科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１２　日現在

機関番号：　１４６０３
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７０００５８

研究課題名（和文）広域環境上における分散制御型マルチサイト仮想クラスタに関する研究

研究課題名（英文）A study on decentralized control multi-site virtual cluster systems in wide area environments

研究代表者
市川　昊平（Ichikawa, Kohei）

奈良先端科学技術大学院大学・情報科学研究科・准教授

研究者番号：９０５１１６７６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円、（間接経費）　　990,000円

研究成果の概要（和文）：本研究は、複数サイトの計算資源からなる仮想的なクラスタ環境（マルチサイト仮想クラスタ）の柔軟かつスケーラブルな構築手法の確立を目的とし、従来の中央集中管理型のクラスタ構成をとらない分散制御型の仮想クラスタ構築・管理機構を実現した。従来型のクラスタ構成をそのままマルチサイトクラスタに適用すると、マスタにクラスタ構成情報、ユーザアカウント情報、ジョブ等の管理が集中し、計算資源の管理・運用面で問題がある。本研究では、各クラスタの構成ノードそれぞれがこれらの情報を分散して管理するアーキテクチャを提案し、スケーラブルで柔軟な分散制御型仮想クラスタを実現した。

研究成果の概要（英文）：In this study, a decentralized control virtual cluster system is designed and implemented. In traditional cluster systems, a centralized control structure has been accepted. In such a centralized control cluster system, all information on the structure of the cluster system, user accounts, and job queues is stored only on a master node of the cluster. This kind of structure is not scalable and not appropriate for widely distributed computing systems. In the proposed architecture, all data related to a cluster system are stored on each compute node in a distributed manner.

研究分野：　総合領域

科研費の分科・細目：　情報学・計算機システム・ネットワーク

キーワード：　仮想クラスタシステム　ハイパフォーマンスコンピューティング　分散システム

１．研究開始当初の背景

地理的に分散する複数の研究機関や大学、データセンタの保有する計算資源上に、動的に仮想計算機を配置し、仮想プライベートネットワーク(VPN)技術等により集約する事によって、複数サイト間にまたがるマルチサイト仮想クラスタを構築する技術に関する研究開発が推進されつつある（図１）。マルチサイト仮想クラスタでは、既存の計算科学アプリケーションを修正することなく、複数のサイトにまたがる計算資源を有効に活用可能であるため、大きく期待が高まっている。

しかし、これまでの仮想クラスタ技術に関する研究は、従来の物理クラスタ構成をそのまま仮想環境上で再現することに焦点を当てて進められてきたため、クラスタの中央集中管理を行うマスタとプログラムの実行を行うその他多数のワーカからなる従来のシステム構成を再現する傾向にあり、スケーラビリティや柔軟性に問題がある。クラスタでは主として、1)クラスタ構成の管理：クラスタを構成する各ノードへのホスト名の割り当てや IP アドレスの割り当て等、クラスタ内の相互接続ネットワークの管理、2)ユーザアカウント情報の管理：クラスタ内で共有すべきユーザのアカウント情報の管理、3)ジョブ管理：ユーザから投入されたジョブへの計算資源の割り当ての管理を行う必要がある。従来型のクラスタ構成をとるマルチサイト仮想クラスタでは、このような管理を集中的に行うマスタを、ある特定のサイトに配備する必要がある。そのため、管理・運用が煩雑で柔軟性が欠如していたり、多数のワーカに関する膨大な情報が１つのマスタに集中するなどスケーラビリティ上の問題がある。仮想クラスタに参加するサイトは本来相互に対等な関係であり、クラスタへの計算資源の参加・脱退は動的に起こり得る。したがって、ある一拠点にマスタを設置する従来型の手法はシステム構成を煩雑にし、耐障害性の面でも有用でない。

