分散 PC グリッドシステムの構築
Construction of Distributed PC Grid System

梅本潤志 * 榎原博之 †
Junji Umemoto Hiroyuki Ebara

1 はじめに

コンピュータのめざましい発展に伴い、家庭や会社で使われている PC でさえ高度な科学技術計算が可能となっている。これらの PC を複数台使えば、ある程度の規模の高速計算が可能であると思われる。さらに、それらの PC は 24 時間稼働しているわけではなく、それらの PC の遊休時間を利用すれば、安価で高速計算が可能となる。特に、PC グリッドシステムは、高性能な PC がその性能をフルに発揮していない点を考慮し、PC の有効活用を目的として、手軽に計算パワーを得ることができるシステムを目指している。

著者らが所属する研究室では、停止中や遊休の PC をグリッドシステムの計算サーバとして利用できるシステムを開発した [4] 。このシステムは、大学内のコンピュータ演習室などにある LAN に繋がった PC の有効利用を想定しており、グリッドサーバが休止している計算機を見つけ、ネットワークを介して起動させる。さらに、マイグレーション機能によって演習室内でユーザが利用を開始すると計算を行っている仮想計算機をサスペンドさせ、計算内容を他の計算機に移行させる機能を持っている。この機能により、長時間ジョブの実行が可能である。

本研究では、PC クラスタをインターネット（WAN）上に分散配置した分散 PC グリッドシステムを提案する。提案システムでは、インターネット（WAN）上に分散配置された PC クラスタを管理サーバであるグリッドエージェントによってつなぎ 1 つのシステムとして機能させる。グリッドエージェント間で通信を行うことにより、効率の良い計算資源配分と通信負荷の軽減を実現する。また、各グリッドエージェントにはグリッドポータルの機能を持たせ、各地点で投入した計算ジョブは地点に関係なく効率良く分散実行できる。

2 グリッドシステム

2.1 PC グリッドシステム

グリッドシステムの定義は、元来、電力系パワーグリッドに例えられるようにネットワークに接続するだけで、誰もが安価にあるいは無料で計算パワーを利用できるシステムである [3][1]。しかし、実際に開発されているグリッドシステムは、専用機にグリッドコンピューティング用ソフトウェア（ミドルウェア）を導入した PC クラスタシステムの延長線上のものをインターネットなどの広域ネットワーク（WAN）に接続したもので、一般ユーザが計算パワーを自由に使えるものとは到底言い難い。

PC グリッドシステムは、家庭などで使われている PC を使って大規模計算を行うシステムと定義されている。しかし、現在のところ、SETI@home プロジェクト [9] に代表されるように、PC 側がサーバにあるソフトウェアをダウンロードしてバックグラウンドで実行するしくみで、PC 起動中に遊休状態になったとき、そのソフトウェアがサーバからデータを取得し、計算し、実行結果をサーバに返すしくみとなっている。このため、独立した小問題に分割できる問題にしか適用できず、プロセス間で通信を行う並列計算には向かない。また、停止中の PC は、計算ジョブを実行できない。

2.1.1 SCore

SCore[2] とは、技術研究組合新情報処理開発機構において開発された PC クラスタのための、HPC(High Performance Computing) 型システムソフトウェアである。HPC クラスタでは、ネットワーク接続する計算機の台数を増やし、計算処理を各計算機に分散させることで全体の計算処理速度を高速にすることを目的としている。SCore は以下のような特徴をもっている。

* 関西大学 大学院 理工学研究科, Kansai University.
† 関西大学 システム理工学部 電気電子情報工学科, Kansai University.

- 高性能通信ライブラリ
- 効率の良いコンピュータ管理
- 高いユーザビリティ

SCore は独自に開発された PM2 高性能通信ライブラリを用いて並列計算を行っている。これにより、Ethernet でつながれた小規模クラスタから、Giga bit Ethernet や Myrinet[7] のようなネットワークで構築された大規模クラスタまで、シームレスに利用することが可能となる。OpenMP や MPI、MPC++ といった並列プログラミング環境も、これらの通信ライブラリの上で実装されている。

2.2 並列プログラミング

並列プログラミングにはいくつかの手法がある。メモリ全体を CPU 同士が共有する共有メモリ型と、各 CPU にメモリが分散している分散メモリ型がある。共有メモリ型と分散メモリ型の並列プログラミングの標準 API として MPI や OpenMP、MPC++ などが挙げられる。

