Vol.2013-IOT-20 No.19
2013/3/14

広域分散ストレージ検証環境におけるI/O性能評価

柏崎 礼生[1,a)] 近堂 徹[2,b)] 北口 善明[3,c)] 楠田 友彦[4,d)] 大沼 善朗[4,e)] 中川 郁夫[1,4,f)]
市川 昊平[5,g)] 棟朝 雅晴[6,h)] 高井 昌彰[6,i)] 阿部 俊二[7,j)] 横山 重俊[7,k)] 下條 真司[1,l)]

概要：大規模災害による危機意識の高まりから災害回復 (Disaster Recover: DR) を実現するための技術として遠隔地データセンターでのバックアップや分散ストレージに注目が集まっている．現在我々はランダムアクセス性能の高さに特徴のある広域分散ストレージ環境を金沢大学，広島大学，NII を中心として構築しており，本研究では本環境の I/O 性能を評価し，この環境の有用性を示す．

Hiroki Kashiwazaki[1,a)] Tohru Kondou[2,b)] Yoshiaki Kitaguchi[3,c)] Tomohiko Kusuda[4,d)]
Yoshirou Onuma[4,e)] Ikuo Nakagawa[1,4,f)] Kouhei Ichikawa[5,g)] Masaharu Munetomo[6,h)]
Yoshiaki Takai[6,i)] Shunji Abe[7,j)] Shigetoshi Yokoyama[7,k)] Shinji Shimojo[1,l)]

1. はじめに

仮想化技術の成熟とともに情報システムを稼働させる物理マシンを仮想化環境へと移行し，さらにはパブリッククラウド事業者が提供する IaaS へと移行する試みが行われている [1]．組織外部のクラウドサービスを使うだけでなく国内の教育・研究機関の情報センターや研究科でのパブリック・プライベートクラウドの構築が行われている．静岡大学はクラウドコンピューティングを全面採用した情報基盤システムを構築した [2]．北陸先端科学技術大学院大学 (JAIST) では仮想デスクトップサービスを提供するためにプライベートクラウドを構築している [3] [4]．佐賀大学は専用線で接続された外注先にプライベートクラウドを構築し，メールサービスの提供を行っている [5]．一方で，東京工業大学の TSUBAME2 に代表されるクラウド型 (スケールアウト型)HPCI や北海道大学アカデミッククラウド [6] など計算能力の大きさに重点をおいたパブリックサービスも提供されている．そもそもクラウドコンピューティングという言葉は，2006 年に開催された Search Engine Strategies Conference で Google の CEO(当時) だった Eric Emerson Schmidt 氏が Danny Sullivan 氏との対談で使ったのが初めてとされる [*1]．クラウドコンピュータの定義は Gartner，UC Berkeley，そして NIST による定義が引用されることが多いが [7] [8] [9]，本稿では「仮想化技術を用いて実現されるスケールアウト可能な基盤の上に構築された，規模を収縮可能なサービス」の意味で用いることとする．

パブリック・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りがちなのは CPU やメモリ資源ではなくストレージである [10]．ストレージ機器との接続が広帯域であることが求められるだけでなく，高い IOPS を実現するコントローラも求められる．その一方で，自然災害による機器の損壊，サービスの中断に対応

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.
5 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
6 北海道大学情報基盤センター
Information Initiative Center, Hokkaido University
7 国立情報学研究所
National Institute of Informatics
a) reo@cmc.osaka-u.ac.jp
b) tkondo@hiroshima-u.ac.jp
c) kitaguchi@imc.kanazawa-u.ac.jp
d) kusuda_tomohiko@cloud.intec.co.jp
e) onuma_yoshiro@cloud.intec.co.jp
f) ikuo@inetcore.com
g) ichikawa@is.naist.jp
h) munetomo@iic.hokudai.ac.jp
i) takai@iic.hokudai.ac.jp
j) abe@nii.ac.jp
k) yoko@nii.ac.jp
l) shimojo@cmc.osaka-u.ac.jp

