Vol.2010-HPC-126 No.31
2010/8/4

# e-サイエンス基盤としての計算機センター POP(Point-of-Presence) 連携

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

ネットワーク分散した種々のリソースを統合して，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-サイエンスが注目されている．e-サイエンス実現基盤として，我々は RENKEI-PoP と名付けた Point-of-Presence 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 e-サイエンスリソースを提供する拠点に設置され，拠点内リソースとは強く結合し，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 は，1)e-サイエンス基盤システムを構成するサービス群の実行環境，および分散システム開発・評価環境を仮想マシンで実現する仮想ホスティングと，2) 拠点間の汎用データ転送・共有環境を提供する．我々は，日本国内 7 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．現状のシステム構成，ネットワーク，ストレージアクセスの評価結果を示す．

# POP(Point-of-Presence) Linkage between Computer Centers as an E-Science Infrastructure

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

As an e-Science infrastructure, We propose a network environment where site resources are connected by a point-of-presence named *RENKEI-PoP*. RENKEI-PoPs are located in sites that provide resources for e-Science, are integrated with site local resources, and relay communications between sites by cooperating with each other using a grid security infrastructure. RENKEI-PoP provides 1) a virtual hosting environment that supports running and developing e-science infrastructure services and 2)a general-purpose data transfer/sharing environment. We installed RENKEI-PoPs in seven sites in Japan and connected them to SINET 10Gbps network. We propose the current RENKEI-PoP system and show its performance of network and storage access.

## 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リソース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リソースを統括的に利用し，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，gsi や 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リソースの連携・管理の一元化にはこれらは重要ではあるが，一方で以下の問題もある．

- 低予算・小規模プロジェクト用の環境準備が困難
- 既存の大規模計算機センターリソースの利用が困難

特に後者は，東京工業大学 TSUBAME 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，ひいては 2011 年稼働予定の RIKEN 次世代スーパーコンピュータなど，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リソース提供は困難となる．我々は e-サイエンス

†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

への小規模プロジェクトの参入障壁，大規模計算機・ストレージリソースの提供障壁が残っていると考えている．

この問題を解決するために，我々は RENKEI-PoP と名付けた Point-of-Presence にて拠点間を接続する e-サイエンスネットワーク環境を提案する．RENKEI-PoP は e-サイエンスリソースを提供する拠点に設置され，拠点内リソースとは強く結合し，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 が提供するサービスには，1)e-サイエンス基盤システムを構成するサービス群の実行環境，および分散システム開発・評価環境を仮想マシンで実現する仮想ホスティングと，2) 拠点間の汎用データ転送・共有環境がある．ソフトウェアの対応も求められるが，RENKEI-PoP の仮想ホスティング上で拠点リソースへの変更が最小限で済むワークフロー，データ管理用システム等を動作させ，それらが利用するデータを汎用データ転送・共有環境で管理する．

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 7 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 は仮想マシン実行支援を持つ，大容量ストレージサーバであり，理論上 1GBps の Disk-to-Disk 転送性能が達成できるように設計してある．本研究では，現状のシステム構成，ネットワーク，ストレージアクセスの評価結果を示す．

## 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 による任意の計算リソースへのジョブ投入，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．プロダクション運用を行うシステムであり，この環境を用いて分散システムの研究開発，評価を行うものではない．

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 リソース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．また，TeraGrid のリソースの一部を用いて，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リソースを用いて，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 ではリソースを slice という仮想単位でユーザ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されていなければならない，という制約がある．

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．タスク実行にはコマンドを直接実行する方法と，バッチシステムを用いた方法が提供されており，Gfarm による拠点間のファイル共有が提供されている．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リソース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リソース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 Reservor 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，我々は拠点間のリソース連携に主眼をおいている．そのため，Data Reservor ではリソース連携サービスのホスティング機能はなく，また，拠点リソースとの連携について具体的な方針が論じられていない．

## 3. RENKEI-PoP 構想

### 3.1 e-サイエンス基盤としての要件

多数の拠点の計算・ストレージリソースを連携する必要性がありうる 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リソース群と研究室リソースのシームレスな連携**

