# 並列データ処理アプリケーション実行時の仮想マシンPCクラスタの動作解析

豊島詩織[†1]　原明日香[*1]　小口正人[†1]

情報量が爆発的に増加している今日においてITコストの増大が問題になっており、データの蓄積や管理を効率よく行なうことが重要になっている。本研究では計算機資源の効率的な運用方法として、各ノードに汎用のパーソナルコンピュータとネットワークを用いたPCクラスタに対し仮想化技術を適用した仮想マシンPCクラスタを構築した。通信には次世代SANとして注目されるIP-SANを導入することで、サーバとストレージ間の広域環境における通信を低コストで実現することができ、遠隔地への通信も容易となる。これにより例えばクラウドコンピューティングの枠組における遠隔計算機リソースの利用が期待できる。本稿では構築したシステム上でデータマイニングの一種である相関関係抽出のアプリケーションを動作させ、システムのモニタを行い、遠隔アクセスを考慮したiSCSI通信をしたときの仮想マシンPCクラスタの振舞を解析する。

# Analysis of system behavior of Virtual PC cluster when parallel date processing application is executed

Shiori Toyoshima,[†1] Asuka Hara[*1] and Masato Oguchi[†1]

The increase of the IT cost becomes the problem today when volume of information increases explosively, and it becomes important to store and manage of data efficiently. In this study , we built the Virtual machine PC cluster in which virtualization is applied to the PC cluster, which used a general-purpose personal computer and network for each node as effective usage of computer resource. We introduced IP-SAN which attracted attention as SAN in the next generation and used a network storage. It is used at low cost communications in the wide area environment between servers and storage, and can be expected that some of the features available in the cloud computing cluster. In this paper, we execute an application of the parallel association rule mining that is a kind of the data mining on the system, which we built and monitor the system and analyze the behavior of the Virtual machine PC cluster when iscsi remote access to communications.

## 1. はじめに

情報通信技術や情報社会の発展に伴い、コンピュータシステムに蓄積される情報量が急速に増えている。利用可能な情報量は増加するが、それを扱うことはますます難しくなっている中でデータを効率的に管理・処理できるシステムが求められている。そこで注目が集まっているのが仮想化技術である。仮想化により物理的な制約を考慮することなく，アプリケーションやサービスのために必要なリソースが必要なだけ利用可能になり、ハードウェアの保有コストや、システムの複雑性やそれに伴う運用管理コストの削減が期待される。

本研究では仮想化の一つであるサーバ仮想化を利用した。従来のIT環境では、新たな業務やサービスが増えるたびに次々とシステムの拡張やサーバの増設を繰り返していた。サーバ仮想化により1台のサーバに独立した複数台のサーバを動作させることができる。ピーク時間が重ならないアプリケーションを特定し、仮想サーバ上に処理を移行させることでサーバの物理的な台数を減らすことができ、システム使用率の向上も図ることができる。本研究では仮想化にXenを使用した[1]。Xenはオープンソースで、複数のOSを動かす為の基盤となるプラットフォームのみを提供する。そのため仮想マシンからリソースへアクセスする際に基本的にはホストOSを介さずに実行が可能で、仮想化による処理の性能低下が比較的小さい。

ストレージにはIP-SANを使用した。IP-SANはIPネットワークを用いてSANが構築でき、現在主に使われているFC-SANよりも通信距離や価格、管理の面でメリットがある。今後はデータセンタのような遠隔ストレージの利用だけではなく、現在注目されているクラウドコンピューティングなどの枠組を用い、計算機リソースもアウトソーシングすることが考えられる。そのため本研究ではIP-SANのプロトコルであるiSCSI(Internet SCSI)を使用し、広域ネットワーク環境での通信を考慮した仮想マシンPCクラスタを構築した。

近年クラスタ上では単なる計算処理よりも、より処理の重いデータ処理アプリケーションを動作させることが期待される。構築したクラスタ上でデータマイニングの一種である相関関係抽出のデータ処理アプリケーションを動作させ、実行時間を測定し、そのときのネットワークトラフィック、CPU使用率、メモリ使用率などの通信状況をモニタリングツール

