Vol.2011-HPC-130 No.67
2011/7/29

広域分散環境を提供するHPCIシステムソフトウェア基盤の設計概要と共有ストレージ構築

實本英之[†1]　建部修見[†2]
佐藤仁[†3]　石川裕[†1]

大規模スパコンである「京」とその他国内主要な計算資源をユーザが容易に利用できる環境として、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) を構築する。基本設計は、ネットワーク、認証基盤、共有ストレージ、ユーザ管理支援、先端ソフトウェア運用基盤に分けて構成され、このうちの共有ストレージについて詳細に紹介する。共有ストレージは、HPCI 各拠点から利用可能な大容量、高信頼、高可用を必須要件とし、かつ持続的な運用が可能であることを重視した検討を行った。実際に運用されている共有ストレージにおける幾つかの利用形態か、求められる特性、運用形態を検討した結果、Gfarm を中核としたシステムを設計し、平成２３年度後期を目処に構築を行っている。

Design Overview of System Software and Shared Global Storage in HPCI Wide Area Distributed Environment

Hideyuki Jitsumoto,[†1] Osamu Tatebe,[†2]
Hitoshi Sato[†3] and Yutaka Ishikawa[†1]

HPCI is constructed as an environment for the user can easily use "K" supercomputer and other highest level computation resource in Japan. We introduce the initial design of shared storage system on HPCI, which includes five working groups: network environment, authentication method, user management, advanced software development/operation environment and our shared storage system. The shared storage system is focused on large capacity, high dependability, high availability, and operational sustainability. In consideration of several operative shared storage systems, we construct our system with Gfarm as the best solution. This system operation will start from the second half of FY2011.

1. はじめに

「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) とこの構築を主導するコンソーシアムのグランドデザイン（平成 22 年 5 月 26 日文部科学省）」[1] に基づき、同年 7 月に計算資源提供機関 25 機関とユーザコミュニティ機関 13 機関の計 38 機関からなる HPCI 準備段階コンソーシアムが発足した。HPCI では、次世代スパコン「京コンピュータ」[2] と全国に存在するスーパーコンピュータセンターを高速ネットワークでつなげるとともに大規模ストレージシステムを導入し、透過的資源アクセスを提供することによりユーザの利便性を高める。

現在このコンソーシアムが主導して、平成 24 年 11 月の運用開始を目途に HPCI の構築とコンソーシアムの形成に向けた検討と準備が進められている。これと平行して、平成 24 年 12 月より、東京大学が代表機関となり、情報・システム研究機構 (NII)、北海道大学、東北大学、筑波大学、東京工業大学、名古屋大学、京都大学、大阪大学、九州大学、理化学研究所計算科学研究機構と共に HPCI のシステムソフトウェアに関する基本設計を行っている。

基本設計では、大きく、ネットワーク、認証基盤、共有ストレージ、ユーザ管理支援、先端ソフトウェア運用基盤の 5 つに分けて検討を行った。システム設計時に重要となるのは、運用ポリシーとメカニズムの分離である。メカニズムが運用ポリシーを規定すべきではないし、運用ポリシーがメカニズムを規定すべきではない。可能な限りの運用ポリシーを提供できるようシステムソフトウェア側は提供していかなければならないとの認識で検討を行った。

本稿では、HPCI において検討中のシステムソフトウェアの全体像を紹介した後、共有ストレージについて述べる。

2. HPCI 概要

HPCI は、次世代スーパーコンピュータ「京」を中核とし、これと国内の計算資源を連携

†1 東京大学
The University of Tokyo
†2 筑波大学
University of Tsukuba
†3 東京工業大学
Tokyo Institute of Technology

