

VXLAN 通信経路の高速切り替えを実現する
マルチキャストツリー集中管理方式西 佑介[†] 坂田 匡通[†] 木下 順史[†][†] 株式会社日立製作所横浜研究所 〒244-0817 神奈川県横浜市戸塚区吉田町 292 番地E-mail: [†] {yusuke.nishi.ef, masayuki.sakata.nm, junji.kinoshita.he}@hitachi.com

あらまし 大規模ネットワーク仮想化技術 VXLAN (Virtual eXtensible Local Area Network) では、ブロードキャストドメイン分割に IP マルチキャスト通信を利用するため、数千を超えるマルチキャストツリーを構築・維持する可能性がある。従来技術である IGMP (Internet Group Management Protocol) では、マルチキャストツリーの数に比例してネットワーク装置の処理負荷やマルチキャストツリー構築時間が増加する。このため、VXLAN 環境では経路障害時のマルチキャストツリー再構築処理に数十秒から数分かかり、その間、通信遅延やパケットロスが発生する。こうした問題に対処するため、本稿では、管理装置がネットワーク構成やサーバ構成を管理し、管理情報を元にあらかじめマルチパス上にマルチキャストツリーを構築するマルチキャストツリー集中管理方式を提案する。提案方式では、経路障害時のマルチキャストツリー再構築が不要となるため、ネットワーク装置に負荷をかけることなくマルチパスの経路を切り替えるだけで IP マルチキャスト通信を再開することが可能となる。実機による評価を行い、本方式はマルチキャストツリーの数に依存せず数百 ms で障害復旧することを確認した。

キーワード マルチキャスト, マルチキャストツリー, マルチパス, VXLAN, 仮想ネットワーク

Centralized Multicast Tree Control for rapid VXLAN path failover

Yusuke NISHI[†] Masayuki SAKATA[†] and Junji KINOSHITA[†][†] Hitachi, Ltd., 292 Yoshida-cho, totsuka-ku, Yokohama-shi, Kanagawa, 244-0817 JapanE-mail: [†] {yusuke.nishi.ef, masayuki.sakata.nm, junji.kinoshita.he}@hitachi.com

Abstract VXLAN (Virtual eXtensible Local Area Network), a large network virtualization technique, utilizes Multicast communication to divide broadcast domain. Therefore thousands of Multicast trees could be created and maintained in data center using VXLAN. IGMP (Internet Group Management Protocol), which has been widely used, increases the processing load and the time for switches to create the multicast trees as the number of tenant increases. In VXLAN environment, it may cause a few minutes of delay in communications or the communication failure of multicast communication when a multicast tree is failed. In this report, we propose a method for configuring a multicast tree on several physical networks (multipath) in advance by a management system which knows the physical network architecture and the information of Virtual Machines. When the physical network failure occurs, the proposal method can continue multicast communication without relearning a multicast tree because there are some secondary multicast routes. We evaluate the merits of the proposal method using experiments. Our prototype recovers from physical network failure in hundreds of milliseconds regardless of the number of multicast trees.

Keyword Multicast, Multicast tree, Multipath, VXLAN, Virtualized network

1. はじめに

近年、迅速なサービスデリバリーや TCO (Total Cost of Ownership) 削減を目的としてクラウドコンピューティングの利用が加速している[1]。クラウド関連市場は年率約 30%増で成長しており、今後も基幹業務など様々なシステムがクラウドへ移行する等、さらにクラウドの利用が進むと予想される。その一方で、クラウドサービスの競争激化によりサービス価格が急速に下落している。このためクラウド事業者にとって、DC (Data

Center) の TCO 削減による収益拡大が課題となっており、同一インフラに、より多くのテナント (顧客システム) を収容する高集約マルチテナント環境の実現や運用コスト削減が必須となっている。

マルチテナント環境では、例えばデータパケットの不正傍受防止や共用ネットワークの効率的な帯域利用を目的として、テナント毎に論理的なネットワークを構築し互いの通信を分離する必要がある。現在のクラウド型 DC では、VLAN (Virtual Local Area Network)

を用いてテナント毎にネットワークを分割しテナントの独立性を確保している。しかし、VLANは識別子数の上限が4094であり、高集約なマルチテナント環境で想定される数千～数万のテナントを収容することは困難である。また、VLANはテナントの追加・削除や仮想サーバの移動の際、ネットワーク装置に設定変更が必要のため運用コストが大きい。