†1 お茶の水女子大学
Ochanomizu University
*1 現在，特許庁
Presently with Japan Patent Office

である Ganglia を用いて観察し、仮想マシン PC クラスタの動作を解析した。

## 2. 仮想マシン PC クラスタ

### 2.1 サーバ仮想化

アプリケーションや用途ごとにハードウェアレベルで別々のサーバを用意することは、運用の安定性向上の面からも、システム管理の単純さを確保するうえでもメリットがある。そのため新たなサービスが増える度に次々とサーバの増設が繰り返され、管理負担の増大や、設置スペース・電力消費・発熱量の増大などが問題となってきている。

これに対しては 1 台のサーバ上に複数の仮想サーバを設け、そこに既存サーバを移行させることが効果的である (図 1)。仮想化技術を用いて、1 つのコンピュータ上で仮想的に複数のコンピュータが稼働しているようにシステムを構築することができる。この擬似的なコンピュータの一つひとつを仮想マシン (Virtual Machine) と呼び、ハイエンドのサーバ環境では仮想マシンを用いサーバ仮想化を実現することが主流になりつつある。サーバ仮想化では 1 台の物理サーバ上にコンピュータがあたかも複数台あるかのように扱え、1 台のコンピュータで別々の OS とアプリケーションを同時に稼動させられる。仮想マシンを利用することにより最新のハードウェアではサポート切れの OS も、最新ハードウェア上で仮想サーバとして動かすことが可能になる (図 2)。さらにシステム使用率のピークが異なる複数のシステムを仮想マシンとして同一サーバ上に移行させることによりシステム使用率の向上が期待される。

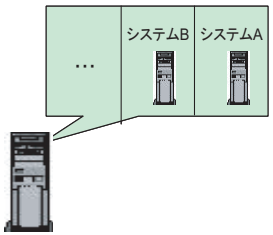


**図 1** サーバ集約

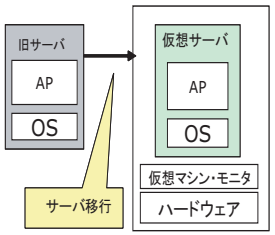


**図 2** サーバ移行

### 2.2 Xen

仮想化技術を用いて仮想マシンを実現するソフトウェアとして、VMware[2] や Virtual PC(Virtual Server)[3] などがある。これらはホスト OS 型と呼ばれ、ハードウェアとの入出力処理にホスト OS の機能を利用する。そのためオーバヘッドが大きく、処理性能が実機よりも低下するといった欠点がある。

一方 Xen はハイパーバイザ型の仮想化ソフトで、図 3 に示すように複数の OS を動かす為の基盤となるプラットフォームのみ提供する。ハイパーバイザ型では仮想化ソフトを通常の OS とは異なる特別な動作モードで動作させることで高いパフォーマンスや柔軟なリソース管理を実現しやすく、仮想化による処理の性能低下も比較的小さい。

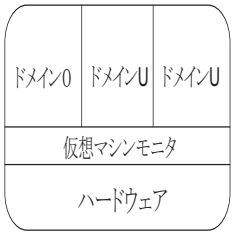


**図 3** xen の階層構造

Xen はイギリスケンブリッジ大学の研究プロジェクトから生まれ、Xen のソースコードはアメリカの XenSource 社という企業が管理しているが、開発はオープンソースコミュニティ主導で行われている。オープンソースとしては非常に高性能で、現在ビジネスユースにおいても用いられるようになっている。仮想マシンモニタが仮想化のための土台となり，その上で動いているのがドメインと呼ばれる仮想マシンである。そして，ホスト OS が動いているのがドメイン 0，ゲスト OS が動いているのがドメイン U である。ドメイン 0 は実ハードウェアへのアクセスやその他のドメインを管理する特権を持つ。

