Point-of-Presence 連携による e-サイエンス分散環境

滝澤 真一朗[†1] 松岡 聡[†1,†2] 友石 正彦[†1]
佐藤 仁[†1] 東田 学[†3]

ネットワーク分散した種々の計算機資源を統合して，科学技術の新発見・融合研究領域の開拓を促進する研究手法である e-サイエンスの実行基盤として，我々は RENKEI-PoP と名付けたストレージサーバにて拠点間を接続するネットワーク環境を提案する．RENKEI-PoP は拠点の 1 ゲートウェイサーバとして働き，1) 拠点間の汎用データ転送・共有環境，2) 仮想マシンによる e-サイエンス基盤サービス群のホスティングを提供する．我々は，日本国内 8 拠点に RENKEI-PoP を設置し，SINET の提供する 10Gbps ネットワークで接続した．RENKEI-PoP の配備状況と，基本性能，活用計画について示す．

An E-Science Distributed Environment Composed by Federated Point-of-Presences

Shin'ichiro Takizawa,[†1] Satoshi Matsuoka,[†1,†2]
Masahiko Tomoishi,[†1] Hitoshi Sato[†1] and Manabu Higashida [†3]

As an e-Science infrastructure, we propose a network environment where site resources are federated by a storage server named *RENKEI-PoP*. A RENKEI-PoP works as a gateway server of a site and it provides 1) a general-purpose data transfer/sharing environment and 2) a virtual machine hosting environment that executes e-science infrastructure services. We installed RENKEI-PoPs in eight sites in Japan and connected them to SINET 10Gbps network. We show the RENKEI-PoP system architecture, its performance and outreach activities.

1. はじめに

高速ネットワーク技術や，グリッドによる異なる組織間の資源連携技術の発展により，ネットワーク接続された高性能計算機，大容量ストレージ，データベース，実験装置などの様々な資源を統括的に利用し，科学技術における新発見や融合研究領域などの新たな研究分野の創出を促進する科学技術研究手法である e-サイエンスを実現するための研究開発が行われている．例えば，高エネルギー物理学分野では LHC (Large Hydron Collider) Computing Grid プロジェクト[1)] において，粒子加速器により生成された莫大なデータの共有・処理環境として，gLite をベースとしたグリッドミドルウェアを用い，全世界 170 拠点以上からなる階層構造を持つ環境を整備している．また，多数の組織が持つ様々な地球観測データを参照し，データ解析やシミュレーションを行う GEO (Global Earth Observation) Grid[2)] では，X.509 証明書と公開鍵暗号技術による認証を実現する GSI や，利用者や利用者の認可属性を管理する VOMS などの標準技術を基盤とした GEO Grid SDK を用いたサービス連携を提供している．

既存 e-サイエンスプロジェクトの共通点として，個別のプロジェクト専用にハードウェア・ネットワークからソフトウェアまでの環境を構築している点，プロジェクト内で一貫した相互運用方針を策定している点がある．e-サイエンス基盤としての資源の連携・管理の一元化にはこれらは重要ではあるが，一方で大規模計算機を所有する計算機センターの資源を組み込んで利用することが困難となる問題がある．例えば，東京工業大学 TSUBAME2.0 スーパーコンピュータや東京大学ら T2K オープンスーパーコンピュータ，RIKEN 京コンピュータなど，プロダクション運用を行うシステムにおいては，安定運用に重点が置かれ，強い相互運用方針を策定しているプロジェクトへの資源提供は

†1 東京工業大学
Tokyo Institute of Technology
†2 国立情報学研究所
National Institute of Informatics
†3 大阪大学
Osaka University

困難となる．

この問題を解決するために，我々は RENKEI-PoP と名付けたゲートウェイサーバにて計算資源を持つ拠点間を接続する e-サイエンスネットワーク環境を提案する．RENKEI-PoP は拠点間のデータ共有と，資源連携のためのグリッドサービスのホスティング機能を持ったアプライアンスである．RENKEI-PoP は e-サイエンス資源を提供する拠点に設置され，拠点内資源とは強く結合し，RENKEI-PoP 間ではグリッド認証により連携して，拠点間通信の中継を行う．RENKEI-PoP により，1) 拠点間の汎用データ転送・共有環境の実現，2)e-サイエンス基盤システムを構成する資源連携サービス群の実験・実行環境を仮想マシン群から構築，が達成される．

我々は提案環境を実現すべく，東京工業大学を含む日本国内 8 拠点に RENKEI-PoP を設置し，SINET の提供する 10Gbps ネットワークで接続した．拠点間の高速汎用データ転送・共有環境を実現するために，RENKEI-PoP 間のネットワークチューニングを行い，Gfarm 分散ファイルシステムを構築した．チューニング前後で平均して 2.6 倍通信帯域が向上したことを確認した．また，資源連携サービス群の実行環境を提供するために，分散配置された仮想マシン (以降 VM) 実行サーバ上での VM 実行管理を行うシステムである RENKEI-VPE を開発し，RENKEI-PoP に導入した．VM により資源連携サービス群を実行するため，ソフトウェア環境の再利用や運用・開発用バージョンの管理を行うことができる．RENKEI-VPE により，約 2 分程度で各 RENKEI-PoP に VM を配置できることが確認できた．

本稿は以下のように構成される．2 章では関連研究について紹介する．3 章では，e-サイエンスネットワーク基盤としてのシステム要件を挙げる．4 章では RENKEI-PoP による e-サイエンス分散環境の提案の詳細を述べ，5 章で RENKEI-PoP の実装，6 章で配備状況を述べる．7 章にて配備環境で行った評価をまとめ，8 章では RENKEI-PoP の活用事例を紹介する．9 章で本稿をまとめる．

