科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ４ 日現在

機関番号：１２６１２
研究種目：若手研究(B)
研究期間：2012～2013
課題番号：２４７０００４６

研究課題名（和文）バック・アノテーションによる密結合型相互結合網の通信シミュレーション

研究課題名（英文）Communication Simulation on Interconnect for Tightly Coupled Cluster Systems Using the Back-annotation

研究代表者
矢崎　俊志（Yazaki, Syunji）

電気通信大学・情報基盤センター・助教

研究者番号：６０４５４１３８

交付決定額（研究期間全体）：（直接経費）　　2,100,000円、（間接経費）　　630,000円

研究成果の概要（和文）：並列処理向け通信シミュレーションの軽量化・高速化を目的として，新たなシミュレーション手法の開発を行った．これまで提案してきた軽量な通信シミュレータMFSの精度を改善すべく，VLSI設計で行われるバックアノテーションによるシミュレーション手法を取り入れ，実機のハードウェア特性を考慮した新しいシミュレーション手法を提案した．本研究ではMFSの通信調停機構を，より実機に近いモデルで実装した．バックアノテーションに用いる物理特性データを得るため，InfiniBnadスイッチの物理特性を測定した．評価のため，既存のシミュレータおよびスーパーコンピュータ実機で全対全通信を比較評価した．

研究成果の概要（英文）：A novel simulation method to provide right-weight and precise simulation for communication on distributed systems was proposed. We proposed a method which uses back-annotation technique which has been used for VLSI design. For back-annotation technique, the physical characteristics of the devices, such as delay and power consumption, will be measured and fed back to the simulation.
This technique enables to make the simulation result more precise. In this research, we first implemented a communication arbitrator modeling the physical communication arbitrator into a flow-based simulator, MFS (Message Flow Simulator). We also measured physical characteristics of an InfiniBand switches to obtain the data which will be annotated to the simulation. We compared all-to-all communication run on both simulator and actual super computer system.

研究分野：総合領域

科研費の分科・細目：計算機システム・ネットワーク

キーワード：ネットワークシミュレータ　並列計算　ハイパフォーマンス・コンピューティング

１．研究開始当初の背景

近年の社会基盤である情報ネットワーク（情報インフラ）の大規模化や科学技術の発展に必要不可欠な高性能スーパー・コンピュータなど，コンピュータ・システムへのより高い性能要求は日々高まっている．一方で集積技術の限界により，コンピュータによる処理の中心的な役割を担う演算装置単体の性能は頭打ちである．この問題を解決するため，近年では打規模システムのみならず，スマートフォンや個人用パソコンなどでも，1 つのコンピュータ・システム（機器）に複数の演算装置を組み込み，並列処理によって演算能力を高めている．

並列処理の方式には様々あるが，大きく共有メモリ型と分散メモリ型に分けられる．前者は，並列処理をしている演算装置が共にアクセスできる共有領域（共有メモリ）に保管された同じデータを参照しながら処理を進める．後者は，各演算装置に，専用の独立したメモリを用意し，様々な種類のネットワークを介して，メモリ間でデータの交換を行いながら処理を進める．

世の中で関心の高い大規模な問題を処理する場合には，分散メモリ方式を選択せざるを得ない場合が多い．共有メモリ方式はシンプルな方法である反面，十分な大きさの共有メモリを用意するのが難しい場合が多いためである．

近年のシステムは，個人用パソコンであっても，4 から 6 個程度の演算装置を搭載していることが多いが，今後は更に多くの演算装置を持つメニーコアプロセッサなどにおいて，NoC (Network on Chip)とよばれるチップ内ネットワークによる分散メモリ並列処理なども想定される．

科学技術分野においては，10 ペタ回/秒（1 秒あたり 10^{15} 回）もの実数計算能力を持つペタスケールの超並列計算機のような，10 万以上にのぼる多数の計算ノード（演算装置）をネットワークで密に相互接続した分散メモリ型の並列システムが数年以内に登場する見込みである．

分散メモリ方式の並列処理において，通信の効率化はシステムの効率的な利用における大きな課題である．これは，一般に，演算装置での処理時間に対してデータ交換にかかる通信時間が長いことから，データ交換の効率が全体の処理能力に大きな影響を与えるためである．

アプリケーションの開発者やシステムの設計者が通信の効率化を行う際には．実機上で多数の実験が必要である．しかしながら大規模システムの利用には様々なコストがかかるため，シミュレータによる事前の検証が必要である．

通信シミュレータとして様々なものが開発されている．従来のシミュレータは，通信中で送受信されるデータの少単位（パケットまたはフリット）を個々に粒として再現するパケット（フリット）ベースのシミュレーションを行っているものが大半である．