### 2.3 仮想マシン PC クラスタ

各ノードが独立して動作する CPU、メモリ、二次記憶を保有し、ノードが必要に応じてネットワークを介し互いに通信することで全体として並列分散処理を実現する分散メモリ型並列計算機において、各ノードに汎用のパーソナルコンピュータとネットワークを用いたものを PC クラスタという。汎用製品をそのまま利用できるため価格対性能比が優れており、また利用用途に応じて規模の拡大が容易である。

PC クラスタの構築および管理の一部自動化を行うクラスタリングソフトウェアとして Rocks がある [4]。UCSD で NPACI プロジェクトの一環として開発され、世界的に広く活用されている。本研究では、オプションとして仮想環境を作り出す Xen や、自動的に各ノー

ドからデータを収集し，それらを視覚的にグラフ化するモニタリングツールである Ganglia などをインストールした [5]。Rocks ではさまざまなサービスが動作するマスタノードである Front-end ノードから計算を行う Compute ノードへジョブの投入を行う。また仮想マシンの起動等も Front-end ノードより操作する。

## 3. ストレージの遠隔利用

### 3.1 遠隔アクセス

コンピュータシステムにおける処理データ量の増大に伴い，効率的にストレージを管理したいという要望が高まっている。現在注目を集めているクラウドコンピューティングは、インターネットを介しストレージを含む計算機リソースを利用できる。ユーザ側で全ての計算機リソースを揃えるより、導入・管理コストの削減が見込まれ今後は遠隔のデータセンタやクラウドコンピューティングの利用が増えると考えられる。その際にはじめから全てを外部のリソースでまかなうのではなく、ユーザのリソースを使いながら、まずはその一部をクラウドコンピューティングなど外部のサービスから借りる形が多用されると考えられる。本研究ではローカルのクラスタに加え、一部遠隔サイトのリソースの利用を想定した環境を構築した。

### 3.2 IP-SAN

HPC 分野では、PC クラスタの記憶装置において、計算ノードーストレージ間のバックエンドのネットワークに SAN を用いることが多くなっている。SAN は、分散したストレージをネットワークで統合し、集中管理とディスク資源の効率的な活用を可能にしている。

現在、SAN で使われる主流のネットワーク技術は高速な専用回線である Fibre Channel を用いる FC-SAN である。しかし FC-SAN では、FC 用のスイッチが高価であることなど、PC クラスタに導入して管理を行なうにはコスト面で障害がある。

SAN の中で次世代 SAN として期待されているのが IP ネットワークを用いた IP-SAN である．IP-SAN は Ethernet インタフェースと TCP/IP 対応ネットワークさえあれば導入でき、通常のネットワーク機器の流用が可能であることから導入コストが安価、管理が容易であるといったメリットがある。また専用網も含め広範囲に IP ネットワークのインフラが整備されているため長距離接続が可能で、広域ネットワークでの利用の期待が高まっている。本研究では iSCSI 接続の仮想マシン PC クラスタを構築し、評価を行なった。iSCSI(Internet SCSI)[6] は、IP-SAN の代表的なプロトコルであり、SCSI コマンドを TCP/IP パケットの中にカプセル化することでブロックレベルのデータ転送を行う。図 4 に iSCSI の階層構造を示す。

