2009－MPS－73 (5)
2009／3／5

仮想計算機を用いたジョブマイグレーションのＰＣグリッドへの適用

森川浩明[†1] 榎原博之[†2]
大西克実[†1] 中野秀男[†1]

PCグリッドは、並列計算システムのひとつとして、遊休PCの有効活用に利用されている。本研究では、仮想計算機のハードウエアに依存しない特徴に着目し、並列計算においてユーザが計算機を利用する場合に別の計算機に仮想計算機ごと計算内容を移行できるマイグレーション機能を実装したPCグリッドシステムの開発をおこなう。また、仮想計算機を利用したマイグレーションにかかる時間は、投入ジョブが利用するメモリサイズで決定されるため、メモリサイズを考慮したジョブ投入を検討する。

Job migration adapting PC grid with virtual machine

Hiroaki Morikawa,[†1] Hiroyuki Ebara,[†2] Katsumi Onishi[†1]
and Hideo Nakano[†1]

PC grid is one of the parallel computing system, and it makes the best use of idle PCs. In this paper, we propose an implementation of a PC grid system with migration function. If a person use the PC running a grid job, the migration function moves it to another idle PC. Furthermore, the time of migration with virtual machine depends on the fixed memory size. So, we propose a method to carry out jobs in considering of the memory size.

1. はじめに

現在、計算機の演算能力・通信機能の発展がめざましく、家庭用計算機複数台で一昔前のスーパーコンピュータに匹敵する計算能力を発揮できる。また、遺伝子解析や素粒子物理学などの研究分野で大規模な計算を必要とする問題が増大している。このような背景から近年の高速なネットワークを介してLANやWAN内に散在する家庭用計算機をつなぎ、PCクラスタシステムやPCグリッドシステムを構築することに注目が集まっている。

現在のPCグリッドシステムでは、SETI@home[3]などのようにある期間中に計算をおこなうプロジェクトが主流である。しかし、これらのプロジェクトでは計算機の使用状況を把握していないため、長時間のジョブ実行に不向きという問題やジョブ間の通信を扱えないなどの問題がある。

これら問題を解決するため、計算機の状態監視をおこない実行時間を考慮したジョブ投入にかかわる曽山の研究[1]や負荷分散とジョブ間の通信を可能にする立薗らの研究[2]などがある。

本研究では、常にユーザが存在する環境で長時間計算ジョブを実行するため仮想計算機を用いたマイグレーション機能を実装し、グリッドシステムの構築をおこなう。マイグレーション機能によってユーザが計算機の利用を開始すると計算内容を別の計算機に移行し、ジョブ実行を途切れることなく長時間のジョブ実行ができる。また、非同期の通信機能を備え、通信が必要なジョブも扱えるようになっている。

本研究では、構築したグリッドシステムの性能評価とマイグレーションの頻度を考慮してジョブ投入プログラムのサイズを変更することで効率的なジョブ投入をおこなうプログラムを提案し、計算機実験により、評価をおこなう。

以下、2章ではマイグレーション機能実装に用いた仮想計算機について述べる。3章では構築システムを4章で構築システムのマイグレーション機能の性能評価とマイグレーションを考慮したジョブ投入法の実験・評価と考察について述べる。

†1 大阪市立大学創造都市研究科
Graduate School for Creative Cities, Osaka City University
†2 関西大学システム理工学部
Faculty of Engineering Science, Kansai University

2. 仮想計算機

仮想計算機は、ホストOS上にメモリやCPU、通信回線などを仮想的に構築し、単一のホストOS上で仮想的に複数の計算機が動作しているかのように見せかけることのできる技術である。代表的な仮想化ソフトとしては、VMWare[4]やXen[5]、Jail[6]などがある。

仮想計算機は、仮想計算機イメージ(VMイメージ)と仮想化層（仮想化ソフト)、ハードウエアから構成されており（図1)、一般的に仮想計算機は仮想化層を通して間接的にハードウエアを操作している。

