実通信パターンを用いた相互結合網の過渡応答の評価

杉森 帝政 † 小川 雅昭 † 横田 隆史 † 大津 金光 † 馬場 敬信 †
† 宇都宮大学工学部情報工学科

1 はじめに

近年、並列計算機は大規模化が進んでおり、その規模が大きくなるにつれてそれらノード間での通信は複雑にまた頻繁に行われる。よって規模の大きい並列計算機ほどその通信性能が全体の性能に大きく影響を与える。通信性能の向上は必要不可欠なものであると言えるだろう。

我々はその通信性能の向上を目的とし、輻輳情報から最適な経路選択を行いルーティングを可能にするアルゴリズム「Cross-Line」[1] を提案した。輻輳情報としてはバーチャルチャネル (Virtual Channel:VC) の状態情報を用い、送信元ノードは自ノードからのびるリンク上にある VC の状態情報から送信方向を決定する。この情報を用いることで輻輳箇所を回避しルーティングを行うことが可能となり最短経路の選択と相互結合網内のトラフィックの分散を両立させ通信性能の向上を実現する。次に我々は実際に Cross-Line の評価を行うため相互結合網シミュレータ Chimera を開発した。Chimera によるシミュレーションはアプリケーションにおける通信部分に限られたものだが容易にアルゴリズムの導入を行える、相互結合網に関する多数のパラメータが設定可能で様々なネットワーク状況によるシミュレーションが可能であるという特徴を持つ。設定出来るパラメータとしてはトポロジ、フロー制御、ルーティングアルゴリズム、ネットワークサイズ、リンク幅、通信パターンがある。

これまで、この Chimera によるシミュレーションを用い、ルーティングアルゴリズムごとの評価、比較が行われてきた。それは確率モデルにおける通信パターンを用いた評価、実際のアプリケーションにおける通信パターンによる評価 [2] と様々なものである。しかしそれらの評価は実行時間全体の比較にとどまっており、それら通信における挙動など詳細なデータは採られていない。そこで本研究ではそれら通信を反映させたシミュレーションからパケットレイテンシを時系列データとして採取し、そのアプリケーション実行に伴う挙動を明らかにすることで、それら通信に適したルーティングアルゴリズムの対策などを容易にするべく評価を行った。

2 実アプリケーションによる通信

2.1 MPI(Message Passing Interface)[3]

各プロセッサがメモリを保持している分散メモリモデルの並列計算機ではデータのやりとりをするために計算機間のメッセージ交換が不可欠であり、MPI はメッセージ交換の仕様である。並列計算に参加するプロセッサはコミュニケータというグループに入り、グループ内固有のランクという番号を割り当てられる。そしてこのコミュニケータとランク番号から行うべき処理を判断し、分散を行う。また MPI では集団通信を行う関数が定義されており、これを使用することにより容易に並列化の実現が可能となる。以下に代表的な集団通信関数を示す。

・`MPI_Alltoall`：コミュニケータ内の全プロセスの送信バッファから全プロセッサの受信バッファにお互いにメッセージの送信を行う。各宛先プロセッサへの送信メッセージの長さは一定で送信バッファの先頭から宛先プロセッサのランク番号が小さい順に送信を行い、送信元プロセッサのランク番号が小さい順に受信バッファの先頭から格納される。

・`MPI_Alltoallv`：`MPI_Alltoall` 関数同様の通信を行うが、送信を行うプロセッサ毎にメッセージ数を変えることができる点で異なる。

・`MPI_Allreduce`：コミュニケータ内の全プロセスの送信バッファのメッセージが通信を行いながら演算され、結果が全プロセスの受信バッファに格納される。

2.2 NAS Parallel Benchmarks 3.1

評価対象とするベンチマークとして並列アプリケーションプログラム NAS Parallel Benchmarks 3.1 を使用。CG 問題、MG 問題の通信について説明を行う。

・CG 問題：正値対称な大規模疎行列の最小固有値を求めるための共役勾配法を行う。全ての通信が 1 対 1 通信 `MPI_send`、`MPI_irecv`、`MPI_wait` により行われ、一回の通信毎に同期がとられている。このアプリケーションの通信は横方向に行う通信パターン 1 (図 1) と対角線方向への通信パターン 2 (図 2) によって構成される。メッセージ数の比で見ると全体の約 84% の通信がパターン 1、約 16% がパターン 2 となっている。

