戦略的創造研究推進事業　CREST
研究領域「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」
研究課題「インタークラウドを活用したアプリケーション中心型オーバーレイクラウド技術に関する研究」

研究終了報告書

研究期間　2015年10月～2021年3月

研究代表者：合田　憲人
（情報・システム研究機構国立情報学研究所　教授）

§1 研究実施の概要

(1)実施概要

クラウドサービスの登場により、簡便かつ短時間で計算資源を利用することが可能となった。しかし、ビッグデータ解析を行うためには、ソフトウェアやネットワーク設定等のソフトウェア環境の構築が必要であり、研究者がそのために非常に長い時間（例えば数週間以上）を要していた。また、データ解析アプリケーションプログラムの中には、プログラムを実行する計算機の OS やライブラリのバージョン等の違いにより、計算結果が異なるものがあり、論文等で報告されている計算機実験の追試を行う上で問題となっている。これらの問題を解決するために、本研究では、ネットワーク接続された複数の異なるクラウド上にアプリケーション毎に最適化されたビッグデータ解析ソフトウェア環境を高速かつ自動的、また再現性をもって構築するための技術を開発するとともに、ゲノム解析ならびに流体音解析等のアプリケーションのソフトウェア環境構築における有効性を実証した。

本研究の中核となる成果は、合田グループが開発した基盤ミドルウェア Virtual Cloud Provider（VCP)である。VCP は、クラウド上の計算資源の選択と確保、ネットワーク設定、ソフトウェアの配備等の処理をユーザに替わって自動的に実行する。また、ユーザが計算資源やアプリケーションの実行状況を監視する機能も提供する。VCP を用いることにより、アプリケーション分野の研究者が従来は数週間から 1 ヶ月以上を要して人手で構築していたデータ解析ソフトウェア環境を自動的かつ迅速（環境の規模に依存するが数分から数十分）に構築することが可能となり、研究者がデータ解析を開始するまでに要する時間を大幅に短縮できるようになった。また、異なるクラウド基盤上に同一のビッグデータ解析ソフトウェア環境を再現して構築することが可能となり、計算機実験の追試を容易に実施することが可能となった。VCP は、国立情報学研究所が進めるクラウド上にソフトウェア環境を構築する実サービスにも活用されている。

ビッグデータ解析環境の構築では、アプリケーション毎に最適な計算資源を選択することが重要となる。棟朝グループでは、多数制約付きの多数目的最適化を実現する手法ならびに最適化エンジンを開発した。例えば、最適な計算資源数と計算資源種別を同時に導出することを実現する、従来手法に比べて多数の目的関数に対して最適解を導出する等、従来手法ではできなかった多数の制約や目的関数を有する複雑な科学技術ワークフローへの最適資源選択を可能とした。これらの技術は、合田グループのミドルウェアから利用可能である。

本研究では、研究開始当初からアプリケーションコミュニティの研究者と共同でビッグデータ解析アプリケーションを開発するとともに、VCP を用いた実証実験を進めた。小笠原グループでは、DNA 塩基配列データをクラウド上で解析するためのツール群を開発するとともに、VCP を用いてデータ解析環境を構築できることを実証した。また、ツール群の実行に必要な計算資源量に関するメトリクスを自動収集する機構を開発し、VCP や棟朝グループの最適化エンジンから利用可能とすることで、実際のゲノム解析アプリケーション向けに最適化された計算資源選択を可能とした。これらのツール群はオープンソースとして公開されている。生命系研究者とのコミュニティ形成を行い、実際にデータを解析するゲノム科学者の意見をツール開発に取り込むとともに、研究成果やノウハウをゲノム科学者に還元する体制を構築することにより、同分野の科学者のクラウドを用いた大規模データ解析の理解を深めた。

また、小野グループでは、マルチスケール・マルチフィジックスな過渡現象など動的で不確定な状況変化に対応するシミュレーション手法を世界に先駆けて開発し、高精度の管楽器の音孔開閉を対象にした流体音数値解析シミュレーションを実現することに成功した。また、同手法において必要となる連成計算を制御するために開発された連成計算機構が他の工業分野でも有効に適用可能であることを示した。

本研究の実証を目的として、本研究に参加する学術研究機関の計算資源、ならびにパブリッククラウドを SINET5 の L2VPN を介して接続し、高速かつ安全な通信が可能なインタークラウド環境を構成した。また、海外の計算資源との連携として、米国の学術向けクラウドサービスである Chameleon Cloud の利用も可能とした。実証実験基盤の整備・運用は、實本グループが中

