L-015

Bloom filter を用いた分散トラフィックログ管理システムの提案
Distributed traffic log data management system using bloom filter

朝倉 浩志 †
Hiroshi ASAKURA

1． まえがき

本稿では，仮想サーバ(VM; virtual machine)・ネットワーク(VN; virtual network)を貸し出すクラウドサービス事業においてトラフィックログを分散保存し，必要なトラフィックデータを素早く参照するシステムについて提案する．

事業者(クラウドプロバイダー)は，設備の運用管理を行い正常動作する VM，VN を提供する必要があるが，貸し出した VM 内部の状態は参照することができない．このため，物理サーバ上で取得可能な情報や，VM のトラフィック情報を参考に，切り分けや正常・異常動作の確認といったトラブルシューティングを行っていく必要がある．本稿ではトラフィック情報に着目し，トラブルシューティングを目的として，必要な情報を即座に参照できるシステムを提案する．

2．データセンタネットワーク

以下にクラウドプロバイダーが構築するデータセンタの典型的なネットワーク構成を示す．通常，木構造構成が用いられる．クラウドの規模によって木構造の高さが異なるが，3 層構造程度となることが多い．

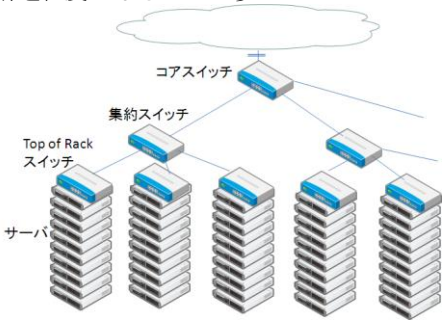

図 1. ネットワーク構成例

物理サーバ上では複数の VM が動作し，それぞれ異なる利用者に貸し出されている．あるタイミングではライブマイグレーションが行われ VM が別の物理サーバに移動する．このため同じディスティネーション IP アドレスを持った通信でも経路が変更される．また，異なる利用者でブロードキャストドメインを分けるため tagged VLAN(IEEE802.1Q, 以下 VLAN)を利用し VN を構成することが多い．このような環境では，IP アドレスと VLAN ID の組で通信を識別することができる．

3. トラフィック情報の活用

トラフィック情報を取得，蓄積するための技術として一般的なのはポート単位のトラフィック量を取得する SNMP，トラフィックの内容を取得できる sFlow, NetFlow, IPFIX といった xFlow 技術やパケットキャプチャがある．ここでは通信の内容を識別できる後者の技術を対象とする．

xFlow は 2 拠点間の通信内容をプロトコル種別まで知ることができ VM や VN の動作をトラフィックの観点から確認することができる．従来，当該技術はマクロ的な視点でシステム全体のトラフィックについてサマリ情報を得るために使われるのが主であった．本稿では，トラブルシューティングの為に，確認したいトラフィックを VLAN ID と IP アドレスの組で指定し，必要なフロー情報を参照するという使い方を実現する．VN では，複数の VM の通信が重畳されており，特定の VLAN ID と IP アドレスで識別される VM が何のプロトコルでどれだけの帯域を消費していたかといった情報はトラブルシューティングに必要である．

パケットキャプチャは従来からトラブルシューティングのために使われてきたが，そのデータ量の多さから問題発生箇所で局所的かつ一時的に使われてきた．本稿では，常時キャプチャを行い，ある一定時間遡り分析するような使い方を実現する．本手段は HTTP 等 L7 プロトコルのヘッダ情報などまで分析することができ，トラブルシューティングの時間短縮が期待できる．

本システムでは，これら二つのトラフィックログをトラブルシューティングに活用できるシステムを提案する．

4. 提案システム

4.1. トラフィックログの取得方法

木構造型のネットワーク構成において，データセンタ内部及び外部へのトラフィック交流を全て取得するためには，全てのホスト OS 内にある仮想スイッチにプローブ(probe)またはフローコレクタ(collector)を接続すれば良い．ここで，ホスト OS は物理サーバを管理する OS を指す．全ての仮想スイッチを監視することで，ライブマイグレーションによりトラフィックの経路が変更された場合でもトラフィックの取得に取りこぼしは発生しない．また，プローブとフローコレクタの機能を VM で実現し，同一物理サーバ内に格納することでハードウエアを設置する際と比べ設備コストを抑制する(図 2).

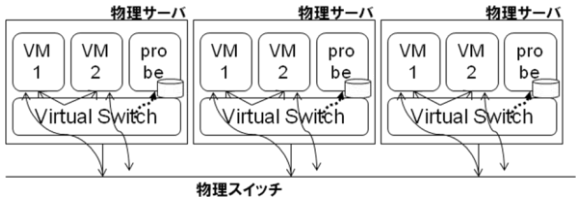

図 2. プローブ・フローコレクタの設置

4.2. トラフィックログの蓄積方法

ログの蓄積については一カ所に集約する場合と、分散して保持する場合が考えられる．文献[1]では 40 台程度のサーバログを一カ所に集約する方法が紹介されている．ログをある条件で参照する目的からは一カ所に蓄積する方が望ましい．しかし，トラブルシューティングで参照するエントリは蓄積されたログ全体のごく一部であり，大半は参照されることなく一定時間経過後破棄される．このため転送して一カ所に集約するよりもログを取得する機器もしくはネットワーク的に近い場所で蓄積する方が良い．特に，パケッ

