Vol.2013-ARC-207 No.8
Vol.2013-HPC-142 No.8
2013/12/16

HPCI共用ストレージの性能評価

建部 修見[1,5,a)]　原田 浩[2]　實本 英之[2]　佐藤 仁[3,5]　平川 学[4]

概要：「京」と全国の大学や研究所などに設置されている主要なスパコンを利用するための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）が整備されている．本研究では，HPCI におけるデータ共有基盤である HPCI 共用ストレージの構成を紹介し，性能評価を行う．HPCI 共用ストレージは Gfarm ファイルシステムを利用し，容量は東拠点 12.6 PB，西拠点 9.5 PB で総計 22.1 PB である．ファイル生成時間はもっとも遠い九州大学で 0.44 msec であった．IOPS はディレクトリ作成で最大 3,693 IOPS，ファイル作成で最大 445 IOPS であった．並列ファイルコピー性能は，北海道大学，京都大学からも 850 MB/s を越える性能であった．今後より少ない並列数で同等以上の性能を達成するために，バルク転送方式を検討している．評価では，京都大学から西拠点への 1 並列のコピー性能が 17 倍となり，有望であることが示された．

1. はじめに

広範な分野において計算科学技術を通じて実施された研究開発が，国際競争力の飛躍的な向上に継る研究成果を世界に先んじて創出するため，「京」と全国の大学や研究所などに設置されている主要なスパコンを利用するための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）が整備され，平成 24 年 11 月より運用が始まっている．

HPCI には HPCI 共用ストレージと呼ばれるデータ共有基盤がある．HPCI 共用ストレージは，HPCI 利用時の大量データの格納，コミュニティでのデータ共有，アーカイブを目的としたストレージである．国内の主要スパコン間での入出力データなどのデータ共有，スパコンで利用価値の高い共有データベースなどを保持する．HPCI 共用ストレージの整備については，HPCI システム整備検討 WG において，利用シナリオ，求められる特性，運用形態などの検討を行い，Gfarm ファイルシステム [6] を中核とするシステム設計などをまとめた [8]．

本論文では，HPCI におけるデータ共有基盤である HPCI 共用ストレージの構成を紹介し，性能評価を行う．

1 筑波大学システム情報系
2 東京大学情報基盤センター
3 東京工業大学学術国際情報センター
4 理化学研究所計算科学研究機構
5 JST CREST
a) tatebe@cs.tsukuba.ac.jp

2. 関連研究

HPCI と同様に，複数の計算センターの資源を共有する大規模な単一仮想システムに XSEDE と PRACE がある．XSEDE [2] は TeraGrid [4] の後継プロジェクトで，SDSC，PSC，TACC，NICS，Indiana 大学などで構成される．HPCI 共用ストレージと同様の複数の拠点からアクセス可能な広域ファイルシステムとしては，XSEDE-Wide ファイルシステム（XWFS）がある．XWFS は GPFS を用いて，物理的には，TACC，NICS，PSC，SDSC，NCSA に配置されたストレージを，単一のファイルシステムとして多くの拠点からマウントしてアクセスできるようになっている．ただし，2013 年 7 月の時点でまだ実運用されていない．

PRACE [3] は DEISA [1] の後継プロジェクトである．PRACE には同様の広域ファイルシステムは現在のところないようであるが，DEISA には RZG, LRZ, BSC, JSC, EPCC, HLRS などで構成される DEISA Global ファイルシステムがあった．DEISA Global ファイルシステムも GPFS を用い，物理的にはそれぞれの拠点に配置されたストレージを単一のファイルシステムとしてマウントしてアクセスできるようになっている．GPFS のクライアントがないプラットフォームからは GridFTP でアクセスする．DEISA Global ファイルシステムでは，パス名に explicit にサイト名が含まれており，格納場所を意識してアクセスすることができる．一方，HPCI 共用ストレージでは，パス名にはサイト名などの場所に関する情報は含まれない．