本研究実施者は現在までに、P2P 技術により構築したオーバレイネットワーク上に、仮想クラスタを構築可能とする技術の研究開発を行ってきたが、情報を分散して管理する典型的な仕組みである P2P 上に、従来の中央集中管理型のクラスタ構成を再現することに矛盾を感じたことも分散制御型の仮想クラスタの着想に至った理由の１つである。

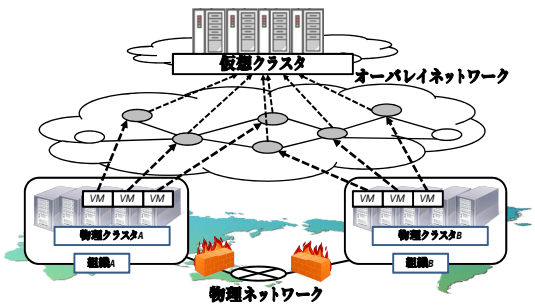

図 1 マルチサイト仮想クラスタの概念図

２．研究の目的

本研究では、中央集中管理を担うマスタを必要とせず、クラスタの構成情報、ユーザアカウント情報、ジョブの管理等を分散して行う分散制御型の仮想クラスタ構築・管理機構の開発を目的とし、それを実現するための要素技術の開発を実施した。

３．研究の方法

本研究は、本研究実施者がこれまでに構築してきた P2P によるオーバレイネットワーク上の仮想クラスタを拡張する形で実施する。具体的には、複数サイトに分散する仮想計算機間の仮想的な通信路をオーバレイネットワーク上に確立し、かつ 1)クラスタの構成情報、2)ユーザアカウント情報、3)ジョブの管理を、構成ノードそれぞれが分散管理する分散制御型のマルチサイト仮想クラスタ構築・管理機構を実現する。これにより、全てのノードが対等に動作し、複数サイトの計算資源上に配備するのに適した、新しいマルチサイト仮想クラスタの構築技術を確立する。

本研究では、本目的を達成するため、マイルストーンとなる以下の 4 つの課題を設定して、実施してきた。

[課題 1] マルチサイト仮想クラスタに特化したオーバレイネットワークの設計と実装

仮想クラスタを実現する上でネットワークを仮想化するオーバレイネットワークは重要な役割を持つ。本研究では、本目的のために当初は P2P 技術を応用したオーバレイネットワークの構築を検討していたが、ネットワーク技術の進展に伴い、近年提唱され始めた Software-Defined Network 技術を基にしたオーバレイネットワークの構築も実施した。

[課題 2] クラスタ構成、ユーザアカウント、ジョブに関する情報の分散管理機能の設計と実装

クラスタが扱う情報はノードのホスト名や IP アドレス、ユーザアカウント等の比較的更新頻度の少ない静的な情報と、ジョブの割り当て状況や負荷等の動的な情報に分けられる。P2P では、ノードの動的な加入や脱退を考慮した情報の冗長化を行うと、情報の一貫性の維持が難しくなる。そのため、本研究では、クラスタで管理する情報の性質に応じた適切な可用性・一貫性のレベルを考慮した分散管理機能を設計する必要があった。

[課題 3] 分散制御型のジョブ割り当て・管理機能の設計と実装

分散制御型クラスタでは、ユーザから投入されたジョブのキューイングや実行、終了状態の管理を各ノード上で分散して実施する。

本研究では、既存のアプリケーションの再利用性を高めるため、今日のクラスタシステムにおいて広く利用されている SGE(Sun Grid Engine)や PBS(Portable Batch System)等のジョブ管理システムを考慮し、既存のジョブ管理システムと共通のインタフェースを有する分散制御型のジョブ管理機能を設計した。

[課題 4] マルチサイト仮想クラスタ構築・管理機構の実証実験による評価

本研究で構築した分散制御型のマルチサイト仮想クラスタ構築・管理機構は実際のアプリケーションに適用し、広域環境上にて実証実験を行い、その実用性・有用性について評価を行った。

