

# 高機能高可用性ストレージシステムのための SDN型ストレージ間通信方式の設計と実装

和泉 諭<sup>1,a)</sup> 江戸 麻人<sup>2,b)</sup> 阿部 亨<sup>1,2,c)</sup> 菅沼 拓夫<sup>1,2,d)</sup>

**概要:** データの複製やバックアップ方式の一つとして、ローカルクラウドと遠隔クラウドを組み合わせたハイブリッドクラウドが注目されている。しかし、東日本大震災において、ローカルクラウドの破損や広域ネットワーク障害による遠隔クラウドへのアクセス不可により、住基情報や医療情報など緊急性・機微性の高いデータにアクセスできず大きな問題となった。そこで、我々はデータを地理的に分散した複数の拠点に効率的にバックアップを行うことで、災害時においても被災直後から必要なデータにアクセスできる高機能高可用性情報ストレージ基盤技術の研究開発を行っている。本研究ではその実現に向けてネットワーク基盤技術を対象とし、データ転送の経路を多重化し、さらにネットワークの利用状況に応じて動的に経路を切り替えるスマートルーティングを提案する。これにより平常時にはデータの高速転送化を図り、被災時には残存機器を活用して最低限の転送性能を確保する。本稿では、Software Defined Network (SDN) を用いてスマートルーティングの設計と実装を行う。さらに、平常時や災害時を想定したシミュレーション実験により、提案手法の効果を検証する。

## 1. はじめに

データの複製やバックアップ方式の一つとして、プライベートクラウドとパブリッククラウドを組み合わせたハイブリッドクラウドが注目されている。これはローカルストレージ（プライベートクラウド）に対して、定期的にデータの複製を行いながら、必要に応じて、インターネット上に存在する遠隔ストレージ（パブリッククラウド）に対してもデータを複製することで、セキュリティを保ちつつ、運用コストを抑えながら、効果的にデータの管理やバックアップを行うことが可能となる。

しかし、東日本大震災において、ローカルストレージの破損により部分的にデータが失われてしまい、さらに、ネットワークの途絶により、遠隔クラウドにアクセスできない状況が発生した [1]。その結果、住基情報や医療情報など緊急性・機微性の高いデータにアクセスできず大きな問題となった。そのため、耐災害性の高い情報ストレージシステムの実現が求められている。

そこで、我々は災害時における機器損壊と広域通信途絶下でも被災直後から必要なデータにアクセスできる高機能高可用性情報ストレージ基盤技術の研究開発を行っている [2]。この目的としては、データを地理的に分散した複数の拠点に効率的にバックアップを行い、災害直後においても残存データの再構成によって、迅速に重要なデータへアクセスすることが可能な対災害性を強化したストレージシステムの実現が挙げられる。

本研究ではこのストレージシステムのためにネットワーク基盤技術に着目する。具体的には、データ転送の高速化・効率化のためにネットワークのトポロジに応じてデータ転送の経路を多重化し、さらにネットワークの利用状況に応じてデータ転送の経路を動的に切り替えるスマートルーティングを提案する。スマートルーティングの実現により、平常時にはデータの高速転送化を図り、被災後においても残存機器を活用して最低限の転送性能を確保する。

本稿では、ネットワークをソフトウェアレベルで柔軟に制御可能な Software Defined Network (SDN) を用いてスマートルーティングの設計と実装を行う。さらに、平常時や被災時を想定したシミュレーション実験を実施し、平常時においてはデータの高速転送化を実現し、災害時においても可能な限りの性能を確保できることを示す。