こうしたVLANの問題に対し、現在様々なベンダや団体により大規模ネットワーク論理分割方式が提案されている[2]-[5]。その一つとしてオーバーレイ方式のネットワーク仮想化技術であるVXLAN (Virtual eXtensible Local Area Network) が注目されている[6]。VXLANは、例えば物理サーバ内に配備された仮想スイッチにVTEP (VXLAN Tunnel End Point) と呼ばれる仮想ネットワーク終端部を設け、VTEPでMAC (Media Access Control) フレームをUDP (User Datagram Protocol) /IP (Internet Protocol) パケットにカプセル化することでVTEP間をトンネリングする。カプセル化の際、VNI (VXLAN Network Identifier) と呼ばれる24bitの識別子を設けることで、約1677万の論理的なネットワークを構築することが可能となっている。

ここで、VXLANにおけるVTEP間通信では、単一ホスト宛のフレームはIPユニキャスト通信に、複数ホスト宛のフレーム (マルチキャストフレーム、ブロードキャストフレーム) または宛先VTEPが不明 (未学習) なフレームはIPマルチキャスト通信にカプセル化し転送する。IPマルチキャスト通信では、あらかじめテナントにマルチキャストアドレスを割り当て、マルチキャストアドレス毎にマルチキャストツリーを構築することで、テナントの論理的なネットワークを実現している。

従来、マルチキャストツリーの構築にはIGMP (Internet Group Management Protocol) が広く用いられてきた。IGMPでは、マルチキャストツリーを構築・維持するため、スイッチが定期的に制御メッセージをやり取りする。VXLAN環境でIGMPを用いると、テナントに対応するマルチキャストアドレス分の制御メッセージを送信する必要があり、マルチテナント環境では大量の制御メッセージがやり取りされる。ここで、経路障害が発生すると、早期にマルチキャストツリーを再構築し通信を再開するために、短時間で大量の制御メッセージを処理する必要がある。制御メッセージをソフトウェアで処理する場合、大量の制御メッセージを処理するとスイッチの負荷が高まり安定した通信が困難となる。また、スイッチの処理負荷を一定に保つと処理できる制御メッセージ数が制限されるため、障害復旧に時間がかかり、IPマルチキャスト通信が利用できない状況が発生する。

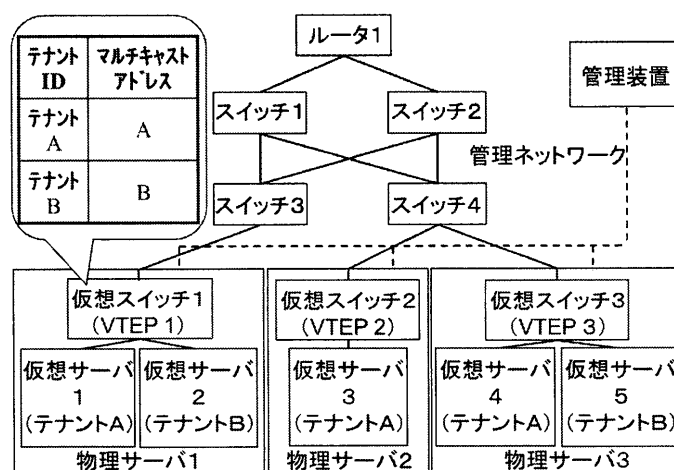


図1 DCネットワークの構成

そこで本研究では、集中管理装置があらかじめマルチキャストアドレス毎に、複数の物理経路 (マルチパス) 上に一つのマルチキャストツリーを構築する方式を提案する。本方式では、経路に障害が起きた場合、マルチキャストツリーの再構築が不要であり、通信に用いる経路を代替経路に切り替えるだけでIPマルチキャスト通信を即時再開することが可能となる。本稿では、実測により経路障害時のスイッチの負荷と障害復旧時間を評価し、提案方式がテナント数に依存せず低負荷で高速な障害復旧が可能であることを示す。

