課題番号 11-IS01

分散クラウドシステムにおける遠隔連携技術

棟朝 雅晴（北海道大学）

概要　本課題では、北海道大学アカデミッククラウドを中心に全国各地に分散配置されたクラウドシステムを連携させ、一体的なインタークラウドシステムとして相互運用するために必要となる技術的課題について検討、検証する。具体的な目標としては、バーチャルプライベートネットワークにより相互接続されたバーチャルマシン群を一つのクラスタとして運用するための技術、分散バーチャルクラスタ上での並列分散処理に関する研究開発、広域ネットワークを使用することに伴う性能評価、異なるクラウドミドルウェア間での相互運用に関する検討、などがあげられる。本年度においては、技術的な選択肢の検討、いくつかの事例における検証実験の実施に加えて、クラウド管理ミドルウェアの API を通したバーチャルマシンの制御や、それらを相互接続するためのバーチャルプライベートネットワークの制御を行うために必要となるインタークラウドマネージャのプロタイプの開発および検証を行った。

1. 研究の目的と意義

本共同研究課題は、地理的に分散配置されたプライベートクラウドシステムを連携させるために必要となる技術的課題について検討、検証することを目的とするものである。学際大規模情報基盤共同利用・共同研究拠点は、大規模計算機とそれを支える汎用計算機資源が大容量ネットワークで接続されており、さらに近年、いくつかの拠点においてプライベートクラウドシステムの構築を計画している。各拠点におけるプライベートクラウドシステムを超大容量ネットワークで接続することにより、広域分散型のクラウドシステムの構築をめざし、主な技術的課題として以下の項目について研究開発を行う。

（１）バーチャルプライベードネットワークとして SINET4 におけるオンデマンド L2/L3VPN サービス等を用いた、分散プライベートクラウドシステムの相互接続に関する検討、検証、接続実験を行う。

（２）使用するネットワークに依存した拠点間の伝送遅延と帯域による影響に関する検証実験を実施する。

（３）分散配置され、相互接続されたバーチャルマシン群を用いたシステム設計法について検討する。特に、大規模分散クラウドシステム上における MapReduce や MPI 等のバーチャルマシンクラスタの構成に関する検討を行い、実験的にクラスタを構成し、その性能について評価を行う。

（４）大規模分散クラウドストレージの実現に必要な、拠点間でのストレージシステムの連携技術について検討、検証する。

（５）ハイパーバイザソフトウェア、クラウドシステム管理ミドルウェアが互いに異なる環境における管理システム間の連携方法に関する検討を行う。

以上の技術的課題について、北海道大学から九州大学にいたる広域分散クラウドシステムのテストベッドを構築することを通して検討、検証を行うことで、日本全体にわたるアカデミッククラウドの連携に必要となる基盤的技術の開発、および運用モデルの確立を目指すものである。

クラウドコンピューティングについて、海外では大規模なクラウドシステムによるサービスが行われているが、ハイパフォーマンスコンピューティングを実現するものはまだ少なく、さらにその利用条件、利用料金、ネットワーク遅延、セキュリティポリシーなどの問題から、国内の大学や研究機関における要求にすべて合致させることは困難である。

このため、各大学・研究機関においてプライベ

ートクラウドシステムの構築が進められているところであるが、予算上の制約等により、それぞれのシステムの規模は比較的小規模なものに留まることが多く、クラウドコンピューティングによるスケールメリットが生かされないという問題が生じる。

そこで、本研究プロジェクトにおいては、各大学においてそれぞれ導入されたプライベートクラウドシステムを連携させ、大規模なアカデミッククラウドシステムを構築することで、各大学のポリシーを生かしつつ、かつスケールメリットを享受できるような分散環境を実現する。そのために必要な技術的課題について、具体的な広域分散クラウドシステムの連携を通して検討、検証するところに、本共同研究の意義が存在する。

