RC-007

大規模グラフ解析を高速化する適応ネットワークシステム

Adaptive Network System for Accelerating Large-Scale Graph Analysis

高田 雅士† 林 真人† 朝 康博† 加藤 猛†
Masashi Takada Masato Hayashi Yasuhiro Asa Takeshi Kato

1. まえがき

近年，ソーシャルネットワークや都市の交通流といった実社会から得られる大規模なデータをグラフとして表現し，そのグラフを解析してデータに潜む有用な知見を得ることへの関心が高まっている．このグラフ解析は，その規模の大きさから一般にサーバクラスタによる並列処理で実現されている．例えば，2010 年に発表されたグラフ処理のベンチマークである Graph500[1]の上位ランキングには，数千から数万サーバによる並列処理システムが登録されている．

これらの並列処理システムではトーラス，メッシュ，Fat tree といった規則的なネットワークトポロジが利用されており，充分な数のリンクを備えることで通信性能を向上させている．例えば，前述の Graph500 ベンチマークで最新(2012 年 11 月)のランキング上位に位置している Blue Gene/Q[2]は 5 次元トーラス，京コンピュータ[3]は 6 次元メッシュ/トーラス，TSUBAME2.0[4]は Fat tree で構成されている．

並列グラフ解析はサーバ間で大量の通信が発生するネットワーク負荷の高い処理であり，今後扱うグラフの規模が増大することを考えると更なる通信性能の向上が必須となる．しかし，要求性能を満たすために充分な数のリンクを追加する従来の方法では，更に多くのリソースを投入していかなければならない．

これに対し，リングのような低次元なトポロジにランダムなショートカットリンクを加えた不規則なトポロジを利用することで通信性能を向上させる方法が鯉渕ら[5]により提案されている．このような不規則なトポロジを利用する方法ではネットワークの効率化が図られるため，リソース追加なしでも通信性能の向上が期待できる．特に，低次元なトポロジであっても問題に合わせて適切なトポロジを選択することができれば，その効果は非常に大きくなると考えられる．

以上の背景を踏まえ，我々は対象とするグラフ毎にサーバクラスタのネットワークトポロジを変更することでグラフ解析の高速化を図る適応ネットワークシステムを提案する．本報告では，まず，適応ネットワークシステムの概要を述べ，その処理フローについて説明する．次に，適応ネットワークの主要技術であるトポロジの探索・評価手法について述べる．その後，サーバクラスタ上でのトポロジ切り替えとグラフ解析の実験結果を示し，適応ネットワークの効果について考察する．なお，本報告では適応ネットワークと従来方式とのグラフ解析処理性能の比較を主とし，トポロジ探索処理性能に関しては今後の課題に譲る．

†(株) 日立製作所 中央研究所
Central Research Laboratory, Hitachi, Ltd.

2. 適応ネットワークシステム

2.1.概要

適応ネットワークはグラフ解析時のサーバ間の通信パターンに合わせたネットワークトポロジを構成する技術であり，様々なグラフに対する通信時間の削減を目的としている．図 1 に適応ネットワークシステムの概要を示す．適応ネットワークの主な処理内容は (1)グラフをクラスタに分割してクラスタ間の通信パターンを見積もること，(2)見積もった通信パターンに対して最適なトポロジを探索すること，(3)サーバクラスタのネットワークをそのトポロジへ切り替えることである．本報告では(2)のトポロジ探索を中心に説明する．

なお，適応ネットワークシステムでは，(3)のトポロジ切り替えの際に手作業によるケーブリングが発生しないよう，光スイッチの利用を想定している．光スイッチはプログラムにより内部の伝送路を切り換えてポート間の接続を変更するものであり，スイッチ間の接続に用いることでネットワークトポロジを高速かつ自由に変更できる．

