計算機アーキテクチャ 132−26
システムソフトウェアとオペレーティング・システム 80−26
ハイパフォーマンスコンピューティング 75−26
（1999. 3. 5）

タイムワープ法並列離散事象シミュレーションにおけるロールバックの雪崩現象

山口 佳紀 本多 弘樹 弓場 敏嗣

電気通信大学大学院情報システム学研究科

概要

並列離散事象シミュレーションをタイムワープ法を用いて実行した場合、ロールバックの雪崩現象が生じるとシミュレーション実行時間に影響を及ぼす。雪崩現象の発生はシミュレーション対象モデルや実行環境に依存する部分が大きい。そこで並列環境において雪崩現象に関する実行時間解析を行い、その発生機構を解明する。併せて、雪崩の発生防止、発生除去、発生からの今後の各技法を議論する。

An avalanche phenomenon of roll-back in Time Warp parallel discrete event simulation

Yoshiki YAMAGUCHI, Hiroki HONDA, and Toshitsugu YUBA
Graduate School of Information Systems
The University of Electro-Communications
E-mail: yoshiki@yuba.is.uec.ac.jp {honda, yuba}@is.uec.ac.jp

Abstract

An avalanche phenomenon of roll-back affects simulation execution time in Time Warp parallel simulation if it were occurred. This phenomenon deeply depends on a simulation target model and execution environments. Using simulation execution time analysis, we investigate the phenomenon on a closed queuing system model on a parallel computer. In addition we discuss some ways of preventing, elimination and countermeasure the phenomenon.

1 はじめに

高速ネットワークや高性能計算機を技術基盤として高度分散システム環境を実現するとき、離散事象シミュレーション技術を用いた設計段階におけるシステムの解析・評価が不可欠になっている。近年の急速な技術の進歩に伴い、高速実行可能な離散事象シミュレーション技術の確立が重要となってきている。この要求を満たすため、並列計算機や分散システム上で実行される並列離散事象シミュレーション (Parallel Discrete Event Simulation: PDES) に関する研究 [1] が行われている。シミュレーション同期プロトコルは、逐次計算機で行うシミュレーション結果と同じシミュレーション結果を PDES で得るために必要となる。

楽観的プロトコルのタイムワープ法 [3] は、メッセージに含まれるタイムスタンプの時刻にかかわらずに事象処理を行うことができ、シミュレーション対象モデルが並列性を多く持つ場合、より多くの並列性が引き出せる。事象処理の因果関係の矛盾を発見すると、シミュレーション状態を因果関係の矛盾が生じていない時刻まで巻戻す処理（ロールバック）が必要となる。他の論理プロセスに因果関係の矛盾がある状態で誤って処理したメッセー

ジを送っている場合、ロールバックの際に取消のための特別なメッセージ、アンチメッセージを送る必要がある。また、シミュレーション終了判定やメモリの有効利用のために、すべての論理プロセスと、論理プロセス間を結ぶ通信リンクの中に存在するメッセージのタイムスタンプの最小値（大域仮想時刻、 Global Virtual Time:GVT) をシミュレーション実行中に求める必要がある。

PDES は、シミュレーション対象モデルの並列実機環境への割り当て方や、シミュレーション同期プロトコルの用い方、さらには並列実機環境のプロセッサ能力および通信遅延などにより、シミュレーション実行時間が変化する。並列実行をしても期待した結果が得られず、逐次計算機で行うシミュレーションよりも遅くなってしまう場合もある。PDES を高速実行するためには、並列環境でどのようにシミュレーションが進行されているかを調べることが重要である。

先の研究 [6] で、並列環境においてタイムワープ法を用いた PDES の実行時間が急激に増大する状態が必ず発生することが確認された。実行時間の増加の要因として、複数の要素プロセッサ (PE) でロールバックが連続して発生するロールバックの雪崩現象が挙げられる。本稿では、ロールバックの雪崩現象をシミュレーションの実行時間解析によって解明する。また、雪崩現象の防止法、雪崩現象が発生した場合の対処法について検討する。

2 ロールバックの雪崩現象

タイムワープ法を用いた PDES の実行には、シミュレーションを正しく進行させるためにロールバックが必要となる。ロールバックは、自論理プロセス内の処理と他論理プロセスへの処理の 2 つに大きく分けられる。自論理プロセス内の処理は、楽観的に行った事象処理によるプロセスの状態を因果関係の矛盾が生じる前の状態に戻すことである。他論理プロセスへの処理は、アンチメッセージの作成及び、送信である。図 1にロールバックの流れを示す。

