広域分散仮想化環境の展開・運用・管理コストの定量的評価

柏崎 礼生[1,a)] 北口 善明[2,b)] 市川 昊平[3,c)] 近堂 徹[4,d)] 中川 郁夫[1,5,e)] 菊池 豊[6,f)]
下條 真司[1,g)]

概要：広域分散アプリケーションの実証実験では，仮想的なネットワークを構築して検証を行うアプローチのほかに，実際にアプリケーションを稼働する基盤を地理的に広域に展開させて検証を行うアプローチがある．後者は現実的な環境での実証実験による説得力を獲得できる一方でその環境構築と維持のコストは無視できない．本稿では国内 11 拠点，海外 1 拠点が提供する計算機資源を SINET5 とインターネットキャリア網を用いて接続した広域分散仮想化環境の設計と運用コストを評価する．

quantitative evaluations of deploymental, operation and administration cost in a wide area distributed virtualization infrastructure

Hiroki Kashiwazaki[1,a)] Yoshiaki Kitaguchi[2,b)] Kouhei Ichikawa[3,c)] Kondo Tohru[4,d)]
Ikuo Nakagawa[1,5,e)] Yutaka Kikuchi[6,f)] Shinji Shimojo[1,g)]

Abstract: There are two approach to demonstrate experiments of wide area distributed appliactions. One is to demonstrate in virtual environments that emulate wide area distributed environments. The other is to demonstrate in real environments that consists of geographically wide are distributed computing resources and communication lines. Although the real environment can be persuasive to evaluate the experiments, the cost to install the environment can not be negligible. This paper shows a design and operational costs of a real wide area distributed environments with 11 domestic regions and one international region.

Keywords: wide area distribution virtualization infrastracture cost evaluation

1. 背景と目的

広域分散システムを代表するインターネットは継続的にその重要性を増している．2015 年の日本におけるインターネットのブロードバンドサービス契約者の総ダウンロードトラフィックは 5.4Tbps，総アップロードトラフィックは 1.1Tbps と推定された[*1]．全世界規模のインターネットトラフィックは 2016 年で平均 34Tbps に及び[*2]，今後 5 年間

[1] 大阪大学
Osaka University
[2] 金沢大学
Kanazawa University
[3] 奈良先端科学技術大学院大学
Nara Institute of Information Science and Technology
[4] 広島大学
Hiroshima University
[5] 株式会社インテック
Intec Inc.
[6] 高知工科大学
Kochi Insitute of Technology
[a)] reo@cmc.osaka-u.ac.jp
[b)] kitaguchi@imc.kanazawa-u.ac.jp
[c)] ichikawa@is.naist.jp
[d)] tkondo@hiroshima-u.ac.jp
[e)] ikuo@inetcore.com
[f)] yu@kikuken.org
[g)] shimojo@cmc.osaka-u.ac.jp

[*1] 総務省による報道資料「我が国のインターネットにおけるトラヒックの集計・試算」2016 年 3 月 2 日
http://www.soumu.go.jp/menu_news/s-news/01kiban04_02000103.html

[*2] Cisco Systems Inc.:“Cisco Visual Networkin Index”, June 1, 2016
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html
月あたり予測値 88.7EB から算出した．

で約 3 倍となることが予測されている．インターネットトラフィックの増大をもたらした要因の一つに携帯電話の普及が挙げられる．2014 年における世界の携帯電話普及率は 96.3%であり，日本や北米・欧州以外の地域における 2000 年と 2014 年の携帯電話の普及率を比較すると 21.7 倍の差 (契約数 2.6 億から 57.1 億) が生じている[*3]．インターネット利用者は広域に分散して存在するようになったが，クラウドコンピューティングの登場によりデータの集積とその解析は集中して行われるように思われた．しかし耐障害性を高める目的やエッジコンピューティング (フォグコンピューティング) への注目度が高まったことにより，データ処理基盤も広域に分散することが予測されている[*4]．

地理的に広域に分散した拠点が計算機資源とネットワーク回線を提供し合い構成される環境を広域分散環境と呼ぶ．この環境を用いて広域に分散した計算機資源を利用して動作するアプリケーションを広域分散アプリケーションと呼ぶ．インターネットの広範な普及に伴い，また自然災害に対する災害回復や有効な事業継続計画のために，広域分散アプリケーションは様々な領域で研究開発が行われている．日本においては WIDE プロジェクト[*5]がマルチメディア通信と分散処理 (DPS) 研究会で様々な研究を 1980 年代後半に発表した [1–4]．また同時期には広域分散環境下における協同作業を支援するシステムの研究も行われていた [5]．広域分散ファイルシステムに関する研究が始まったのもこの頃である [6–8]．1990 年代前半から中盤にかけてプラントのプロセス制御用計算機システムへの自律分散システムの適用に関する研究や [9]，マルチメディアを対象とした研究が登場した [10,11]．1990 年代後半から 2000 年代前半において広域分散環境が一部の大学や研究機関だけのものでなくなり，地域 IX を巻き込んだ相互接続に関する研究が登場する [12–14]．同時期にグローバルコンピューティング (グリッド) の研究開発が盛んになり [15]，広域分散ファイルシステムとして著名な Gfarm もこの頃に研究開発が開始された [16]．

