様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　２５年　５月２０日現在

機関番号：３２６９２
研究種目：基盤研究（Ｃ）
研究期間：2010～2012
課題番号：２２５００００５２
研究課題名（和文）　超並列コンピュータ向け高効率通信アルゴリズム開発・評価手法の研究

研究課題名（英文）　A study of development and evaluation method for efficient communication algorithm used in massively parallel computer

研究代表者
石畑　宏明（ISHIHATA HIROAKI）
東京工科大学・コンピュータサイエンス学部・教授
研究者番号：９０４６８８８５

研究成果の概要（和文）：本課題では，超並列コンピュータ向けの通信アルゴリズム開発環境として，高性能なネットワークシミュレータとその結果の可視化ツールを開発し，通信アルゴリズム開発環境のプロトタイプシステムを構築した．大規模なネットワークでも少ないリソースで高速にシミュレーション可能となり，シミュレーション結果は，通信ネットワークのトポロジ上にマップして分かりやすく表示した．

研究成果の概要（英文）：In this study, we developed network simulator and visualization tool for efficient communication algorithm development. The system enables high speed simulation of large scale network and provides easy understanding of communication behavior by mapping statistic data of network on 3D view of network topology.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,100,000	330,000	1,430,000
２０１１年度	700,000	210,000	910,000
２０１２年度	600,000	180,000	780,000
年度			
年度			
総　計	2,400,000	720,000	3,120,000

研究分野：総合領域
科研費の分科・細目：情報学，計算機システム・ネットワーク
キーワード：ハイパフォーマンスコンピューティング，ネットワーク，可視化

１．研究開始当初の背景

並列化されたアプリケーションで頻繁に使用される全対全通信などの通信アルゴリズムを設計する際に，従来はネットワークのトポロジ毎に輻輳の影響を考慮するようなことはなかった．将来の超並列コンピュータでは，数万～数十万のノードコンピュータを通信ネットワークで接続する必要がある．このため，Mesh や Torus といった低コストではあるが通信時に輻輳が多く発生しやすい通信ネットワークを選択せざるを得ない．このような性能の低い通信ネットワークを効率よく利用するために，ネットワーク中での輻輳の影響を考慮に入れた通信アルゴリズムの設計が望まれる．このためには，設計した通信アルゴリズムに従ってノード間通信を行ったときに，メッセージがネットワーク中でどのように振る舞うのかを正しく認識し，通信時間に与える影響を精度よく評価することが必要となる．

２．研究の目的

本研究では，図１に示すようなネットワークシミュレータと可視化ツールを組み合わせた通信アルゴリズム評価環境を開発・構築し，その効果を実証する．

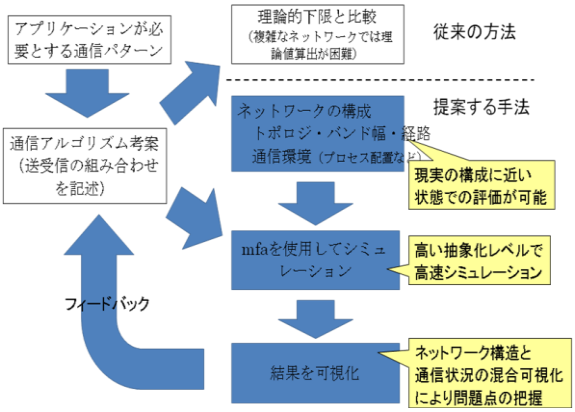

図１．通信アルゴリズム開発環境

３．研究の方法

（１）これまでの研究

Mesh や Torus に対しては，古くから全対全通信などの特定の通信パターンに対して種々の通信アルゴリズムが提案されている．我々も Mesh 上で理論的下限の時間で全対全通信が実行できるアルゴリズムを提案している．

従来，考案した通信アルゴリズムの評価は，通信経路のバンド幅とそこを通過するデータ量から通信時間の理論的下限を解析的に求め，考案したアルゴリズムでの通信時間と比較することで行われてきた．理論値との比較を行うため，正方形の Mesh や Torus などの通信時間の理論値を算出しやすいシンプルな構成のネットワークを対象にしてきた．

しかし，現実に使用されるネットワークは，長方形の Mesh や Torus であったり，Torus でも特定の次元だけが Mesh になっていたり，一部にバイパス経路を持つなど，単純なものばかりではない．これまで提案された通信アルゴリズムは，このような構成での性能評価は対象外であった．