2. 前提となるDCネットワークシステム

本章では、前提となるDCネットワークの構成と、VXLANにおけるVTEP間のIPマルチキャスト通信について説明する。

2.1. DCネットワーク構成

図1に前提となるDC構成を示す。物理サーバには仮想サーバが配備されており、仮想サーバからの通信は仮想スイッチを経由して宛先に送信される。仮想スイッチには、VXLANにおけるVTEPが実装され、MACフレームのカプセル化やデカプセル化を行う。管理装置は管理ネットワークを通じてすべての物理サーバと接続しており、仮想サーバの追加・削除・移動を指示する。仮想サーバはそれぞれテナントに属し、同一テナントにのみ通信を行う。物理サーバ間は、STP (Spanning Tree Protocol) によるシングルパス構成に代わって主流となりつつある、複数の物理経路を同時に通信へ利用するマルチパス構成で接続する。図1では、例えばVTEP1とVTEP2が通信を行う場合、スイッチ3ースイッチ1ースイッチ4のパスと、スイッチ3ースイッチ2ースイッチ4のパスが存在しマルチパスとなる。スイッチ3やスイッチ4では、例えばレイヤ2ネットワークではイーサネットファブリック、レ

イヤ3 ネットワークでは ECMP (Equal Cost Multi Path) といった技術を用いて通信に利用するパスを選択する。

2.2. VXLAN におけるマルチキャスト通信

VTEP は、あらかじめテナントの ID とテナントの ID に対するマルチキャストアドレスをリストとして管理する。VTEP が仮想サーバから複数ホスト宛のフレームまたは宛先アドレスが不明なフレームを受信した場合、リストを参照して、仮想サーバのテナント ID からマルチキャストアドレスを特定する。その後、特定したマルチキャストアドレスを宛先アドレスとして受信フレームをマルチキャストパケットにカプセル化し、スイッチへ転送する。スイッチ間では、あらかじめ学習されたマルチキャストツリーに沿ってマルチキャストパケットが転送され、宛先 VTEP に送信される。マルチキャストパケットを受信した VTEP は、パケットをデカプセル化し、デカプセル化したフレームの宛先アドレスに従って、配下の仮想サーバへフレームを送信する。

以上に示すように、VXLAN では、テナント毎に割り当てられたマルチキャストアドレスを宛先とするマルチキャストパケットが VTEP 間で転送される。このため VTEP 間を接続するスイッチは、あらかじめこれらのマルチキャストアドレスのマルチキャストツリーを学習しておく必要がある。

3. 従来方式：IGMP と IGMP Snooping

一般に、レイヤ2 ネットワークにおけるマルチキャストツリーの学習には、IGMP と IGMP Snooping が用いられる。本章では、VXLAN 環境において IGMP と IGMP Snooping を用いた場合のマルチキャストツリー構築手順と課題を説明する。

3.1. マルチキャストツリーの構築

VXLAN 環境において、IGMP はルータと VTEP 間でやり取りされる。ルータは IGMP Query を用いて各 VTEP に受信を希望するマルチキャストアドレスを問い合わせ、IGMP Query を受信した VTEP は IGMP Join Report を返答し、配下にある仮想サーバのテナント ID に対応するマルチキャストアドレスをルータに通知する。

このとき各スイッチは、IGMP Snooping 機能により、IGMP Query と IGMP Join Report をスヌーピング（盗み見）し、転送すべきマルチキャストアドレスとその転送ポートを学習する。この結果、VTEP とルータを接続する物理経路において、マルチキャストアドレス毎にマルチキャストツリーが構築される。

3.2. 障害復旧時間

冗長化されていないマルチキャストツリーに障害が起きた場合は、レイヤ2 ネットワークの物理経路を

正常系に切り替えた上で、IGMP Query と IGMP Join Report をやり取りし、新たにマルチキャストツリーを再構築する必要がある。ここで、通常 IGMP Query は、ルータから 125 秒（デフォルト値）間隔で定期的を送信される。また IGMP Join Report は、すべての VTEP が同時に返信すると帯域を圧迫するため、IGMP Query を受信した後、10 秒（デフォルト値）以内のランダム時間を待ってルータへ送信される。従って、IGMP による障害復旧時間は下記となる。

障害復旧時間 =

レイヤ2 ネットワーク 切り替え時間 +

IGMP_Query送信待ち時間 +

IGMP_Join_Report送信待ち時間

3.3. 課題

VXLAN は約 1677 万のネットワーク識別子を持つため、最大で同等数のテナントを収容することが可能である。このとき、VXLAN では、テナントに割り当てたマルチキャストアドレス毎にマルチキャストツリーを構築するため、IGMP を用いるとルータと VTEP の間で大量の IGMP Join Report が送信される。ここで、経路に障害が起きた場合は、通信を再開するために、多数のマルチキャストツリーを早期に再構築する必要がある。このとき、IGMP Query の送信間隔や IGMP Join Report の送信待ち時間を短くして障害復旧時間を一定時間以内に保つと、テナント数の増加に伴いスイッチの単位時間当たりの IGMP Join Report の処理負荷が高まって安定した通信が困難となる。これに対し、IGMP Query の送信間隔や IGMP Join Report の送信待ち時間を調整してスイッチの処理負荷を一定以下に保つと、スイッチが単位時間当たりに処理する IGMP Join Report 数が制限され、障害復旧に時間がかかる。