Vol.2011-HPC-130 No.67
2011/7/29

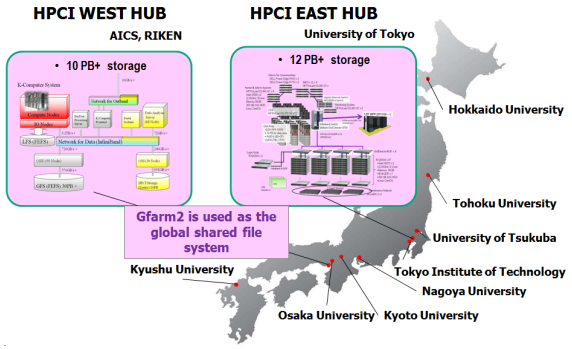

図 1 初期環境におけるストレージ資源

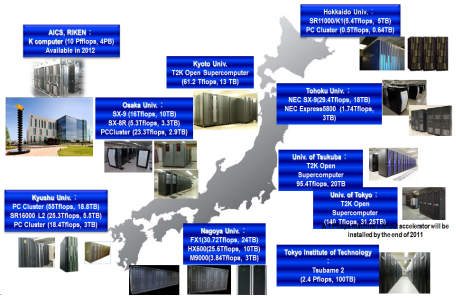

図 2 初期環境におけるコンピューティング資源

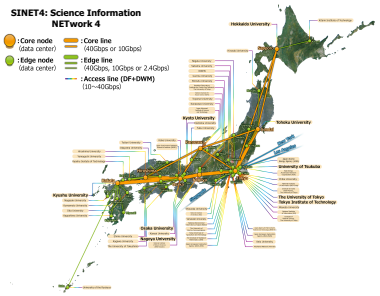

図 3 ネットワーク資源− SINET4

して大規模に利用するための計算基盤である。この基盤は、コンピューティング資源、ネットワーク資源、ストレージ資源、人的資源の連携により達成される。

初期環境では、コンピューティング資源として大学の計算センター、理化学研究所計算科学研究機構をはじめとする各研究所のスーパーコンピュータが所属し、ネットワーク資源として学術情報ネットワーク SINET4[3] を用いこれらを接続する (図 2, 3)。またストレージには東拠点として東京大学柏キャンパスに設置された 12PB ストレージ、および西拠点として理化学研究所計算科学研究機構の 10PB ストレージが利用される (図 1)。

3. 共有ストレージシステム

HPCI で整備する共有ストレージ（以下、「HPC ストレージ」）は、HPCI の各資源提供

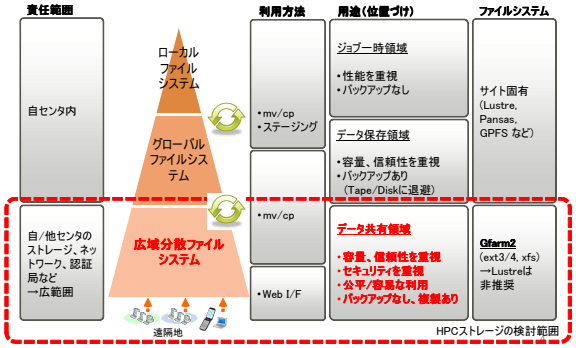

図 4 ファイル階層における HPC ストレージの位置づけ

機関に属する国内の研究者間、研究グループ内で効率よくデータ共有することを目的として構築し、研究業務の効率化や利便性の向上を図る。共有ストレージ基盤は持続的なシステム増強が図られることを念頭に、提供するサービスの目標は以下である。

- HPCI 上のどの拠点からでも利用可能な大容量ストレージ基盤の提供
- HPCI 上のスパコン間での効率的なファイル共有機能の提供
- 常に安定して稼働するための高い RAS 性を備えた共有ファイルシステムの提供（高信頼、高可用、常時安定して稼働する共有ストレージの提供）

この目標を達成するため、HPC ストレージは、既存のスパコンのファイルシステムとは別に、どの拠点からもアクセス可能な広域の分散ファイルシステムとする構成を考える。図 4 に HPC ストレージと各拠点スパコンのファイルシステムの関係を示す。この図では、拠点スパコンのファイルシステムは、ジョブ一時領域としての超高速/小容量のローカルファイルシステムと、データ保存領域としての高速/大容量のグローバルファイルシステムに分かれているが、拠点によってはローカルファイルシステムを利用しないでグローバルファイルシステムだけで構成される拠点もありうる。

