課題番号: jh130019-MD03

広域分散ファイルシステムに基づく「ビッグテーブル」型の超大規模データ処理系の構築と機能および性能評価

東田 学（大阪大学）

概要　HPCI-JHPCN システムとして提供される多拠点 VM ホスティング環境である先端ソフトウェア運用基盤システムと、HPCI 共通運用システムとして整備されている HPCI 共有ストレージの基盤技術である Gfarm を活用し、Hadoop/HBase によるいわゆる「ビッグテーブル」型の超大規模データ処理系を地理的に分散させて構築し、その機能および性能評価を行っている。さらに、具体的な運用シナリオの遂行を通して、その有効性の検証を行なった。

1. 研究の目的と意義

本研究の目的: HPCI-JHPCN システムとして提供される多拠点 VM ホスティング環境である先端ソフトウェア運用基盤システムと、HPCI 共通運用システムとして整備されている HPCI 共有ストレージを活用し、Hadoop/HBase によるいわゆる「ビッグテーブル」型の超大規模データ処理系を地理的に分散させて構築し、その機能および性能評価を行う。それと同時に、具体的な運用シナリオの遂行を通して、その有効性の検証を行う。

具体的には、まず、北海道大学、東京大学、東京工業大学、九州大学で整備されている RENKEI-VPE による VM ホスティング環境上に、Hadoop によるビッグデータ処理系および HBase によるビッグテーブル処理系を構築する。通常、Hadoop による MapReduce 処理に際して分散データ蓄積を行う HDSF（Hadoop Distributed Filesystem）は、各処理ノードのローカルストレージ上に構築するが、本研究課題においては、広域分散ファイルシステムである Gfarm 上に構築する。その際、Gfarm をユーザファイルシステムとしてマウントする方式ではなく、筑波大で開発され阪大で機能拡張を行った Hadoop-Gfarm プラグインを用いて Hadoop から Gfarm API 呼び出して入出力を行う。その際の導入・構築性の検証とそれぞれの拠点内での性能検証を行うと同時に、Gfarm のファイル複製機能を活用し、拠点間でのデータ連携性や移行性を検証する。

これらの検証結果を元に、Hadoop-Gfarm プラグインの機能・性能改善を行った上で、実運用を想定したシナリオ検証として、HPCI 共通運用システム整備において開発を進めている HPCI アカウント集計システムを導入し、テータの投入・解析・表示機能を多拠点に分散させた際の運用性の検証を行う。これらの評価、検証を通じて、超大規模数値計算システムと連携可能な大規模データ処理システムとしての有用性を検討する。

本研究の意義: 超大規模数値計算に伴う超大規模データ処理技術の整備と利用技術の普及が求められている。

超大規模データ処理分野においては、いわゆる「ビッグデータ」処理系として Google BigTable を初めとして、そのクローンである Hadoop などが開発され、既に商業データセンターにおいて PaaS のようなサービス形態で提供されている。しかし、超大規模数値計算によって入出力される数 TB 規模のデータをこれらの商業データセンターに逐次転送し相互に解析処理を行う事は、ネットワーク転送にかかるコストを勘案すると現実的ではない。一方、学術分野のスーパーコンピュータセンターにおける超大規模データ処理環境は、依然として整備途上であり、利用技術の啓蒙を踏まえて整備を加速させる必要がある。

一方、HPCI-JHPCN システムとして提供されている先端ソフトウェア運用基盤のように仮想 OS をホスティングし IaaS 型サービス提供する環

境の整備も行われている。しかし、仮想 OS 上で超大規模数値計算や超大規模データ処理を行うには、仮想化に伴う入出力ボトルネックをひとつひとつ丁寧に解消する必要があり、これもいまだ普及には至っていない。

本研究では、これまで超大規模数値計算環境や超大規模データ処理環境の構築に携わってきた研究者の支援を受け、個々の要素技術の整備とそれらの連携に際して発生する問題点を明らかにするとともに、それらを解消し、地理的に離れた拠点間に超大規模データ処理系を構築する運用技術の確立を目指し、具体的な運用シナリオを遂行するために必要となる機能や性能を明確にする。さらに、多拠点間で大容量のデータ共有を行うためのネットワーク環境の検証も同時に行う。この過程で得られる知見と整備された環境を、既存の超大規模数値計算環境と連携させることによって、拠点間を横断する研究開発の加速を促進する。

