Vol.2012-HPC-135 No.29
2012/8/2

Infinibandを用いたファイルアクセスの高速化

大辻 弘貴[1,3]　建部 修見[2,3]

概要：大規模な共有ファイルシステムを構成する際，アプリケーションはネットワークを経由してファイルアクセスを行うが，ネットワークがボトルネックとなり，記憶装置の性能を十分に活かせない場合がある．特にレイテンシに関しては，通常の Ethernet 等を利用した場合，様々なレイヤが障壁となり，昨今のフラッシュメモリを用いたストレージの性能を引き出すことが出来ない．そこで，低遅延・高帯域を特長とする Infiniband を，共有ファイルシステムのネットワークとして使用し，その性能を最大限に活用するための検討および評価を行った．Infiniband の特長の一つである RDMA(Remote Direct Memory Access) を，ハードウェアに近い低レイヤーの API 群を介して利用した結果，ローカルストレージに対するアクセスに迫る性能を引き出す事が出来た．本稿ではその実装と評価，考察について述べ，ネットワークが記憶階層の中で性能面からどのような位置付けにあるか示し，今後の性能向上について検討する．

1. 序論

今後のデータ量増加に備えるため，複数のコンピュータに渡った大容量ストレージシステムに対する高速なアクセス手法が求められている．特に，データインテンシブコンピューティングや e-サイエンスの発展により，扱われるデータ量は数年後にはエクサバイト単位になると見込まれており，従来のアーキテクチャやシステムでは要求を十分に満たせない可能性がある．現状では，このような要求に対して，ネットワークで接続された分散ファイルシステムなどが用いられており，その構成は広域に分散したものや集中型のものがある．前者については，ファイル複製 [1] といった高速化手法が採られている．これは，広域分散環境ではネットワークの遅延が大きいことから，より近くにデータを移動してからアクセスしたほうが性能面で有利であることが背景にある．また，筆者らのこれまでの研究では，このような環境に対して，性能を向上するための手法を提案してきた．一方で一カ所の拠点に集中してシステムを構築する場合，ボトルネックがネットワークそのものの性能や，OS・ドライバレベルの遅延といった部分に移動するため，従前のような方法では性能向上が望めない．本稿では，これまでの研究についても紹介すると共に，高速なネットワークの一つである Infiniband[2] を用い，その重要な機能である RDMA (Remote Direct Memory Access) を活用した低レイテンシ・高帯域のネットワークファイルアクセスについて検討する．具体的には**図 1** に示すような構成において，40〜100Gbps の帯域，1〜10 μ 秒のレイテンシをめざし，様々なレイヤーやデバイスにおける性能特性の把握および，最適化手法の可能性を探ることを目標としている．

[1] 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
[2] 筑波大学システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba
[3] 独立行政法人科学技術振興機構 CREST
JST CREST

2. 関連研究

本研究の対象はネットワークを経由したファイルアクセスであり，その関連研究は多岐にわたる．遠隔ファイルアクセスには 2 つの種類に大分することができる．この 2 つとは，RPC によって逐一リクエストを発行する方式と，HTTP 等に見られるファイル全体を転送する方式である．前者には Gfarm[3] ファイルシステムや NFS[4]，Lustre[5] などを例として挙げられる．後者の方式には，AFS[6] やその後継である Coda[7]，FTP をグリッド環境向けに拡張した GridFTP[8] がある．また，これらの性能最適化にあたって，アクセスパターンの分類などが必要になることがあるが，[9] はその記述法に関しても言及している．本稿においては，逐次リクエストを発行する前者の方式がターゲットであり，これまでの研究である通常のネットワーク環境を対象とした最適化についての紹介および，より高速なネットワークを利用したファイルアクセスについて述べる．高速なネットワークには，RDMA 機能を備えた物があ