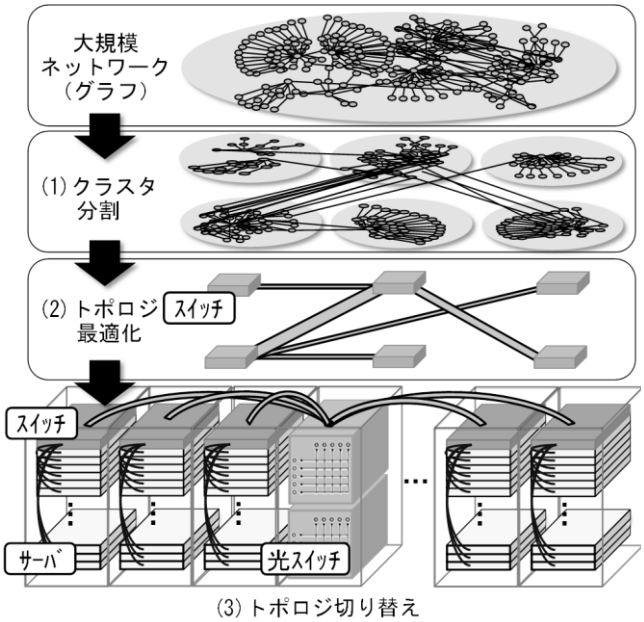

図 1 適応ネットワークシステム概要

2.2.処理フロー

適応ネットワークシステムにおけるグラフ解析は，グラフに適したトポロジを事前に発見するトポロジ探索フェーズとトポロジを切り替えて実際に解析を進めるグラフ解析フェーズから構成される(図 2). 探索フェーズは解析フェーズよりも一般に処理時間オーダが大きいこと，解析フェ

ーズはグラフの特徴を調べるために複数回実行されることから，フェーズ毎に独立な処理フローとしている．なお，両フェーズの処理はサーバクラスタ上で実行されることを想定している．

トポロジ探索フェーズでは，まず，(1)解析対象となるグラフデータとサーバ数を入力としてグラフ頂点をどのサーバに配置するかを決定する．グラフ解析時間を短縮するには，各サーバの計算時間が均等になり，かつ全体のサーバ間通信時間が少なくなるようなクラスタ分割が望ましい．本研究では，クラスタ分割手法として，グラフ処理フレームワークの Pregel[6]にも採用され，広く利用されているサイクリック分割を用いる．サイクリック分割は，ランダムな順番でグラフ頂点に連続した整数番号をつけ，番号をクラスタ数で割った際の余りで割当先を決定する手法である．クラスタ分割後は配置結果をクラスタファイルとして出力し，これに基づいてサーバ間通信量を見積もる．通信量はサーバ間を跨ぐグラフ辺数×辺あたりのメッセージサイズとして算出する．次に，(2)通信量見積もりに対してグラフ解析時間が最小となるトポロジを探索する．この際，スイッチ数，スイッチのポート数，リンク数，リンク帯域といったハードウェア構成を制約条件とする．探索結果のトポロジはトポロジファイルへ出力される．

グラフ解析フェーズでは，(3)トポロジファイルに基づいて光スイッチを制御しネットワークトポロジを変更した後，グラフデータとクラスタファイルを入力としてグラフ解析を実行する．

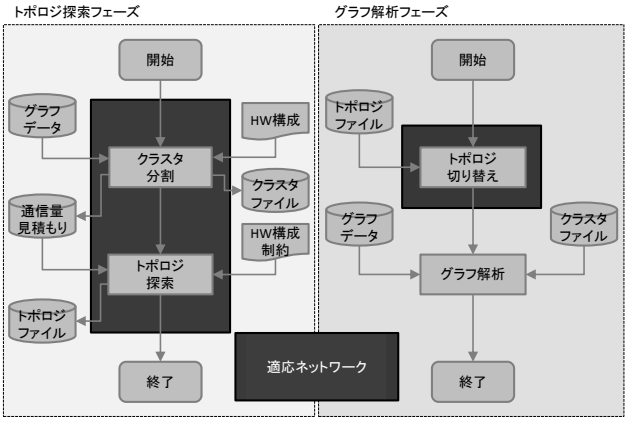

図 2 適応ネットワーク処理フロー

2.3. トポロジ探索

サーバ間通信量見積もりに対して通信時間が最小となるネットワークトポロジを探索することは NP 困難な問題である．そこで，本研究ではメタヒューリスティックな手法である遺伝的アルゴリズム(GA)を利用してその近似解を求める．

