2007－HPC－111 (39)
2007／8／3

クラスタシステムにおけるIP-SANを用いたI/O処理の並列ベンチマークによる評価

神坂 紀久子[†1] 山口 実靖[†2]
小口 正人[†1] 松岡 聡[†3,†4]

スーパーコンピュータや大規模なHPCクラスタシステムにおいて，ノード間通信とストレージアクセスのネットワークの統合が実現され始めている．このような統合ネットワークを提供することにより，大規模なクラスタシステムにおいても構成を単純化し，高いストレージアクセス性能を実現しながら運用管理コストを削減できる．しかし，ノード間通信とストレージアクセスのネットワークを統合することがシステム全体の性能にどのように影響するかについては明らかにされていない．本稿では，ノード間通信とストレージアクセスの統合ネットワークの一つとして，IPをベースにしたSANであるIP-SANを用い，I/Oを伴う並列ベンチマークによる評価を行った．その結果，iSCSI接続のIP-SAN統合型クラスタは，ローカルストレージのみを用いたクラスタの並列処理性能とほぼ同程度となった．このことより本評価においては，IP-SANにおけるネットワーク転送処理ではなく，ノード間通信を含む並列処理もしくはストレージのI/O処理がシステム全体の性能を決めるボトルネックとなっていることがわかった．

Evaluation of I/O Performance of IP-SAN on Cluster System using Parallel Benchmark

Kikuko Kamisaka,[†1] Saneyasu Yamaguchi,[†2] Masato Oguchi[†1] and Satoshi Matsuoka[†3,†4]

In Supercomputing and large scale HPC clusters, cluster system integrating connection networks of computing nodes and storage is beginning to be realized. Such cluster system simplifys network composition and reduce its costs. However, it is not clarified how the integration affects total performance of the system. In this paper, as one of the integrated networks, the cluster system connected with IP-SAN is evaluated by using parallel benchmark with I/O. In consequence, the performance of IP-SAN integrated cluster is about the same as that of cluster using local storage. According to the result, the bottleneckof en tire system's performance should be parallel processing and/or I/O processing of storage, rather than data transfer processing of networkstorage.

1. はじめに

近年，プロセッサの性能向上やネットワークの高速化，コモディティ化が急速に進み，HPC分野において，クラスタ型のコンピュータシステムが一般的となった．スーパーコンピュータや大規模なクラスタシステムで取り扱うデータ量は年々大規模化し，データ処理を扱うアプリケーションやデータ管理をするストレージの重要性が増してきている．このようなハイエンドなHPCクラスタシステムでは，ストレージアクセスには専用の高速アクセスネットワークを使用し，これをノード間通信のネットワークに加えて構築することが多い．しかし，その場合には，ネットワーク構成が複雑になるだけでなく，高速なネットワークを別に構築するためコストがかかる．

そこで，ノード間通信とストレージアクセスのネットワークを統合することにより，大規模なクラスタシステムにおいても構成を単純化し，高い性能を実現しながら構築，運用コストを下げることができる．例えば，東京工業大学のクラスタ型スーパーコンピュータ

†1 お茶の水女子大学
Ochanomizu University
†2 工学院大学
Kogakuin University
†3 東京工業大学
Tokyo Institute of Technology
†4 国立情報学研究所
National Institute of Infomatics

TSUBAMEでは，ペタバイト級のストレージサーバが，Single Laneの10Gbps InfiniBandで接続され，トータルピーク性能85TFlopsの655台の計算ノードが，Dual Laneの10Gbps InfiniBandで接続されており，InfiniBandによる統合ネットワークを実現している．将来的には，HPC向けのハイエンドなクラスタシステムは，計算ノードとストレージが統合ネットワーク接続される構成が増えていくと考えられる．LUSTRE, GPFSなどの並列ファイルシステムも，そのようなハードウェア環境において最も高い性能を発揮するように設計されている．

しかし，そのような大規模なクラスタシステムでは，解析するためのログやデータ量が膨大であるため，I/Oとノード間接続のトラヒックがどのようであるかなど，ノード間通信とストレージアクセスのネットワークを統合することによる相互影響については明らかにされておらず，解析は容易でない．

そこで本稿では，ノード間通信とストレージアクセスの統合ネットワークとして，IPをベースにしたSANであるIP-SANを用いて，その相互影響について評価を行った．IP-SANで統合したクラスタの構成の一つとして，iSCSIを用いてサーバ機とストレージを1対1接続したクラスタの並列処理性能を評価した．