このように、既存のスパコンのファイルシステムとは別の広域分散ファイルシステムとして HPC ストレージシステムを構成することにより、既存のスパコンシステムに対する設定変更を少くすることができ、また、ファイルシステムとして構成することにより、既存のアプリケーションからのアクセスも可能となり、利用者の研究業務の効率化や利便性の向上を図ることができる。

また、複数拠点、多数の利用者によるデータ共有において、共有の仕組みや信頼性の確保、セキュリティ面での考慮は特に重要となる。公平なグループ単位の容量制限や容易な利用（アクセス方法）の整備も求められる。

4. HPC ストレージに求められる課題

前節で挙げた、目標達成のための課題を明確化するため、HPCI において実行される可能性のある幾つかの計算について、HPC ストレージに求める仕様を検討した。

4.1 ユースケース

格子 QCD（アンサンブルデータ） 次世代スーパーコンピュータ戦略プログラムである格子 QCD において、様々の解析に用いられるアンサンブルデータはデータ生成に膨大な計算機資源が必要とされ、非常に貴重なデータとされる。データサイズは 100TB 規模の大きさで、複数の研究機関の研究者がデータを共有して解析を進めたいニーズが挙げられている。

格子 QCD では京コンピュータや拠点スパコンで長時間実行して得られたゲージ配位アンサンブルデータを HPC ストレージに格納する。ゲージ配位アンサンブルデータは、各拠点のスパコンで様々な物理量を計算するために複数の研究者により利用される。

ライフサイエンス（シークエンスデータ） ゲノムシークエンスデータは数 TB 規模の大きさが想定される。また、Bio-Mirror.net[4] のような DNA データバンクを利用するニーズもある。また、一時的に特定ユーザにのみデータアクセスを許可したい、許可したユーザからのファイルアクセスがあったことを監査したいなど、セキュリティの固有ニーズが挙げられている。ライフサイエンスのシークエンスデータをそれぞれの拠点スパコンのストレージにコピーした上で解析が進められる。扱うデータは DNA データバンクのような共有データもあれば、取り扱いが厳しく、公開範囲が限定されるデータも含まれる。ライフサイエンスの種類によってはアセンブリング後のデータを共有するニーズもある。

スパコンでのシミュレーション解析（大規模解析データ、実験データ） 大規模シミュレーションの解析結果のデータ保管庫としての利用ニーズがある。データサイズは 1PB 規模の大きさが想定される。ログインノードだけではなく、計算ノードからもファイルアクセスしたい、どの拠点からでも同じパス名でアクセスしたいなどのニーズが挙げられている。

セキュアストレージ（医学・製薬系コミュニティの個人データ） データの取り扱いに特に厳しいポリシを持つ” コミュニティ” において、利用者自身で段階的にアクセスコントロールする仕組みを設けることにより、安全かつ容易に利用可能なセキュアストレージの利用ニーズが挙げられている。

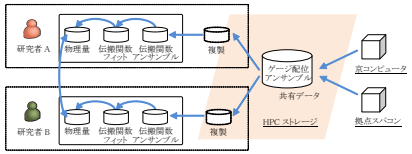

図 5 格子 QCD における共有データフロー

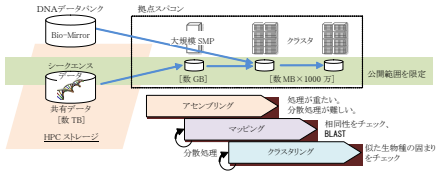

図 6 ライフサイエンスにおける共有データフロー例

ユースケースから、HPC ストレージに求められる利用シナリオをまとめると以下の２つが挙げられる。

データ共有 HPC ストレージにファイルを保存し、ファイルアクセスの権限を適切に設定することにより、特定のコミュニティ（HPC 課題グループ）内、あるいは HPCI 全体について再利用性の高いデータを共有する。共有されたデータは HPCI 全拠点において、更に複数の利用者により大規模に利用することができるものとする。これらのデータは再生成の難しい大規模中途データや、グローバルなデータベースを利用して効率的な計算を行うためのデータ複製を想定する。