ネットワークの輻輳の通信性能への影響に関する研究は，従来からネットワークシミュレータを利用して行われてきた．複雑なネットワーク上での通信アルゴリズムの評価にもこれを活用することができる．しかし，これまでのネットワークシミュレータは、ノード間の通信メッセージを複数のパケットに分割しその動きを追跡する方式をとっていた．パケットの移動ごとにイベントを発生させ処理するため，数十万ノード規模でのシミュレーションでは，長大な時間がかかる．

（２）本研究の学術的な特色・独創的な点

ネットワークシミュレータ mfa は，通信メッセージを構成する個々のパケットを追跡するのではなく，与えられた通信パターンの通信を実行したときの流量から通信時間を求める．この方法によって，シミュレーション時間と扱える規模の問題を解決し，従来のパケットレベルシミュレーションでは，その規模の大きさから対応できなかったような大規模な通信ネットワークを解析可能にする．

メッセージの流量に着目し，ネットワークの解析を行うアイディアは，極めて独創的である．同様のアプローチとしては、TCP 通信のシミュレーションの例があるが，スーパーコンピュータで使用されるような，大規模・低レイテンシのネットワークに対して適用したものは未だない．

通信アルゴリズムを改善していく過程では，通信のどの部分によって性能が劣化しているかを迅速に判断することが必要となる．与えられた通信アルゴリズムの輻輳発生時の振る舞いを明らかにするためには，通信の状況を可視化できることが重要である．通信状況の可視化に関しては，メッセージ通信の送信元・受信先の関係を時系列に表示する機能が種々開発・利用されている．通信のボトルネックがネットワークトポロジ上のどの場所で発生しているかを把握するためには，ネットワークのトポロジと輻輳の状況を関連付けて三次元表示する可視化機能が望まれる．

４．研究成果

（１）新方式のネットワークシミュレータの開発

研究代表者は，ノード間通信のフローに着目した新しい方式に基づいたネットワークシミュレータ（mfa)を提案した．mfa は，従来から行われていた通信パケットを追跡する方法と異なり，「システム内の全ノードについて２ノード間のメッセージのフローを算出し，システム全体を重ねあわせる」という方式に基づいており，大規模なネットワークを高速にシミュレーションできる．１万ノード以上の大規模な構成で，従来型のシミュレータの 1～2%の処理時間とメモリ使用量でシミュレーションが行え，大規模なネットワークのシミュレーションが現実的時間で行えることを示した．

また提案した方式の精度を確認するため，mfa と，従来から使われているパケット方式のシミュレータ Booksim を統計的手法により比較を行なった．mfa でネットワークの性能を見積もるうえでは，FatTree などの通信のホップ数の小さいネットワークでは比較的精度の高い結果が得られるが，辺の長さの長

い mesh・torus では，誤差が大きくなることが示された．ネットワーク中のスイッチでのアービトレーションモデルの違いによる差は，アービトレーションの段数が多くなるほど大きくなるためであると判明した．

この誤差を小さくする方法として，新たに「ポートアービトレーションモード」を検討した．この方法は，ネットワーク中のリンクを通過するメッセージのバンド幅を決定するのに，従来はそのリンクの最大バンド幅を，そこを通過するメッセージ数で割って近似していたものを，通常のスイッチが行うように送信元から順次メッセージを中継しリンクごとにそのバンド幅を決定していくようにしたものである．mfa の欠点である「通信のホップ数の小さいネットワークでは比較的精度の高い結果が得られるが，ホップ数の大きいトポロジでは，スイッチ内のアービトレーションの影響による誤差が大きくなる」を改善できる．この方法のプロトタイプを実装した．

（2）ネットワークシミュレーション結果の可視化

シミュレーション結果の可視化に関しては，ネットワークトポロジを比較的自然に表示する事が出来るメッシュ・トーラス系のトポロジに加え，FatTree トポロジの三次元表示手法を考案した．FatTree をそのままトポロジ表示すると，著しく横長になり全体の様子を把握するのが困難になる．そこで，FatTree の各階層をグループ化し，グループごとにまとめて 3 次元空間中に配置する方法を考案した．トポロジ上で近い場所にあるものは近く，遠いものは遠くに配置され，全体の状況把握を容易にした．

「ネットワークシミュレーションの結果を，通信ネットワークのトポロジ上にマップして表示する機能」の開発を行った．図２のように，Mesh や Torus・Fattree トポロジのネットワークを 3 次元表示し，通信パターンの進みに合わせて，ネットワーク内の個々のリンクの状態（通過メッセージ数や通信実効バンド幅）を表示する．これにより，映像から直感的に把握した通信の輻輳状況を数値で定量的に表示させることができるようになった．

