C-043

Approximate Hybrid HPC ネットワークの研究
Study on Approximate Hybrid HPC Networks

藤原 一毅 [1]　藤木 大地 [2]　石井 紀代 [3]　松谷 宏紀 [2]　天野 英晴 [2]　鯉渕 道紘 [1]
Ikki Fujiwara, Daichi Fujiki, Kiyo Ishii, Hiroki Matsutani, Hideharu Amano, Michihiro Koibuchi

1. 概要

我々は、スーパーコンピュータやデータセンターなどの HPC システム向けに、超広帯域・低遅延である代わりにエラーフリー伝送を保証しない approximate ネットワークを提案している。HPC システムの相互結合網に用いられる InfiniBand や 100G Ethernet は規格上、10^{-12}以下という極めて低いビット誤り率（BER: Bit Error Rate）を要求する。これ以上の広帯域化には前方誤り訂正（FEC: Forward Error Correction）の導入が不可欠だが、FEC はバッファリングを要するため 1 ホップ毎に 100 ナノ秒程度の遅延増加を伴う[3]。InfiniBand QDR スイッチの遅延が現行製品でも 100 ナノ秒程度であることから、FEC に伴う遅延増加は並列処理性能に大きな影響を及ぼすと考えられる。一方、数万ノードからなる大規模 HPC システムでは、メモリやストレージにおける偶発的ビット反転など、計算結果に影響を及ぼす誤差が無視できない頻度で発生する[1,2]。このため、並列アプリケーションにおいて誤差の影響を避ける方法や、逆に誤差を許容（approximate computing）して高速化・低消費電力化する方法が広く研究されている。このような状況でネットワークだけがエラーフリーに固執する必然性はなく、むしろ帯域幅と誤り率のトレードオフを探る余地があると我々は考えた。すなわち、アプリケーションが伝送エラーを許容することを前提として、多値変調を導入して広帯域化を図り、なおかつ FEC を省略して通信遅延の増加を避ける。これを本報告では approximate ネットワーク（以下 Approx NW）と呼ぶ。図 1 に従来型ネットワークと Approx NW の構成を示す。

2. 背景と前提条件

光通信では、光の強度を 2 値で変調する OOK (On-Off Keying) 方式が広く使われている。100G Ethernet 規格では OOK 信号を用いて 10Gbps×10 レーンまたは 25Gbps×4 レーンの構成が採用されている。これを本報告では従来型ネットワーク（Conventional NW）と呼ぶ。一方、長距離系の 100G OTN (Optical Transport Network) では位相偏移を 4 値に変調する QPSK (Quadrature Pulse Shift Keying) 方式に加えて FEC による誤り訂正が採用され、エラーフリー通信が達成されている。これを本報告では FEC Perfect NW と呼ぶ。

本報告では、物理層は電気スイッチ同士をアクティブ光ケーブルで接続したネットワークを想定し、電気スイッチチップの広帯域化は今後も続くと考える。ネットワークはアプリケーションから渡されるビット列（浮動小数点数であれば IEEE 754 形式）をグレイコードに変換してシンボルにマッピングする。