これに対して研究代表者はこれまでに，通信を流体の流れ(フロー)として再現する新しいフローベース方式を提案し，その実装である MFS (Message Flow Simulator) を研究・開発してきた．

MFS は，従来の数十から数百分の一の実行時間とメモリ消費量の実現により，家庭用パソコン程度の性能の PC でも，大規模な通信シミュレーションを実行可能とする．一方で，MFS はネットワーク中で通信の流れを制御するネットワークスイッチで複数の通信が衝突した場合に通信の優先度を決定する調停機構やスイッチの物理動作時間が完全に再現できないため，従来手法よりシミュレーション精度が約 40%低くなるという問題があり，この改善が大きな課題であった．

２．研究の目的

本研究では，これまで MFS に関する研究で培ってきた通信シミュレーション高速化技術を基点として，シミュレーションの実行時間の増加をおさえつつ，より高精度な結果を得る新しいシミュレーション技術の開拓を目的とする．

３．研究の方法

本研究では，大規模集積回路（VLSI , Very Large Scale Integrated circuit）の設計工程中の回路シミュレーションで用いられているバック・アノテーション（Back annotation, 注釈書き）シミュレーションを応用した通信シミュレーション手法を提案する．

回路設計の初期段階においては，通常，回路シミュレーションにより，演算結果の正しさを検証する．しかし，正しい結果を得る回路であっても，それが電気的に動作可能かどうかをさらに検証する必要がある．そのためには，回路を構成する素子を電気が通過する時間（遅延時間，Delay）や，それにかかる電力（消費電力，Power）を明らかにする必要がある．これらを見積もり，正しく動作する回路を設計するために,図 1 に示すように,回路を構成する素子の電気的な特性に関する情報を回路シミュレーションにフィードバックするシミュレーション手法がある．この手法により，より実機の動作に近い精度の回路シミュレーションが可能である．

本研究では，同様の手法を通信シミュレーションに適用する．通信シミュレーションにおける構成素子はネットワークスイッチである．この手法では，図 2 に示すように，ネットワークを構成するスイッチの重要な物理特性である平均通信遅延時間（Delay）や消費電力（Power）を実機で測定し，MFS のよ

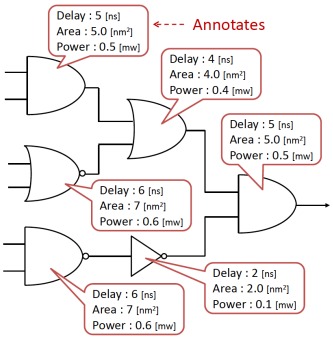

図 1 回路シミュレーションにおけるアノテーション

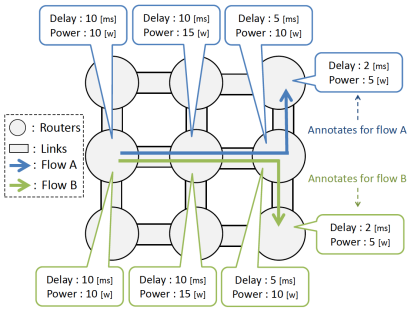

図 2 通信シミュレーションにおけるアノテーション

うなフローベース通信シミュレーションにフィードバックする．これにより，フローベースシミュレーションでは難しかった物理特性を考慮したより精密な通信シミュレーションを実現する．

バック・アノテーションに用いる情報を得るため，測定環境を構築する．測定環境の構築においては現実の機器を想定し，大規模並列計算機の国際的な性能ランキング Top500 に記載されている約 40%のシステムで使われているものと同じ InfiniBand と呼ばれるネットワークアーキテクチャを用いる．InfiniBand を構成するネットワークスイッチや複数のコンピュータを使い，ネットワークスイッチにかかる通信のコストを測定し，これをシミュレータにフィードバックする．また，比較のために，通常のインターネットで利用されている TCP/IP による通信を可能とするため，Gigabit Ethernet によるネットワークも構築した．

作成した通信シミュレータの評価においては，実用的な通信パターンを評価のベンチマークとする必要がある．本研究では，スーパー・コンピュータ上で行われるような大規模並列処理で問題となっている全対全と呼ばれる通信パターンで評価を行う．全対全通信は，通信に参加する演算装置（プロセス）が他のすべてのプロセスと通信を行う．N 個のプロセスで全対全通信を行うと，ネットワーク上で N^2 回の通信が行われるため，ネットワークに高い負荷をかける．この通信パターンは様々な分野で多くの重要なアプリケーションに利用されている高速フーリエ変換（FFT，Fast Fourier Transform）を行う場合などに発生するため，効率化を図った様々な通信手法が提案されている．本研究でも，実用的な通信のシミュレーションを精密に評価できるかどうかを確認するために，この全対全通信をベンチマークとして用いる．

