大規模シミュレーションデータベース用パラレルI/O制御エージェント

松岡 有希 *　松山 仁美 *　小金山 美賀 *　上島 豊 †　城 和貴 *
yuchan@ics.nara-wu.ac.jp
* 奈良女子大学理学部情報科学科　* 奈良女子大学大学院人間文化研究科
† 日本原子力研究所 関西研究所 光量子科学研究センター

概要

大規模シミュレーションを支援する統合管理システムにおいて，現状ではシミュレーション計算によって出力されるデータは，データサーバ上の複数のディスクサイズを考慮することなくセーブされている．そのため，データサーバ上の各ディスクサイズが不均等になり，データサーバの運用が非効率的である，という問題が生じている．そこで我々は，自律的に各ディスク上のデータを監視し，状況に応じてユーザへの問い合わせやデータファイルの移動を行うことによって，ディスクサイズをバランスよく保つデータベース用パラレルI/O制御エージェントのプロトタイプの設計と実装を行う．

Parallel I/O Control Agent for Large-scale Simulation Database

Yuki Matsuoka *　Hitomi Matsuyama *　Mika Koganeyama *　Yutaka Ueshima †　Kazuki Joe *
* Nara Women's University, Faculty of Science, Department of Information and Computer Sciences
* Nara Women's University, Graduate School of Human Culture
† JAERI, KANSAI Research Estabilishment, Advanced Photon Research Center

Abstract

An integrated management system for large-scale simulation saves the output data of simulations without consideration of the disk size on the data server. The present problem is an inefficient operation of the data server because each disk size on the data server is unbalanced. We design and implement a prototype of the parallel I/O control agent for large-scale simulation database which supervises the data server, makes suggestions to users and moves data files to the other disks to keep each disk size well-balanced.

1　はじめに

シミュレーション実験は，ある程度の処理能力を持つ計算機があれば，様々なパラメータで何度も実験を行うことができ，実験量に対する相対的なコストもよい．そのためシミュレーション実験は，あらゆる分野に対して有用性が高く，必要とされている．かつては，シミュレーション実験を行うことができる計算機環境は十分とは言えず，シミュレーション実行に要する計算時間が長く，また，満足な精度の結果を得ることも困難であった．近年の計算機ハードウェアの発展による高性能計算機の普及や，シミュレーション結果を評価するための有効な手段である可視化手法の確立によって，シミュレーション実験を短時間で効率よく行うことができるようになった．さらに，LANやWANなどの普及により，ネットワーク環境も充実し，ネットワークで接続された複数の計算機を用いた大規模なシミュレーション実験も試みられている．

シミュレーション環境が充実するとともに，シミュレーションの規模が大きくなり，手順も複雑になってきている．それにもかかわらず，パラメータの設定，使用する計算機の選択などの各種設定は，ユーザが手動で行うのが現状である．そのため，従来は別々に行っていたシミュレーション・データ転送・可視化・データ解析というシミュレーションサイクルを統合管理し，一括処理するシステムを用いるシミュレーションもある．

現在，日本原子力研究所 関西研究所 光量子科学研究センターでは，超並列計算機上で動作する超並列プラズマ粒子コードを用いて，時空間的に大規模，高精度なシミュレーションを行っている．この大規模シミュレーションを効率よく運用するために「Pキューブ支援システム」という統合管理システムが利用されている．しかし，現行のPキューブ支援システムを効果的に利用するためには，各種パラメータなどの設定を行う際に，シミュレーション実行に関するある程度の知識や複雑な操作を必要とするため，ユーザの負担は非常に大きい．また，シミュレーション計算結果のデータサーバへのセーブや，他サーバへのデータ転送が非効率的であることも問題となっている．

このような背景から，我々は，ユーザの操作補助や適切なアドバイスを行う能動的なエージェントシステムの開発に着手している [1]. エージェントとは，自らの意思決定原理に基づき，合理的な動作をするシステムである．その主な特徴として，ユーザの細かい指示がなくても，目標達成のためのプランを自ら考え，行動する自律性がある．

