情報処理学会論文誌 Vol.62 No.1 321–331 (Jan. 2021)

ベアメタルサービスのためのVXLANオーバコミット方式

木下 順史[1,2,a)] 薦田 憲久[2] 藤原 融[2]

受付日 2020年4月6日, 採録日 2020年10月6日

概要：VXLAN（Virtual eXtensible Local Area Network）等のネットワーク仮想化を用いて利用者ネットワークを分離するベアメタルサービス基盤において，スイッチに設定するVXLANの数が設定上限を超過するという課題を解決するために，VXLANを使用状況に基づいて削除することで，見かけ上のVXLAN設定数を増加させるVXLANオーバコミットを提案する．VXLANのMACアドレス情報のみを用いて使用頻度を判定するモデルを複数提案し，シミュレーション実験により提案方式の効果や影響を評価することで，オーバコミットの実現性を確認した．

キーワード：ベアメタル，VXLAN，オーバコミット，データセンタ

Method of VXLAN Overcommitting for Bare Metal Service

Junji Kinoshita[1,2,a)] Norihisa Komoda[2] Toru Fujiwara[2]

Received: April 6, 2020, Accepted: October 6, 2020

Abstract: In Bare metal service infrastructure where user networks are isolated with network virtualization like VXLAN (Virtual eXtensible Local Area Network), the number of VXLAN can exceed the configuration limit of switches. To solve the problem, we propose VXLAN over-committing by deleting VXLAN based on actual usage and increasing the number of VXLAN provided to users. Five VXLAN replacement models only with MAC addresses usage are proposed. The evaluated effectiveness and impact of the proposed models by simulation shows feasibility of VXLAN over-committing.

Keywords: bare metal, VXLAN, overcommit, data center

1. はじめに

クラウドや機械学習，HPC（High Performance Computing）等の利用拡大にともない，物理サーバをオンデマンドで利用するベアメタルサービスが進展している．ベアメタルサービスを支えるデータセンタ（以下，DC：Data Center）においては，サービス利用者ごとに物理サーバのネットワークを複数ラックにまたがって分離する必要がある．分離手法として，たとえばVLAN（Virtual LAN）[1]等の様々な手法が使われているが，利用者がVLAN番号について独自の採番ルールをすでに有する場合には運用手順を変更しなければならない等，柔軟性が低下する等の問題があった．

既存の解決手法としてスイッチにおけるVXLAN（Virtual eXtensible Local Area Network）[2]等のネットワーク仮想化を用いることで，利用者ごとに独立したVLAN空間を提供することが考えられてきた．しかしハードウェアやOS（Operating System）の制約により，設定可能なVXLANの数に上限があるため，利用者が多数のVXLANを利用すると設定上限を超過してしまう．

上記課題を解決すべく，設定上限を超えてVXLANを用いるオーバコミットが考えられるが，具体的なオーバコミット方式やその効果，通信への影響が明らかではない．

そこで本論文では，スイッチで観測可能な情報（MACアドレス等）を用いたVXLANのオーバコミット方式を検討し，シミュレーション実験により効果や通信への影響

[1] 株式会社日立製作所研究開発グループ
Hitachi Ltd., Research & Development Group, Kokubunji, Tokyo 185–8601, Japan
[2] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technbology, Osaka University, Suita, Osaka 565–0871, Japan
[a)] junji.kinoshita.he@hitachi.com

を評価する．スイッチで観測可能な情報のみを用いることで，利用者の様々な通信に対応することができる．

本論文の以降の構成は次のとおりである．2 章で現状のベアメタルサービスの概要と課題を説明し，3 章で VXLAN のオーバコミット方式を説明する．4 章で実験と評価をふまえて効果や影響を確認する．最後に 5 章で本研究のまとめを述べる．

2. ベアメタルサービスの概要と課題

2.1 ベアメタルサービスの概要

ベアメタルサービスは**図 1** に示すとおり，利用者に物理サーバ，および，物理サーバ間を接続する利用者ネットワークとして VLAN を提供する．ある利用者の物理サーバ群は物理的には複数のラックに収容されている可能性があるため，複数ラック間にまたがって VLAN を分離する必要がある．

仮想マシン貸しを行う IaaS（Infrastructure as a Service）の分野においては，仮想マシンのネットワーク分離手法として様々な方法が研究されてきた．具体的には，物理サーバ上のソフトウェア仮想スイッチにおいて，VLAN，あるいは VXLAN や NVGRE（Network Virtualization using Generic Routing Encapsulation）[7]，STT（Stateless Transport Tunneling）[8] 等のネットワーク仮想化技術 [9] を用いてネットワークを分離する．しかし，物理サーバ貸しを行うベアメタルサービスにおいては，**図 2** に示すように，物理サーバが接続された ToR（Top of Rack）スイッチ等の物理スイッチにおいてネットワークを分離する必要がある．

物理スイッチにおける既存のネットワーク分離手法には，VLAN をラック間にまたがってそのまま用いる手法や，通信キャリアで用いられている手法，ネットワークファブリック，OpenFlow 等がある．VLAN をそのまま用いると，ラック間にまたがる共通の VLAN 空間から個々の利用者に VLAN を採番することとなり，利用者は VLAN 番号を選べない．利用者がそれぞれ異なる組織に所属している場合は利用者間での調整は困難であり，利用者が VLAN 番号について独自の採番ルールをすでに有する場合には，利用者側で手順変更が必要となりうるため柔軟性が低下する．さらに，ラック間ネットワークも L2（Layer 2）で構成する必要がある等の制限がある．PBB（Provider Backbone Bridge）[3] や Q-in-Q [4]，MPLS（Multiprotocol Label Switching）[5] 等の通信キャリアで用いられている手法は，ブロードキャストドメインを利用者タグで分離できない，キャリア向けスイッチでしかサポートされていないため DC 向けスイッチでは利用できない等の制限がある．ネットワークファブリック [6] や OpenFlow は，ラック間ネットワークまでリプレースが必要になりうるため，導入の敷居が高い．

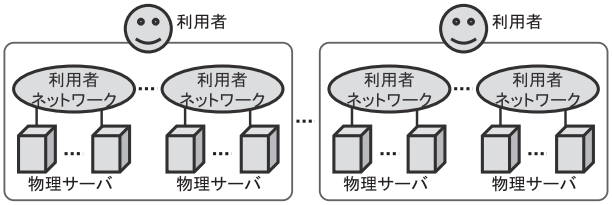

図 1 ベアメタルサービスの概要
Fig. 1 Overview of bare metal service.

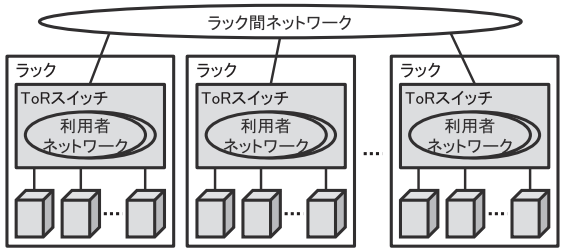

図 2 ベアメタルサービス基盤の物理構成
Fig. 2 Physical configuration of bare metal service.

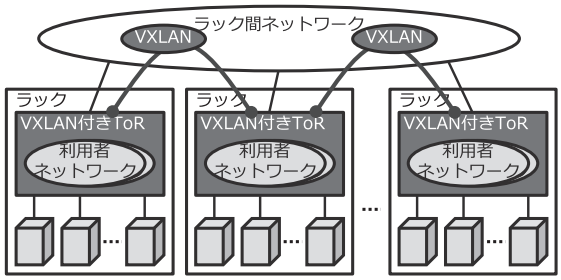

図 3 VXLAN による利用者ネットワークの分離
Fig. 3 User network isolation with VXLAN.