2000 年代は，それまで培われた広域分散アプリケーションに関する研究をもとにして多様化した 10 年と言える．2011 年 3 月 11 日には東北地方太平洋沖地震が発生した．当時情報処理学会会長であった白鳥則郎先生は「情報処理」の Vol.52 No.7 における記事「被災記:再生を目指して -大震災から 50 日」において「災害にも強い広域分散」と記している [17]．この災害を一つの契機として災害時にサービスを継続するためのシステムとして広域分散ストレージシステムを災害回復用途に使う研究が増大する [18,19]．

[*3] 総務省: “平成 28 年度情報通信白書”http://www.soumu.go.jp/johotsusintokei/whitepaper/h28.html
[*4] IDC Japan 株式会社: “2016 年 国内エッジ／フォグコンピューティング市場分析” http://www.idcjapan.co.jp/Report/ICT/jpj40600816.html
[*5] http://www.wide.ad.jp/

拠点間を接続するネットワーク回線の特性は広域分散アプリケーションの品質に影響を及ぼす．単一の拠点が提供する計算機資源を利用して動作するアプリケーションと比較し，広域分散アプリケーションは広域分散環境での品質評価を行うことが求められる．広域分散アプリケーションの品質評価には 2 つの方法論がある．一つは拠点間を接続するネットワーク回線の特性を模倣する実際的なネットワークを単一の拠点が提供する計算機資源上で構築し，この実際的なネットワーク上で評価する方法．もう一つは，現実に構築された広域分散環境で評価する方法である．ネットワーク回線の特性の中には時間の経過とともに変化する要素があるため，現実の広域分散環境での評価はより強い説得力を持つ．複数の独立した組織からなる広域分散環境では，その構築と運用に要するコストは無視できない．

本稿では国内 11 拠点，国外 1 拠点からなる広域分散仮想化環境の設計を紹介する．この設計は拠点の拡大に対応可能なスケールアウト指向である点に特徴がある．展開と運用コストを低減する手法の有効性を定量的に評価する．

2. これまでの取り組み

著者らはこれまで国内外の大学や研究組織を相互接続して広域分散アプリケーションを評価する取り組みを行ってきた．1 つは広域分散ストレージと，この広域分散ストレージを共有ストレージとして利用する広域分散仮想化環境 “distcloud” である．もう 1 つは広域分散環境の拠点間に意図的なネットワーク障害を発生させることにより耐災害性・耐障害性を検証し，評価・反映を行うプラットフォーム “DESTCloud” である．本章ではこの 2 つの取り組みを紹介し，広域分散仮想化環境を用いた実証実験の有効性とその問題点について明らかにする．

2.1 distcloud

データセンターを利用した災害回復において，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，この方式による災害回復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．

そこで distcloud 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することで広域分散型の仮想化基盤を実現した．2013 年には国内三拠点 (広島大学，金沢大学，国立情報学研究所 (NII)) 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した．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 つ, L3VPN サービス 1 つ）を利用している．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．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．management 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 (VM Monitor: VMM) のためのセグメントである．management LAN は VMM の管理用セグメントとなり，migration LAN は VMM 上で動作する 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．

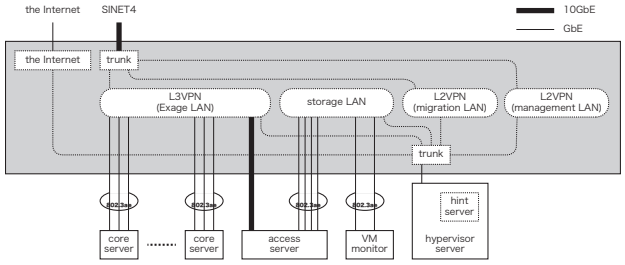

図 1 distcloud 広島大学拠点のネットワーク構成図 (2013 年)
Fig. 1 Distcloud network diagram of Hiroshima University in 2013

図 1 は SINET4 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．各拠点には，拠点内の core server，access server，および VMM の死活監視と統計情報を収集する hint server を設置する．hint server は VM であり，hypervisor server 上で動作する．各拠点では access server 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，access server 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．access server は 1 ポートの 10GbE および 4 ポートの GbE のリンクアグリゲーション (IEEE802.3ad)，core server は 3 ポートの GbE のリンクアグリゲーションにより集約スイッチに接続する．また，access server を NFS マウントする VM monitor は 2 ポートの GbE 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．

distcloud で構築した広域分散ストレージの性能を評価するために iozone[*6] を用いた計測実験を行った．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，広域分散に対応していない方式と広域分散対応方式の両方で性能を評価した．distcloud 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．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した．

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 [19].