ユーザのプログラム開発・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，ユーザ利便性を向上するには，研究室計算リソースから大規模計算機センター計算リソースへのプログラム・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．

また，ワークフローを構成するタスクが特定のアーキテクチャを前提としたプログラムを実行する場合，あるいは特定のソフトウェアライセンスを必要とする場合，個々のタスクを異なる計算機センターで実行し，その結果を集約する必要がある．各計算機センターで実行されるタスク，および利用されるファイルは計算機センター毎に異なる認証基盤により管理されるため，タスク実行，データの統合的かつ容易な管理を実現するためには，計算機センター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リソースをネットワーク接続する必要があるため，ユーザ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，東工大 TSUBAME 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リソース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，ユーザの負担削減，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-サイエンスリソース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サービス運用を行っており，システム構成変更が困難であり，それが原因で e-サイエンス普及に歯止めをかけられないからである．

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などのリソース連携サービスを実行するためのホスティング環境を構築することが我々の目的である．

### 3.2 PoP 連携による e-サイエンス実証環境

我々は図 1 に示す，Point-of-Presence （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，本研究を補助するプロジェクト名（RENKEI プロジェクト）に由来する．

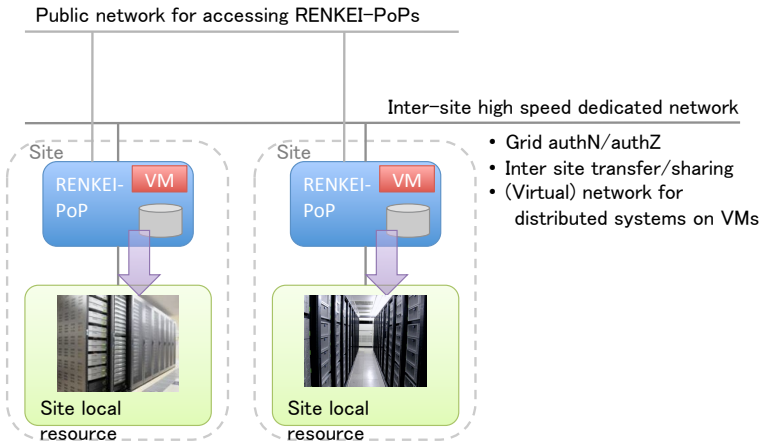



**図 1** RENKEI-PoP が構成する e-サイエンス環境
Fig. 1 An e-Science infrastructure enabled by RENKEI-PoP

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，拠点計算機センターの 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・ストレージリソースへのアクセスが可能，UNIX ユーザ ID を統一するなど，RENKEI-PoP と拠点内リソースの結合は比較的強い．一方で，RENKEI-PoP 間ではグリッド認証基盤を用いた認証認可を行う．拠点内のリソースはプライベートネットワークにあることが多く，拠点間リソース同士の直接的な通信ができないこと，また，拠点内リソースがグリッド認証基盤を用いるとは限らないため，拠点をまたぐリソースへのアクセスには対象拠点の RENKEI-PoP を介して行うことになる．

RENKEI-PoP の共通仕様を表 1 にまとめる．RENKEI-PoP はいわば，10Gbps 高速ネットワーク対応，仮想マシン実行支援を備えた大容量ストレージサーバである．RENKEI-PoP は，その上の仮想マシンに各種 e-サイエンス基盤ソフトウェアを導入し，アプリケーション実行環境を提供すると共に，アプリケーション実行に必要なデータの転送・共有環境を提供する．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．

**表 1** RENKEI-PoP 共通仕様
Table 1 RENKEI-PoP Specification

| Component | Spec. |
|---|---|
| CPU Speed | Enough speed to sustain 10Gbps network |
| CPU Function | Virtulization technology |
| Memory | Enough size to run multiple VMs |
| Network Bandwidth | 10Gbps |
| Storage | Enough size and speed to store temporal transmitted data |

このPoPによるe-Science支援ソフトウェアの仮想ホスティング環境と，大容量ストレージからなるRENKEI-PoP構想により，上記の要件を以下のように満たす．