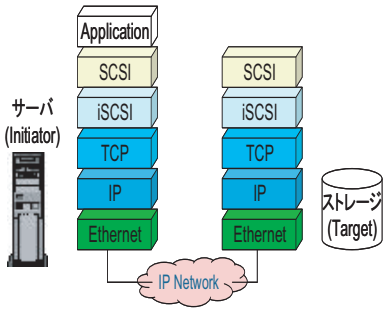


**図 4** iSCSI の階層構造

## 4. 研究内容

### 4.1 既存研究

文献 [4] では PC クラスタの記憶装置において、計算ノード-ストレージ間のバックエンドのネットワークに IP-SAN のプロトコルである iSCSI を使用し、そのフロントエンドとバックエンドのネットワークを同一の IP ネットワークに統合した IP-SAN 統合型 PC クラスタを構築している。IP ネットワークを使用することで安価にクラスタが作成でき、またフロントエンドとバックエンドが同じ IP ネットワークを使用することから構築および管理コストの削減が期待されるが、ノード間通信とストレージアクセスで同じネットワークリソースを使用するため、互いに衝突し、性能が低下する可能性が懸念される。このシステム上で相関関係抽出のアルゴリズムである Apriori アルゴリズムをハッシュ関数を使用して並列化した HPA(Hash Partitioned Apriori)[8] と、FP-growth アルゴリズムを並列化した PFP(Parallelized FP-growth)、バイオインフォマティクスにおいて用いられる科学技術計算の一種を並列化した mpiBLAST を動作させ、 IP-SAN 統合型クラスタの詳しい振舞を明らかにしている。評価を行なった範囲では iSCSI のネットワークを統合してもネットワークバウンドにはならないということが分かっている。

### 4.2 研究概要

本研究では、クラスタリングソフトウェアである Rocks を用いて、計算ノード数が 4 の仮想マシン PC クラスタを構築した。iSCSI のストレージ (Target) も 1 ノード用意し、同

一 IP ネットワークで接続する IP-SAN 統合型 PC クラスタとした実験環境を図 5 に示す。

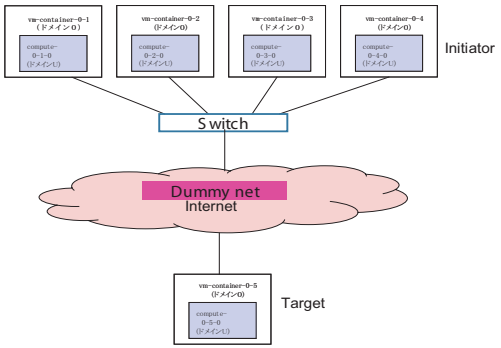


**図 5** 実験環境

Rocks の仮想マシンクラスタでは Compute ノードを vm-container と呼び、これが Xen におけるドメイン 0 となる。また vm-container の中にはそれぞれ compute と呼ばれる仮想マシンを一つずつ作成した。これが Xen におけるドメイン U となる。バージョンは仮想化に対応した Rocks5.0 を使用した。５台の PC は CPU が Intel(R) Xeon(TM) 3.60GHz でそれらを Gigabit Ethernet で接続した。メインメモリが 4GB、OS が Linux 2.6.18-53.1.14.el5xen(CentOS 5.0) である。計算ノードのメモリの振り分けは Rocks が自動で行なったものを用い、ドメイン 0 が 3GB、ドメイン U が 1GB である。

ストレージはローカルストレージに加え iSCSI を用いたネットワークストレージを使用し、Initiator をつなぐ Switch と Target 間には広域ネットワークを想定した人工的な遅延装置である Dummynet を挿入した。

そのクラスタ上で、まず基本性能測定としてハードディスクベンチマークの Bonnie++[8] によりローカルおよび iSCSI アクセス、また 10msec の遅延を含む iSCSI アクセスのディスクアクセス性能を測定した。

次にデータマイニングの並列アプリケーションを動作させた。

データベースの膨大なデータから、その中に潜む項目間の相関関係やパターンなど技術有用な情報を抽出するデータマイニングの一種として相関関係抽出が知られている。有益な規則性や関係を抽出するために、データ集合の中から高頻度で発生する特徴的なパターンを見つける。あるパターンが現れる頻度 (サポート値) が多ければ、そこから得られる関係は相関があると見なされる。相関関係抽出で扱うデータは巨大であることが多く、効率の良い処理には並列化が不可欠であるといえる。相関関係抽出の代表的なアルゴリズムとして Apriori アルゴリズムが挙げられる。Apriori アルゴリズムは信頼度（confidence）と支持度（support）に基づいてルールを評価し、「パンとバターを購入した取引の 90 %がミルクも購入している」などといった、大量のデータの中に隠れた規則や関係を抽出する。信頼度とは、X が発生したときに、Y が起こる割合を示す。この数値が高いほど、X が発生したときに Y が起こるというルールが強いことを意味する。一方、支持度とは、X と Y を同時に満たすトランザクションが全トランザクションに占める割合をいう。つまり、ルールそのものの出現率である。信頼度と支持度に閾値を設け（ユーザーが指定)、それを超える信頼度と支持度を持つルールを相関があると見なし、候補アイテムセット（ルールとして抽出される候補）から頻出アイテムセットを抽出するという動作を繰り返し行なう。Apriori をベースにした並列相関関係抽出のアルゴリズムはいくつか提案されているが、本研究ではハッシュ関数を使用して Apriori を並列化する HPA を用いる。HPA は候補アイテムセットを各ノードに分割し、その後全ノード対全ノード通信により頻出アイテムセットを繰り返し検索していく。ノード間通信や、繰り返し計算のためネットワーク通信量は多く、ディスク I/O も繰り返し行なわれるといった特徴を持つ。