2.2 DESTCloud

distcloud での検証をもとに筆者らは「広域分散仮想化環境が災害回復において役立つことを実証するためにはどうすればいいか」についての議論を開始した。

2012 年 8 月に内閣府が発表した「南海トラフの巨大地震

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

による津波高・浸水域等 (第二次報告) 及び被害想定 (第一次報告)」によると，想定されるケース「『四国沖』に『大すべり域+超大すべり域』を設定」および「『四国沖〜九州沖』に『大すべり域+超大すべり域』を設定」において高知県幡多郡黒潮町および土佐清水市は最大津波高 (満潮位・地殻変動考慮) において国内最大の 34m と推定されている[*7]．高知県では県内の高等学術機関 5 組織が協働し，高知 IX，高知 PoP，南国 PoP を連携したさせた冗長構成でインターネットや相互接続を実現している．既存の ICT システムに意図的に障害を起こすことにより，ICT システムの冗長性や障害対策の機能および ICT 関係者間での連絡体制等を確認・検討し，実際に機能する事業継続計画を策定することを目的として，TEReCo4[*8] プロジェクトは 2013 年度にネットワーク防災訓練を行った[*9]．この取り組みでは様々な障害要因をロジックモデル手法で可視化しており，3 つの障害パターンでの検証が行われた．その結果，本来不通になるはずの障害パターンにおいてもインターネットへの導通が確認されたことにより，運用者が把握していない冗長構成の存在が発覚するなど，防災訓練を行うことで，耐災害性・耐障害性を向上させるのみならず，本来の目的以外の効果が現れた事例の報告が行われている [20, 21]．

このような活動は，以下を確認し，課題がある場合は改善するための材料とすることを目的として行われる．

- ICT システムの障害に対する機能が正しく機能するか
  - 冗長性によって機能維持をする場合には，予備系が正しく機能するか
- ICT システムの障害発生時に，システムが障害を通知し，管理者が正しく障害を認識し，原因特定・復旧にいたる手段を実施できるか
  - 直ちに復旧できない場合に代替手段を講じるようになっている場合には，それを実施できるか

これまで ICT システム管理者を対象としてこのようなネットワーク防災訓練の実現が提案されてきたが[*10]，その反応は，活動の意義や効果は認めるものの，実施することに対する困難が大きいというものであった．また，ユーザが 365 日 24 時間利用しうるため，機能停止を伴う活動が困難であるという声も大きかった．ネットワーク防災訓練を実際に実施することで，ネットワーク防災訓練の有効性を実証することができ，また，ネットワーク防災訓練の方法論を確立して実施の手続きを整備し，小さな手間で実施することが可能となる．高圧以上の受電をしている組織が，電力設備の法定点検を義務付けられているのと同様に，一定以上規模の ICT システムを運用している組織に対する法令による点検義務が課せられるようになれば，特に地震の発生件数の多い環太平洋地域の国々において有益であると考えられる．

高知県でのネットワーク防災訓練ではネットワーク障害の発生は人為的に，かつ実際に人間の手による手動操作で行われたものである．しかし例えば組織内においてネットワークトラフィックの少ない時間帯である深夜から朝にかけての時間帯において障害を発生させようとすると，必然的に人力で訓練のための障害を発生させるコストを要する．前述の取り組みでも障害がもたらす影響を加味して，1 月 5 日の午前 5 時から訓練を開始しているため，定期的に行うことは困難であることが考えられる．我々は，この訓練が定常的に高い頻度で，しかも多様な障害で行われることが必要であると考えている．人力で行うためには多様な障害シナリオを記述するコストが必要であり，なおかつ高い頻度で障害を発生させ，その評価を行い，検証するコストを算出すると，人力で実現することは現実的ではない．また先の検証においては複数箇所で同時に発生する障害を人為的に作り出すことや，障害後のネットワーク情報を収集すること，そして障害発生後に元の状態へ戻すことの困難さを明らかにしている．

情報システムの中でも特に分散システムは防災訓練を行うために複数のステイクホルダーの了承を得ることが求められるために困難であったが，それと同時に高い耐災害性・耐障害性が求められるシステムでもある．このような動機付けのもと，障害を形式的に記述し，かつ障害時にネットワーク情報を収集することが可能であり，評価後にネットワーク状態を元に戻すことが可能な，分散システムの耐災害性・耐障害性の検証・評価・反映を行うためのプラットフォーム “DESTCloud” が開発された．