**大規模計算機センターリソース群と研究室リソースのシームレスな連携**

リソースの連携はRENKEI-PoPの仮想ホスティング環境に導入する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・共有環境が有効になっている．拠点間でデータを送受信する際には，異なる拠点内リソース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．中継のためのデータの一時的なバッファスペースとして，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ユーザ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て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，

- RENKEI-PoPを拠点リソース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
- 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の仮想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，拠点計算リソース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リソースを用いる場合には，拠点計算リソース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りうる．

**表 2** RENKEI-PoP 設置拠点
Table 2 Locations where RENKEI-PoPs are installed

| Location | Host Name | Remarks |
|---|---|---|
| 東京工業大学 | titech, titech2 | Two RENKEI-PoPs are installed |
| 大阪大学 | osaka | |
| 国立情報学研究所（NII） | nii | Installed in Chiba annex |
| 高エネルギー加速器研究機構（KEK） | kek | |
| 名古屋大学 | nagoya | |
| 筑波大学 | tkb | |
| 産業技術総合研究所（AIST） | aist | Installed in Tsukuba |

## 4. RENKEI-PoP 配備・展開状況

RENKEI-PoPは2010年7月現在，表2からなる，日本国内7拠点に設置されている．東京工業大学に2台あることを除き，各拠点1台のRENKEI-PoP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．最初期に導入されたKEKのRENKEI-PoPと，最新機にアップデートした東京工業大学のRENKEI-PoPの仕様を表3に示す．10Gb Ethernetと8台のSSDからなるRAID，あるいは16台のHDDからなるRAIDを用いることで，理論上1GBpsのDisk-to-Diskのボトルネックレス転送が可能となる．全RENKEI-PoPのストレージ容量の合計は192.5TB (raw)である．具体的には，NII，KEKのRENKEI-PoPでは容量256GB (raw)のSSD RAIDを，東京工業大学，大阪大学，名古屋大学，筑波大学，AISTのRENKEI-PoPでは容量32TB (raw)のHDD RAIDを用いている．将来的にはNII，KEKのRENKEI-PoPも大容量のHDD RAIDに交換することを検討している．

RENKEI-PoPのネットワーク接続状況を図2に示す．個々のRENKEI-PoPは最大2つのネットワークに接続されている．1つは広域インターネット網であり，ユーザがRENKEI-PoPにアクセスするために使用する．計算機センター運用の都合上，インターネットには接続できていない拠点があり，接続している拠点でもセキュリティの都合上，http, ssh, ftp等必要なポートのみ開けている．

**表 3** KEK と東京工業大学の RENKEI-PoP の仕様
Table 3 Specification of RENKEI-PoPs in KEK and Tokyo Tech

| Component | KEK | Tokyo Tech |
|---|---|---|
| CPU | Intel Core i7 965 Extreme (3.2GHz) | Intel Core i7 975 Extreme (3.33GHz) |
| Memory | DDR3 12GB (2GB x6) | DDR3 12GB (2GB x6) |
| Network Card | Chelsio 10Gb Ethernet | Myricom 10Gb Ethernet |
| Storage | SSD Raid 256GB (32GB x8) | HDD Raid 32TB (2TB x16) |

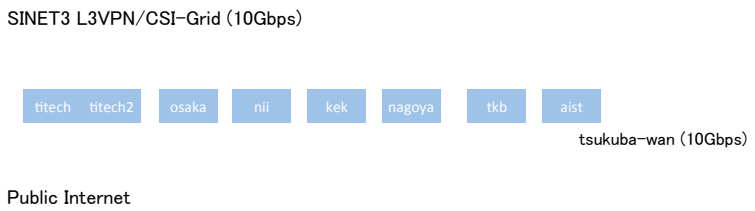



**図 2** RENKEI-PoP 間ネットワーク
Fig. 2 Network connection between RENKEI-PoPs

もう 1 つは拠点間高速ネットワークとして使用している SINET3 L3VPN/CSI-Grid（以降 CSI-Grid）である．CSI-Grid は大学・研究所間を接続する学術ネットワーク SINET が提供する閉域ネットワーク網で回線速度は 10Gbps である．ネットワーク配線の都合上接続されていない拠点もあり，また，最大 10Gbps での通信を行えるが，ハードウェア仕様や拠点内経路等の都合により，後述のように十分な性能が出ていない拠点もある．閉域網であるため，ファイアウォールは設置していない．東京工業大学や筑波大学等，大規模計算リソースを持つ拠点では，RENKEI-PoP からこの 2 つのネットワークのどちらかで計算リソースへのアクセスが可能である．また，東京工業大学では RENKEI-PoP とスーパーコンピュータ TSUBAME でアカウントを統一し，TSUBAME ストレージ・テープライブラリとの透過的な連携を目指した設計を進めている．