本エージェントシステムを実現させるために必要不可欠な機能の一つに，パラレル I/O 制御機能がある．P キューブ支援システムでは，シミュレーション計算結果のセーブの際に，データサーバ上の複数のディスクサイズを考慮していない．そのため，各ディスクサイズが不均等で，データサーバの利用が非効率的であることが多い．そこで，データサーバを監視し，ディスクサイズが均等になるよう自律的に処理を行うパラレル I/O 制御エージェントを構築する．

本稿では，日本原子力研究所 関西研究所 光量子科学研究センターで行われる大規模シミュレーションに用いるデータサーバ用パラレル I/O 制御エージェントの設計と，そのプロトタイプ実装に関して報告する．

2 データサーバの構成

2.1 システム環境

パラレル I/O 制御エージェントが管理するサーバは，データサーバとグラフィックサーバである．本稿ではデータサーバのみを対象とするエージェントを構築する．以下に日本原子力研究所 関西研究所におけるデータサーバの構成を述べる．

日本原子力研究所 関西研究所には，大規模計算用の開発，実行環境である超並列計算環境と中・小規模計算用の開発計算環境がある．超並列計算環境のノード構成は，データサーバ，グラフィックノード，プロダクティブノード等，5 種類のノードから成る．開発計算環境のノード構成は，データサーバ，アプリケーションノード，開発環境用プロダクティブノード等，8 種類のノードから成る．両計算環境にあるデータサーバは，ジョブの入出力に使用されている．両計算環境のデータサーバは，それぞれ図 1 のような構成になっている．

超並列計算環境のデータサーバは，ノード名 sscmpp2 から sscmpp19 までの 18 台の AlphaServerES40-4 から構成されている．それぞれのノードには FibreChannel 経由で高速ディスク装置 Compaq StorageWorks MA8000 が接続されている．ディスク装置は RAID5 で構成され，全体で約 15TB の容量がある．ディスク領域は，/work01 ～/work90 となっている．

開発計算環境のデータサーバは，ノード名 sscmppl2 から sscmppl5 までの 4 台から構成されている．ディスク装置の接続，構成は超並列計算環境と同様である．ディスク領域は，/work01～/work16 となっている．

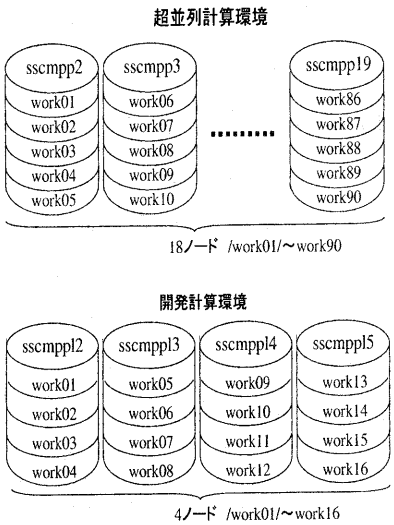

図 1: データサーバ構成

2.2 データサーバ利用における問題点

データサーバの利用者は，P キューブ支援システムの使用者だけでなく，複数の研究グループで複数のユーザ (約 120 人) が使用している．各ユーザは全ディスク領域を使用することができるが，ユーザによっては，常に同じディスクにセーブしている場合がある，またユーザによってディスク使用方法にばらつきがある．そのため，各ディスクの使用量が不均等な状態であることが多く，P キューブ支援システムを使用する際にも以下のような問題が生じている．

P キューブ支援システムを用いたシミュレーション計算は，複数の計算ノードで分割して行われ，各計算結果はデータサーバ上の複数のディスクにパラレルにセーブされる．しかし，データセーブ時は，各ディスクのサイズを考慮していないため，セーブ前にディスクサイズが不均等な状態でも，ほぼ同サイズの計算結果データを複数のディスクにセーブしてしまう．そのため，上記ディスクサイズの偏りを助長し，1 つのディスクに大きな負荷がかかることがある．ディスク容量を一定以上超えると，データ保存に失敗し，シミュレーション実行が不可能となり，処理が中断してしまう可能性もある．

そこで我々は，ディスクサイズを常に監視し，ユーザの代わりに負荷の大きなディスクから，負荷の小さいディスクへデータ移動を行い，データサーバ上のディスクサイズを均等な状態に保つエージェントを設計，実装する．