これら柔軟性の問題に対する解決手法として**図 3** に示すように，DC 向けのスイッチでも利用可能な VXLAN を用いて VLAN をカプセル化して分離することで，利用者ごとに VLAN 空間を分離することができる [10], [11], [12]．ネットワーク仮想化により，利用者は任意の VLAN 番号を用いることができ，ラック間のネットワークも通常の L3（Layer 3）でよく，個々の物理サーバから送受信されるタグなしあるいはタグつきの VLAN を ToR スイッチにおいてカプセル化し，複数ラック間にまたがって柔軟に搬送できる．

ただし，個々の物理サーバで送受信する VLAN を利用者ごとの VLAN 空間に対応付けるために，ポートと VLAN の組合せを VXLAN にマッピング可能な機能を有する ToR スイッチを用いる必要がある．本論文では VXLAN 機能を有する DC 向けスイッチの中でも，特に当該機能を有するスイッチを想定する．

2.2 VXLAN 利用時の課題

DC 向けスイッチにおいて VXLAN を利用するためには，**図 4** に示すように，VXLAN 識別情報（VNI：Virtual

Network Identifier），および，VXLAN と利用者 VLAN のマッピング情報をスイッチに設定する．DC 向けスイッチにおいては，これら識別情報やマッピング情報の設定数に上限がある．これは，ソフトウェアである仮想スイッチとは異なり，DC 向けスイッチで使われている ASIC（Application Specific Integrated Circuit）にはハードウェア制約が存在するためである．設定上限は製品に依存するが，たとえば VXLAN 機能を有する市販品の VNI 数の設定上限は，数百～6,000 程度である [13], [14]（ポートと VLAN の組合せで VXLAN にマッピングする機能を有しているとは限らない）．これは VXLAN の標準仕様上の上限値（4 k × 4 k = 約 1,600 万）をはるかに下回る．ベンダによってはこれらの上限値を公開していないが，ASIC で保持できる情報量には限りがあるため，設定上限を考慮する必要がある．

国内の一般的なラックは 42 U（ユニット）であるため，1 ラックあたりの物理サーバ数は 1 U サーバであれば 40 台，2 U4 ノード等の高密度サーバであれば 80 台が限界である．一方，一般的なベアメタルサービスは利用者ごとに VLAN を 1～64 程度割当て可能としている．仮に，あるラック内の個々の物理サーバを異なる利用者に割り当て，個々の利用者が VLAN を一般的なベアメタルサービス仕様上の上限値の 30 個まで利用しようとすると，40 × 64 = 2,560 個あるいは 80 × 64 = 5,120 個の VXLAN が ToR スイッチごとに必要となる．サービス仕様を超える VLAN 数を個別に要求されると，VXLAN の標準仕様上の上限値までは不要だが，スイッチの設定上限である数百～6,000 を超過しうる．

設定上限を超過しないためには，利用者の VLAN 数の制限や，ラックに収容する利用者数や物理サーバ数の制限，あるいはラック内のスイッチ増設によるスイッチあたりの物理サーバ数の削減等を行う必要があり，サービスレベルの低下やコスト増を招き，サービスの競合力が低下する．

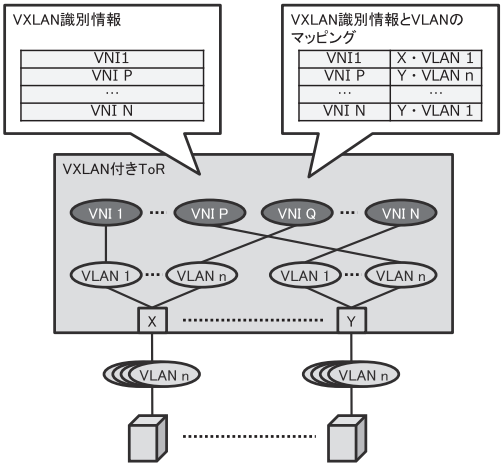

図 4 VXLAN 設定
Fig. 4 VXLAN configuration.

そのためスイッチにおいてより多くの VXLAN を設定できることが望ましく，本論文では ToR スイッチにおける VXLAN 設定数の増加を対象とする．

3. VXLAN オーバコミット

3.1 オーバコミットの概要

利用者の VLAN はつねに使われているとは限らず，実際には使用されていない，あるいは使用頻度が低いものがありうる．たとえば，筆者らが運用している研究開発用のベアメタルサービスにおいて，平日のある 1 日に 10 秒ごとに通信量を測定した結果を**表 1** 示す．通信が発生していない，あるいは極端に通信が少ない利用者ネットワークが全利用者ネットワークのそれぞれ約 3 割あるいは 4 割を占めている．

そこで，利用者 VLAN に対応する VXLAN の設定（識別情報およびマッピング情報）を，利用者向けのサービスポータル等においては設定済みと表示しつつ，その裏側では設定を削除し，利用者が実際に VLAN 内で通信を行う際に VXLAN を動的に再設定することで，設定上限を超える VXLAN を利用者に提供する（オーバコミット）．VXLAN の削除や再設定は，当該 VXLAN における通信断や通信開始遅延等を引き起こし，利用者の通信に影響を与えうる．そのため，削除や再設定による通信への影響を最小限に抑える必要がある．また，削除や再設定により，通信できる利用者と通信できない利用者が偏る可能性がある．そのため，削除や再設定による影響が VXLAN 間で差がない（公平性がある）ことが望ましい．

VXLAN の削除においては，ランダムに削除する方式と，使用予測に基づいて削除する方式が考えられる．また，削除後の再設定のために，利用者の VLAN 内での通信開始を検出する必要がある．

3.2 ランダム削除方式

本方式においては，スイッチにおける VXLAN の設定数を定期的に測定し，VXLAN が設定上限を超過する場合には，超過する数だけの VXLAN をランダムに選んで削除することで，VXLAN 数を設定上限範囲内に抑える．VXLAN の実際の使用状況を考慮しないため実装が簡易である．

ランダムに削除することにより，削除の確率の点では公平性がある一方，実際に使用している VXLAN も使用していない VXLAN も等しく扱うため，実際に使用している VXLAN にとっては公平性を欠く．

表 1 利用者ネットワークの使用状況事例
Table 1 Example of user network usage.

通信量が常時 0 Byte	通信量が常時 1K Byte 未満
全体の 32%	全体の 40%

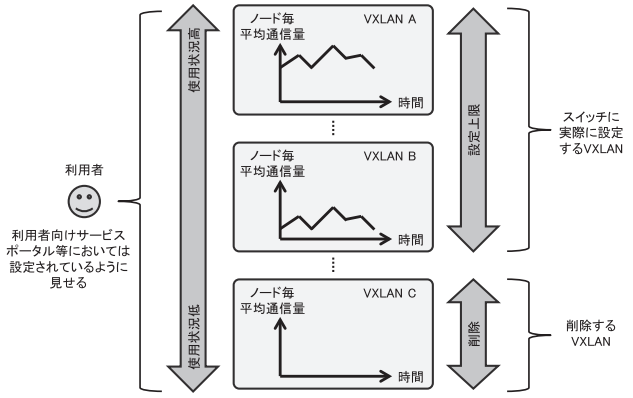

図 5 使用状況に基づく VXLAN の削除
Fig. 5 Usage based VXLAN deletion.