2. 関 連 研 究

複合領域研究のためのネットワーク基盤として既に存在，あるいは提案されているシステムを紹介する．

DEISA[3] はヨーロッパ 12 拠点の計算機センターを 10Gbps 専用ネットワークで接続した HPC(High Performance Computing) 環境である．globus や UNICORE による任意の計算資源へのジョブ投入，および MC-GPFS によるファイル共有環境を提供し，銀河の構成シミュレーションや気象モデリングと言った応用計算に用いられている．我々の提案でも同様な環境の実現を目指しているが，資源連携を行うサービス群を VM で徹底して管理する点で異なる．

TeraGrid[4] は米国の 11 拠点の計算機センターを 10Gbps 専用ネットワークで接続した分散環境を提供し，そのソフトウェアは Coordinated TeraGrid Software and Services (CTSS) というパッケージ単位で管理されている．コアパッケージ等，全ての TeraGrid 資源に必須のパッケージもあるが，これらは拠点間でバージョンが異なっている．この理由のひとつとして，我々は管理の困難さがあると考えている．基本ソフトウェアの設定がされた VM を再利用することで，我々はこの問題の解決を目指す．また，TeraGrid の資源の一部を用いて，グリッド・クラウドのための分散システム・アプリケーション研究用のテストベッドを提供する FutureGrid プロジェクト[5] が 2009 年 10 月より開始されている．大規模計算機センター資源を用いて，仮想化技術によりシステム・アプリケーションを実行するサイエンスクラウドを提供する点など，我々の提案と共通する．

PlanetLab[6] は全世界 507 拠点 1091 ノードから構成される分散システム開発・実験のためのテストベッド環境である．PlanetLab では資源を slice という仮想単位で利用者に利用権限を与える．この slice の管理や分散システム開発の目的上，PlanetLab の各ノードは，特別な OS・システムソフトウェアを導入する必要があり，また，ファイアウォールのないグローバルネットワークに接続しなければならず，拠点資源の運用方針と衝突する問題がある．

情報爆発プロジェクトの実験評価環境として整備されている InTrigger[7] は全国 17 拠点に設置されたクラスタより構成されている．Intrigger では全ての拠点で Unix アカウントが統合されており，どの拠点のサーバにも同じアカウントでログイン可能である．Intrigger は 1 つの完結した分散システムとして仕様策定・構築されているため，計算機センター等の運用システムの資源を組み込むことは難しい．また，グリッド認証基盤を持たないため，数多く開発されているグリッド認証を用いる資源連携技術を利用できない．

Data Reservoir[8] は高遅延高バンド幅環境での，科学技術研究のための大量データ共有システムとして提案されている．ストレージサーバである Data Reservoir 間でストレージブロックレベルのデータ共有を行うことで，低オーバーヘッドで高効率のネットワーク利用

を実現している．Data Reservoir と我々の RENKEI-PoP はアーキテクチャ面で似ているが，前者はネットワークの効率的利用に主眼がおかれているのに対し，我々は拠点間の資源連携に主眼をおいている．そのため，Data Reservoir では資源連携サービスのホスティング機能はなく，また，拠点資源との連携について具体的な方針が論じられていない．

3. e-サイエンス基盤としての要件

多数の拠点の計算・ストレージ資源を連携する e-サイエンス実証環境を構築するための要件として，以下の 3 点を挙げる．

大規模計算機センター資源と外部資源の連携　利用者のプログラム開発・実行・評価サイクルとして，開発中のデバッグや問題サイズの小さいデータセットを用いた実行には，研究室にある比較的小規模な計算機システムを用い，プログラム完成後の大規模実行には大型計算機センターにあるスーパーコンピュータを用いることがある．この一連のサイクルにおいて，利用者の利便性を向上するには，研究室計算資源から大規模計算機センター計算資源へのプログラム・データセットのアップロード，結果のダウンロードを簡易化するソフトウェア環境，ネットワーク環境が必要となる．このとき計算機システム毎に異なる認証基盤により利用者・データが管理されているため，透過的な連携には計算資源間をまたがる統一認証基盤が求められる．

データ転送・共有環境　e-サイエンスを実現するには多数の拠点の資源をネットワーク接続する必要があるため，利用者の研究室から計算機センターへのデータ入出力に限らず，拠点間でのデータ転送・共有環境の整備が重要となる．従来はプロジェクト個別にデータ転送・共有環境を用意していたが，東京工業大学 TSUBAME2.0 や T2K システムなどのスーパーコンピュータが汎用的な計算資源を提供しているので，それら計算機センター間での汎用的なデータ転送・共有環境を構築することで，利用者の負担削減，利便性向上に繋がる．

既存環境への最小限の変更　上記の環境を構成するにあたり，e-サイエンス資源提供拠点となりうる計算機センターへのソフトウェア・ネットワーク的な構成変更は最小限にとどめる必要がある．計算機センターは課金を伴う高信頼サービス運用を行っており，セキュリティポリシ・システム構成変更が困難だからである．

以上の要件を満たすネットワーク基盤，および，ワークフローシステム等の資源連携サービスを実行するための環境を構築することが我々の目的である．資源連携サービスそのものを提案する研究ではない．

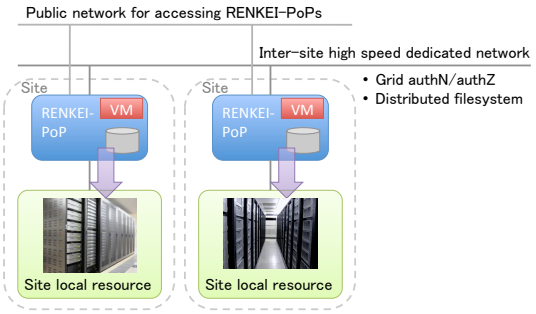

図 1　RENKEI-PoP が構成する e-サイエンス環境
Fig. 1　An e-Science infrastructure enabled by RENKEI-PoP