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拠点名および役割分担

- 北海道大学: RENKEI-VPE 運用基盤提供
- 東京大学: RENKEI-VPE 運用基盤提供
- 東京工業大学: RENKEI-VPE 運用基盤提供と管理
- 大阪大学: 検証システム構築と評価
- 九州大学: RENKEI-VPE 運用基盤提供

(2) 共同研究分野

- 超大規模データ処理系応用分野
- 超大容量ネットワーク技術分野
- 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

これまで、超大規模データ処理環境を実現するための個々の要素技術開発は行われているが、大規模数値計算と連携させることによって、より大規模な研究を促すような、実運用を想定した機能評価や性能評価の取り組みが遅れている。また、仮想 OS 上で仮想化に伴って生じたオーバヘッドを解消し、実システムと遜色ない性能を実現するための機能評価や性能評価も遅れている。本研究を通じて、各研究拠点が提供する資源を活用し、個々の要素技術を連携させることによって生じる課題を明らかにし、参加する共同研究者の協力を得て即時解消することにより、環境整備を加速することができると考えている。

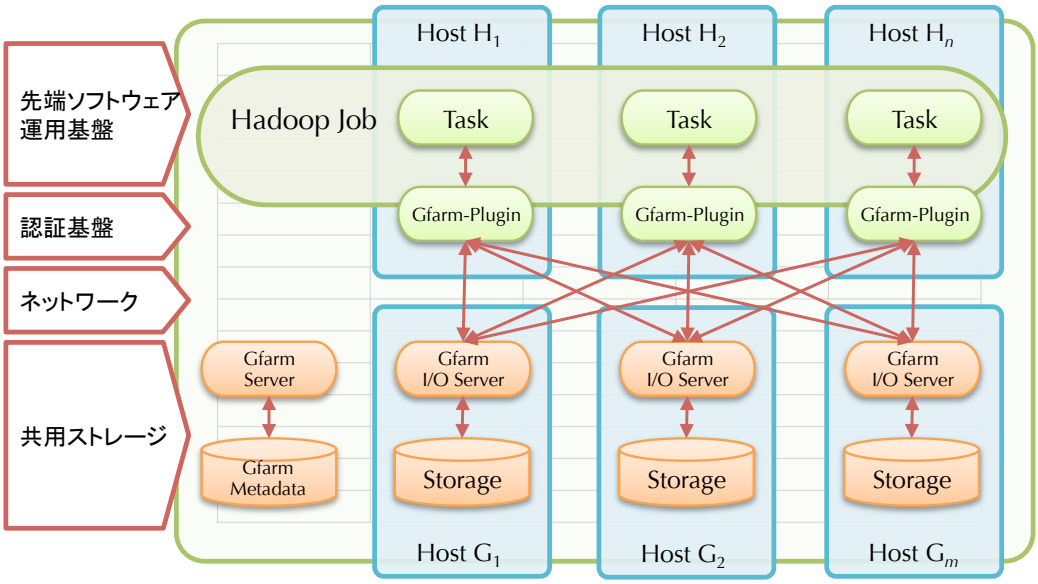

図 1: 本研究開発における基本アーキテクチャ

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

本研究グループでは、我が国における e-サイエンスを活用した研究を促進することを目指し、学際大規模情報基盤共同利用・共同研究拠点および革新的ハイパフォーマンスコンピューティングインフラ（HPCI）システム構成機関に設置された計算機システム、およびこれらを接続する学術情報ネットワークである SINET4 から構成される実用的なグリッド基盤を構築・運用する技術を確立することを目的として、認証基盤、ストレージ共有および先端的なソフトウェア運用基盤に関する研究開発を行ってきた（図 1)。

認証基盤に関連しては、HPCI システム構成機関で管理されるユーザアカウント管理システムと国立情報学研究所が運用するグリッド認証システムを Shibboleth 認証連携技術により連携し、さらに Grid Security Infrastructure (GSI)を用いて、グリッド基盤にシングルサインオンする認証基盤を設計および構築した。また，認証基盤の本格運用に向けた検証を行うため、認証基盤の実用性を重視した実証実験を行い、本認証基盤の有効性を確認した。