4．研究成果

本研究においては，まず，提案している MFS の改良を行った．1．で述べたように，これまで MFS では，ネットワークスイッチで複数の通信が衝突した場合に通信の優先度を決定する調停機構を精密に再現していなかった．具体的には，ネットワークスイッチを介して複数の通信のフローが合流するような場面において，本来ならば合流する各通信の発信元を考慮してフローの流量を計算しなければならない．しかし，これまでの実装では，すべてのフローの流量を公平にあつかってきた．このような通信調停を実物のネットワークで行うためには，特殊な条件が必要となる．この問題を解決するため，より現実のネットワークにちかい振る舞いとなるようにシミュレータの実装を変更し，ごく小規模の単純な通信パターンにて動作を確認した．大規模な通信シミュレーションによる評価は今後の課題である．

提案している通信シミュレーション手法を評価するための準備として，既存のパケットベース・シミュレータを用いて大規模な全対全通信のアルゴリズムの性能評価を行った．シミュレータ自体の性能を比較するためには，評価に用いる通信パターンを，比較対象となる他のシミュレータでも評価しておく必要がある．他のシミュレータとしてオープンソースのパケットベース・シミュレータである Booksim を用いた．Booksim はネットワーク自体の性能を評価する目的で実装されているため，特定のパターンでの通信アルゴリズムを評価する機能はない．本研究ではこのための機能の追加実装も行った．

実験においては，最大 3,375 個のスイッチで構成される 3 次元メッシュおよびトーラス型のネットワーク上の全対全通信を評価した．メッシュ・トーラスネットワークは，近年のスーパー・コンピュータシステムで利用されているネットワーク形態の一つである．メッシュは格子状，トーラスはドーナッツ型である．通信アルゴリズムとしては，実際のスーパー・コンピュータ上で使われているアルゴリズムだけでなく，研究代表者らが別の研究プロジェクトで提案している独自の高効率アルゴリズムを用いた．

この実験から，パケットベース・シミュレーションにおいて比較基準となるベンチマーク通信の性能をいくつかのパターンで確認することができた．この実験結果の一部は 5．の[1]で示す国際会議にて公表している．

次に，バック・アノテーションに必要なネ

ットワークスイッチの物理的な特性に関する情報を収集するための環境構築を行った．３．で述べたように，ネットワークアーキテクチャとしては，InfiniBand を用いた．構築した環境において，並列計算機向けのネットワーク性能評価用ベンチマークプログラムを用いて性能測定実験を行った．実験により,実機ネットワークの性能や振る舞いを特徴付ける遅延情報等を確認することができた．これらを提案している MFS に組み込むための実装は現在も継続中である．

シミュレータとスーパー・コンピュータ実機との振る舞いの違いを調べた．シミュレータの評価においては，実機との比較が望ましいが，大規模並列処理において実機のデータを測定するのは容易ではない．そこで，研究協力者から，「京」コンピュータのほぼ全システムを用いて実行された全対全通信の通信記録（ログ）の提供受けた．比較においては，数値のみの比較にとどまらず，他の研究プロジェクトにて提案している通信可視化ツール CLV (Communication Log Viewer) を使って視覚的な比較も行った．

提案しているフローベース・シミュレータ MFS のバック・アノテーション機能の実装がこの時点では未完であったため，今回の実験においては，シミュレータには，MFS ではなく，改良したパケットベース・シミュレータである Booksim を用いた．

実験においては，京コンピュータに実装された独自の全対全通信アルゴリズムと，世の中で一般的に使われている通信ライブラリである OpenMPI に実装された全体全通信アルゴリズムを比較した．本実験に関する詳細な報告は本報告書作成時点では論文として投稿中である．この実験により，実機とパケットベース・シミュレータとの比較結果を得ることができた．この結果を用いて，今後は提案している MFS と実機との比較を行う．

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）

〔学会発表〕（計１件）

[1] Syunji Yazaki, Haruyuki Takaue, Yuichiro Ajima, Toshiyuki Shimizu and Hiroaki Ishihata, ``An Efficient All-to-all Communication Algorithm for Mesh/Torus Networks," Proceedings of 10th IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA2012), pp.277-284, Madrid, Spain, 10-14 Jul. 2012.

６．研究組織

(1)研究代表者

矢崎　俊志（YAZAKI, Syunji）
電気通信大学・情報基盤センター・助教
研究者番号：60454138

(2)研究分担者

(3)連携研究者