４．研究成果

本研究では、前節で挙げた研究目的を達成するために解決が必要な各課題に応じて、要素技術の開発や研究を実施し、それぞれ成果を達成してきた。以下、各項目にしたがって、説明する。

(1) マルチサイト仮想クラスタに特化したオーバレイネットワークの設計と実装

研究当初は P2P 技術を応用したオーバレイネットワークの構築を検討していたが、本研究では近年提唱され始めつつある Software-Defined Network（SDN）技術を基盤にマルチサイト間を接続するオーバレイネットワークを構築する技術を開発した。

P2P 技術をベースにしたオーバレイネットワークではその通信のルーティング手法は P2P のアルゴリズムに依拠することとなり、マルチサイト間の通信路を最適化することは困難であった。一方で、SDN 技術では各ルーター間の通信路の選択を完全にソフトウェアから動的に制御できるため、マルチサイト間の通信の最適化が容易である。本研究では、SDN 技術を実装する上で標準的に用いられている OpenFlow 技術により、オーバレイネットワークを構築した。OpenFlow ネットワークは、OpenFlow プロトコル対応のスイッチ（OpenFlow スイッチ）とそれらを制御するプログラマブルコントローラからなり、プログラマブルコントローラ上で計算したルーティング方針に従い、OpenFlow スイッチの動作全てを制御することで SDN を実現する。

本研究では、OpenFlow スイッチとして、OepnFlow の機能をソフトウェア的に実装した、Open vSwitch を用いることにした。Open vSwitch は Linux 標準の Bridge デバイスと互換性のある実装を有し、パフォーマンスも優れている。また、異なるサイト間において既存の L3 ネットワーク上に仮想的な L2 のトンネリングリンクを作成する GRE プロトコルに対応していることも特徴の一つである。

本研究では、この Open vSwitch を各拠点に配備し、拠点間を GRE で相互に結合することで、オーバレイネットワークを構築した（図 2）。そして、拠点間の通信は OpenFlow のプログラマブルコントローラが動的に制御した。具体的には、プログラマブルコントローラが拠点間を結合するネットワークトポロジを常に把握し、動的に変化するトポロジに応じて、各拠点間を結ぶ最短経路を常に算出し、ある拠点から拠点への通信を必要とするプログラムに対し、その最短経路を割り当てるように制御する技術を開発した。これにより、従来のマスタノードを中心とするネットワーク構成ではなく、個々のサイトが対等に最適路で通信を確立できるようになった。

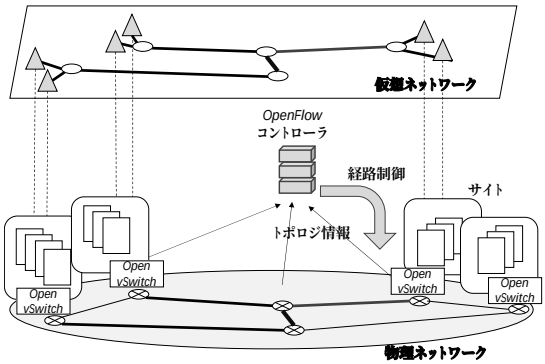

図 2 OpenFlow に基づくオーバレイネットワークの概要

(2) クラスタ情報の分散管理

本研究では、クラスタ内の情報を従来のクライアントのマスタノードのみで集中的に管理する手法を改め、各サイトの計算機において分散管理手法を提案した。具体的には Apache ZooKeeper と呼ばれる情報管理フレームワークを基盤に、このフレームワーク上にマルチサイトクラスタを構築する上で必要な情報の管理と、それらを操作するための独自のインタフェースの実装を行った。本研究では、この ZooKeeper 上の情報管理システムにおいて、比較的静的な管理情報と動的なジョブに関する情報を共に管理し、前節の課題 2、3 の両方の解決を実施した。