セキュアストレージ 機密性の高いデータについて、複数の認証手法を選択することにより適切なセキュリティ強度を実現する。基本仕様では GSI（Grid Security Infrastructure）によるユーザ認証、ACL によるユーザ/グループ単位のより細かなアクセス制限（ACL）、ユーザレベルでのファイル暗号化によるデータストアを想定する。

4.2 HPC ストレージと各拠点の運用ポリシ

HPC ストレージを利用するにあたり、各資源提供機関の既存システムにおいては、最低限の環境変更や HPC ストレージの運用に則った運用ポリシーの見直しが必要となる。それによって既存システムの運用および HPC ストレージの運用が阻害されることのないよう、アクセス手法、認証について検討した。

アクセス手法 一般的な計算センターでは、グローバルネットワークに配置されたログインノードを経由して、プライベートネットワークにある計算資源を利用することが多い。このため、計算ノードから直接 HPC ストレージを利用することは各資源提供機関に大きな構成変更を強いる可能性がある。このため、HPC ストレージへはログインノー

ドからアクセスすることを基本とする。また、利便性を保つため、ユーザは HPC ストレージをログインノードでマウントして利用する。マウントされた HPC ストレージはいかなる拠点においても同じ名前空間（パス名）でファイルアクセスできることが望ましい。

更に補足事項として、HPC ストレージに保存されたデータについては、仮りにそのデータが各資源提供機関に分散して保存されていようと、ユーザはその分散配置を意識することなくアクセスできることが望ましい。その場合、効率的なアクセスが可能となるよう、データの再配置、複製作成などについても、ユーザが意識することなく、ストレージシステムが適切に対応することが望ましい。

認証 HPCI では、ユーザの利便性のため、シングルサインオンによるアクセスが認証基盤 WG で検討されている[5]。現在、そのためにそれぞれのユーザには一意の HPCI アカウントの提供が検討されている。HPC ストレージに対する認証においても、このシングルサインオンの機構を利用することが望ましい。また、ファイルアクセスの権限は、各拠点におけるローカルアカウントの権限ではなく、HPCI アカウントの権限であることが望ましい。これにより、どの拠点のログインノードからも、仮りに異なるローカルアカウント、ローカルグループでログインしていた場合でも、HPCI アカウントの権限で同様にアクセスできるようになる。

HPCI 資源の利用はプロジェクト課題毎に資源利用が認可されるため、HPC ストレージもこの課題グループのグループメンバシップにより HPCI アカウントの管理を行う必要がある。

5. 共有ストレージの構築

以上の目標、要求を考慮した上で、HPC ストレージの実装を検討した。

5.1 広域分散ファイルシステム

データ共有の利便性を図るためには、HPC ストレージは各拠点のログインノードでマウント可能なファイルシステムである必要がある。さらには、どの拠点からアクセスしても効率的にアクセスできるような仕組が必要である。

分散したストレージに対する透過なアクセスを実現する分散ファイルシステムとして Andrew File System (AFS)[6] がある。AFS は、遠隔ファイルサーバに対するアクセス性能を向上させるため、ファイルをアクセスするときにクライアント側にキャッシュする。ただし、HPCI においてデータ共有するような大規模ファイルについては、クライアント側にキャッシュすることができず、効率的なアクセス性能を実現することが難しいと考えられる。また、認証基盤 WG で検討しているシングルサインオンの機構をどのように利用するのか、特定ユーザ、特定グループに対するアクセス制御をどのように実現するのか、など問題点も多い。