ファイルは自動的に近いサイトに配置され，また遠いサイトに複製を作成することもできるため，柔軟性，耐故障性，性能において有利であると考えられる．

Lustre ファイルシステムを遠隔のサイトからアクセスする研究は TeraGrid で多くなされた．その延長上の研究に，100 Gbps のネットワークを用いて，広域からアクセスする研究 [5] がある．理想的な環境では，ペンディングリクエストの数を調節することにより，高い性能が示されるが，実環境では複雑さが増え，単純には性能向上しないことが示されている．Lustre ファイルシステムは複数の遠隔拠点にストレージを分散させることができないため，ネットワーク障害やストレージの配置されている拠点のメンテナンス時などに利用できなくなる．また，Lustre ファイルシステムの場合は，各拠点からストレージ拠点までのネットワーク帯域を十分確保する必要があるが，HPCI 共用ストレージの場合は各拠点からどれかのストレージ拠点までのネットワーク帯域を確保すればよいという点で有利であると考えられる．

3. HPCI 共用ストレージ

HPCI 共用ストレージは，HPCI 利用時の大量データの格納，コミュニティでのデータ共有，アーカイブを目的としたストレージである．主な機能としては以下がある．

- 大容量のデータを格納し，場所を気にせずアクセス可能
- 複製やバックアップなどによる高信頼性の保持
- 継続的な保守とシステム増強が可能

これらの目的，機能を基に，HPCI システム整備検討 WG において，利用シナリオ，求められる特性，運用形態などの検討を行い，Gfarm ファイルシステム [6] を中核とするシステム設計などをまとめた [8].

3.1 特徴

HPCI 共用ストレージは，HPCI 参加機関をはじめどこからでもマウント可能な広域ファイルシステムである．主な特徴としては以下がある．

- Grid Security Infrastructure (GSI) によるシングルサインオン
- グローバルユーザ名とグローバルグループ
- 単一障害点なし

HPCI では，マシンにログインするために GSI を利用する [7]．HPCI 共用ストレージでは，広域に分散する百規模のストレージサーバに対し，クライアントーサーバ間，またサーバ間で通信を行うこととなるが，その認証，認可をシングルサインオンで安全に行うために GSI の権限委譲を利用する．

HPCI 共用ストレージは複数の拠点でマウントすることが想定されるが，それぞれの拠点においてローカルアカウント名は異なるため，グローバルなユーザ名へのマッピングが必要となる．HPCI 共用ストレージでは，各ユーザの GSI ユーザ証明書の ID をキーとして，HPCI-ID をグローバルユーザ名として利用する．これにより，ユーザはどこにいても同一ユーザとして扱うことができる．また，グローバルグループについては，グローバルユーザ名を用いて，HPCI 共用ストレージで管理する．

ストレージは，アクセスできなくなるとジョブ実行ができなくなるなど, 多大な影響を及ぼす．HPCI 共用ストレージは HPCI 全体についてのストレージであるため，その影響はさらに大きくなる．システムに単一障害点があると，そのリスクが大きくなってしまう．HPCI 共用ストレージは，メタデータサーバを 4 重化し，ファイルデータについてはデフォルトで二重化を行っている．またストレージ自体も RAID6 相当による冗長化とコントローラの冗長化がなされている．

3.2 構成

東京大学（東拠点）と AICS（西拠点）にそれぞれ 12.6 PB，9.5 PB の容量のストレージが配置され，全体で 22.1 PB の容量をもつ．**図 1** に論理的な詳細を示す．

HPCI 共用ストレージは Gfarm ファイルシステムで構成される．Gfarm ファイルシステムはメタデータサーバ（gfmd）とファイルシステムノード（gfsd）で構成され，gfmd は，東拠点，西拠点それぞれ 2 ノードずつ準備されている．そのうち，1 ノードがマスタ gfmd となり，同じ拠点の 1 ノードが同期型スレーブ gfmd，異なる拠点の 2 ノードが非同期型スレーブ gfmd となる．同期型スレーブ gfmd は，どのタイミングをとってもマスタ gfmd と等価なメモリ状態をもつことが保証され，いつでもフェイルオーバ可能である．非同期型スレーブ gfmd は，マスタ gfmd における更新が必ずしも反映されているとは限らないが，ディザスタリカバリを想定して準備されている．非同期スレーブ gfmd が 2 ノードあるのは対称性を保つためであり，どちらの拠点がマスタ gfmd となったとしても，同期型スレーブ gfmd を準備できるようになっている．

