並列離散事象シミュレーションにおける論理プロセスの移送による通信最適化

3Z－5

蔡 晟蔚, 日高宗一郎, 相田仁, 齊藤忠夫
東京大学　工学部

1　はじめに

並列離散事象シミュレーション (Parallel Discrete Event Simulation:**PDES**) の研究の歴史が十数年にもなるにもかかわらず、DES の効率よい並列化には、まだハードウェアやプロトコル同期機構等の低レベルの知識とプログラミングの習熟が必要なのが現状である。

本稿は PDES ユーザがシステムごとに異なるネットワークトポロジを意識せずにプログラムをしても、最適なマッピングを得て高い転送性能が実現できるライブラリの開発を目的とする。

2　並列計算機の通信ネットワーク

現在主流である分散メモリ型並列計算機は、数百から数千台のプロセッサをネットワークで結合した構成をとる。データの授受はネットワークを介してメッセージパッシングにより行うため、ネットワークの転送性能がシステム全体の性能に影響を与える。現在よく使われているのはハイパキューブネットワーク (以下 **HCB** と呼ぶ)、トーラスネットワーク (以下 **TORUS** と呼ぶ) そしてハイパクロスバネットワーク (以下 **HXB** と呼ぶ) である。

東大の大型計算機センターに設置されている SR2201(日立) は 3 番目の **HXB** で構成されている。

3　LP の移送による通信最適化

PDES プログラムは外部イベントの交換を頻繁に行うため極めて通信集約的であり、論理プロセス (Logical Process:以下 **LP** と呼ぶ) 間通信が全体の性能に与える影響は大きい。更に、LP 間の通信パターンと、物理プロセッサ (Physical Processor:以下 **PP** と呼ぶ) 間の結合トポロジとの間で整合が取れていないと、PP 間の結合ネットワークに対する負荷を増大させ、性能低下を招く場合がある。

以上の問題をハードウェアですべて解決するのは困難だと思われる。もし、ソフトウェアのレベルでマッピングがプログラムの走っている間に臨機応変に変えられたら、問題は解決できる。そこで、実行時に LP の移送による通信の最適化を試みる。LP 間通信と PP 間通信結合網の不整合性を具体的に数値として表すために"通信コスト"という概念を導入する。LP 間通信の性能を低下させる原因はマップされた PP 間の相互結合網上の距離による遅延と、PP 間の経路の競合による遅延に分けることが出来る。例えば、トーラスの場合以下のようなコスト関数が考えられる。

$$(1 + \text{通信距離} \times \text{係数 } A) \times \text{通信量} \times \text{係数 } B^{\text{軸乗り換え数}}$$

集中アルゴリズムによる移送のシーケンスは以下のようになる。なお、通信において、メッセージの追い越しはないと仮定する。

1. 移送決定、通信の中断を全 PP に放送 (ID=0)
2. PP 間通信チャンネルのクリア
   (a) 送信の停止、受信バッファの中のメッセージが空になるまで繰り返し受信 (ID=0)
   (b) 通信中断の放送を受信すると送信を停止, 受信を繰り返す。ID=0 に Ack を返す。全 PP に相手 PP 宛ての外部イベントがないことを知らせる。(ID!=0)
   (c) 自分宛ての外部イベントがもうないと分かったら、ID=0 に準備完了を知らせる。(ID!=0)
   (d) すべての PP から準備完了のメッセージが届くまで待つ。もちろん自分宛ての外部イベントがないことを確認 (ID=0)
3. 全 PP に移送開始命令を放送 (ID=0)
4. 各 PP の LP カーネルを通して通信量の集計、コスト関数とコスト最小化アルゴリズムに従って、LP 移送候補を決める (ID=0)
5. LP 移送候補に移送の知らせ。(ID=0)
6. LP の直列化および転送 (ID!=0　移送候補)
7. LP の復元、ID=0 に転送終了を知らせる (ID!=0 移送候補)
8. 転送終了を受けたら、マッピング情報の更新 (ID=0)
9. 通信の再開を全 PP に放送する (ID=0)

4　バニヤンスイッチモデルへの適用

4.1　HXB 上の最適マップ及び性能評価

図 1 のようなマッピングパターンで、それぞれシミュレーションを行なった。SR2201 では X 軸優先のルーティ

"Optimization of communication process by logical process transfer in PDES"
Shengwei Cai,Soichiro Hidaka,Hitoshi Aida,Tadao Saito
Faculty of Engineering,The University of Tokyo