ある論理プロセスにおける事象処理が他の複数の論理プロセスに必ず影響を与えるようなシミュレーション対象モデルの場合、1 回の誤った事象処理の影響が複数の論理プロセスに伝搬してしまう。アンチメッセージによる事象処理の取消がその伝搬に追いつけない場合、さらにアンチメッセージの作成、送信といったロールバックも伝搬することになる。特に 1 回のロールバックで 2 個以上のアンチメッセージが他の PE 上に存在する論理プロセスに送信される場合、2 箇所以上の PE でロールバックが時間的に続いて起こる可能性がある。このように、1 回のロールバックが 2 箇所以上の PE でのロールバックを連続的に導くような現象をロールバックの雪崩現象と呼ぶ。

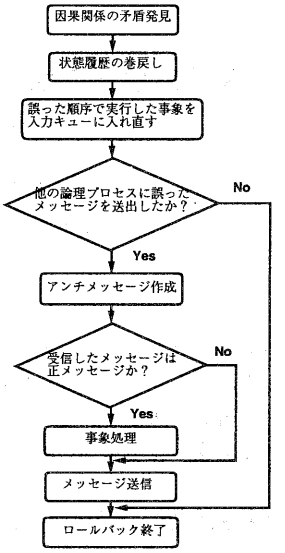

図 1: ロールバックの流れ

雪崩現象が発生するための条件として、1 回のロールバックで 2 箇所以上の他の PE にアンチメッセージを送信し、かつそれらアンチメッセージを受信した PE でロールバックが発生することが挙げられる。

このような現象をモデルを用いて解析した研究 [4] も行われている。しかし、雪崩現象の発生はロールバックに費やす時間やメッセージの通信時間などの並列実機環境や、メッセージをどのように扱うかなどシミュレーション対象モデルに依存

するところが大きい。そこで、実際に雪崩現象が生じるようなシミュレーションを実行し、その実行時間解析をもとに並列実機環境やシミュレーション対象モデルとの関係を調べることが必要となる。

3 並列計算機へのタイムワープ法の実装

今回の解析では、以下に述べるシミュレーション対象モデルにあわせたPDESシステムを並列環境に実装した。

3.1 シミュレーション対象: 閉鎖的待ち行列システム

シミュレーション対象として、4×4の16個のサーバが2次元トーラス状に接続された、閉鎖的待ち行列システムを用いる（図2)。このシステム内には一定数の客が存在する。客は、サーバに到着すると、あるサービス時間を経て、次のサーバにつながる2つの出力枝の一方から出発し、移動時間0で次のサーバに到着し、サービスを受けることを繰り返す。

3.2 PDESの実装環境

実験のハードウェア環境として、分散記憶型並列計算機Cenju-3[5]を用いた。Cenju-3は、要素プロセッサ(PE)にVR4400を用い、PE数の最大値は256台である。PE間の通信ネットワークは4×4のクロスバスイッチ4段による多段接続網を用い、その転送速度は40MB/secである。また、全てのPEへ等距離のネットワークを保証している。ソフトウェア環境としては、主な部分をC言語で記述し、通信部分は、Cenju-3のMPICHライブラリ[2][7]を用いて記述した。MPIで用いた通信関数は、ブロッキング送信関数MPI_Sendとブロッキング受信関数MPI_Recを標準モードで用いた。また、実行時解析のためにMPEライブラリ[2]を用いてメッセージの通信状況のログを収集した。

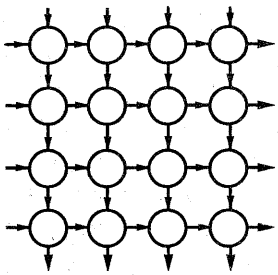

図2: 2次元トーラス接続の閉鎖的待ち行列システム

3.3 シミュレーションの実行

閉鎖的待ち行列システムの挙動のシミュレーションを行うために、まずシミュレーション対象のモデル化を行う。シミュレーション対象での時刻は、仮想的な時刻、仮想時刻にモデル化される。サーバはリンクで結ばれた論理プロセスに対応する。客がサーバに到着して何らかのサービスを受けた後、出発する事象は、客の出発時刻（客の移動時間が0のため、次のサーバでの到着時刻でもある）のタイムスタンプを到着時刻から計算し、サービスの内容に応じて論理プロセスの状態変化を行い、タイムスタンプを含んだ（正）メッセージを作ることに対応する。論理プロセスのPEへの割り当ては1対1とし、16台のPEを使用する（以下、論理プロセスとPEを同一とする)。システムに存在する客の総数は、常に一定とする。シミュレーションにおいては、客の総数と同じ数のメッセージを初期状態としてシステムに与える。