1: 国立情報学研究所
2: 慶應義塾大学
3: 産業技術総合研究所

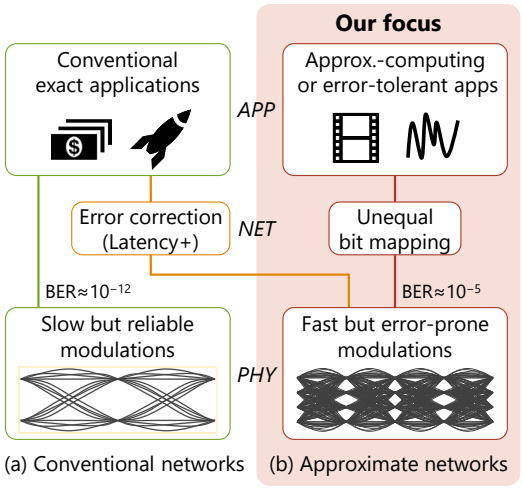

図 1　従来型ネットワークと approximate ネットワーク

3. Approximate ネットワークの構成

変調方式として、光の強度と位相偏移を多値変調する直角位相振幅変調 (QAM: Quadrature Amplitude Modulation) を採用する。QAM は 1 シンボルあたりの状態数を任意に設定することで帯域幅とノイズ耐性のトレードオフを調整できる。例えば、16-QAM では 1 シンボルあたり 4 ビット、256-QAM では 1 シンボルあたり 8 ビットを伝送できるが、状態数が増えるほど誤り率も高くなる。本報告では 25Gbps×4 レーンの 100G Ethernet と同じ変調レートを想定し、1 シンボルあたり 10 ビットを伝送する 1024-QAM を用いて、250Gbps×4 レーン＝最大 1Tbps のリンクを構成する。本報告では 1024-QAM の BER を10^{-5}と想定する。

帯域幅と誤り率のトレードオフは、1 シンボルあたりの状態数を間引くことで調整できる（図 2）。このとき変調方式そのものを切り替える必要はなく、ビット列からシンボルへのマッピングを変更するだけでよい。とはいえ、アプリケーションにとっては誤り率よりも数値誤差を制御できる方が望ましい。そこで、アプリケーション側に Approximate データ型[5]を追加し、保護されるべきビット幅 p をデータ型ごとに指定できるよう通信ミドルウェアを

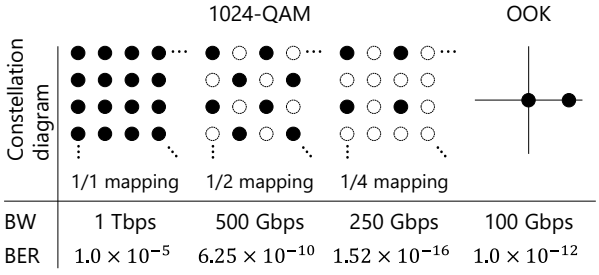

図 2　帯域幅と誤り率のトレードオフ（1024-QAM の信号空間ダイアグラムは左上の 16 個を抜き出して示す）

拡張する。ネットワーク側では、当該ビット列の上位 p ビットまでをエラーフリーで（間引いたマッピングで）伝送する。例えば、倍精度浮動小数点数型は上位 16 ビットを 1/4 に間引き、残り 48 ビットを間引かずに伝送する。これにより、数値の伝送精度を保証しつつ、帯域幅の減少を最小限に抑えることができる。

4. 評価

表 1 に示す 3 種類のネットワーク上で走る並列アプリケーションの性能をシミュレーションによって評価する。シミュレータは SimGrid (v3.12) [4] に前述のビット保護機構を追加し、さらに光リンク上の加法性ホワイトガウスノイズを模擬するよう改造したものを用いる。通信ミドルウェアは SimGrid 組み込みの MVAPICH2 を用いる。ネットワークは 256 スイッチからなる Dragonfly トポロジ、スイッチ遅延は 60 ナノ秒／台、ケーブル遅延は 25 ナノ秒／本、ルーティングは最短経路とする。各スイッチには 500 GFlops の計算ノードが接続されている。

アプリケーションは MPI 版 NAS Parallel Benchmarks (v3.3.1) の CG と FT、および K-means クラスタリングアルゴリズムを用いる。CG は MPI_DOUBLE 型のメッセージ 21,015,549 個のうち 399,360 個を approximate 伝送し、上位 16 ビットを保護する。FT は MPI_DOUBLE_COMPLEX 型メッセージ 2,882,519 個を approximate 伝送し、実数部・虚数部それぞれ上位 16 ビットを保護する。K-means は 9 次元空間にある 11,500,000 個の点を 11 個のクラスタに分類する問題であり、MPI_FLOAT 型のメッセージ 95,805,592 個を approximate 伝送し、上位 12 ビットを保護する。

図 3 は、各ベンチマークの実行時間（従来型ネットワークを 1 とした相対値）である。Approx NW の利用により、従来型ネットワークに比べて CG で 54%、FT で 66%、K-means で 43%、それぞれ実行時間が減少した。FEC Perfect NW と比較すると、実行時間が CG で 23%、FT で 33%減少した反面、K-means では実行時間が 66%増加した。このように、通信遅延に敏感でない（うまく隠蔽されている）アプリケーションや、帯域幅の増加がより重要なアプリケーションでは、Approx NW の利点が表出しないと考えられる。なお、Approx NW は FEC と共存可能であり、アプリケーションの性質によって切り替えて使うこともできる。