心となって全グループと共同で進めている。また、實本グループでは、スーパーコンピュータとクラウドの計算資源を連携させる技術についての開発を進めた。従来のスーパーコンピュータは内部の計算ノードが外部と強く切り離されており、クラウドとの連携利用が困難であった。本研究では、内部計算ノードと外部のクラウドとの安全な通信手段の確立、コンテナ実行環境の構築、計算ノードの効率的な共有手法を開発し、スーパーコンピュータとクラウドとの効率的な連携利用を可能とした。また、研究成果を東京工業大学が運用するスーパーコンピュータ(TSUBAME 3.0)上での実サービスとして展開した。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1. クラウド上でのビッグデータ解析ソフトウェア環境構築手法の開発
概要:
本研究が開発した基盤ミドルウェアを用い、クラウド基盤上にゲノム解析ならびに流体音解析シミュレーションのビッグデータ解析ソフトウェア環境を自動的かつ再現性を持って構築できることを示した。従来技術では、適用可能なクラウド基盤が限定される等、異なるクラウド基盤間で同一のソフトウェア環境を再現して構築することが難しいという問題があるが、本研究では異なるクラウド基盤上に再現性を持ってソフトウェア環境を構築することを可能とした。本研究成果は、国際会議 IEEE CLOUD 2016 等で発表された。

2. 多数制約・多数目的最適化によるクラウド上での最適資源選択手法の開発
概要:
アプリケーションの実行に最適なクラウド上の計算資源選択を目的として、多数制約付きの多数目的最適化を実現する手法ならびに最適化エンジンを開発した。例えば、最適な計算資源数と計算資源種別を同時に導出することを実現する、従来手法に比べて多数の目的関数に対して最適解を導出する等、従来手法ではできなかった多数の制約や目的関数を有する複雑な科学技術ワークフローへの最適資源選択を可能とした。本研究成果は、国際会議 IEEE Big Data 2016 等で発表された。

3. マルチスケール・マルチフィジックスな過渡現象を対象とした連成計算機構の開発
概要:
マルチスケール・マルチフィジックスな過渡現象など動的で不確定な状況変化に対応するシミュレーション手法を世界に先駆けて開発し、高精度の管楽器の音孔開閉を対象にした流体音数値解析シミュレーションを実現することに成功した。また、同手法において必要となる連成計算を制御するために開発された連成計算機構が、不確定性・想定外対応を実現する方法として他の工業分野でも有効に適用可能であることを示した。

＜科学技術イノベーションに大きく寄与する成果＞

1. クラウド上でのデータ解析環境自動構築サービスの展開
概要:
国立情報学研究所の事業サービスとして、本研究が開発した基盤ミドルウェアを活用し、クラウド上にソフトウェア環境を自動的に構築するサービスを提供している。本サービスでは、ゲノム解析ソフトウェア環境をはじめとして、HPC ソフトウェア環境や LMS (Learning Management System) といった教育ソフトウェア環境の構築も可能であり、クラウドを活用した研究教育の促進につながるものである。

2. クラウドを活用するゲノム解析ツール開発ならびにコミュニティ形成

概要：
DNA 塩基配列データをクラウド上で解析するための環境を自動的かつ再現性をもって構築するためのツール群を開発し、公開した。生命系研究者とのコミュニティ形成を行い、実際にデータを解析するゲノム科学者の意見をツール開発に取り込むとともに、研究成果やノウハウをゲノム科学者に還元する体制を構築することにより、同分野の科学者のクラウドを用いた大規模データ解析の理解を深めた。特にデータ解析環境の再現性の議論や実践は欧米にも先んじている。

3．スーパーコンピュータ・クラウド連携技術の開発
概要：
従来のスーパーコンピュータは内部の計算ノードが外部と強く切り離されており、クラウドとの連携利用が困難であった。本研究では、内部計算ノードと外部のクラウドとの安全な通信手段の確立、コンテナ実行環境の構築、計算ノードの効率的な共有手法を開発し、スーパーコンピュータとクラウドとの効率的な連携利用を可能とした。また、研究成果を東京工業大学が運用するスーパーコンピュータ（TSUBAME 3.0）上での実サービスとして展開した。

＜代表的な論文＞
1．Shigetoshi Yokoyama, Yoshinobu Masatani, Tazro Ohta, Osamu Ogasawara, Nobukazu Yoshioka, Kai Liu, Kento Aida, ″Reproducible Scientific Computing Environment with Overlay Cloud Architecture″, Proc. of the International Conference on Cloud Computing (CLOUD 2016), 2016
概要：
本研究の目標であるアプリケーション中心型オーバーレイクラウド技術のフィーシビリティスタディとして、インタークラウド実証実験基盤のプロトタイプを構築するとともに、ゲノム解析アプリケーションを用いた予備実験を実施した。この結果、本研究が開発を進めている基盤ミドルウェアを用いて仮想クラウド環境（ソフトウェア環境）を異なるクラウド基盤上に構築可能であり、かつこれらの仮想クラウド環境上でのアプリケーション実行結果の再現性も得られることを確認した。

