広域分散仮想化基盤のストレージ評価の最新動向

柏崎 礼生[1,a)] 北口 善明[3,b)] 近堂 徹[2,c)] 中川 郁夫[1,4,d)] 下條 真司[1,e)]

概要：複数の研究拠点をネットワークで接続し，広域に分散した環境でも高いランダムアクセス性能を示すことが可能なストレージ技術について研究開発を進めている．既存の分散ストレージは仮想マシンでの利用に実用的な性能を提供することが困難であり，あるいはマイグレーション後の性能劣化が発生していたが，本技術ではこれらの問題を克服し，広域分散仮想化環境を実現可能であることを示唆している．本発表ではマイグレーション前後のパフォーマンス評価の結果，およびストレージノードの増強後の性能向上について評価する．

キーワード：仮想化基盤, クラウドコンピューティング, 障害回復 (DR), 広域分散, 広域ライブマイグレーション

The latest results of a distributed storage system for a widely distributed virtualization infrastracture

Kashiwazaki Hiroki[1,a)] Yoshiaki Kitaguchi[3,b)] Kondo Tohru[2,c)] Ikuo Nakagawa[1,4,d)]
Shinji Shimojo[1,e)]

Abstract: The authors have been researching and developing widely distributed storage environment that is characterized by high performance of random I/O access. It is difficult for conventional distributed storage to serve practicable performance for using virtualized machines, especially on live migration. In this paper, the authors show the latest result of performances of I/O on proposed distributed algorithm.

Keywords: Virtualization infrastracture, cloud computing, disaster recovery, widely distributed, wide area live migration

1. はじめに

日本では 2011 年 3 月 11 日に発生した東日本大震災以来，自然災害による機器の損壊，回線の切断などを要因とするサービスの中断に対応することが切実な問題として表面化した事により，災害回復 (Disaster Recovery: DR) や事業継続計画 (Business Continuity Plan: BCP) を実現する手法が求められている．この手法として遠隔地データセンターの利用と一部システムあるいは基幹システム全ての移行というアプローチを京都教育大学や京都大学が採用している [*1][*2] ．組織の本拠点とデータセンターが同時に一つの自然災害により損壊する確率は低いが，本拠点もデータセンターも人的災害や各種要因によりサービスの中断が発生することがあるため，他一拠点にデータの複製やバックアップを確保することは十分な対策とは言えない．その一方で複数拠点のデータセンターを利用することはコスト

1 大阪大学サイバーメディアセンター
Cybermedia Center, Osaka University
2 広島大学情報メディア教育研究センター
Information Media Center, Hiroshima University
3 金沢大学総合メディア基盤センター
Information Media Center, Kanazawa University
4 株式会社インテック
Intec Inc.
a) reo@cmc.osaka-u.ac.jp
b) kitaguchi@imc.kanazawa-u.ac.jp
c) tkondo@hiroshima-u.ac.jp
d) ikuo@inetcore.com
e) shimojo@cmc.osaka-u.ac.jp

*1 http://pr.fujitsu.com/jp/news/2011/06/28.html
*2 http://pr.fujitsu.com/jp/news/2013/01/10.html

の面で困難が生じる．

データセンターを利用した DR において，組織の本拠点と同じ構成のシステムをデータセンター側でも稼働させホットスタンバイ方式で稼働させる場合，本拠点からデータセンターまでの遅延による影響を受けるためストレージのパフォーマンスが距離に応じて低下する．プライベートクラウドの構築に当たって性能向上のボトルネックとなるのは CPU やメモリ資源ではなくストレージであることが指摘されており [1]，この方式による DR の実現には費用対効果の困難さがある．仮想化基盤においては，仮想マシン (Virtual Machine: VM) で稼働する OS やサービスを停止させることなく他のハイパーバイザサーバ上で稼働させるライブマイグレーションが利用される．ライブマイグレーションを利用するためには複数のハイパーバイザサーバが共有するストレージが必要となるが，広域環境で共有ストレージを利用すると前述のホットスタンバイ方式での問題同様，遅延の影響を受けストレージへの I/O パフォーマンスが劣化する．一方，共有ストレージを利用せずに VM イメージを拠点間で移動させるストレージマイグレーションも利用されているが，共有ストレージを利用したライブマイグレーションに比べてサービス断時間が長くなる問題を解決しなければいけない [2]．