[*1] http://www.google.com/press/podium/ses2006.html

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がある．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．複数拠点のデータセンターを利用することはコストの面で困難が生じる．

そこで本研究では複数拠点でデータの複製やバックアップを行いたい組織がストレージ資源を提供し合うことにより広域に分散されたストレージを実現する手法を提案する．広域分散型のストレージとして Gfarm [11] や Google の GFS [12]，および HDFS[*2] が挙げられる．これらのストレージは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．DR や BCP を対象とする時，仮想マシンのイメージファイルを複数拠点で参照可能にし，広域でライブマイグレーションをする利用が想定される．そのため各拠点からは POSIX ファイルシステムとして利用可能であり NFS のような一般的なインターフェースプロトコルで利用可能であり，ランダムアクセスにおいて十分な性能を発揮することが求められる．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．

2. ストレージアーキテクチャ

本研究で利用するストレージアーキテクチャは株式会社インテックがクラウドコンピューティング技術を応用して開発したスケールアウト型の並列分散処理プラットフォームフレームワーク「EXAGE」を用いて実装された分散ストレージである EXAGE/Storage を用いる．EXAGE/Storage は，NFS，CIFS，iSCSI といった一般的なインタフェースプロトコルに対応したユニファイドストレージであり，分散ストレージでありながら，ファイルに対するランダムアクセスにも対応している．EXAGE/Storage では，読み書きされるファイルを複数の細かな単位に分割し並列分散処理を行うことで，高いスループットを実現するとともに，クラウドコンピューティング技術に準じたデータや処理の冗長化，フェイルオーバーの技術により高い信頼性を提供する．また，EXAGE/Storage はネットワークに接続された複数のコンピュータにインストールされるソフトウェアである．EXAGE/Storage がインストールされたコンピュータは，ネットワークを通じ相互に論理的に接続されることによって，巨大なストレージ空間を提供することが可能である．ネットワーク接続が可能であれば地理的に遠隔地にあるデータセンタを接続してより巨大なストレージ空間を提供することも可能である (図 1)．

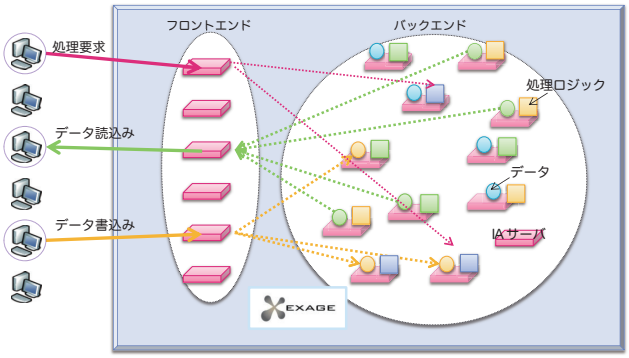

図 1　EXAGE/Storage 概念図
Fig. 1　EXAGE/Storage Conceptual Diagram

EXAGE/Storage ではフロントエンドサーバ (アクセスサーバ) がクライアントに対するインターフェイスプロトコルを提供する．クライアントはフロントエンドサーバに対して処理要求を送信し，フロントエンドサーバは IA サーバによって構成されるバックエンドサーバ (コアサーバ) 群にメタデータ (管理データ) およびユーザデータ (実データ) を保存する．EXAGE/Storage は広帯域低遅延環境での動作を想定していたが，広帯域高遅延環境での動作を実現させるためにコアサーバのうちアクセスサーバの近傍とそうでないものとを区別する．

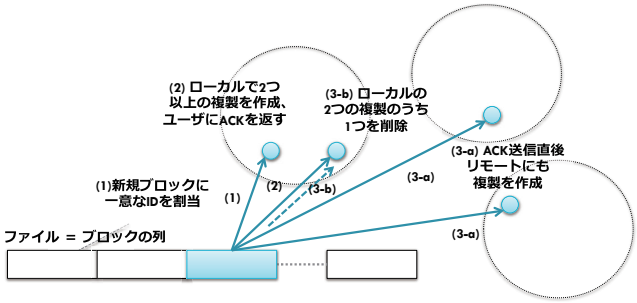

図 2　EXAGE/Storage における分散型の複製管理
Fig. 2　Distributed Replication Management on EXAGE/Storage