GA を適用するにあたり，トポロジを図 3 に示す染色体として表現した．各遺伝子はスイッチ間の接続情報であり，1 本の染色体が 1 個のトポロジを表す．図中の src は接続元のスイッチ番号を，dst は接続先のスイッチ番号を意味する．例えば src 番号 0 の位置にある dst 番号 1,2 の遺伝子は，スイッチ 0 がスイッチ 1 とスイッチ 2 の間にリンクを持つことを示す．リンクは双方向であり，スイッチ間の接続を表現するには接続元と接続先の情報は不要である．しかし，ルーティングパスの決定にスイッチのポートへの接続順を利用するルーティングアルゴリズムがあるため，接続元と接続先の情報からポートの接続順を表現した．例えば，src 番号 1 の位置にある dst 番号 3,0 という遺伝子から，スイッチ 1 は 1 番ポートでスイッチ 3 へ接続し，2 番ポートでスイッチ 0 へ接続する，という情報が得られる．この遺伝子表現の結果，1 本のリンクは src 番号と dst 番号を入れ替えた形で染色体上に 2 回現われるため，遺伝子の数がリンク数の倍となっている．

なお，この染色体では他スイッチとのリンクを持たない孤立したスイッチや自スイッチのポート間を繋ぐループといった不適切な解も表現できてしまうため，初期世代の生成，交叉，突然変異といった遺伝子操作の際にスパニングツリーを含み，かつ dst 番号が src 番号と一致しないよう調整し，解空間の探索効率を低下させないようにしている．

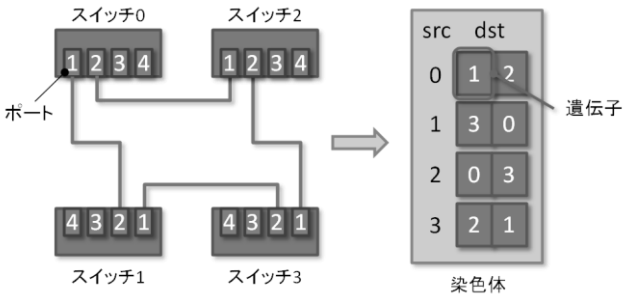

図 3 トポロジの染色体表現

染色体の評価には式(2.1)で計算される最大通信時間の見積もり値 f(t)を用いる．式(2.1)の $data_i$ はリンク i を通るデータの総量であり，全サーバ間の通信量とルーティングテーブルから計算される．なお，本見積もり式は，ネットワーク性能に起因したボトルネックが発生する際の通信時間を表現したものであり，CPU 性能などの他の要因に起因したボトルネックが発生する場合はその影響を加味する必要がある．

$$f(t) = \max_{link_i \in link} \frac{data_i}{link_i} \tag{2.1}$$

$link_i$: リンク i の帯域[Gbps]
$data_i$: リンク i を通るデータ総量[Gbit]

3. 性能評価

3.1. 通信時間の見積もり精度の評価

GA で探索したトポロジは通信時間の見積もり式(2.1)に基づいて優劣が判定されるため，その見積もり値が実際のグラフ解析における通信時間を適切に反映していなければならない．そこで，見積もりの精度を知るために，幅優先探索を行う Graph500 ベンチマーク(MPI 実装版)を用いて実測との比較を行った．評価環境として 64 台のサーバ，18 台のスイッチ，1 台の光スイッチから成るサーバクラスタを用い，ネットワークトポロジは 2 次元トーラスと Fat tree を切り替えて測定した(図 4)．ネットワークのリンク帯域は 2.5Gbps とし，そのルーティングアルゴリズムには不規

則なトポロジにも適用可能な LASH(LAyered SHortest path routing)[7]を用いた．測定に使用したデータは Graph500 ベンチマークで生成した Scale27(グラフ頂点数 2^{27})のグラフである．

