仮想マシンPCクラスタにおける並列データ処理アプリケーション実行時のストレージアクセスに関する一検討

豊島　詩織†　　原　明日香††　　小口　正人†

† お茶の水女子大学　〒 112-8610 東京都文京区大塚 2-1-1
†† 特許庁　〒 100-8915 霞が関三丁目４番３号
E-mail: †{shiori,asuka}@ogl.is.ocha.ac.jp, ††oguchi@computer.org

あらまし　高度 IT 社会の進展により利用可能なデータの量が増大し、それに伴いデータの管理や IT コストの問題が深刻になっている。計算機の効率的な運用として、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 クラスタがある。本研究ではその PC クラスタの個々の計算ノードに仮想マシンを配置した仮想マシン PC クラスタを構築した。ストレージアクセスには接続インターフェイスとして IP ネットワークを利用する 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行なうことが期待される。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を仮想マシン PC クラスタから利用することが期待できる。data-intensive アプリケーションに対する、仮想化した PC クラスタの評価を行なうため、本稿ではデータベースベンチマークである OSDL-DBT3 を動作させ、遠隔アクセスを含む iSCSI ストレージアクセスを行なったときの仮想マシン PC クラスタの振舞を解析する。

A study on virtual PC cluster when parallel date processing application is executed

Shiori TOYOSHIMA†, Asuka HARA††, and Masato OGUCHI†

† Ochanomizu Univ　　2-1-1 Otsuka, Bunkyouku Tokyo 112-8610 JAPAN
†† Japan Patent Office　　4-3 Kasumigaseki, Tiyodaku Tokyo 100-8915 JAPAN
E-mail: †{shiori,asuka}@ogl.is.ocha.ac.jp, ††oguchi@computer.org

Abstract　The amount of the data that can be used by an advanced information technology society increases, and the problem of data management and the IT cost is serious along with it. 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 uses IP network as connection interface for communication. 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 OSDL-DBT3 which is a database benchmark and analyze the behavior of the Virtual machine PC cluster with iscsi remote access for communications.

情報発信の増加やネットワーク上へのデータ蓄積が進み、利用可能な情報が増えており、今後もさらに増え続けると考えられる。その膨大な情報の管理や処理に対して、手元のクラスタを使用しながら、必要な部分だけクラウドから利用するようなアプローチを支援するシステムを考える。

現在話題になっているクラウドコンピューティングはユーザがハードウェア、ソフトウェア、データなどを、インターネットの向こう側からサービスとして使用するという考え方である。現在、手元のクラスタを用いているシステム利用者がクラウドを使うことを考えた場合、当初はクラスタにおいてはクラウド内のストレージ機能のみを使うことが考えられるが、次第にストレージに加え、計算処理を行なうサーバもクラウドで利用する段階へ進むと思われる。

汎用の PC とネットワークを用いて構築した PC クラスタに、PC を一つのコンピュータリソースとみなし、ユーザから要求があった場合や障害時などに必要なときに必要な分だけ動的に別のシステムに割り当てることを可能にする仮想化技術を取り入れ仮想マシン PC クラスタを構築した。仮想マシンを用いる

ことでサービスの要求やシステム負荷に応じてシステムリソースを提供できるようなり、柔軟なインフラ管理が可能となる。

サーバとストレージを結ぶネットワークには IP-SAN を使用した。IP-SAN は、TCP/IP ネットワークで SAN を構築する次世代の SAN で、FiberChannel を用いる従来の FC-SAN に比べ導入コストが安価、管理が容易、長距離通信が可能といったメリットがあるため、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。

構築したクラスタにおいて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の HPA(Hash Partitioned Apriori)[1] を動作させ、実行時間を測定した [2]。この場合はストレージのみを遠隔から使用するようにした。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、ノード間通信やストレージアクセスといった処理でなく、計算処理が主に行なわれているため CPU の負荷が重くなる。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 local で実行する場合と iSCSI を用いた遠隔アクセスには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。

本稿ではデータベースクエリのようなデータ処理が重いアプリケーションを、遠隔アクセス環境下で仮想化した PC クラスタのストレージ環境がどのように扱えるかを評価する。Initiator 対 Target が 1 対 1 の場合と 1 対 4 の場合でデータベースベンチマークである OSDL-DBT3(Open Source Development Labs Database Test3)[3] を動作させ、仮想マシン PC クラスタの動作を解析した。

1. 仮想マシン PC クラスタ

1.1 サーバ仮想化

サーバの仮想化とは CPU やメモリなどのコンピュータリソースを抽象化する技術で、単一の物理サーバのリソースを分割し、あたかも複数のサーバが動作しているように見せたり、複数の物理サーバのリソースを論理的に一つのリソースと見せることができる。これにより処理量を正確に予測することが困難である場合などにシステム負荷やサービスの需要の状況に応じてシステムリソースを即座に調達・融通することが可能となる。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、仮想マシンを用いて仮想的なハードウェアを複数用意することで、同時に複数の OS を実行することが可能となり、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 (図 1)。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。