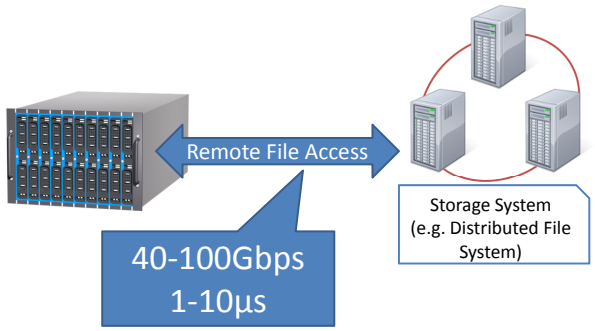

図 1 システム構成・目標とするレイテンシと帯域幅

り，これらは CPU を介さずにリモートメモリ転送を行う事が出来る．RDMA を活用して通信を行う事により，通常の IP による通信に比べて低遅延・高帯域の通信が実現出来る．既に，いくつかの分散ファイルシステム上で通信に RDMA を使用した例が存在し，例えば [10] は，PVFS[11] で RDMA 転送を適用し，低 CPU 負荷および高性能化を達成している．[12] は，Infiniband RDMA て転送を行う際，ゼロコピーを実現するための手法について述べている．

3. 遠隔ファイルアクセスと Infiniband

3.1 遠隔ファイルアクセスの方法

前述の通り，本研究は RPC によって逐次ファイルアクセスを行う場合を前提としている．**図 2** は RPC による遠隔ファイルアクセスを行う場合のシステム構成と動作を示している．最下段はクライアントコンピュータ上で動作するアプリケーションを示しており，ファイルアクセスのリクエスト元である．リクエストは中段のファイルシステムのクライアントに送られ，ネットワークを介してファイルサーバに対して実際のデータアクセスを行う．その際，ファイルシステムのクライアントは上段のファイルサーバに対して，RPC を発行してファイルを要求する．サーバはこの RPC に応じる形でデータをクライアントに返す．このデータは一旦バッファに書き込まれ，最下段のアプリケーションはそこからデータを得る．

この一連の手順により，遠隔ファイルアクセスが実現される．これらは POSIX に準拠するインターフェスでアプリケーションに提供されており，通常のローカルファイルシステムと同じように利用できるのが特徴である．本稿では，このクライアントとサーバのネットワーク部分に Infiniband を用い，転送に RDMA を利用する．ネットワークを介して実行される RPC には大きく分けて，非同期型と同期型の 2 種類 (**図 3**) があるが，本稿の評価では全て同期型を用いている．

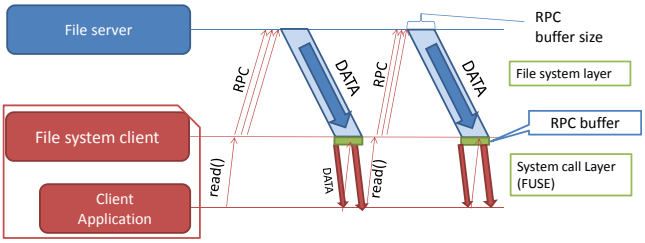

図 2 RPC による遠隔ファイルアクセス

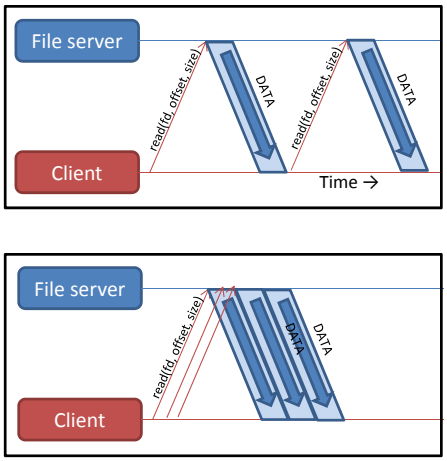

図 3 同期型 RPC(上) と非同期型 RPC(下)