その際，ストレージへのI/Oを伴う並列ベンチマークを使用して評価し，ローカルストレージのみで構築したクラスタと比較した．その結果，iSCSI接続のIP-SAN統合型クラスタは，ローカルストレージのみを用いたクラスタの並列処理性能とほぼ同程度となり，ボトルネックはIP-SANにおけるネットワーク転送処理部分ではなく，ノード間通信を含む並列処理もしくはストレージのI/O処理の部分であるということがわかった．

2. IP-SAN統合型クラスタ

SANでは，現在，高速な専用回線であるFibre Channelを用いるFC-SANが主流である．HPC分野におけるクラスタにおいても計算ノード–ストレージ間のバックエンドのネットワークにFC-SANが利用されることが多い．しかし，クラスタなどを構築する場合には，FC用のスイッチが高価，管理者が少ないなど，導入コストや管理コストの面で障害がある．

IP-SANは，TCP/IPネットワークで構築する次世代のSANであり，安価なコストでクラスタのストレージを導入，運用ができる．そのIP-SANのプロトコルには，iFCP(Internet Fibre Channel Protocol)，FCIP(Fibre Channel over IP)，iSCSI(Internet Small Computer Systems Interface)[1]があるが，中でも2004年4月にIETFによりRFC化されたiSCSIが急速に普及し始めている．iSCSIは，SCSIコマンドをTCP/IPパケットの中にカプセル化することでブロックレベルのデータ転送を行う．また，iSCSIには，iSCSI HBA(Host Bus Adapter)にiSCSIプロトコルを実装したハードウェア実装やソフトウェアによりiSCSIを実装したものがある．本稿では，安価なコストと多数のノードにおける構築のしやすさの観点から，ソフトウェア実装を用いている．iSCSIの階層構造は，図1のようになっている．

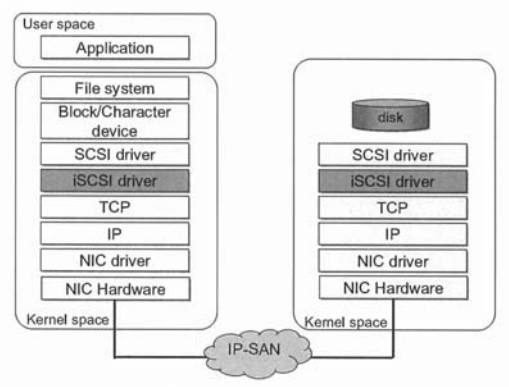

図1 iSCSIの階層構造

クラスタにおいては，ディスクへのI/O処理を行うストレージアクセスは転送データ量が多く，ノード間通信と比べてバースト性が高い．そのため，ハイエンドなHPCクラスタなどでは，計算ノード(サーバ)–ストレージ間のストレージトラヒックに高速なFC-SANを用い，ノード間通信を行うMPIトラヒックと切り離してクラスタを構築することも多い．しかし，ギガビットイーサネットが普及し，10ギガビットイーサネットが登場し始めたことにより，コモディティベースのクラスタシステムでも，iSCSIを使用することで高性能なクラスタ環境を提供できるようになった．

また，iSCSIを使用して，ストレージとMPIのネットワークをTCP/IPとEthernetを用いたコモディティなネットワークに統一することにより，ネットワーク構築コストの削減と運用管理の効率化が容易に可能となる．図2は，それら双方のネットワークを統一したIP-SAN統合型クラスタの構築例である．それぞれのiSCSIターゲットに対してiSCSIイニシエータが接続する場合や，一つのiSCSIターゲットに対して複数の

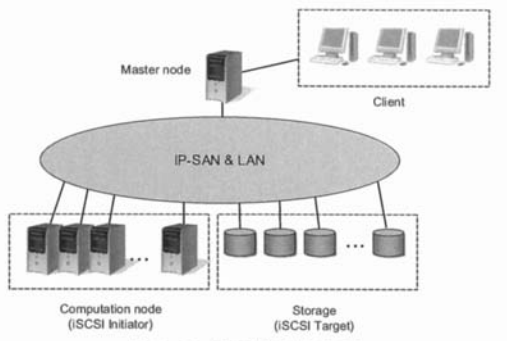

図 2 IP-SAN 統合型クラスタ

iSCSI イニシエータが接続する場合などがある．

このような双方のネットワークを統一した IP-SAN 統合型のクラスタでは，ノード間通信とストレージへのデータ転送がリソースを競合することにより，並列分散処理性能への影響の懸念される．そのため，それら双方のトラヒックを統合したことによる性能への影響については，まだ明らかにされておらず，そのトラヒックの実態を把握し，解析する必要がある．

