

ON/OFF リンクにおける通信開始遅延を低減するための プリウェイクアップ手法の提案

松山 朋樹[†] 三輪 忍[†] 八巻 隼人[†] 本多 弘樹[†]

[†]電気通信大学

1 はじめに

近年のスーパーコンピュータは大量の電力を消費することが問題となっており、消費電力の削減が求められている。スーパーコンピュータの各ノードを接続するインターコネクション・ネットワーク（以下ネットワーク）における消費電力はシステム全体の最大 30%と言われており [1]、データ通信を行っていないリンクを低電力モードにすることで消費電力を削減する ON/OFF リンクが注目されている。

ON/OFF リンクでは、低電力モードのリンクにデータが到着した場合、リンクを通常モードへモード遷移させる時間分、データ通信の開始が遅延してしまう。これに対し、リンクオフスレッシュホールドを有する ON/OFF リンクが提案されているが、エネルギー削減効果が低下してしまう。

そこで、本研究では、リンクオフスレッシュホールドを有しない ON/OFF リンクで、上記の遅延がアプリケーション性能に与える影響を緩和するため、低電力モードのリンクを通信要求に先立って通常モードにし（プリウェイクアップ）、データ到着後直ちに通信を開始できるようにする方法を検討する。

2 ON/OFF リンク

2.1 概要

ネットワークの通常のリンクは、リンクの両端にある PHY がリンクの接続状態を確認するための信号を定期的に送信しているため、データ通信を行っていない状態でも一定の電力を消費する（図 1(a)）[2]。

一方、ON/OFF リンクでは、通信を行っていないリンクを省電力モードにすることでリンクの省電力化を図る（図 1(b)）。しかし、低電力モードのリンクにデータが到着した場合、通信を開始するためには PHY を低電力モードから通常モードにする必要があり、モード遷移にかかる時間により、通常のリンクに比べて通信性能が低下することが知られている。

2.2 リンクオフスレッシュホールドを有する ON/OFF リンク

ON/OFF リンクの問題点である通信性能低下を回避するため、先行するデータ通信終了後からリンクオフスレッシュホールド分の時間経過した後にリンクを低電力モードにする手法が提案されている（図 1(c)）[1]。リンクオフスレッシュホールド内に次のデータが PHY に到着した場合は、モード遷移することなく直ちにデータ通信を開始できる。

HPC アプリケーションにおける通信は、通信フェーズ（先行するデータの通信終了後から次のデータまでの間隔が短く、リンクに頻繁にデータが到着するフェーズ）と計算フェーズ（全く通信を行っておらずリンクにデータが到着しないフェーズ）に分かれる。リンクオフスレッシュホールドを有する ON/OFF リンクでは、通信フェーズでの頻繁なデータ到着によるリンクのモード遷移時間分の通信性能低下は抑制される。一方、計算フェーズでは、データの到着間隔が十分長い（リンクオフスレッシュホールド以上の長さである）ため、その間低電力モードにすることで、電力を削減できる。

2.3 先行研究

リンクオフスレッシュホールドを有する ON/OFF リンクでは、リンクオフスレッシュホールドをどのように設定するかが問題となる。そこで、先行研究 [1] では、ON/OFF リンクを採用した HPC システムを対象とし、クラスタシミュレータ Dimemas を用いて、アプリケーションの通信トレースを解析し、リンクオフスレッシュホールドとエネルギー消費量の関係から最適なリンクオフスレッシュホールドを 50μ と決定している。その結果、リンクオフスレッシュホールドを有する ON/OFF リンクでは、リンクの消費電力を 70% 削減でき、この際のアプリケーション性能の低下率は 2% であったと報告されている。

3 研究の目的

リンクオフスレッシュホールドを有する ON/OFF リンクでは、通信終了後リンクオフスレッシュホールド期間内の電力を消費することができない。そこで、本研究ではリンクオフスレッシュホールドを有していない ON/OFF リンクで、低電力モードのリンクを通信要求に先立って通信モードに遷移させ、データの到着後直ちに通信を開始することを可能とし、リンクの起動時間分の遅延を隠蔽するプリウェイクアップ手法（図 2d）を検討する。

Pre-wake up method for reducing wake up delay in ON/OFF Links

Tomoki Matsuyama[†], Shinobu Miwa[†], Hayato Yamaki[†] and Hiroki Honda[†]

[†]The University of Electoric Communications
182-8585, Tokyo, Japan
m1424080@edu.cc.uec.ac.jp