広域分散型のストレージとして Gfarm [3]，分散ファイルシステムとしては Google の GFS [4]，および HDFS[*3] が広く利用されている．Gfarm ではデータの保存はファイル単位であり，ファイルの任意の位置の修正においてもファイル全体へのアクセスが必要となってしまう．一方，GFS(HDFS) はファイルをブロック分割して保存するものの，Write-Once-Read-Many(書き込みは一度で読み出しを何度も行う) モデルに基づいたデータアクセスを前提とした設計であるため，POSIX の要件を緩和しており，ファイルの任意の位置の修正や複数の単一ファイルへの同時書き込みはできない．以上のことより，シーケンシャルアクセスに対しては十分な性能を発揮する一方で，ファイルの部分的な更新といったランダムアクセス性能については十分な性能を提供することが困難である．現在広く用いられている複数の仮想化ハイパーバイザの実装は POSIX 準拠のファイルシステムに対応している．また，仮想化ハイパーバイザは VM のイメージファイルに対してランダムアクセスする．これまでの POSIX 準拠の広域分散型のストレージはランダムアクセス性能がローカルストレージに比べて低いため，仮想化基盤のためのストレージとして利用することが困難である．そのため仮想化基盤のためのストレージは POSIX 準拠であり，かつ広域分散型であってもローカルストレージと同程度のランダムアクセス性能を示す必要がある

[*3] http://hadoop.apache.org/docs/hdfs/current/hdfs_user_guide.html

そこで本研究では，スケールアウト型の分散ストレージを地理的に広域に分散した複数拠点に配備し，広域分散型の仮想化基盤を実現するための広域分散ストレージ構築手法について提案する．本稿では国内三拠点で広域分散ストレージ環境を構築し，その I/O 性能を評価するとともに，拠点間ライブマイグレーションの評価実験を通して，本提案手法が広域分散仮想化基盤の実現に有効であることを示す．

2. 評価実験

本研究では広域分散対応を行った分散型ストレージを利用し，その評価を行っている．本稿では評価実験の構成とストレージの I/O 性能の結果を示す．

2.1 広域分散ストレージ環境の構成

現在構築を進めている広域分散ストレージ環境は，広島大学，金沢大学，国立情報学研究所 (以下，NII) の 3 拠点の接続が完了している．拠点間は NII が提供する学術情報ネットワーク SINET4 を利用して 10Gbps で接続し，用途に応じた 3 つの VPN サービス（L2VPN サービス× 2, L3VPN サービス× 1）を利用している．以下に，それぞれについて説明する．

EXAGE-LAN(L3VPN) は，分散ストレージ内部の分散処理用セグメントである．このセグメントは各拠点がそれぞれ独立した L3 ネットワークで構成され，各 L3 ネットワークが SINET4 の L3VPN サービスで相互接続されている．これは前節でも述べた通り，分散ストレージのアーキテクチャ上，ブロックの配置アルゴリズムがネットワーク単位で決まるためである．

管理 LAN(L2VPN) と MIGRATION-LAN(L2VPN) は，本ストレージをデータストアとする仮想計算機モニタ (VMM) のためのセグメントである．管理 LAN は仮想計算機モニタの管理用セグメントとなり，MIGRATION-LAN は仮想計算機モニタ上で動作する仮想マシン (VM) が接続するセグメントである．このセグメントに接続される VM は，本分散ストレージを OS イメージのデータストアとして利用する．各拠点には，拠点内のコアサーバ (CS)，アクセスサーバ (AS)，および仮想計算機モニタの死活監視と統計情報を収集するヒントサーバ (HS) を設置する．