我々は，4 ノードで構築される小規模な iSCSI 接続の IP-SAN 統合型クラスタを用いて，並列分散処理性能について評価を行い，非統合型のクラスタと比較を行った[2]．しかし，PCI バスなどを含めたノードの性能の影響により，ストレージアクセスがボトルネックとなり，非統合型と統合型のクラスタでは性能差がみられないという結果を得ている．本稿では，I/O 性能のボトルネックを小さくする目的で，より高性能なノードを使用して IP-SAN 統合型クラスタを構築し，I/O 処理を伴う並列分散処理性能の基礎的な評価を行う．

3. IP-SAN 統合型クラスタの性能比較

本稿では，IP-SAN 統合型クラスタの一つの構築例として，まず，それぞれの iSCSI ターゲット (ストレージ) に対して iSCSI イニシエータ (計算ノード) が接続する場合を想定し，評価を行っている．その際の基礎評価として，NAS Parallel Benchmark(NPB)[3] の並列計算ならびに I/O を実行するベンチマークアプリケーションを使用し，並列演算性能を評価し，ローカルストレージを使用した場合と比較した．

3.1 実験環境

本実験における IP-SAN 統合型クラスタは，東京工業大学松岡研究室の PrestoIII を用いて構築している．実験に用いた計算機を表 1 に示す．

表 1 測定環境：使用計算機

OS	Debian GNU/Linux
Kernel	2.6.18
CPU	Opteron250 2.4GHz × 2
Main Memory	2GB
HDD	68GB SCSI HDD
Network	1000BASE-T

IP-SAN 統合型クラスタの場合には，これらの計算機の 16 台を iSCSI イニシエータ (ノード) に，16 台を iSCSI ターゲット (ストレージ) として構築している．そのため，各 iSCSI イニシエータは特定のストレージデバイスに接続する 1 対 1 接続の環境となっている．iSCSI イニシエータと iSCSI ターゲットは，それぞれ DELL PowerConnect5224 のギガビットイーサネットスイッチに接続されており，ノード間通信もストレージアクセスも，同一のネットワークを介してデータが転送される．一方，ローカルストレージを使用したクラスタでは，それぞれのノードに接続されたストレージに対してデータが読み書きされる．

iSCSI のソフトウェア実装には，イニシエータ側では Open-iSCSI ver. 2.0-865[4] を使用し，ターゲット側では iSCSI Enterprise Target ver. 0.4.15[5] を使用している．また，Open-iSCSI と iSCSI Enterprise Target などの iSCSI ソフトウェアでは，iSCSI で接続することにより，ターゲット側の指定したストレージデバイスがイニシエータ側でみえるようになり，ファイルシステムを介してファイルを読み書きすることができる．本実験では，iSCSI で使用するファイルシステムは ext3 としている．また並列分散処理に使用する MPI ライブラリには，MPICH2 Ver. 1.0.3[6] を使用している．

3.2 ネットワーク性能とシーケンシャル I/O 性能

まず，基本性能として，PrestoIII におけるノード間のネットワーク性能を iperf ver.2.0.2 のベンチマークを用いて測定した．その結果，スイッチ内もスイッチ間もともにバンド幅が 949 Mbits/sec(転送量 1.11 GBytes) であった．

また，iSCSI イニシエータから iSCSI ターゲットに対して，Bonnie++ ver. 1.0.3 を用いてシーケンシャル I/O の性能を測定した．本稿の実験すべてに使用している iSCSI の設定パラメータは表 3 のようになっている．Bonnie++は，ファイルシステムを介したディスクアクセスの性能を測定するベンチマークである．ファイルサイズ 4GB に設定し，Bonnie++を使用して測定した結果を表 2 に示す．Bonnie++による測定

表 2 Bonnie++測定結果

	local SCSI	iSCSI
Sequential Block Read	52183KB/sec	55404KB/sec
Sequential Block Write	55206KB/sec	55867KB/sec

表 3 iSCSI の設定パラメータ

InitialR2T	No
ImmediateData	Yes
MaxRecvDataSegmentLength	32KB
FirstBurstLength	256KB
MaxBurstLength	16776192Byte