さらに，実際に使用しているにもかかわらず同一の VXLAN を偶然に連続して削除し，利用者が安定したサービスを享受できない可能性がある．そこで，VXLAN を再設定する際には一定期間の優先権を与え，優先権が与えられている当該期間（以降，優先権付与期間と称する）は当該 VXLAN を削除しない．半面，あまり通信をしない VXLAN に無駄に資源を割り当てる可能性もある．

3.3 使用予測に基づく削除方式

本方式においては，**図 5** に示すように，スイッチにおける VXLAN の設定数と個々の VXLAN の使用状況を定期的に測定し，VXLAN が設定上限を超過する場合には，個々の VXLAN の過去の使用状況から今後それぞれが使用される確率を算出し，今後使用される確率で順位を付けることで，設定上限を超過している VXLAN（順位が低いもの）を削除する．

計算機資源の制約を緩和するための手法として，たとえば OS の仮想記憶における LRU（Least Recently Used）アルゴリズムを用いた低使用頻度データの削除等，様々な研究が行われている．同様の考え方を VXLAN のオーバコミットに適用するためには，使用頻度の推定方法の選定が必要であるが，次節で述べるように，ベアメタルサービスにおいて測定できる情報に制限があるため，適当な指標を見つけることが課題となる．

具体的な使用予測方法は次節で述べるが，例として図 5 は VXLAN ごとの通信量を用いた場合の使用予測を示している．図中のすべての VXLAN は利用者からは設定されているように見えているが，VXLAN の総数はスイッチの設定上限を超えている．通信量が多い VXLAN ほど今後も通信する確率が高いと判断し，通信量が多い順に VXLAN を並べると，VXLAN A > B > C の順となる．VXLAN 設定上限に収まるのは VXLAN B までであるため，それより下位の VXLAN を削除する．

ネットワーク通信は定常的に発生しているとは限らず，間欠的な通信も発生しうるため，ある瞬間の使用状況だけをふまえると偶然に連続して使用されていないと判断されうる．そこで，使用予測においては，過去一定期間の通信頻度を考慮する（以降，使用状況振り返り期間と称する）．

ある瞬間に削除する VXLAN は，削除が原因でその後も順位が低下し，再接続され難い，あるいは，再接続と切断を繰り返しうる．また，利用者による利用者 VLAN の作成あるいは削除にともない VXLAN は動的に生成あるいは削除されるため，ある瞬間に新たに生成された VXLAN は既存の VXLAN と比べて相対的に使用状況が少なく，順位が低くなりうる．そこでランダム削除と同様に，VXLAN の再設定や新規生成の際には優先権付与期間を与え，そのほかの VXLAN よりも順位を引き上げる．

3.4 ネットワーク情報を用いた使用予測

ベアメタルサービスにおいては，利用者が物理サーバ上で動かすワークロードやソフトウェアは様々である．HPC や機械学習に利用する場合もあれば，仮想マシンやコンテナを用いて IaaS や VDI（Virtual Desktop Infrastructure），PaaS（Platform as a Service）を運営して別の利用者に貸す場合，SaaS（Software as a Service）等の Web サービスを運営する場合もある．また，利用者が物理サーバ上でソフトウェアベースの VXLAN を使用することで，利用者 VLAN の中に VXLAN 通信が流れ，結果として ToR スイッチでの VXLAN と多段になる場合がありうる等，通信内容も複雑であり，さらには利用者が通信を暗号化する場合もある．そのため，使用予測のために，様々なワークロードやソフトウェア，通信タイプに対応し，利用者の物理サーバの内部や通信の内部から情報を得ることは，サービス規約の観点だけでなく技術的にも実現困難である．

そこで，スイッチや周辺環境（温度や電力等）等，物理サーバの外側で得られる情報のみを用いて使用頻度を判定し，使用予測を行う．本論文ではスイッチにおいて得られるネットワーク情報を用いる．

スイッチで得られるネットワーク情報として，たとえば VLAN ごとに観測される MAC アドレスや通信量（VLAN インタフェースにおける送受信量），通信統計情報（NetFlow や sFlow）がある．VLAN インタフェースごとの送受信量や通信統計情報はスイッチによっては利用できない場合があることから，本論文では以下，どのスイッチでも利用できる MAC アドレスに着目する．

具体的には，利用者 VLAN の MAC アドレスをスイッチの MAC アドレステーブルから定期的に収集し，VLAN に対応する VXLAN 内の通信頻度を導出し，使用予測に用いる．ある利用者の物理サーバや VLAN は複数のスイッチにまたがって接続されている場合があるため，全スイッチから MAC アドレスを収集し，VXLAN 単位で集約する．ただし，スイッチによっては MAC アドレス観測状況をフォワーディングテーブル全体でしか取得できず，VLAN 単位

で取得できない場合があるため，提案方式の実現のためには，コモディティな DC 向けスイッチのなかでも，特に当該機能を有するスイッチを用いる必要がある．

ある VXLAN において MAC アドレスを観測する場合，当該 VXLAN にノード（物理サーバや仮想マシン等）が接続されていて通信が発生していることを示しうる．物理サーバ上でソフトウェアベースのネットワーク仮想化やルーティングを行う場合もあるため，スイッチで得られる MAC アドレスが実際のノードのものであるとは限らないが，当該 VXLAN 配下に何らかのノードが存在して通信を行っていると判断することができる．一方，定期的なアップデート確認やノード間の制御通信等，ノードは何らかの重要度の低い通信を継続的に行いうる．さらに，スイッチは観測した MAC アドレスを MAC アドレステーブルに一定期間保持する．そのため当該 VXLAN の実際の使用有無にかかわらず MAC アドレスが観測されうる．そこで，単純な MAC アドレスの有無だけではなく，それら MAC アドレスをもとに当該 VXLAN の使用頻度の判定を行うモデルが必要になりうる．

本論文では，MAC アドレスのみを用いて VXLAN の使用頻度を判定するモデルとして，以下の 5 種類を提案し，次章で評価を行う．MAC アドレスのみを用いた VXLAN の使用頻度判定については既存研究は存在しない．いずれのモデルにおいても，個々の VXLAN において観測される MAC アドレス情報を定期的に測定し（時間 Δt ごと），過去一定期間（使用状況振り返り期間 ΔL）に測定した MAC アドレス情報を用いて当該 VXLAN の使用頻度を判定する．ΔL は Δt の 2 倍以上の倍数値とする．ある VXLAN x の時刻 $t = T$ における使用頻度 $Ux(T)$ とし，x 内にノードが n 個（MAC アドレスが n 個）存在して，それぞれのノードが通信しうるとする．時刻 $t = T$ におけるノード i の MAC アドレスの有無を $Mi(T)$ とする（当該 MAC アドレスが観測されれば $Mi(T) = 1.0$，観測されなければ $Mi(T) = 0.0$）．

（model1）**MAC アドレス観測確率**：Δt ごとに MAC アドレスの有無を測定し（1 つでも MAC アドレスが存在すれば 1.0，存在しなければ 0.0），使用状況振り返り期間 ΔL における平均値を使用頻度とする．

$$Ux(T) = \frac{\sum_{t=T-\Delta L}^{T}(\max\{Mi(t), i = 1, 2, \ldots, n\})}{\Delta L}$$