Lustre ファイルシステムに対する遠隔からの効率的なアクセスに関する研究[7] もある。ただし、この研究は、複数拠点に分散したストレージに対する透過なアクセスを実現するのではなく、特定拠点の Lustre ファイルシステムに対する高速アクセスを実現するものであり、複数拠点にストレージを分散配置する HPC ストレージの構成とはあわない。

米国の TeraGrid では、GPFS-WAN (Global Parallel File System-Wide Area Network) を利用している。GPFS-WAN はファイルシステムは、物理的には全てのストレージが SDSC (San Diego Supercomputer Center) に配置されており、複数拠点にストレージを分散配置する HPC ストレージの構成とはあわない。

広域に分散するストレージに対する透過なアクセスを実現する、オープンソースの広域分散ファイルシステムとして、Gfarm ファイルシステム[8] がある。Gfarm ファイルシステムは、分散ストレージに対してファイル複製を作成することができ、各拠点からのアクセス性能の向上が可能なだけではなく、耐障害性も向上させることが可能である。また、運用を停止することなく、ストレージの追加を行うことが可能であり、持続的なシステム増強も可能である。また、単一障害点がなく、高い RAS 性も備えている。このように、HPC ストレージにおけるほとんど全ての要求を満たす可能性をもっている。そのため、以後、Gfarm ファイルシステムを念頭において HPC ストレージの構築を検討する。

5.2 ローカルアカウントとストレージアカウントの関係

HPC ストレージへのアクセスは HPCI アカウントで行う必要がある。それぞれの HPCI アカウントに対し GSI のユーザ証明書が発行されるため、HPC ストレージへのアクセスに対し GSI のユーザ証明書の Subject DN（Distinguished Name）をユーザの identity として利用すればよい。Gfarm ファイルシステムでは、GSI 認証の場合、ファイルシステムのユーザは GSI のユーザ証明書の Subject DN により決めることが可能である。

拠点のログインノードから HPC ストレージにアクセスする場合、同一ユーザであっても拠点毎にユーザ名が同じとは限らないが、Gfarm ファイルシステムで Subject DN によるグローバルなユーザ、グループ名を管理することで、各拠点から HPC ストレージ上のファイルにこのグローバルユーザ名で一意にアクセスすることが可能である。ローカルアカウントと HPCI アカウントの関係を図 7 に示す。

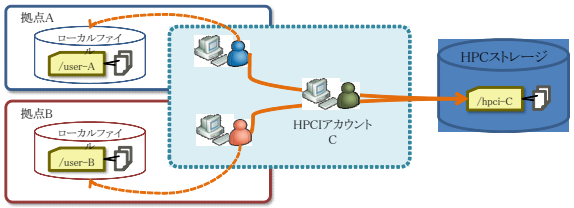

図 7 ローカルアカウントと HPCI アカウントの関係

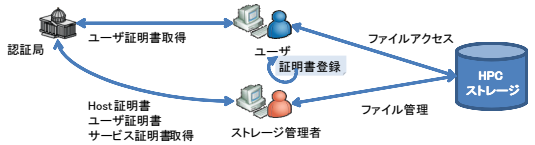

図 8 gsi_auth 手法によるユーザ認証イメージ

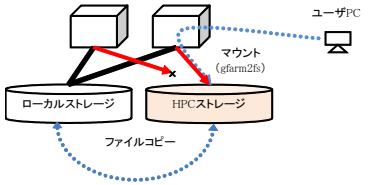

図 9 HPC ストレージへのアクセス方法

5.3 認証方法

Gfarm ファイルシステムでは、ユーザの認証方式として、sharedsecret 認証、GSI を用いた gsi_auth 手法、GSI 認証の 3 種類をサポートしている。データ共有、データアーカイブシナリオでは性能を重視し、gsi_auth 手法によるユーザ認証を基本とする。gsi_auth 手法はユーザ認証時に GSI 認証を用い、その後のデータ通信はデータに対する署名や暗号化保護のない生データの転送を行うため、セキュリティ強度は落ちるが、高速なデータ転送が可能である。