<sup>1</sup> 東北大学サイバーサイエンスセンター  
Cyberscience Center, Tohoku University

<sup>2</sup> 東北大学大学院情報科学研究科  
Graduate School of Information Sciences, Tohoku University

a) izumi@ci.cc.tohoku.ac.jp

b) asato@ci.cc.tohoku.ac.jp

c) beto@cc.tohoku.ac.jp

d) suganuma@cc.tohoku.ac.jp

## 2. 関連研究と課題

### 2.1 高機能高可用性情報ストレージ基盤技術の開発

本節では我々が取り組んでいる高機能高可用性情報ストレージ基盤技術の開発について説明する。本ストレージ基盤技術はデータを地理的に分散した複数の近隣拠点に効率的にバックアップを行い、災害直後においても残存データの再構成によって、迅速に重要なデータへアクセス可能な対災害性を強化したストレージシステムを実現するための基盤となる技術である。これにより、半数の機器損壊と広域通信途絶下でも被災直後から必要な情報にアクセスできることを目指している。近隣ストレージを活用することで、地域ネットワークがダウンしている、もしくは地域ネットワークに接続していない拠点の場合でも、近隣の複製先拠点へ出向いて業務を再開することが可能となる。

本ストレージ基盤技術の実現に向けて以下の項目について、研究開発を行っている。

- 高可用性の耐災害性強化ストレージシステム [3], [4]
 

大規模災害の被災地において、情報サービスを提供可能なストレージシステムの実現に向けて、地理的条件などをもとにして、同時被災リスクが低いと判断された近隣ストレージにデータのバックアップを行う機能の開発を行っている。さらに、各拠点の被災・障害状況に応じて、残存拠点ストレージから代替拠点へデータのリストアを行う機能の開発を行っている。
- ディスク内及びディスク並列化によるストレージ機器の高速データ転送
 

ディスク装置内部で複数トラックを並列に処理し、さらに分散ファイルシステムによるディスク装置の並列化によりデータ転送を高速化する機能の開発を行っている。
- 高速ストレージ間通信方式
 

ストレージ間の通信を効率化・高速化するために、回線のトラフィック状況や優先度に応じて、高速に転送できる経路に動的に切り替えるネットワーク基盤技術の開発を行っている。
- 高機能プログラミングフレームワーク
 

高可用ストレージを柔軟かつ安全に使えるプログラミング環境を実現するために、高機能プログラミングフレームワークの開発を行っている。
- 投薬情報システムを用いた高可用ストレージ実証実験 [5]
 

高機能高可用性情報ストレージ基盤上に投薬情報システム（電子お薬手帳）を開発し、実証実験を通してストレージ基盤技術の評価を行っている。

本研究では、上記の研究開発項目のうち、高速ストレージ間通信方式を対象とする。

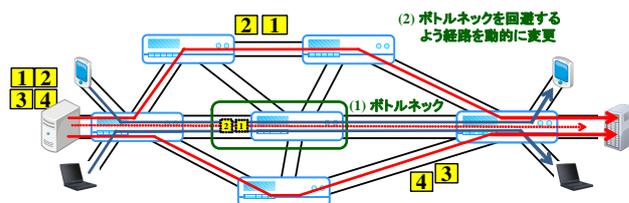


図 1 スマートルーティングの概要

### 2.2 高速ストレージ間通信に関する関連研究

ストレージ間通信の高速化のためのアプローチの一つとして、送信元から送信先までの複数の経路のうち、高速な回線を有する経路を選択して利用することが挙げられる。例えば、データセンターのボトルネックを回避して広帯域な経路を選択する手法 [6] や、経路長が最少かつ全体の帯域が最大となるようヒューリスティックに経路を選択する手法 [7] などが挙げられる。

また、利用可能な複数の経路を同時に利用する（経路多重化）ことにより、ストレージ間通信を高速化するアプローチも存在する。例えば、アプリケーション毎の伝送品質を考慮して、多重経路を選択する手法 [8] や送信先を複数設定し、それぞれに対してデータを分割して、多重経路でデータを送信する手法 [9] などが挙げられる。

### 2.3 課題

上記で述べた手法では、予め決められたネットワークのトポロジや帯域のみを考慮して経路を選択している。しかし、実際のネットワークにおいては、様々なトラフィックが流れており、ネットワークの利用状況は絶えず変化している。また、災害時においては、ネットワーク上のリンクやスイッチの途絶・復旧により、利用可能なネットワークのトポロジも変化することが考えられ、それにより、利用可能な経路も絶えず変化する。そのため、絶えず変化するネットワークのトポロジや利用状況に応じて動的に経路を選択することが困難である。