中間報告の時点においては、北海道大学と国立情報学研究所、東京工業大学などとの連携試験の実施、2011 年 11 月にサービスを開始した北海道大学アカデミッククラウドにおける大規模バーチャルマシンクラスタの検証実験の結果について主に報告したが、最終報告においては、それらに加えて、異なるクラウド管理ミドルウェアを相互運用するために必要となるインタークラウドマネージャの構築についても当初計画以上の成果が上がったため、その内容についても報告する。

2. 当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した大学名と研究体制

北海道大学、東京工業大学、国立情報学研究所、九州大学、東京大学、東京藝術大学、広島大学

(2) 共同研究分野

超大規模情報システム関連研究分野

(3) 当公募型共同研究ならではという事項など

全国規模での分散クラウド連携プロジェクトとして、ネットワーク型拠点に分散配置されたクラウド資源を活用した研究開発であることが特徴的である。

3. 研究成果の詳細と当初計画の達成状況

(1) 研究成果の詳細について

本研究課題の本年度における主な研究成果は以下の通りである。

1. 北海道大学⇔国立情報学研究所間でのクラウドシステムのバーチャルプライベートネットワークによる相互接続試験および Hadoop による性能評価試験の実施
2. 北海道大学アカデミッククラウドにおける Hadoop による大規模バーチャルマシンクラスタの性能評価試験の実施
3. 互いに異なるハイパーバイザソフトウェア、クラウドミドルウェアにより管理されるクラウドシステムの連携方式の検討
4. 学術クラウド連携向きのインタークラウドマネージャのプロトタイプ開発

3.1 北海道大学⇔国立情報学研究所間での相互接続および性能評価試験

北海道大学情報基盤センターと国立情報学研究所（一ツ橋）との間をバーチャルプライベートネットワークにより相互接続することで、図 1 に示す分散プラベートクラウド環境を構築した。

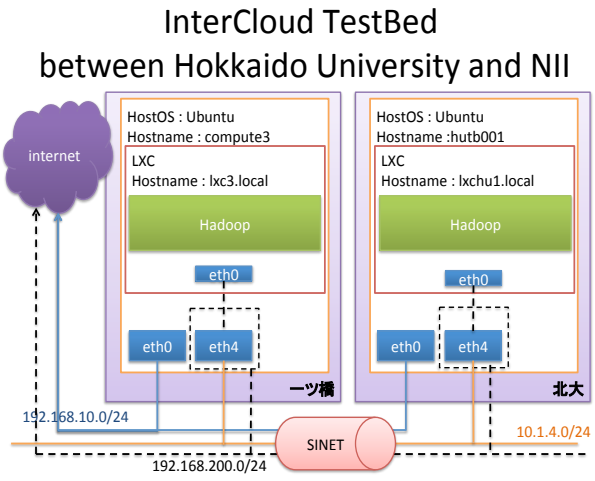

図 1 北大–NII 間での InterCloud 試験環境

本試験システムでは、Linux コンテナをベースとした CaaS (Cluster-as-a-Service)と呼ばれるクラウド環境を構築し、遠隔接続のために、SINET の L2-VPN（帯域幅：1Gbps）機能を用いてネット

ワークを構成した。図の赤線の部分のネットワークがバーチャルプライベートネットワークとしてサイト間で共有されているネットワークである。それぞれのサイトにおける計算資源は以下の通りである。

・国立情報学研究所：DELL PowerEdge R715 x 2 台（AMD Opteron 6128 2GHz x 2, Mem: 32GB, HDD: 500GB）
・北海道大学：DELL PowerEdge R200 x 2 台
（Intel Xeon E3110 3GHz, Mem: 2GB, HDD: 160GB）

・VPN: SINET-4 L2VPLS（1000Base-T 接続）
・ソフトウェア環境：Ubuntu 11.04、LXC で Hadoop をインストールし、クラスタを構成

北海道大学と国立情報学研究所との間で構成されたバーチャルプライベートネットワークを用いて、それぞれのサーバ間での転送実験を行った。測定にあたっては Linux コンテナとして実現されたサーバから、Linux コンテナが動作している物とは別の物理マシンへのファイル転送を行った場合のスループットを測定している。測定を行った結果を図２に示す。縦軸はスループット（Mbps/s）を示している。