4. Point-of-Presence 連携による e-サイエンス分散環境

我々は図 1 に示す，Point-of-Presence（以降 PoP）サーバにより接続された計算機センター群からなる e-サイエンス実証環境を提案する．この PoP を RENKEI-PoP と名付ける．RENKEI は REsources liNKage for E-scIence の略であり，本研究を補助する RENKEI プロジェクト[⋆1]の名に由来する．

RENKEI-PoP は 1) 広域インターネット網，2) 拠点間高速ネットワーク，3) 拠点内プライベートネットワークに接続し，拠点計算機資源の 1 ゲートウェイとして働く．個々の拠点には最低 1 つの RENKEI-PoP があり，拠点内計算・ストレージ資源へのアクセスが可能，UNIX ユーザ ID を統一するなど，RENKEI-PoP と拠点内資源は強結合する設計とした．一方で RENKEI-PoP 間では，RENKEI-PoP 毎に同一利用者でもアカウントが異なる場合を想定し，グリッド認証を用いたシングルサインオン認証基盤を提供する．拠点内の資源はプライベートネットワークにあることが多く，異なる拠点の資源間で直接通信ができないこと，また，拠点内資源がグリッド認証基盤を用いるとは限らないため，拠点をまたぐ資源へのアクセスには対象拠点の RENKEI-PoP を介して行う．

RENKEI-PoP には大容量ストレージと VM 実行支援機能が搭載されている．これにより，拠点間でのデータ転送・共有，および，e-サイエンス基盤を構築するための資源連携サービスの VM によるホスティングを提供する．RENKEI-PoP 上でサービスを実行する

⋆1 http://www.e-sciren.org/

ことで，既存環境への変更は不要となる．利用シナリオとしては，RENKEI-PoP 上の VM にグリッドワークフローツールを導入し，それにより実行されるワークフローが処理するデータの移動を RENKEI-PoP 間データ転送・共有機能を用いて行う，ことを想定している．また，e-サイエンス基盤ソフトウェア自体の開発・実験評価環境も RENKEI-PoP 上でホストすることで，開発版・運用版の両立を図り，開発版から運用版へのシームレスな移行実現を目指す．

この PoP による拠点間データ転送・共有と e-サイエンス基盤ソフトウェアの VM ホスティング構想により，3 章の要件を以下のように満たす．

大規模計算機センター資源と外部資源の連携 資源連携は RENKEI-PoP の VM ホスティング環境で実行する e-サイエンス基盤ソフトウェアが行うことになるが，RENKEI-PoP は計算機センター間や計算機センターと個々の研究室間を接続するネットワーク・認証環境を提供する．

データ転送・共有環境 RENKEI-PoP 間ではグリッド認証基盤を用いたデータ転送・共有環境が提供される．拠点間でデータを送受信する際には，異なる拠点の資源同士では直接的には通信できないことを仮定し，RENKEI-PoP が中継サーバとなる転送方式を採用する．具体的には，拠点のストレージから拠点内 RENKEI-PoP にデータを転送し，RENKEI-PoP 間のデータ共有・転送環境を用いて，転送先拠点の RENKEI-PoP からその拠点のストレージに転送する．sshfs 等で RENKEI-PoP をマウントしてファイルコピーを行うことで，転送を意識せず，透過的な転送も可能となる．中継のためのデータの一時保存領域として，RENKEI-PoP には大容量ストレージを搭載している．

RENKEI-PoP のデータ転送・共有環境に個々の研究室の利用者がデータを入出力する方法として，最寄りの RENKEI-PoP に，1)scp 等の一般的なデータ転送技術を用いて転送する方法，2)RENKEI-PoP 間と同一のグリッド認証を研究室サーバで設定し，グリッド認証を行うデータ転送技術を用いて転送する方法，などがある．

既存環境への最小限の変更 RENKEI-PoP 構想を実現するにあたり，各拠点で行うべきことは，1)RENKEI-PoP を拠点資源にアクセス可能なネットワーク及び広域ネットワーク網に接続，2)RENKEI-PoP との通信用ファイアウォール設定の変更，だけである．ただし，RENKEI-PoP の VM ホスティング環境で実行する e-サイエンス基盤ソフトウェアの実装・種類次第では，拠点計算資源への変更も必要となりうる．例えば，RENKEI-PoP 上でジョブスケジューラを動かし，そのジョブ実行サーバとして拠点計算資源を用いる場合には，拠点計算資源にスケジューラ実行エンジンのインストールが必要になる場合がある．

表 1 東京工業大学，北海道大学の RENKEI-PoP 仕様
Table 1 Specification of RENKEI-PoPs in Tokyo Tech and Hokkaido Univ.

CPU	Intel Xeon W3565 (3.2GHz, 4core)
Memory	DDR3 24GB (4GB x6)
Network Card	Myricom 10Gb Ethernet x1 (Equipped on Tokyo Tech)
	Intel 1Gbps Ethernet x2
Storage	HDD Raid 32TB (SATA 2TB x16)

5. RENKEI-PoP の実装

RENKEI-PoP の実装について，ハードウェア，システムソフトウェア，データ転送・共有環境，VM ホスティング環境についてそれぞれ述べる．

5.1 ハードウェア

RENKEI-PoP のハードウェアは以下の基本方針に従い設計した．

- 10Gbps 通信を維持するために十分な CPU 性能
- VM 実行支援機能を搭載
- 複数 VM を実行するのに十分なメモリ量
- 広域 10Gbps ネットワークを活用できるネットワークインターフェース
- 広域データ転送・共有用，複数 VM の OS イメージ保存用に十分な量と性能のストレージ

既に複数台の RENKEI-PoP を構築済みであり，構築時期により若干構成が異なるが，標準的な構成を表 1 に示す．32TB の HDD を搭載するモデルでは，16 台の HDD で RAID5 を構成しており，iozone によるバッファキャッシュを用いない read/write ベンチマークではそれぞれ 720MBps(Byte per second), 690MBps を記録している．