発生区分	障害要因	症状	実装する機能
制御・運用・ソフトウェア	通信規制制御	輻輳	遅延発生+n% パケロス
			トラフィックシェープ
	不正な経路伝搬	経路ループ	RIB/FIB 強制書換
		経路フラップ	
		経路障害 (宛先不達)	
ネットワーク機器	装置故障 (全体)	通信断 (全体)	インターフェイスダウン
	装置故障 (部分)	通信断 (部分)	
	リソース過負荷	パケットロス	n% パケロス
		遅延増大	遅延追加
通信回線	拠点間通信ケーブル断	通信断 (部分)	インターフェイスダウン +100% パケットロス
	中継器・交換機故障		
	トラフィックの集中	輻輳	遅延発生+n% パケロス
			トラフィックシェープ
設備環境	局舎損壊	通信断 (全体)	インターフェイスダウン +100% パケットロス
	電源喪失		
	空調故障		
		通信断 (部分)	

図 2 DESTCloud におけるネットワーク障害の分類
Fig. 2 Classification of network disorder in DESTCloud

*7 南海トラフの巨大地震による津波高・浸水域等 (第二次報告) 及び被害想定 (第一次報告) 資料 1-2 都府県別市町村別最大津波高一覧表＜満潮位＞
http://www.bousai.go.jp/jishin/nankai/taisaku/pdf/1_2.pdf

*8 Traffic Engineering for Regional Communities, version 4

*9 福本昌弘ら「災害時に事業継続性を発揮する情報通信インフラのための運用計画改善手法および冗長化技術の研究開発」総務省地域 ICT 振興型研究開発 (平成 25 年度〜26 年度, 四国総合通信局)
http://www.soumu.go.jp/main_content/000284013.pdf

*10 菊池豊「防災訓練! 本当に切るとしたら何処を切りたい何を知りたい?」裏ジャノ 2013@ミクシィ，など．

2.2.1 設計と実装

災害がもたらすネットワーク障害の表現として，自然災害に起因する障害から装置故障等に起因する障害まで様々な要因が考えられる．障害パターンの影響範囲，空間的な変化の有無，時間的な推移などの検討も必要となる．総務省「大規模災害等の緊急事態における通信確保の在り方に関する検討会」で示されている災害事象[*11]や「情報通信ネットワーク安全・信頼性基準」[*12]の内容をもとに災害時における通信設備等に対する障害に焦点を絞って，各事象に対してネットワーク装置に適用する制御について検討した (図 2)．通信障害は主に 2 つの原因に大別される．ひとつはトラフィック集中による輻輳，もうひとつは回線や機器等のハードウェア・設備障害である．図 2 はこれらを細かな区分に分類し，それぞれの障害要因と具体的な症状の対応付けを行い，各々について本プラットフォームで実装する機能についてまとめたものである．通信断であっても，ネットワーク機器自体が故障する場合と，中継機・交換機が故障する場合では実際の通信において観測される症状が異なることが予想される．本プラットフォームでは，このような症状も模倣できるパターンの分類を検討した．

2015 年に行われた初歩的な実装では，図 2 で定義したネットワーク障害のうち，“インタフェースダウン”，“100%パケットロス” および “経路表強制書き換え” という 3 つの障害を実装した．“インタフェースダウン” と “100%パケットロス” は，それぞれルータの “shutdown コマンド” と “ACL による 100%パケットロス” と同様の制御を行う実装としている．また，“経路表強制書き換え” は，CLI における “static route の追加” と同様の制御を行うことで，動的経路に対して administrative distance が小さい経路を設定し，経路の上書きを実現している．以下に，実装したコマンドとその引数を記す．

インタフェースダウン

```
linkctl.py {down|up} <router> <ifid>
```

100%パケットロス

```
pktloss.py {100|0} <router> <ifid>
```

経路表強制書き換え

```
dcroute.py {add|dell}\
<static_route> <router> <ifid>
```

これらのコマンドは制御用サーバにて実装しており，各障害をコマンドラインで実行可能としている．そのため，複数の障害を時系列に発生させる障害シナリオをプログラムとして記述することができる．さらに，再現性を有する障害処理を繰り返し実行することが可能となる．図 3 に，構築した障害発生プラットフォーム構成を示す．このプラットフォームは，JGN-X[*13]上に配置された onePK 対応ルータを利用して構築し，札幌，仙台，大手町，名古屋，大阪，岡山，広島，福岡の各アクセスポイント (AP) で利用可能とした．配置されているルータの構成は拠点毎で異なっており，大手町 AP および大阪 AP の 2 カ所が Cisco ASR9006，その他の AP では Cisco ASR1004 となっている．このような JGN-X の環境を利用し，全国 5 拠点のユーザセグメント (広島大，高知工科大，大阪大，NAIST，金沢大) とルータを結ぶ論理パスをそれぞれ構築し，複数の論理パスで各ルータを結ぶネットワークとして構築した．金沢大のみ JGN-X への直接の接続性を持たないため，SINET4[*14]による L2VPN サービスを経由した接続形態としている．

図 3 JGN-X と SINET4 を用いた DESTCloud の論理パス構成
Fig. 3 Logical topology of DESTCloud on JGN-X and SINET4