具体的には、ZooKeeper のサーバプロセスを各拠点にそれぞれ配備し、クラスタを構成する各計算ノードは各拠点の ZooKeeper とやり取りすることで情報の取得・更新を実施する仕組みを実装した。このようにすることで、中央集中的なサーバ管理を排除することができ、各拠点が対等に分散して情報を管理する仕組みが構築された。

各計算ノードのアドレスや構成情報、ユーザアカウントなどの比較的静的な情報の管理のためには、分散ロックなど排他的な情報更新・閲覧を実装する必要がなく Zookeeper を通して、各拠点が任意のタイミングで情報を更新できるようにしておき、他の拠点は情報の更新があったことが ZooKeeper より通知されると、自身の各計算機ノードの内の

/etc/hosts情報などをZooKeeper上の更新情報に応じて更新するというモデルをとった。

一方で、ジョブなどの動的な情報管理に関してはZookeeperがサービスとして有しているシーケンスノードの生成機能や、その生成されたシーケンスノードの削除は一つのクライアントからしか出来ないという機能を利用し、排他的にジョブキューを管理できる仕組みを分散環境上に構築した（図3）。

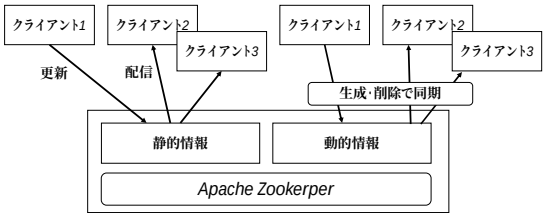

図 3 Zookeeperを基盤とした情報管理

(3) 提案システムの実証実験

実証実験として、提案システムを用いて、グリッド及びクラウドコンピューティングの国際的な連携コミュニティである PRAGMA の参画機関の提供する計算資源上にマルチサイト仮想クラスタシステムを構築した。

具体的には、創薬の初期段階で実施する候補薬物のスクリーニングに用いられる分散計算型の科学アプリケーションである DOCK を実行する仮想クラスタを実行するための仮想クラスタを構築した。仮想クラスタを配備するため、DOCK をインストールした仮想ディスクイメージを各拠点に配備し、それらから起動した仮想マシンをオーバレイネットワークで結合し、正常に DOCK が実行できることを示した。この結果は PRAGMA が実施する国際ワークショップにおいて、ライブデモンストレーションにより実証した。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① Daiki Morita, Kohei Ichikawa, Hirotake Abe, Susumu Date and Shinji Shimojo, “Implementation and Evaluation of Multiple Deduplication Methods for VM Disk Images Composing a Virtual Cluster,” INFORMATION - An International Interdisciplinary Journal, Vol.16, No.8(B), pp.6055-6068, August 2013, 査読有.
② 多田大輝, 市川昊平, 伊達進, 阿部洋丈. 下條真司. オーバレイネットワークを用いたマルチサイト仮想クラスタ構築システム. 情報処理学会論文誌：コンピューティングシステム 第40号, vo.5, no. 5, pp. 76-89, 2012年10月, 査読有.

