様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　１１　日現在

機関番号：　８２６２６
研究種目：　若手研究(B)
研究期間：　2012～2014
課題番号：　２４７０００４０

研究課題名（和文）性能可搬性を提供する仮想計算機マイグレーション技術の研究

研究課題名（英文）Study of a VM migration mechanism with performance portability

研究代表者

高野　了成（Takano, Ryousei）

独立行政法人産業技術総合研究所・情報技術研究部門・研究グループ長

研究者番号：１０５０９５１６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：クラウドの利用が拡大しているが、利用者が性能や価格をもとにクラウドを使い分けることは容易ではない。本研究では、アプリケーション実行環境を一度作れば、どんなクラウド上でも実行できる（「Build Once, Run Everywhere」）という設計思想による仮想クラスタ型計算機の実現に向けて、EthernetやInfiniBandといった異なる通信インタフェースをもつクラスタ計算機間での仮想計算機ライブマイグレーション機構を実現した。さらにその成果の一部をプライベートクラウドAIST Super Green Cloudに適用し、その実用性を確認した。

研究成果の概要（英文）：Cloud users cannot easily use clouds based on the price, the performance, and other factors. We have proposed a multi-cloud deployment method in which a customized development and execution environment for applications can be deployed across the HPC platforms, in a ``build once, run everywhere'' manner. To achieve the goal, This study proposes an interconnect-transparent migration mechanism to simultaneously migrate multiple co-located virtual machines between Clouds equipped with different interconnect devices such as Ethernet and InfiniBnad. The part of the proposed mechanism has been deployed on a private cloud platform on our recently introduced supercomputer system, the AIST Super Green Cloud (ASGC).

研究分野：　オペレーティングシステム

キーワード：　ソフトウェア　計算機システム　ネットワーク　仮想計算機　高性能計算

１．研究開始当初の背景

クラウドコンピューティングは、計算資源を抽象化して運用する手段として近年その利用が拡大している。さらに高性能計算（HPC）用途での利用にも注目も高まっており、物理的にはネットワークで接続された複数のスーパーコンピュータにより構成されるインフラを、クラウドコンピューティングのように容易に利用できる仕組みの実現が期待されている。

HPC 向け計算資源の仮想化はユーザ側における実行環境整備の省力化の点でメリットが大きい一方で、仮想化のオーバヘッドなどによる性能の低下が問題となる。さらに、あらかじめ実行環境のハードウェア詳細を仮定できないので、サービスが提供されるサイトで利用可能な最善の通信インタフェースを自動的に選択し実行できる、性能可搬性を提供することが課題となる。

以上の背景を踏まえ、我々はユーザの利便性と HPC 用途に耐えうる性能の追求を目指した、HPC クラウドの「Build once, Run Everywhere」化を構想している。まず、ユーザが手元の計算機でアプリケーションを含む仮想計算機イメージを作成、テストする。これを任意のサイトに配備し、ユーザの要求に応じた規模の仮想化 HPC クラスタをオンデマンドに構築する。そして計算機センタ内のジョブスケジューリングや故障の検出等の状況変化に応じて、仮想計算機はマイグレーションを続け、ジョブ実行を継続する。さらに、マイグレーション前後で計算機の通信インタフェースが異なる場合でも、アプリケーションプログラムから透過に性能可搬性を提供することを目指す。

２．研究の目的

本研究は、図１に示すように、Ethernet と InfiniBand などそれぞれ異なる通信インタフェースによって接続されたクラスタ計算機を跨ぐ仮想計算機マイグレーションを実現し、アプリケーションプログラムから透過に、かつ計算機環境の変化に適応して最大の通信性能を達成する性能可搬性を実現することを目的とする。具体的には、基本技術から応用にかけて、次の３項目の研究開発を行う。

（１）InfiniBand を利用する仮想計算機のライブマイグレーション

（２）InfiniBand から Ethernet のように異種のネットワークを跨ぐ仮想計算機ライブマイグレーション

（３）資源管理システムと連携したライブマイグレーションの制御

３．研究の方法

（１）に関して、PCI パススルーなど OS バイパス型デバイスを用いた際のライブマイグレーションの研究は存在するものの[1]、一般に入手できる実装はなく、その実現方式

