネットワークの消費電力予測

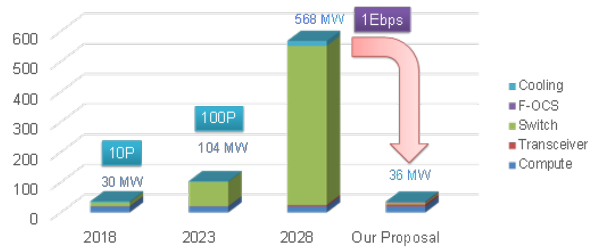

図５　提案手法による電力削減

そこで、並木らが開発するμ秒の速さで切り替え可能なシリコンフォトニクス光スイッチを多数用いてエクサビット級の総容量を実現するスイッチシステムを検討し、これと既存ネットワーク技術と併用することで大幅に電力を削減できることが分かった（図５）。
このような検討をまとめ、将来のデータセンターでは、スイッチ速度がμ秒の回線交換網をデータセンターシステム全体で如何に有効に活用できるかが今後の重要な課題であることを主張すべく、ヨーロッパの主要国際会議 ECOC2018 のワークショップでパネリストとして報告した。

(7) 光ネットワーク技術、光ネットワークアーキテクチャ検討

光スイッチを用いた光回線交換ネットワークのデータセンタ/HPC 内ネットワーク適用に関する二大ボトルネック、すなわち、光スイッチポート数制約と光スイッチスピード制約について技術動向調査及び技術検討を行った。一般に、光スイッチポート数と光スイッチスピードとの間にはトレードオフの関係が成り立つ。このトレードオフを解消するために、関連研究では、波長空間を使用しポート数の拡大を図る（この場合、ポートあたりの帯域とポート数とがトレードオフとなる）構成や、多数の光スプリッターと高速小規模光スイッチとを組み合わせる構成（この場合、ポート数と光信号品質とがトレードオフとなる）などが検討されている。また、高速光スイッチを使いこなすには、高速な光スイッチ制御プレーンが必要となるため、同期制御方式や分散制御方式が検討されているが、スケーラビリティや実装の観点で課題が残る。

本検討では、μ秒の速さで切り替えが可能であり、数十ポートのスイッチチップが既に実現されているシリコンフォトニクス光スイッチを用いたシステム構成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を多段に用いる構成により、エクサビット級の総容量の実現が可能であること、多段に光スイッチを通過する際のレベルダイヤの検討により信号品質に問題がないこと、および、光スイッチの挿入損失低減によりスイッチステム全体の消費電力低減が可能であるとの試算結果を得た。また、光スイッチの制御プレーンについて、一般的な Linux OS 上でのソフトウェア実装により 200μ秒以下での光スイッチ設定更新間隔が実現可能であることを確認した。

５．主な発表論文等

〔雑誌論文〕（計　2　件）

① Yuma Sakakibara, Shin Morishima, Kohei Nakamura, Hiroki Matsutani, Hardware-Based Caching System on FPGA NIC for Blockchain, IEICE Transactions on Information and Systems 査読有, Vol.E101-D, No.5, 2018, 1350-1360, DOI:10.1587/transinf.2017EDP7290
② Ryuta Kawano, Ryota Yasudo, Hiroki Matsutani, Michihiro Koibuchi, Hideharu Amano, HiRy: An Advanced Theory on Design of Deadlock-free Adaptive Routing for Arbitrary Topologies, Proc. of the IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS'17) 査読有, 2017, 664-673, DOI: 10.1109/ICPADS.2017.00091

〔学会発表〕（計　27　件）

① Ryousei Takano, A Disaggregated Accelerator Cloud Data Center, Disaggregation in data centers and metro networks Workshop（招待講演）, 2018
② 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
③ Shu Namiki, Opportunity and Challenges of Silicon Photonics Switches for Future Data Centres, ECOC2018（招待講演）, 2018
④ Tomohiro Kudoh, Revisiting computing and communication: New landscapes in wide area, data centers and computing systems, The Sixth International Symposium on Computing and Networking (CANDAR '18)（招待講演）, 2018
⑤ Tomohiro Kudoh, High Performance Computing in Japan and the Role of Optics in the Future, OFC 2019（招待講演）, 2018
⑥ Shu Namiki, Tomohiro Kudoh, Christian Koos, Chris Cole, Will Optical Switching Drive Data Center Design in 2028? - Team A Presentation, OFC2018 Workshop（招待講演）, 2018
⑦ 赤沼 領大、高野 了成、工藤 知宏, CNN の学習におけるチャネル方向並列化の提案, 情報

処理学会ハイパフォーマンスコンピューティング研究会，2018
⑧ 工藤知宏、並木周，将来のデータセンター像と光インターコネクション・光デバイスへの期待，2018 年電子情報通信学会総合大会（招待講演），2018
⑨ Kiyo Ishii, Takashi Inoue, Shu Namiki, Toward Exa-scale Optical Circuit Switching Interconnect for Future Computing Systems, PHOTONICS: Photonics-Optics Technology Oriented Networking, Information and Computing Systems （招待講演）, 2017
⑩ Kohei Nakamura, Ami Hayashi, Hiroki Matsutani, An FPGA_Based Low-Latency Network Processing for Apark Streaming, 4th IEEE International Converence on Big Data /1st Workshop on Real-Time and Stream Analytics in Big Data, 2016
⑪ 工藤知宏，異種プロセッシング装置を構想区な通信で柔軟に組み合わせるポストムーア時代のアーキテクチャ，第 15 回情報科学技術フォーラム(FIT)（招待講演）,2016
⑫ 石井紀代，インターコネクト向け光ネットワーク技術，第 15 回情報科学技術フォーラム(FIT)（招待講演）,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〕

６．研究組織

(1)研究分担者

研究分担者氏名：高野　了成

ローマ字氏名：(TAKANO, Ryousei)

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：情報・人間工学領域

職名：研究グループ長

研究者番号（8 桁）：10509516

研究分担者氏名：並木　周

ローマ字氏名：(NAMIKI, Shu)

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：副研究部門長

研究者番号（8 桁）：30415723

研究分担者氏名：鯉渕　道紘

ローマ字氏名：(KOIBUCHI, Michihiro)

所属研究機関名：国立情報学研究所

部局名：アーキテクチャ科学研究系

職名：准教授

研究者番号（8 桁）：40413926

研究分担者氏名：天野　英晴

ローマ字氏名：(AMANO, Hideharu)

所属研究機関名：慶應義塾大学

部局名：理工学部（矢上）

職名：教授

研究者番号（8 桁）：60175932

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

(2)研究協力者

なし

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。