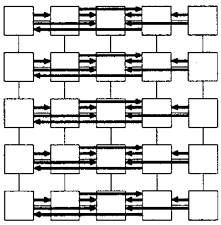

図 1: 通信パターン 1

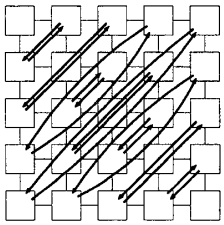

図 2: 通信パターン 2

・MG 問題：3 次元ポアソン法を簡略化したマルチグリッド法で解く。通信は `MPI_Allreduce` と 1 対 1 通信により構成されている。全体の通信の約 5% が `MPI_Allreduce`、約 95% が 1 対 1 通信となっている。1 対 1 通信おける到着パケット数は全てのノードで等しいが、各ノード決まったノードと頻繁に通信を行うため全ノードから全ノードに均一にパケット転送が行われているわけではない。

3 評価

NAS Parallel Benchmarks 3.1 の実通信パターンと確率モデル、ユニフォームランダム通信の過渡応答を用いたルーティングアルゴリズムの評価を行う。実通信パターンの評価については Chimera にてシミュレーションを実行する際、パケット毎の到着クロックと生成クロックからレイテンシを計算し、そのデータを用いた評価を行う。またシミュレーション時の設定パラ

Evaluation of Transient Behavior of Interconnection Network with Practical Communication Patterns

† Tadamasa Sugimori, Ogawa Masaaki, Takashi Yokota, Kanemitsu Ootsu and Takanobu Baba
Department of Information Science, Faculty of Engineering, Utsunomiya University (†)

メータについては表 1 に示す。

表 1: 設定パラメータ

トポロジ	2 次元メッシュ網
フロー制御	ストアアンドフォアード方式
ノードサイズ	32 × 32 (1024 ノード)
パケットサイズ	64byte (過渡応答)128byte
リンク幅	4byte/clock

3.1 CG 問題

図 3 は CG 問題のシミュレーション結果であり、実行開始から終了までの 100 クロック内に到着したパケットのレイテンシの平均と実行時間の推移を示している。また図 3 内、右上のグラフはルーティングアルゴリズム Cross-Line 適用時の CG 問題における到着パケットのレイテンシを時系列で採取したものである。平均レイテンシのデータを見ると両ルーティングアルゴリズムともに一度は高いレイテンシを示すが Cross-Line はその収束が早く実行終了も早いことからこの問題に関し適したルーティングをしていると読み取れる。

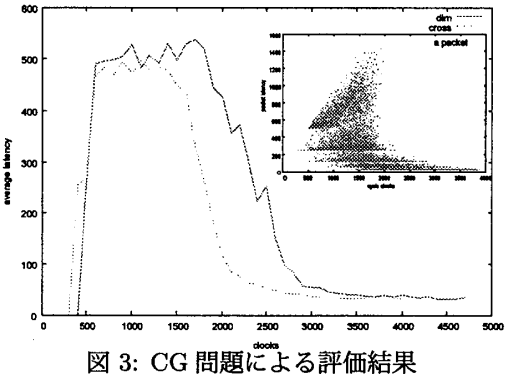

図 3: CG 問題による評価結果

3.2 MG 問題

図 4、図 5 は MG 問題の評価結果である。図 4 は開始から終了までの 100 クロック毎の平均レイテンシと実行時間の推移の関係、図 5 は開始から 40000 クロック経過までの様子を取り出したものである。MG 問題での通信は `MPI_Allreduce` と 1 対 1 通信が交互に行われる。`MPI_Allreduce` では全ノードのデータを使用して計算を行い、計算結果を全ノードに保持させるため大きく負荷がかかりその通信処理の時、非常に大きなレイテンシを記録している。また図 5 を見ると Cross-Line は Dimension-order に比べ、若干ではあるがレイテンシを低く保ち全体的な処理が先を行っていることから、MG 問題に関し Cross-Line は Dimension-order より適しているといえる。