gfsd は論理的には東拠点に 64 サーバ，西拠点に 30 サーバある．物理的には，各ノードは 10Gbit Ethernet あるいは Infiniband QDRx4 で接続されるが, それぞれのノード上で 1〜4 の gfsd が起動している．それぞれのストレージに対するネットワークバンド幅は平均 2.5Gbps から 32Gbps と幅があり均一ではない．また，東拠点，西拠点のアップリンクはそれぞれ 30 Gbps，40 Gbps である．

クライアントは，北海道大，東北大，筑波大，東京大，東工大，名古屋大，京都大，大阪大，AICS，九州大の各計算センターに置かれる．現時点では，北海道大，東京大，東工大，京都大，AICS が 10Gbit Ethernet で接続されている．

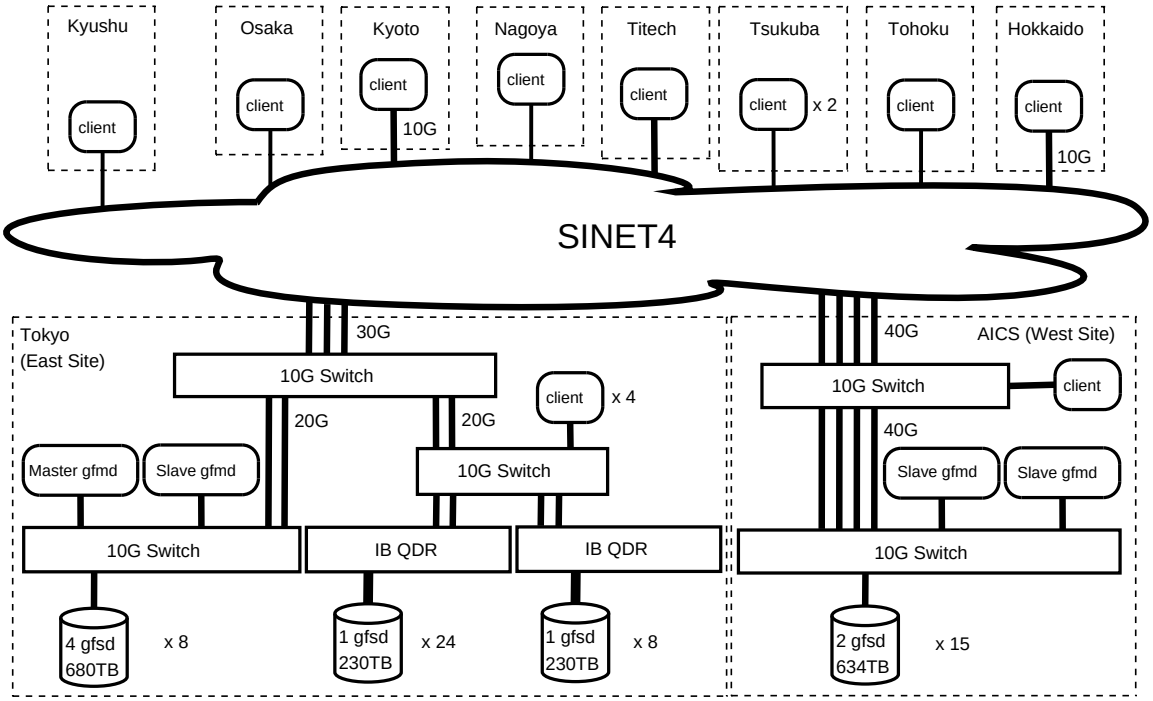

図 1 HPCI 共用ストレージの構成.