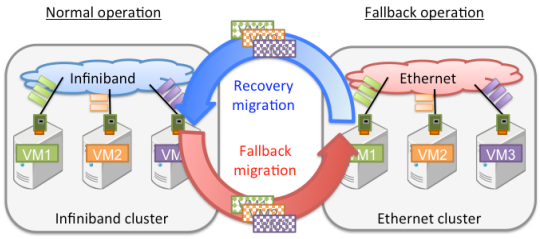

図１　ネットワークインタフェースに透過な仮想クラスタマイグレーション

も自明ではない。本研究では、ゲスト OS のユーザレベルに実装するネットワーク抽象化ライブラリと PCI ホットプラグ（活線挿抜）機能を組み合わせることで実現する。本ライブラリはパススルーで使用する物理デバイスと仮想化 IO デバイスをアクティブ・スタンバイ構成で動的に切り替えることが可能であり、仮想計算機上でも物理性能に匹敵する性能とマイグレーション中の接続性を実現する。通常時は物理デバイスを使用し、マイグレーション中はそのデバイスを一時的に切り離して、仮想化 IO デバイスを使用する。マイグレーション完了後、再度デバイスを認識、初期化する。本研究期間にて、提案システムを実装、評価し、その設計の妥当性および有効性を明らかにする。

（２）に関して、既存の仮想計算機マイグレーションは、仮想計算機モニタの違いに対応する研究[2]はあるものの、移送元と移送先のハードウェア構成（PCI パススルーデバイス）が異なる場合に対応していない。この問題に対しても、（１）と同様に、マイグレーション中は一時的に対象デバイスを切り離し、再度接続することで解決できると考える。通信インタフェースごとの API の違いは、MPI ライブラリなどで吸収することで、アプリケーションからの透過性を維持する。

（３）に関して、その制御方式は運用ポリシに依存するため、本研究では必要最低限の API の抽出とその実装に注力する。

４．研究成果

（１）InfiniBand を利用する仮想計算機のライブマイグレーション

高性能計算分野でデファクト標準となっている MPI (Message Passing Interface)プログラムを対象にシステムの設計を行い、InfiniBand を PCI パススルーで用いた場合のライブマイグレーション機能を仮想計算機モニタ KVM 上に実装した[雑誌論文⑤、⑧]。PCI パススルーを利用することで仮想計算機でも物理計算機に匹敵する通信性能を得られていたが、従来技術では仮想計算機をマイグレーションすることができなかった。本技術により、PCI パススルーの性能上の利点を損なうことなく、仮想計算機のマイグレーションおよびチェックポイント・リスタートを実現できた。これには、高性能計算分野など、高性能と耐故障性の両立が必要となる分野

においても仮想化技術が適用可能になるという意義がある。

（２）異種のネットワークを跨ぐ仮想計算機ライブマイグレーション

（１）の仮想計算機マイグレーション技術を発展させ、異なる通信インタフェースをもつクラスタ計算機を跨ぐマイグレーション機構 Ninja migration を新たに設計・開発した[雑誌論文⑥、⑦]。本技術により、従来実現できていた仮想クラスタ環境における高性能 I/O に加えて、アプリケーションの実行を止めることなく、要求性能やコストを基準に、実行環境を選択、移行する自由度が高まった。これはクラウド環境における柔軟な運用やディザスタリカバリに資する機能である。ディザスタリカバリ応用については、大規模災害時でも遠隔地に仮想計算機をマイグレーションすることで、サービスを継続するための技術を開発し、その予備評価を実施した。クラスタの通信インタフェースに依存しない仮想計算機マイグレーションを実現することで、InfiniBand クラスタを商用クラウドに退避できるなど、避難先の選択肢が増加し、災害時のサービス継続性の向上が期待できる[学会発表⑧]。

準仮想化デバイスの代わりに PCI パススルー技術を用いて仮想計算機にデバイスと直接割り当てた場合、仮想計算機モニタはデバイスのレジスタに書き込まれたデータ（InfiniBand の Local ID や Queue Pair 番号など）を保存・復帰できないので、マイグレーションは動作しない。一方、PCI ホットプラグ機能を使えば、一時的に PCI パススルーデバイスを仮想計算機から取り外すことでマイグレーションは可能になるが、仮想計算機モニタが安全にデバイスを取り外すタイミングを知ることは困難である。メッセージ送信中にデバイスを取り外すと、メッセージの欠損やアプリケーションの異常終了を引き起こす可能性がある。そこで、MPI ランタイムと仮想計算機モニタが連携するための SymVirt 機構を適用ことで、全ての仮想計算機を送信中のメッセージがない安全な状態にしてから、デバイスを取り外す（図２、３）。

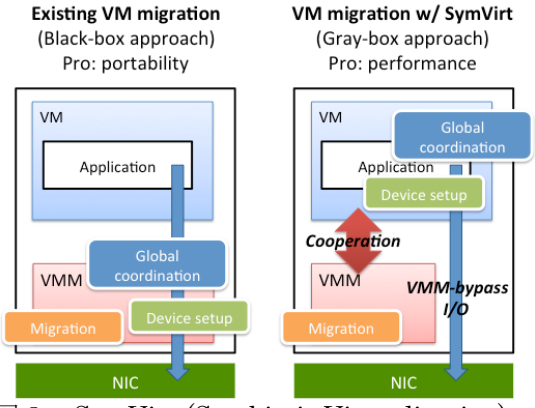

図２　SymVirt (Symbiotic Virtualization)