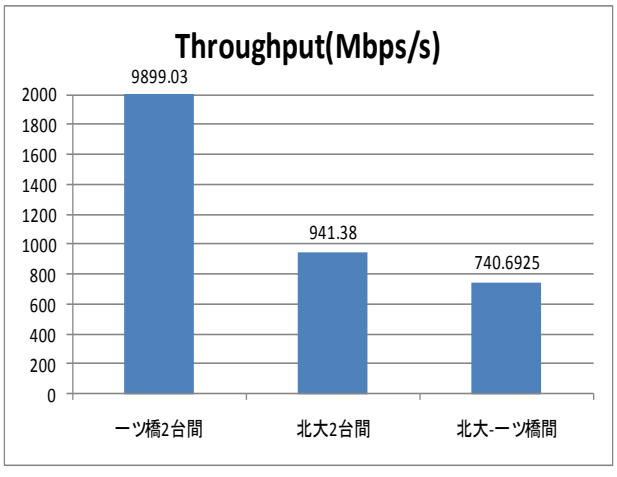

図２ 北大⇔NII 間での帯域幅測定

国立情報学研究所（一ツ橋）内では 10GbE で直接接続されているため、理論値に近いスループットがでており、北大内においても GbE の理論値に近い値が出ている。北大⇔国立情報学研究所間では L2-VPN によるオーバーヘッドが出ているため、直接接続した場合と比べて 21%程度の帯域幅の低下が見られていることが分かった。

また、ネットワーク遅延については、VPN のオーバーヘッドもあり、ping のレスポンスが 20ms 程度となったため、図３に示されるトランザクションの性能（秒あたり処理できるトランザクション数）については、大幅な性能劣化が見られた。