次に RENKEI-PoP に導入されているソフトウェアの説明をする．RENKEI-PoP の OS には CentOS 5.4 を使用している．RENKEI-PoP で有効な認証手段は Unix ユーザ認証と globus[9] gsi 認証の 2 つである．Unix ユーザ認証は利用者が研究室から，あるいは拠点リソースから最寄りの RENKEI-PoP にアクセスする際に使用される．RENKEI-PoP 間では，Unix ユーザ認証も同様に可能であるが，gsi 認証が設定されている．gsi 認証を用いることで，

- アカウント名が異なる RENKEI-PoP 間でのログイン・データ転送の透過性実現
- 別プロジェクトで進められている，NII，東工大を含む全国 9 大学で配備を行っている計算機センター間グリッドの認証基盤の利用

が可能となる．

RENKEI-PoP 間でのデータ転送には，標準的な scp，ftp 等に加えて，gsi-enabled ssh，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している．また，CSI-Grid に接続された拠点間では Gfarm[10] 分散ファイルシステムによる，最大 110TB のデータ共有環境が構築されている．Gfarm でも gsi 認証を利用可能である．現状，拠点間転送を行うには，利用者が手動で RENKEI-PoP へのデータ入出力を行う必要があり，自動化，および RENKEI-PoP 上の一時データの管理を今後の課題としている．

RENKEI-PoP 上での仮想ホスティング環境として，現状では kvm と libvirt による仮想マシン実行環境を提供している．

## 5. 評　　価

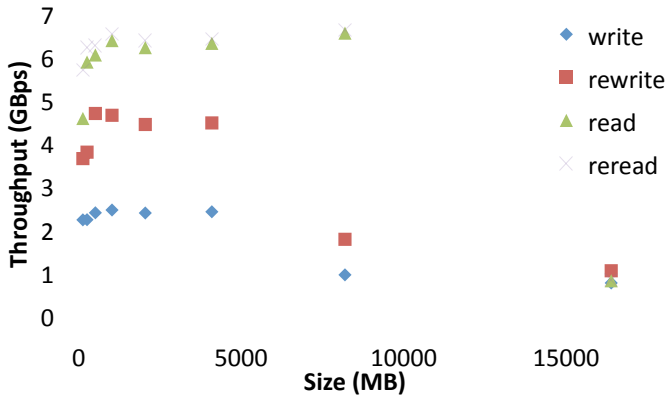
評価として，RENKEI-PoP のファイルシステム IO 性能，RENKEI-PoP 間のネットワーク遅延（RTT: ラウンドトリップタイプ）・バンド幅の計測，アプリケーションデータを用いた RENKEI-PoP 間のデータ転送性能の計測を行った．

### 5.1 ファイルシステム IO 性能

HDD RAID ストレージデバイスを搭載している東京工業大学の RENKEI-PoP におけるファイルシステム IO 性能を iozone ベンチマークにより計測した．ストレージデバイスはデータを安全に保持するため RAID 5 でフォーマットされており，ファイルシステムには xfs を用いている．

結果を図 3 に示す．RENKEI-PoP には 12GB のメモリが搭載されており，8GB までの結果はキャッシュに乗るためメモリアクセス性能に依存し，16GB の場合にはストレージデバイスアクセス性能が結果に大きく影響を与える．16GB の結果では Re-Write でしか目標性能を達成できていないが，これは RAID 5 で構成したが原因だと考えている．実際，RAID 0 で構成したときのデータでは，Write が Re-Write と同程度の 1.03GBps に，Read，Re-Read も若干の向上が得られていた．