2.2.1 MPI

MPI（Message Passing Interface）[5] とは、メッセージパッシング方式に基づいた仕様で分散メモリ型の並列計算を行うために複数のプロセス間でのデータのやりとりに用いられるライブラリである。メッセージパッシング方式とは、あるプロセスから他のプロセスへデータを明示的に送る方法である。各プロセッサ上で並列に動作するプロセスがそれぞれ独立したアドレス空間を持っている分散メモリ型のクラスタ環境では、並列化インターフェースとしてメッセージパッシング方式が用いられている。SCore では MPICH2[6] という形で、MPI が実装されている。

2.2.2 OpenMP

OpenMP[8] とは、OpenMP Architecture Review Board (ARB) によって業界標準規格に規定された、共有メモリ型で並列計算を行うためのライブラリである。MPI ではプロセス間でのデータのやり取りをプログラム中に明示的に記述しなければいけないのに対し、OpenMP では逐次プログラムから段階的に並列化することが可能である。また、OpenMP をサポートしない環境では OpenMP を無効にし逐次処理で実行することができる。そのため、OpenMP で共有メモリ型で並列化されたものを、MPI でさらに並列化させることが可能である。

2.3 仮想マシン

仮想マシンとは、コンピュータ上に CPU やメモリ、通信回線などを仮想的に構築し、コンピュータ内に複数の仮想的なコンピュータを構築する技術である。仮想マシンは、ハードウェア上に仮想化層を構築し、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。

2.3.1 Xen

Xen は、高い性能と機能を持ちオープンソースソフトウェアで構成されている、ハイパーバイザタイプの仮想化ソフトウェアである。ハイパーバイザタイプとは仮想化層にハイパーバイザと呼ばれる小さなプログラムを動かし、すべての OS（ゲスト OS）はそのハイパーバイザ上で動作する。また、Xen の仮想マシン環境はマルチプロセッサにも対応し、各ドメインに対して資源割り当て量を動的に変更することができる。メモリ量も同様に動的に割り当てることができるので、資源を効率的に管理することが可能となる。仮想マシンを物理マシン間で移動させるライブマイグレーションという機能もある。

3 分散 PC グリッドシステム

3.1 概略

本研究で提案する分散 PC グリッドシステムは、インターネット（WAN）上に分散配置されている PC クラスタをつなぎ、それらを一つのシステムとして機能するものである。PC クラスタ内の各 PC の状態を把握し、PC クラスタ間の情報交換のために、各 PC クラスタにグリッドエージェントと呼ぶ管理サーバを設置する。グリッドエージェントはグリッドポータルの機能も有し、各地点で計算ジョブの投入が可能である。

具体的には、図 1 に示すように、3 地点に PC クラスタとそれを管理するグリッドエージェントを設置し、グリッドエージェント間で通信することにより、効率の良い計算資源配分を実現する。

ユーザは、各地点のグリッドエージェント（グリッドポータル）に計算ジョブを投入する。投入された計算ジョブは、グリッドエージェント間の情報交換により、最も適した PC クラスタで実行される。投入された地点の PC クラスタで実行されるとは限らない。さらに、一つの PC クラスタでの実行が難しい場合は、複数の PC クラスタをまたがった実行も可能である。この実行は、プロセス間通信の必要が無いパラメータスイープ型のジョブだけでなく、プロセス間通信が必要なジョブでも実行できる。ただし、複数の地点での実行は、通信遅延などの問題が起こる可能性があるため、ユーザ側であらかじめ、PC クラスタを分けることができる範囲を指定し、PC クラスタ内通信と PC クラスタ間