3.2 Infiniband RDMA

序論でも述べた通り，Infiniband では RDMA を利用することができる．これは，リモートコンピュータのメモリに対して直接転送を行える仕組みである．メリットとしては非常に低レイテンシ (RTT 2 μ sec) であることや，バンド幅を最大限活用できる点が挙げられる．

Infiniband(RDMA) を利用する方法はいくつかあり，それぞれ実装コストや性能に差がある．以下に 3 つの特性と説明を記す．

最も簡単な方法は IP over IB(IPoIB)[13] を用いる方法である．これは，IP を Infiniband 上に流す方法であり，通常の Ethernet と同じように利用することが出来る．しかしながら，最もレイテンシが大きく（約 30〜100 μ sec），帯域も出にくい．これは，Infiniband がフロー制御などの機能をハードウェアレベルで備えているにも関わらず，OS の IP スタックなどを利用しているため，オーバーヘッドが生じている事に起因している．

次に挙げられる方法は Socket Direct Protocol(SDP) を用いる方法である．この方法も IPoIB と同じく，プログラムに対してほとんど変更が要らず，ライブラリの差し替えまたはソケットオプションの変更で簡単に利用できる．SDP は IPoIB に比べるとパフォーマンス面で優れて

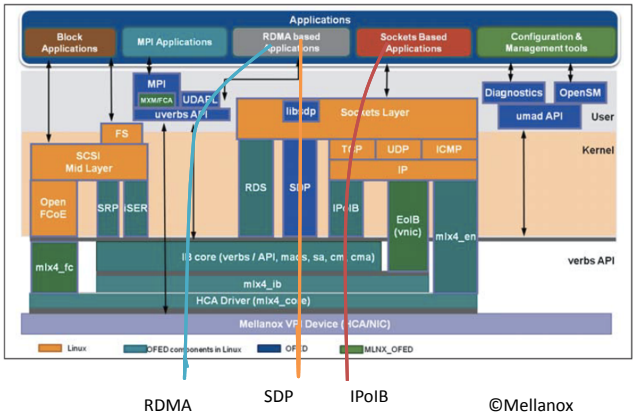

図 4 OFED コンポーネントの構成図 (Mellanox[15] のドキュメントより引用)

おり，レイテンシは約 6 μ sec と大幅に小さくなる．これは，IPoIB でオーバーヘッドの原因となっていたカーネルによる制御が省かれるためである．

最後に，RDMA を低いレイヤの API で利用する方法がある．Infiniband を利用するためのソフトウェア群である OFED[14] では，Verbs と呼ばれる API 群が提供されている．上記 2 つの方法とは異なり，これは全く異なるインターフェースで通信を行う．そのため，通常の Socket プログラムをそのまま利用することは出来ず，通信に関わる部分を全て書き直す必要がある．一方でパフォーマンスの面では非常に優れており，2 μ sec の転送時間，QDR で 25Gbps を超える通信速度を実現できる．実装にあたっては，RDMA に利用するメモリ領域を予め登録し，転送命令を発行することで RDMA 書き込み/読み込みを行う．やりとりに際してはいくつかの方式があるが，本稿では RC (Reliable Connection) を使用しており，通信の到達性・順序が保証されている．

従って，実装コストを抑えながらもパフォーマンスを出すためには SDP が最適で，RDMA による恩恵を最大限に受けるためには最後の方法を用いるのが良い．図 4 は，OFED のコンポーネントマップにそれぞれの方法に対応した線を引いたものである．この図からも，IPoIB が多くの層を通過し，SDP では一部がバイパス，Verbs API では多くの部分が素通しになっている事が分かる．

3.3 Infiniband のネットワーク性能

RDMA を使用した Infiniband ネットワークの性能を調べるため，予備評価を行った．

図 5 は，前節で述べた 3 つの方法によるレイテンシ比較である．x 軸は転送データサイズを表し，y 軸はレイテンシ (μ sec) を表している．IPoIB が最もレイテンシが大きく，RDMA(verbs) の場合が最もレイテンシが小さくなっている事が分かる．この傾向はいずれの転送サイズにおいても変わりない．