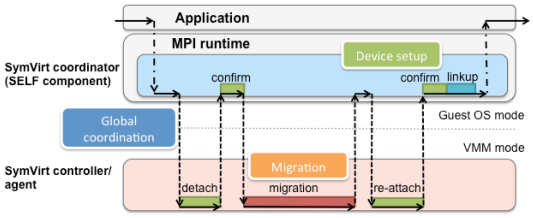

図３　Ninja migration の実装

さらに Infiniband クラスタと Ethernet クラスタなど、計算ノード間のインターコネクトが異なるクラスタ間でのマイグレーションは下記のように実現した。MPI ランタイムでは複数の通信デバイスを利用するために通信層が抽象化されている。また、チェックポイント・リスタートに対応するため、全プロセスが送信途中のメッセージが存在しない状態を作り、かつリスタート時にコネクションを再確立する機能を有している。この機能を流用することで、チェックポイントの代わりにマイグレーションを行い、マイグレーション後に、MPI ランタイムが利用可能な通信デバイスを検出し、アプリケーションが利用するコネクションを張り直す。これはアプリケーションからは透過に実行されるので、マイグレーション中に通信デバイスが切り替わったとしても、アプリケーションの再実行は不要である。

図４に Ninja migration のオーバヘッドを示す。通常実行時のオーバヘッドは無視でき、マイグレーション時間はメモリ転送量に比例することがわかる。また、InfiniBand のリンクアップ時間のオーバヘッドが大きく、その短縮が今後の課題である。

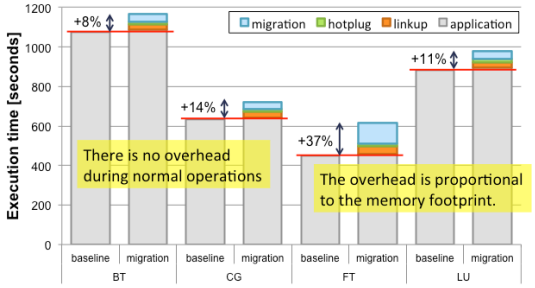

図４　Ninja migration のオーバヘッド
（NPB　64 プロセス、クラス D）

（３）資源管理システムと連携したライブマイグレーションの制御

クラウドミドルウェアとして普及が進む Apache CloudStack へ、仮想計算機のゲスト OS から物理計算機の PCI デバイスに直接アクセスする PCI パススルー機能および SR-IOV（Single-Root I/O Virtualization）を対応させ I/O 性能の改善を図った[雑誌論文③、④、学会発表①]。本成果の一部は産総研内のプライベートクラウドである AIST Super Green Cloud (ASGC)の運用で利用されている

[雑誌論文①、②]。また、ASGC 上で動作する仮想クラスタを必要に応じて Amazon EC2 でも再構築できるようにした。これにより利用者は InfiniBand の高い入出力性能を損なうことなくクラウド環境の利便性を享受できることを確認した。

現在の Ninja migration の実装では、仮想計算機モニタの改変を要するが、ホスト OS のデーモンプログラムとして再実装することで、実サービスの運用で利用しやすくすることが今後の課題である。

＜引用文献＞

[1] Wei Huang 他, “Nomad: Migrating OS-bypass Networks in Virtual Machines”, ACM SIGPLAN/SIGOPS Conference on Virtual Execution Environments, 2007.
[2] P. Liu 他, “Heterogeneous Live Migration of Virtual Machines”, International Workshop on Virtualization Technology (IWVT), 2008.