	ファイル生成 [sec]	東拠点 [msec]	西拠点 [msec]
北海道大	0.34	15.2	23.4
東北大	0.26	14.9	28.7
筑波大	0.11	6.1	15.8
東京大	0.02	0.16	10.7
東工大	0.04	3.2	12
名古屋大	0.27	9.0	6.8
京都大	0.26	14.2	6.0
大阪大	0.26	10.7	3.1
AICS	0.16	11.1	0.6
九州大	0.44	22	14.3

表 1 各拠点のファイル生成時間と各拠点から東拠点，西拠点までのネットワーク往復遅延時間.

4. 性能評価

HPCI 共用ストレージの性能評価にあたり，ファイル生成時間，ファイル操作の IOPS，並列ファイルコピーのバンド幅，ファイル複製性能を評価する．なお，全ての評価は HPCI 共用ストレージで運用中に行った．複数回計測を行い最良値を示しているが，さまざまな外乱の影響は完全には排除されていない.

4.1 ファイル生成時間

各拠点におけるファイル生成時間，また各拠点から東拠点，西拠点までのネットワーク往復遅延時間（RTT）を**表 1** に示す.

各拠点から東拠点までの RTT は，東工大は 3.2 msec，京都大，北海道大は 14〜15 msec，九州大は 22 msec であった．西拠点までの RTT は，大阪大は 3.1 msec，筑波大，九州大は 14〜15 msec，北海道大は 23.4 msec，東北大は 28.7 msec であった.

ファイル生成は，各拠点で HPCI 共用ファイルシステムをマウントし，単一のマウントポイントに対し，逐次的に 3 バイトファイルを 100 ファイル作成した．表 1 のファイル生成時間は，1 ファイル作成あたりの平均値を示している．ファイルの生成時間は，大きくメタデータ更新，書き込むファイルシステムノードのスケジューリング，ファイルデータの書き込みからなる．従って，東拠点に置かれたマスタ gfmd までの RTT と，東拠点，西拠点のファイルシステムノードまでの RTT の影響を受ける．東拠点から遠い北海道大学では 0.34 秒，九州大学では 0.44 秒であった．このファイル生成時間は，単一マウントポイントに対し逐次的にファイル作成した場合であるが，ファイル操作は並列に行うことができ，次節で並列にファイル操作したときの IOPS を評価する.

4.2 ファイル操作の IOPS

ファイル操作の IOPS (I/O operations per second) を評価するために，Gfarm ライブラリ API を並列に発行する gfiops ベンチマークを用いた．gfiops は Gfarm ファイルシステムのソフトウェアのディストリビューションに含まれ，各種 Gfarm API を並列に実行することにより，IOPS を計測する.

ディレクトリ作成を並列に行ったときの IOPS の評価結果を**図 2** に示す．東京大学，AICS のいずれも並列発行数

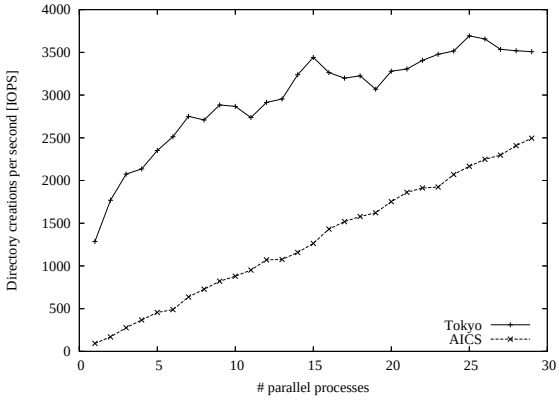

図 2 東京大学と AICS におけるディレクトリ作成性能.

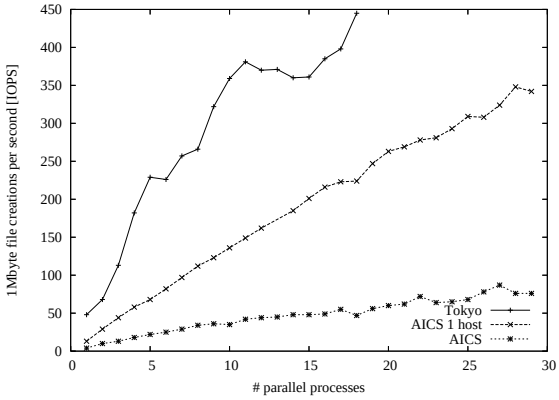

図 3 東京大学と AICS における 1M バイトファイルの作成性能.