## 3. スマートルーティングの設計

### 3.1 概要

上述した課題を解決し、ストレージ間通信の高速化を実現するために本研究ではスマートルーティングを提案する。本手法は、これまで静的・固定的に割り当てられていたネットワーク経路について、経路の利用状況をリアルタイムに把握し、高速に転送できる複数の経路を動的に設定する。本手法は経路探索機能と経路選択機能から構成される。

経路探索機能は送信元から送信先への多重経路表を計算する。具体的には、送信元から送信先までの間の不要な経路を全て取り除き、利用可能な経路のみを抽出しながら、各ノード（スイッチ）が持つ経路表を計算する。

経路選択機能は経路探索機能によって求めた多重経路に対して、効果的にデータを送信するように、適切な経路の選択を行う。具体的には、ネットワークの利用状況をリアルタイムに収集し、それに基づいて経路を選択することでデータ伝送の高速化を図る。さらに各ノードの位置情報を基にリスク値を算出し、その値を重み付けすることで、可用性の向上を実現する。

以下、各機能の詳細設計について説明する。

### 3.2 経路探索機能の設計

送信元から送信先の多重経路表を計算する経路探索機能のアルゴリズム（経路探索アルゴリズム）の設計を述べる。本アルゴリズムでは送信元から送信先までの間の不要な経路を取り除き、利用可能な全ての経路を抽出する。ここで「不要な経路」とは送信先がない袋小路や複数ノード間でのループを指す。この「袋小路」および「ループ」を検出し排除することで、送信元からのデータは送信先へ届くことが保障される。

経路探索アルゴリズムの設計にあたって以下のパラメータを定義する。

- $p$  : 親ノード（探索するノード）
- $A = a_1, a_2, \dots, a_m$  : 祖先ノード群（探索が終了したノードの集合）
- $C = c_1, c_2, \dots, c_n$  : 子ノード群（親ノードに隣接するノードの集合）
- $Path(x)$  : 標準経路（ノード  $x$  から送信先へ最短ホップ数で到達する次段ノード）

経路探索アルゴリズムの動作手順を以下に示す。送信元から送信先への通信セッションが開始されたとする。ここで、初期条件として送信元であるホストを親ノード  $p$  とする。 $p$  では隣接するスイッチの集合を子ノード群  $C$  として経路探索アルゴリズムを実行する。経路探索アルゴリズムでは、あるノードが子ノードを探索する際に、その子ノードの各要素が、自身の子ノードに対して再帰的に経路探索アルゴリズムを実行する。ここで、あるノードが経路探索アルゴリズムを実行するまでに経由した一連のノードを祖先ノード群  $A$  と呼ぶ。

経路探索アルゴリズムのフローチャートを図2に示す。まず、 $p$  が送信先か判定し、送信先である場合は  $p$  に完了を伝達して処理を終了する。 $p$  が送信先でない場合は  $A$  に  $p$  を追加し、 $C$  に対して以下の除外条件を適用する。

- $C$  から  $A$  の要素を除外
- $Path(c_x)$  が  $p$  である場合、 $c_x$  を  $C$  から除外

ここで、 $Path(x)$  は標準経路（ノード  $x$  から送信先へ最短ホップ数で到達する次段ノード）とする。そして、 $C$  に子ノードの要素が残っていない場合、 $p$  に完了を伝達して終了する。 $C$  に子ノードの要素が残っている場合、 $C$  の各要素に対して、経路アルゴリズムを再帰的に実行する。