## 5. 基本性能測定

### 5.1 ストレージアクセス性能

ハードディスクベンチマークツールの Bonnie++[9] を用いて、domain0、domainU について local および iSCSI アクセス、また iSCSI アクセスで片道遅延 10msec の場合のディスク I/O 性能測定を行なった。Write の結果を図 6 に示す。

まず domain0 では性能のよい順に local、iSCSI、片道 10msec 遅延 iSCSI の順になった。domainU では local よりも iSCSI のほうが早くなった。遅延のない時には iSCSI はかなり高いアクセス性能であったが、遅延を入れると性能が大きく落ちることが分かる。

## 6. 並列データマイニング実行

HPA アルゴリズムについてアイテム数を 1000 とし、トランザクション数が 1M、8M、10M、20M、最小支持度を 0.7 %に設定し、クラスタの Compute ノードのドメイン 0 だけにジョブを与えた場合（以下システム A）とドメイン U にだけジョブを与えた場合（シ

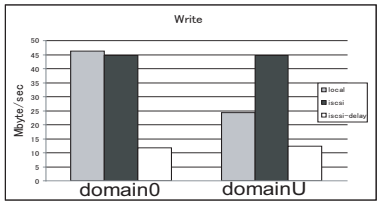


**図 6** ディスクアクセス性能 (Write)

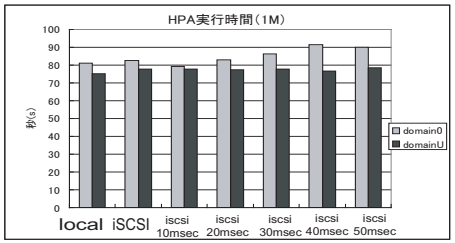


**図 7** HPA 実行時間 (1M トランザクション)

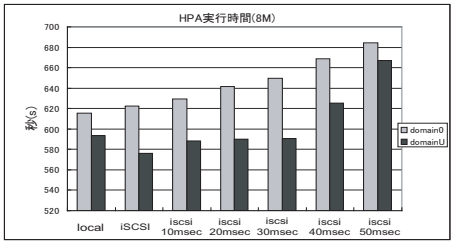


**図 8** HPA 実行時間 (8M トランザクション)

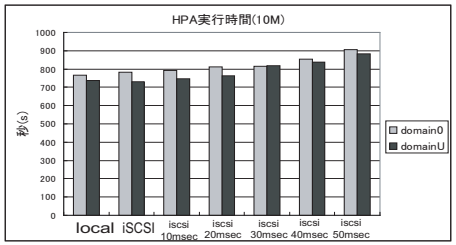


**図 9** HPA 実行時間 (10M トランザクション)

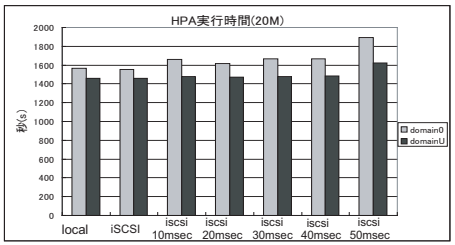


**図 10** HPA 実行時間 (20M トランザクション)

ステム B）で、それぞれ local および iSCSI ストレージを使用し、iSCSI は遅延なしおよび 10msec〜50msec の遅延を含む場合の実行時間を測定する。1 トランザクションは約 50Byte、計算ノード数が 4 台より、トランザクション数が 1M、8M、10M、20M のときの 1 ノードあたりのデータサイズはそれぞれ 12.5MByte、100MByte、125MByte、250MByte となる。