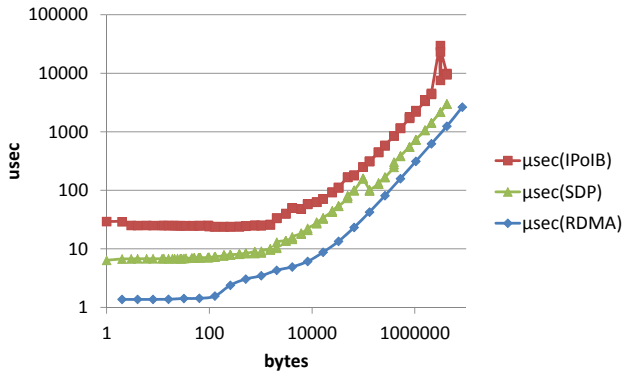

図 5 IPoIB，SDP，verbs(RDMA) の転送時間比較　x 軸は転送サイズ y 軸は時間

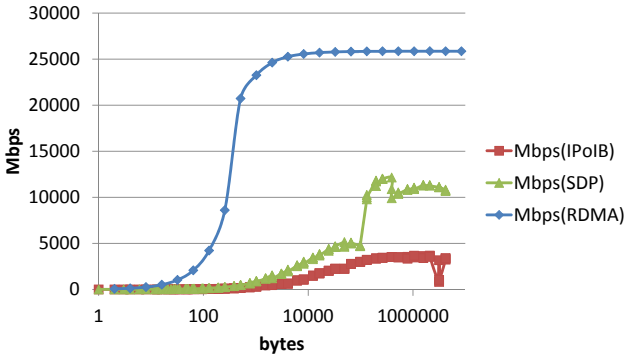

図 6 IPoIB と SDP，verbs(RDMA) のバンド幅比較　x 軸は転送サイズ y 軸はスループット

図 6 は，同じように 3 つの方法による帯域幅比較である．x 軸は同じく転送データサイズ，y 軸はスループット (Mbps) を表している．これは定義としては前の図の逆数であるが，特に注目すべき点は，RDMA(verbs) において，2KB の転送であっても，バンド幅をほぼ完全に使い切れている事である．これまで高遅延環境（msec オーダ）に対して高速化を行ってきたが，それらのボトルネックが十分に解消することになる．

3.4 遠隔ファイルアクセスへの適用

Infiniband RDMA の低レイテンシ・高帯域通信を遠隔ファイルアクセスに適用するにあたって，ストレージシステムの階層構造において，ネットワーク性能が相対的にどのように位置付けられるかを検討した．特にレイテンシは遠隔ファイルアクセスにおいて重要な要素であり，**図 7** は現在用いられている記憶装置と各種ネットワークをその観点から比較したものである．今回取り扱う Infiniband は高速なフラッシュメモリストレージのレイテンシに匹敵する事が分かる．フラッシュメモリベースのストレージは今後も普及が進むことが期待されるため，このような高速なネットワークで接続し，性能を引き出す事には大きな意義がある．