では，ローカルストレージと iSCSI を使用したストレージを比較し，双方ともにノードとストレージを再起動した後に行っている．ローカルストレージの書き込み性能が 53.91MB/sec であり，iSCSI の書き込み性能 54.56MB/sec を下回り，数回の測定を行ったが同様の結果が得られている．Bonnie++は，ファイルシステムの上から負荷を与えるため，ファイルシステムやブロックデバイスのキャッシュは有効になり，このような結果が得られたものと考えられる．

3.3 NAS Parallel Benchmark I/O

本実験で使用した並列ベンチマークは，MPI ベースのソースコード実装を用いている NPB I/O Ver.2.4 である．NPB は並列計算のベンチマークであるが，このバージョンでは大量の I/O を行う並列計算ベンチマークアプリケーションを実行する NPB I/O (BTIO)[7] が使用できる．ただし，NPB I/O は対象問題 BT (Block Tri-diagonal) に対してのみ実行可能であるため，本実験では BT を対象として NPB I/O を使用した．NPB I/O には，I/O と性能測定の方法によって異なる実行オプションがあり，本実験で使用した実行オプションは次の 2 つである．

- “full” : MPI I/O without collective buffering
- “epio” : parallel I/O

オプション “full” では，各ノードのメモリに分散したデータが単一のプロセッサ上に集められ，単一のファイルとして書き出される．その際，ファイルとして書き出される前にデータが再構成される．“epio” は，各ノードに分散したデータは集められることなく，各々のノードが所有するディスクに並列に I/O を行う．なお，NPB I/O の “epio” では，I/O は並列計算と並列に実行される．

また，実行した NPB I/O の Class，配列サイズ，ストレージに書き出される合計データサイズを表 4 に示す．本実験では，Class A，B，C について測定を行った．表 4 のデータサイズは，オプション “full” を実行した場合には，単一のノードのストレージに書き出されるデータサイズを示す．オプション “epio” の場合には，並列 I/O が実行されるため，表 4 のサイズをノード数で割った値が各ノードのストレージに書き出される．

表 4 NPB の Class と問題サイズ

Class	Size	Mbytes written
A	64 × 64 × 64	419.43
B	102 × 102 × 102	1697.93
C	162 × 162 × 162	6802.44

3.4 NPB I/O による実験結果と考察

本実験では，NPB の実行時間と Mops(Million Operations Per Second) 値を測定した．Mops 値は 1 秒間あたりの 100 万演算数である．Class A，Class B，Class C を実行した場合の実行時間を図 3〜5 に示す．また，Class A，Class B，Class C を実行した場合の Mops 値を図 6〜8 に示す．ただし，Class C の 1 ノードの場合には，NPB のコンパイルができなかったために結果からは省略している．測定は 3 回行い，キャッシュの影響を小さくするため，それぞれの測定 1 回ごとにすべてのノードとストレージは再起動をして測定を行っている．

問題サイズが Class A の場合には，書き出されるデータサイズも 419MB であるため，ローカルストレージと iSCSI あるいはノード数によってばらつきがみられる．しかし，Class C の場合には，やや iSCSI の方が性能が良いという結果が得られた．オプション “epio” に比べ，“full” の方が実行時間が短く，Mops 値が高いのは，単一のプロセッサにデータが集められる際に，プロセッサ間でデータ転送していることが原因である．“epio” は単に並列 I/O が実行され，各ノードのデータを一つのノードへ転送する処理が省略される．また，プロセッサ数を増加させても，それに比例した性能向上は見られない．

全体として，基礎実験として行った NPB I/O の結果では，iSCSI 接続の IP-SAN 統合型クラスタは，ローカルストレージの並列演算処理性能とほぼ同程度であるという結果が得られた．我々は，これまで小規模なクラスタにおける評価を行っていたが 3)，本稿ではより性能の高い計算機をノードとして使用し，ノード数を増やした IP-SAN 統合型クラスタ環境において基礎的な測定を行った．環境を変えたにも関わらず，実験結果として IP-SAN 統合クラスタのベンチマーク