ストレージ共有に関連しては、広域分散ファイルシステム Gfarm を用い、HPCI システム構成機関に対して 20PB を超える共有ストレージサービスを提供している。HPCI の利用者は、シングルサインオンによって一様にこの大規模共有ストレージを利用できる。さらに、Hadoop MapReduce アプリケーションから Gfarm ファイルシステムを利用するためのプラグインの設計と実装を行っており、入出力ファイルを専用の HDFS にステージングすることなく共有することが可能になると同時に、入出力性能も勝っていることを確認した。

さらに、グリッド技術の普及、特に資源管理者のグリッドサービス管理コスト削減と、利用者による多拠点資源間でのデータ共有の容易化を実現するための先端ソフトウェア運用基盤として、RENKEI-VPE と呼ぶ分散環境における VM ライフサイクル管理システムを開発し、運用・配備を進めている。資源管理者は RENKEI-VPE を用いて、VM としてワークフローツールやスケジューリングシステム等のグリッドサービスを実行することができる。利用者は、そのグリッドサービスと多拠点間でのデータ共有を高速かつ透過的に実現する分散システムである RENKEI-PoP が提供する広域データ共有機能を用いてグリッドアプリケーションを実行できる。

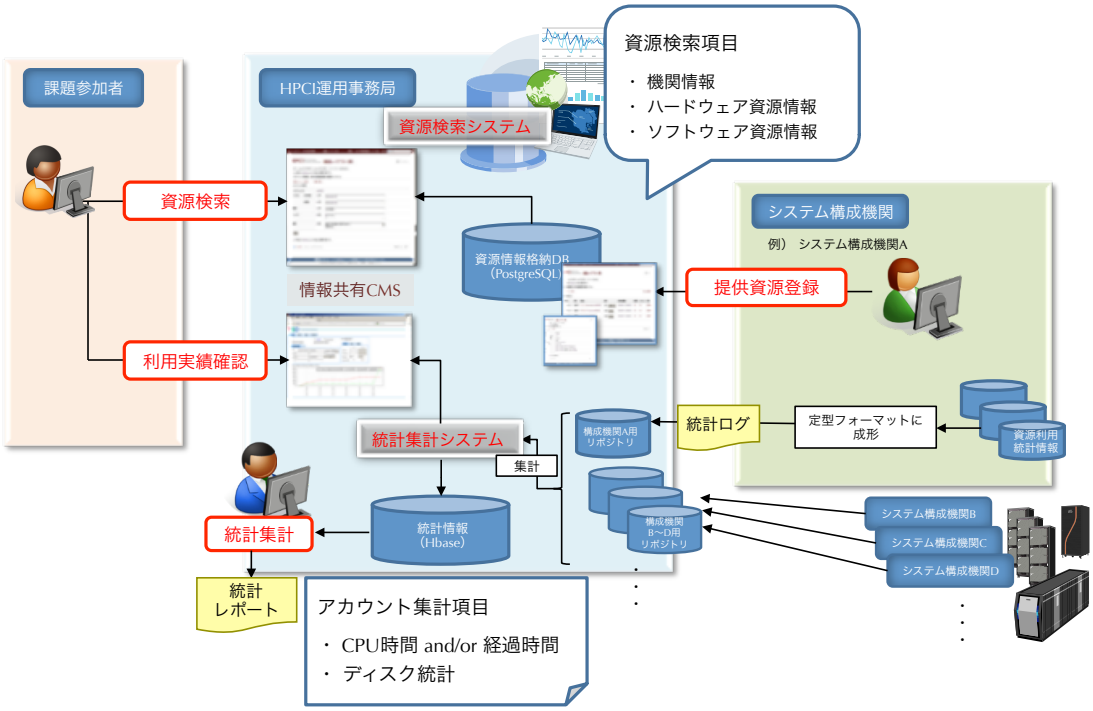

図 2: HPCI アカウント集計システムの概略

これまでの HPCI 共通運用システム整備において、HPCI アカウント集計システムとして、Hadoop/HBase によって実現される「ビッグテーブル」型データベースに時系列データを蓄積し、統計処理した結果をグラフ表示するシステム開発を行った。当該システムにおいては、時系列データとしては、HPCI システム上で実行される個々のジョブの CPU 時間やメモリ容量、また、日々のディスク利用量を逐次蓄積し統計処理を行い、HPCI システムの利用者や運用管理者に、随時、任意期間の集計結果をグラフ表示可能にする（図 2）。