以上に示す通り、VXLAN 環境において多数のテナントが存在する場合、IGMP では安定した通信と高速な障害復旧を同時に実現することが困難である。VXLAN 環境では、テナント数に依存せず安定した通信が可能なマルチキャストツリー構築方式が必要となる。

4. 提案方式

本章では、前章で示した課題を解決するため、管理装置を用いてマルチパス上に一つのマルチキャストツリーを構築するマルチキャストツリー集中管理方式を説明する。

従来方式では、障害発生後にマルチキャストツリーの再構築を行うため、安定した通信と高速な障害復旧の実現が困難となっていた。そこで、あらかじめ冗長なマルチキャストツリーを構築し、経路障害時に冗長

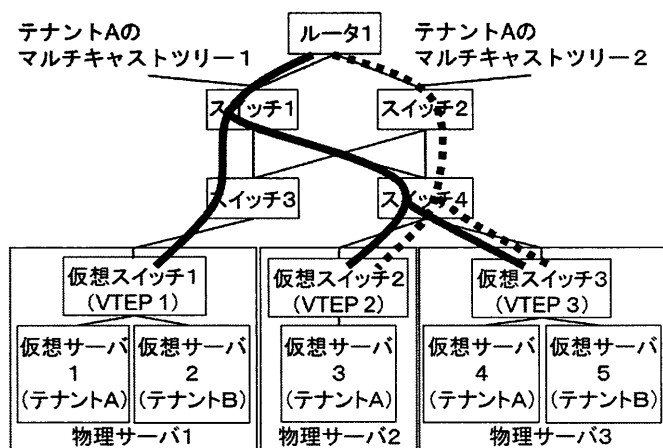


図2 複数マルチキャストツリーによる冗長化

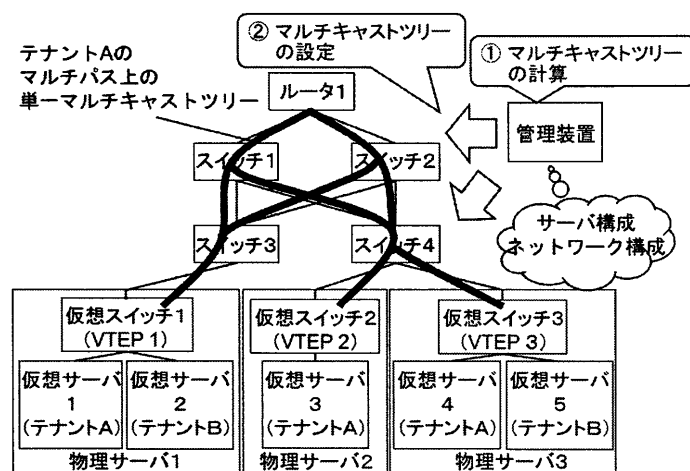


図3 単一マルチキャストツリーによる冗長化

系に切り替えることでスイッチに負荷をかけることなく高速に障害復旧を行う方式を検討する。

冗長なマルチキャストツリーの構築には、図2に示すように、マルチキャストアドレス毎に複数のマルチキャストツリーを構築する方法が考えられる。しかし、この方法ではスイッチのマルチキャストツリー学習量が増えるため、スイッチのリソースを消費し収容可能なテナント数を制限する可能性がある。さらに、IGMPのような自律分散型のマルチキャストツリー構築方式では、ネットワークの全体構成を把握していないスイッチが、すべてのマルチキャストアドレスに対し冗長なマルチキャストツリーを構築することは負荷が高い。

そこで、本稿では図3に示すように、管理装置を用いてマルチパス上に、マルチキャストアドレス毎に一つのマルチキャストツリーを構築する方式を提案する。障害発生時は、スイッチの転送機能によりマルチパスの経路を切り替えるだけでマルチキャストツリーが切り替わり、マルチキャストツリーを再構築することなく障害から復旧する。