2.2.2 検証

distcloud のような広域分散アプリケーションにおいて，アプリケーションを構成する広域分散拠点のトポロジが 2 つ以上の独立した領域に分断されるとスプリットブレインシンドロームと呼ばれる状態に陥り，メタデータのトランザクションを停止しなければ完全性を保つことができなくなる．広島 AP がダウンすることで広島拠点が切り離される障害を想定し，広島 AP にあるルータに対してインタフェースダウンの障害を発生させる．この障害ではルータのインタフェース 2 つに対して同時に障害を発生させた．今回の構成では，スプリットブレインの検出に用いるキャッシュサーバを 7 台設定し，金沢拠点に 3 台，広島と NAIST 拠点に 2 台づつ配置している．上記の障害により，

[*11] 総務省「大規模災害等緊急事態における通信確保の在り方に関する検討会」
http://www.soumu.go.jp/main_sosiki/kenkyu/saigai
[*12] 総務省「情報通信ネットワーク安全・信頼性基準」
http://www.soumu.go.jp/menu_seisaku/ictseisaku/net_anzen/anshin
[*13] http://www.jgn.nict.go.jp
[*14] http://www.sinet.ad.jp

金沢拠点と NAIST 拠点が接続される側が多数派となることから，広島拠点のコアサーバ群が強制停止 (フェンシング) 対象となる．

検証の結果，少数派となる広島拠点がフェンシング処理により強制停止されることが確認でき，広島拠点が切り離されている状態においても，広島拠点で作成されたファイルを読み出すことができた．これは，ブロックデータの多重度が 3 であることから，すべての拠点においてレプリケーションデータを保持していることに起因している．また，障害復旧後にフェンシングからの復旧 (再起動) を実施した際のデータベース競合も発生することなく，読み出しが可能であることも確認している．ただし，金沢拠点から広島拠点で作成したデータを読み出す性能は，49.5 MB/s (定常時: 65.9MB/s) に落ち込む結果となった．加えて，読み込み中に障害を発生させた場合を確認したところ，読み出し処理に 2 分以上要する結果も確認したが，読み出し不能とはならなかった．次に，拠点間通信断による耐障害性検証を同じ障害シナリオにて三度繰り返し行ったところ，その一回において対象ファイルが読み出し不能となる事象が確認された．この現象は，現在の実装において想定外の挙動であったことから，詳細なログ解析とコード解析が必要となった．

広域分散ストレージの性能評価として，通信経路の変動が書き込み性能に与える影響を評価する実験を行った．名古屋 AP と大手町 AP 間に通信障害 (名古屋 AP のルータに対して 100%パケットロス) を発生させ，BGP により金沢大–奈良先端大の経路に対して迂回経路を取らせ，通信経路の変動による影響を観測する．なお，書き込み処理の影響を観測するため，金沢拠点においてファイルを作成中に障害を発生させている．検証の結果，障害発生により BGP の経路再計算が行われ，その間拠点間の通信断により書き込み処理が中断されるが，障害復旧後は処理が継続されることが確認できた．経路変更により金沢拠点と奈良拠点間が通信不能となった時間は 22.1 秒であった．障害に伴う経路変更が実施された後の書き込み性能を計測すると 54.3MB/s となり，定常時の 59.5MB/s と比較して若干の低下が観測された．

2.3 問題点と課題

distcloud が広域分散仮想化環境と名乗っているのは distcloud が提供する共有ファイルシステムを利用して地理的に広域に分散した拠点で同じ VM を起動することができる，という特性に着目したからである．その具体的な応用事例が長距離ライブマイグレーションの実証実験だが，共有ファイルシステムの構築に力点が置かれ，その上で動作する仮想化基盤，およびその上で動作するアプリケーションを支える VM については重視しない構成であったと言える．そのため，広域分散環境を用いた耐災害性・耐障害性検証プラットフォームである DESTCloud は distcloud とは別のネットワークとして構築する必要があり，広域分散仮想化環境としての特性を活用することが難しかった．

distcloud と DESTCloud の取り組みを通して得られた広域分散仮想化環境の構築とその運用における問題点は以下の通りである．

- 1 つの物理計算機に 1 つの OS を入れ，広域分散仮想化環境を構成する 1 ノードとしているため，セットアップに要するコストが大きい．
- 各拠点が保有する計算機資源の内容や量が多様であり，各拠点に応じた設計をしなければならない．
- 研究用の回線と接続するための組織内ネットワークの手続きが各拠点によって異なるため，構成に時間を要し，他拠点は手助けすることが困難である．