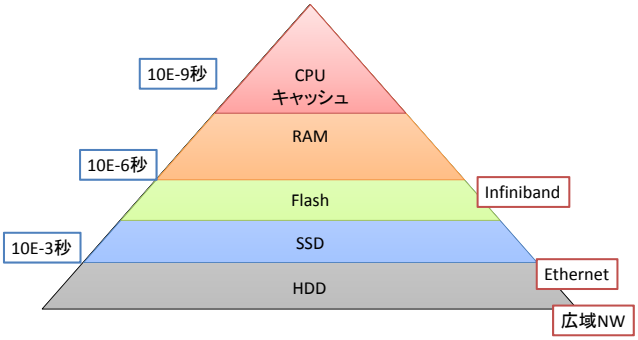

図 7　記憶階層とネットワークのレイテンシ

4. 性能評価

4.1 RDMA を利用した遠隔ファイルアクセスの評価

RDMA を利用した遠隔ファイルアクセスの性能を評価するために，**図 8** に示す評価システムを構成した．クライアント側のコンピュータはサーバに対して，ファイルデスクリプタとオフセットを指定してデータを要求する．サーバはクライアントの要求に応じて，対応するデータをストレージから読み込み，メモリに格納後クライアントに対して RDMA 転送を行う．

評価は, Mellanox 社製 Infiniband QDR アダプタ・Fusion-io ioDrive(160GB) を搭載した 8 コアのマシン 2 台を用いて行っている．アクセスする対象のファイルは ioDrive 上に配置している．また，評価を行う度にカーネルのキャッシュをクリアしている．

リモートアクセスとローカルアクセスそれぞれについて，それぞれ比較評価を行った．

結果を，**図 9** と**図 10** に示す．図 9 はシーケンシャルアクセスを行った際のスループット評価である．横軸は，一度の RDMA で転送したサイズ（バイト）を表し，縦軸はスループットである．赤い線が Verbs API を利用した RDMA によるスループット，青い線はローカルの ioDrive に対してアクセスした場合である．ローカルのスループットでは約 760MB/s が最大であった．一方，RDMA によるスループットは最大で約 600MB/s であった．また，その最大性能は，転送サイズが 16KB から 128KB の間に見られた．

図 10 は，あるサイズのデータを一定間隔のシーク幅 (64KB) で転送した場合に，1 秒間に何回アクセスできたか (IOPS) を示したものである．横軸が読み込みサイズで RDMA 転送サイズと等しく，縦軸は IOPS を表している．赤は Verbs API による RDMA 転送のスループットを示し，緑はローカルの ioDrive に対してアクセスした場合である．いずれの場合も，転送サイズが最も小さい 2KB の場合に最大性能を示し，その後サイズが増加すると共に徐々に数値が低下する．