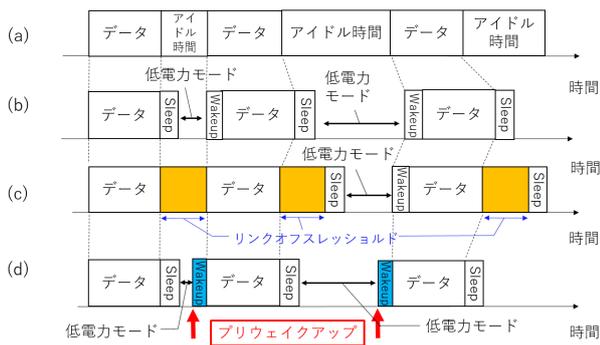


図 1: (a) 通常のデータ通信 (b)ON/OFF リンクでのデータ通信 (c) リンクオフスレッシュホールドを有する ON/OFF リンクでのデータ通信 (d) プリウェイクアップによる ON/OFF リンクでのデータ通信

4 プリウェイクアップ手法の事前評価

4.1 シミュレーション方法

プリウェイクアップの有用性を事前評価するため、HPC アプリケーションを対象として、並列実行のシミュレーションを行った。シミュレーションには、並列計算機シミュレータである SimGrid-3.11[3] を用いた。今回使用したアプリケーションは、Nas Paralell Benchmarks[4] である。NPB のうち、CG、FT、LU、MG を用い、入力クラスを W、A、B に、ノード数を 4 ノード、8 ノードに設定した。また、シミュレーションの各ノードの計算性能は 307.2GFLOPS、500GFLOPS、1000GFLOPS とし、1Gbps のネットワークである。シミュレーション対象の計算機システムは 1 台のスイッチと複数台の計算ノードから構成されている。スイッチに各ノードが接続されており、各ノードは、ノードからスイッチに向かう UP リンクと、スイッチからノードに向かう DOWN リンクで接続されている。シミュレーションにより得られた通信のトレースファイルを解析し、通信終了後から次の通信開始までの時間を調査した。

4.2 分析結果

各アプリケーションに対し、リンクのアイドル時間を解析した結果、ほとんどのアプリケーションの実行時間のうち 90%以上をアイドル時間が占めていた。

次に、リンクオフスレッシュホールドとリンクの消費エネルギー、通信性能の関係から、各システムの最適なリンクオフスレッシュホールドを求めた (表 1)。

表 1: 各システムの最適なリンクオフスレッシュホールド

ノード数	4	4	4	8	8	8
FLOPS 数	307.2G	500G	1000G	307.2G	500G	1000G
最適値	10 μ s	10 μ s	10 μ s	500 μ s	1ms	100 μ s

次に、プリウェイクアップを実現した際のリンクの消費エネルギーを見積もった。図 2 に各システム構成における、通常のリンク、各システムの最適なリンク

オフスレッシュホールドを有する ON/OFF リンク、プリウェイクアップ手法を用いた際のリンクの消費エネルギーを示す。プリウェイクアップ手法を用いることで、最適なリンクオフスレッシュホールドの場合に比べ消費エネルギーを削減できることが分かった。特にノード数 8、計算性能 500GFLOPS の場合、エネルギー削減率は最大約 47%となっている。プリウェイクアップ手法を実現することでさらなる電力削減が期待できる。

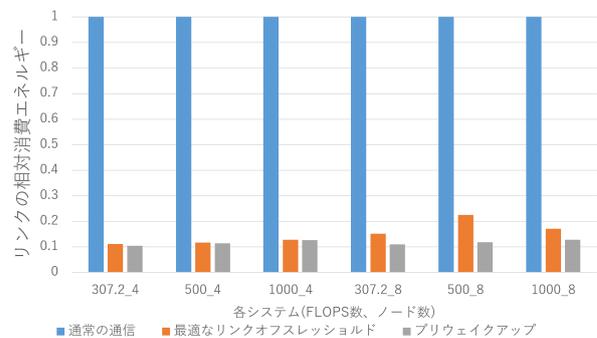


図 2: 各システム構成におけるリンクの消費エネルギーの関係

5 おわりに

NPB の通信トレースを解析し、最適なリンクオフスレッシュホールドを用いた場合でも、ON/OFF リンクが多くのエネルギーを浪費していることを明らかにした。今後、プリウェイクアップを実現する手段として、アプリケーションを書き換えて、本来のデータ通信が行われる直前にダミーデータ通信を行う方法を検討している。

参考文献

- [1] 三輪忍, 會田翔, 安島雄一郎, 清水俊幸, 安里彰, 中村宏 “実 HPC 環境における EEE の電力/性能評価” 情報処理学会論文誌コンピューティングシステム, Vol7, No.4, pp.67-83 (2014)
- [2] K.P.Saravanan, P.M.Carpentop, and A.Ramirez “Power/performance evaluation of energy efficient ethernet(EEE) for high performance computing” In Proceedings of the 2013 IEEE International Symposium on Performance Analysis of Systems and Software, pp. 205-214 (2013).
- [3] SIMGRID, <http://simgrid.gforce.inria.fr/>
- [4] NASA Advanced Supercomputing Division, <https://www.nas.nasa.gov/publications/npb.html>