５．主な発表論文等

〔雑誌論文〕（計　８件）

①Nuttapong Chakthranont, Phonlawat Khunphet, Ryousei Takano, Tsutomu Ikegami, "Exploring the Performance Impact of Virtualization on an HPC Cloud," Proceedings of 2014 IEEE 6th International Conference on Cloud Computing and Science (CloudCom), 査読有, pp.426-432, 2014, DOI: 10.1109/CloudCom.2014.71
②高野了成, 谷村勇輔, 竹房あつ子, 広渕崇宏, 田中良夫, "高性能かつスケーラブルな HPC クラウド AIST Super Green Cloud", 情報処理学会, 査読無, 2014-HPC-145, pp.1-6, 2014.
③Pawit Pornkitprasan, Vasaka Visoottiviseth, Ryousei Takano, "Engaging Hardware-Virtualized Network Devices in Cloud Data Centers", Proceedings of 3rd ICT-ISPC 2014, 査読有, pp.1-2, 2014, DOI: 10.1109/ICT-ISPC.2014.6923234
④ Tanasak Janpan, Vasaka Visoottiviseth, Ryousei Takano, "A Virtual Machine Consolidation Framework for CloudStack Platforms", Proceedings of 28th International Conference on Information Networking (ICOIN), 査読有, pp.28-33, 2014, DOI: 10.1109/ICOIN.2014.6799494
⑤Ryousei Takano, Hidemoto Nakada, Takahiro Hirofuchi, Yoshio Tanaka, and Tomohiro Kudoh, "Cooperative VM Migration: a Symbiotic Virtualization Mechanism by Leveraging the Guest OS Knowledge", IEICE Transactions on Information and Systems, 査読有, Vol.E96-D, No.12, pp.2675-2683, 2013.
⑥高野了成, 中田秀基, 広渕崇宏, 田中良夫, 工藤知宏, "異種クラスタを跨がる仮想マシンマイグレーション機構", 情報処理学会, 査読無, 2013-OS-126, pp.1-6, 2013.
⑦Ryousei Takano, Hidemoto Nakada, Takahiro Hirofuchi, Yoshio Tanaka, and Tomohiro Kudoh, "Ninja Migration: An Interconnect-transparent Migration for Heterogeneous Data Centers", High-Performance Grid and Cloud Workshop (HPGC), in conjunction with IEEE IPDPS, 査読有, 2013, DOI: 10.1109/IPDPSW.2013.114
⑧Ryousei Takano, Hidemoto Nakada, Takahiro Hirofuchi, Yoshio Tanaka, and Tomohiro Kudoh, "Cooperative VM Migration for a Virtualized HPC Cluster with VMM Bypass I/O devices", Proceedings of IEEE 8th International Conference on eScience 2012, 査読有, pp.1-8, 2012, DOI: 10.1109/eScience.2012.6404487

〔学会発表〕（計　９件）

①Ryousei Takano, Yusuke Tanimura, Akihiko Oota, Hiroki Oohashi, Keiichi Yusa, and Yoshio Tanaka, "AIST Super Green Cloud: lessons learned from the operation and the performance evaluation of HPC cloud," Internet Symposium on Grids and Clouds 2015, 2015 年 3 月.
②Nuttapong Chakthranont, Phonlawat Khunphet, Ryousei Takano, Tsutomu Ikegami, "Exploring the Performance Impact of Virtualization on an HPC Cloud," Proceedings of 2014 IEEE 6th International Conference on Cloud Computing and Science (CloudCom), 査読有, 2014 年 12 月.
③高野了成, 谷村勇輔, 竹房あつ子, 広渕崇宏, 田中良夫, "高性能かつスケーラブルな HPC クラウド AIST Super Green Cloud", 情報処理学会, 査読無, 2014-HPC-145, 2014 年 8 月.
④Pawit Pornkitprasan, Vasaka Visoottiviseth, Ryousei Takano, "Engaging Hardware-Virtualized Network Devices in Cloud Data Centers", Proceedings of 3rd ICT-ISPC 2014, 査読有, pp.1-2, 2014 年 3 月
⑤ Tanasak Janpan, Vasaka Visoottiviseth, Ryousei Takano, "A Virtual Machine Consolidation Framework for CloudStack Platforms", Proceedings of 28th International Conference on Information Networking (ICOIN), 査読有, pp.28-33, 2014 年 2 月.
⑥高野了成, 中田秀基, 広渕崇宏, 田中良夫, 工藤知宏, "異種クラスタを跨がる仮想マシンマイグレーション機構", 情報処理学会, 査読無, 2013-OS-126, 2013 年 8 月.
⑦Ryousei Takano, Hidemoto Nakada, Takahiro Hirofuchi, Yoshio Tanaka, and Tomohiro Kudoh, "Ninja Migration: An Interconnect-transparent Migration for Heterogeneous Data Centers", High-Performance Grid and Cloud Workshop (HPGC), in conjunction with IEEE IPDPS, 査読有, 2013 年 5 月.
⑧Ryousei Takano, Hidemoto Nakada, Takahiro Hirofuchi, Yoshio Tanaka, and Tomohiro Kudoh, "Preliminary Evaluation of Disaster Recovery based on Interconnect-transparent VM

Migration", PRAGMA Workshop 24 Poster session, 査読有, 2013 年 3 月.
⑨Ryousei Takano, Hidemoto Nakada, Takahiro Hirofuchi, Yoshio Tanaka, and Tomohiro Kudoh, "Cooperative VM Migration for a Virtualized HPC Cluster with VMM Bypass I/O devices", Proceedings of IEEE 8th International Conference on eScience 2012, 査読有, 2012 年 10 月.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

○取得状況（計　0 件）

〔その他〕
https://github.com/AIST-ITRI/ninja

６．研究組織
(1)研究代表者
高野　了成　（TAKANO, Ryousei）
産業技術総合研究所・情報技術研究部門・研究グループ長
研究者番号：10509516

(2)研究分担者
なし

(3)連携研究者
なし