一方、セキュアストレージシナリオでは GSI 認証を利用することを推奨する。この手法ではユーザ認証もデータ通信時にも暗号化通信が可能であり、gsi_auth に対して通信性能は劣るが高いセキュリティを確保することができる。どの認証を行うかは自身のホームディレクトリに Gfarm 設定ファイル（ /.gfarm2rc）を作成することで、ユーザ個々に認証方法を切り替え、利用することが可能である。

5.4 HPC ストレージへのアクセス方法

HPC ストレージへのアクセス方法を図 8 に示す。アクセス手法は長期的な検討事項ではあるが、第 1 段階としてはログインノードからの利用を考える。HPC ストレージにアクセスしたい HPCI 上の計算資源のログインノードにログインし、HPC ストレージをマウントすることで、ローカルストレージと同様に透過的にファイルアクセスを行う。または、Gfarm が提供するコマンドを利用し、ファイルアクセスを行う。

第 1 段階においては、ログインノードまたは public IP アドレスに対して接続が可能なクライアントからのアクセスを考え、public IP アドレスに対し直接接続できない計算ノードからの HPC ストレージへの直接アクセスは考慮しない。HPC ストレージのマウントはユーザ自身で gfarm2fs コマンドを実行することでマウントされる。このマウント操作はユーザ権限で行われ、システム側で予め静的にマウントしておくことはできないが、一度マウントすれば、ユーザ PC からもログインノードを介し、HPC ストレージにデータ転送できる。

5.5 ファイル容量制限

Gfarm ファイルシステムの quota 機能を用いてユーザ/グループ単位にファイル数、ファイルサイズを制限することが可能である。Gfarm ファイルシステムでは、ファイル複製を任意数作成可能であるが、ファイル複製をふくめたファイル数、ファイルサイズでも制限することが可能である。

5.6 効率的なファイル共有・複製方法

Gfarm ファイルシステムに格納されたデータは、いずれかの拠点のファイルシステムノードに作成、格納されるが、ユーザは格納場所を意識することなく、ファイル共有が可能である。

ファイルを新規作成する際は自拠点（ファイル作成を行った拠点）のファイルシステムノードが優先的に選択されるため、ユーザは意識することなく、高速なファイルアクセスが可能である。Gfarm ファイルシステムにおいて、どこにファイルが作成されるのか、どのファイル複製が参照されるかの基本ルールはユーザでも認識すべき事項のため、以下に補足する。

- 新規ファイルを作成する場合、自拠点のファイルシステムノードに十分な空きがあれば書き込む。十分な空きがないと判断された場合、よりネットワーク的にアクセス元に近く、負荷の低いファイルシステムノードに書き込む。
- 既存ファイルへアクセスする場合、ファイル複製があれば自拠点のファイルシステムノードへアクセスする。近くのファイルシステムノードにファイル複製がない場合、ファイル複製を持つ、よりネットワーク的にアクセス元に近く、負荷の低いファイルシステムノードへアクセスする。
- ユーザ自身でファイル複製を操作することも可能である。任意ファイルの複製を、指定されたホスト群に、指定された複製の数だけ、作成することができる。また、このファ

イル複製はファイル容量制値以下になるように制限も働く。それ以外に、ファイル複製数、どのファイルシステムノードに複製が存在するかなどを確認するコマンドも利用できる。

5.7 アクセス制限

セキュアストレージなどの機密性の高いデータについて、また特定ユーザ、特定グループに対して公開したいデータについては、コミュニティ内で安全にファイル共有するためのアクセス制限機能として、Gfarm ファイルシステムでは、POSIX 1003.1e DRAFT 17 に準拠した拡張アクセス制御リスト（ACL）が利用可能である。ACL エントリには所有者（owner）、指定ユーザ（named user）、所有グループ（group）、指定グループ（named group）、その他（other）があり、これらを組み合わせたアクセス権（rwx）を設定することが可能である。拡張 ACL の設定変更は所有者自身で行うことができる。