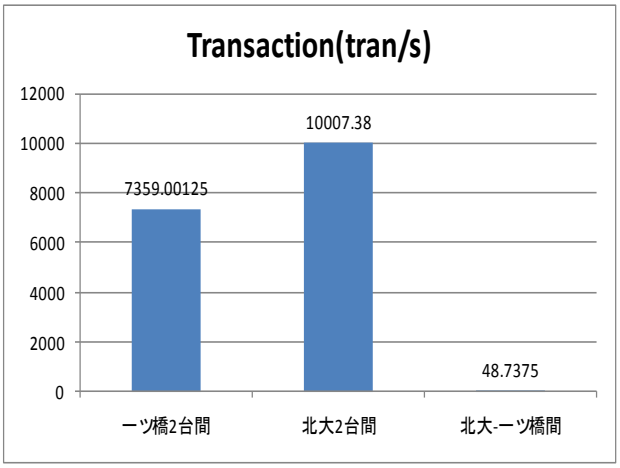

図３ 北大⇔NII 間でのトランザクション性能評価

さらに、北大⇔NII 間で実現された InterCloud 試験環境（図１）で実施した Hadoop の試験結果の一例を図５に示す。

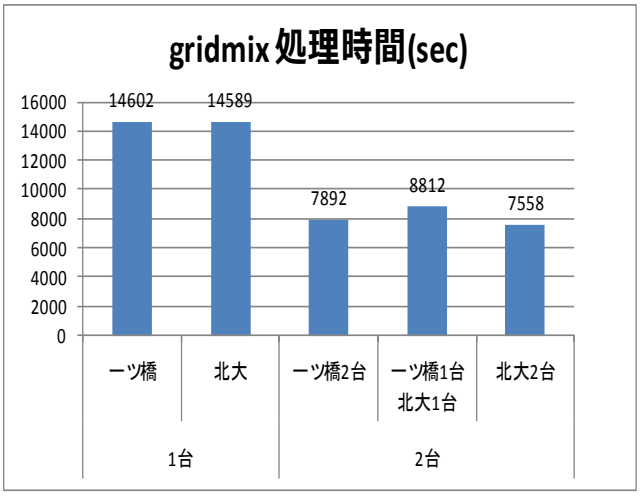

図５ 北大⇔NII 間での Hadoop 試験結果

この結果によると、同一サイト内で行った結果に比べ、北大、NII の両方のサーバをつないで行

った結果が若干劣るものの、その性能劣化が 10%程度に抑えられていることが分った。

3.2 北海道大学アカデミッククラウドに置ける大規模バーチャルマシンクラスタの性能評価

2011 年 11 月にサービスを開始した北海道大学アカデミッククラウドを用いて、Hadoop をインストールしたバーチャルマシンから構成される大規模なクラスタシステムに関する性能評価を行った。本性能評価については、当該クラウドシステムの負荷試験の一部として実施したものであり、システムの仕様で定められた上限である 2,000 のバーチャルマシンを同時にクラスタ化して Hadoop を実行した場合の性能についても試験している。

試験に使用したハードウェア、ハイパーバイザ、ソフトウェア等の環境は以下の通りである。

・ハードウェア：HITACHI BladeSymphony BS2000 A1（Intel Xeon E7-8870 2.4GHz x 4 (10 x 4 = 40 cores), Mem: 128GB, HDD: FC-SAN 20TB, Network: 10GbE x 2 per node)
・ハイパーバイザ：XenServer 5.6
・ソフトウェア：Hadoop 0.20.2
・マスターノード：Xeon E7-8870 4 コア(VM) Mem:8GB, HDD:100GB, CentOS 5.5 (64bit)
・スレーブノード：Xeon E7-8870 1 コア(VM), Mem: 3GB, HDD:100GB, CentOS 5.5 (64bit)

図４に評価試験の結果を示す。ここでは、Hadoop のベンチマーク問題である Teragen と Terasort について試験を行っている。横軸はタスク数（使用したバーチャルマシンの数）であり、縦軸は単位時間あたりに処理できたデータのサイズ（MB/s）を示している。

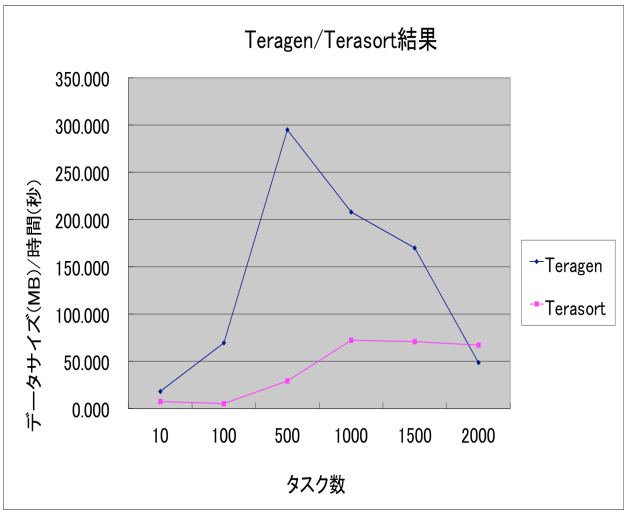

図４ 北大クラウドにおける大規模 Hadoop クラスタの性能評価結果

この結果によると、Teragen の場合で 500 まで、Terasort の場合で 1000 まで性能向上が見られる。それを超えると性能低下が見られるが、この理由としては、物理ノードの数が 100 程度であり、ディスク I/O でのオーバーヘッドが無視できなくなるためであると考えられる。当該クラウドでは、すべての利用者で最大 2,000 のバーチャルマシンを共有して利用することから、実利用において一利用者が構成可能なクラスタとしては高々200 程度のバーチャルマシンであるものと予想されるため、ここで示された結果は、実運用上妥当なものであると考えられる。

3.3 異なるハイパーバイザソフトウェア、クラウドミドルウェア環境における連携方式の検討

異なるクラウド環境の間での相互接続、連携方式の検討のため、CloudStack と OpenNebula の互いに異なるクラウドミドルウェアを用いた小規模なインタークラウド試験システムの構築を行った。構築中の試験システムの構成図を図６に示す。