（model2）**平均 MAC アドレス数**：Δt ごとに MAC アドレスの数を測定し，使用状況振り返り期間 ΔL における平均値を使用頻度とする．

$$Ux(T) = \frac{\sum_{t=T-\Delta L}^{T}\left(\sum_{i=1}^{n} Mi(t)\right)}{\Delta L}$$

（model3）**ユニーク MAC アドレス数**：Δt ごとに MAC アドレスを測定し，使用状況振り返り期間 ΔL 中に観測された一意な MAC アドレス数の合計値を使用頻度とする．

$$Ux(T) = \sum_{i=1}^{n}(\max\{Mi(t), t = T - \Delta L, \ldots, T\})$$

（model4）**最大観測期間**：Δt ごとに MAC アドレスを測定し，使用状況振り返り期間 ΔL 中に観測された一意な MAC アドレスごとに観測された期間を算出し，その最大値を使用頻度とする．

$$Ux(T) = \max\left\{\sum_{t=T-\Delta L}^{T} Mi(t), i = 1, 2, \ldots n\right\}$$

（model5）**平均観測期間**：Δt ごとに MAC アドレスを測定し，使用状況振り返り期間 ΔL 中に観測された一意な MAC アドレスごとに観測された期間を算出し，その平均値を使用頻度とする．

$$Ux(T) = \frac{\sum_{i=1}^{n}\left(\sum_{t=T-\Delta L}^{T} Mi(t)\right)}{n}$$

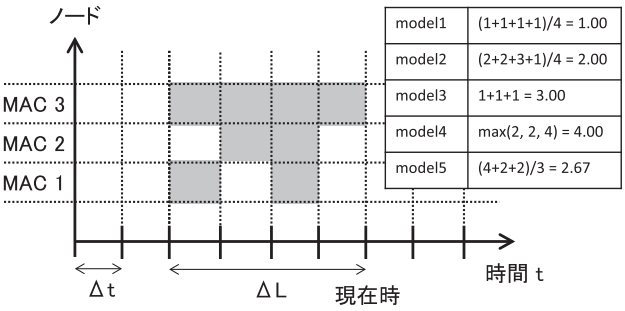

図 6 使用頻度判定モデル
Fig. 6 Usage model.

ネットワーク情報を用いた使用予測においては，これらのモデルに基づく個々の VXLAN の使用頻度を算出して VXLAN の優先順位を決定する．

これらのモデルに基づく使用頻度の算出方法を**図 6** の例を用いて説明する．図においては，ある VXLAN 内に MAC 1〜3 という MAC アドレスで表される 3 つのノードがある場合に，Δt ごとに MAC アドレスを観測し，それぞれの MAC アドレスが網掛けで示された時間に観測されたことを示している．現在時において使用状況振り返り期間 $\Delta L = 4$ に観測された MAC アドレスを基に 5 種類のモデルで使用頻度を算出した結果を図中の表に示している．

3.5 ネットワーク情報を用いた利用者通信開始の検出

利用者が VLAN 内で通信を開始しようとする際，対応する VXLAN が削除されている場合がある．VXLAN を再設定するためには通信開始を検出する必要がある．

通信開始の検出においては，前節と同様にスイッチで得られるネットワーク情報を用いることができる．具体的には，物理サーバが通信を開始する際に，スイッチの VLAN インタフェースで観測される MAC アドレスを用いる．利

用者 VLAN において MAC アドレスが観測された場合は，対応する VXLAN の再設定を行う．

ただし，MAC アドレスの取得や VXLAN の再設定の間に通信破棄やリトライが発生するため，提案方式はミッションクリティカルな利用者アプリケーションには適さない．

3.6 オーバコミット手順

3.2〜3.5 節をふまえた具体的な手順は以下のとおりである．定期的（時間 Δt ごと）に以下の手順を繰り返す．

[手順 1]：VXLAN ごとに，対応する利用者 VLAN において観測された MAC アドレスを収集し，当該 VXLAN の MAC アドレスを集約する．

[手順 2]：対応する利用者 VLAN 内で通信が観測された場合，かつ，VXLAN が削除されている場合，当該 VXLAN の優先権付与期間を ΔP とする．

[手順 3]：使用予測に基づく削除の場合，VXLAN の MAC アドレスを基に，前述のモデルを用いて使用状況振り返り期間 ΔL における使用頻度 tr を算出する．

[手順 4a]：ランダム削除の場合，設定上限範囲内の VXLAN と範囲外の削除候補 VXLAN をランダムに選ぶ．

[手順 4b]：使用予測に基づく削除の場合，tr を用いて VXLAN を降順に並べ，設定上限数の範囲内の VXLAN を接続候補とし，範囲外の VXLAN を削除候補とする．

[手順 5]：削除候補の中に優先権を与えられた VXLAN が存在する場合は，当該 VXLAN を接続候補の先頭に追加し，接続候補の最後尾の VXLAN を削除候補に移す．

4. 提案方式の評価

4.1 シミュレーション実験の概要

前章で提案した VXLAN オーバコミット方式を評価するために，シミュレーション実験を行った．シミュレーションは，VXLAN 数が x であり，かつ，個々の VXLAN 内のノード数が m のときに，前章で述べたオーバコミット手順を一定期間 Δt ごとに t_n 回実施し，全期間 $T\ (= \Delta t \times t_n)$ におけるオーバコミットの影響 y（後述）を測定した．VXLAN の設定上限数は V_{limit} であり，x を $1 \sim V_{\mathrm{max}}\ (1 < V_{\mathrm{limit}} < V_{\mathrm{max}})$ まで順次増加させて測定を行った．さらにこれらを N 回繰り返し，測定結果を平均した．また，ランダム削除に基づく削除の場合は優先権付与期間 ΔP を，使用予測に基づく削除の場合は使用状況振り返り期間 ΔL と優先権付与期間 ΔP を変えて結果を比較した．さらにノード数も変えて結果を比較した．シミュレーション実験の環境は**表 2**，実験におけるそれぞれの条件値は**表 3** に示すとおりである．実環境における V_{limit} は 2.2 節で示したとおり数百〜6,000 であるが，シミュレーション実験は V_{limit} に対するオーバコミットの度合いを算出することを目的としているため，シミュレーション実験に要する時間の短縮のために，V_{limit} を 100 とした．

表 2 シミュレーション環境
Table 2 Simulation environment.

OS	ソフトウェア	言語
Ubuntu 16.04	Jupyter 4.4.0	Python 3.6.6

表 3 前提条件
Table 3 Conditions.

V_{limit}	V_{max}	Δt	t_n	T（$=\Delta t \times t_n$）	N
100	200	10 秒	8640 回	86400 秒 (24 時間)	4 回

表 4 VXLAN のタイプ
Table 4 Types of VXLAN.