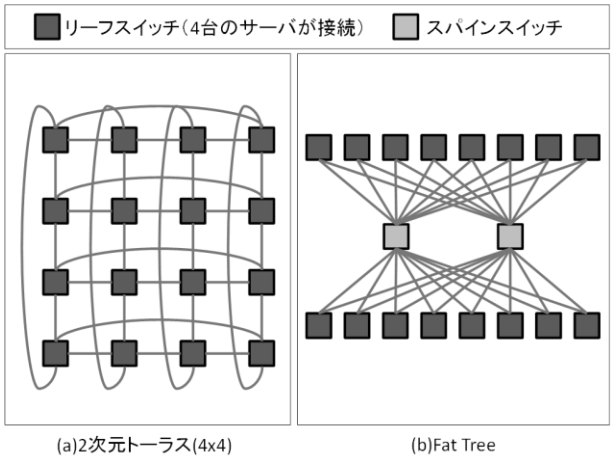

図 4 評価に利用したネットワークトポロジ

図 5 に Graph500 ベンチマークの通信時間の見積もりと実測を示す．2 次元トーラスでは見積もりの 15.1 秒に対して実測が 23.7 秒，Fat tree では見積もりの 12.2 秒に対して実測が 19.8 秒という結果が得られた．実測に対する見積もりの割合は各々63.8%，61.5%である．式(2.1)による簡易な見積もりのため，実測と見積もりの差は大きいが，異なるトポロジ間での見積もりのばらつきは小さい．以上から，見積もりと実測でトポロジ間の優劣が逆転する可能性は低く，本見積もりに基づいてトポロジの優劣を判定することに問題はないと言える．

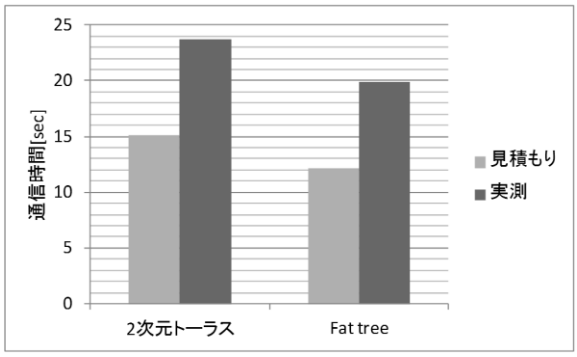

図 5 Graph500 ベンチマークの通信時間の見積もりと実測

3.2.GA で発見したトポロジの評価

図 6 に 3.1 節で述べた評価環境，データセットに対して，2 次元トーラス，Fat tree，GA で発見したオリジナルトポロジを切り替えて Graph500 ベンチマークを動作させた結果を示す．本実験ではトポロジの差異による通信時間への影響のみを評価するために，上記トポロジはスイッチ間のリンク数を同一の本数(32 本)で構成した．GA で発見したオリジナルトポロジの通信時間は 17.1 秒であり，2 次元トーラスの 23.7 秒に対して 27.9%，Fat tree の 19.8 秒に対して 14.0%削減することができた．このことから，GA によって Scale27 のグラフに適したトポロジを発見することができたと考える．

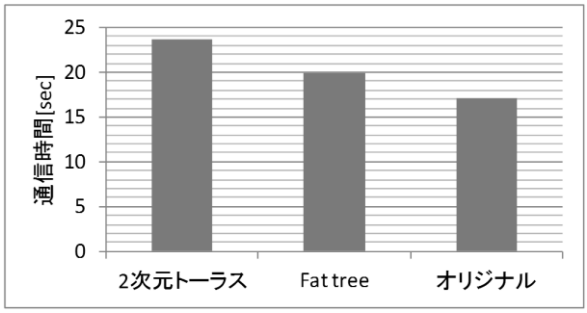

図 6 Graph500 ベンチマークの通信時間

なお，トポロジ探索に用いた GA の設定と遺伝子操作のパラメータは以下の通りである．

- スイッチ数…16，リンク数…32，リンク帯域…2.5Gbps
- 1 世代の染色体数 … 40
- 選択 … エリート選択 + ルーレット選択
- 交叉 … 2 点交叉
- 突然変異率 … 2%
- 終了条件 … 最良の染色体が 100 世代更新なし