初期条件として、論理プロセスの2つの入力リンクにあらかじめメッセージを均等に配置する。各論理プロセスにおいてメッセージを受けとると、メッセージに含まれるタイムスタンプをもとに事象処理を行い、その後、新たなメッセージを1つ生成する。生成されたメッセージは、出力リンクの一方から隣接する論理プロセスに送信される。

MPICHライブラリを用いた通信でメッセージの受信は、MPIの受信バッファにメッセージが存在するか否かを判定するprobe関数MPI_Iprobeを用いて、メッセージが存在すれば受信関数MPI_Recvを実行するようにした。また、メッセージの送信は

1つのメッセージの送信毎に送信関数MPI_Sendを実行し、複数のメッセージをまとめて送ることはしていない。

4 実行時間解析

シミュレーション系内に存在するメッセージ数が異なる場合と、ロールバックの際の巻戻しの数が多くなる場合で雪崩現象の発生があるか否かを調べる必要がある。シミュレーションを以下の4つの初期条件下で行い、シミュレーション実行時間、ロールバック回数（アンチメッセージ受信によるロールバック回数は含まれない）と全メッセージ数に対するアンチメッセージ数の割合を測定し、同時に雪崩の発生を調べる。なお、シミュレーションの終了はGVTが100,000仮想時刻に達するまでとした。

1. すべての論理プロセスの2つの入力リンクに均等に2個のメッセージを配置
2. すべての論理プロセスの2つの入力リンクに均等に16個のメッセージを配置
3. 1個の論理プロセスの2つの入力リンクに均等に2個のメッセージを配置
4. 1個の論理プロセスの2つの入力リンクに均等に16個のメッセージを配置

初期条件1や3のような初期に配置するメッセージが少ない場合、事象処理を行う回数が少なくなる。また、初期条件1、2に比べ条件3、4は、メッセージが特定の論理プロセスの入力リンクに溜る傾向が増える。

4.1 シミュレーション結果

表1にシミュレーション結果を示す。初期条件の4番目のシミュレーションでロールバックの雪崩現象が発生したが、それ以外の条件では雪崩現象は発生していなかった。

表1: シミュレーション結果

	条件1	条件2	条件3	条件4
（秒）	1.78	1.94	1.94	2.12
（回）	160	302	102	1710
（%）	0.6	0.9	7.0	8.7

4.2 雪崩の発生状況

メッセージ通信のログ可視化ツールnupshot[2]を用いた、初期条件4のロールバックの雪崩発生時のシミュレーション実行時のスナップショットを図3に示す。図の縦軸はPE番号を示し、横軸はシミュレーション実行の経過時間（秒）を示している。

PE2の1.233秒辺りで因果関係の矛盾に起因する（今まで処理したメッセージのタイムスタンプよりも小さな値を持ったメッセージが到着）ロールバックが始まっている。自PE内の状態の巻戻しとアンチメッセージの作成、ロールバックを引き起こしたメッセージの事象処理を約600マイクロ秒で終え、他PEへの6つのアンチメッセージと1つの正メッセージの送信関数を計約2200マイクロ秒かけて実行している。PE3は、1.234秒辺りからPE2からのアンチメッセージを受信し、その結果5つの新たなアンチメッセージ送信を伴うロールバックが行われる。また、PE3からのアンチメッセージを受信したPE7も1.236秒辺りから4つの新たなアンチメッセージ送信を伴うロールバックが行われ、ロールバックの雪崩現象が生じている。

図3で着目したPE2での1回のロールバックは結果的に、23個のアンチメッセージを生成し、最後のアンチメッセージが消去されるまでに約8000マイクロ秒かかっていた。