を増やすとディレクトリ作成性能は向上している．ディレクトリ作成はメタデータ更新だけのファイル操作であるため，マスタ gfmd が近い東京大の立ち上がりが早く，最高で 3,693 IOPS を達成している．AICS は 29 並列で 2,495 IOPS を達成しているが，性能はまだ飽和していないと考えられる．

1M バイトファイルを並列に作成したときの IOPS の評価結果を**図 3** に示す．この評価は，ディレクトリ作成と違い，メタデータの更新と，ファイルデータの書込が必要となる．ファイルデータの書込については，94 ファイルシステムノードから書き込み先のスケジューリングを行い，実際に書き込みを行う．スケジューリングは，クライアントからネットワーク距離とファイルサーバの CPU 負荷情報を利用して行う．これらの情報は各クライアントが実行時に計測するが，一度計測した情報はデフォルトで 10 分間キャッシュされる．東京大学の性能については，IOPS が高く，外乱の影響が表れやすくなるため，性能は安定していない．18 並列で 445 IOPS を達成している．AICS については，並列数を増やすと性能は向上しているものの，プロセスあたり 3〜5 IOPS の向上となっている．プロセスあたりの性能が落ちている大きな原因は gfmd までのネットワーク遅延と，ファイルシステムノードのスケジューリングオーバヘッドが考えられる．特に，HPCI 共用ストレージではファイルシステムノード数が 94 と多いため，スケジューリングのオーバヘッドが大きくなる．この影響を調べるために，書き込みホストを 1 ホストに限定して計測した．結果を図中の AICS 1 host に示す．プロセスあたり 12〜15 IOPS となり，性能が向上している．ただし，書き込みホストを 1 ホストに限定することによりスケジューリングオーバヘッドはなくなるが，一方で並列クライアントから 1 ホストへファイル書き込みが集中するため，ファイル書込性能は落ちていると考えられる．

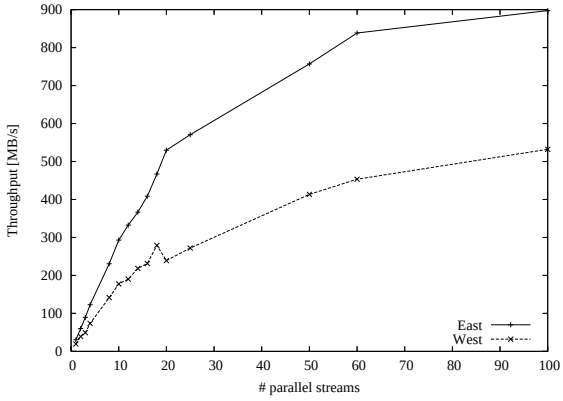

図 4 北海道大学から HPCI 共用ストレージへの並列ファイル書き込み性能.

4.3 並列ファイルコピーのバンド幅

HPCI 共用ストレージは，各センターのスパコンで利用する大容量データの共有が主目的である．スパコンで利用するためには，スパコンのファイルシステムと HPCI 共用ストレージ間でのファイルコピーが必要となる．ファイルコピーは，HPCI 共用ストレージをマウントして，UNIX の cp コマンドでもできるが，この方法ではファイルコピーが逐次的になされ性能を出すことができない．並列にファイルコピーを行うためには gfpcopy コマンドを利用する．

北海道大学から HPCI 共用ストレージへ並列ファイルコピーする性能を**図 4** に示す．東拠点への書き込みは 20 並列までは 1 並列あたり 30 MB/s で性能が伸びているが，その後性能の伸びが鈍っている．100 並列では 898 MB/s であった．西拠点へは 10 並列までは 1 並列あたり 20 MB/s で性能が伸びているがその後鈍っている．100 並列では 532 MB/s であった．