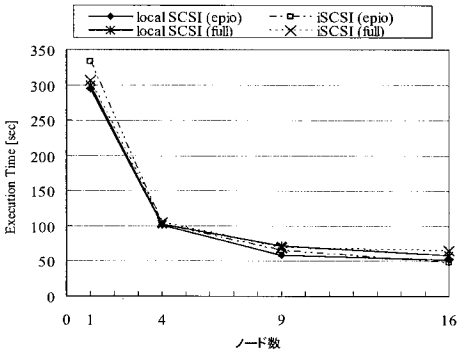

図 3　NPB I/O の実行時間：Class A

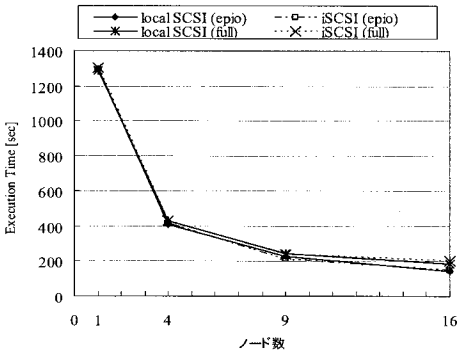

図 4　NPB I/O の実行時間：Class B

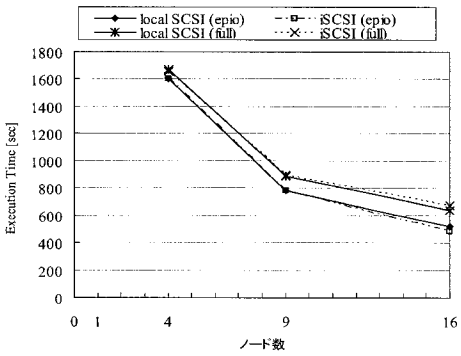

図 5　NPB I/O の実行時間：Class C

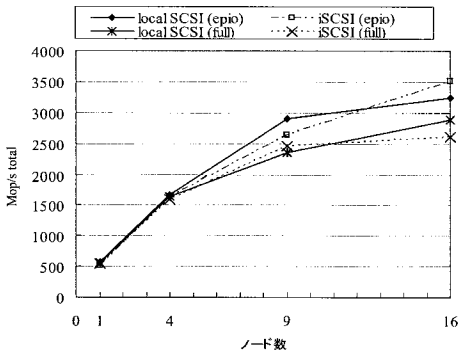

図 6　NPB I/O の Mops 値：Class A

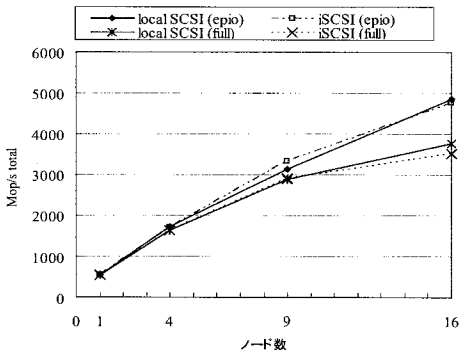

図 7　NPB I/O の Mops 値：Class B

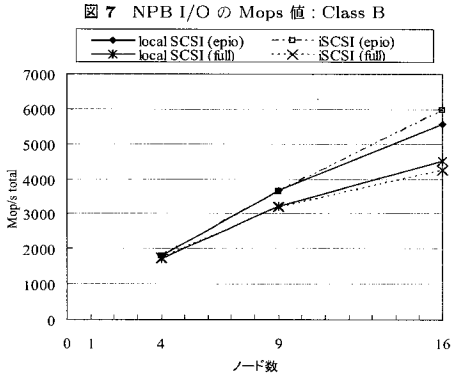

図 8　NPB I/O の Mops 値：Class C

実行性能がローカルストレージを使った場合とほぼ同じであるという結果は変わらないものとなった．このことからボトルネックは IP-SAN におけるネットワーク利用ではなく，並列処理の実行およびその際に発生するノード間通信のためのネットワーク利用，もしくはローカルストレージの場合にも動いており IP-SAN の場合はターゲットで動作する PCI-X バスなどを介したストレージの I/O 処理の部分であるということがわかる．いずれにしても，ネットワークの帯域が使い切られて IP-SAN のネットワークを介したアクセスがボトルネックといった状況にはなっていない．

ストレージの I/O 処理に関しては，今後例えば RAID 0 構成にするなどにより，その影響の変化を見ることができる．ストレージ I/O の性能を向上させることにより，ネットワークの帯域を使い切るようになる場合もあるかもしれないと考えている．一方，

並列処理の実行と I/O の関係については，本実験で使用している NPB I/O は最大でも単一のノードに約 6.6GB の書き込みを行うのみである．従って単一のベンチマークだけでなく，複数のベンチマークあるいは大量の I/O 処理を行うベンチマークを併用するなどして評価を行う必要があると考えている．