なお，Write と Re-Write で大きく性能が異なる理由は，Re-Write では既に存在するファイルへの書き込みを行うのに対して，Write では新規ファイルを作成するため，ディレクト




| Size | Write (GBps) | Re-Write (GBps) | Read (GBps) | Re-Read (GBps) |
|---|---|---|---|---|
| 128M | 2.21 | 3.63 | 4.46 | 5.68 |
| 256M | 2.22 | 3.78 | 5.86 | 6.20 |
| 512M | 2.37 | 4.67 | 6.03 | 6.25 |
| 1G | 2.44 | 4.63 | 6.36 | 6.51 |
| 2G | 2.37 | 4.42 | 6.20 | 6.37 |
| 4G | 2.40 | 4.46 | 6.30 | 6.40 |
| 8G | 0.94 | 1.76 | 6.53 | 6.61 |
| 16G | 0.75 | 1.03 | 0.80 | 0.81 |

**図 3** ファイルシステム IO 性能
Fig. 3 Filesystem Performance

リ情報の更新，ディスクスペースの確保等のオーバーヘッドが生じているためである．

### 5.2 ネットワーク性能

CSI-Grid 10Gbps ネットワーク網に接続されている 7 つの RENKEI-PoP 間での RTT・バンド幅計測を行った．

RTT の計測結果を表 4 に示す．全 RENKEI-PoP 間で双方向通信の遅延がほぼ等しいことがわかる．また，関東圏にある RENKEI-PoP から大阪大学の RENKEI-PoP への遅延が大きく，最大で 14.39ms であることが確認できる．

ここで得られた RTT を参考に，RENKEI-PoP の転送バッファサイズの調整を行った．転送バッファサイズを含む，CentOS 5.4 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 5 に示す．転送バッファサイズ（Maximum buffer size）は RTT と最大理論バンド幅である 10Gbps と掛け合わせた帯域遅延積とした．その他のパラ

**表 4** RENKEI-PoP 間の RTT (単位は millisecond)
Table 4 Round trip time between RENKEI-PoPs (millisecond)

| From\To | titech | titech2 | osaka | nii | kek | nagoya | tkb |
|---|---|---|---|---|---|---|---|
| titech | | 0.24 | 9.80 | 4.27 | 5.87 | 6.30 | 5.71 |
| titech2 | 0.16 | | 9.62 | 4.27 | 5.86 | 6.28 | 5.71 |
| osaka | 9.62 | 9.59 | | 12.84 | 14.39 | 4.27 | 14.26 |
| nii | 4.27 | 4.30 | 12.86 | | 9.08 | 9.56 | 8.92 |
| kek | 5.83 | 5.89 | 14.37 | 9.06 | | 11.12 | 0.81 |
| nagoya | 6.32 | 6.33 | 4.28 | 9.51 | 11.06 | | 11.00 |
| tkb | 5.71 | 5.71 | 14.25 | 8.96 | 0.83 | 10.98 | |

**表 5** RENKEI-PoP のネットワークパラメータ
Table 5 Network parameters for RENKEI-PoPs

| | titech | titech2 | osaka | nii | kek | nagoya | tkb |
|---|---|---|---|---|---|---|---|
| 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 | ← | ← | ← | ← | ← | ← |

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．後者に該当する具体例としては，NII の RENKEI-PoP で TCP segmentation offloading を有効にするとパケットロスが頻発し著しい性能低下が起きたこと，KEK の RENKEI-PoP に搭載されている NIC が adaptive interrupt coalescing に対応していないことがある．

iperf によるバンド幅の計測結果を表 6 に示す．近距離の titech - titech2 間では目標性能の 1GBps を達成できているが，拠点間をまたぐ通信では，最大でも titech2 - nii 間の 770Mbps である．特に問題となっているのが，nagoya との通信であり，10Gbps ネットワークに接続されているにもかかわらず，数十 MBps の性能しか出ておらず，経路上のボトルネック箇所の確認調査を行っている．

さらなる性能向上を見込めるが，今回は環境の制約により適用できなかったパラメータとして次の 2 つがある．1 つは MTU である．RENKEI-PoP に搭載されている NIC は全てジャンボフレーム対応しているが，今回の計測ではデフォルトサイズである 1500 を用いた．拠点内スイッチがジャンボフレーム対応していないことが原因である．例えば，東京工業大学では SINET スイッチのポートでジャンボフレーム設定しておらず，対応させるに