クライアントがブロック作成要求をアクセスサーバに対して行うと，アクセスサーバは近傍のコアサーバ上にブロックとその複製を作成し，この時点でクライアントに ACK を返す (図 2)．コアサーバ上にある処理ロジックはこの情報の複製を近傍以外の拠点に対して作成する．近傍の判定にはネットワークセグメントを利用し，異なる拠点は異なるネットワークセグメント上にコアサーバを配置す

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

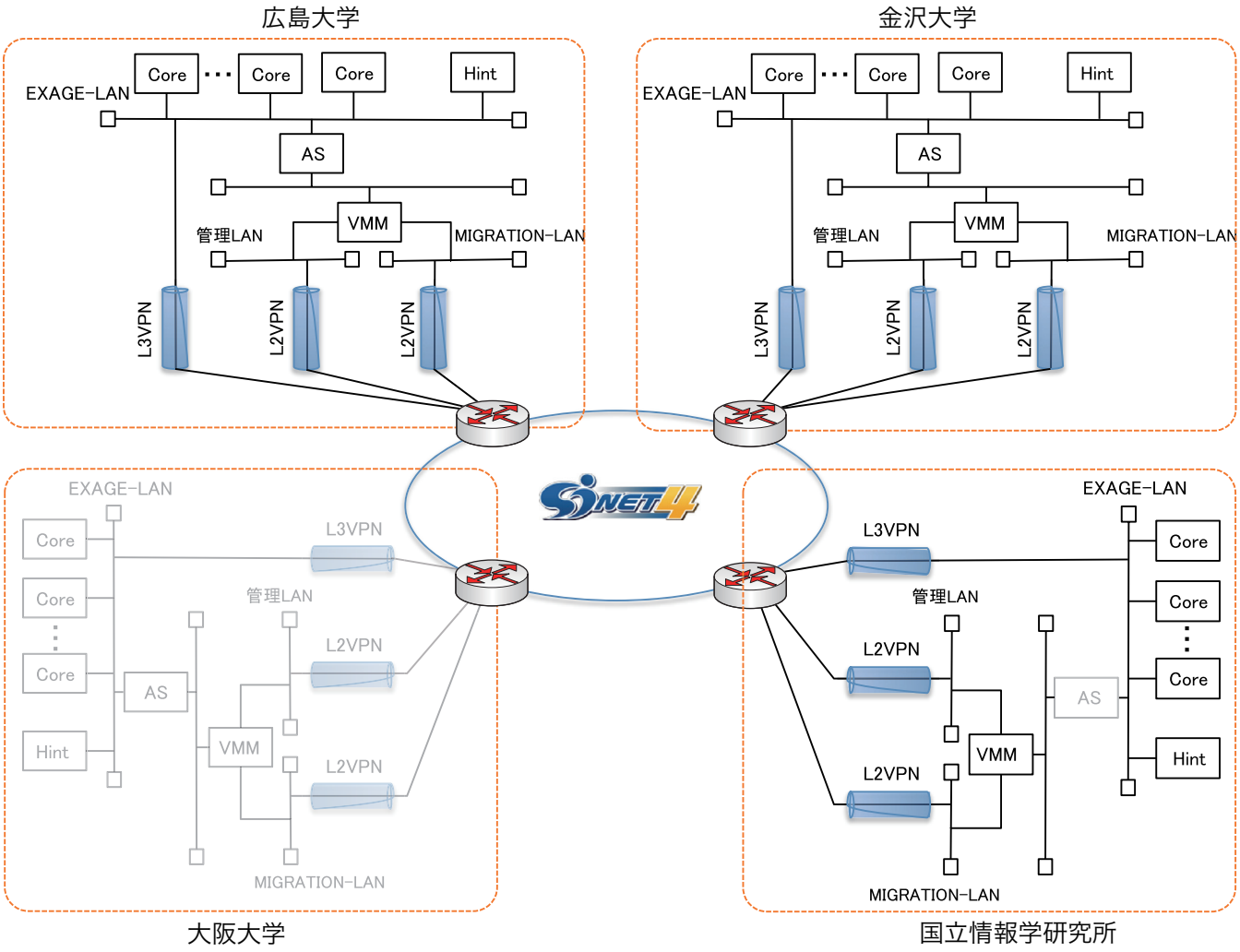

図 3　拠点間構成図
Fig. 3　Participating Institutions Diagram

る．複製を作成する自拠点以外の拠点数 (多重度) は設定可能であり，指定された多重度を満足する複製が作成されるまでコアサーバは他拠点への複製を繰り返す．これによりクライアントは他拠点への複製の作成を意識することなく，他拠点への複製を実現することができる．

3. 広域分散ストレージ環境

3.1 拠点間接続の構成

現在構築を進めている広域分散ストレージ環境の構成図を図 3 に示す．原稿執筆時点では，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おり，現在，大阪大学，北海道大学等の拠点拡大に向けた作業を進めている．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