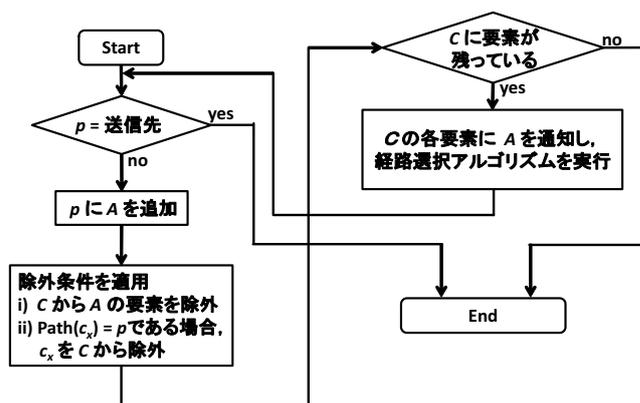


図2 経路探索アルゴリズムのフローチャート

これを  $p$  が送信先となるまで再帰的に繰り返す。最終的に送信元が経路探索アルゴリズムを完了した時点において、各ノードはある送信元から送信先への使用可能な次段経路を保持している。ここで作成した経路を保持するテーブルを多重経路表と呼ぶ。

### 3.3 経路選択機能の設計

経路選択機能のアルゴリズム（経路選択アルゴリズム）は多重経路表とネットワークの利用状況をもとに次段経路を選択し、データを転送する。その際、実際に送出したデータを用いて次段経路ごとの転送効率を計測し、それに基づいて各次段経路の流量を調整する。

具体的には、まず送信データサイズを送信元から送信先への利用可能な経路数に分割し、それぞれのデータを各経路を通じて転送する。データを転送する通信セッションが開始されると、各ノードの次段経路との間の利用可能帯域と次段経路への単位時間あたりのデータ送信量を計測する。この情報を基にして各ノード間のネットワーク利用率を算出する。これを送信元から送信先への利用可能な全ての経路についてネットワーク利用率を計算し、ネットワーク利用率が最も低い経路を優先的に利用するように経路表を更新する。

この計算を一定周期で行うことで、空いている経路を効果的に利用し、混雑している経路は徐々に利用しなくなるという適応的なルーティングが可能となる。経路選択機能によりネットワークトラフィックの負荷を分散することで、ネットワークの高速転送を実現する。

## 4. 実装

設計した各機能の実現には、ネットワークの利用状況の観測や柔軟なネットワーク構成の変更が可能な技術が必要である。そこで、本研究ではネットワーク基盤技術として、ソフトウェアによりネットワークをプログラミング可能なソフトウェア定義型ネットワーク（Software Defined Network: SDN）の実装形態の一つである OpenFlow を用

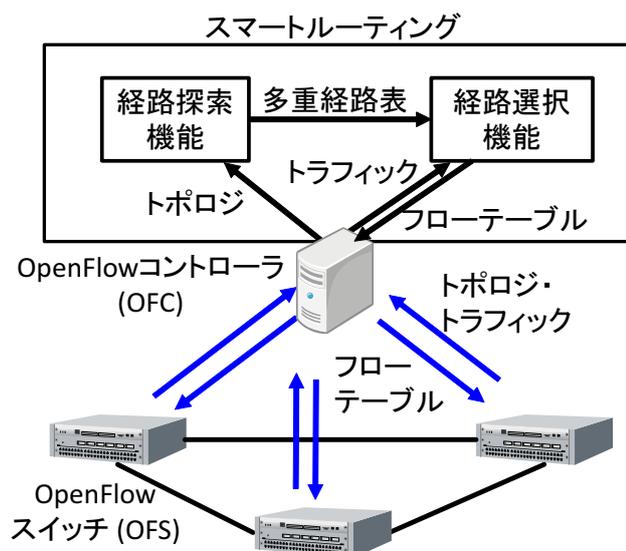


図3 OpenFlowを用いたスマートルーティングの実装

いて、本手法の設計・実装を行った[10]。その構成を図3に示す。

経路探索機能と経路選択機能はOpenFlowコントローラ(OFC)に組み込み、実装した。経路探索機能はOFCがOpenFlowスイッチ(OFS)からネットワークトポロジの情報を収集し、その情報に基づき送信元から送信先の経路を計算する。経路選択機能はOFCによりOFSから定期的に収集されたポート毎の帯域や各フローのトラフィック量、多重経路表をもとにして、各経路のネットワーク利用率を計算し、ネットワーク利用率の低い経路にフローを送信するようにフローテーブルを生成する。それをOFC経由で各OFSに設定することで、経路の切り替えを実現する。