**表 6** RENKEI-PoP 間のバンド幅 (単位は MBps)
Table 6 bandwidth between RENKEI-PoPs (MBps)

| From\To | titech | titech2 | osaka | nii | kek | nagoya | tkb |
|---|---|---|---|---|---|---|---|
| titech | | 1120 | 381 | 770 | 451 | 29 | 174 |
| titech2 | 1120 | | 418 | 761 | 450 | 19 | 172 |
| osaka | 530 | 526 | | 466 | 319 | 65 | 205 |
| nii | 557 | 555 | 484 | | 466 | 67 | 490 |
| kek | 412 | 413 | 299 | 376 | | 48 | 748 |
| nagoya | 14 | 12 | 16 | 10 | 7 | | 5 |
| tkb | 357 | 433 | 341 | 415 | 623 | 43 | |

は全学に影響を与えるネットワーク瞬断が起こりうるため，対応困難である．もう 1 つは KEK の RENKEI-PoP の PCI Express (PCIe) 転送性能である．KEK の RENKEI-PoP では PCIe x8 対応の NIC を用いているが，マザーボード側に十分な数の PCIe x8 のスロットが無く PCIe x4 スロットに接続したため，理論上最大でも 8Gbps となる．この問題はマザーボードの変更だけで，簡単に解決できる．

### 5.3 データ転送性能

最後に Disk-to-Disk のデータ転送性能を確認するために，Montage[11] の 6756 個，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．用いた Gfarm のバージョンは 2.3.0 であり，Gfarm metadata server（以降 Gfarm MDS）は東京工業大学に RENKEI-PoP とは異なるサーバで構築し，nagoya と osaka を除く CSI-Grid に接続された 5 台の RENKEI-PoP を Gfarm filesystem server とした．nagoya は 5.2 章のバンド幅計測の結果より極端に帯域が狭かったため，osaka は証明書発行が済んでいないために除外した．測定には gfrep コマンドを用い，titech2, nii, kek, tkb から titech 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 のデータセットの 6756 個のファイルは，最大サイズ 27MB，最小サイズ 292B，平均サイズ 2MB，標準偏差 400KB である．

結果を図 4 に示す．凡例 individual は 6756 個のファイルを逐次に個別転送した場合の結果，archived は 6756 個を 1 つのファイルにまとめて（非圧縮）転送した場合の結果である．両者では Gfarm MDS へのアクセス回数が大きく異なる．総じて individual の結果が archived の結果よりも劣り，titech2 → titech への転送では 4 倍，tkb → titech では 6 倍の性能比がある．原因は小容量のデータを送信しているためネットワークが有効活用されていないことによるのか，Gfarm MDS のアクセス性能によるのか，完全な究明は行えていない

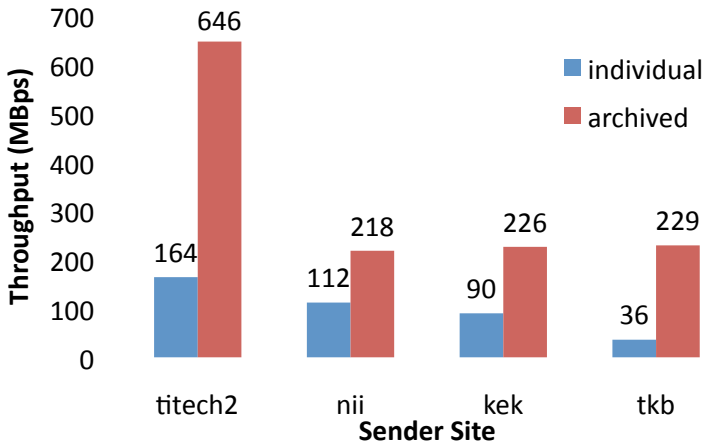



**図 4** titech への Gfarm データコピー性能
Fig. 4 Data copy performance to *titech* by Gfarm

が，少なくとも titech2 → titech では後者が原因だと考えている．実際，titech2 - titech 間で iperf により 2MB のデータを転送したときのスループットは 627MBps を記録しており，この差分は Gfarm MDS へのアクセスオーバーヘッドであると考えられる．また，長距離 Gfarm 転送の性能向上のためには，ソケットバッファサイズの変更が有効であると確認できており，将来的にはこの修正を行うことで individual，archived 共に性能向上を図る．