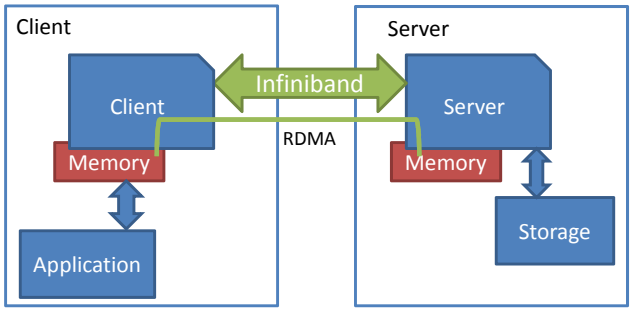

図 8　RDMA による遠隔ファイルアクセスの評価システム構成

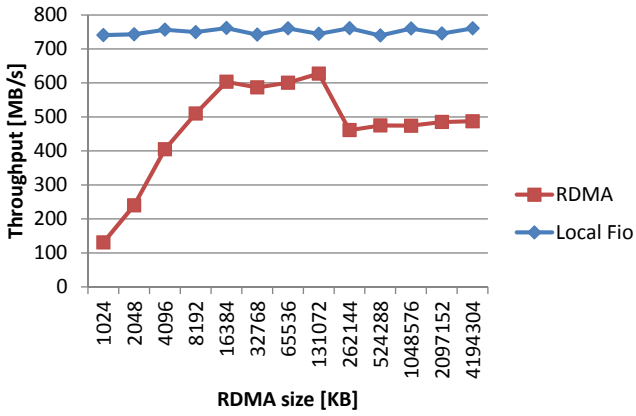

図 9　ioDrive に対する RDMA 転送とローカルアクセスのスループット比較（シーケンシャルアクセス）

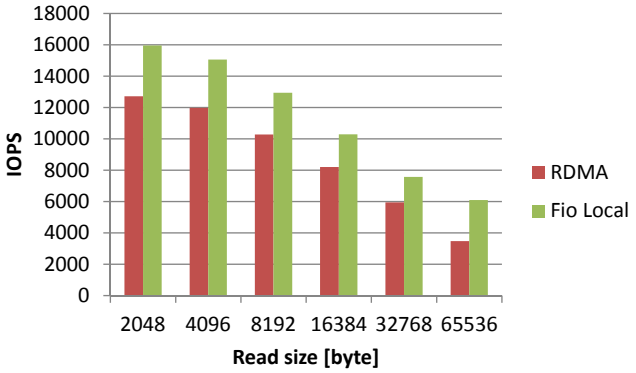

図 10　ioDrive に対する RDMA 転送とローカルアクセスの IOPS 比較（ストライドアクセス）

4.2 考察

前節の評価について考察する．まず，図 9 については，ローカルアクセスの性能から，ioDrive の最大スループットは約 760MB/s であると考えられる．RDMA によるリモートアクセスは最大で約 620MB/s であった．性能差が最も最小であったのは転送サイズが 128KB の場合で，RDMA がローカルアクセス比 84.3 サイズ毎の性能の推移で，1KB から 128KB までは性能が向上している．これはオーバーヘッドの影響が徐々に緩和された結果であると考えられる．しかしながら，256KB 以降は性能が低下している．RDMA のみの転送性能の評価ではこのような現象は起きていない

ことから，ローカルアクセスとメモリコピーの間でこのような性能低下が生じていると考えられる．今回の評価プログラムは，一旦転送サイズ分のデータを読み込み，メモリへコピーしてから転送している．従って，ある程度転送サイズが大きくなると，前回の読み込みで生じた OS のページキャッシュが活用されにくくなり，ローカル I/O と転送のオーバーラップ比率が減少している可能性がある．これに関しては，まだ検討の余地が大いにあるが，ボトルネックがネットワークからローカル I/O に移動していることは注目すべき点である．これまではネットワークに対しての最適化が性能に与える影響が多くを占めていたが，このような状況下では，ローカルの I/O を含めて様々な部分で最適化を図らなければならないことを意味している．

図 10 については，転送サイズが大きくなると IOPS が単調に減少している事から，転送時間の影響を受けている事が分かる．しかしながら，転送サイズと IOPS が必ずしも比例しておらず，特に小さな転送サイズの場合は変化が少ないことから，レイテンシの影響を大きく受けていると考えられる．この比較では，いずれも実効バンド幅は図 9 の数値に達していないことも，その理由の一つである．この遅延は．ioDrive と RDMA 双方により生じている．そのどちらが大きな影響を持っているかについては，RDMA の予備評価（**図 5**）とこの図から分かる．RDMA は最小 6 μ 秒程度のレイテンシであるのに対して，ioDrive は最大 IOPS が 16,000 すなわち 1 回あたり約 62 μ 秒を要しており，ioDrive のレイテンシが支配的である．しかしながら，ローカルと RDMA による転送の場合で一定の差が存在し続けていることから，ioDrive からメモリの転送などの部分について，最適化の必要がある．この事を確かめるため，以下に示す条件で性能のシミュレーションを行った．**図 11** にその結果を示す．グラフの軸やデータの種類は図 10 と同じである．これは，ストレージの遅延を 60 μ s，ネットワークの遅延を 7.5 μ s，ストレージバンド幅を 800MB/s，RDMA バンド幅を 1200MB/s と仮定した場合の性能をシミュレーションしたものである．1 回の操作にかかる時間を，ローカルの場合はストレージ遅延とデータサイズに応じた転送時間，RDMA の場合はそれに加えてネットワークのレイテンシを加算し，その逆数を IOPS として計算した．傾向としては図 10 に近く，以上に考察した状況が実際と近いことが予想される．