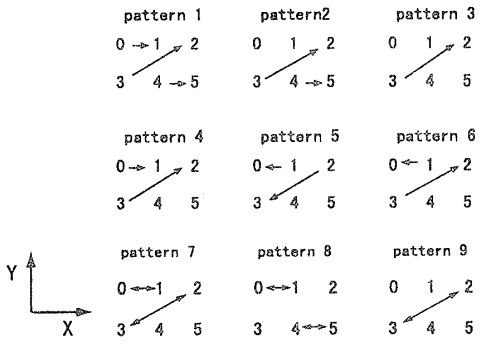

図 1: プロセッサ間通信衝突の測定に用いたパターン

ングが行われる。実際にパターン 1、2、5、7 では、ノード輻輳が生じているのが確認された。以上の実験はコスト関数を作るヒントとなる。

コスト関数理論を応用して、3 段バニヤンスイッチネットワークの最適マップを全解探索を用いて求めてみた。結果は図 2 になる。

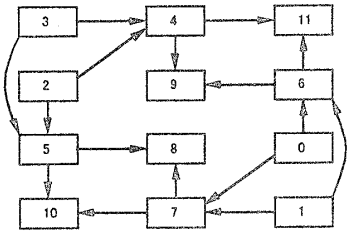

図 2: 3 段バニヤンスイッチの最適マップ

そして、最適マップを使って、実際にシミュレーションプログラム (VT : 0〜2.0e + 07) を走らせた結果、表 1 のようになった。

結果からみると SR2201 の上では最適マッピングと従来のマッピングの差はあまりはっきり現れていない。それは SR2201 ではかなり贅沢なほどリンクを張っていて、スイッチング方式とルーティング方式の面でもかなり工夫されているのが原因だと思われる。

負荷 0.8	従来のマッピング	最適マッピング
1 回目	150 秒	143 秒
2 回目	138 秒	142 秒
3 回目	158 秒	146 秒
4 回目	139 秒	144 秒
5 回目	166 秒	150 秒
平均	150.2 秒	145.0 秒
標準偏差	10.8 秒	2.82 秒

表 1: 三段バニヤンスイッチ最適マッピングの実験結果

4.2 コスト最小化アルゴリズム

LP 移送による通信の最適化において、コスト関数の作り方は重要であるが、それと同時にコスト最小化アルゴリズムもまた重要である。探索時間を短くするために以下のように工夫した。

1. **方針**
   (a) コストのパラメータは三つ、通信の量と距離そして軸乗り換え数
   (b) 隣同士の LP しか交換できない
2. **プログラムのアルゴリズム**
   (a) 2 次元なら X,Y 軸、3 次元なら X,Y,Z 軸の順にループに入って、LP 交換ペア (隣同士) を見つける。
   (b) 各 LP は隣の LP と仮に互いに交換したら、全体にどのくらいのコスト値の変化が出るのかを計算する。
   (c) 全体に対して、一番利益の出る二つの LP を互いに交換させる
   (d) 前の交換によりコスト値が変わるので、全体のコスト値を更新。そして、繰り返し前の操作をする。
   (e) 全体に対して、もうプラスの利益がなかったら、終了

5 結論

本稿では、PDES における論理プロセス間の通信と物理プロセッサの結合網の不整合によるデータ転送性能の低下を招く問題について述べ、その解決方法の 1 つとして—論理プロセスの移送による方法を提案した。

コスト関数理論を使って、3 段バニヤンの場合、従来のマッピングの 4 分の 1 に、4 段バニヤンの場合 3 分の 1 に通信負荷を抑えることが出来た。今後はもっと複雑なコスト関数モデルおよび最小化アルゴリズムの検討を行なう。そして実装レベルでの問題について述べる。

参考文献

[1]Yoshiko Yasuda, et al. "Architecture and Performance of the Hitachi SR2201 Massively Parallel Processor System." In Proceedings of the 11th International Parallel Processing Symposium, pp. 233-241, April 1997

[2] 東京大学大学院電子情報工学専攻　齊藤・相田研 日高 宗一郎 : "並列離散事象シミュレーション環境の研究" 学位論文 (博士),February 1999