5.2 システムソフトウェア

RENKEI-PoP では Red Hat Enterprise Linux 互換である CentOS 5.5 を OS として用いている．Globus Toolkit[9] が導入され，GSI 認証が設定されている．GSI 認証は主に RENKEI-PoP 間，RENKEI-PoP と外部資源との連携に利用され，RENKEI-PoP と拠点内部資源との連携には Unix 認証を用いる．RENKEI-PoP の監視システムとして，CPU，メモリ等のサーバ資源利用量の推移監視に Munin[10] を，

RENKEI-PoP の死活・サービスの動作整合性監視に INCA[11] を，リアルタイム資源利用量監視に VGXP[12] を導入している．監視結果は RENKEI-PoP 群を管理する管理サーバ (以降，RENKEI-PoP 管理サーバ) が提供する Web インターフェースから確認できる．

5.3　データ転送・共有環境

RENKEI-PoP 間でのデータ転送には，標準的な scp に加えて，gsiscp，gridftp 等のグリッド認証基盤を用いるデータ転送手段を提供する．また RENKEI-PoP 上では Gfarm 2.3.2[13] による分散ファイルシステムが構築されている．Gfarm でも GSI による認証が利用可能である．各 RENKEI-PoP には Gfarm のストレージサーバ機能を，RENKEI-PoP 管理サーバには Gfarm のメタデータサーバ機能を導入した．

5.4　VM ホスティング環境

RENKEI-PoP 上での VM ホスティングには，libvirt による qemu/kvm VM 実行環境を提供している．これを用いて RENKEI-PoP 上に個別に VM を実行することは可能であるが，広域分散環境上での VM 用 OS イメージの管理，VM のライフサイクル管理を行うシステムである RENKEI-VPE (Virtual Private Environment) を開発し，RENKEI-PoP に導入した．RENKEI-VPE は VM ネットワークの設定，起動，アクセスまでを提供し，VM への個別ソフトウェア設定は利用者が行う．

VM のライフサイクルを管理する既存システムとして，IaaS(Infrastructure as a Service) クラウド管理システムである Eucalyptus[14] や Nimbus[15]，OpenStack[16] 等があるが，これらを採用せず新規に開発を行った．RENKEI-PoP における VM ホスティングの目的は e-サイエンス基盤構築のための資源連携サービスを実行することであり，このためには拠点内の各種計算・ストレージ資源が接続されたネットワーク環境を変更せずに，VM を接続できる必要がある．これら既存システムでは外部接続用として登録できるネットワークセグメント数に制限があることや，ネットワークセグメンドの柔軟な追加・削除が行えないため，不採用とした．

RENKEI-VPE のアーキテクチャを図 2 に示す．RENKEI-VPE は 1 台の管理サーバ，1 台以上のホスティングサーバ，管理サーバとホスティングサーバからアクセスできる 1 つの OS イメージストレージからなる．それぞれの役割は以下の通りである．

管理サーバ　利用者情報管理と，利用者からの VM・OS イメージ操作リクエストの受理，VM 実行・イメージ操作を行う．VM 実行要求に対しては，利用者が指定したサイトのホスティングサーバを 1 つ選択し，そこで実行する．OS イメージ操作に対しては，操作内容に対して，OS イメージストレージへのイメージの格納・取得・削除，属性情報の更新などを行う．RENKEI-VPE を利用する際には，管理サーバに ssh や gsissh でログインし，コマンドラインで操作を行う．RENKEI-VPE は VM・OS イメージ操作用の xmlrpc を提供しているため，外部サーバからの操作も可能である．

RENKEI-PoP 管理サーバが本機能を提供する．

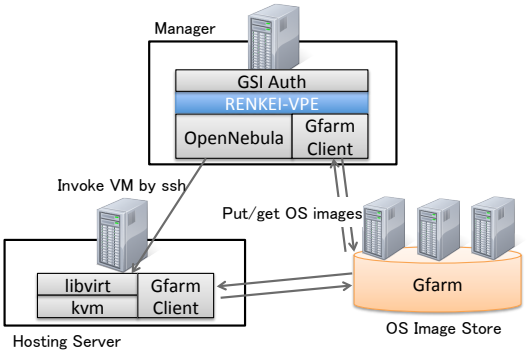

図 2　RENKEI-VPE の構成図
Fig. 2　Architecture of RENKEI-VPE

ホスティングサーバ　分散配置された VM 実行機能を持つサーバ，あるいはクラスタ構成の計算機資源である．管理サービスからの VM 実行要求を受け取った後に，OS イメージストレージから OS イメージを取得し，VM を実行する．利用者の VM を接続するための各種ネットワーク回線が VLAN 等で引き込む必要があるが，一度引き込めば任意の VM を接続できる．

各 RENKEI-PoP が本機能を提供する．

OS イメージストレージ　VM が使用する OS イメージが格納される，全ホスティングサーバからアクセスされるストレージである．OS イメージにはシステム管理者が用意したものだけでなく，利用者による登録も可能であり，利用者による世代・役割管理が可能である．

RENKEI-PoP 上に構築した Gfarm による共有ファイルシステムが本機能を提供する．

RENKEI-VPE での VM ライフサイクルは図 3 に示すとおりである．

(1)　管理サーバに VM 起動を要求
(2)　管理サーバは利用者の要求に基づき，ホスティングサーバを選択
(3)　選択されたホスティングサーバは OS イメージ