京都大学から HPCI 共用ストレージへ並列ファイルコピーする性能を**図 5** に示す．東拠点への書き込みは北海道大学の場合と同様に 20 並列までは 1 並列あたり 30 MB/s で性能が伸びているが，その後性能の伸びが鈍っている．100 並列では 847 MB/s であった．西拠点へもほぼ同様の傾向であり，100 並列では 828 MB/s であった．

東京大学から HPCI 共用ストレージへ並列ファイルコピーする性能を**図 6** に示す．東拠点への書き込みはマス

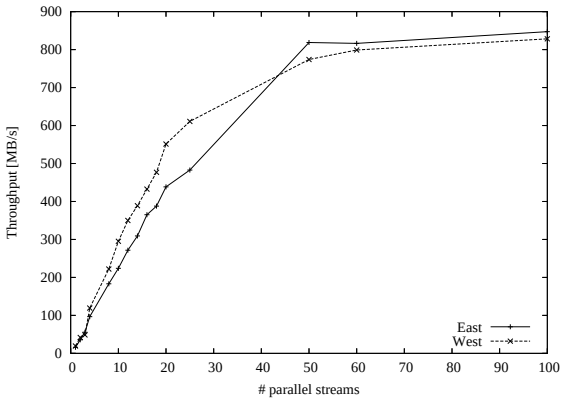

図 5 京都大学から HPCI 共用ストレージへの並列ファイル書き込み性能.

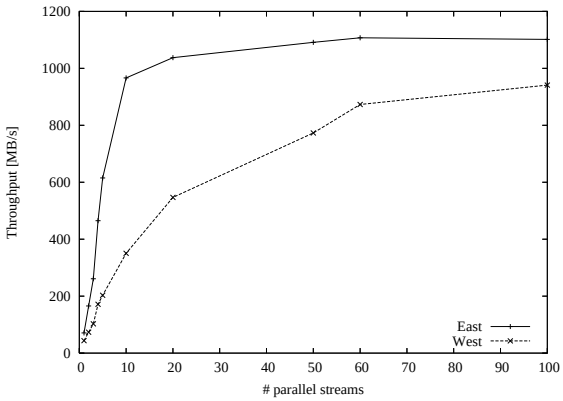

図 6 東京大学から HPCI 共用ストレージへの並列ファイル書き込み性能.

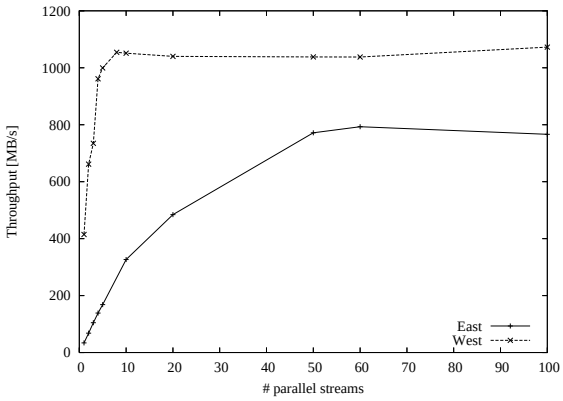

図 7 AICS から HPCI 共用ストレージへの並列ファイル書き込み性能.

ター gfmd およびファイルシステムノードと同じ拠点ということもあり，立ち上がりが早くなっている．1 並列あたり 70〜120 MB/s の性能で，最高では 1,107 MB/s であった．西拠点へは 20 並列までは 1 並列あたり 30〜40 MB/s で性能が伸びているがその後若干鈍っている．100 並列では 941 MB/s であった．

AICS から HPCI 共用ストレージへ並列ファイルコピーする性能を**図 7** に示す．西拠点へは 10 並列までは 1 並列

	東拠点		西拠点	
	RPC	バルク転送	RPC	バルク転送
北海道大	30.5	60.7	19.3	81.0
京都大	15.9	63.4	19.3	326

表 2 1M ブロックの RPC とバルク転送方式による各拠点からのファイル転送性能 [MB/s].