4.3 処理の実行時間

雪崩現象の発生は事象の巻戻し数、アンチメッセージ送信を含めたロールバックに要する時間と関係がある。巻戻す事象数が少ないと、事象数に比例した自PE内でのロールバック（状態の巻戻しとアンチメッセージ作成）に要する時間が短い。また、アンチメッセージの送信数も少ないので、すべてのアンチメッセージを送信し終えるまでの時間も短い。逆に巻戻す事象数が多いと、その数に比例したロールバックの時間がかかるため、特にアンチメッセージの送信が遅くなる傾向が高くなる。そこで、シミュレーション実行中の各処理の実行時間を調べた。

正メッセージの事象1個の処理時間は約55マ

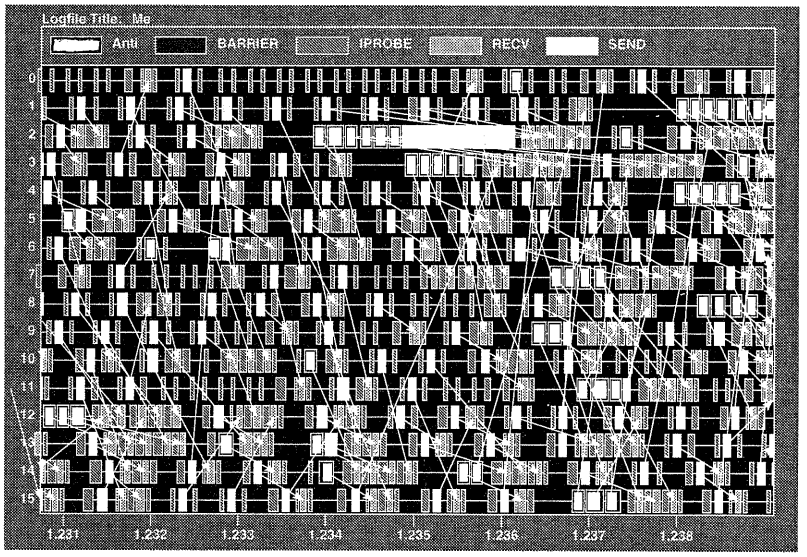

図 3: ロールバックの雪崩発生状況

イクロ秒、アンチメッセージ作成を含めた事象1個の巻戻し処理時間は約100～250マイクロ秒である。1つのメッセージをMPI受信バッファから2つの入力リンクのどちらかの入力キューに入れるための時間（MPI_Iprobe関数、MPI_Recv関数の実行と正メッセージとアンチメッセージの取消処理を含む）は約120マイクロ～200マイクロ秒、1つのメッセージを送信するための時間（MPI_Send関数が完了するまでの時間）は平均約85マイクロ秒である。ここで平均の値を示したのは、MPI_Send関数が完了するまでの時間が長くなることがCenju-3上で確認されたからである。

今回使用した1個のメッセージの大きさは24バイトで、MPICHの実装[7]によるとshortプロトコルと呼ばれる通信プロトコルを用いて送受信される。このプロトコルを用い、同じ送信先PEに対してMPI_Send関数を複数回連続して実行した場合、送信先PEがMPIの受信バッファに対するなんらかの処理を行わないと送信元のMPI_Send関数の5つ目でブロックしてしまう。つまり、送信先PEの状態によってMPI_Send関数の実行時間が影響を受ける。このブロック状態は、図3で示す雪崩現象の始まりの7個目のメッセージ送信の際(PE2の時間1.235秒辺り）に見られる。

shortプロトコルを用いた1個のメッセージの通信時間（送信先PEのMPIの受信バッファにメッセージが到着するまでの時間）の最大値はMPI_Send関数の平均的な実行時間（約85マイクロ秒）である。また、次の事象処理の対象となるメッセージがMPIの受信バッファに存在するか確認する間隔の最低値は約98マイクロ秒である。

4.4 雪崩の発生機構

初期条件の1,2のすべてのPEがシミュレーション初期に均一にメッセージを持つ場合、雪崩が起きなかった。これは、使用したCenju-3のPEの能力、通信時間が均一であること、また、メッセージの到着を確認する間隔よりも通信時間が短いため、全PE内に存在するメッセージの数が多い場合、MPIの受信バッファに常にメッセージが存在する確率が高くなり、アンチメッセージが正メッセージを取り消す確率が増えるためである。表1の結果より条件1、2において全メッセージに占めるアンチメッセージの割合は1%未満と小さい。