3.3.光スイッチのリンク切り替え時間の評価

ネットワークトポロジの切り替えでは，図 7 に示すように，①テンポラリリンクの追加，②旧トポロジ用リンクの削除，③新トポロジ用リンクの追加，④テンポラリリンクの削除といった多数のリンク操作が発生する．テンポラリリンクはトポロジを切り替える際にスイッチがネットワークから切断されて制御不能に陥るのを防ぐためのものである．これらのリンク操作は適応ネットワークのオーバヘッドとなるため，その処理時間を把握しておく必要がある．

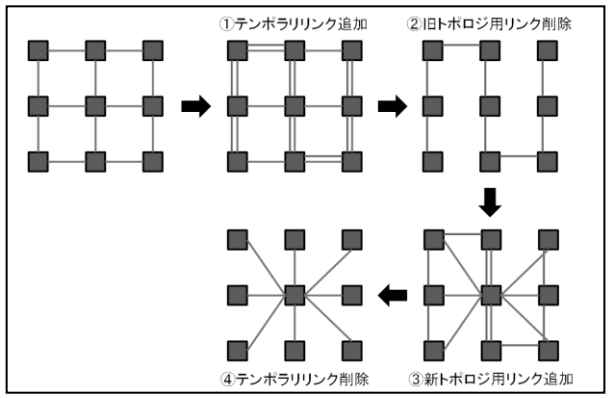

図 7 トポロジ切り替えにおけるリンク操作

図 8 に本評価で用いた光スイッチについて，16 本から 80 本までの範囲でリンクの追加と削除にかかる時間をプロットした結果を示す．リンクの切り替え時間は，光スイッチへのリモートログインなどに要する時間とリンクの追加や削除に要する時間から成る．前者は操作するリンクの本数には依存せず 1.3 秒である．後者は操作するリンクの本数に比例し，1 本あたり追加時には 200 ミリ秒，削除時には 80 ミリ秒であった．

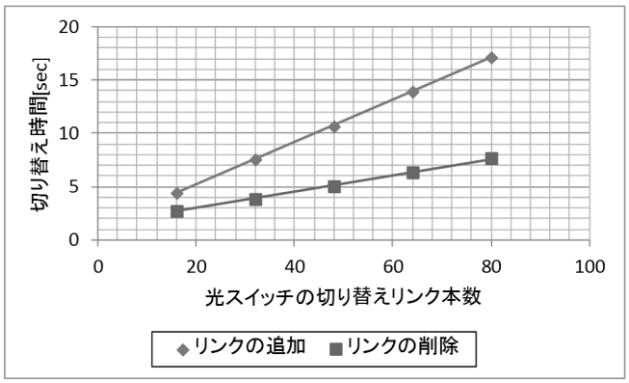

図 8 光スイッチのリンク切り替え時間

3.4.適応ネットワークの評価

3.1 節で述べた評価環境を用いて 3 種類のグラフに対し連続して Graph500 ベンチマークを実行する実験を行った．本実験では，Graph500 ベンチマークで生成したグラフ(g500)の他に，SNAP[8]で生成したグラフを用いた．SNAP は Stanford 大学が開発したグラフ解析・操作用のツールであり，入力のパラメータを指定することで異なる形状のグラフを生成することができる．ここでは文献[9]に記載のパラメータを利用して 2 種類のグラフを生成した(flickr, gnutella-30)．これらのグラフの規模は Scale26～28 とした．Graph500 ベンチマーク実行時のトポロジは，(a)2 次元トーラス， (b) Fat tree，(c)各グラフに適したトポロジへその都度切り替える適応ネットワークとした．これらのトポロジのスイッチ間のリンク数は 3.2 節と同様に全て 32 本である．

図 9 の測定結果から，適応ネットワークを利用することで 2 次元トーラスに対して 28.5～30.1%，Fat tree に対して 14.7～20.7%グラフ解析時間が削減され，適切なトポロジを選択することはグラフ解析時間の削減に有効であることが分かった．なお，1 回のトポロジ切り替え時間は 16.2 秒であり，これはグラフ解析で削減した時間より長く，全体の処理時間が削減されるという期待した結果は得られなかった．しかし，グラフの規模が 2 倍になる(Scale が 1 増加する)とグラフ解析時間もほぼ 2 倍になるという傾向から，規模増大に伴いトポロジ切り替えのオーバヘッドは相対的に小さくなる．本実験ではサーバクラスタのメモリ搭載量の制限上試すことができなかったが，より大規模なグラフを解析する際に適応ネットワークは有効な技術であると言える．