2．Katunori Miura, Tazro Ohta, Courtney Powell, Masaharu Munetomo: Intercloud Brokerages based on PLS Method for deploying Infrastructure for Big Data Analytics, Workshop of Big Data for Cloud Operation Management, Proceedings of the 2016 IEEE International Conference on Big Data (IEEE Big Data 2016), 2016
概要：
本論文では、等価変換理論に基づくクラウド資源選択法を提案した。提案手法では、クラウドブロケージ方式の確立に等価変換アルゴリズムを採用することにより、システム要件に対する計算基盤の必要十分性の証明を可能とする。その結果、アプリケーション最適実行に必要な任意の仮想マシン台数およびネットワーク構成から成る様々な計算基盤の発見を可能にした。これは SAT ソルバを基とする従来方式では難しい解探索である。

3．Ohta, T., Tanjo, T., & Ogasawara, O. (2019). Accumulating computational resource usage of genomic data analysis workflow to optimize cloud computing instance selection. GigaScience, 8(4).
概要：
大規模ゲノム解析では処理速度とコストの両面で最適な計算機を選択することが非常に重要である。本研究では多種類の解析ソフトの実行制御に用いられる Common Workflow Language 処理系を利用して最適な計算資源を選択するためのメトリクス収集システムを開発した。これを用いて複数の異なるワークフローを比較し、それぞれに最適なクラウド資源を提案できることを示した。成果はオープンソース化された。

§2 研究実施体制

(1)研究チームの体制について

（1）「合田」グループ

① 研究代表者：合田　憲人　（情報・システム研究機構　国立情報学研究所、教授）

② 研究項目

- 実行環境再構成技術に関する研究
- 基盤ミドルウェア開発
- 実証実験基盤の整備

（2）「棟朝」グループ

①主たる共同研究者：棟朝　雅晴　（北海道大学情報基盤センター、教授）

② 研究項目

- 多数目的最適化アルゴリズムに関する検討および実装
- システム構成仕様記述方式に関する検討および実装

（3）「小笠原」グループ

① 主たる共同研究者：小笠原　理　（情報・システム研究機構 国立遺伝学研究所、特任准教授）

② 研究項目

- ゲノム解析ワークフローに関する研究
- ゲノム配列自動アノテーションに関する研究

（4）「小野」グループ

① 主たる共同研究者：小野　謙二　（九州大学情報基盤研究開発センター、教授）

② 研究項目

- 連成計算管理機構に関する検討・実装
- 不確定要素対応に関する検討・設計

（5）「實本」グループ

① 主たる共同研究者：實本 英之　（東京工業大学学術国際情報センター、助教）

② 研究項目

- スーパーコンピュータとクラウドリソースとの連携に関する研究
- インタークラウド環境上のデータ保存・アクセス方式に関する研究
- インタークラウド環境下における HPC ビッグデータ解析の適用に関する研究

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について

ノースカロライナ大学の Yufeng Xin 研究員と協力し、米国の学術向けクラウドサービスである Chameleon Cloud 上に VCP を利用してビッグデータ解析ソフトウェア環境を構築可能な国際的なインタークラウド実験環境を構築した。インタークラウド実験環境の構築にあたっては、国際的なネットワーク実験基盤である AutGOLE の協力も得ながら日米にまたがる VPN をオンデマンドに作成することに成功した。

国際的なオープンソースソフトウェア・プロジェクトである Common Workflow Language (CWL) Project[1] に、合田グループ 丹生智也と小笠原グループ 大田達郎の 2 名がコミッタとして参加している。さらに、本プロジェクトで主催した国内コミュニティからさらに 2

[1] https://commonwl.org/

名がコミッタとして参加している。コミッタらは CWL の開発や意思決定の議論、広報や教材の開発などに従事している。

また、小笠原グループで開発したソフトウェアはゲノム医科学分野の産学連携国際団体である Global Alliance for Genomics and Health (GA4GH) で開発された規約に基づいた開発を行っている。プロジェクト期間中に開催された国際会議 BioHackathon などを通じて GA4GH のワーキンググループと議論や意見交換を行っており、本研究における開発によって得られた知見をフィードバックしている。