本方式では、管理装置は、スイッチの接続構成やサ

ーバ情報（サーバのID、所属テナント、各テナントのマルチキャストアドレス、接続するVTEP）を管理する。管理装置がマルチキャストツリーを構築する際は、まず管理する情報を用いて、マルチパス上にマルチキャストツリーを構築するようスイッチと転送ポートを特定する。その後、管理装置は特定した転送ポートをスイッチや仮想スイッチに通知することで、ネットワーク全体にマルチキャストツリーを構築する。

ここで、管理装置がスイッチや仮想スイッチにマルチキャストツリーを通知する方法として、次の二つの方法を検討した。一つは、管理装置が転送すべきマルチキャストアドレスと転送ポートを通知し、スイッチや仮想スイッチが通知に従ってマルチキャストパケットの転送テーブルを書き換える方法である。この方法は、例えばOpenFlowといった、転送ルールをソフトウェアで管理するシステムと親和性が高い。二つ目は、管理装置が、特定したマルチキャストツリーを元に、IGMP Join Reportの転送ポートをスイッチや仮想スイッチに設定し、その後、VTEPにIGMP Join Reportを送信するよう指示を与える方法である。指示を受けたVTEPがIGMP Join Reportを送信し、各スイッチがIGMP Snoopingによりマルチキャストパケットの転送経路を学習する。この方法は、従来技術であるIGMPを用いるため、転送ルールをソフトウェアで管理するシステムに対応していない場合でも用いることが可能である。どちらの方法においてもマルチパス上にマルチキャストツリーを構築することが可能であり、運用システムに応じて方法を選択する。

5. 実験による評価

提案方式に基づきマルチパス上のスイッチ（実機）にマルチキャストツリーを設定し、通信経路に障害を発生させ、IPマルチキャスト通信再開に伴うスイッチの負荷と障害復旧時間を評価した。

5.1. 評価環境

図4に評価構成を示す。スイッチを4台、物理サーバを2台とし、物理サーバ間は「スイッチ1－スイッチ2－スイッチ4」のパスと、「スイッチ1－スイッチ3－スイッチ4」のパスでマルチパス構成とした。

実験では、スイッチに複数ポートからIGMP Join Reportを送信できるよう機能を付加した。さらに、付加機能の利用を集中管理によって指示することを模擬するために、事前にCLI（Command Line Interface）によってスイッチにIGMP Join Reportの転送ポートを複数設定した。その後、物理サーバにIGMP Join Reportを送信させることでマルチパス上にマルチキャストツリーを構築した。

評価では、常に物理サーバ2から物理サーバ1へ

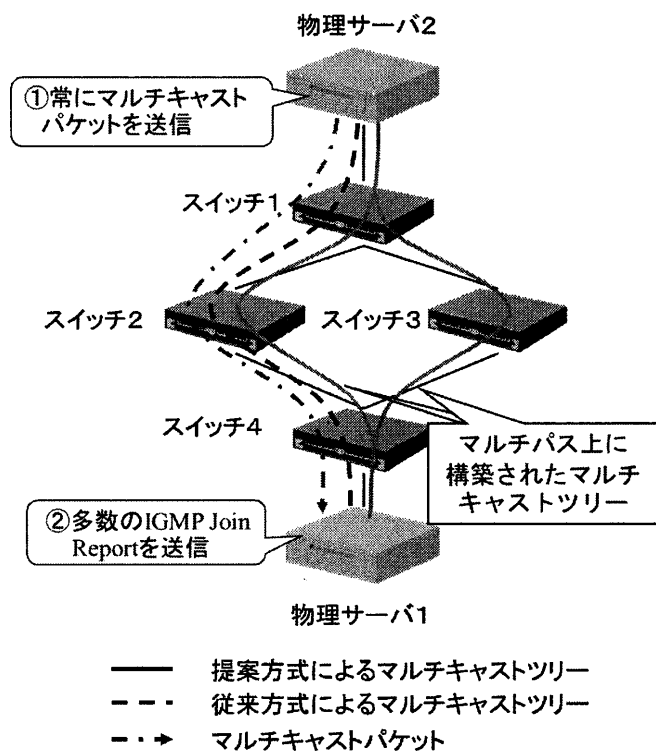


図4 評価環境

「スイッチ1－スイッチ2－スイッチ4」のパスを使ってマルチキャストパケットを送信した。この状態からスイッチ2に障害を発生させ、「スイッチ1－スイッチ3－スイッチ4」のパスによってマルチキャストパケットの転送が再開するまでのスイッチの負荷（CPU使用率）と、時間（障害復旧時間）を測定した。