CG の解は従来型ネットワークと Approx NW とで同じ値に収束した。FT の解の質は保護ビット幅に依存する（図 4）。保護ビット幅が大きいほど実行時間が長く、結果の誤差が小さくなることが確かめられた。特に、浮動小数点数の符号ビットを含む上位 8 ビットを保護することは必須と言える。K-means は従来型ネットワークと Approx NW とで 337 個（0.003%）の点が異なるクラスタに分類された。

	Bandwidth	Sw. Latency	BER
Conv. NW	100 Gbps (OOK)	60 ns (w/o FEC)	Error-free
FEC Perfect NW	1 Tbps (1024-QAM)	160 ns (w/ FEC)	Error-free
Approx. NW	≤1 Tbps (1024-QAM)	60 ns (w/o FEC)	$\geq 10^{-5}$

表 1　評価パラメータ

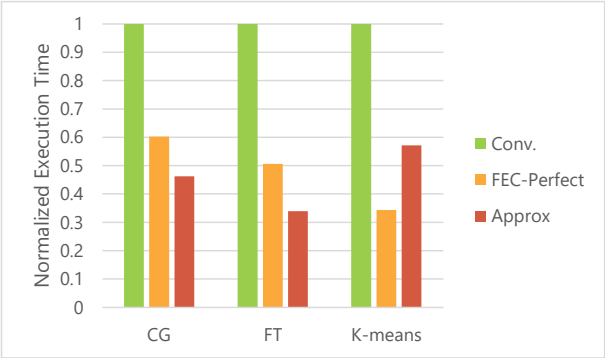

図 3　実行時間の比較

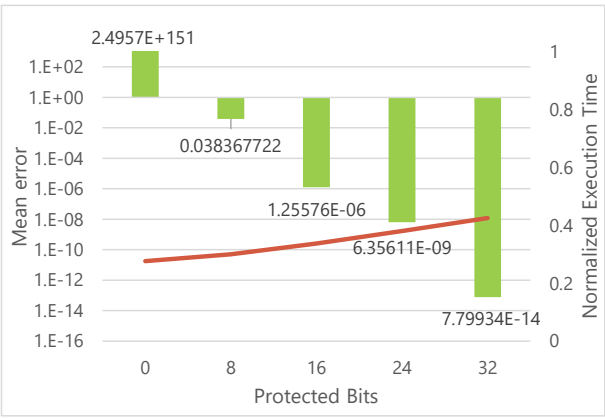

図 4　倍精度浮動小数点数の保護ビット幅（横軸）と FT の平均誤差（緑棒、左目盛り）および実行時間（赤線、右目盛り、従来型ネットワークを 1 とした相対値）

5. まとめ

本報告では、超広帯域・低遅延である代わりにエラーフリー伝送を保証しない approximate ネットワークの提案とシミュレーションによる評価を行った。その結果、帯域幅とエラー率のトレードオフというコンセプトが実現できる見通しを得た。今後は FPGA を用いて approximate NIC を試作し、ハードウェア的な実現性の立証を図りたい。

謝辞

本研究の一部は総務省 SCOPE、JSPS 科研費 16H02816 の支援による。

参考文献

[1] F. Cappello, A. Geist, W. Gropp, S. Kale, B. Kramer, and M. Snir, “Toward exascale resilience: 2014 update,” Supercomputing frontiers and innovations, vol. 1, no. 1, 2014.
[2] A. Dixit and A. Wood, “The impact of new technology on soft error rates,” in IEEE International on Reliability Physics Symposium (IRPS), 2011, pp. 5B.4.1–5B.4.7.
[3] M. Andrewartha, B. Booth, and C. Roth, “Feasibility and Rationale for 3m no-FEC server and switch DAC,” http://www.ieee802.org/3/by/public/Sept15/andrewartha_3by_01a_0915.pdf, 2015.
[4] SimGrid: Versatile Simulation of Distributed Systems, http://simgrid.gforge.inria.fr/.
[5] A. Sampson, W. Dietl, E. Fortuna, D. Gnanapragasam, L. Ceze, and D. Grossman, “EnerJ: Approximate data types for safe and general low-power computation,” in ACM SIGPLAN Notices, vol. 46, no. 6. ACM, 2011, pp. 164–174.