広島大学を例に拠点内ネットワーク構成を説明する．図 1 は，SINET アクセスポイント配下の広島大学拠点の構成を示したものである．各拠点ではアクセスサーバが広域分散ストレージのインタフェースとなる．利用するクライアントは，アクセスサーバに対して NFS マウントすることで POSIX 準拠のファイルシステムとして参照することができる．アクセスサーバは 10Gbps および 1Gbps × 4 のリンクアグリゲーション，コアサーバは 1Gbps × 3 のリンクアグリゲーションにより集約スイッチに接続し，ヒント

表 1　各拠点の機器構成
Table 1　Equipment Configuration on Each Facility

拠点名	サーバの種類	台数
広島大学	アクセスサーバ	1 台
	ヒントサーバ	1 台
	コアサーバ	4 台
金沢大学	アクセスサーバ	1 台
	ヒントサーバ	1 台
	コアサーバ	8 台
NII	ヒントサーバ	1 台
	コアサーバ	4 台

サーバは仮想マシンで用意している．また，アクセスサーバを NFS マウントする VMM は 1Gbps × 2 のリンクアグリゲーションで集約スイッチと接続する構成としている．なお，各拠点の機器構成を表 1 に示す．

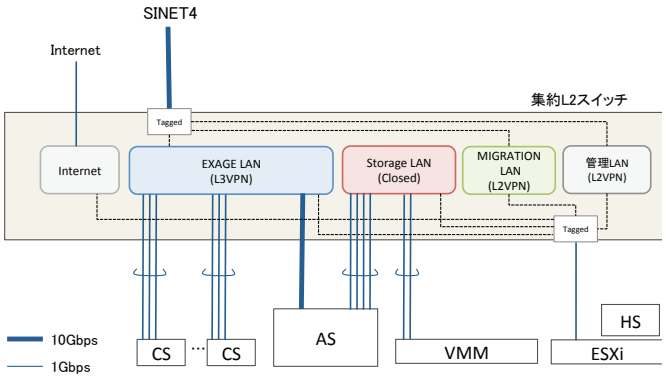

図 1　広島大学のネットワーク構成図
Fig. 1　Network Diagram of Hiroshima University

2.2　ストレージ性能評価

[*4]EXAGE / Storage の I/O 性能を評価するために iozone[*5] を用いて計測した (図 2)．広島大学の拠点に設置した VMM は Intel Xeon (E5-2640) を 2 基，64GB のメモリを搭載し，CentOS 6.3 がインストールされている．この VMM 上で iozone を実行し，従来方式と広域分散対応方式の両方で性能を評価した．EXAGE / Storage のインターフェイスプロトコルは NFS とし，close コールを含めた時間を計測する．検証環境の NFS クライアントの実装はキャッシュを保持している．NFS の write 時はキャッシュに対して行われ，fsync によりキャッシュが書き出される．また read 時はキャッシュ上のファイルと NFS サーバ上にあるファイルの mtime およびファイルサイズを比較し，同一である場合にはキャッシュ上にあるデータを返す．そのため flush(fsync コール) に要する時間を含めた処理時間を計測することで，キャッシュによる性能への影響を排除し，ストレージの性能を直接的に評価する．また Direct IO を利用し，open システムコールがカーネル空間のページキャッシュを利用しないように指定する．アクセスパターンは write, rewrite, read, reread, random read, random write, bkwd read, record rewrite, stride read, fwrite および fread を指定する．ブロックサイズは 4MB とし，4MB から 32GB までのファイルサイズでスループットを計測する．

従来方式では 30〜40MB/sec にピークが存在し，平均スループットは 58.5MB/sec である．一方，提案する広域分散対応方式では 30〜40MB/sec のピークが 40〜50MB/sec に移動し，また 110〜120MB/sec の頻度は 48.7% 増大している．平均スループットは 71.2MB/sec であり，従来方式より 21.7%の性能向上を実現している．この結果はこの VMM サーバと同一セグメントに配置された同スペックのサーバが持つローカルストレージへの NFS によるアクセスと同等のパフォーマンスを示している