そこで広域分散アプリケーションの検証実験を行おうとするユーザに対して強いる導入コストを低減した次世代 distcloud を設計する．

3. 次世代 distcloud の設計

distcloud の実証実験を開始した 2012 年当時，distcloud を構成する core server は要求する I/O 性能が高く，当時としては比較的大容量のメモリ空間を必要とした．ストレージの I/O 要求に対する仮想化オーバーヘッドによる性能低下といった外乱要因を排除したい要望もあり，また大容量のメモリを調達する投資対効果に対する疑念もあり，実験を開始した 2012 年度は物理計算機による構成を採用した．その後の検証により，準仮想化技術を用いたストレージへの直接的な I/O 要求 (KVM における virtio など) を用いることにより物理計算機と同等の I/O 性能を確保できることが定量的に示された．また物理計算機では CPU の使用率がそれほど高くないこと，多くのコアを利用しないことも計測から把握することができた．半導体ドライブ (Solid State Drive: SSD) やメモリ価格の下落に伴い投資対効果も認められるに至った．そこで次世代の distcloud では，構成するノードを全て仮想化する指針を定めた．VM の雛形を作ることにより，distcloud の core server のように複数台必要となる VM を少ない工数で大量に作ることも仮想化の利点として想定することができる．

3.1 仮想化ホストの仕様

DESTCloud の研究を開始する際，2014 年度に大阪大学拠点に次世代 distcloud のための計算機資源を導入した．1 つの物理ホストで 8〜16VMs 程度を提供できる計算機資源を想定した．スペックは表 1 の “2014” の項目に示す．

このサーバ製品ではドライブを購入しなければドライブベイにドライブを設置できないため 8 基の 250GB HDD を備えた構成で購入したが，この 8 基の HDD はバルクで購入した 1TB の SSD と換装された．8〜16VMs 程度の稼働

＼	2014	2014 増強	2015	2016 以降
RU	2	2	2	1
CPU	10C 2.4GHz 2 sockets	10C 2.4GHz 2 sockets	14C 2.0GHz 2 sockets	16C 2.3GHz 2 sockets
memory	96GB	192 GB	128 GB	256 GB
storage	1TB x8	1TB x8	1TB x8	1TB x8
network	10GbE x2 + GbE x2	10GbE x2 + GbE x2	10GbE x2 + GbE x2	10GbE x2 + GbE x2

表 1 次世代 distcloud における仮想化ホストのスペック変遷
Table 1 Changes of spec for virtualization host machines on distcloud next generation

を想定した計算機であったがこのスペックで 16VMs を稼働させる場合，1VM あたり 2 物理コア以上を専有することはできず，全ての VM が 2 物理コア以上を要求する場合は CPU のオーバーサブスクリプションが発生する．サーバ集約のための仮想化ホストの設計においては CPU のオーバーサブスクリプションを許す構成とすることで CPU リソースの有効活用を目指すものがある．しかし研究用途における仮想化ホストにおいては CPU のオーバーサブスクリプションは仮想化に伴うオーバーヘッドとして評価結果に影響を及ぼす可能性があるため，1 物理コアと 1 仮想 CPU の組み合わせとなる構成が望ましい．1VM あたりの平均メモリ容量は 6GB であり，これは研究用サーバ用途では大きな容量とは言い難い．SSD は複数の VM で共用しなければならない．

そこで表 1 の “2014 増強” の項目に示すスペックで 2015 年度に 2014 年度調達機器を増強した．これにより，1VM で平均 12GB のメモリを確保することができるようになり，また 1TB の SSD を専有して扱うことができるようになった．しかし 20 物理コアという点に CPU 性能の上限があり，これは解消されなかった．このサーバ製品が 2015 年に EOL を迎え，その後の CPU リリースに追従しないことが発表されたことに起因する．そのため 2015 年度に distcloud 接続拠点を増やした際にはこの上限を考慮し，より物理コア数の多い CPU を搭載できるサーバ製品を調達した．2015 年度設置モデルのスペックは表 1 の “2015” の項目に示す．このスペックに従い，2015 年度では琉球大学拠点，九州大学拠点，および東北大学拠点が新たに増設された．

2014 年度から 2015 年度の DESTCloud 研究推進に伴う distcloud 拠点の拡大において得られた知見として，研究用途のサーバでは 1VM あたり 32GB のメモリ容量が求められることもあり，1 仮想化ホストあたり 16VM の収容は現実的ではないという点が挙げられる．このため，1 の “2016 以降” で挙げられているように，今後の distcloud 拠点では 1RU のサーバ製品として，8VMs の収容を想定した構成とする．16 物理コア以上の CPU を 2 ソケット搭載することにより 1VM あたり 4 仮想 CPU を割り当てても CPU のオーバーサブスクリプションは発生しない．メモリは 256GB 以上を搭載するため 1VM あたり 32GB のメモリを確保することができる．