4. 関連研究

iSCSI を用いた IP-SAN の性能評価の関連研究として，文献[8]においては，Sarkar らによる iSCSI のソフトウェア実装と TOE (TCP Offload Engine) や HBA (Host Bus Adapter) を用いた iSCSI ハードウェア実装の比較に関する研究を行っている．この文献では，ハードウェア実装は，CPU の負荷を軽減させることはできるが，総合的にはソフトウェア実装の方が性能が高くなることが実証されている．本稿では CPU 負荷を軽減するだけでなく，並列分散処理における I/O 処理の総合的な評価を対象としているため，iSCSI のソフトウェア実装を用いている．

文献[9]において，藤田らは iSCSI ターゲットに着目して実装手法を考慮し，2 種類の iSCSI のオープンソースソフトウェアを用いて，iSCSI の性能評価を行っている．同文献では，OS のカーネルが提供する標準機能を利用する手法よりも，変更を加えたカーネルが提供する iSCSI ターゲットに最適化された機能を利用する手法の方が性能が優れていることを確認している．同文献ではクラスタにおいて iSCSI を使用するものではないので，本研究と異なる．

谷村らの文献[10]では，ストレージやネットワークを含めた仮想クラスタを事前予約に基づいて提供することを目的として，Xen による仮想計算機上の NFS や iSCSI のアクセス性能について評価している．クラスタにおいて iSCSI を使用するという点においては本稿と同じであるが，本稿では並列処理性能に着目して評価しているため，本研究とは異なる．

5. まとめと今後の課題

大規模なクラスタシステムやスーパーコンピューティングでは，クラスタの構築および運用管理コストが削減するため，ノード間通信とストレージアクセスのネットワークの統合が実現され始めている．このようなクラスタでは，MPI とストレージ双方のトラヒックを統合したことによる性能への影響については明らかにされていない．本稿では，統合ネットワークを使用したクラスタの一つとして，iSCSI を用いた IP-SAN 統合クラスタについて，ストレージへの I/O を伴う並列ベンチマークである NPB I/O を使用して基礎的な評価を行った．

その結果，基礎実験として行った NPB I/O の結果では，iSCSI 接続の IP-SAN 統合型クラスタは，ローカルストレージの並列演算処理性能とほぼ同程度であるという結果が得られた．それは，ボトルネックが IP-SAN におけるネットワーク転送処理部分ではなく，ノード間通信を含む並列処理もしくはストレージの I/O 処理の部分にあるといえる．

今後の課題として，ストレージを RAID0 構成にし，本稿では 1 対 1 接続であったクラスタシステムを一つの iSCSI ターゲットに複数の iSCSI イニシエータが接続するなど，様々なシステム構成に構築し，について評価を行っていきたい．単一のベンチマークによる評価だけでなく，複数のベンチマークや実アプリケーションを使用し，より大規模なクラスタ環境において MPI とストレージのトラヒックがどのように影響するのかを詳細に解析していく．

参考文献

1) iSCSI RFC: `http://www.ietf.org/rfc/rfc3722.txt`.
2) 神坂紀久子，山口実靖，小口正人，喜連川優：iSCSI を用いた PC クラスタにおけるバックエンドネットワーク統合による性能への影響評価，電子情報通信学会コンピュータシステム研究会 (CPSY), CPSY2006-46, pp. 49–54 (2006).
3) NAS Parallel Benchmark (NPB): `http://www.nas.nasa.gov/Software/NPB`.
4) Open-iSCSI: `http://www.open-iscsi.org/`.
5) iSCSI Enterprise Target: .
6) MPICH2: `http://www-unix.mcs.anl.gov/mpi/mpich`.
7) NPB-MPI 2.4 I/O: `http://www.nas.nasa.gov/News/Techreports/2003/PDF/nas-03-002.pdf`.
8) Sarkar, P., Uttamchandani, S. and Voruganti, K.: Storage over IP: When Does Hardware Support help?, *Proc. FAST 2003, USENIX Conference on File and Storage Technologies*, pp. 231–244 (2003).
9) 藤田智成，小河原成哲：iSCSI ターゲットソフトウェアの解析，情報処理学会誌コンピューティングシステム，Vol. 46, SIG3 (ACS 8), pp. 38–50 (2005).
10) 谷村勇輔，小川宏高，中田秀基，田中良夫，関口智嗣：仮想クラスタに対する IP ストレージの提供方法の比較，情報処理学会研究報告，HOKKE2007, pp. 109–114 (2007).