なお本評価では、多数のテナントを収容しているVXLAN環境を想定し、スイッチに多数のIGMP Join Reportが送信される状態を、物理サーバ1が多数のIGMP Join Reportを送信することで代替した。VXLAN環境においてテナント毎に異なるマルチキャストアドレスを割り当てるとすると、マルチキャストアドレス毎にIGMP Join Reportが送信されるため、本評価構成では、物理サーバ1が送信するIGMP Join Reportの数はスイッチ4の配下に収容しているテナント数を表す。

5.2. 評価結果

5.2.1. 障害復旧時のパケット処理負荷

本項では、従来方式（IGMP）と提案方式について、障害復旧時のスイッチのCPU使用率を比較する。

図5に、障害発生後に物理サーバ1が送信するIGMP Join Report数すなわちスイッチ4が収容するテナント数を変化させた場合の各スイッチのCPU使用率を示す。従来方式におけるIGMP Join Report送信待ち時間はデフォルト値の10秒とし、この間のCPU使用率を測定した。提案方式は、障害発生時から通信再開までのスイッチのCPU使用率を測定した。

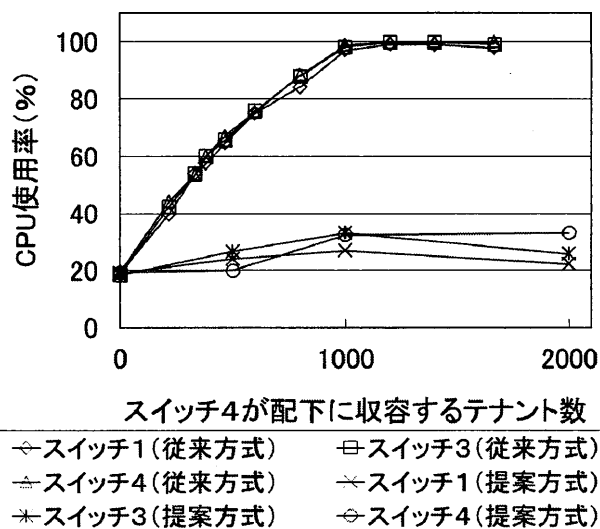


図5 スイッチの処理負荷

従来方式について、図5より、テナント数の増加に伴いCPU使用率が増加することが分かる。これは、テナント数が増加すると、マルチキャストツリー再構築のためにスイッチが処理する単位時間当たりのIGMP Join Report数が増えるためと考えられる。スイッチ1、スイッチ3、スイッチ4のCPU負荷が同等であるが、三者とも同数のIGMP Join ReportについてIGMP Snoopingを行ったためと考えられる。また、テナント数が1000以上の場合はCPU使用率が100%となり、IGMP Snoopingが正常に機能せず、IGMP Join Reportを送信してもマルチキャストツリーが構築されないマルチキャストアドレスが存在することを確認した。この結果から、従来方式ではスイッチのCPU使用率を100%以下に抑える必要があり、収容するテナント数を制限する必要があることが分かる。

提案方式について、図5より、テナント数に関わらずスイッチのCPU使用率はほぼ一定であることが分かる。本方式ではあらかじめマルチパス上にマルチキャストツリーを構築しているためマルチキャストツリーを再構築する処理が不要であり、テナント数が増加してもCPU使用率が変化しないと考えられる。スイッチ1、スイッチ3、スイッチ4のCPU使用率が常に30%程度で同等であるが、これは定常状態におけるCPU使用率と考えられる。以上の結果から、本方式は障害復旧時にスイッチの負荷を高めることがなく、収容するテナント数に制限を受けないことが分かる。

5.2.2. 障害復旧時間

本項では、従来方式（IGMP）と提案方式について、障害が発生してから再び物理サーバ1がマルチキャストパケットを受信するまでの時間、すなわち障害復旧

表 1 障害復旧時間 (秒)

方式 テナント数	従来方式	提案方式
1	67.92	0.15
500	67.19	0.11
1000	67.83	0.13
2000	— (測定不可)	0.21
5000		0.11
10000		0.10
20000		0.22