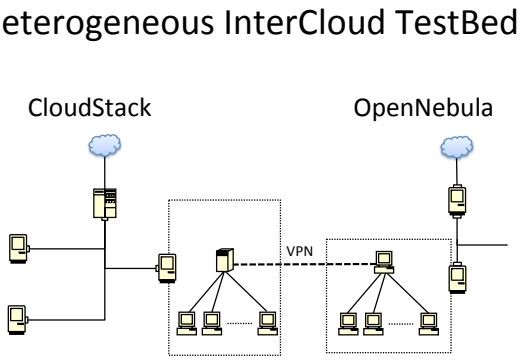

図 6 CloudStack と OpenNebula による InterCloud 試験環境

本試験システムは、CloudStack および OpenNebula の標準インストール環境を実装するとともに、それらの間でバーチャルプライベートネットワークを自動的に設定し、異なるクラウドミドルウェアの管理下にあるバーチャルマシン群を一つのクラスタとして運用するための試験をおこなうものである。

クラウドミドルウェアにおいては、その API を REST 形式により呼び出すことで、バーチャルマシンやストレージ、ネットワーク等の資源を確保することが可能となっている。図 7 に CloudStack におけるバーチャルマシンを確保するための API 呼び出しの例を示す。

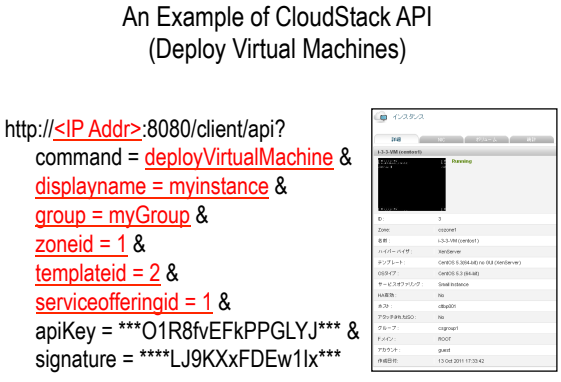

図 7 CloudStack における API 呼び出し例

バーチャルプライベートネットワークについて、CloudStack では専用のソフトウェアスイッチを標準で装備しており、その API を以下のような形式で呼び出し、VPN アクセスのためのユーザ設定を行い、クライアント側から当該ユーザ、パスワードで呼び出すことで、L2TP ベースの VLAN を設定することが可能となっている。

http://<IPaddress>:8080/client/api?
command = addVpnUser &
username = <vpn-user> & password = <pw> &
apiKey= *******O1R8fvEFkPPGLYJ******* &
signature = *******pL81VMNaUbLoV*******

このような API 呼び出しを用い、バーチャルマシンおよびバーチャルプライベートネットワークを設定し、バーチャルマシンのクラスタを自動的に設定するために必要な技術的課題について検討を行っている。図 8 に OpenNebula と CloudStack 間での VPN 接続の試験環境を示す。

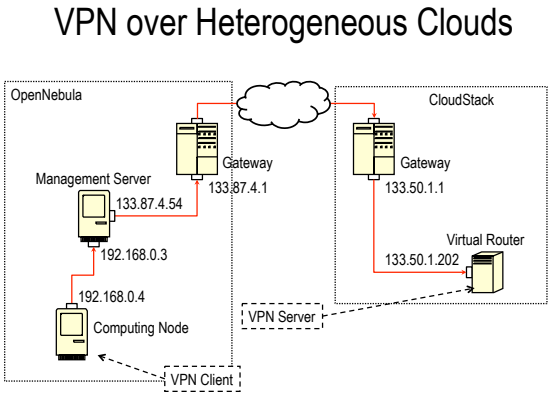

図 8 OpenNebula と CloudStack 間での VPN 接続

OpenNebula 側では CloudStack のようなソフトウェアスイッチが標準で用意されていないため、VPN クライアントソフトウェアを特定のバーチャルマシンにインストールし、そこから CloudStack 側のソフトウェアスイッチを呼び出す方式について検討し、そのために必要となるクライアントソフトウェアを開発した。