この一連の処理を定期的に行うことで、トラフィック状況に応じて動的に経路を切り替え、データの高速転送化を実現すると共に、災害などでトポロジが変化した場合でも、その状態を認識し、経路表を再計算することで、データ転送を継続することが可能となる。

## 5. 実験と評価

### 5.1 実験環境

スマートルーティングの有効性を検証するためにシミュレーション実験を行った。本節ではシミュレーション実験環境について説明する。本実験では、仮想環境上にOFC、OFSのソフトウェア実装と、仮想ネットワーク環境を導入し、ネットワークエミュレーション環境を構築した。OFCにはOpenDaylight[11]、OFSにはOpen vSwitch[12]を用いた。また仮想ネットワーク環境としてMininet[13]を使用した。

図4に構築したネットワーク環境を示す。Mininetを用いてホストが8台(h1(10.0.0.1)~h8(10.0.0.8))、OFSが20台(s1~s20)から構成される仮想ネットワーク環境を構

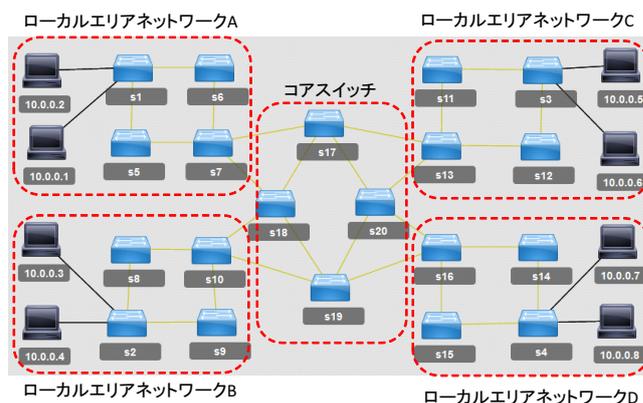


図4 実験で用いるネットワークトポロジ

築した。本研究で開発しているストレージシステムは地域ネットワークある程度、限られたネットワーク内での利用を想定していることから、ここではホスト2台とネットワークスイッチ4台からなるローカルエリアネットワークを4つ用意し、それらの間を4つのコアスイッチが接続している構成とした。また、スイッチ間のネットワーク帯域は100Mbps、スイッチ-ホスト間の帯域は1000Mbpsとした。図中には示されていないが、各OFSは1台のOFCに接続されており、各OFSで観測したネットワークの利用状況がOFCに送られ、それに応じて、OFCがデータ転送に利用する経路を決定し、OFSにその情報を送信する。

### 5.2 実験1: 平常時におけるデータの高速転送化の検証

#### 5.2.1 実験内容

まずは実験1として、平常時におけるデータの高速転送化の検証を行う。ここでは、ホストh1からホストh8へバックアップのためにデータを高速転送することを想定したシミュレーション実験を行った。具体的には、ホストh1からホストh8へ利用可能な多重経路を経路探索アルゴリズムにより計算する。そして、データのバックアップのためのファイル転送として、ホストh1からホストh8に1GbytesのデータをTCPで送信する。その10秒後にホストh2からホストh5に50MbpsのトラフィックをUDPで50秒間送信する。そして、実験開始から20秒後にホストh6からホストh3に50MbpsのトラフィックをUDPで50秒間送信し、実験開始から30秒後にホストh4からホストh7に5MbpsのトラフィックをUDPで50秒間送信する。各トラフィックの標準の経路は図5のように設定した。

この時、ホストh1からホストh8のスループットをリアルタイムで計測し、80Mbpsを下回った場合に、経路選択機能により経路の切替を実施する。そして、従来手法として、単一経路を固定した場合、多重経路を固定した場合と、提案手法として多重経路を動的に変更した場合のデータ転送時間とh8のスループットをそれぞれ比較する。