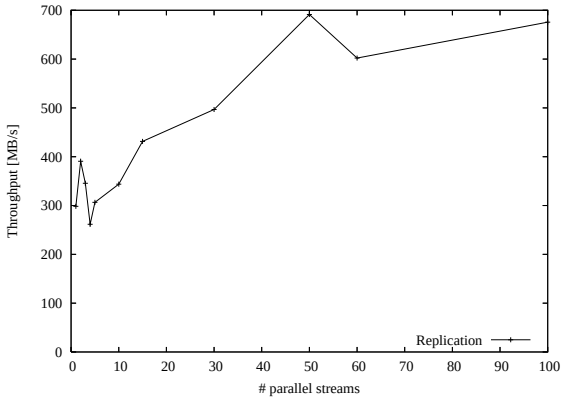

図 8 東拠点から西拠点へのファイル複製作成性能.

あたり 30〜40 MB/s で性能が伸びているがその後若干鈍っている．100 並列では 766 MB/s であった．西拠点への書き込みはファイルシステムノードと同じ拠点ということもあるが，立ち上がりは東京大学より早くなっている．1 並列あたり 200〜400 MB/s の性能で，8 並列で 1,000 MB/s を越え，最高では 1,073 MB/s であった．

HPCI 共用ストレージでは，現在 Gfarm ファイルシステム 2.5.8 を利用しており，gfpcopy のファイルコピーは 1M バイトブロックごとのブロッキング RPC で実現されている．この方式では，ネットワーク遅延が大きくなったときに 1 並列あたりのファイル転送性能が落ちてしまう．実際，北海道大学から西拠点へのファイルコピーでは 1 並列あたり 20 MB/s 程となっている．この性能を改善するため，ブロッキング RPC ではなく，FTP のデータ転送と同様のバルク転送を利用した場合の性能を**表 2** に示す．バルク転送では，送信側はデータを連続して送信することが可能であるため，RPC における ack 待ちがない．また，連続して送信できるため，TCP の輻輳ウィンドウサイズも大きくなる．北海道大学では，東拠点への性能は 2.0 倍，西拠点への性能は 4.2 倍となった．京都大学では，東拠点へは 4.0 倍，西拠点へは 17.0 倍と大幅に向上した．

4.4 複製作成性能

ファイル複製はファイル生成時などに自動的に作成されるが，本節では，その複製作成性能を評価する．自動的に作成される場合は，東拠点，西拠点のどのファイルシステムノードに作成されるか分からないが，評価においては，東拠点から西拠点に複製を作成する性能を評価する．

並列数を変えてファイル複製性能を評価したものを**図 8**

に示す．並列数が 10 まではファイル複製作成性能は 300〜400 MB/s 程であるが，50 並列になると 700 MB/s 程の性能となり，最高で 691 MB/s であった．ファイル複製は前節におけるバルク転送のプロトコルを用いており，1 並列でも高い性能となっている．並列数を増やしても性能が伸びていないことについては，外乱など複雑な要因が考えられるが，詳細は今後調査予定である．

5. まとめ

HPCI 共用ストレージの構成と性能についてまとめた．HPCI 共用ストレージは Gfarm ファイルシステムを利用している．メタデータサーバは東拠点と西拠点にそれぞれ 2 ノードあり，4 重に冗長化を図っている．ファイルシステムノードは東拠点に 64 サーバ，西拠点に 30 サーバで構成され，容量は東拠点 12.6 PB，西拠点 9.5 PB で総計 22.1 PB である．ファイルデータはデフォルトで二重化されている．

本研究では，ファイル作成時間，IOPS，並列ファイルコピー性能，複製作成性能について評価を行った．ファイル生成時間はもっとも遠い九州大学で 0.44 msec であった．IOPS はディレクトリ作成などメタデータ操作だけのものは最大 3,693 IOPS，ファイル作成などファイルデータ操作も含むものは最大 445 IOPS であった．クライアントが遠隔のときは，並列数をあげるとこの値に近付くと考えられる．並列ファイルコピー性能は，東拠点，西拠点から遠い北海道大学，京都大学からも 850 MB/s を越える性能であった．この性能は HPCI 共用ストレージにとりとても重要な性能であり，今後より少ない並列数で同等以上の性能を達成することが望まれる．複製作成性能は東拠点から西拠点への複製で最高 691 MB/s であった．ただし，複製作成性能については並列数をあげても性能が単純に向上しないため，より詳しい調査が必要である．