クライアントソフトウェアは、Python の Django フレームワークを用いて開発し、xl2tpd と openswan をもとに構成している。図 9 に開発した

VPN クライアントソフトウェアの概要を示す。

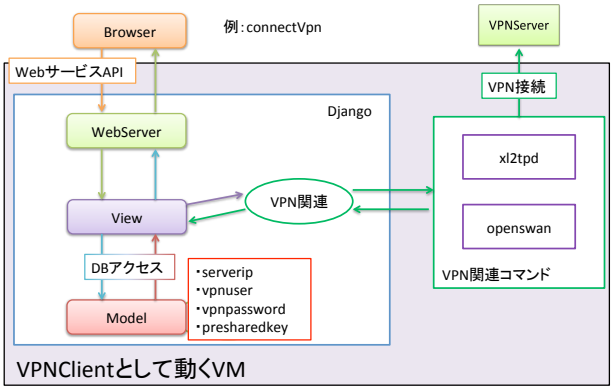

図 9 開発した Web サービス対応 VPN クライアントソフトウェアの概要

これにより、VPN の設定をすべて Web サービスにより一元的に制御することでき、異なるクラウドミドルウェア間での VPN 接続を可能とした。

3.4 学術クラウド連携向けインタークラウドマネージャのプロトタイプ開発

以上の検討をふまえ、異なるクラウドミドルウェアでの管理下にあるクラウドが相互連携したインタークラウド環境において、図 10 に示すようなバーチャルマシンクラスタを自動設定するためのクインターラウドマネージャーのプロトタイプを開発した。

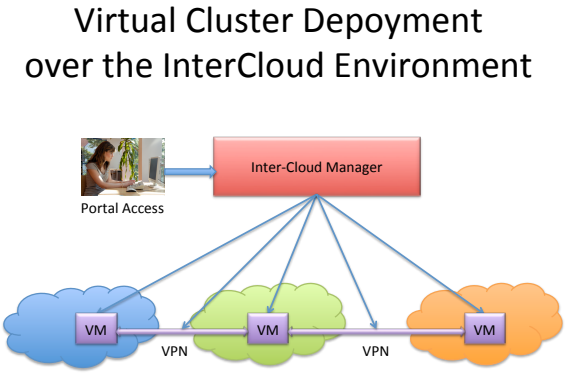

図 10 インタークラウドでの仮想クラスタの構成

本インタークラウドマネージャにおいては、利用者がそれぞれのクラウドミドルウェアのクラウド API を操作するための証明書を事前に登録しておき、それらを Single Sign On ポータルから利用することを可能とする。

ポータルシステムから登録された証明書にしたがって、異なるサイト上に分散配置された資源にたいして、利用者の要求に応じたバーチャルマシンクラスタの資源確保および設定を自動的に行える。自動構成の動作概念図を図 11 に示す。

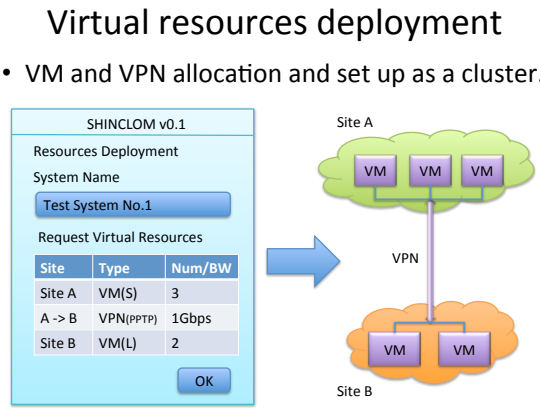

図 11 仮想クラスタの自動構成

ここでは、Site A からバーチャルマシン（サービスレベル S）を 3 台、Site B からバーチャルマシン（サービスレベル L）を 2 台、それらを相互接続する VPN（PPTP）を 1Gbps の帯域幅で確保するという例を示しており、ポータルからの指示にしたがって、クラウドマネージャーが自動的にそれぞれのクラウドミドルウェアの API を事前に登録された証明書によりアクセスして呼び出すことで、分散配置されたバーチャルマシンクラスタを自動的に構成することができる。