以上の結果より，ファイルはまとめて送信することが望ましいとわかる．ただ，個別転送の性能向上のために，並列データ転送をサポートすることを検討している．

## 6. RENKEI-PoP 活用計画

RENKEI-PoP はまだ配備・設定を進めている最中ではあるが，RENKEI-PoP を活用した研究プロジェクトの話が出てきているので，いくつか紹介する．

静岡県三島市にある国立遺伝学研究所，東京大学柏キャンパス，宮崎大学等に設置された新型 DNA シーケンサにより生成された多数の小容量（数百 MB）ゲノムデータの転送基盤として RENKEI-PoP を用いるプロジェクトがある．実際，国立遺伝学研究所と東京工業大学に RENKEI-PoP を設置し，国立遺伝学研究所で生成されたゲノムデータを東京工業大学 TSUBAME スーパーコンピュータで解析する計画を立てている．

乱流データ解析の共同研究プロジェクトが，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おり，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．

本研究を補助するRENKEIプロジェクトでは，他のチームがジョブ実行，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を活用することになっている．

東京工業大学では，上述した，計算機センター間グリッドを構成するためのソフトウェアをRENKEI-PoP上の仮想ホスティング環境に乗せることを検討している．このグリッドではNAREGIグリッドミドルウェア[12]を用いており，各センターでは以下の3種類のサービスを動かしている．

(1) 拠点内計算リソースの情報を収集する情報サービス
(2) 拠点内スケジューラへジョブ投入するGridVM Scheduler
(3) 拠点内計算リソース管理のため，計算リソースに導入されるGridVM Engine

このうち，(3)のGridVM Engineは無くても，機能制限はされるが，拠点計算リソースをグリッドに接続することは可能であり，(1)，(2)の2つのサービスのみをRENKEI-PoPの仮想ホスティング環境で動作させることを検討している．

## 7. まとめ

e-サイエンス実証ネットワーク基盤環境として，我々はRENKEI-PoPにより連携された拠点間ネットワーク環境を提案した．RENKEI-PoP上の仮想ホスティング環境でe-Science基盤ソフトウェアを実行し，ユーザ研究室内のリソース，RENKEI-PoP設置拠点のリソース連携を実現すると同時に，拠点間の高速データ転送をサポートするネットワーク環境である．ネットワーク，ストレージ性能評価の結果，特に長距離通信性能が低く，ネットワーク性能チューニングが今後の課題であることが確認された．

今後の計画として以下を考えている．まず，データ共有環境に関しては，現状ではユーザが手動でRENKEI-PoP経由のデータ転送をする必要があったが，透過的に行える手法の提供を予定している．また，拠点のストレージ資源（並列ファイルシステム，テープライブラリ）と連携した，階層的なデータステージング実現のための技術検討も行っている．仮想ホスティング環境に関しては，OpenNebula等，仮想マシン・ネットワーク環境を柔軟に構成するためのフレームワークを導入する．また，e-サイエンス基盤ソフトウェア実験・評価用途で使うネットワークエミュレータや，通信性能解析のための監視機能を導入する．さらに，これらサービスの充実に平行して，RENKEI-PoP設置拠点の追加も計画している．

**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 (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 TestGrid, *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 (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* (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) Tatebe, O., Soda, N., Morita, Y., Matsuoka, S. and Sekiguchi, S.: Gfarm V2: A Grid File System that Supports High-Performance Distributed and Parallel Data Computing, *the 2004 Computing in High Energy and Nuclear Physics* (2004).
11) : Montage, http://montage.ipac.caltech.edu/.
12) Matsuoka, S., Shimojo, S., Aoyagi, M., Sekiguchi, S., Usami, H. and Miura, K.: Japanese Computational Grid Research Project : NAREGI, *IEEE*, Vol.93, No.3 (2005).