通信を区分する必要がある。

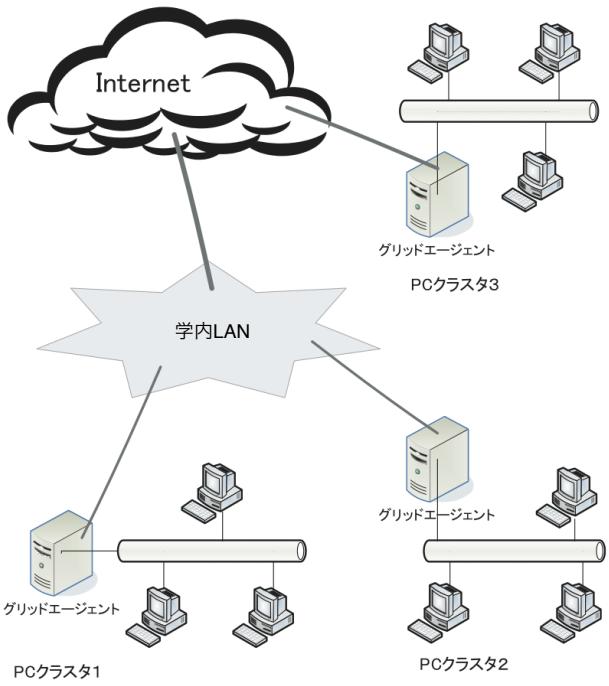

図 1　分散 PC グリッドシステム

これらを実現するために、グリッドエージェントは以下のような機能を備えている。

- 各 PC のコア単位での負荷状況の把握
- 計算ジョブの転送
- プロセス間通信の仲介
- マイグレーションされるジョブイメージの転送
- グリッドポータル機能

3.2　特徴

提案する分散 PC グリッドシステムは、PC クラスタ間でのプロセス間通信を実現することにより、PC クラスタをまたがった効率の良い計算資源配分が可能となる。さらに、PC の CPU がマルチコア化していることから、コア単位での PC の負荷状況を把握し、コア単位での計算ジョブの投入や PC 単位でのコア分散を考慮した並列計算ジョブの投入を可能とする。また、プロセス間通信は、暗号化によりセキュリティを保証する。

本システムは、仮想マシン上で構築するため、計算ジョブのマイグレーションが可能である。ジョブ実行中の PC を途中で中断させ、他の PC へマイグレーションさせることにより、他のジョブを実行させたり、ユーザが単に PC として利用したりすることが可能である。この機能により、大学などの計算機室の PC をオープン利用時でも計算サーバとして利用することが可能となる。結果として、長時間のジョブも実行可能である。

提案する分散 PC グリッドシステムの特徴を以下に列挙する。

- 効率の良い計算資源配分
- マルチコアを考慮したジョブ投入
- 暗号化によるグリッドエージェント間通信
- 計算ジョブのマイグレーション
- ジョブ実行中 PC の途中中断
- マイグレーションによる長時間ジョブの実行
- 休止中 PC の電源投入
- グリッドエージェント間は対等
- ジョブイメージの効率の良い転送
- 分散配置されたグリッドポータル

3.3　分散 PC グリッドシステムの開発

提案する分散 PC グリッドシステムの構成は、複数のグリッドエージェントがインターネット上に分散配置されており、各グリッドエージェントは Ethernet Switch を介して PC クラスタを構成している。分散 PC グリッドシステムのそれぞれの機能の開発状況について以下に述べる。

3.3.1　グリッドエージェントと PC クラスタ

管理サーバであるグリッドエージェントは、ローカル PC クラスタ内の各仮想計算機の計算資源やジョブの実行をコア単位で管理し、クラスタ内全体の計算資源を効率よく利用できるように計算資源の配分を行う。計算資源を配分するとき、各 PC クラスタ内で仮想計算機がマイグレーションするための空き PC を一定数確保することも考慮に入るようにする必要がある。グリッドエージェント同士は、クライアント-サーバ方式のような上下関係を持たずに、P2P（Peer to Peer）方式同様の対等な関係で密に通信を行い、それぞれの利用状況などの計算資源情報を常に共有する。グリッドエージェントと各仮想計算機には、OpenMP や MPI、MPC++ などの並列計算ライブラリをインストールして並列コンピューティング環境を構築する。グリッドエージェント同士が各 PC クラスタ間の通信を仲介することで、複数クラスタ環境でのジョブの実行や計算ジョブの転送が可能となる。

3.3.2　計算ジョブのマイグレーション

各計算機は仮想マシンソフトである Xen を利用し、準仮想化で仮想マシンモニタと仮想計算機の 2 つの仮想マシンを動作させる。Xen の準仮想化技術を利用することにより、エミュレーションのオーバヘッドを最小限に抑えるこ