更に取り扱いの厳しいデータについては、ユーザ自身でファイルを暗号化して HPC ストレージに格納するものとする。ユーザが暗号化したファイルはパスワードや鍵が漏れない限り、管理者であっても復号化することは困難である。

6. おわりに

平成 24 年 11 月を目途に運用開始される HPCI についての概要、およびその中核となる HPC ストレージにおける検討・設計について述べた。また、HPC ストレージの仮運用における構築手法について詳細に述べた。

これらの検討・設計は本運用に向けて更に進めていく予定であり、以下のような内容を検討中である。

アクセス方法の拡充　スパコンのログインノードからだけではなく、CIFS/WebDAV などで PC からも容易にデータアクセスしたいニーズが考えられる。また、ファイル操作を行うポータルサービスの構築も利用者の利便性向上につながる。これらに関しては、システムのセキュリティポリシや管理工数の増加にもつながるため検討が必要である。

計算ノードからのデータ参照　計算ノードから HPC ストレージ上のファイルカタログ（ファイル所在情報）を参照する機能、計算ノード〜HPC ストレージ間のファイルステージング機能、計算ノードから直接、HPC ストレージ上のファイルにアクセスする機能、拠点スパコンと HPC ストレージ間の自動ファイル同期機能など、多様なユーザニーズが挙げられているが、これらのユーザニーズの強さ、実現可否、費用対コスト、実現する上での課題有無などを検討する。

以上の検討に加え、HPC ストレージの信頼性や性能向上、管理者やユーザの利便性向上を目的とした開発を行う。具体的には、Gfarm ファイルシステムについて、1) HPC ストレージと拠点のスパコンのファイルシステム間の高速ファイルコピー、拠点間の複製作成機能、2) メタデータサーバ冗長化（ホットスタンバイ）、3) 自動ファイル複製機能の高度化、4) 監視機能の高度化といった拡張を予定している。また、Gfarm の性能評価および負荷テストツールの開発も行う。

謝辞　本稿をまとめるにあたり御議論頂きました「HPCI の基本仕様に関する調査検討」委員の皆様、ならびに、HPCI システム WG 委員の皆様に感謝致します。また、本研究の一部は、文部科学省委託事業「HPCI の基本仕様に関する調査検討」および「HPCI の詳細仕様に関する調査検討」によります。

参 考 文 献

1) 革新的ハイパフォーマンス・コンピューティグ・インフラ (HPCI) : . http://hpcic.riken.jp/.
2) 次世代スーパーコンピュータの開発・整備 : . http://www.nsc.riken.jp/.
3) 学術情報ネットワーク SINET4 : . http://www.sinet.ad.jp/.
4) Bio-mirror.net: . http://bio-mirror.net/.
5) 合田憲人，　東田学，漆谷重雄，天野浩文，坂根栄作，小林克志，青木道宏，柴山悦哉，　石川裕：広域分散環境を提供する HPCI ネットワーク・認証・ユーザ管理支援基盤の設計，*2011-HPC-130* (2011).
6) Howard, J.H., Kazar, M.L., Menees, S.G., Nichols, D.A., Satyanarayanan, M., Sidebotham, R.N. and West, M.J.: Scale and performance in a distributed file system, *ACM Transactions on Computer Systems*, Vol.6, No.1, pp.51–81 (1988).
7) Simms, S.C., Pike, G.G. and Balog, D.: Wide Area Filesystem Performance using Lustre on the TeraGrid, *Proceedings of the TeraGrid 2007 Conference* (2007).
8) Tatebe, O., Hiraga, K. and Soda, N.: Gfarm Grid File System, *New Generation Computing*, Vol.28, No.3, pp.257–275 (2010).