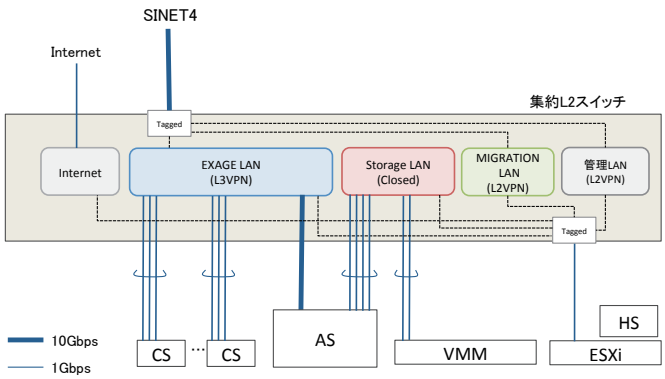

図 4　広島大学のネットワーク構成図
Fig. 4　Network Diagram of Hiroshima University

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

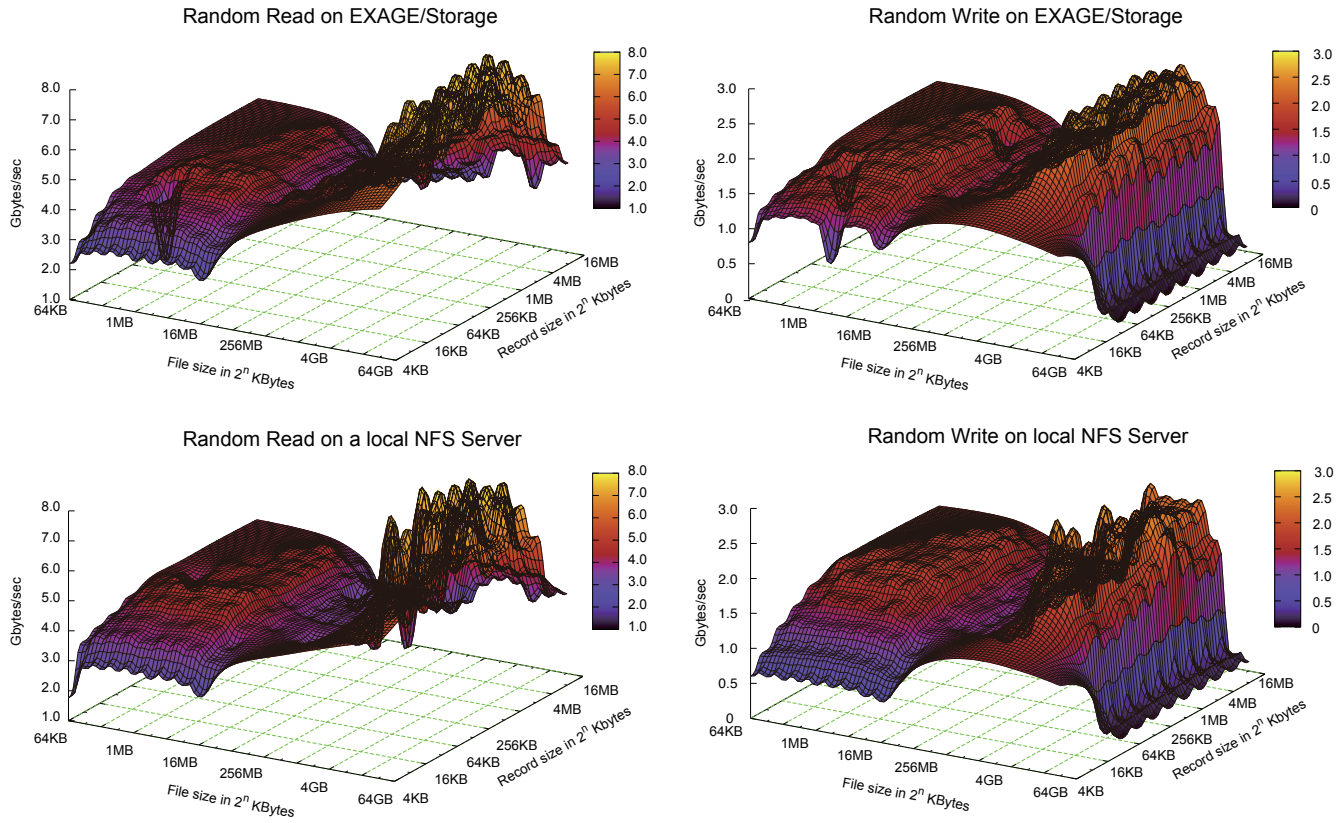

図 5　EXAGE/Storage と NFS ストレージの random read/write パフォーマンス
Fig. 5　EXAGE/Storage and a NFS Storage Random Read/Write Performance

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

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 (VMM) 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．管理 VLAN および MIGRATION-LAN を利用した，拠点間の長距離ライブマイグレーションの検証およびアプリケーション構築は今後実施を予定している．

3.2　拠点内の構成

広島大学を例に拠点内ネットワーク構成を説明する．図 4 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 × 3 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 マウントするアプリケーションサーバ (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．

4.　性能評価

EXAGE/Storage の I/O 性能を評価するために，特に VM のマイグレーションを拠点間で行う際に問題となる Random Read および Random Write の性能を iozone[*3] を用いて計測した．広島大学の拠点に設置した x86 サーバは Intel Xeon (E5-2640) を 2 基，64GB のメモリを搭載