またドメイン 0 のみを用いたシステム A、ドメイン U のみを用いたシステム B においてそれぞれ最もデータ処理量が多いトランザクション数が 20M、片道遅延時間が 50msec のときの CPU 使用率、ネットワークトラフィック、メモリ使用率をモニタリングツールである Ganglia を用いて観察し、そのときの、システムの振舞を解析する。

図 7〜10 にトランザクション数が 1M〜20M の HPA アルゴリズムの実行時間を示す。1M〜20M まで local、iSCSI、iSCSI 通信で 10msec〜50msec と遅延をいれるとそれに伴い実行時間が徐々に長くなることが分かる。iSCSI のアクセス性能がかなり高いことから、遅延がない iSCSI 通信は local と実行時間はほぼ変わらない結果になった。また HPA は全体の処理の中では I/O の重みがそれ程大きくないため、遠隔にしても local と iSCSI ではあまり大きな差は見られないということが分かった。また domain0 と domainU を比べると domainU のほうが速いという結果になった。

図 11、12 に domain0、domainU それぞれで実行した際の CPU 使用率を示す。20M とトランザクション数が多いためどちらの場合も CPU をほぼ 100 %使用していることが分かる。HPA は候補アイテムセットから頻出アイテムセットを生成することを繰り返す。この比較演算処理のため計算量が多くなると考えられる。

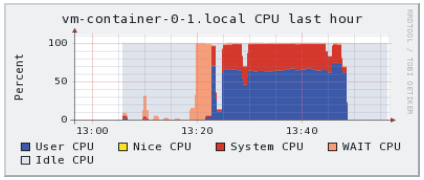


**図 11** domainU で実行した際の CPU 使用率

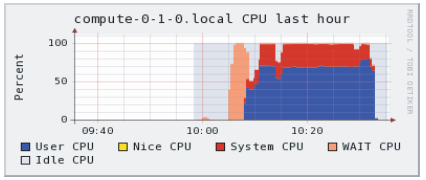


**図 12** domainU で実行した際の CPU 使用率

図 13、14 に domain0、domainU それぞれで実行した際のメモリ使用率を示す。どちらの場合も割り当てられたメモリは使い切っておらず、メモリ不足が原因となる性能低下は起こっていないと予測される。

図 15〜17 にこのときのネットワーク使用率を示す。domain0 で実行した際の最大ネットワーク帯域は約 60MByte/sec、domainU で実行した際の domain0 と domainU の最大ネットワーク帯域はそれぞれ約 45MByte/sec と約 24MByte/sec で合計は 69MByte/sec となり、ここでも domainU で実行した場合のほうが通信効率がよくなっていることが分かる。