開発中のプロタイプシステムの画面例を図 12, 13, 14 に示す。本システムは国立情報学研究所の CSI 経費による援助を受けつつ開発され、JHPCN において実装されたクラウド資源上に実現された物であり、CloudStack, OpenNebula, Eucalyptus, AWS という異なるクラウド管理ミドルウェアに対応して、バーチャルマシンおよびバーチャルプライベートネットワークの設定を一括して行うことが可能となる。

図 12 インタークラウドマネージャプロトタイプシステムのログイン画面

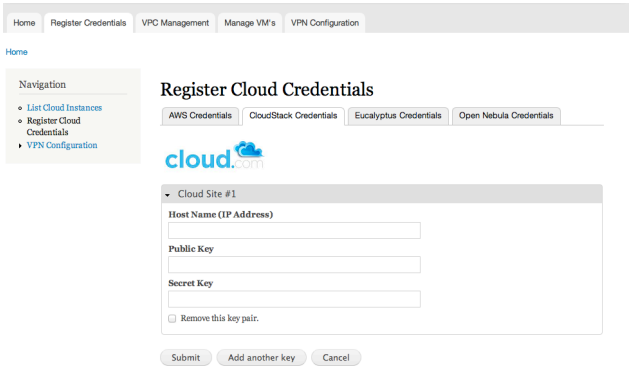

図 13 クラウド管理ミドルウェア制御のための証明書情報の登録画面

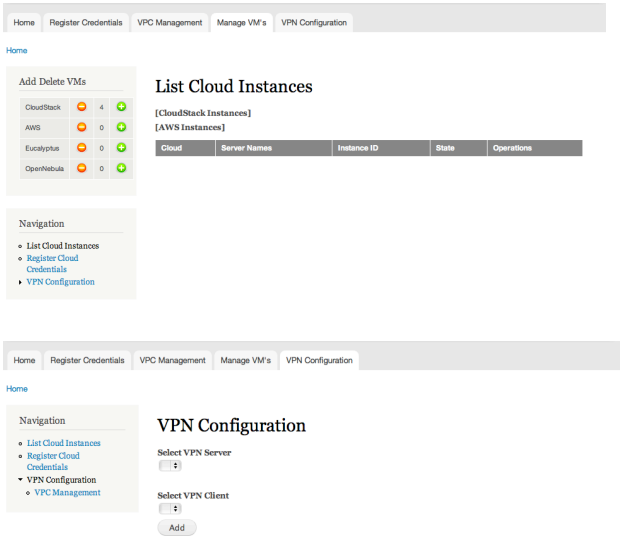

図 14 VM および VPN の設定画面

図 12 は、利用者がログインするシングルサインオンポータルの画面であり、図 13 に示す証明書情報の登録画面から、さまざまなクラウド管理ミドルウェアの制御に必要となる API キーの登録ができる。登録された API キーをもとにして、図 14 に示される画面から、異なる管理ミドルウェアの配下にあるバーチャルマシンおよびバーチャルプライベートネットワークを一括して設定することが可能となった。

(2) 当初計画の達成状況について

当初計画の達成状況について、分散クラウドの相互接続に関する検証実験、ネットワークに関する検証実験、バーチャルマシンを分散配置したクラスタの構成試験については、ほぼ予定通り実施することができた。さらに、異なるクラウド管理ミドルウェア間での連携方式の検証については、当初計画以上に進めることができ、検討にとどまらずインタークラウドマネージャのプロトタイプ開発まで実施することができた。分散ストレージシステムに関する検証については、基盤の構築を優先させたため、現在のところ検討段階にとどまっている。