並列ファイルコピー性能を向上させるために，バルク転送を用いることが有効であることが分かった．評価では，京都大学から西拠点への書き込みでは性能が 17 倍となった．本方式は，今年度リリース予定の Gfarm ファイルシステム 2.6 で導入される予定である．

謝辞 本研究を遂行するにあたりご議論頂きました「HPCI の詳細仕様に関する調査検討」，ならびに HPCI システム整備検討 WG 委員の関係各諸氏に感謝いたします．

参考文献

[1] Distributed European Infrastructure for Supercomputing Applications. http://www.deisa.eu/.
[2] Extreme Science and Engineering Discovery Environment. https://www.xsede.org/.
[3] Partnership for Advanced Computing In Europe. http://www.prace-ri.eu/.
[4] Catlett, C., Allcock, W. E., Andrews, P., Aydt, R., Bair, R., Balac, N., Banister, B., Barker, T., Bartelt, M., Beckman, P., Berman, F., Bertoline, G., Blatecky, A., Boisseau, J., Bottum, J., Brunett, S., Bunn, J., Butler, M., Carver, D., Cobb, J., Cockerill, T., Couvares, P. F., Dahan, M., Diehl, D., Dunning, T., Foster, I., Gaither, K., Gannon, D., Goasguen, S., Grobe, M., Hart, D., Heinzel, M., Hempel, C., Huntoon, W., Insley, J., Jordan, C., Judson, I., Kamrath, A., Karonis, N., Kesselman, C., Kovatch, P., Lane, L., Lathrop, S., Levine, M., Lifka, D., Liming, L., Livny, M., Loft, R., Marcusiu, D., Marsteller, J., Martin, S., McCaulay, S., McGee, J., McGinnis, L., McRobbie, M., Messina, P., Moore, R., Moore, R., Navarro, J., Nichols, J., Papka, M. E., Pennington, R., Pike, G., Pool, J., Reddy, R., Reed, D., Rimovsky, T., Roberts, E., Roskies, R., Sanielevici, S., Scott, J. R., Shankar, A., Sheddon, M., Showerman, M., Simmel, D., Singer, A., Skow, D., Smallen, S., Smith, W., Song, C., Stevens, R., Stewart, C., Stock, R. B., Stone, N., Towns, J., Urban, T., Vildibill, M., Walker, E., Welch, V., Wilkins-Diehr, N., Williams, R., Winkler, L., Zhao, L. and Zimmerman, A.: TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications, *High Performance Computing and Grids in Action* (Grandinetti, L., ed.), Advances in Parallel Computing, Vol. 16, IOS Press, pp. 225–249 (2008).
[5] Henschel, R., Simms, S., Hancock, D., Michael, S., Johnson, T., Heald, N., William, T., Berry, D., Allen, M., Knepper, R., Davy, M., Link, M. and Stewart, C.: Demonstrating Lustre over a 100Gbps Wide Area Network of 3,500km, *Proceedings of IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC12)* (2012).
[6] Tatebe, O., Hiraga, K. and Soda, N.: Gfarm Grid File System, *New Generation Computing*, Vol. 28, No. 3, pp. 257–275 (2010). DOI: 10.1007/s00354-009-0089-5.
[7] 合田憲人，　東田学，坂根栄作，天野浩文，小林克志，棟朝雅晴，江川隆輔，建部修見，鴨志田良和，滝澤真一朗，　永井亨，岩下武史，　石川裕：高性能分散計算環境のための認証基盤の設計，情報処理学会論文誌コンピューティングシステム (ACS)，Vol. 5, No. 5, pp. 90–102 (2012).
[8] 實本英之，建部修見，　佐藤仁，　石川裕：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，研究報告ハイパフォーマンスコンピューティング (HPC)，Vol. 2011-HPC-130, No. 67, 情報処理学会，pp. 1–6 (2011).