タイプ	通信の特性
unused	個々のノードは通信しない．
random	個々のノードがランダムに通信する．
periodical	個々のノードが一定間隔で一定期間通信する．間隔と期間はノード毎にランダムである．
flat	個々のノードが常に通信する．

筆者らが運用している研究開発用のベアメタルサービスにおける VXLAN の通信量を測定した結果（表 1）をふまえ，VXLAN については**表 4** に示す 4 タイプのネットワークを想定した．unused は通信が発生していないネットワーク，random はノードが通信をランダムに発生させているネットワーク，periodical はノードが通信を周期的に発生させているネットワーク，flat はノードが通信をつねに発生させているネットワークをそれぞれ表している．シミュレーション実験においては，VXLAN を V_{max} まで追加する際に，表 1 をふまえて，unused タイプの VXLAN を約 3 割（31%）の確率で生成し，それ以外の 3 タイプの VXLAN を約 7 割（それぞれ 23%で合計 69%）の確率で生成した．また，使用状況振り返り期間 ΔL と優先権付与期間 ΔP の効果を同条件で測定するために，あるノード数 m のときに，個々の VXLAN における全期間 T の通信を N 回分，あらかじめ生成して測定を行った．

ノードが VLAN 内で通信を発生させる際に，対応する VXLAN が削除されている場合には，3.5 節で示したように通信発生を検出して当該 VXLAN を再設定する．その際，再設定された回数をオーバコミットによる影響 y として測定し，全測定回数 t_n に対する割合を算出した．本論文ではこれを Affected Rate と呼ぶ（単位は%）．たとえばある VXLAN の Affected Rate が 1%の場合，当該 VXLAN 内のノードが通信をしようとした際に当該 VXLAN が削除されていたために再設定された期間が，全期間 T の 1%存在していたことを意味する．次節のシミュレーション結果，およびその分析においては，Affected Rate を VXLAN のタイプごとに平均した．

利用者にはオーバコミットを知らせず，使用状況に基づいて VXLAN を動的に再設定するため，たとえば Affected

Rate が 1%であってもサービスが利用できなかった期間が 1%（すなわちサービスレベルが 99%）という意味ではない．しかし最悪のケースにおいては，再設定を待つ間に利用者がサービスを利用できないと認識する可能性があるため，Affected Rate はサービスレベルと同程度に低いことが望ましい．一般的なクラウドサービスのサービスレベルが最低でも 99.9%であることをふまえ，本論文では Affected Rate の目標値を 0.01%とし，Affected Rate が 0.01%に達するときの VXLAN 数 $V_{a=0.01}$ を基にオーバコミットの達成度合いを評価する．

VXLAN オーバコミットにともなうネットワーク情報の取得や VXLAN の削除および再設定に要する時間を実機で測定したところ，約 2.5 秒であった．3.5 節で示したとおり，実環境では Δt ごとにこの時間が発生する．ノードが通信をリトライする間に再設定が完了するのに十分な短さであるが，リトライが発生することからミッションクリティカルなアプリケーションには適さない．また今回のシミュレーション実験においては，個々のノードの通信を VXLAN タイプに沿った MAC アドレスの出現パターンのみで模擬しており，通信のリトライを含めた実際の通信の挙動を模擬していないため，削除および再設定に要する時間がシミュレーション結果に影響しないことから，評価においては削除および再設定に要する時間は無視している．

4.2 シミュレーション結果（ランダム削除）

ランダム削除によるオーバコミットにおいて，例としてノード数 $m = 2$，優先権付与期間 $\Delta P = 1$（$\Delta t \times 1$ を意味する，以降同様）とした場合の Affected Rate の測定結果を**図 7** に示す．VXLAN 数が V_{limit} 未満の場合は全タイプの VXLAN で Affected Rate が 0 であるため，図では省略している（以降の図も同様）．VXLAN 数が V_{limit} を超えると，unused 以外の全タイプの VXLAN が影響を受け始め，Affected Rate が上昇する．特に全期間で常時通信が発生するする flat が影響を受けやすく，ランダムあるいは周期的に通信する random と periodical が続く．V_{limit} を超えて VXLAN 数が 101 になると Affected Rate が 0.01 に達するため，$V_{a=0.01}$ は 101 である．

$m = 2$，$\Delta P = 15$ とした場合の結果を**図 8** に示す．$\Delta P = 1$ の場合と比較すると，VXLAN 数が V_{limit} を超えた際に unused 以外の全タイプの VXLAN において Affected Rate が上昇する傾向は同じだが，全体的にはその値が改善されている．ΔP を増加させると同じ傾向のまま全体的に値が改善されていくが，$\Delta P = 20$ 程度で改善が頭打ちとなる．また，ΔP を増加させると全体的に Affected Rate は改善されるが，$V_{a=0.01}$ は変わらず 101 のままである．

あるノード数 m（$m = 1, 2, 5, 10, 15, 20$）のときに，ΔP を増加させて（$\Delta P = 1, 2, 3, 4, 5, 10, 15, 20, 25, 30$）$V_{a=0.01}$ をそれぞれ測定すると，ランダム削除においては m や ΔP によらず $V_{a=0.01}$ は 101 で変わらない．

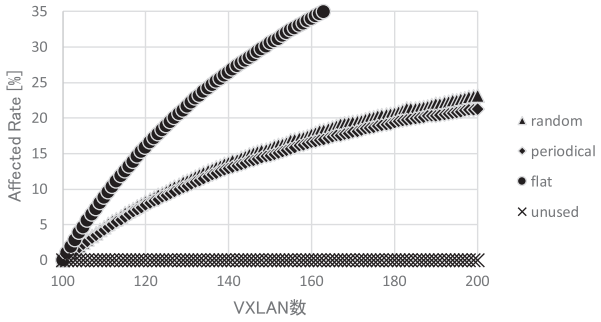

図 7 ランダム削除方式（$m = 2$，$\Delta P = 1$）
Fig. 7 Random method ($m = 2$, $\Delta P = 1$).

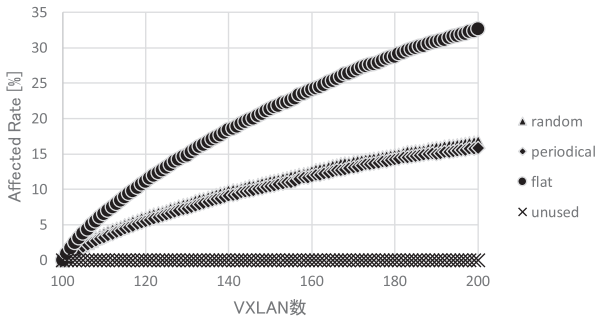

図 8 ランダム削除方式（$m = 2$，$\Delta P = 15$）
Fig. 8 Random method ($m = 2$, $\Delta P = 15$).

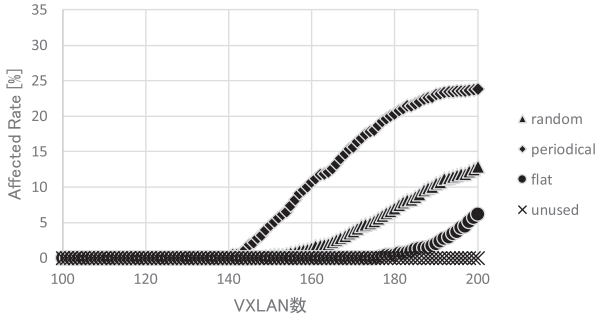

図 9 使用予測に基づく削除方式（$m = 5$，$\Delta L = 10$，$\Delta P = 5$，model3）
Fig. 9 Usage-based method ($m = 5$, $\Delta L = 10$, $\Delta P = 5$, model3).