3.　おわりに

本稿では DC 内で完結する低遅延環境を対象としたスケールアウトストレージシステムを高遅延環境において適用可能にするためのアーキテクチャを再設計し，国内 3 拠点からなる広域分散ストレージのための検証環境を構築して I/O パフォーマンスとライブマイグレーションを評価した．拠点内に存在する NFS サーバと本提案手法を実装した広域分散ストレージのパフォーマンスを比較し，提案手法は他拠点へブロックを複製しながらも拠点内に存在する NFS サーバと同等の I/O 性能を示すことを明らかにした．この結果から各拠点は DR のためのストレージと自拠点の仮想化基盤のためのストレージとを区別することなく利用できることを示した．

また通信遅延 (RTT) が 18 msec の環境において，拠点内に存在する NFS サーバを用いたライブマイグレーションと提案手法を用いたライブマイグレーションは同等の性能となることを確認した．また，マイグレーション後の I/O 性能について，拠点間の通信遅延による影響が一拠点の NFS サーバを複数拠点で共有利用する場合と比較して，小さいことを確認した．日米間の通信では 100〜200 msec 程度の遅延が発生するなど，世界規模のグローバルな通信では遅延が大きな問題になり得る．コアサーバ数を増加させることによりパフォーマンスを向上させ，高遅延環境における検証をすることが今後の課題である．

謝辞　本研究は平成 24 年度北海道大学情報基盤センター共同研究「インタークラウドをより拡張するための地域間相互接続の調査検証」，平成 24 年度国立情報学研究所共同研究「“Trans-Japan Inter-Cloud Testbed” の構築に向けたネットワーク基盤に関する検討」，平成 24 年度学際大規模