[*3] http://www.iozone.org

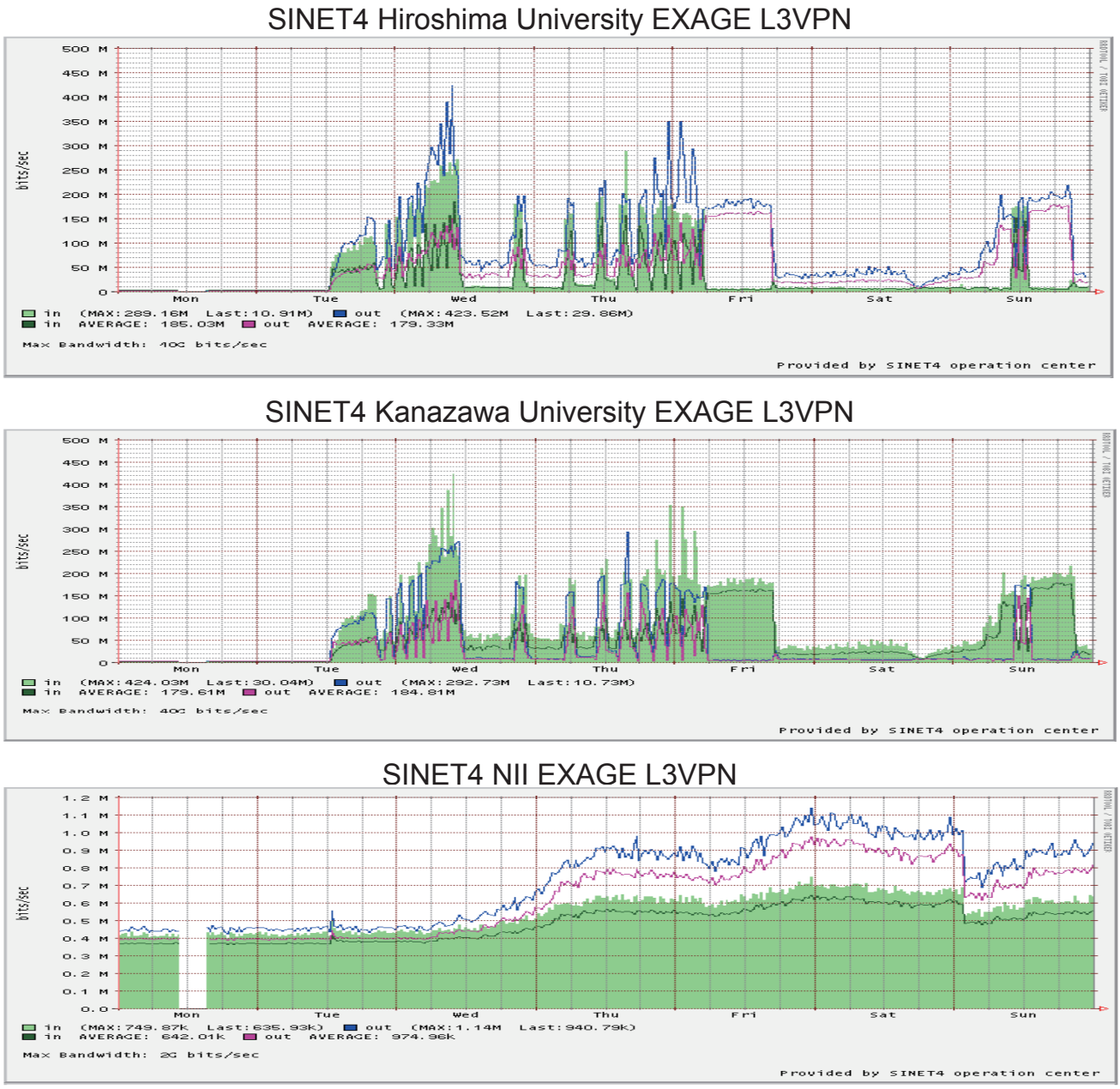

図 6　SINET4 L3VPN のトラフィック状況
Fig. 6　Traffic Condition of L3VPN on SINET4.

し，CentOS 6.3 がインストールされている．record size を 4KB から 16MB まで段階的に変化させ，ファイルサイズは 64KB から 128GB まで変化させる．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．また flush(fsync,fflush コール) に要する時間を含めた時間を計測する．EXAGE/Storage への NFS アクセスと，この x86 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 アクセスとを比較した計測結果を図 5 に示す．

Read および Write ともに 4GB 以上のファイルサイズにおいてパフォーマンスの向上が観測される．また x86 サーバのローカルストレージへの NFS アクセスと遜色のないパフォーマンスを示している．この書き込みによって広島大学からだけでなく NII，および金沢大学からも同一のファイルが存在することを確認することができる．他拠点から見える同一ファイルへの Read/Write のパフォーマンス試験については今後の試験を通して公開する．

コアサーバが複製を作成するため，広島拠点から NII および金沢大学に対して SINET4 L3VPN を通してブロックの作成が行われる．SINET4 で観測された該当する VPN のトラフィック量の変化を図 6 に示す．多重度を 1 としているため，複製は金沢大学のみに作成されており，広島拠