4.3 シミュレーション結果（使用予測に基づく削除）

使用予測に基づく削除によるオーバコミットにおいて，例としてノード数 $m = 5$，使用状況振り返り期間 $\Delta L = 10$，優先権付与期間をその半分（$\Delta P = 5$）とした場合の，model3 における結果（VXLAN 数に対する Affected Rate）を**図 9** に示す．VXLAN 数が 140 程度まではどのタイプの VXLAN も影響を受けない．全 VXLAN に占める unused タイプの VXLAN の割合が約 3 割であることを考慮すると，unused タイプの VXLAN の分だけ VXLAN の設定上限 V_{limit} を超えて正しくオーバコミットできている．VXLAN がさらに増加すると，periodical タイプの VXLAN，random

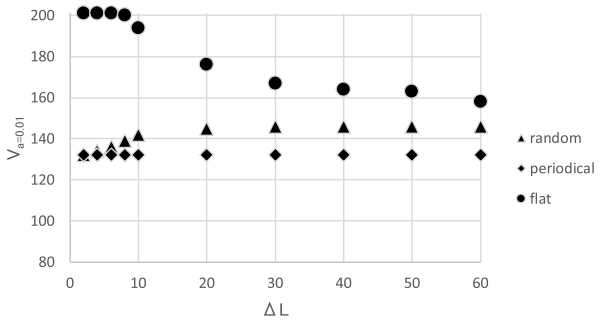

図 10 ΔL に対する $Va = 0.01$ の変化（$m = 10$，model3）
Fig. 10 $Va = 0.01$ and ΔL ($m = 10$, model3).

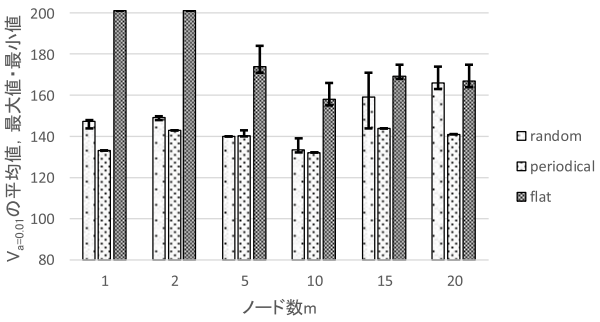

図 11 ノード数 m に対する $V_{a=0.01}$ の変化（model1）
Fig. 11 $V_{a=0.01}$ and the number of node m (model1).

タイプの VXLAN，flat タイプの VXLAN がそれぞれ影響を受け始め，Affected Rate がそれぞれ上昇する．Affected Rate が上昇し始める VXLAN 数や $V_{a=0.01}$，どのタイプが先に上昇するか，どの程度上昇するか，については，ノード数 m や使用状況振り返り期間 ΔL，優先権付与期間 ΔP，どのモデルで使用予測をするかによって異なる．

$m = 10$ および model3 のときに，ΔL と ΔP を増加させて（$\Delta L = 2, 4, 6, 8, 10, 20, 30, 40, 50, 60$，$\Delta P$ はそれらの半分）$V_{a=0.01}$ をそれぞれ測定した結果を**図 10** に示す．VXLAN 数が $V_{\max} = 200$ を超えても Affected Rate が 0.01 に到達しない場合は便宜上 $V_{a=0.01} = 201$ としている．ΔL を増加させると，random タイプの VXLAN は影響を受け難くなり $V_{a=0.01}$ が増加するが，flat タイプの VXLAN は影響を受けやすくなり $V_{a=0.01}$ が低下する．$\Delta L = 40$ 程度で $V_{a=0.01}$ はそれ以上変化しなくなるが，VXLAN のタイプによっては VXLAN 数が 140 を超えてオーバコミットできている．このような ΔL に対する $V_{a=0.01}$ の変化は，m やモデルによって異なる．

そこで，それぞれのモデルにおいて m を増加させて（$m = 1, 2, 5, 10, 15, 20$），ΔL に対する $V_{a=0.01}$ を測定した結果を**図 11**，**図 12**，**図 13**，**図 14**，**図 15** に示す．図中の棒グラフと誤差範囲線は，unused 以外の VXLAN タイプにおける，ΔL に対する $V_{a=0.01}$ の平均値と最大値・最小値を示す．いずれのモデルにおいても VXLAN のオーバコミットは正しく行われているが，モデルによってオーバコミットの度合いや VXLAN タイプごとの公平性が異なる．

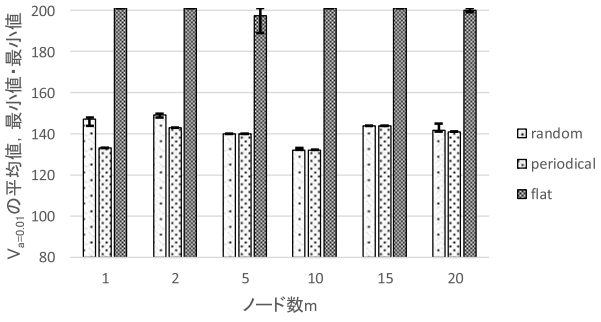

図 12 ノード数 m に対する $V_{a=0.01}$ の変化（model2）
Fig. 12 $V_{a=0.01}$ and the number of node m (model2).

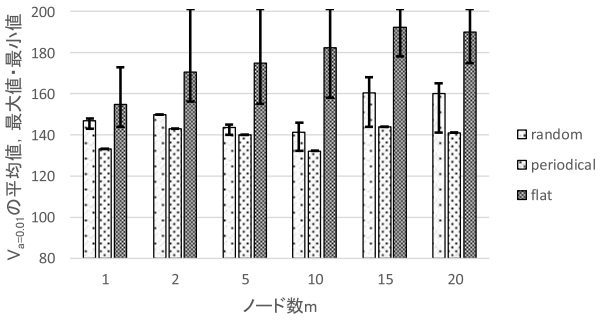

図 13 ノード数 m に対する $V_{a=0.01}$ の変化（model3）
Fig. 13 $V_{a=0.01}$ and the number of node m (model3).

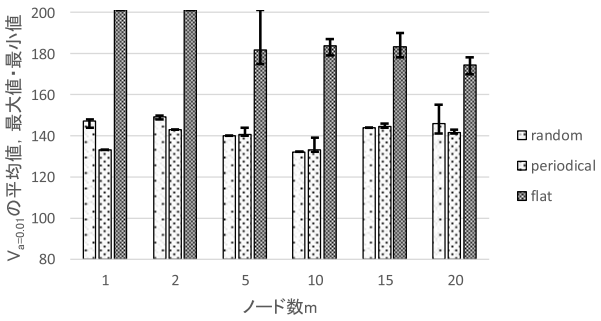

図 14 ノード数 m に対する $V_{a=0.01}$ の変化（model4）
Fig. 14 $V_{a=0.01}$ and the number of node m (model4).