初期条件3の結果より、1つのPEにだけメッ

セージを配置してもメッセージ数が少ない場合雪崩は発生しない。この理由は、ロールバックにおいて巻戻す事象の数が少なく複数のPEへのアンチメッセージの送信が少ないためである。条件3において全メッセージに占めるアンチメッセージの割合は表1の結果から7%と条件1、2と比べて大きい。これは特定のPEの入力リンクにメッセージが溜っていて、その他のPEを伝搬するアンチメッセージは伝搬する回数が増えるという初期のメッセージの配置から導かれる性質であることに関係する。

初期条件4においては、表1の結果より多数のロールバックが発生し、アンチメッセージの数も多いことがわかる。ロールバック回数が多いことは、雪崩のきっかけを誘発する機会が増えることを意味する。また、アンチメッセージの数が多いことは、事象の取消が追い付いていないことを意味する。アンチメッセージ送信を含めたロールバックの時間が増えると、その時間、他のPEでは多くの事象処理が行なわれることになる。これらの事象がロールバック対象になると雪崩現象が発生する。実際、図3に見られる雪崩現象で、通常の事象処理時間に比べてロールバックに費やす時間が大きいことが確認できる。

5 雪崩現象の対処

雪崩の始まるきっかけは、巻戻す事象数（生成するアンチメッセージ数）とロールバックに費やす時間に依存している。巻戻す事象数が多いことは論理プロセスの事象の進行に差があることが原因の1つである。LVT進行の差は、シミュレーション対象モデルと並列実機環境との関係に影響を受ける。モデルの論理プロセスの並列実機環境への割り当て方を適切にすることで巻戻す事象数を小さくすることが可能となる。次にロールバックの処理時間を小さくするには、複数のアンチメッセージの送信を高速に行う必要がある。このためには同じ送信先宛の複数のアンチメッセージを1つにまとめて送信する方法が考えられる。雪崩が発生している状況をシミュレーション実行中に確認できれば、雪崩の進行を押えるなんらかの機構を働かせることができる。それには、アンチメッセージに雪崩の情報を付加することや、自PEの送受信関数の実行時間を測定しその変化から雪崩が発生している状況を確認する技法が考えられる。

6 まとめ

ロールバックの雪崩現象が生じている状態を並列実機環境での実行時間解析を行った。その結果、ロールバックにおける事象の巻戻しの数が多い場合、アンチメッセージの送信を含めたロールバックの処理時間が増え、アンチメッセージによる誤った処理の取消が遅れることが雪崩発生につながる原因と分かった。今後は、雪崩現象が生じた時の対処法の具体的な提案、評価を行う。また、OSによるシミュレーション実行への介在や不均一なCPU資源、通信網が存在する計算機クラスタ環境において並列実機環境の違いによる雪崩現象の発生を調査する。

謝辞 並列計算機Cenju-3の利用環境をご提供頂いた、NEC C&C研究所並列処理センターに感謝致します。

参考文献

[1] Ferscha, A., "Parallel and distributed simulation of discrete event systems," in A. Y. H. Zomaya (ed.), *Parallel and Distributed Computing Handbook*, Chapter 35. McGraw-Hill, 1995.

[2] Gropp, William and E. Lusk, "User's Guide for mpich, a Portable Implementation of MPI," Technical Report ANL/MCS-TM-ANL-96/6, Argonne National Laboratory, 1996.

[3] Jefferson, D. R., "Virtual Time," ACM Transactions on Programming Languages and Systems, Vol. 7, No. 3, pp.404-425, July 1985.

[4] Tay, S. C., Y. M. Teo and R. Ayani, "Performance Analysis of Time Warp Simulation with Cascading Rollbacks," Proc. of 12th Workshop on Parallel and Distributed Simulation(PADS'98), pp.30-37, May. 1998.

[5] 広瀬, 細見, 丸山, 加納, "並列コンピュータCenju-3のプロセッサ間通信方式とその評価," 情報処理学会論文誌, Vol.37, No.7, pp.1378-1387, July 1996.

[6] 山口 佳紀, 本多 弘樹, 弓場 敏嗣, "並列離散事象シミュレーションの実行時間解析," 計測自動制御学会, 第23回 離散事象システム研究会, pp.9-16, Dec. 1998.

[7] http://www.ccrl-nece.technopark.gmd.de/~mpich/