7G-02

ラック間をまたぐ計算コンポーネント間通信への光無線パスの割り当て手法

原　弘明† 尾崎　友哉‡ 森島　信‡ 松谷　宏紀†

慶應義塾大学　理工学部† 慶應義塾大学大学院　理工学研究科‡

1. はじめに

CPU・メモリ、ストレージ、GPU などの計算機を構成するハードウェアリソースの利用率向上を目的として、必要なハードウェアリソースをソフトウェア側で定義して利用しようとする計算機アーキテクチャが考えられている。このようなアーキテクチャを実現するために、各ハードウェアリソースをラック毎にまとめて設置し、それらのラックをネットワークで接続して利用することで、ラック間をまたいだハードウェアリソースの利用率向上を図ることができる。各ラックを接続するネットワークは有線ケーブルで構成されており、大規模なものを構成すると複数のスイッチを介することによる通信遅延が問題となる。

ここで、本研究では光無線リンクを動的に割り当てることでショートカットリンクを構築し、通信遅延の問題を改善することを提案する。

2. 関連研究

2.1. ラックスケールアーキテクチャ

図 1(a)に示すようなラックスケールアーキテクチャでは、それぞれのハードウェアリソースを大量に 1 つのラック内に格納し、それらをネットワークで接続することでラック全体を 1 つの計算機として考える[1]。こうすることで各アプリケーションにラック内のハードウェアリソースを自由に割り当てられる。さらに、図 1(b)に示すように、異なるラックに格納されたハードウェアリソースをラック間ネットワークで接続して 1 つの計算機として扱うこともできる。本論文ではこのようなラック間をまたぐアーキテクチャを先に述べたようなものと区別するため”インターラックスケールアーキテクチャ”と呼ぶこととする。

2.2. NEC ExpEther

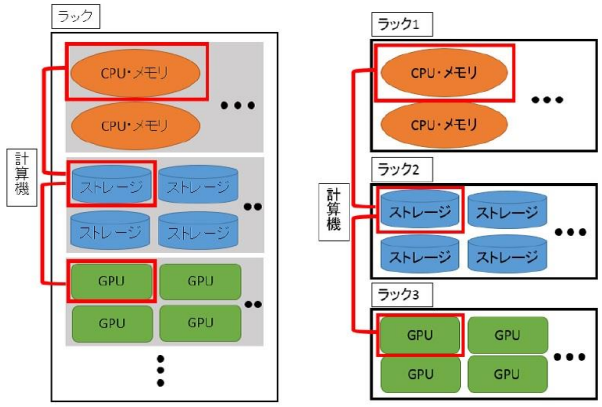
ExpEther は、Ethernet を利用して PCI Express を延長する技術である[2]。本論文では各種ハードウェアリソースが ExpEther 技術を利用して Ethernet で接続されているものと想定する。

2.3. 光無線通信

光無線通信は、10Gbps 以上の高いスループットを光電変換なしで提供できる[3][4]。サーバラック上部にコリメータレンズを設置し、ラック間通信に光無線リンクを用いることを想定している。

3. 提案内容

インターラックスケールアーキテクチャにおいて、計算機内での通信遅延の問題を解決するために、動的に光無線リンクを割り当てることでショートカットリンクを構成することを提案する。



(a) (b)

図 1. (a)ラックスケールアーキテクチャと(b)インターラックスケールアーキテクチャの概念図

3.1. 全体のアーキテクチャ

計算機を構成するハードウェアリソースとして CPU・メモリ、ストレージ(SSD)、GPU の 3 種類が存在し、それぞれがラックごとに分類して格納されている図 1(b)に示すようなインターラックスケールアーキテクチャを想定する。ラック間の有線ネットワークトポロジとして、3 階層のスイッチからなる Fat-tree を考える。この場合、リソース間のホップ数は 4 から 6 ホップとなる。このようなラック間ネットワークにおいて、各ラックに光無線デバイスを設置し光無線リンクによるショートカットリンクを構成すると、光無線リンクを経由した場合のホップ数は 3 ホップとなる。