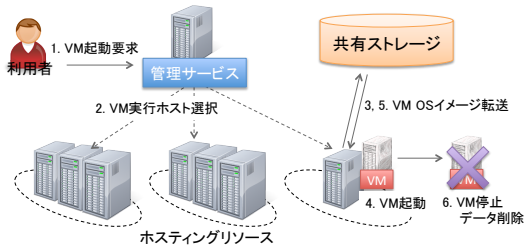

図 3　RENKEI-VPE の VM 処理の流れ
Fig. 3　Flow Chart of VM Life Cycle on RENKEI-VPE

ストレージより，利用者が指定した OS イメージを取得

(4)　ホスティングサーバ上で，利用者のアクセスキーや指定した OS イメージを用いて VM を起動し，指定されたネットワークに VM を接続

(5)　利用者は任意のタイミングで VM の OS イメージを OS イメージストレージに保存可能

(6)　利用者が VM 停止を管理サーバに要求することで，VM は停止され，ホスティングサーバ上から関連データが消去される

RENKEI-VPE の実装の詳細を述べる．RENKEI-VPE では，ホスティングサーバ上の VM 管理には IaaS クラウド管理システム OpenNebula 2.0[17] を用い，OS イメージストレージには Gfarm を用いる．RENKEI-VPE 自体が xmlrpc サーバ/クライアント実装となっており，VM 管理要求には RENKEI-VPE で VM 定義ファイル作成，OS イメージ，ネットワークセグメント選択等の前処理を行った後に OpenNebula の xmlrpc を呼び出し VM 管理を依頼する．VM を起動するホスティングサーバの選択ポリシは次のように実装されている．1) 利用者が VM を起動する拠点を指定する，2) 拠点内に複数の RENKEI-PoP が存在する場合，負荷に応じてラウンドロビンで VM を実行する．RENKEI-VPE からホスティングサーバへの VM 管理要求と，利用者の ssh 公開鍵や VM 設定情報などの VM 個別設定は ssh にて送られる．RENKEI-VPE において，OpenNebula に対して拡張した主な機能は以下になる．

- 利用者毎の利用可能拠点，起動可能 VM 数制御の資源制限
- VM 構成の共有，再利用
- VM 定義の自動生成，マルチホーム対応 VM の構築
- 利用者への IP アドレスの事前静的割当
- Gfarm の統合

OS イメージストレージに用いる Gfarm には，RENKEI-PoP 上に拠点間データ転送・共有を目的に構築したものを用いる．管理サーバ，ホスティングサーバ共に Gfarm クライアント機能を備えており，内部的に Gfarm へのデータ登録コマンド (gfreg)，データ取得コマンド (gfexport) を実行する．

表 2　RENKEI-PoP 設置拠点 (導入順)
Table 2　Locations where RENKEI-PoPs are installed (installed order)

Location	Host Name
東京工業大学	titech1, titech2
大阪大学	osaka
国立情報学研究所	nii1, nii2
高エネルギー加速器研究機構	kek
名古屋大学	nagoya
筑波大学	tkb
東北大学	thk
北海道大学	hkd

6. RENKEI-PoP の配備展開状況

RENKEI-PoP は 2011 年 7 月現在，表 2 からなる，日本国内 8 拠点に設置されている．東京工業大学と国立情報学研究所に 2 台あることを除き，各拠点 1 台が設置されている．RENKEI-PoP の導入は 2009 年より進めており，導入時期により仕様が若干異なる．具体的には nii2，kek には容量 256GB(raw) の SSD RAID が搭載されているに対して，それ以外には容量 32TB(raw) の HDD RAID が搭載されている．メモリ容量も titech1，titech2，nii1，nii2，hkd には 24GB 搭載されているが，それら以外は 12GB である．また，nii2，nagoya，thk，hkd は 10Gbps ネットワークには対応していない．

RENKEI-PoP のネットワーク接続と Gfarm の構成を図 4 に示す．RENKEI-PoP 群へのソフトウェア，アカウント管理を行う RENKEI-PoP 管理サーバは東京工業大学にある．各 RENKEI-PoP は RENKEI-PoP 間で最大 10Gbps の高速データ転送を行うために SINET 4 L3VPN に接続し，一部の RENKEI-PoP は外部からのアクセスのために広域インターネット網に接続している．東京工業大学の RENKEI-PoP では以前はスーパーコンピュータ TSUBAME とネットワーク・アカウントを統合した運用を行っていたが，2010 年 11 月に TSUBAME2.0 へとシステムが更新された際にその運用が一時停止した．ただ，現在でも広域インターネット網側を介した RENKEI-PoP と TSUBAME2.0 間データ共有，SINET 4 L3VPN を介した TSUBAME2.0 グリッドサービスから RENKEI-PoP ストレージへのアクセスは可能である．現在，TSUBAME2.0 の内部ネットワークと RENKEI-PoP の接

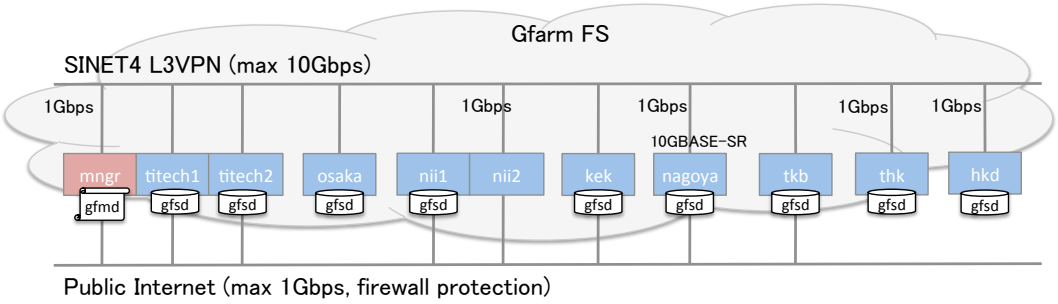

図 4　RENKEI-PoP のネットワーク接続と Gfarm 構成
Fig. 4　Network Connection and Gfarm Structure of RENKEI-PoPs

続，アカウント統一に向けた作業を行っている．これが達成すると，RENKEI-PoP と TSUBAME2.0 間でのデータ転送の高速化が実現できる．

RENKEI-PoP 上の Gfarm の構成は次の通りである．RENKEI-PoP 管理サーバでは Gfarm メタデータサーバを実行し，nii2 を除く各 RENKEI-PoP が Gfarm ファイルサーバとなり，最大 215TB の Gfarm ストレージが構成されている．Gfarm ファイルサーバでは，gridftp による Gfarm 領域へのデータ書き込みをサポートする．全ての RENKEI-PoP は Gfarm クライアントの機能を持つ．