トキャプチャの場合はキャプチャされたデータ分のトラフィックが転送時に発生することから，一カ所への集約は現実的ではない．

このため，プローブやフローコレクタと同じ VM 内にログを蓄積することとする．

4.3. トラフィックログの参照方法

サーバやネットワークを管理する際に問題解決で必要となる分析は、仮説に基づき原因となるログ情報が存在するかどうかを確認する作業とも言える。このため，本稿ではある時間(問題発生時間)において発生した通信(VLAN ID, IP アドレスによって識別)を参照するというケースを考える．

4.4. トラフィックログの管理方法

分散保存されたログの管理については，インデックスサーバを用意し高速に参照できるようにする．フローコレクタまたはプローブは xFlow やキャプチャデータを受信した際に，最低限(VLAN ID, IP，時刻,自 IP)を含んだ要約情報を作成し，インデックスサーバに登録する．登録のタイミングは任意の時間とする. インデックスサーバは 4.3.で述べた問い合わせに対し, 蓄積されている場所(IP アドレス)を高速に返す．アーキテクチャとしては，P2P ネットワーク分野でのハイブリッドタイプとも言えるが，既存の方法ではログを対象としたものはない．このため，ログの特徴にあった効率の良い索引方法を導入する．分散蓄積されたデータの管理については，分散ストレージの分野で研究が行われているが，永続性やコンテンツの発見等に主眼が置かれており，今回の課題とは異なる．

トラフィックログ情報の特徴として挙げられるのは，

- 追加は随時行われる
- 更新はない
- 削除は一定時間が過ぎた場合のみ行われる
- ログのデータ量が巨大(キャプチャの場合は，全体で数テラバイト/日を仮定)

これらの特徴を踏まえた上で効率の良い索引を適用する必要がある．本稿では Bloom Filter を用いる．

4.5. Bloom Filter を用いた索引

Bloom Filter (以下 BF)は空間効率の良いデータ構造であり，ある要素が集合に属するかどうかを判定することができる．様々なネットワークアプリケーションに使われている[2]が，トラフィックログの索引としては使われていない．このデータ構造はいくつかの特徴を持つが，以下の点において今回扱うデータと親和性がある。

(1) 他の索引と比べ空間効率が良い
トラフィックログは全体で 1 日当たりテラバイトオーダーの容量であり，これらを索引付けするためには空間効率が良いデータ構造が必要である．

(2) 基本的に要素(索引)の追加しかできない
ログ情報は常に追加され更新が行われないことから親和性がある．

(3) 偽陽性があるが偽陰性はない
集合に属さない要素に対して「集合に含まれる」と誤判定する場合があり，これを偽陽性と呼ぶ。これに対し、集合に属する要素に対し、「集合に含まれない」と誤判定する偽陰性は無い。このため取りこぼしがなく探したいログ情報を取り出す為には問題はない．

一方、課題となる点もある．BF をそのまま適用した場合、一定期間経過しログが廃棄されると，索引と保持している実データに齟齬が生じる．索引は要素の追加は可能であるが削除が不可能であるため，索引を作り直さなければならず問題である．このため，一定時間毎に BF を作成し，BF のキューとして保持する(図 3)．そして，ログが破棄された場合には, 破棄された時間分の BF をキューから削除することで再索引を防ぐことができる．

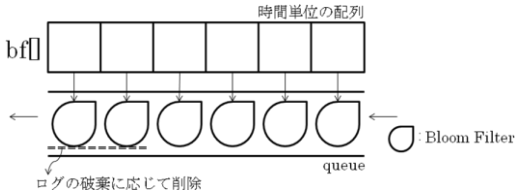

図 3. 索引構造

また，問い合わせに軸間を含むことから，時間単位の配列をもつ．BF は各フィルタの論理和で集約可能な性質をもつためため，指定された時区間の BF に対し論理和を計算し、問い合わせを行えば，時区間が指定された場合の処理が行える．

5. 比較検討

従来の監視製品では, xFlow を用いて主に統計処理を行い，ネットワーク全体の分析に用いていたが，本システムでは，具体的な通信を特定して参照することができ，特定のトラフィックが正常に流れていたか等のトラブルシューティングに利用できる．また，パケットキャプチャも従来は局所的にアナライザのような形で利用していたものを，データセンタ全体の監視装置として利用できるようになる．

6. まとめ

クラウド向けデータセンタ内のトラブルシューティングを目的として分散トラフィックログの管理システムを提案した．特に複数の利用者が設備を共有することから従来と比較し，問題は複雑化する．このような環境においてトラフィックログは一つの重要な情報であり，全体を統合的にカバーし個別のログを参照できることは重要である．

本システムではライブマイグレーションによる経路変化への対応ができること，大規模なデータセンタにおいても分散してログを蓄積し，高速に蓄積箇所を特定できること等が特徴として挙げられる．今後は，蓄積部分で高速にトラフィックログを取り出す方法についての検討や，索引の大きさに関する見積もり，問い合わせの拡張，実装を含めた処理能力の評価等を行う予定である．

[1] 安井 真伸, 横川 和哉, ほか, “サーバ/インフラを支える技術～スケーラビリティ、ハイパフォーマンス、省力運用”, 技術評論社, 2008

[2] Andrei Broder, Michael Mitzenmacher, Network Applications of Bloom Filters: A Survey, Internet Mathematics, 1(4): 485-509, 2004

† 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所， NTT Information Sharing Platform Laboratories.