4. 今後の展望

今後の展望としては、インタークラウドマネージャの開発の継続、機能拡張、および、それを用いた分散クラウド連携試験の実施が重要な課題としてあげられる。インタークラウドマネージャについては、シングルサインオン、証明書の登録、バーチャルマシンおよびバーチャルプライベートネットワーク設定など、最小限の機能は実装できているが、バーチャルマシンクラスタの自動設定等、今後必要とされる機能について実装を進めて行く予定である。

大規模分散クラウドストレージについては、現在検討をすすめている段階であり、OpenStack swift などオブジェクトストレージサーバの構築および連携について検討をさらに進める。

ディザスタリカバリに関する検討については、今後の重要な課題となる。本研究課題により本年度実現された、分散配置されたバーチャルマシン

クラスタを構成する技術を元に、それらを構成する一部のバーチャルマシン、バーチャルプライベートネットワークが使用できなくなった場合を想定することで、クラスタシステム全体としての可用性を最大化するために必要となる技術開発を行う予定である。

5. 研究成果リスト

(1) 学術論文（投稿中のものは「投稿中」と明記）

[1] 棟朝雅晴，高井昌彰：北海道大学アカデミッククラウドにおけるコンテンツマネジメントシステムの展開，情報処理学会第 10 回情報科学技術フォーラム論文集（査読付）（2011）

(2) 国際会議プロシーディングス

[2] Omar Abdul-Rahman, Masaharu Munetomo and Kiyoshi Akama: Multi-Level Autonomic Architecture for the Management of Virtualized Application Environments in Cloud Platforms, Proceedings of the IEEE 4th International Conference on Cloud Computing (IEEE CLOUD 2011), pp.754-755 (2011)
[3] Mohamed Wahib, Asim Munawar, Masaharu Munetomo and Akama Kiyoshi: A Framework for Cloud Embedded Web Services Utilized by Cloud Applications, Proceedings of the IEEE 2011 World Congress on Services Computing (IEEE SERVICES 2011), pp.265-271 (2011)

(3) 国際会議発表

(4) 国内会議発表

[4] 棟朝雅晴：北海道大学アカデミッククラウドの構築，アカデミッククラウドシンポジウム 2011＠北海道大学（2011）
[5] 日下部茂：九州大学システム情報科学府・研究院でのキャンパスクラウドの活用，アカデミッククラウドシンポジウム 2011＠北海道大学（2011）
[6] 横山重俊：教育クラウド edubase Cloud の利用事例と運用，アカデミッククラウドシンポジウム 2011＠北海道大学（2011）
[7] 西村浩二：組織間連携による分散ファイル管理システムの開発，アカデミッククラウドシンポジウム 2011＠北海道大学（2011）
[8] 滝澤真一朗：RENKEI-PoP による広域分散 VM ホスティングの構築，アカデミッククラウドシンポジウム 2011＠北海道大学（2011）
[9] 實本英之：VM を考慮したジョブスケジューリングシステムの開発，アカデミッククラウドシンポジウム 2011＠北海道大学（2011）
[10] 小林泰三：大規模広域分散システム:管理者と利用者の視点から，アカデミッククラウドシンポジウム 2011＠北海道大学（2011）
[11] 横山重俊，長久勝，吉岡信和：クラウド基盤構築フレームワーク Dodai について，第 30 回インターネット技術第 163 委員会研究会（2011）
[12] 棟朝雅晴：北海道大学アカデミッククラウドの構築と運用について，グリッド協議会第 33 回ワークショップ，秋葉原コンベンションホール（2011）
[13] 棟朝雅晴：「分散クラウドシステムにおける遠隔連携技術」の取り組みについて，アカデミッククラウドワークショップ 2012＠広島（2012）

(5) その他（特許，プレス発表，著書等）

[14] “Hokkaido University Builds Japan’s Largest Academic Cloud Using Cloud.com” : Business Wire (http://www.businesswire.com/) 他多数（2011）

謝辞

本研究の実施にあたっては、学際大規模共同利用・共同研究拠点共同研究に加えて、北海道大学情報基盤センター、同共同研究経費、国立情報学研究所の援助を受けている。