〔学会発表〕（計 14 件）

① Daniel Li, Brian Tsui, Charles Xue, Jason Haga, Kohei Ichikawa and Susumu Date, “Protein Structure Modeling in a Grid Computing Environment,” 9th IEEE International Conference on eScience, October 23, 2013, Beijing, China.
② Pongsakorn U-chupala, Kohei Ichikawa, Luca Clementi, Nadya Williams, Philip Papadopoulos, Yoshio Tanaka, Akihiko Ota and Weicheng Huang, “PRAGMA VC Sharing Automation Phase 4,” Pragma25 Workshop, October 18, 2013, Beijing, China.
③ Kohei Ichikawa, Kevin Lam, Karen Rodriguez, Wen-Wai Yim, Jason Haga, “Deployment of Virtual Clusters for Molecular Docking Experiments on the PRAGMA Cloud,” Pragma25 Workshop, October 18, 2013, Beijing, China.
④ Pongsakorn U-chupala, Putchong Uthayopas, Kohei Ichikawa, Susumu Date and Hirotake Abe, “An implementation of a multi-site virtual cluster cloud,” The 10th International Joint Conference on Computer Science and Software Engineering (JCSSE'13), pp. 155-159, May 29, 2013, Khon Kaen, Thailand.
⑤ Kohei Ichikawa, “International Clouds using OpenFlow,” PRAGMA Cloud Computing and Software-Defined Networking (SDN) Technology Workshop, March 20, 2013, Bangkok, Thailand.
⑥ Kohei Ichikawa, Taiki Tada, Susumu Date, Shinji Shimojo, Hirotake Abe, Nawawit Kes, Putchong Uthayopas, Bong Zoebir, Lim Teck Leong Derrick, Francis Lee Bu Sung, Cindy Zheng and Philip Papadopoulos, “Network throughput-aware routing for Pragma Cloud,” PRAGMA24 Workshop, March 22, 2013, Bangkok, Thailand.
⑦ Daiki Morita, Kohei Ichikawa, Hirotake Abe, Susumu Date and Shinji Shimojo, “Implementation and Evaluation of Multiple Deduplication Methods of VM Disk Images Composing a Virtual Cluster,” The 3rd International Workshop on Ubiquitous Computing & Applications (IWUCA 2012), December 22, 2012, Hong Kong.
⑧ 多田大輝, 市川昊平, 伊達 進, 阿部洋丈, 下條真司, "オーバレイネットワークを用いたマルチサイト仮想クラスタ構築システム", 先進的計算基盤システムシンポジウム(sacsis2012), pp. 219-226, 2012年5月18日, 兵庫県神戸市.

⑨ 森田大希，市川昊平，阿部洋丈，伊達進，下條真司,“仮想クラスタを構成する複数ディスクイメージの効率的移送手法”,第121回システムソフトウェアとオペレーティング・システム研究会，pp. 1-9，2012年5月8日，沖縄県恩納村.
⑩ Taiki Tada, KoheiIchikawa, Susumu Date, Shinji Shimojo, Yoshio Tanaka, Akihiko Ota, Tomohiro Kudoh, Cindy Zheng and Philip Papadopoulos, “An implementation of OpenFlow based virtual network for virtual clusters on the PRAGMA testbed,” PRAGMA22 Workshop, April 19, 2012, Melbourne, Australia.
⑪ Kohei Ichikawa, Taiki Tada, Susumu Date, Shinji Shimojo, Yoshio Tanaka, Akihiko Ota, Tomohiro Kudoh, Cindy Zheng and Philip Papadopoulos, “An Openflow based virtual network environment for Pragma Cloud virtual clusters,” The 22th PRAGMA Workshop, April 18, 2012, Melbourne, Australia.
⑫ Pongsakorn U-chupala, Kohei Ichikawa, Hirotake Abe, Susumu Date, and Shinji Shimojo, “A Virtual Cluster Manager using a Hierarchical Management Model for Cloud Infrastructure,” The 6th International Conference on Ubiquitous Information Technologies and Applications, pp.223-228, December 16, 2011, Seoul, Korea .
⑬ Susumu Date, Taiki Tada, Kohei Ichikawa, and Shinji Shimojo , “Towards multi-site virtual cluster deployment,” The 6th International Conference on Ubiquitous Information Technologies & Applications, pp.29-32, December 15, 2011, Seoul, Korea.
⑭ Kohei Ichikawa, Susumu Date, Yasuyuki Kusumoto, and Shinji Shimojo, “A P2P-based Virtual Cluster Computing using PIAX”, The 2011 IEEE Conference on Granular Computing (GrC 2011), pp.294-299, November 10, 2011, Kaohsiung, Taiwan.

６．研究組織

(1)研究代表者

市川 昊平（ICHIKAWA, Kohei）
奈良先端科学技術大学院大学・情報科学研究科・准教授
研究者番号：90511676

(2)研究分担者

なし

(3)連携研究者

なし