3 パラレルI/O制御エージェント

3.1 エージェントの構成

パラレルI/O制御エージェントは，以下の3つのエージェントによって図2のように構成される．

- データサーバ監視エージェント
  ディスクサイズが上限値を超えているかどうかを動的に調べる．
- ディスクサイズ均一化エージェント
  最適なディスクサイズになるように移動対象データの選択や移動後のディスクサイズの計算を行う．
- ユーザ返答確認エージェント
  ユーザからの要望を受け取り，要望に応じた適切な対応や行動をする．

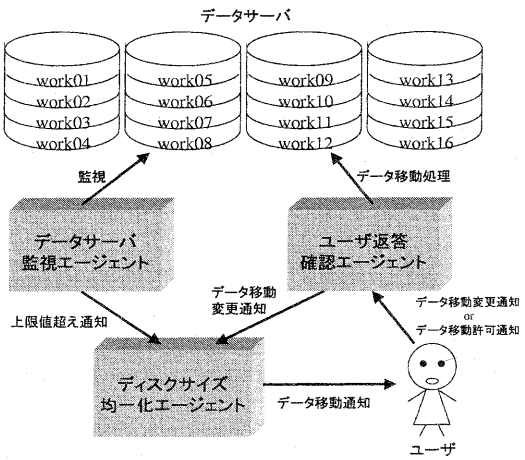

図 2: パラレルI/O制御エージェント構成

本稿で設計するパラレルI/O制御エージェントは，実際のデータサーバへの導入前のプロトタイプである．したがって図2のデータサーバは，実際のデータサーバ上のworkディスク領域に見立てたディレクトリを仮想的に作成したものである．各ディレクトリにはグループおよびユーザ用のサブディレクトリを作成する．

以下の節で，図2の各エージェントの機能や動作について詳しく説明する．

3.2 データサーバ監視エージェント

データサーバ監視エージェントは，データサーバを一定間隔で監視し，ディレクトリサイズが上限値を超えていないかを調べる．上限値を超えるディレクトリがある場合，ディスクサイズ均一化エージェントに通知する．監視対象は，work用ディレクトリとユーザー用ディレクトリである．上限値はkbyte単位で任意に指定することができる．

本エージェントが調べる情報は，以下のとおりである．

- work用ディレクトリ情報
  - ディレクトリサイズ
    ・上限値を越えているかどうかを判断するために用いる
    ・各ディスク使用量の平均値を求めるために用いる
    ・データ移動後のディレクトリサイズを計算するために用いる
- ユーザ用ディレクトリ情報
  - データサイズ
    各ユーザのwork用ディレクトリ使用率を調べるために用いる
  - ファイルサイズ
    データ移動対象ファイルを選択するために用いる
  - ファイル名
    データ移動時のコマンドの引数として用いる

上記の情報は，unixのコマンドであるduコマンドとlsコマンドを用いて調べる．

3.3 ディスクサイズ均一化エージェント

ディスクサイズ均一化エージェントは，データサーバ監視エージェント，およびユーザ返答確認エージェントから通知を受け取り，ユーザ要求を反映しながら，最適なディレクトリサイズを保つように移動対象データの選択とデータ移動後のwork用ディレクトリサイズの計算を行う．

データサーバ監視エージェントから，あるディレクトリのサイズが上限値を超えたという通知を受けると，現在のwork用ディレクトリサイズの平均値を求める．各ディレクトリサイズが求めた平均値に近付くように，平均値以下のサイズであるディレクトリにデータを移動させる．移動の対象となるデータは，上限値を超えたディレクトリ使用率が高いユーザが保持するデータから候補にあげられる．候補にあげられたデータ移動後の移動先ディレクトリサイズを計算し，平均値以下であれば，移動対象データとして確定する．

移動候補に選ばれたデータを保持するユーザに，上限値を超えたディレクトリにおける該当ユーザが保持するデータの占有率や，データの移動先をメー

ルで通知する．ユーザは，メールで通知されたURLにアクセスすることによって，現在の各work用ディレクトリサイズや，ユーザが保持するデータサイズの詳細情報を参照することができる．さらに，ディスクサイズ均一化エージェントが提示したデータ移動候補先ディレクトリ，移動データサイズを変更できる．