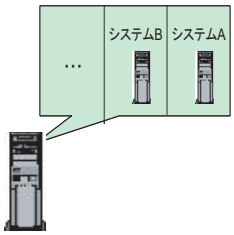

図 1　サーバ集約

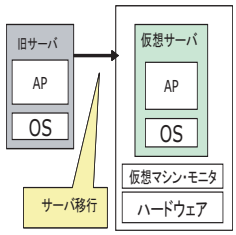

図 2　サーバ移行

1.2 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[4] や Virtual PC(Virtual Server)[5] などがある。本研究ではハイパーバイザ型の仮想化ソフトウェアである Xen を使用した [6]。Xen の特徴はオープンソースであることに加え、Xen の仮想マシンにインストールするゲスト OS をあらかじめ Xen 用に修正する準仮想化と呼ばれる技術を採用しており、仮想マシンのオーバヘッドを少なくし物理マシンと同程度の性能が発揮されるよう工夫されている。

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 として動いているものがドメイン 0，ゲスト OS として動いているものがドメイン 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 (図 3)。

図 3　xen の階層構造

2. ストレージの遠隔利用

2.1 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ユーザーがその所在や内部構造を意識することなく、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検討を行なった。

2.2 IP-SAN

HPC 分野では、PC クラスタの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する。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。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。

本研究では IP-SAN のプロトコルである iSCSI(Internet SCSI)[7] 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。iSCSI の構造を図 4 に示す。iSCSI は 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。Gigabit Ethernet/10Gigabit Ethernet が広く普及していくであろうことを考慮すると、IP-SAN をバックエンドに持つ PC クラスタ多くが使用されるようになると考えられる。

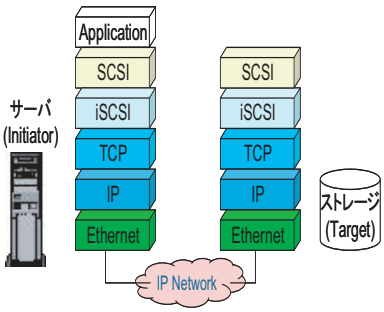

図 4　iSCSI の階層構造

3. 研究内容

3.1 既存研究

文献 8 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 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た。このことより、HPA や PFP、mpiBLAST など大量のデータ処理を行なうが 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ている。

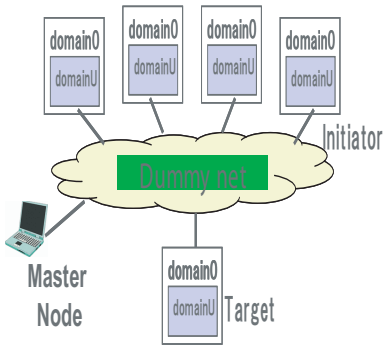

図 5　実験環境

3.2 研究概要

本研究では、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。

各計算ノード内には domainU を一つずつ配置した。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 domain0 と domainU をそれぞれ 2GB とした。

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。

このシステム上でデータベースベンチマークの OSDL-DBT3 を動作させた。

OSDL-DBT3(Open Source Development Labs Database Test3)[8] は、Linux やオープンソースソフトウェアのために、データベーステストキットとして開発された。これは意思決定支援システムの評価を目的したベンチマークテストである TPC-H[9] を単純化したものである。TPC-H はロードテストとパフォーマンステストによって構成される。ロードテストはパフォーマンステストを実行するために、データベースの構築、インデックスの構築などが行なわれる。パフォーマンステストではパワーテストとスループットテストが行なわれ、パワーテストはユーザが一人の場合をシミュレートし、スループットテ

ストはパワーテストの複数のインスタンスで、どれだけ多くのクエリを少ない時間で実行できるか、スケールファクタにより 2 から 10 の同時接続数で行なう。TPC-H に準じ、OSDL-DBT3 も連続的に DB にデータの追加、22 個の意思決定支援を行なうクエリの問い合わせ、DB からのデータの削除を行なうため、頻繁なディスク I/O が行なわれ、システムメモリ内でデータベースバッファキャッシュの競合をもたらすと考えられる。

本稿では 1 台の Target に接続する Initiator の数が 1 台と 4 台でデータベースベンチマークである OSDL-DBT3 を動作させ、ストレージの遠隔アクセスを含む 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し、モニタリングツールである Ganglia[10] により振舞をモニタリングした。