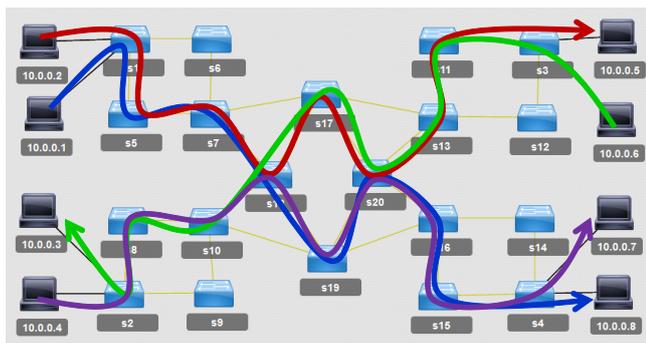


図 5 各トラフィックの標準の経路

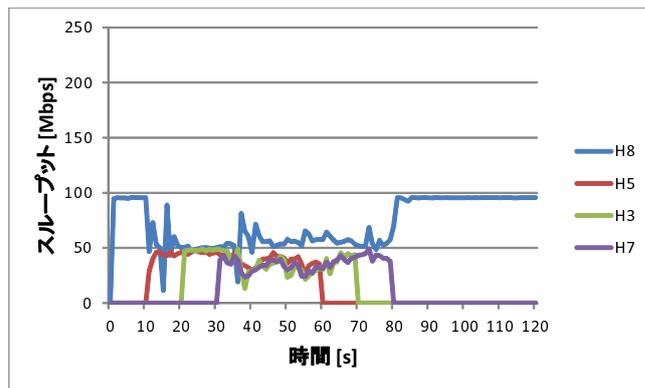


図 6 既存手法（単一経路固定）のスループット

表 1 実験結果：データ転送時間と h8 の平均スループット

	データ転送時間	h8 の平均スループット
既存手法		
単一経路固定	120.5 秒	71.3Mbps
多重経路固定	70.6 秒	121Mbps
提案手法		
経路多重化動的変更	63.9 秒	134Mbps

### 5.2.2 実験結果

図 6 に既存手法（単一経路固定）の各ホストのスループットのグラフを、図 7 に既存手法（多重経路固定）の各ホストのスループットのグラフを、図 8 に提案手法（多重経路動的変更）の各ホストのスループットのグラフをそれぞれ示す。また、表 1 に各手法の 1GByte のデータの転送時間と、ホスト h2 の平均スループットを示す。

図 6 のグラフから単一経路のみを固定して用いた場合は、他のトラフィックが発生した際に、それらが同一経路上に流れることで、h8 のスループットが低下したことが確認できる。多重経路を固定して利用した場合、図 7 のグラフから単一経路のみを利用する場合と比べて、スループットの向上が確認できたが、他のトラフィックが発生するにつれて、徐々に h8 のスループットが低下した。

多重経路を動的に変更した提案手法においては、図 8 のグラフから他の様々なトラフィックが発生した場合でも、他のネットワーク利用率が低い経路に切り替えることにより、スループットの低下が抑制された。以上の結果から提案手法により、経路を動的に切り替えることで、転送時間を最大約 50%削減することができた。

これは、データ転送のために利用中の経路に他のフローが流れると、OFS がそれを検知し、OFC によりデータ転送の経路が利用されていない他の経路に動的に切り替えたためである。それによりホスト h8 のスループットも上昇したことが確認できる。