これまでのグリッドミドルウェアの多くは Ganglia を用いており、時系列情報が RRD 形式に保存されていた。各々のデータが個別の固定容量ファイルに蓄積されており、時間の経過に伴って蓄積データの再サンプリングが行われ粒度が荒くなる、また、値を取り出すためのインターフェイスが別途必要であるという課題があった。OpenTSDB では、Hadoop/HBase をバックエンドとして時系列データを一括投入しているために、さまざまな形態で値の呼び出しが可能であり、利用者向けまたは管理者向けのグラフ表示ができると同時に、ビックデータ型の傾向分析も可能である。これらの情報を参照することによってより効率的なジョブスケジューリング機能を実現できると考えられる。これまでも様々なメタスケジューラが提案されるが、実際の業務フローに適合することが難しく、実用化されないという悪循環があった。HPCI の運用実績が蓄積することによって、その知見を活用する実用的なメタスケジューラを提案することが可能になったと考えている。

このプロトタイプシステムを、HPCI-JHPCN システムとしても資源提供されている阪大の Express 5800/53Xh において、仮想 OS 上の Hadoop/HBase をソフトウェア基盤として構築し、10 月から評価運用を開始した。当該クラスタシステムは、ベース OS として Windows 7、ハイパーバイザとして Parallels Desktop Workstation Extreme を採用している。Hadoop 環境は、仮想 OS 実行環境の CentOS 上に構築されている。Hadoop 分散ファイルシステムは、各計算ノードが有する SSD 上に構築されている。総容量は 60TB 以上で、仮想 OS 実行環境においてもノードあたり数百 MB/s でのデータ入出力が可能である。時系列データを効率的に HDFS に格納するために、HBase (http://hbase.apache.org/) による分散 KVS (Key-Value Store) を構成し、さらに、HDFS へのログ収集機能を提供する Flume