4. データベースベンチマークの実行

OSDL-DBT3 において、同時接続数であるスケールファクタを 1、スループットテストで同時に実行するトランザクション数を 2 とし DBT-3 ワークロードを実行した。このときクラスタの Compute ノードのドメイン 0 だけにジョブを与えた場合とドメイン U にだけジョブを与えた場合の性能を比較し、ストレージは local のストレージを用いる場合、iSCSI 通信によるストレージを用いる場合、さらに 4 対 1 の通信の場合は遠隔アクセス環境を想定し、iSCSI 通信に Dummynet により片道遅延 2msec、4msec、8msec、16msec、32msec の遅延を挿入した環境で実験を行なった。図 6 に Initiator 対 Target が 1 対 1 の場合、図 7 に Initiator 対 Target が 1 対 4 の場合の実行時間を示す。

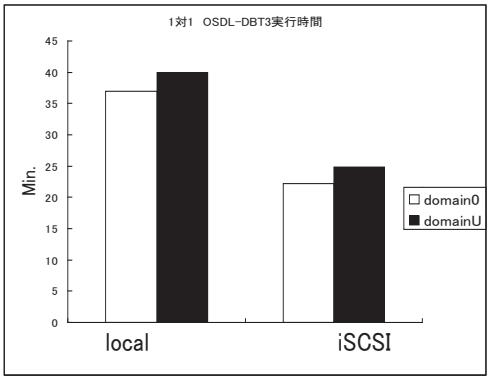

図 6　Initiator 対 Target：1 対 1 の実行時間

1 対 1 の実験ではまずストレージが local のときより、iSCSI 通信を用いるほうが実行時間が早くなっている。これは iSCSI 利用時にキャッシュが利いているためで、自身のディスクキャッシュとともにストレージノードのキャッシュ領域も使用できているためと考えられる。

4 対 1 の実験では local、iSCSI、iSCSI 通信で 2msec〜32msec と遅延を大きくするに伴い実行時間が長くなっている。特に遅延が 32msec となった場合の実行時間の増加が著しい。図 8 はデータマイニングの一種である相関関係抽出の並列データ処理アプリケーション HPA をデータの大きさ 20M で実行した際の実行時間、図 9 はそのときの CPU 使用率である。図 9 より

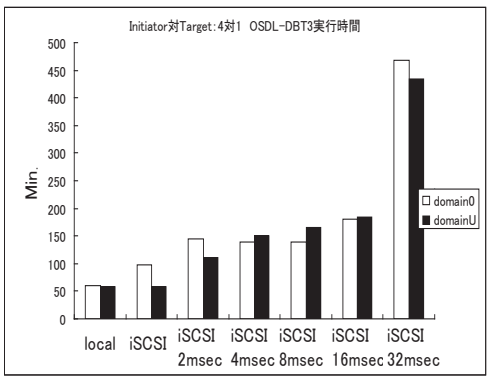

図 7　Initiator 対 Target：1 対 4 の実行時間

CPU をほぼ 100 %使用していることが分かる。

HPA は Apriori をベースにした並列相関関係抽出のアルゴリズムをハッシュ関数を使用して Apriori を並列化したアプリケーションで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。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実行時間にそれほど大きな差が見られなかった。

これに対し OSDL-DBT3 は連続的な DB へのアクセスが発生するため I/O インテンシブとなり、遅延が大きくなるにつれて実行時間に大きな差が見られる。

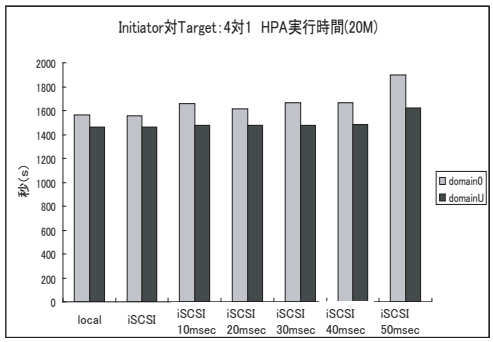

図 8　並列データマイニング (HPA) 実行時間

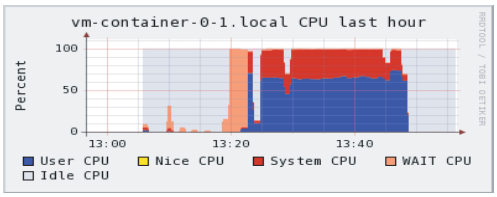

図 9　Initiator の CPU 使用率 (HPA)

図 10〜15 に片道遅延 8msec の iSCSI 通信で、domain0 にジョブを与えた際の、Initiator、Target それぞれの CPU 使用率、メモリ使用率、ネットワーク帯域を示す。

CPU 使用率は余裕があることが分かる。HPA を動作させたときは CPU をほぼ 100 %使用していた。メモリ使用率は Initiator も Target もキャッシュも含めほぼ 100 %使用してる。