*4　【3.1.15 対応】
*5　http://www.iozone.org

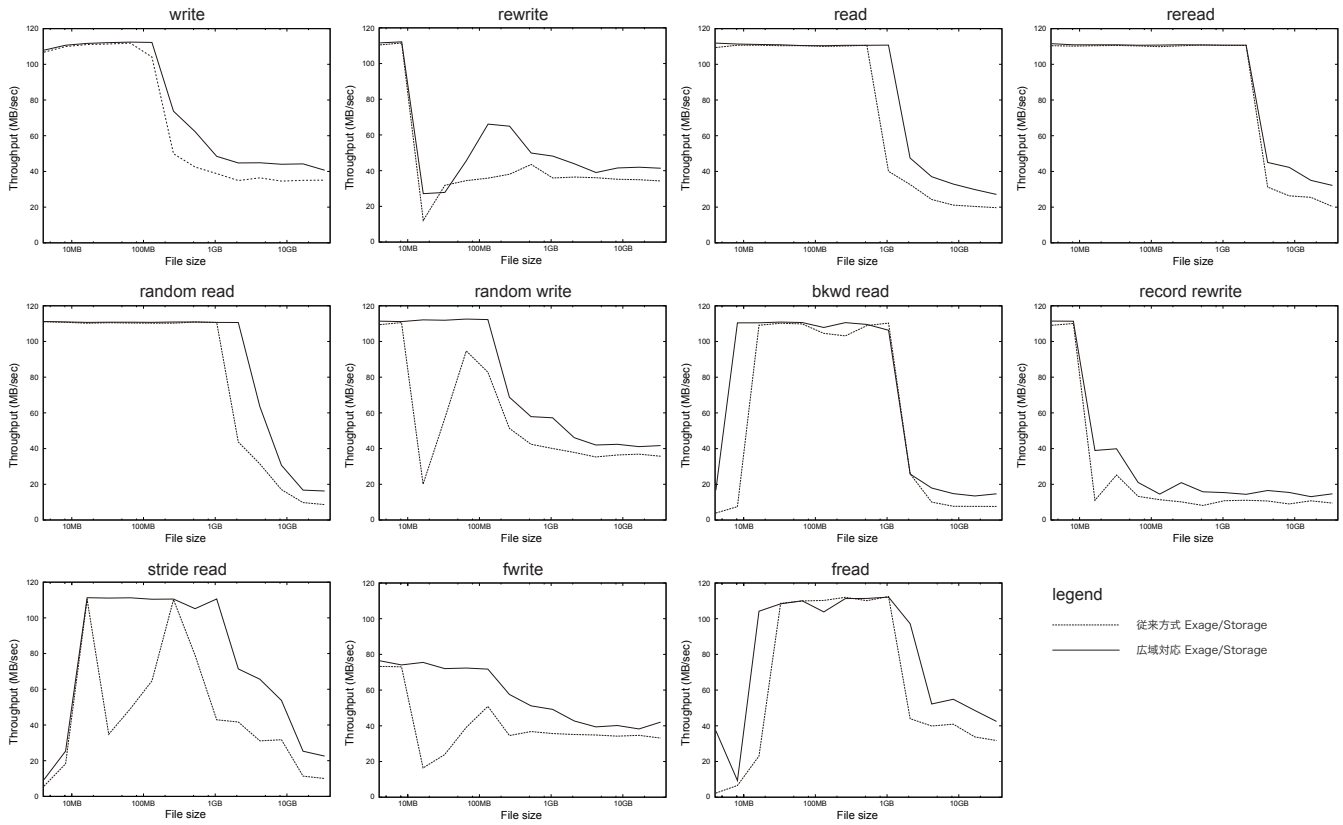

図 **2** EXAGE / Storage の Read/Write パフォーマンス
Fig. 2 EXAGE / Storage Read/Write Performance

情報基盤共同利用・共同研究拠点公募型共同研究「分散クラウドシステムにおける遠隔連携技術」による支援を受けました．本研究の実証実験にあたり，コンピュータリソースのご提供をいただいた各大学, JGN-X の回線をご提供いただいた独立法人情報通信研究機構, SINET4 の回線をご提供いただいた国立情報学研究所, および, クラスタストレージ技術である EXAGE / Storage をご提供いただいた株式会社インテック，および，アクセスサーバとして UCS をご提供いただいた Cisco Systems 合同会社に感謝します．

参考文献

[1] Jeffrey Shafer: I/O virtualization bottlenecks in cloud computing today, Proceedings of the 2nd conference on I/O virtualization (WIOV'10), pp.5–5 (2010).
[2] 関谷勇司: 広域分散クラウドへの挑戦と課題, 電子情報通信学会信学技報, Vol. 111, No. 375, IA2011-63, pp. 49-54 (2012).
[3] S. Mikami, K. Ohta, O. Tatebe: Using the Gfarm File System as a POSIX Compatible Storage Platform for Hadoop MapReduce Applications, Grid Computing (GRID), 2011 12th IEEE/ACM International Conference on, pp.181-189 (2011).
[4] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung: The Google file system, Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP '03), pp.29-43 (2003).
[5] A. Azagury, V. Dreizin, M. Factor, E. Henis, D. Naor, N. Rinetzky, O. Rodeh, J. Satran, A. Tavory, and L. Yerushalmi: Towards an object store, Proceedings of the 20 th IEEE/11 th NASA Goddard Conference on Mass Storage Systems and Technologies (MSS'03), pp.165 (2003).
[6] Factor, M.; Meth, K.; Naor, D.; Rodeh, O.; Satran, J.: Object storage: the future building block for storage systems, Local to Global Data Interoperability - Challenges and Technologies, pp.119-123 (2005).
[7] Ikuo Nakagawa, Kenichi Nagami: Jobcast - Parallel and distributed processing framework Data processing on a cloud style KVS database, Journal of Information Processing, Vol.21, No.3 (2013).
[8] 首藤一幸: "key-value ストアの基礎知識", Software Design, 2010 年 2 月号 (2010).
[9] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati,Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels: Dynamo: Amazon's Highly Available Key-value Store, Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles (SOSP'07), pp.205-220 (2007).