仮想計算機では、仮想化層によってハードウエアから切り離させているために以下のような特徴を備えている。

- 複数の仮想計算機を起動できる
  仮想化層によるハードウエアの排他的な使用によって1台の計算機が複数台の計算機であるかのように見せかけられる。サーバなどでは、資源の有効活用のため導入されている。
- ハードウエアに依存しない
  ハードウエアの操作は仮想化層でおこなわれるため、仮想計算機にはハードウエアの影響が少ない。このため、異なるハードウエア環境にVMイメージを転送しても仮想化ソフトが同じであるならば動作可能である。
- ホストOSからの独立性
  仮想計算機はホストOSが何であっても動作可能であり、一部の仮想計算ソフトではホストOSが存在しない環境であってもゲストOSが起動できる。

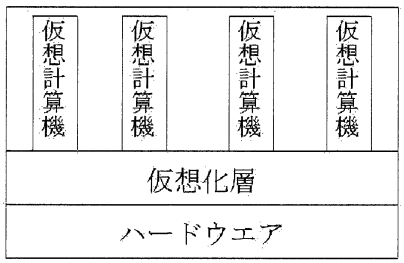

図1 仮想計算機の構造

2.1 マイグレーション機能

マイグレーション機能は、現在ある計算機上で実行中の計算内容や計算環境を別の計算機に移行させる機能である。マイグレーション機能は移動するデータなどによって以下の分類ができる。

(1) チェックポイントマイグレーション
一定期間、もしくは特定のアクションごとに現在実行中の状態（メモリ内容、レジスタ内容など）を保存（チェックポイント）し、障害発生時などに別計算機で保存した状態を展開する手法

(2) プロセスマイグレーション
メモリ内容などのデータを移動させるのではなくプロセス自体を別の計算機に移動させる手法しかし、移動した計算機で計算環境そのものが異なるとジョブ実行がうまくいかないという問題がある。

(3) ライブマイグレーション
計算機を停止させずメモリ内容などを少しずつ別計算機に移動させることで、外見上はある計算機が計算を停止したと同時にその計算機で実行していた内容を別計算機が引き継いで実行しているように見せかけることができる手法

現在、マイグレーション機能は主に仮想計算機によって実現されており、本研究でもマイグレーション機能の実装にVMware Serverを利用している。

3. 構築システム

3.1 構築システム概要

構築したグリッドシステムは関西大学と富士通研究所が共同開発したシステムでマスタサーバ1台と計算用PC7台を用いて、サーバ・クライアントからなるスター型のネットワーク構造を成している(図2)。また、これらシステムで利用している計算用PCにはそれぞれ使用者が存在し、研究や業務に使用している。本システムではユーザが存在する環境において効率的なジョブ投入をおこなうため、VM管理テーブルを利用したジョブ管理機能と仮想計算機によるマイグレーション機能を実装している。

グリッドシステムでは高速にマイグレーション機能を利用するため、VMイメージをコピーして送信するのではなく、マスタサーバが保持しているVMイメージを各計算機がネットワークドライブとして直接操作する。

3.2 マイグレーション機能

本システムのマイグレーション機能は、仮想計算機のサスペンド機能を用いて行われ、以下の条件のときに発生する。

- ユーザがマウス・キーボードを使用する（ログオンする）
- 障害などによるシャットダウン

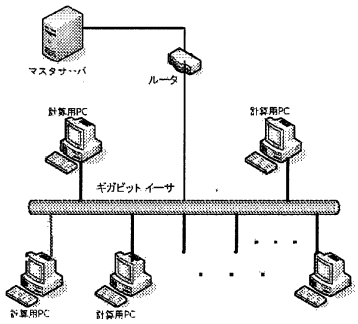

図 2 構築システムの全体図

上記の条件が発生すると、集中的に計算機を管理するマスタサーバにジョブを実行するゲスト OS のホスト OS が使用状況の変化をマスタサーバへ送信すると同時にログイン中のホスト OS のバックグラウンドでゲスト OS をサスペンドさせる。マイグレーションしたジョブはジョブキューの最後に登録されるため、VM 管理テーブルが満たされるかマイグレーションしていないジョブが全て終了した後ジョブの再投入がおこなわれる。