点と金沢拠点のトラフィックを数百 Mbps 占有していることが分かる．複製の通信はユニキャストで行われるため多重度を上げるとそれだけ帯域を占有することとなる．本手法はコアサーバなしのアクセスサーバのみでも拠点として成立するが，コアサーバを設置した拠点については多重化の設定によっては 1Gbps の帯域を逼迫することが懸念される．現在，広島および金沢は 10Gbps で接続されている．

接続する拠点が SINET4 に参加する組織のみとは限らないため，今後は JGN-X への接続性を持つ拠点もこの広域分散ストレージプロジェクトに参加できるよう，大阪拠点は SINET4 と JGN-X のルーティングを行う予定である．SINET4 と JGN-X の双方への接続性を持つ組織も複数考えられ，潤滑かつ効率的な運用のためにマルチホーム環境でのトラフィックエンジニアリングを行う必要性が考えられる．

5. おわりに

本稿では広帯域低遅延を対象としたスケールアウトストレージシステムを広帯域高遅延環境に適用するべく国内 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，I/O パフォーマンスの計測結果を示した．ローカルストレージへの NFS アクセスと本システムのパフォーマンスを比較し，本システムは他拠点への複製処理を行いながらもローカルストレージへの NFS アクセスと遜色ない R/W 性能を示すことを明らかにした．ある拠点で書き込まれたデータを他拠点から読み書きする際のパフォーマンス計測や，数十台規模の VM のグローバルマイグレーションをこの広域分散ストレージを用いて行った際の評価実験については今後の課題である．

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に感謝します．

参考文献

[1] 柏崎礼生: スモールスタートで始める大学の仮想化基盤の構築と運用の実情, インターネットと運用技術シンポジウム 2012 論文集, pp.94-101 (2012).
[2] 坂田智之, 長谷川孝博, 水野信也, 永田正樹, 井上春樹: 情報セキュリティの観点からみた静岡大学の全面クラウド化, 情報処理学会研究報告, 2011-IOT-14, Vol.7, pp.1 (2011).
[3] 松原義継, 大谷誠, 江藤博文, 渡辺健次, 只木進一: プライベートクラウドによる電子メール管理コストの低減とサービスレベルの改善 —佐賀大学の事例—, 情報処理学会研究報告, 2011-IOT-14, Vol.8, pp.1-6 (2011).
[4] Shikida Mikifumi, Miyashita Kanae ,Ueno Mototsugu, Uda Satoshi: An evaluation of private cloud system for desktop environments, Proceedings of the ACM SIGUCCS 40th annual conference on Special interest group on university and college computing services (SIGUCCS '12), pp.131-134 (2012).
[5] 宮下夏苗, 上埜元嗣, 宇多仁, 敷田幹文: 大学におけるプライベートクラウド環境の構築と利用, 第 3 回インターネットと運用技術シンポジウム, pp.17-24 (2010).
[6] 棟朝雅晴, 高井昌彰: 北海道大学アカデミッククラウドにおけるコンテンツマネジメントシステムの展開, 第 10 回情報科学技術フォーラム 情報科学技術レターズ pp.15-18 (2011).
[7] Daryl C. Plummer, Thomas J. Bittman, Tom Austin, David W. Cearley and David Mitchell Smith: Cloud Computing: Defining and Describing an Emerging Phenomenon, Gartner Research, G00156220 (2008).
[8] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski Gunho Lee, David A. Patterson, Ariel Rabkin, Ion Stoica and Matei Zaharia: Above the Clouds: A Berkeley View of Cloud Computing, UCB/EECS-2009-28 (2009).
[9] Lee Badger, Tim Grance, Robert Patt-Corner, Jeff Voas: DRAFT Cloud Computing Synopsis and Recommendation, NIST Special Publication 800-146 (2012).
[10] 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).
[11] 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).
[12] 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).