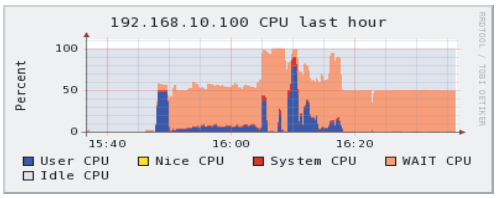

図 10　Initiator の CPU 使用率

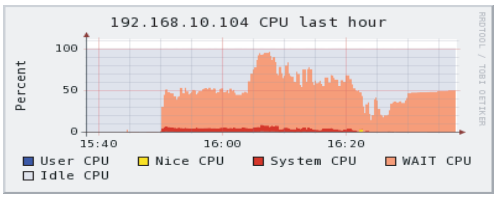

図 11　Target の CPU 使用率

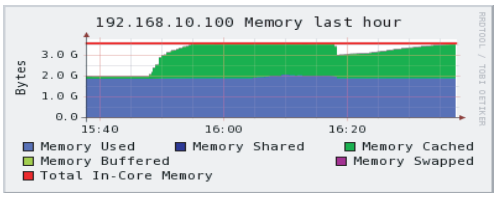

図 12　Initiator のメモリ使用率

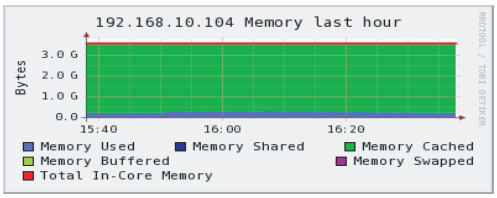

図 13　Target のメモリ使用率

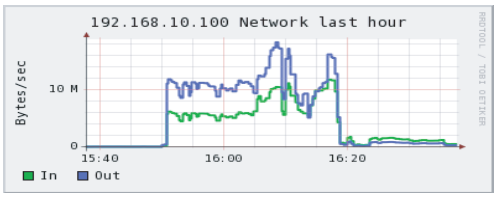

図 14　Initiator のネットワーク帯域

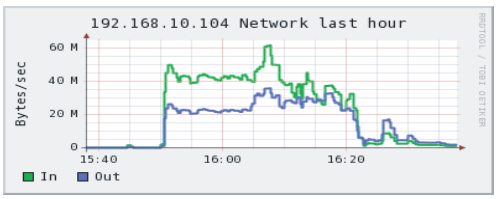

図 15　Target のネットワーク帯域

これにより Initiator が自身のディスクキャッシュとともに、Target のメモリもキャッシュ領域として使用していることが分かる。ネットワークは GigabitEthernet と帯域の広いネットワークを使用しているためまだ余裕がある。以上のモニタリングから OSDL-DBT3 は I/O インテンシブであることが分かる。

5.　まとめと今後の課題

HPA のような CPU の負荷が重いが I/O バウンドではない並列データ処理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ている。そのため data-intensive なアプリケーションに対する仮想化した PC クラスタの評価を行なうため、頻繁な I/O が発生すると考えられるデータベースベンチマークである OSDL-DBT3 を動作させ、遠隔アクセスを含む iSCSI 通信を行い仮想マシン PC クラスタの振舞を解析した。その結果、データベースクエリのようなデータ処理が重いアプリケーションでは、遅延が大きくなるにつれて実行時間に大きな差がでることが確認された。

システム利用形態として今後はクラウドコンピューティング内のストレージのみを使うという段階からストレージに加え計算処理を行なうサーバもクラウドで利用するという段階へ発展していくと考えれるため、ストレージのみを遠隔に配置するのではなく、計算ノードも遠隔におくようなシステムを構築する。そして仮想マシンのマイグレーションなどの機能を使用しストレージにアクセスするサーバを遠隔のクラウドに動的に持っていき、負荷分散を行なうようなシステムを構築し、振舞を解析する。

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。

文　　献

[1] 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
[2] 豊島詩織、原明日香、小口正人: ”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”, DICOMO2009,2009 年 7 月
[3] OSDL-DBT3:http://ldn.linuxfoundation.org/
[4] VMware : http://www.vmware.com/jp/
[5] Virtual PC : http://www.microsoft.com/japan /windows/products /winfamily/virtualpc/default.mspx
[6] Xen : http://www.xen.org/
[7] iSCSI RFC: http://www.ietf.org/rfc/rfc3722.txt
[8] 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
[9] TPC-H:http://www.tpc.org/tpch/
[10] Ganglia Monitoring System:http://www.ganglia.info/
[11] Aravind Menon,Alan L.Cox,Willy Zwaenepoel: ”Optimizing Network Virtualization in Xen”, USENIX Annual Technical Conference,2006 年
[12] 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
[13] 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