さらに，通信の輻輳状況をより容易に把握できるようにするために，複数の表示方法を組み合わせて表示する機能を付加した．上記のネットワークトポロジ上に通信状況をマップして表示する方法に加えネットワーク中の通信状況の時系列表示（図３），ノード間の通信関連図（図４）などを統合して表示・操作できるようなシステムのプロトタイプを作成した．

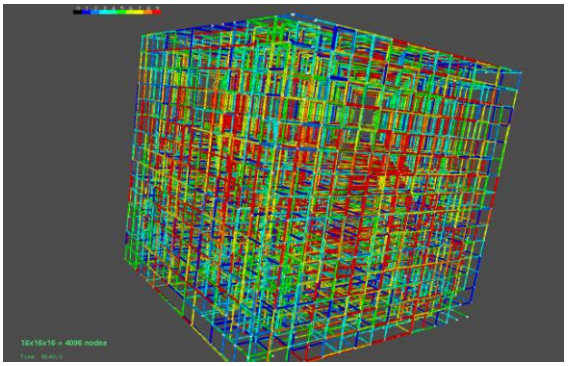

図２　通信状況のトポロジ上への表示

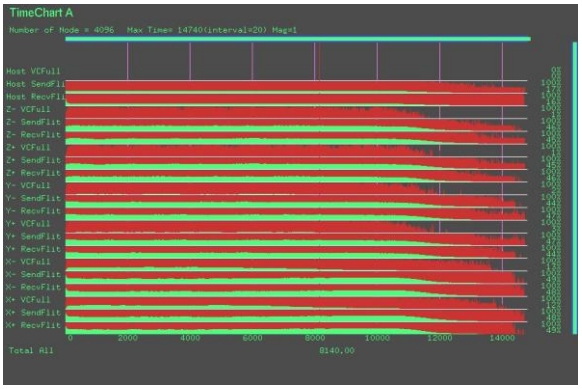

図３　ネットワーク通信状況の時系列表示

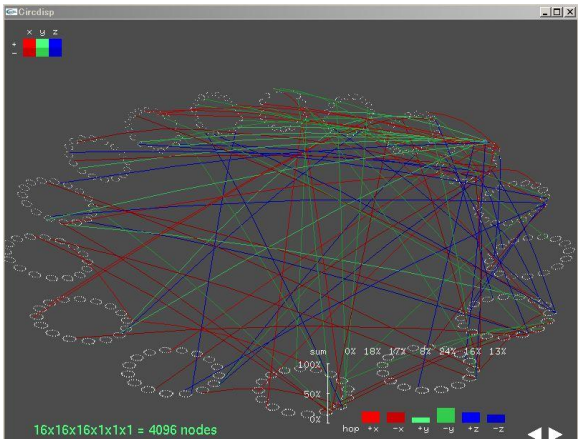

図４　ノード間通信関連図

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　５件）
①S. Yazaki, H. Takaue, Y. Ajima, T. Shimizu, and H. Ishihata, An Efficient All-to-all Communication Algorithm for Mesh/Torus Networks, The 10th IEEE International Symposium on Parallel and Distributed Processing with Application, 査読有, 1巻, 2012, pp.10-13
②鈴木 遼平, 石畑 宏明, 大規模並列計算機向け通信アルゴリズム開発環境の構築, 情報処理学会研究報告(HPC), 査読無, HPC-136, 2012, pp.1-6
③矢崎俊志, 石畑宏明, メッセージフローに基づくネットワークシミュレータMFSの評価, 情報処理学会論文誌(ACS), 査読有, Vol.4, No.3, 2011, pp.47-55
④矢崎 俊志, 石畑　宏明, 通信アルゴリズム評価用メッセージフローシミュレータの開発, 情報処理学会論文誌(ACS), 査読有, Vol.3 No.2, 2010, pp.76-87

〔学会発表〕（計　３件）
①R. Suzuki and H. Ishihata, Visualization Tool for Network Topology Aware Communication Algorithm Development, Supercomputing 2012 Research Poster, Nov. 11-17, 2012, Salt lake city , USA
②鈴木遼平, 石畑 宏明, 並列計算機の通信ネットワークトポロジの3次元表示手法, 第10 回情報科学技術フォーラム (FIT2011), Sep. 7, 2011, 函館大学(北海道)
③石井 省吾, 石畑　宏明, 大規模並列コンピュータを対象とした通信アルゴリズムの可視化ツールの設計, 情報処理学会第 73 回全国大会, Mar. 20, 2011, 東京工業大学（東京都）

〔その他〕
ホームページ等
http://www2.teu.ac.jp/his/

６．研究組織
(1)研究代表者
石畑　宏明　（ISHIHATA HIROAKI）
東京工科大学・コンピュータサイエンス学部・教授
研究者番号：90468885