以上より、スマートルーティングにより、経路の利用状況を観測して、それに応じて適切に経路を切り替えることで、平常時におけるデータの高速転送化を実現できた。

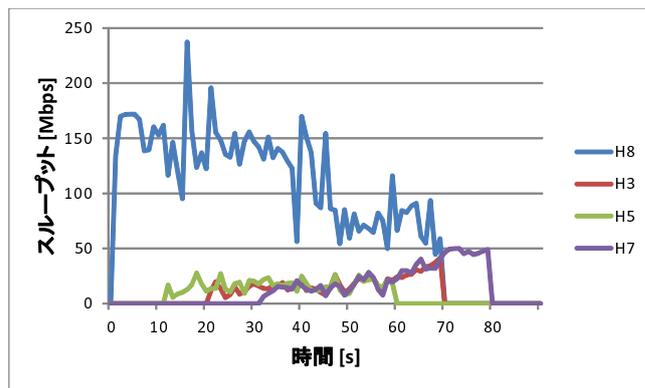


図 7 既存手法（多重経路固定）のスループット

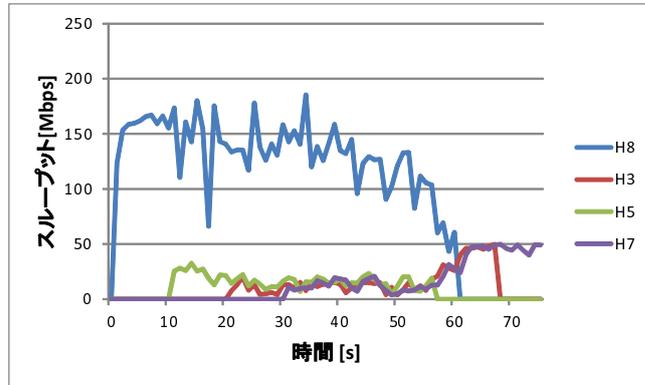


図 8 提案手法（多重経路動的変更）のスループット

### 5.3 実験 2：災害直後におけるデータ転送の検証

#### 5.3.1 実験内容

次に、災害時におけるデータ転送の検証を行う。ここでは、災害直後に、バックアップされていないデータをホスト h1 からホスト h8 へ転送することを想定したシミュレーション実験を行った。具体的には、ホスト h1 からホスト h8 へ利用可能な多重経路を経路探索アルゴリズムにより計算する。そして、データのバックアップのためのファイル転送として、ホスト h1 からホスト h8 に 1Gbytes のデータを TCP で送信する。今回は経路多重化は行わず、単一経路のみを利用してデータを送信する。その 10 秒後にホスト h2 からホスト h5 に 50Mbps のトラフィックを UDP

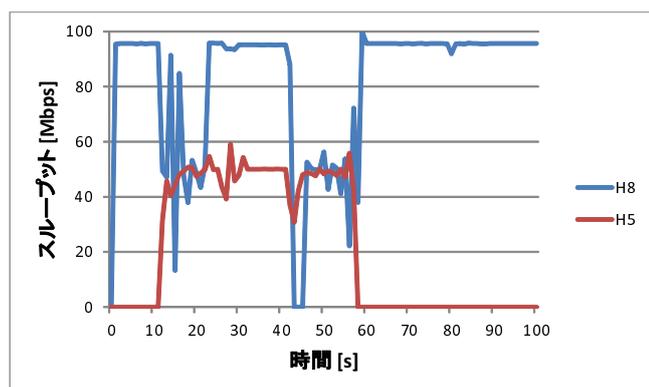


図9 実験結果2：各ホストのスループット

で50秒間送信する。この時、ホストh1からホストh8のスループットをリアルタイムで計測し、80Mbpsを下回った場合に、経路選択機能により経路の切替を実施する。そして、実験開始から45秒後に災害時を想定して、スイッチs1とスイッチs6の間のリンクを切断する。この時の各ホストのスループットを計測する。

### 5.3.2 実験結果

各ホストのスループットのグラフを図9に示す。実験開始10秒後の時点で、ホストh2からホストh5へのトラフィックが発生し、h8のスループットが低下した。しかしその後、経路選択機能により、経路を切り替えることで、h8のスループットが向上した。

また、実験開始45秒の時点で、スイッチs1とスイッチs6の間のリンクが切断されたことで、h8のトラフィックが停止した。しかし、提案手法により、リンクの切断を検出し、トポロジ情報をOFCが受け取り、経路表を素早く再計算することで、h8のトラフィックが再開し、データ転送を継続することができた。