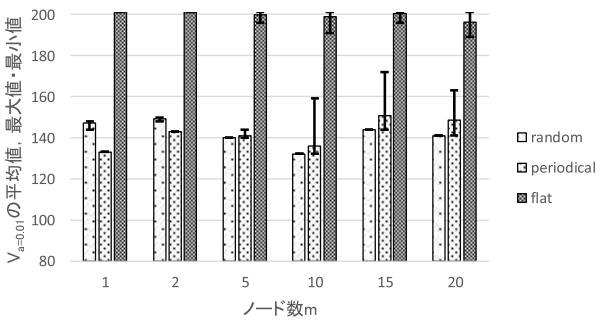

図 15 ノード数 m に対する $V_{a=0.01}$ の変化（model5）
Fig. 15 $V_{a=0.01}$ and the number of node m (model5).

model1（図 11）はノード数が少ない場合は flat タイプの VXLAN が優先されやすく，ノード数が増えると $V_{a=0.01}$ が増加するとともに，タイプ間の $V_{a=0.01}$ の差異が小さくなり公平性が増す．ノード数 20 の場合の $V_{a=0.01}$ の平均値は，

random タイプが 166.1，periodical タイプが 141，flat タイプが 166.8 に達する．一方，ΔL の増加に対する $V_{a=0.01}$ の変化はノード数によって異なる．たとえばノード数 15 の場合，ΔL を増加させると random タイプの $V_{a=0.01}$ は増加し，flat タイプの $V_{a=0.01}$ は減少するが，ノード数 20 の場合，ΔL を増加させると random タイプと flat タイプともに $V_{a=0.01}$ は減少する．model2（図 12）はノード数の影響を受けにくいが，$V_{a=0.01}$ が他のモデルと比較して高いわけではなく，またつねに flat タイプが優先されて公平性が低い．model3（図 13）は，ノード数が増加するに従って flat タイプの VXLAN が優先されやすくなるが，ΔL を増加させと flat タイプの優先度が下がり random の優先度が上がるため公平性を改善することができる．また $V_{a=0.01}$ も他のモデルと比較して高くなる傾向にあり，ノード数 20 および $\Delta L = 60$ の場合には，random タイプの $V_{a=0.01}$ は 165，flat タイプの $V_{a=0.01}$ は 178 に達する．model4（図 14）と model5（図 15）は model2 と類似の傾向にあるが，ノード数の増加にともない公平性が改善される．特に model5 は他のモデルと異なり，ノード数が増加すると periodical タイプの $V_{a=0.01}$ が増加する．

4.4 シミュレーション結果の考察

ランダム削除方式は VXLAN の設定上限を超えると全タイプの VXLAN で Affected Rate が上昇し，目標値の 0.01 を超えてしまう．優先権付与期間を設けることで Affected Rate は改善させるが，Affected Rate が 0.01 に到達するときの VXLAN 数は変わらないため有効ではない．

使用予測に基づく削除方式は，いずれのモデルにおいても，少なくとも未使用の VXLAN 分のオーバコミットを達成している．さらに model1 および model3 は，ノード数が増加すると，特に random タイプと flat タイプにおいてさらなるオーバコミットを実現できる．model5 も同様に，ノード数が増加すると，特に periodical タイプと flat タイプにおいてさらなるオーバコミットを実現できる．

model1，model3，model5 について，オーバコミットの限界と比較すべく，VXLAN の通信を先読みできると仮定した場合について測定を行い，上限を導出した．具体的には，事前に生成した個々の VXLAN の通信を基に，次に通信が発生する VXLAN を特定し，通信を発生させるノードが多い順に VXLAN に優先順位を付け，設定上限を超える VXLAN を削除することで測定した．

測定した結果を**図 16** に示す．ノード数が増えると VXLAN 間の優先順位がつきにくくなることから，先読みの効果が現れにくい．そのため，ノード数 1 の場合の測定結果に先読みの効果が最も現れており，VXLAN のタイプごとの $V_{a=0.01}$ の上限値を示していると考えることができる．すなわちオーバコミットの上限は，random タイプで $V_{a=0.01}$ が約 167，periodical タイプで約 178，flat タイプで約 186 である．

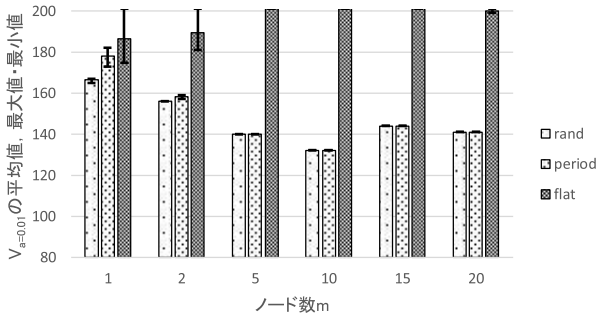

図 16 先読みに基づく削除方式
Fig. 16 Theoretical result.

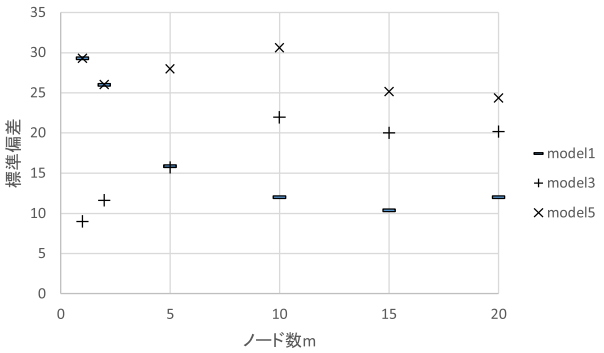

図 17 個々のモデルにおける $V_{a=0.01}$ の標準偏差
Fig. 17 Standard deviation of $V_{a=0.01}$.

model1 は，ノード数が少ない場合は上限値に及ばないが，ノード数が増加すると random タイプにおいて上限値と同程度のオーバコミットを実現できる．一方，使用状況振り返り期間の増加に対するオーバコミットの度合いはノード数によって異なる．これは，model1 が MAC アドレスの有無しか考慮していないため，ノード数が少ない場合は個々の VXLAN タイプの通信特性の影響が出やすく，ノード数が多い場合は影響が出にくくなり，使用状況振り返り期間の効果が変わるためである．model3 は，ノード数が多い場合に random タイプと flat タイプにおいて上限値に近づく．model5 は，ノード数が多い場合に periodical タイプと flat タイプで上限値に近づく．以上のことから，モデルごとに，上限値に近いオーバコミットを実現可能な通信タイプが異なる．

次に model1，model3，model5 の公平性について比較すべく，個々のモデルにおいて，VXLAN タイプごとの $V_{a=0.01}$ の平均値（図 11，図 13，図 15 の棒グラフ）の標準偏差を算出した．標準偏差が小さいほど，VXLAN タイプ間での $V_{a=0.01}$ の差異が少ないため，VXLAN タイプ間の公平性が高いと考えることができる．算出結果を**図 17** に示す．model1 は，ノード数が多い場合に VXLAN タイプ間で Affected Rate の差異が少なくなり公平性が増す．具体的には，$V_{a=0.01}$ の平均値の標準偏差が小さくなり，値が 11 前後となる．これは，model1 が個々の VXLAN におけ

表 5 シミュレーション結果
Table 5 Simulation result.