3.2. 光無線リンクの割り当てアルゴリズム

光無線リンクは動的にリンクを構成することができるため、通信状況に応じてリンクを切り替えることでより効果的にネットワーク性能を改善することができる。ネットワーク性能の改善とはネットワーク全体における平均ホップ数の削減であり、平均ホップ数の削減により通信遅延の改善が見込める。

本論文では”割り込みなし”と”割り込み付き”の簡単な 2 つのアルゴリズムを用いて光無線リンクを割り当てる。割り込み付きアルゴリズムでは通信要求のあったデータ転送量に応じて通信中の光無線リンクの切り替えを許可する。

4. 評価

4.1. 複数ホップのアプリケーション性能

4.1.1. 実機評価環境

メインマシン(CPU)と SSD を、ExpEther を用いて 10Gbit Ethernet(10GbE)で接続し、10GbE スイッチの代わりに NetFPGA10G と呼ばれる 10GbE インターフェースを 4 個有する FPGA ボード[5]に Reference Switch Learning Lite と呼ばれる 10GbE スイッチ機能を焼き込んで使用した。FPGA ボードを複数用いて複数ホップのネットワーク接続を再現する。

An FSO Link Allocation Strategy for Inter-Component Communication across Racks

†Faculty of Science and Technology, Keio University

‡Graduate School of Science and Technology, Keio University

4.1.2. 評価結果

SSD への書き込みを行った際のスループットを図 2 に示す。ホップ数が増加するにしたがってスループットが減少している。これは dd コマンドのオプションによって 1MB ずつのブロック単位で 5.2GB 分のファイルを書き込んでいるため、ブロック毎に通信遅延の影響を受けることが原因だと考えられる。

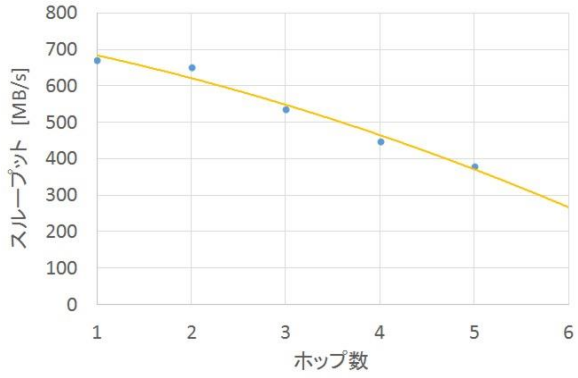


図 2. SSD への書き込み性能の評価

4.2. 提案手法のシミュレーション

4.2.1. シミュレーション環境

C 言語で作成したシミュレータによってインターラックスケールアーキテクチャを構成するネットワークの平均ホップ数を算出することで、光無線リンクの割り当てによる効果を評価する。本論文で対象とするインターラックスケールアーキテクチャは 48 ポートスイッチを用いた 3 階層の Fat-tree 構造として、全ラックの数は 1152 台、ハードウェアリソースの内訳は CPU・メモリ、ストレージ、GPU それぞれ 384 台とした。以下にシミュレータの動作内容について記す。データ転送を含む処理時間が 500～5000 サイクルかかるタスクがランダムに発生する。タスクの発生時刻は 1～5000 サイクルの間でランダムとする。タスク発生場所は CPU・メモリのラックのうちでランダムに発生し、ストレージまたは GPU とのデータ転送を伴う処理をするものとする。1 回のシミュレーション中に発生するタスクの数は 100～10000 個としてそれぞれシミュレーションを行う。発生するタスクが多いほどネットワークのワークロードは大きくなる。各タスク毎に処理にかかったホップ数を記録して全タスクで平均を取り、ネットワーク全体の平均ホップ数を算出する。各ラックに構成される光無線リンクの数は 0～4 本で変更してシミュレーションを行う。

4.2.2. 評価結果

各アルゴリズム毎に平均ホップ数の変化を図 3,4 に示す。ワークロードは 1 回のシミュレーション中に発生するタスクの数で表している。光無線リンクの数が多くなるほど平均ホップ数の改善が見られる。また、割り込み付きアルゴリズムのほうが、より通信量の大きなタスクに対して光無線リンクを割り当てることができるため、ホップ数を削減できている。