以上より、スマートルーティングにより、災害時においても、リンクやノードの切断を認識し、経路を動的に切り替えることで、可能な限りの性能を確保することができた。

## 6. おわりに

本研究では、対災害性を強化したストレージシステムのために、ネットワークの利用状況に応じて動的に経路を選択するスマートルーティングを提案し、SDNを用いてその設計と実装を行った。さらに、様々な状況を想定したシミュレーション実験により、平常時にはデータの高速転送化を実現し、災害時においても可能な限りの転送性能を保持することを示した。

今後は、経路選択機能をより詳細化することで、さらなるデータの高速転送化を目指すと共に、実機のOFSを用いた実験により、その実用性を検証する。さらに、高可用性を目指し、各経路やスイッチについて損壊のリスクをモデル化し、それを考慮した経路選択アルゴリズムの拡張を検討する[14]。

## 参考文献

- [1] 田中博：災害時と震災後の医療IT体制 そのグランドデザイン、情報管理, Vol.5, No.12, pp.825-835 (2012).
- [2] 高機能高可用性情報ストレージ基盤技術の開発, (入手先 (<http://www.it-storage.riec.tohoku.ac.jp/>)) (2015.07.03).
- [3] Matsumoto, S., Nakamura, T., and Muraoka, H.: Risk-based Method for Data Redundancy Determination to Improve Replica Capacity Efficiency, *Proc. of the 3rd Asian Conference on Information Systems (ACIS2014)*, pp.529-536 (2014).
- [4] Nakamura, T., Matsumoto, S., and Muraoka, H.: Discreet Method to Match Safe Site-Pairs in Short Computation Time for Risk-Aware Data Replication, *IEICE Transactions on Information & Systems, Vol.E98-D, No.8*, pp.529-536 (2014).
- [5] 宗形聡, 宋チュウ, 手塚大, 村岡裕明: 薬歴データへのアクセスを想定した大規模災害時の高可用ストレージ基盤の耐災害性評価, 第77回情報処理学会全国大会予稿集, Vol.4, pp.451-451 (2015).
- [6] Ferraz, L., Mattos, D., and Duarte, O.: A Two-phase Multipathing Scheme based on Genetic Algorithm for Data Center Networking, *Proc. of the IEEE Global Communications Conference (GLOBECOM2014)*, pp.2270,2275 (2014).
- [7] Katrinis, K., Guohui, W., and Schares, L.: SDN control for hybrid OCS/electrical datacenter networks: An enabler or just a convenience?, *IEEE Photonics Society Summer Topical Meeting Series*, pp.242-243 (2013).
- [8] Weiyang, M., Jun, H., Karbassian, M., Wissinger, J., and Peyghambarian, N.: Situation-Aware Multipath Routing and Wavelength Reassignment in a Unified Packet-Circuit OpenFlow Network, *Proc. of the Optical Fiber Communication Conference and Exposition (OFC/NFOEC2013)*, pp.1-3 (2013).
- [9] Nagata, A., Tsukiji, Y., and Tsuru, M.: Delivering a File by Multipath-Multicast on OpenFlow Networks, *Proc. of the 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS2013)*, pp.835-840 (2013).
- [10] Izumi, S., Edo, A., Abe, T., and Suganuma, T.: Design and Implementation of SDN-based Smart Routing for Disaster-resistant File Transfer, *Proc. of the 3rd Asian Conference on Information Systems (ACIS2014)*, pp.560-565, 2014.
- [11] OpenDaylight - A Linux Foundation Collaborative Project (available from (<http://www.opendaylight.org/>)) (2015.07.03).
- [12] Open vSwitch, (available from (<http://openvswitch.org/>)) (2015.07.03).
- [13] Mininet : An Instant Virtual Network on your Laptop (or other PC), (available from (<http://mininet.org/>)) (2015.07.03).
- [14] 江戸麻人, 和泉諭, 阿部亨, 菅沼拓夫: 災害リスクを考慮したスマートルーティングの設計と実装, マルチメディア, 分散, 協調とモバイル (DICOMO2015) シンポジウム, pp.1520-1524 (2015).