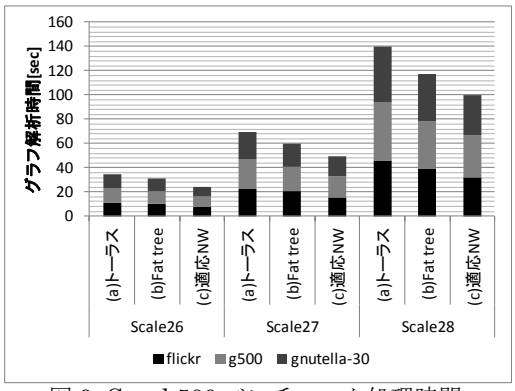

図 9 Graph500 ベンチマーク処理時間

4. まとめ

グラフ毎にサーバクラスタのネットワークトポロジを変更することでグラフ解析の高速化を図る適応ネットワーク技術を提案し，クラスタ間の通信パターンから通信時間を見積もり，トポロジを探索する手法を検討した．光スイッチを用いてネットワークトポロジの可変なサーバクラスタを構成し，適応ネットワークの効果を評価した結果から，大規模グラフの解析において，その通信時間の削減に有効な技術であるという結論を得た．

今後の課題は，クラスタ分割手法の検討，探索フェーズを含めた適応ネットワーク処理フロー全体の評価，小規模な問題やグラフ以外の問題への適用などである．例えば文献[10]のような数ナノ秒程度の切り替え時間の光スイッチを利用すれば，本技術のオーバヘッドは大幅に削減できる見込みである．

参考文献

[1] Graph500. http://www.graph500.org/
[2] Dong Chen, Eisley, N. A., Heidelberger, P., Senger, R. M., Sugawara, Y., Kumar, S., Salapura, V., Satterfield, D. L., Steinmacher-Burow, B. and Parker, J. J.: The IBM Blue Gene/Q interconnection network and message unit. International Conference for High Performance Computing, Networking, Storage and Analysis (SC), Nov. 2011.
[3] Ajima, Y., Takagi, Y., Inoue, T., Hiramoto, S. and Shimizu, T.: The Tofu Interconnect. IEEE 19th Annual Symposium on High Performance Interconnects (HOTI), pages 87–94, 2011
[4] Maruyama, N., Nomura, T., Sato, K. and Matsuoka, S.: Physis: An implicitly parallel programming model for stencil computations on large-scale GPU-accelerated supercomputers. International Conference for High Performance Computing, Networking, Storage and Analysis (SC), Nov. 2011.
[5] Koibuchi, M., Matsutani, H., Amano, H., Hsu, D. F. and Casanova, H.: A Case for Random Shortcut Topologies for HPC Interconnects. Proc. of the International Symposium on Computer Architecture (ISCA), pages 177–188, 2012.
[6] Grzegorz M., Matthew H. A., Aart J. C. B., James C. D., Ilan H., Naty L. and Grzegorz C.: Pregel: A System for Large-Scale Graph Processing. Proc. of ACM SIGMOD International Conference on Management of data, pages 135–146, 2010.
[7] Lysne, O., Skeie, T., Reinemo, S.-A. and Theiss, I.: Layered routing in irregular networks. IEEE Transactions on Parallel Distributed System, vol. 17, pages 51–65, 2006.
[8] SNAP. http://snap.stanford.edu.
[9] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, Z. Ghahramani: Kronecker Graphs: An approach to modeling networks, Journal of Machine Learning Research (JMLR), Vol. 11, pages 985–1042, 2010.
[10] Joris V. C., William M. G., Solomon A. and Yurii A. V.: Low-power, 2x2 silicon electro-optic switch with 110-nm bandwidth for broadband reconfigurable optical networks. Optics Express, Vol. 17, pages 24020–24029. 2009.