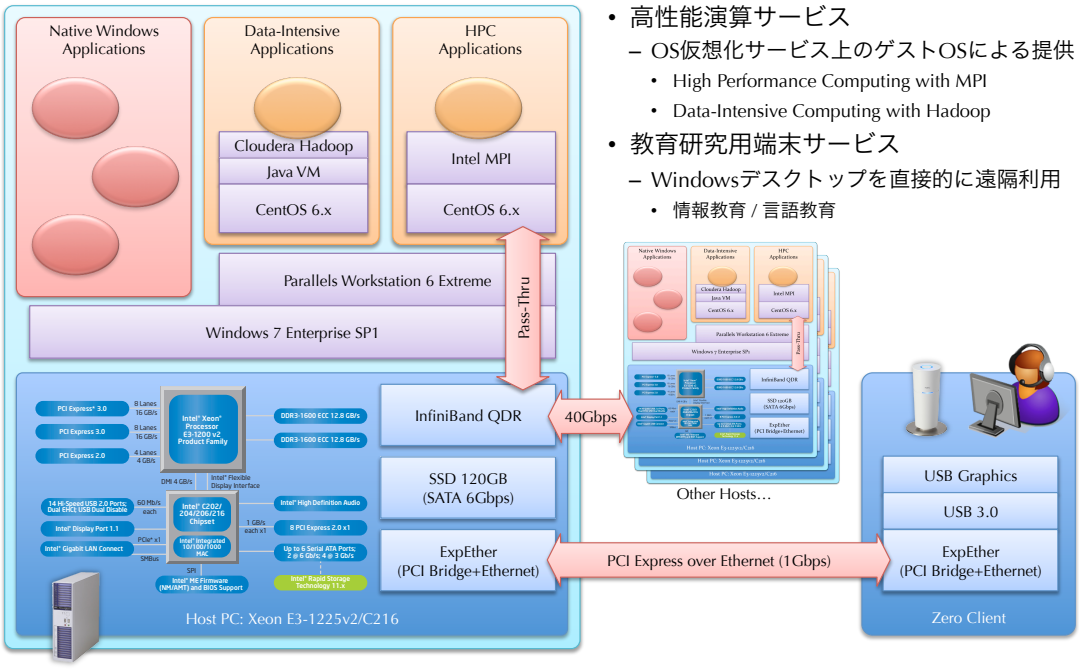

図 3: 阪大クラスタ型汎用コンピュータシステムの構成

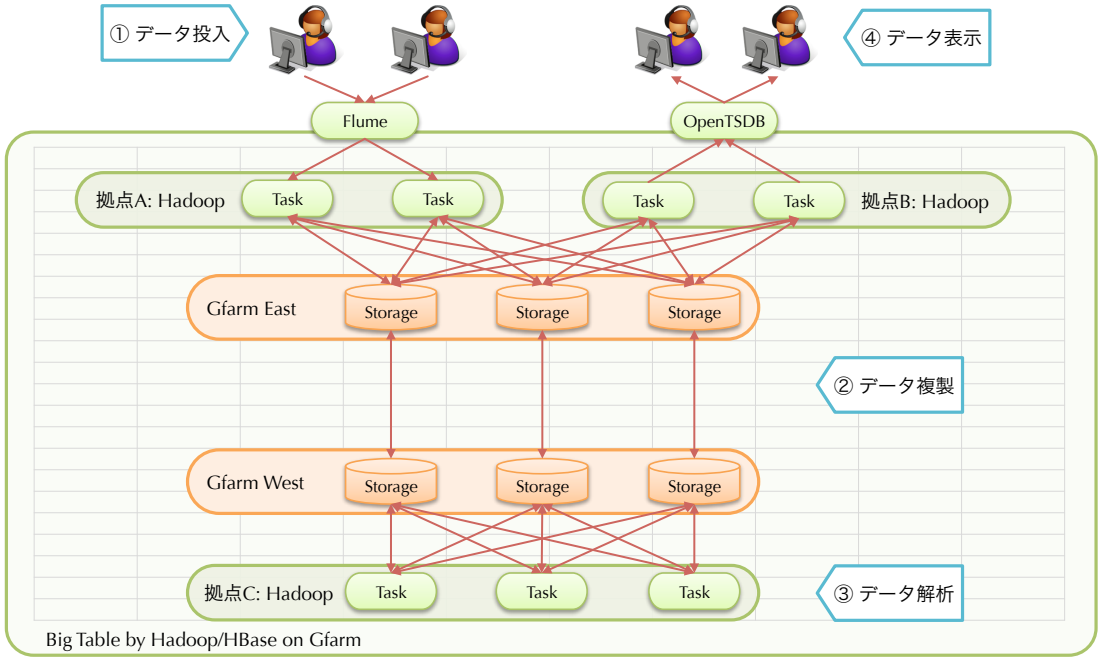

図 4: 検証シナリオ

(http://flume.apache.org/) を介して、外部ソースから時系列データを投入できる環境を構成している。さらに、ロボットから時系列データを継続的に投入するフロントエンドサービスを開発しており、OpenTSDB (Time-Series Database - http://opentsdb.net/) を組み合わせることで、スケーラブルなデータ蓄積から逐次的なグラフ表示まで可能であることを検証している。

本研究で運用シナリオとして採用した HPCI アカウント集計システムの開発も進めている。HPCI システム構成機関が提供する計算資源の利用統計情報を、Hadoop/HBase によって提供されるビッグテーブル型のデータベースに逐次的に蓄積し統計処理を行い、HPCI システムの利用者や運用管理者に、随時、任意期間の集計結果をグラフ表示することができる。このアカウント集計システムに、上述の Hadoop-Gfarm プラグインを適用し、実運用シナリオを想定した機能および性能評価を行った。具体的には、以下のような手順である

(1-1) 構築性および機能・性能評価

本研究課題では、まず、各拠点の RENKEI-VPE による仮想 OS ホスティング環境への Hadoop/HBase 環境の導入性を検証し、各拠点内および拠点間の Gfarm I/O サーバとの通信性能の検証を行う予定であった。上述の検証運用の際に準備した Parallels 用の仮想 OS イメージを、そのまま RENKEI-VPE に移転することも検討したが、Cloudera による Hadoop Distribution の構築性・運用性に配慮し、Cloudera Manager によって管理が可能な、より可搬性の高い仮想 OS イメージを再構成することとした。