時間を比較する。

本評価では、物理サーバ2からマルチキャストパケットを2ms間隔で物理サーバ1へ送信し続け、障害が発生してから再びマルチキャストパケットを受信するまでの時間を、障害発生後に物理サーバ1が送信するIGMP Join Report数すなわちスイッチ4が配下に収容するテナント数を変化させて測定した。IGMP Queryの送信間隔およびIGMP Join Reportの送信待ち時間はそれぞれデフォルト値の125秒、10秒とした。なお、従来方式の障害復旧時間には、3.2で述べた通り、「障害が発生してからIGMP Queryが送信されるまでの時間」と「VTEPがIGMP Queryを受信してからIGMP Join Reportを返信するまでの時間」が含まれる。前者は障害発生時のタイミングに依存するため、机上計算によりIGMP Queryの送信間隔125秒の1/2である62.5秒かかるとし、後者を実測して結果をまとめた。

表1に、各方式の障害復旧時間を示す。従来方式について、表1より、テナント数が1～1000のとき、67.5秒程度で障害復旧することが分かる。これは、VTEPがIGMP Queryを受信してからIGMP Join Reportを返信するまでに、平均して、IGMP Join Reportの送信待ち時間10秒の1/2である5秒かかったためと考えられる。テナント数が2000以上の場合は、5.2.1で述べた通りCPU負荷が高まってIGMP Snoopingが正常に機能しないため、IGMP Join Reportを応答してもIPマルチキャスト通信が再開されないマルチキャストアドレスが発生する。このため、障害復旧時間の測定を行っていない。

提案方式においては、表1より、テナント数に依らず0.15秒程度で障害復旧することが分かる。この0.15秒はスイッチ1におけるマルチパスの切り替え時間、すなわちスイッチ1が通信に利用するパスをスイッチ2へのパスからスイッチ3へのパスに切り替える時間である。これにより、提案方式ではレイヤ2ネットワークが切り替わると同時にIPマルチキャスト通信が再開していることが分かる。この結果から、提案方式

では、障害復旧時間はスイッチにおけるレイヤ2ネットワークの切替時間に依存し、一般的なスイッチでは、レイヤ2ネットワークが切り替わる1秒未満でIPマルチキャスト通信を再開できると考えられる。

6. まとめ

本研究では、大規模ネットワーク論理分割方式であるVXLANの実現に向け、IPマルチキャスト通信の低負荷・高速障害復旧方式を検討・評価した。具体的には、従来のIGMPを用いたマルチキャストツリー構築方式の代替として、あらかじめ複数の物理経路上に一つのマルチキャストツリーを構築しておくことで経路障害時にマルチキャストツリー再構築処理が不要な方式を提案した。提案方式を実機により評価し、収容するテナント数によらず、経路障害時のスイッチの負荷が定常状態と同等であること、および障害復旧時間がレイヤ2ネットワークの切替時間（スイッチの性能に依存）に等しいことを確認した。

以上の結果から、提案方式はテナント数によらず経路障害時にスイッチの負荷を一定に保ち、かつ高速に障害から復旧できるため、DCにおける高集約なマルチテナント環境の実現に寄与できると考える。

今後の課題は、管理装置がスイッチにマルチキャストツリーを設定する処理を実装し、高集約なマルチテナント環境で本方式の有効性を確認することである。

文 献

- [1] Luiz Andre Barroso, et al., "The Datacenter as a Computer", MORGAN&CLAYPOOL PUBLISHERS, <http://www.morganclaypool.com/doi/pdf/10.2200/S00193ED1V01Y200905CAC006>, 参照 June 28, 2009.
- [2] R. N. Mysore, et al., "PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric", ACM SIGCOMM Computer communication review, Vol. 39, Issue 4, pp.39-50, Oct. 2009.
- [3] A. Greenberg, et al., "VL2: A Scalable and Flexible Data Center Network", ACM SIGCOMM Computer Communication Review, Vol. 39, Issue. 4, pp.51-62, Oct. 2009.
- [4] Amazon.com, Inc., "Using Security Groups", <http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/using-network-security.html>, 参照 June 28, 2009.
- [5] N. McKeown, et al., "OpenFlow: Enabling Innovation in Campus Networks", Open Networking Foundation, <http://www.openflow.org/documents/openflow-wp-latest.pdf>, March 14, 2008.
- [6] M.Mahalingam, et al., "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", IETF, <http://tools.ietf.org/id/draft-mahalingam-dutt-dcops-vxlan-01.txt>, 参照 June 28, 2009.