5. まとめ

本稿では，高速なストレージデバイスを，Infiniband の RDMA 機能によってリモートアクセスする際の性能および最適化の余地について検討した．第 3 章では，Infiniband のネットワークとしての性能の予備評価を行った．

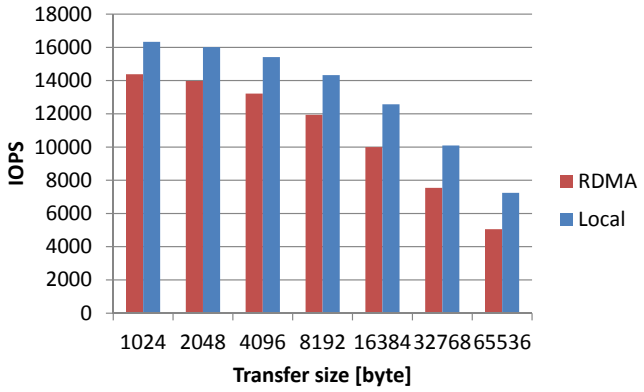

図 11 ローカルアクセス対 RDMA 転送の性能シミュレーション (IOPS)

第 4 章では，実際にストレージに対して遠隔ファイルアクセスを行い，その性能評価と性能特性についての考察を行った．

今後の課題としては，評価において明らかになった性能低下を改善するための最適化，ベンチマークプログラム以外の実環境における性能評価に取り組むことが挙げられる．

謝辞 本研究の一部は，JST CREST「ポストペタスケールデータインテンシブサイエンスのためのシステムソフトウェア」による．

参考文献

[1] Chervenak, A. L., Foster, I. T., Kesselman, C., Salisbury, C. and Tuecke, S.: The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets, *JOURNAL OF NETWORK AND COMPUTER APPLICATIONS*, Vol. 23, pp. 187–200 (1999).
[2] Infiniband Trade Association: Infiniband, *http://www.infinibandta.org/*.
[3] Tatebe, O., Hiraga, K. and Sod, N.: New Generation Computing, Ohmsha, Ltd. and Springer, *Gfarm Grid File System*, Vol. 28, No. 3, pp. 257–275 (2010).
[4] Callaghan, B., Pawlowski, B. and Staubach, P.: NFS Version 3 Protocol Specification, *RFC 1813* (1995).
[5] Braam, P. J.: Lustre, *http://www.lustre.org/*.
[6] Howard, J. H.: Scale and performance in a distributed file system, *ACM Trans. Computer Systems*, Vol. 6, No. 1, pp. 51–81 (1988).
[7] Satyanarayanan, M.: Coda: A Highly Available File System for a Distributed Workstation Environment, *IEEE Trans. Computers*, Vol. 39, No. 4, pp. 447–459 (1990).
[8] Allcock, W.: GridFTP: Protocol Extensions to FTP for the Grid, *Global Grid Forum Draft* (2003).
[9] Jun He, e. a.: Pattern-Aware File Reorganization in MPI-IO, *PDSW* (2011).
[10] Wu, J., Wyckoff, P. and Panda, D.: PVFS over InfiniBand: Design and Performance Evaluation (2003).
[11] Carns, P. H., Iii, Ross, R. B. and Thakur, R.: Pvfs: a parallel file system for linux clusters, *In ALS 00: Proceedings of the 4th annual Linux Showcase and Confer-*

ence (2000).
[12] Koop, M. J., Sur, S. and P, D. K.: Zero-Copy Protocol for MPI using InfiniBand Unreliable Datagram.
[13] Kashyap, V.: IP over InfiniBand (IPoIB) Architecture, *RFC 4392* (2006).
[14] OpenFabrics: OFED, *http://beany.openfabrics.org/*.
[15] Mellanox Technologies: ., *http://www.mellanox.com*.