4.3. シミュレーション結果に基づくアプリケーション性能

4.1.2 節で示したシミュレーション結果に基づくホップ数の削減が、4.2.2 節で示した実機でのアプリケーション性能にどのくらい影響するかを検討する。ホップ数の変化による SSD への書き込み性能の改善を図 5 に示す。図 5 は改善率の最も高かった光無線 4 本/ラックを想定したときの評価を示している。ワークロードが小さく、光無線デバイスの数が多いほど光無線リンクを十分に割り当てることができるため、改善率は高くなる。

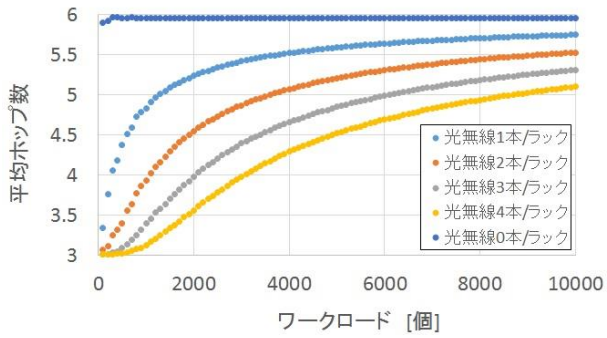


図 3. 割り込みなしアルゴリズムの評価

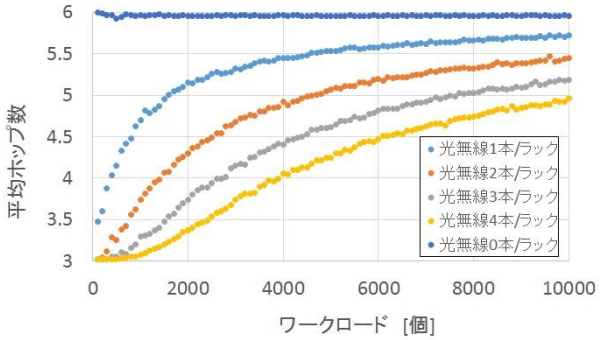


図 4. 割り込み付きアルゴリズムの評価

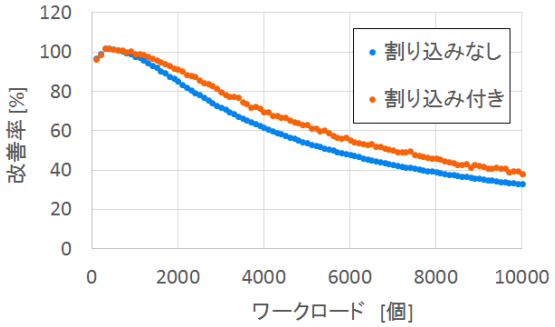


図 5. 光無線 4 本/ラックを想定したときの SSD への書き込み性能の改善率

5. まとめ

本研究ではインターラックスケールアーキテクチャを想定した時に光無線を用いて動的に光無線リンクを割り当てることで、ネットワークのショートカットリンクとして機能させる手法を提案した。これによりラック間ネットワークのホップ数を削減できた。また、ホップ数の削減により通信遅延が減少し、SSD への書き込み性能が光無線デバイスの数に応じて最大 40%改善するという結果が得られた。

参考文献

[1]J. Kyathsandra and E. Dahlen, “Intel Rack Scale Architecture Overview，" Interop，May 2013．

[2]J. Suzuki, Y. Hayashi, M. Kan, S. Miyakawa, and T. Yoshikawa, “End-to-End Adaptive Packet Aggregation for High-Throughput I/O Bus Network Using Ethernet," International Symposium on High-Performance Interconnects, pp.17-24, Aug. 2014.

[3]”光無線通信システム推進協議会," http://j-photonics.org/icsa/ .

[4]”ビル間高速光空間通信網推進協議会," http://www.obn.ne.jp/ .

[5] “NetFPGA Probject," http://netfpga.org/.