とができ、物理ハードウェア上での動作時と遜色の無い性能を引き出すことが可能となる。また、Xen のマイグレーション機能を使用することで、ジョブ実行中の仮想計算機を途中中断し、他の PC に移動させることで、引き続きジョブを行うことができる。もし起動中の PC がすべて利用中でありマイグレーションできない場合も、休止中の PC を Wake on Lan で起動することで新たにマイグレーション先を確保し、ジョブの実行が途中で終了してしまわないようにすることができる。

3.3.3 暗号化によるグリッドエージェント間通信

PC クラスタ内では、高速にかつ密に通信を行う MPI などが十分に性能を発揮できるように、セキュリティレベルは低く設定しなければならない。しかし、インターネットを介する際には、他者からの介入や安全保持のためにセキュリティレベルは高く維持する必要がある。そのため、本システムでは通信に暗号や認証技術を利用して安全にリモートコンピュータと通信できる、SSH（Secure Shell）を使用することで通信経路の安全を確保する。SSH によって通信経路全体の暗号化を行うことで、通信内容の漏洩を防ぐ。

また、各グリッドエージェントへの外部からの通信は、SSH が使用するポート以外は遮断し PC クラスタ内の安全を維持する。しかし、これではグリッドエージェント間での通信も遮断されてしまい、PC クラスタ間で相互に通信することができない。そのために、各 PC クラスタ間で通信が必要となるときには、SSH の機能の一つであるポートフォアリングを用いてグリッドエージェントの TCP 接続を暗号化し、グリッドエージェントを介して PC クラスタ間の通信を行う。これにより、必要最小限のポートのみ解放することで、インターネット側のセキュリティを高く保持したまま通信をすることが可能となる。

3.3.4 グリッドポータル機能

各グリッドエージェントは、リモートアクセスや Web を利用してポータルの役割を果たす。グリッドポータルでは、使用する計算資源、例えば CPU コアやメモリ量などを指定してジョブを投入する。グリッドポータルから投入されたジョブは、各グリッドエージェントが管理している PC クラスタの状況に合わせてジョブを実行する PC クラスタを決定し、使用する計算資源を確保する。このとき、ジョブの実行先はブラックボックス化されているため、ユーザがジョブを投入したグリッドエージェントが管理している PC クラスタ以外の遠隔の PC クラスタで実行されることもある。しかし、計算資源を 1 つのクラスタで確保することができず、複数のクラスタにまたがってしか実行環境を確保できない場合やそのような環境で実行したい場合などは、WAN を介した実行環境となるため通信遅延が起こる。そのため、複数クラスタの使用については、ユーザ側で PC クラスタの範囲の指定や実行を許可するかをゆだねる必要がある。

4 まとめ

提案する分散 PC グリッドシステムは、管理サーバであるグリッドエージェントによってインターネット上に分散している複数の PC クラスタをつなぎ 1 つのシステムとして機能する。分散 PC グリッドシステムでは、1 つの PC クラスタ内だけでなく複数のクラスタにわたるような環境でも計算資源を確保し、ジョブを実行することができる。グリッドエージェントがそれぞれの PC クラスタの計算資源情報及び利用状況を管理・把握し、グリッドエージェント同士で情報を共有することで複数クラスタをまたぐ資源管理が可能となり、効率的な計算資源配分をすることができる。また、仮想環境で構築することから、遊休 PC を計算資源として活用することもでき、マイグレーション機能によって計算ジョブをクラスタ内で移動させて長時間のジョブを実行することが可能である。ユーザはグリッドエージェントのポータル機能を用いることで、ローカル PC クラスタの利用状況によらず計算資源を確保しジョブを投入することができる。

参考文献

[1] 合田憲人, 関口智嗣編著. グリッド技術入門. コロナ社, 2008.

[2] PC Cluster Consortium. http://www.pccluster.org/.

[3] Ian Foster and Carl Kesselman. *The Grid 2.* Morgan Kaufmann, November 2003.

[4] 森川浩明, 榎原博之, 大西克実, 中野秀男. 仮想計算機を適用した PC グリッドの開発と性能評価. *Vol.J93-D,No.8,Aug. 2010.* 掲載予定.

[5] MPI. http://www.mpi-forum.org/.

[6] MPICH. http://www.mcs.anl.gov/research/projects/mpich2/.

[7] Myrinet. http://www.myri.com/.

[8] OpenMP. http://openmp.org/wp/.

[9] SETI@home. http://setiathome.berkeley.edu/.