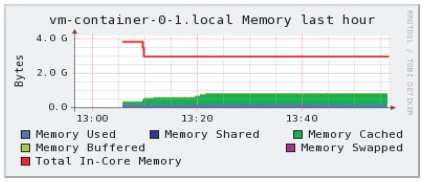


**図 13** domainU で実行した際のメモリ使用率)

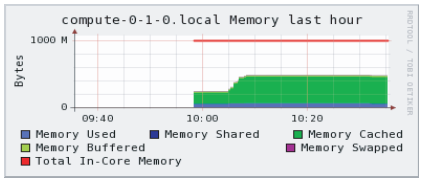


**図 14** domainU で実行した際のメモリ使用率

これは Xen において domainU を使用することで I/O や通信処理などのシステムの最適化が行なわれているためと考えられる [10],[11]。

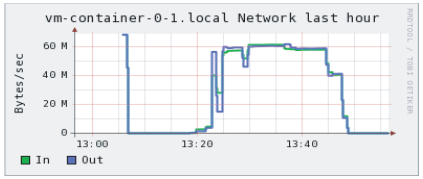


**図 15** domain0 で実行した際のネットワーク帯域

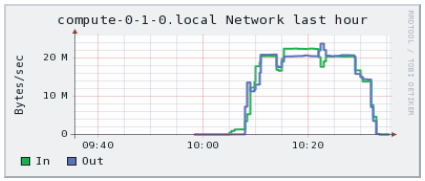


**図 16** domainU で実行した際の domain0 の NW 帯域

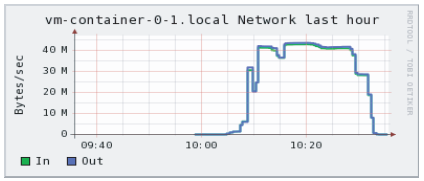


**図 17** domainU で実行した際の domainU の NW 帯域

## 7. まとめと今後の課題

Rocks を用いて仮想マシン PC クラスタを構築した。そのクラスタ上で基本性能測定としてディスク I/O 性能測定を行なった。次に並列データマイニングのアプリケーション HPA を domain0、domainU において local、iSCSI、iSCSI 通信において片道 10msec〜50msec の遅延をいれて振舞を調べた。その結果、HPA は大量のトランザクションデータを処理するデータマイニングではあるが、CPU の負荷が重くなることから I/O バウンドなアプリケーションではないため遅延が大きくなってもそれほど大きな差が見られなかった。このことより、HPA のように I/O バウンドではない並列アプリケーションの場合、PC クラスタにおいてストレージを遠隔サイトに配置しても、十分実用的な性能を発揮できることが分かった。また domain0 と domainU を比較すると domainU で実行したほうが実行時間が早くなるという結果になった。これは domainU を動作させるとシステムの最適化が行なわれ、そのため全体の性能がよくなったと考えられる。

今後は I/O 処理が重いと考えられる DB 系のベンチマークを動作させ、iSCSI 通信を用いた場合の仮想マシン PC クラスタの動作を解析する。また今回はストレージのみを遠隔に配置したが、計算ノードも遠隔におくようなシステムも構築していき、その際のシステムの最適化なども検討する。

## 謝　　辞

本研究は一部、文部科学省科学研究費特定領域研究課題番号 18049013 によるものである。

## 参 考 文 献

1) Xen : http://www.xen.org/
2) VMware : http://www.vmware.com/jp/
3) Virtual PC : http://www.microsoft.com/japan/windows/products /winfamily/virtualpc/default.mspx
4) Rocks Cluster:http://www.rocksclusters.org/
5) Ganglia Monitoring System:http://www.ganglia.info/
6) iSCSI RFC: http://www.ietf.org/rfc/rfc3722.txt
7) 原明日香、神坂紀久子、山口 実靖、小口正人: ”並列データマイニング実行時の IP-SAN 統合型 PC クラスタのネットワーク特性解析”, DEIM2009,2009 年 3 月
8) 小口正人、喜連川優:”ATM 結合 PC クラスタにおける動的リモートメモリ利用方式を用いた並列データマイニングの実行”, 電子情報通信学会論文誌,Vol.J84-D-I, No.9, pp.1336-1349, 2001 年 9 月
9) Bonnie++ : http://www.coker.com.au/bonnie++/
10) Aravind Menon,Alan L.Cox,Willy Zwaenepoel: ”Optimizing Network Virtualization in Xen”, USENIX Annual Technical Conference,2006 年
11) Jose Renato Santos,Yoshio Turner,G.(John)Janakiraman,Ian Pratt:”Bridging the Gap between Software and Hardware Techniques for I/O Virtualization”, USENIX

Annual Technical Conference,2008 年
12) 谷村勇輔、小川宏高、中田秀基、田中良夫、関口智嗣: ”仮想クラスタに対する IP ストレージの提供方法の比較”, 「ハイパフォーマンスコンピューティングとアーキテクチャの評価」に関する北海道ワークショップ (HOKKE),2007 年