RENKEI-VPE の構成を図 5 に示す．RENKEI-PoP 管理サーバを RENKEI-VPE 管理サーバとして構築し，RENKEI-PoP を VM ホスティング用サーバとして構築している．VM への割当用 IP アドレス数の制約により，現在では東京工業大学，国立情報学研究所，北海道大学の RENKEI-PoP でのみ RENKEI-VPE は有効になっている．OS イメージストレージには RENKEI-PoP 上に構成した Gfarm を用いた．

7. 評　　価

評価として，ネットワーク帯域，Gfarm によるデータ転送性能，RENKEI-VPE による VM 起動時間の評価を行った．RENKEI-PoP は既に各種プロジェクト支援のために用いられていること，ネットワーク回線が占有線では無いため，他への影響を考慮し，計測は夜間・休日に行った．

7.1　ネットワーク性能

SINET 4 L3VPN 10Gbps に接続されている 6 台の RENKEI-PoP 間での帯域を測定した．広帯域・高遅延ネットワーク上での通信となるため，測定にあたり，各種ネットワークパラメータを調整した．CentOS 標準パラメータから変更した，RENKEI-PoP のネットワークパラメータを表 3 に示す．転送バッファサイズ（Maximum buffer size）は各 RENKEI-PoP から他 RENEI-PoP への RTT を測定し，その最大値と理論バンド幅 10Gbps と掛け合わせた帯域遅延積とした．その他のパラメータは実計測した値，および RENKEI-PoP 設置環境の制約を基に設定している．例えば MTU は拠点毎の SINET 接続スイッチにて 1500 で設定されていたため，ジャンボフレーム対応は不可能であった．

iperf による帯域の計測結果を表 4 に示す．表中括弧内の数値はネットワークパラメータ調整前の値である．調整前後で，例えば nii1 → osaka で約 4.3 倍，tkb → kek で約 22 倍の性能向上があり，平均で 2.6 倍の性能向上が確認できた．拠点間通信では，最大で nii1 → titech1 の 960Mbps を示してはいるが，非対称性が目立つ．特に titech1 ↔ tkb では 3 倍近い差がある．また titech1, titech2 → tkb の様に調整の結果性能が低下した箇所もあり，これらは現在改善に向けて調査中である．

7.2　Gfarm データ転送性能

Disk-to-Disk のデータ転送性能を確認するために，Montage[18] の約 14GB のデータセットを RENKEI-PoP 間で Gfarm を用いて転送した際の性能を示す．測定には gfrep コマンドを用い，titech2, osaka, nii1, kek, tkb から titech1 へデータをコピーした際の時間を計測し，スループットを計算した．なお，Montage のデータセットはファイル数 6756 個であり，最大サイズ 27MB，最小サイズ 292B，平均サイズ 2MB，標準偏差 400KB であった．平均的にファイルサイズが拠点間通信時の帯域遅延積よりも小さく，またファイル毎に遠隔地にあるメタデータサーバへのアクセスを行うため，転送効率を上げるために 1 つのファイルにまとめて（非圧縮）転送した．実際，個別にファイルコピーを行ったところ，nii1 → titech1 において 112MBps とネットワーク性能 960MBps の 12%程度

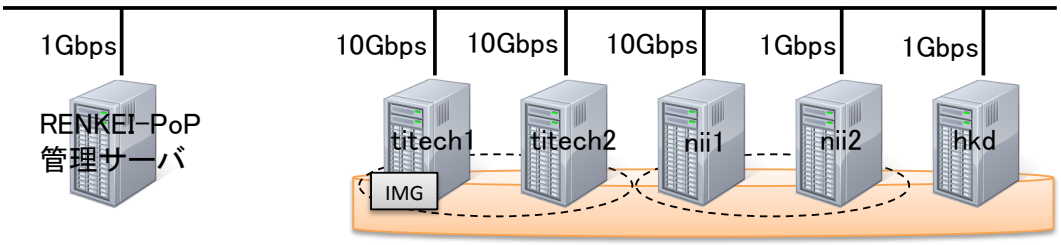

図 5　RENKEI-PoP 上の RENKEI-VPE 構成
Fig. 5　RENKEI-VPE Structure on RENKEI-PoPs

表 3　RENKEI-PoP のネットワークパラメータ
Table 3　Network parameters for RENKEI-PoPs

	titech1	titech2	osaka	nii1	kek	tkb
MTU	1500B	←	←	←	←	←
Maximum buffer size	12MB	←	18MB	←	←	←
Interface queue length	10000	←	←	←	←	←
Packet queue length	30000	←	←	←	←	←
TCP segmentation offloading	on	←	off	←	on	off
Adaptive interrupt coalescing (rx)	on	←	←	←	off	on
disable caching route metrics	on	←	←	←	←	←

表 4　RENKEI-PoP 間のバンド幅 (単位は MBps)
Table 4　bandwidth between RENKEI-PoPs (MBps)

From\To	titech1	titech2	osaka	nii1	kek	tkb
titech1		1120(1120)	415(162)	559(516)	428(319)	173(322)
titech2	1120(1120)		417(161)	592(513)	437(331)	210(362)
osaka	526(187)	529(186)		416(167)	327(126)	331(99)
nii1	960(509)	956(508)	474(110)		443(236)	264(215)
kek	454(365)	464(367)	289(61)	374(263)		747(736)
tkb	493(387)	512(381)	363(101)	521(264)	627(28)	