4. 構築グリッドシステムの性能評価とメモリ分割法

4.1 マイグレーション機能の性能評価

本研究では以下のスペックの計算用 PC(ホスト OS) 4 台と各計算用 PC に構築するゲスト OS、マスタサーバ 1 台を用いてシステムの評価と実験をおこなう。

表 1 マスタサーバ

OS	RedHatEnterPriseLinux4
CPU	Intel Xeon 1.86GHz
Memory	2GB

表 2 ホスト OS

OS	Windows XP Professional SP2
CPU	Intel Core2 Duo 2.4GHz
Memory	2GB
仮想計算機	VMware Server1.0.5
分散環境	Systemwalker Cyber GRIP

本システムのゲスト OS では、メモリと CPU をホスト OS の半分しか使用していないが、これはユーザがログオンしてきた際に、バックグラウンドで実行するマイグレーション処理によってユーザのプログラム実行に影響を与えないためである。

表 3 ゲスト OS

OS	CentOS 4.4
CPU	1Unit
Memory	1GB
ネットワーク設定	NAT

マイグレーション時間はジョブ投入する計算機の探索時間と仮想計算機のサスペンド時間、再開時間の 3 つからなる。マイグレーション時間の取得では、n 次正方行列の乗算と同一計算で消費するメモリ領域（実メモリとスワップ領域の和）の実メモリ領域を確保する malloc プログラムを用いて、2048 次正方行列から 4096 次正方行列までのプロセスがマイグレーションした際にかかる時間（表 4）を実験により求め、実際の投入ジョブで消費するメモリ・スワップ領域からマイグレーション時間を得る。

表 4 メモリ・スワップ領域ごとのマイグレーションにかかる時間 (単位:sec)

問題 (メモリ) (スワップ)	2048 (80 (100)	2986 (145) (199)	3547 (212) (297)	4096 (280) (395)
malloc	209	218.5	265	264.8
行列演算	392.2	416.8	420.2	436.0
差	183.2	198.3	155.2	171.2

表 4 により、スワップ領域を利用した場合、スワップを利用しなかったプログラムよりも 180 秒程度マイグレーションにかかる時間が延びていることがわかる。

本システムのマイグレーション時間は表 4 より実システム上で 200 秒から 450 秒かかるが、標準のマイグレーション機能よりも高速なマイグレーションができる。加えて、障害によって計算機がシャットダウンするときでも仮想計算機のサスペンド状態の遷移の報告には 90 秒程度で済むため、障害対策としても十分機能する。

4.2 メモリ分割法

マイグレーションにかかる時間は、仮想計算機のサスペンド時間と VM イメージ取得時間、仮想計算機の再開時間によって決定される。このうち、サスペンド時間と再開時間は表 4 のように消費するメモリ領域（スワップも含む）で変化する。

メモリ分割法では、並列プログラムが消費するメモリ領域でのジョブ分割を自動でおこなう。消費するメモリ領域での分割では、消費するスワップ領域の削減によってマイグレーションにかかる時間の削減も可能である。加えて、投入ジョブが小さくなることで実行時間の長いジョブが停滞して他の並列計算に影響を与えることを最小限にできる。

反面、最終結果を取りまとめる統合演算やファイル分割による通信オーバーヘッドの増大といった問題を考慮しなければならない。ユーザの利用頻度（マイグレーション頻度）が低い状況においてメモリ分割法を利用した場合、ジョブ分割数によっては計算用PCの初回投入遅延によって通常よりも実行時間が延びる。