また，ユーザ返答確認エージェントから，ユーザーの移動先変更通知を受け取ると，データ移動後の移動先ディレクトリサイズを計算し，平均値以下であれば，ユーザの変更要求を反映した情報をメールで通知する．平均値以上であれば他の移動先ディレクトリ候補を挙げてメールで通知する．

3.4 ユーザ返答確認エージェント

ユーザ返答確認エージェントは，ユーザからの返答に対して，適切な対応を行う．

ユーザーからデータ移動先変更通知を受け取ると，ディレクトリ均一化エージェントに移動先ディレクトリ，移動データサイズを知らせる．

ユーザーからデータ移動許可通知を受け取ると，以下の手順でデータ移動を行う．

1. 現在のディレクトリから，該当ファイルを移動する際，移動先のディレクトリに同じ名前のファイルが存在する場合は，該当ファイル名を変更してから移動させる．

2. 元のファイルから移動先へシンボリックリンクを張る．

3.5 プロトタイプ実装

3.2節から3.4節で説明したデータサーバ用パラレルI/O制御エージェントを実装した．実装環境はOS windows2000, 開発言語はjavaでj2sdk1.4.1_01を使用している．メール送信に使用するライブラリはJavaMail ver1.3, JAF(JavaBeans Activation Framework) ver1.0.2である．

プロトタイプで用いる仮想work用ディレクトリは16個，各ディレクトリにはグループ用ディレクトリが3個，1グループにつきユーザ用ディレクトリが3個，合計9ユーザという設定である．ファイルは乱数を用いて適当なファイルサイズになるように作成し，ユーザ用ディレクトリにランダムに配置する．

エージェントの動作は3.2節から3.4節で説明した通りである．図3は，ディスクサイズ均一化エージェントがユーザへ通知したURL表示の例である．

実装の結果，プロトタイプを正確に動作させることが確認できた．このプロトタイプは，パラレルI/O制御エージェントフレームワークであり，今後は様々な機能を付け加え，実際のデータサーバ上での運用を目指す．

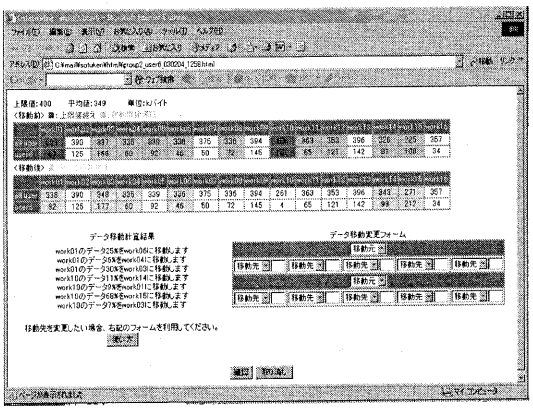

図 3: 詳細情報，ユーザ要求入力フォーム

4 まとめ

本稿では，日本原子力研究所 関西研究所 光量子科学研究センター の超並列計算機上で行われる大規模シミュレーションで用いるデータベース用パラレルI/O制御エージェントの設計と，そのプロトタイプ実装について報告した．実装の結果，パラレルI/O制御エージェントに必要な最小限の機能を実現することができた．

今後，より最適なパラレルI/O制御を行うエージェントを開発するためには，2つの課題があげられる．第一の課題は，各ユーザに応じてデータ移動先ディレクトリの選択などを行うために，各ユーザに関する知識ベースを構築することである．そのためには，各ユーザの使用データサイズや，どのディレクトリにセーブしているかなどを調査し，各ユーザ情報をプロファイル化する必要がある．

第二の課題は，最適なパラレルI/O制御を行うために，どのように各ユーザに関する知識ベースを利用するかというアルゴリズムを定義することである．

以上の課題を解決したパラレルI/O制御エージェントを完成させ，実際のデータサーバ上で運用させる予定である．

参考文献

[1] 松山仁美，松岡有希，小金山美賀，上島豊，城和貴: ”大規模シミュレーションサイクルを統治する知識自動獲得型エージェントプロトタイプの設計と実装”，情報処理学会 第42回 MPS 研究会, 2002-MPS-42, Vol.2002, No.114, pp.17-20 (2002).