の性能に留まった．

結果を図 6 に示す．凡例 Gfarm は Gfarm によるデータ転送性能，Network は表 4 の各 RENKEI-PoP から titech1 へのネットワークバンド幅，Storage は 5.1 節で示した RENKEI-PoP ストレージ Read 性能 720MBps を表す．今回転送した Montage データセットは titech1 のバッファキャッシュに収まりきるサイズであるため，Gfarm による転送性能は 1) 送信元 RENKEI-PoP のストレージ Read 性能，2)RENKEI-PoP 間の通信帯域，のいずれかに律速される．図より (1) による性能律速を受けるのは titech2，nii1，(2) による律速を受けるのは osaka，kek，tkb とわかる．

nii1 と titech2 の結果の差分は，Gfarm メタデータアクセス性能が影響していると考えている．同一拠点にあるため titech2 から Gfarm メタデータサーバへは平均 0.16ms の RTT でアクセス可能であるが，nii1 からは平均 4.27ms 要することが確認できている．osaka，kek，tkb → titech1 の場合はネットワークアクセスが Gfarm 性能を律速するため，Gfarm 性能は

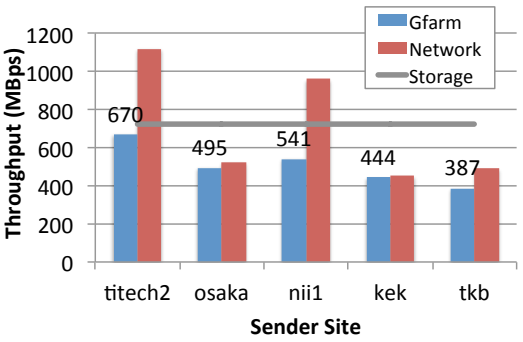

図 6　titech1 への Gfarm データコピー性能
Fig. 6　Data copy performance to *titech1* by Gfarm

Network 性能との差分が小さい．tkb の場合は差分が大きいが，これは titech1 ↔ tkb のバンド幅の非対称性も原因の 1 つであると考えている．

7.3　RENKEI-VPE による VM 起動時間

titech1 上の Gfarm 領域上に格納した OS イメージを用いて，titech1，titech2，nii1，nii2，hkd の各

RENKEI-PoP 上で 1CPU，1GB メモリを持つ VM を起動するのに要する時間を測定した．VM 用 OS イメージには QCOW2 フォーマットされた物理サイズ 653MB の最小構成の CentOS5.5 イメージを用いた．

表 5 に結果を示す．表中の単位は秒であり，「Time to Boot」は VM 起動要求受信から VM 起動完了 (VM で全 init スクリプト実行完了) までに要した時間である．「Time to Prepare」は RENKEI-PoP 上での VM 起動準備に要する時間であり，管理サーバや Gfarm からのファイル転送時間，及び，VM 用 Swap ファイルの作成時間を含む．「Time from Prepared to Boot」は RENKEI-PoP 上での VM 起動完了までの時間であり，「Time to Prepare」との和が「Time to Boot」となる．「Time for gfexport」は Gfarm から RENKEI-PoP 上に OS イメージをコピーする際の時間であり，「Time to Prepare」に含まれる．

VM 配置の性能は RENKEI-PoP 上での VM 起動準備に要する時間に影響して変化することが確認できる．VM 起動準備で行う作業の内，大容量のデータ転送を行う gfexport コマンドによる OS イメージの取得は，表よりネットワーク帯域，地理的距離に応じて RENKEI-PoP 毎に大きく異なることがわかる．その他，RENKEI-PoP 毎の VM 起動準備に要する時間の違いを生じる要素として，RENKEI-PoP 管理サーバからの数 KByte のデータ転送が考えられるが，管理サーバでは各種監視サービスや Web サービスも提供しているため，これらのサービスによるジッタの影響により，差として現れづらいと思われる．運用時にはアクセス性能向上，および，データ保護を考慮して，OS イメージは複数の複製を作成して管理するため，短距離でのデータ転送が中心となり，2 分程度の起動時間となると見込んでいる．

8. RENKEI-PoP 活用計画

ゲノム支援プロジェクト[⋆1]にて，ゲノムデータの転送基盤として RENKEI-PoP を利用すべく配備を開始した．具体的には，静岡県三島市にある国立遺伝学研究所に RENKEI-PoP の設置が完了し，現在国立情報学研究所で生成されたゲノムデータを東京工業大学 TSUBAME2.0 に転送し，解析するための環境を整備中である．今後，東京大学柏キャンパスへの設置も計画している．

名古屋大学太陽地球環境研究所では，RENKEI-PoP データ共有環境上に格納したデータを用いた地球磁気圏の解析を行っている．また，名古屋大学の流体力学研究室と筑波大学の統計力学研究室間で計画されている乱流データ解析の研究プロジェクトにて 2 拠点間のデータ共有の手段の 1 つとして RENKEI-PoP の活用を検討している．

本研究が補助を受けている RENKEI プロジェクトでは，ジョブ実行，分散ファイルシステム，データベース連携等の e-サイエンス基盤ソフトウェアを開発している．これらソフトウェアの実証評価環境として RENKEI-PoP が利用されている．具体的には国立情報学研究所の RENKEI-PoP 上で NAREGI グリッドミドルウェア[19] 管理サービス群が VM により実行されており，この NAREGI グリッドから東工大 TSUBAME2.0 や国立情報学研究所で管理しているクラスタ群へのジョブ投入が実施されている．また，このグリッドと研究室の計算資源を連携させるためのワークフローツールや，グリッド間連携システムなど RENKEI プロジェクトの各種成果物の配備が始まっている．

さらに，平成 24 年度秋からの運用が計画されている HPCI[⋆2]における，スーパーコンピュータ間共有ストレージや，分散環境ホスティングサービスの設計に RENKEI-PoP は大きく影響している．