3.2 ネットワークの設計

研究用途で用いるノードは様々なネットワーク構成を要求される．それに対応するため distcloud では SINET4 の L2VPN，L3VPN サービス，および JGN-X の L2 接続性，IP 仮想化サービスを用いた．しかし研究プロジェクトごと必要な VLAN をこれらのサービスに対して申請し，導通の確認を行うことが広域分散仮想化基盤を構成する各々の拠点に求められるため，この申請コストの大きさが拠点担当者にとって負担となっていた．そこで 2016 年度は研究プロジェクトの追加が発生しても拠点の申請が増大しないよう，SINET5 の L2VPN は 2 つ，L3VPN を 1 つ申請するものと固定し，L2VPN のトラフィックを仮想化ホスト上で動作する Linux ベースのネットワークオペレーティングシステムである Vyos[*15]で集約する構成とした．Vyos では IEEE802.1ad (QinQ) を扱い，研究プロジェクトごとに必要とされる VLAN は広域分散仮想化環境内のみで唯一性のある VLAN として取り扱うこととした．このように簡潔な構成とすることにより，一定水準のネットワーク技術，サーバー技術や継続的な運用を行うことのできる人材がいなくとも distcloud 拠点として新たに参画することができるようになった．

これまでの SINET4/JGN-X 利用において研究プロジェクトのネットワーク構成の決定から実証実験までに要する時間は図 4 のように構成されていた．研究プロジェクトごとに VLAN ID，IPv4 アドレス帯の認識を共有し割当を決定し，これをもとに各拠点が SINET4/JGN-X に申請手続きを行う．申請が受理されてから拠点と接続性のあるネットワーク機器に設定が投入されるまで 2 週間以上かかり，この間は作業待ちが発生する．先方から指定された日付時間で導通試験が行われ，要求した仮想ネットワークサービスが開始される．この後，実証実験のための環境構築が行われ，実証実験が行われる．この間，計算機資源とネットワーク資源の分離が行われていないため，相互の研究プロジェクトの実証実験の影響を除外するため他の実証実験は行うことができない．また，新たな研究プロジェクトを始める場合，以前行った SINET4/JGN-X の仮想ネットワークサービスを再利用することも可能だが，構成によっては再度申請する必要がある．申請手続きと作業待ちが前回同様発生し，導通試験と環境構築を経て実証実験を行う．

一方で提案する次世代 distcloud では広域分散仮想化基盤を構成する拠点は既に述べられたネットワーク構成に従

[*15] https://vyos.io/

い拠点の機器を設定すればよく，申請手続きに必要な書類の雛形も既に用意されている．仮想ネットワークサービスが提供されるまでの作業待ちは発生するが，導通試験まで終われば拠点担当者の作業は終了となる．あとは実証実験を行いたいユーザが広域分散仮想化環境内で VLAN ID やアドレス帯がバッティングしないようにネットワーク構成の認識合わせを行い，必要な VM を確保し，実証実験を行う．計算機資源とネットワーク資源は各プロジェクトごとに分離されており互いに影響を及ばさないため，資源がある限りにおいては複数の実証実験を並行して進めることが可能である．

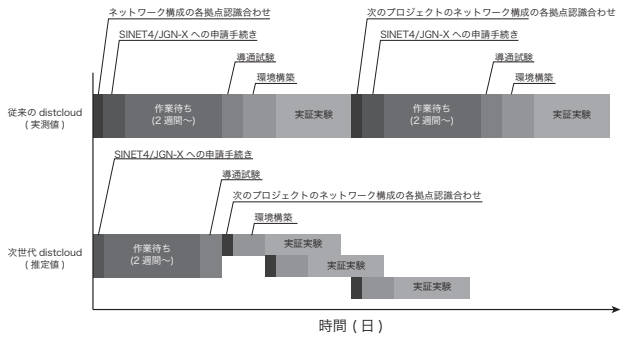

図 4 新旧 distcloud の運用コスト比較

Fig. 4 Comparison of operation cost between old and next generation distcloud

実測値として，2014 年の distcloud 論文 [19] で広域分散ストレージの検証試験を行うために要した時間を distcloud ML で交わされたメールから計測したところ，認識合わせから実証実験の開始まで 98 日間を要している．一方，2016 年に行われた DESTCloud 検証で異なる広域分散ストレージの検証試験を行うために要した時間は認識合わせから実証実験の開始まで 33 日間であった．

4. まとめと今後の課題

本稿では広域分散仮想化基盤における検証実験の重要性を主張し，運用コストの問題点について指摘した．運用コストを低減するための次世代 distcloud の定量的な設計を示し，実際に実証実験を行うまでに要する時間の低減に寄与したことを示した．今後は統合環境を用いて実証実験をより容易に実現する手法の研究開発を推進する．