これに際しては、まず、最新の Fedora をハイパーバイザとし、仮想 OS は Ubuntu を改めて採用することを検討した。Fedora は、安定性を多少損なっても積極的に最新のライブラリが導入されている。Ubuntu は、本研究課題代表者が JHPCN で遂行している別の研究課題「計測融合オペレーション (13-MD05)」で採用している ROS (Robot Operating System) のリファレンスプラットフォームであり、次年度以降の研究開発に有用と判断したためである。Hadoop 環境も、最新の CDH (Cloudera's Distribution including Apache Hadoop) を採用し、それぞれ最新の環境による評価の準備を進めた。しかし、RENKEI-VPE は、CentOS 系の仮想 OS ホスティングに特化した実装がされており、期間内に Ubuntu を仮想 OS として起動することが困難と判断されたため、中途、仮想 OS を CentOS へ切り替えた。仮想 OS イメージの準備に時間は取られ

図 5: 時系列データの表示

たものの、上位ソフトウェア層は、Cloudera Manager で管理していたために、移行作業は円滑であった。

この移行作業と同時に、HPCI 運用システムが行っているように、事前評価を行うための試験環境と、実運用を行う本番環境を準備した。試験環境は、実験用の PC サーバ (Xeon E3-1225v2, 4-cores/3.2GHz, メモリ 16GB) を準備し、本番環境は東工大で運用されている RENKEI-VPE 上に準備した。ベース OS としては、RENKEI-VPE での運用性を考慮してホスト OS と仮想 OS ともに CentOS6.5 を採用した。さらに、Hadoop については、継続的な運用性を考慮して、CDH (Cloudera Distribution for Hadoop) 4.6 を採用した。これまでの事前開発では、RPM パッケージによって配布されているモジュールから導入を行ない、手動で運用管理を行っていたが、運用状況の詳細なモニタが可能である Cloudera Manager を導入した。Hadoop の構成情報の管理も Cloudera Manager を介して行ない、プロビジョニング支援が可能なように検証を行った。同時に Gfarm 2.5.8 の導入を行い、ローカルな Gfarm ストレージを整備し、ローカルストレージと HPCI 共用ストレージとの比較、マイグレーションの検証が可能な導入を行った。

これらの基本モジュールがあらかじめ導入された仮想 OS イメージを準備し、DHCP や DNS 構成情報に応じて複数の仮想 OS として起動可能な準備を行った上で、検証環境と本番環境への仮想 OS イメージの展開を行った。ただし、Cloudera Manager の制約から、Hadoop の構成情報は、仮想 OS 起動後に Cloudera Manager から改めて投入する運用とした。また、RENKEI-VPE では、仮想 OS イメージに対して独自の設定投入やスクリプトの導入が必要なため、そのカスタム化を行った仮想 OS イメージを用意して展開している。

まず、試験環境にて、本研究課題で開発した Hadoop/HBase-Gfarm プラグインを導入し、Gfarm ストレージ上に HBase のテーブルを作成し、OpenTSDB によって時系列データを格納し統

計処理が可能であることを検証した。具体的には、6 千万点の時系列データ（14 地点で毎分取得したデータを 2 ヶ月分に相当）をダウンサンプルして表示するのに 20 秒程かかることを確認した。通常想定している監視においては、キャッシュヒットによって 1 秒以内のレスポンスが得られており、実運用にたり得ることを確認した。

(1-2) 運用シナリオに基づく有効性の確認

図 4 のシナリオを検証するために、別の HPCI 課題で資源提供を受けている共用ストレージ領域に格納されているデータセットを用い、試験環境から東工大の RENKEI-VPE 上に構築した本番環境へのマイグレート可能であることを確認した。このマイグレートに際しては、Gfarm ストレージの複製は必要なく、試験環境のマウントポイントを本番環境へそのままマウントしている。だたし、この際、HBase のファイルシステム I/O レイテンシが大きくなると HBase マスターサーバが不整合を検知し予期しない停止する事象を確認している。これを防止するためには、構成情報のチューニングや Hadoop/HBase-Gfarm プラグインの改修が必要であるが、本研究では開発期間の制約でそのチューニングはペンディングとしている。現時点では、データの書き込みを伴う運用においては、ローカルの Gfarm ストレージへデータをダンプ・レストアすることが望ましいと判断している。これには、HBase のダンプ・レストア操作を行うか、もしくは、Gfarm のファイル複製を行う必要がある。