9. ま と め

e-サイエンス基盤環境として，我々は Point-of-Presence により計算資源が連携された拠点間ネットワーク環境を提案した．その Point-of-Presence は GSI 認証をサポートする VM 実行支援を持ったストレージサーバであり，RENKEI-PoP と名付けた．RENKEI-PoP は拠点の 1 ゲートウェイサーバとして働き，拠点間の高速データ転送・共有を支援すると共に，e-サイエンス基盤ソフトウェア群を VM としてその上で実行し，RENKEI-PoP 設置拠点間での資源連携，利用者の研究室内資源との連携を実現する e-サイエンスネットワーク環境である．拠点間データ共有には Gfarm を用い，VM ホスティングには OpenNebula や libvirt, kvm 等を基盤に開発した RENKEI-VPE を用いた．

今後の計画として，e-サイエンス基盤としての活用のために，TSUBAME2.0 スーパーコンピュータとのシステム統合を実施する．特に RENKEI-PoP による広域データ共有環境と，並列ファイルシステムやテープライブラリ等の拠点内ストレージ資源と連携した透過的かつ階層的なデータ管理の実現を目指す．また，

⋆1 http://www.genome-sci.jp/

⋆2 http://hpcic.riken.jp/

表 5 RENKEI-PoP 上での RENKEI-VPE による VM 起動時間 (単位は秒)
Table 5 Time for VM creation on RENKEI-PoPs using RENKEI-VPE(second)

RENKEI-PoP	Time to Boot	Time to Prepare	Time from Prepared to Boot	Time for gfexport
titech1	116	19	97	0.549
titech2	111	14	97	1.47
nii1	118	21	97	7.58
nii2	141	44	97	23.2
hkd	164	68	96	60.3

ネットワーク性能評価の結果，通信の非対称性があるため，これらを解析すると共に，さらなる広帯域を実現するための調整を行うことを予定している．

謝辞 本研究の一部は，文部科学省の科学技術試験研究委託事業による委託業務：次世代ＩＴ基盤構築のための研究開発「e-サイエンス実現のためのシステム統合・連携ソフトウェアの研究開発」の補助による．

参 考 文 献

1) Worldwide LHC Computing Grid: `http://lcg.web.cern.ch/LCG/`.
2) 田中良夫，小島　功，山本直孝，横山昌平，谷村勇輔，関口智嗣：GEO Grid: 地球観測グリッドの設計と実装，情報処理学会研究報告 2007-HPC-112，pp.7–12 (2007).
3) Riedel, M., Memon, A., Memon, M., Mallmann, D., Streit, A., Wolf, F., Lippert, T., Venturi, V., Andreetto, P., Marzolla, M., Ferraro, A., Ghiselli, A., Hedman, F., Shah, Z.A., Salzemann, J., Costa, A.D., Bloch, V., Breton, V., Kasam, V., Hofmann-Apitius, M., Snelling, D., vande Berghe, S., Li, V., Brewer, S., Dunlop, A. and Silva, N.D.: Improving e-Science with Interoperability of the e-Infrastructures EGEE and DEISA, *Proceedings of the MIPRO* (2008).
4) Beckman, P.H.: Building the TeraGrid, *The Royal Society*, Vol.363, No.1833, pp.1715–1728 (2005).
5) Future Grid: `http://futuregrid.org/`.
6) Chun, B., Culler, D., Roscoe, T., Bavier, A., Peterson, L., Wawrzoniak, M. and Bowman, M.: PlanetLab: an overlay testbed for broad-coverage services, *ACM SIGCOMM Computer Communication Review*, Vol.33, No.3, pp.3–12 (2003).
7) 斎藤秀雄，鴨志田良和，澤井省吾，弘中　健，高橋　慧，関谷岳史，頓　楠，柴田剛志，横山大作，田浦健次朗：InTrigger：柔軟な構成変更を考慮した多拠点にわたる分散計算機環境，情報処理学会研究報告 2007-HPC-111，pp.237–242 (2007).
8) Hiraki, K., Inaba, M., Tamatsukuri, J., Kurusu, R., Ikuta, Y., Koga, H. and Zinzaki, A.: Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research, *Conference on High Performance Networking and Computing, Proceedings of the 2002 ACM/IEEE conference on Supercomputing*, pp.1–9 (2002).
9) Foster, I.: Globus Toolkit Version 4: Software for Service-Oriented Systems, *IFIP International Conference on Network and Parallel Computing*, pp.2–13 (2006).
10) Munin: `http://munin-monitoring.org/`.
11) Smallen, S., Ericson, K., Hayes, J. and Olschanowsky, C.: User-level Grid Monitoring with Inca 2, *High Performance Distributed Computing, Proceedings of the 2007 workshop on Grid monitoring*, pp.29 – 38 (2007).
12) 鴨志田良和，田浦健次朗，近山　隆：多拠点に渡る分散計算機環境を効率的にモニタリングするための情報収集と表示，電子情報通信学会技術研究報告. CPSY，pp.7–12 (2007).
13) Tatebe, O., Hiraga, K. and Soda, N.: Gfarm Grid File System, *New Generation Computing*, Vol.28, No.3, pp.1–6 (2010).
14) Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L. and Zagorodnov, D.: The Eucalyptus Open-source Cloud-computing System, *9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, pp.124–131 (2009).
15) Globus Nimbus Homepage: `http://www.nimbusproject.org/` (2010).
16) OpenStack Open Source Cloud Computing Software: `http://www.openstack.org/` (2011).
17) Sotomayor, B., Montero, R.S., Llorente, I.M. and Foster, I.: Virtual Infrastructure Management in Private and Hybrid Clouds, *IEEE Internet Computing*, Vol. 13, No. 5, pp. 14–22 (2009).
18) Montage - Image Mosaic Software for Astronomers: `http://montage.ipac.caltech.edu/`.
19) Matsuoka, S., Shimojo, S., Aoyagi, M., Sekiguchi, S., Usami, H. and Miura, K.: Japanese Computational Grid Research Project : NAREGI, *IEEE*, Vol. 93, No. 3, pp. 522–533 (2005).