参考文献

[1] 村井純, 田中啓介: 広域分散環境における資源管理, 情報処理学会研究報告マルチメディア通信と分散処理 (DPS), Vol. 1099, No. 11, pp. 73–80 (1988).
[2] 村井純, 加藤朗, 佐藤智満, 楠本博之, 山口英: 大規模広域分散環境 WIDE の構築, 情報処理学会研究報告マルチメディア通信と分散処理（DPS）, Vol. 1989, No. 41, pp. 55–64 (1989).
[3] 本田和弘, 加瀬直樹, 尾上淳, 中島達夫, 所真理雄, 村井純: WIDE 上の X.25 機能の設計と実装, 情報処理学会研究報告マルチメディア通信と分散処理（DPS）, Vol. 1989, No. 41, pp. 65–72 (1989).
[4] 嵯峨和幸, 石河裕子, 海野英俊, 佐藤智満 , 村井純: WIDE 資源管理機構, 情報処理学会研究報告マルチメディア通信と分散処理（DPS）, Vol. 1989, No. 41, pp. 73–80 (1989).
[5] 鳩野逸生, 上田鉄雄, 阪田史郎: グループ協同作業支援のためのマルチメディア在席対話システム, 情報処理学会論文誌, Vol. 30, No. 4, pp. 527–535 (1989).
[6] Spector, A.Z., Kazar, M.L.: Wide Area File Service and the AFS Experimental System, Unix Review, Vol. 7, No. 3 (1989).
[7] Cate, V.: Alex - a global filesystem, Proceedings of the Usenix File Systems Workshop, pp. 1-11 (1992).
[8] 鈴木重治, 小泉一郎, 内田昭宏: 広域分散ファイルシステム MUFS とその適用, マルチメディア通信と分散処理ワークショップ論文集, pp.153–160 (1993).
[9] 土井幸一, 佐久間信晴, 堀真司, 森欣司, 鈴木靖雄: 自律分散システムの鉄鋼への適用事例, 第 42 回全国大会講演論文集, pp.311–312 (1991).
[10] 寺西裕一, 松浦敏雄, 下條真司, 谷口健一: 広域分散型マルチメディアシステムのための情報間同期機構の実現, 全国大会講演論文集, Vol.49, No. データ処理, pp.343–344 (1994).
[11] 田中裕之, 吉川耕平, 岡村耕二, 荒木啓二郎: 分散環境におけるマルチメディア処理機構の実装と評価 -情報提供システム Cosaic-, マルチメディア通信と分散処理ワークショップ論文集, Vol. 1994, No.1, pp.131–140 (1994).
[12] 中川郁夫, 林英輔, 樋地正浩, 八代一浩, 菊池豊, 西野大: ギガビットネットワークを用いた地域間相互接続の試み, 情報処理学会研究報告インターネットと運用技術（IOT）, Vol.1999, No.77, pp.7–12 (1999)
[13] 中川郁夫, 米田政明, 安宅彰隆: 地域 IX による安定した地域内通信環境の実現と評価, 情報処理学会論文誌, Vol.42, No.12, pp.2887–2896 (2001).
[14] 中川郁夫, 江崎浩, 菊池豊, 永見健一: MPLS を用いた広域分散 IX の実現, 情報処理学会論文誌, Vol. 43, No. 11, pp. 3519–3529 (2002).
[15] 竹房あつ子, 合田憲人, 松岡聡, 中田秀基, 長嶋雲兵: グローバルコンピューティングのスケジューリングのための性能評価システム, 情報処理学会論文誌, Vol. 41, No. 5, pp. 1628–1638 (2000).
[16] 建部修見, 森田洋平, 松岡聡, 関口智嗣, 曽田哲之: 広域大規模データ解析のための Grid Datafarm アーキテクチャ, 情報処理学会研究報告ハイパフォーマンスコンピューティング（HPC）, Vol. 2001, No. 77, pp. 177–182 (2001).
[17] 白鳥則郎: 被災記: 再生を目指して -大震災から 50 日, 情報処理, Vol. 52, No. 7, pp. 765-767 (2011).
[18] 石津晴崇, 永岡孝, 大西健司, 高杉英利, 建部修見: 耐障害性を高めた分散ストレージシステムの開発とその評価, Vol. 2011-HPC-130, No. 36, pp. 1–8 (2011).
[19] 柏崎礼生, 北口善明, 近堂徹, 楠田友彦, 大沼善朗, 中川郁夫, 阿部俊二, 横山重俊, 下條真司: 広域分散仮想化環境のための分散ストレージシステムの提案と評価, 情報処理学会論文誌, Vol.55, No.3, pp.1140–1150 (2014).
[20] 岡村健志, 菊池豊, 福本昌弘, 豊永昌彦, 佐々木正人, 今井一雅, 山田覚, 風間裕, 一色健司, 名和真一, 高畑貴志: 地域 IX における人為的障害による耐災害性の検証, マルチメディア，分散，協調とモバイル (DICOMO2014) シンポジウム, pp.485-489 (2014)
[21] 菊池豊, 岡村健志, 福本昌弘, 豊永昌彦, 佐々木正人, 今井一雅, 山田覚, 風間裕, 一色健司, 名和真一, 高畑貴志, 栢分正人, 井上望美, 柴田祐輔: 地域 IX で恣意的な障害を発生させることによる耐障害性の検証, ITRC technical report 2013 (2014)