本実験では8192次正方行列の乗算を4台の計算用PCを用いて並列計算する。行列演算は$O(n^3)$の問題であり、対象となる行列をx分割すると$O(x^3)$*$O(\left(\frac{n}{x}\right)^3)$ + $O(n^2)$となり、行列演算の最終処理に$O(n^2)$の演算がかかる。

本実験ではマイグレーション機能を効率的に使用するため、前節で述べた実行ジョブのメモリ分割法を計算用PCの使用状況に応じて投入する。メモリ分割法は4096から1024次正方行列に分割した場合の最小の実行時間を選択する。実行ジョブをユーザの使用状況を考慮せずただ計算用PC数と同数だけ分割したジョブを投入した場合（4分割）とメモリ分割法を比べて、実行ジョブの終了時間にどの程度影響が出るか実験をおこなう。

グリッドシステムを利用し、メモリ分割法とマイグレーションを考慮しないジョブ分割でマイグレーション頻度を変化させながら自作プログラムによる実験おこない、実験結果を図3に示す。

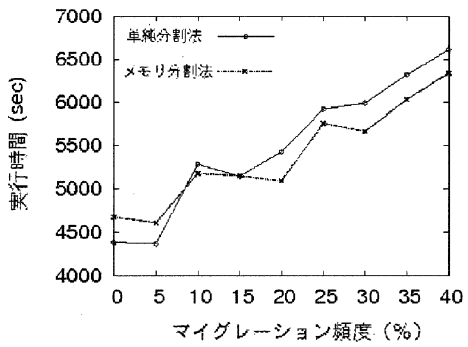

図3 ユーザの利用を考慮しないジョブ投入とメモリ分割法の比較

結果より、マイグレーション頻度の増加に比例して実行時間が増加したことがわかる。メモリ分割法においてマイグレーションが発生しにくい環境では、実行時間が考慮なしと比べて多くなったが、マイグレーションが発生しやすい環境では、実行時間が考慮なしよりもよい結果となった。

結果よりスワップ領域を多く消費するジョブであるほど計算機の使用状況に応じてジョブ分割をおこなったほうが、PCグリッドを利用するにあたり有効であると考えられる。このため、並列計算におけるスワップ領域の使用はあらかじめおこなわないのが賢明であるが、メモリ領域の利用はOSやコンパイラに依存する場合が多く、あらかじめスワップ領域をオフにする行為もマルチスレッドプログラムの普及などで実行が困難である。

5. おわりに

本研究では、遊休計算機を有効活用するために計算途中のジョブを他の計算機にマイグレーションできる機能を有したPCグリッドシステムの構築をおこない、その評価をおこなった。その結果、高速なマイグレーションが実現できるグリッドシステムであることが示された。また、マイグレーション機能を考慮してジョブ分割をおこない評価をおこない、その結果から効率のよいジョブ投入ができた。

謝　辞

本研究をおこなうにあたり、多くの人々にご協力をいただいた。研究のシステム構築に携わってくださった富士通研究所の皆様、研究の場を提供してくださった関西大学ソシオネットワーク戦略研究センターの皆様に感謝の意を表す。

また、本研究は、平成20年度関西大学重点領域研究助成金において、研究課題「休止中のコンピュータを有効利用するグリッドシステムの構築とその応用」として研究費を受け、その成果を公表するものである。

参考文献

1) 曽山豊: 企業におけるグリッド・コンピューティングの活用とその成果, グリッド協議会セッション,Grid World 2006(2006).
2) 立薗真樹, 中田秀基, 松岡聡: 仮想計算機を用いたグリッド上でのMPI実行環境, SACSIS 2006, (2005).
3) SETI@home: http://setiathome.berkeley.edu/.
4) VMware: http://www.vmware.com/.
5) Xen: http://www.xen.org/.
6) FreeBSD jail http://www.onlamp.com/pub/a/bsd/2003/09/04/jails.html/.
7) 合田憲人・関口智嗣 編著: グリッド技術入門, コロナ社, (2008)
8) ITpro 編: すべてわかる仮想化大全 2009, 日経BP, (2008).