モデル	オーバーコミットの度合い	公平性
model1	○	△
model3	○	○
model5	△	×

る使用状況振り返り期間中の MAC アドレスの有無しか考慮しておらず，ノード数が増加するといずれの VXLAN タイプも差異がなくなるためである．model3 は，model1 とは逆に，ノード数が多くなると $V_{a=0.01}$ の平均値の標準偏差が大きくなり，値は 20 前後となる．ただし，使用状況振り返り期間を変化させた場合の $V_{a=0.01}$ の最大値・最小値に注目すると，使用状況振り返り期間が長いときの $V_{a=0.01}$ の標準偏差は小さくなる．たとえばノード数が 20 のときの標準偏差を，使用状況振り返り期間を長くすると約 14 に抑えることができる．これは，model3 が個々の VXLAN における使用状況振り返り期間中の一意な MAC アドレス数を考慮しているため，使用状況振り返り期間を長くすることで VXLAN のタイプ間での Affected Rate の差異を少なく抑えることができるためである．使用状況振り返り期間を増減させることで公平性をコントロールすることができるため，model1 よりも実環境に適用しやすい．model5 は，ノード数が多いと periodical タイプのオーバコミット度合いが上昇するが，基本的には flat タイプのオーバコミット度合いが高い．図 17 で示すとおり，標準偏差はノード数にかかわらず 25〜30 であり，model1 や model3 と比較して高く，公平性を欠く．以上のことから，ノード数が少なければ model3 が，ノード数が多い場合は model1 が公平性の観点で優れている．さらに，model3 は使用状況振り返り期間を増加させることでノード数が多い場合の公平性を改善することができるため，ノード数の変化にかかわらず全般的に優れている．

model1，model3，model5 について，オーバコミットの度合いおよび公平性の比較結果を**表 5** にまとめる．

5. おわりに

本論文では VXLAN オーバコミットの具体的な方式を示し，実環境をふまえたシミュレーション実験によりその効果や影響を評価し，オーバコミットの実現性を確認した．MAC アドレスのみを用いて VXLAN の使用頻度を判定する 5 つのモデルを提案し，そのうち 3 つのモデルについて，未使用の VXLAN 分を超えたオーバコミットを行えることを確認した．

今後は実環境に適用してその効果や影響を評価する必要がある．たとえば，シミュレーションにおいては，実環境の通信特性をふまえた VXLAN のタイプを定義し，VXLAN のタイプに沿って MAC アドレスの発生を模擬したが，実環境においては VXLAN の削除や再設定が通信断や通信リトライ等を引き起こし，その後の通信特性に影響を及ぼすことで，MAC アドレスの発生パターンに影響しうるためである．

また，本論文の実機評価においては，VXLAN 機能を有する ToR スイッチにログインして設定を編集・再読み込みするという人手作業の自動化を行ったため約 2.5 秒を要しており，その間に通信遮断が発生する．これは利用者においても認識可能なレベルであり短縮が必要である．よりオーバヘッドの少ない API 等を ToR スイッチ側に設ける等の対策を検討する必要がある．

また今回のシミュレーション結果をふまえると，VXLAN の使用頻度の判定モデルが同じであっても，ノード数や VXLAN のタイプ，使用状況振り返り期間によって，オーバコミットにより個々のノードが受ける影響や公平性は異なる．今回のシミュレーションではノード数を VXLAN 間で等しくし，未使用以外の VXLAN タイプの割合も等しくしたが，個々の VXLAN のノード数や割合をふまえて適用するモデルを動的に変える，あるいは機械学習を用いる等することで，使用頻度の判定精度を向上させうる．これらの改善については今後の課題とする．

さらに今回のシミュレーション実験では個々のノードの利用者には着目していないが，個々のノードが VXLAN 削除により受ける影響と個々の利用者が VXLAN 削除により受ける影響は一致するとは限らない．たとえば個々のノードを異なる利用者が使用している可能性がある一方で，個々のノードを同じ利用者が使用して，たとえばクラスタを構築している可能性がある．個々のノードが VXLAN 削除により受ける影響を抑えても，後者では実際の影響は大きくなりうる．これら，利用者視点での公平性の評価についても今後の課題とする．

参考文献

[1] IEEE 802.1Q - Virtual LANs, available from ⟨http://www. ieee802.org/1/pages/802.1Q.html⟩ (accessed 2020-03-25).
[2] Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks, IETF, RFC7348 (2014), available from ⟨https://tools.ietf.org/html/rfc7348⟩.
[3] IEEE 802.1ah - Provider Backbone Bridges, available from ⟨http://www.ieee802.org/1/pages/802.1ah.html⟩ (accessed 2020-03-25).
[4] IEEE 802.1ad - Provider Bridges, available from ⟨http://www.ieee802.org/1/pages/802.1ad.html⟩ (accessed 2020-03-25).
[5] Multiprotocol Label Switching Architecture, IETF, RFC3031 (2001), available from ⟨https://tools.ietf.org/html/rfc3031⟩.
[6] Cisco: FabricPath Encapsulation, available from ⟨https://www.cisco.com/en/US/docs/switches/datacenter/sw/5_x/nx-os/fabricpath/configuration/guide/fp_switching.html#wp1790848⟩

(accessed 2020-03-25).
[7] NVGRE: Network Virtualization Using Generic Routing Encapsulation, IETF, RFC7637 (2015), available from ⟨https://tools.ietf.org/html/rfc7637⟩.
[8] A Stateless Transport Tunneling Protocol for Network Virtualization (STT), IETF (2016), available from ⟨https://tools.ietf.org/html/draft-davie-stt-08⟩ (accessed 2017-03-18).
[9] Pfaff, B., Pettit, J., Kopenen, T., Amidon, K., Casado, M. and Shenker, S.: Extending Networking into the Virtualization Layer, *ACM SIGCOMM Workshop on Hot Topics in Networking* (*HotNets*) (2009).
[10] Cisco Application Centric Infrastructure Fundamentals, available from ⟨http://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/1-x/aci-fundamentals/b_ACI-Fundamentals/b_ACI_Fundamentals_BigBook_chapter_0100.html⟩ (accessed 2017-03-18).
[11] Kinoshita, J., Maeda, K., Yabusaki, H., Akune, K., Noumi, M. and Komoda, N.: Implementation and Evaluation of VXLAN Gateway-based Data Center Network Virtualization, *Studies in Informatics and Control Journal*, Vol.25, No.3, pp.313–322 (2016).
[12] Kinoshita, J., Maeda, K., Yabusaki, H., Akune, K. and Komoda, N.: Virtualizing Service Infrastructure with Hardware Gateway in Data Center, *Proc. 7th Int. Conf. on Data Communication Networking* (*DCNET 2016*), pp.95–98 (2016).
[13] VXLAN Scale, available from ⟨https://docs.cumulusnetworks.com/display/DOCS/VXLAN+Scale⟩ (accessed 2020-03-25).
[14] AX3660S Manual, available from ⟨https://www.alaxala.com/jp/techinfo/archive/manual/AX3660S/HTML/12_1_J/CFGUIDE/0025.HTM⟩ (accessed 2020-03-25).
[15] ネットワーク構成（ベアメタルサーバー），入手先 ⟨https://doc.cloud-platform.kddi.ne.jp/service/network/network-baremetalserver/⟩（参照 2020-03-25）.
[16] ベアメタルサーバー，入手先 ⟨https://ecl.ntt.com/en/documents/service-descriptions/rsts/server/baremetal-server.html⟩（参照 2020-03-25）.

木下 順史

2000 年京都大学大学院工学研究科機械工学専攻修士課程修了．同年（株）日立製作所入社．現在，IoT プラットフォームの研究に従事．

薦田 憲久

1974 年大阪大学大学院工学研究科電気工学専攻修士課程修了．同年（株）日立製作所入社．1991 年大阪大学工学部助教授，1992 年同大学教授．2015 年大阪大学名誉教授，コーデソリューション（株）顧問．2016 年より大阪大学招へい教授．工学博士．技術士（情報工学）．IEEE，電気学会の終身会員．

藤原 融 （正会員）

1981 年大阪大学基礎工学部情報工学科卒業．1986 年同大学大学院基礎工学研究科博士課程修了．工学博士．同年同大学助手．1997 年より同大学教授．現在，情報科学研究科所属．符号理論，情報セキュリティの研究に従事．電子情報通信学会フェロー，IEEE 等の会員．