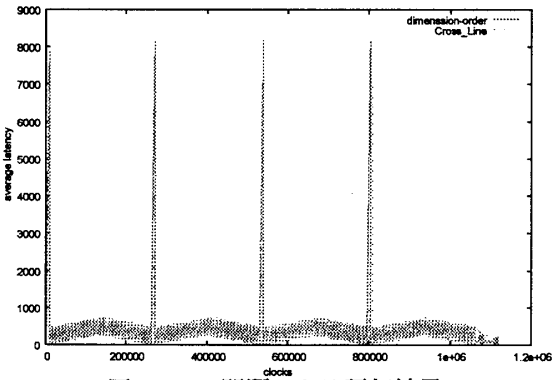

図 4: MG 問題による評価結果

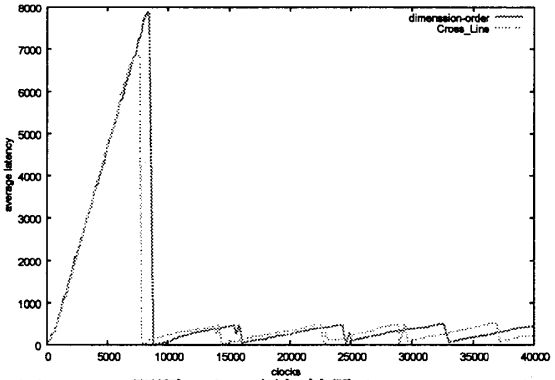

図 5: MG 問題による評価結果 (0-40000clocks)

3.3 ランダム通信

以下はランダム通信の評価である。この評価ではシミュレーション中、開始 20000clocks でパケット生成インターバルを引き上げ、40000clocks で元に戻した。縦軸は送信先到着パケットのレイテンシ、横軸は経過クロックを表す。この評価では Cross-Line は他のルーティングアルゴリズムに比べやや高いレイテンシを示している。また全てのルーティングアルゴリズムに関し 40000clocks 後に一時的に急激にレイテンシが伸びている原因はインターバル上昇中に滞っていたパケットが目的地に到着し始めたものと考えられる。

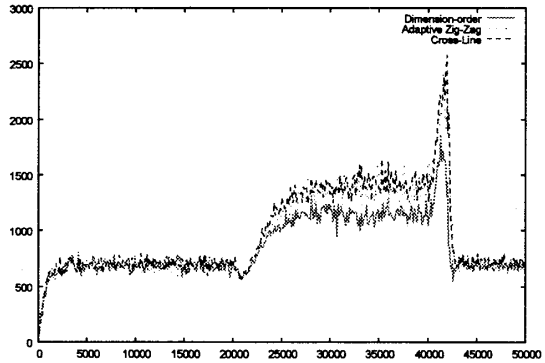

図 6: ランダム通信による過渡応答の評価

4 おわりに

本稿ではルーティングアルゴリズム毎の様々な通信について相互結合網用シミュレータ Chimera を用い、評価を行った。その結果としてシミュレーション中にその通信がどのような挙動をとるかについてデータをとることが出来た。今後の課題としては本稿で評価を行ったベンチマーク以外のものに対し同様の評価を行いその挙動の明らかにすることが挙げられる。

謝辞 本研究は、一部日本学術振興会科学研究費補助金 (基盤研究 (B)18300014、同 (C)19500037、若手研究 (B)17700047) および宇都宮大学重点推進研究プロジェクトの援助による。

参考文献

[1] 横田 隆史、西谷 雅史、大津 金光、古川 文人、馬場 敬信、"大域的な情報を用いる相互結合網方式 Cross-Line"、情報処理学会論文誌:コンピューティングシステム、Vol.46、No.SIG 16 (ACS-12)、 pp.28-42、 2005.12.

[2] 鈴木剣之介、"実アプリケーションにおけるルーティングアルゴリズムの評価、" 宇都宮大学大学院工学研究科情報工学専攻、修士論文、2007.3.

[3] 青山幸也、"並列プログラミング入門 MPI 版、"理化学研究所情報基盤センター、`http://accc.riken.jp/HPC/training/mpi/mpi_all_2007-02-07.pdf`.