(2) 当初計画の達成状況について

以上の通り、広域に分散した超大規模情報システム構築と評価を行ない、HPCI アカウント集計システムの移行を検証するため疑似環境を構築し、データ投入・複製・解析・表示という実運用に即したシナリオ遂行を通じて、地理的に離れた多拠点間での運用性の検証を行った。

なお、本研究課題で開発した Hadoop/HBase-Gfarm プラグインは GitHub にて公開している（研究成果リスト 5-1 の成果物の公開）。

4. 今後の展望

ここで得られた知見を活かし、JHPCN の別プロジェクトとして研究開発を進めてきた「計測融合オペレーション（13-MD05）」で得られた知見も踏まえ、大阪大学サイバーメディアセンターIT コア棟における計測融合オペレーション実現に際して、時系列データ蓄積のための基盤としてこれらの技術を活用する計画である。

これと同時に、HPCI 認証基盤や HPCI ネットワーク基盤との連携性や相互運用性の検証も同時に行う。特に、認証基盤との連携に際しては、HPCI が採用している二要素認証（Shibboleth と GSI）との連携性を検証し、多拠点での Gfarm によるデータ共有に際して親和性が高く、かつ HPCI 運用に際する個人情報を共有するために適切な認証と機密保持方式を改めて検討している。具体的には、HPCI で運用している電子証明書の操作は、1 年間の有効期間があり、頻繁に発行する必要がないため、Web インターフェイスが適していると考えられる。一方、代理証明書は、利用に先立って都度引き換える必要があり、有効期間が最長 2 週間に設定されているため、頻繁に再発行する必要がある。代理証明書発行をコマンドライン・インターフェイス（CLI）で行う事が要望として挙がっているが、Shibboleth は Web 認証基盤であり、Web 認証を CLI 化するためのブリッジ技術の開発が必要となっている。HPCI と同じように、Shibboleth 認証によって電子証明書発行を行っている米国 CyberInfrastructure プロジェクトでは、Web 認証ポータルである CILogon に対して SAML-EC（Enhanced Client）と呼ばれているブリッジ技術の導入を進めている。

本研究課題では、CILogon 用の SAML-EC 対応クライアントを HPCI 認証ポータル向けに移植する検討を進めている。さらに、大阪大学サイバーメディアセンターでは、スーパーコンピュータ・システムの認証に Kerberos 認証を導入している。昨秋からは、Hadoop 環境の提供に際しても、Kerberos 認証を取り入れている。SAML-EC は、GSS-API にも

実装が進みつつあり、cURL ライブラリによって認証方式をブリッジすることが可能になりつつある。これによって、Hadoop 分散ファイルシステム(Kerbero 認証) と HPCI で運用されている Gfarm (Shibboleth+GSI 認証) による共用ファイルシステムの連携が可能となることを期待しており、そのための要素技術の研究開発を進めている。これらの検証に関しては、大学 ICT 推進協議会(AXIES) 年次大会 HPCテクノロジー企画セッションにて口頭発表を行っている (研究成果リスト 4-1)。

5. 研究成果リスト

(1) 学術論文

(2) 国際会議プロシーディングス

(3) 国際会議発表

(4) 国内会議発表

(4-1) 東田 学, “HPCI 申請支援システムの実装と運用、そして将来像”, 大学 ICT 推進協議会(AXIES) 年次大会 HPC テクノロジー企画セッション, 2013 年 12 月.

(4-2) 東田 学, “広域分散ファイルシステムに基づく「ビッグテーブル」型の超大規模データ処理系の構築と機能および性能評価」および「大規模計算機空気冷却風速場の高解像度解析と適応的クラウドロボット技術による実効的な計測融合オペレーション」”, 学際大規模情報基盤共同利用・共同研究拠点 第 1 回ネットワーク型学際研究シンポジウム, 2014 年 3 月.

(5) その他 (特許, プレス発表, 著書等)

(5-1) 成果物の公開

https://github.com/moyhig/gfarm_hbase