L-014

VM 間のトラフィック交流を考慮した仮想サーバの効率的な配置方法の提案
An Effective Arrangement of Virtual Machines Based on traffic matrix between VMs

朝倉 浩志 † Hiroshi ASAKURA　倉上 弘 † Hiroshi KURAKAMI　山田 博司 † Hiroshi YAMADA

1． まえがき

本稿ではデータセンタにおける仮想サーバ(VM; virtual machine)の効率的な配置方法を提案する．

サーバ仮想化技術の進展により仮想化されたサーバやネットワークインフラの提供を行う IaaS 事業が盛んである．クラウドプロバイダーと呼ばれる事業者が，「クラウド環境」を利用者に提供する．クラウド環境の特徴として利用者自身で自由にシステム構築を行うことができること，要求性能の変化に対し柔軟な変更が可能なことがある．VM の CPU 資源やメモリ容量の割り当てを増減させるスケールアップ/ダウン，VM 数を増減させシステム全体としての処理能力を増減させるスケールアウト/インといった手法がある．物理サーバを更改，増減設していた時に比べ，要する時間が格段に減少したため一つの魅力となっている．

このような柔軟な変更要求に対応するためには，サーバ・ネットワーク双方の資源を統合的に管理し，自動的に資源を割り振っていく運用が必要である。

2．データセンタ内ネットワークの構成

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，物理サーバを収容する ToR(Top of Rack; 通常サーバラックの上部に配置されることからそう呼ばれる)スイッチ，ToR スイッチを束ねる集約スイッチ，集約スイッチを束ねるコアスイッチ等から構成される．

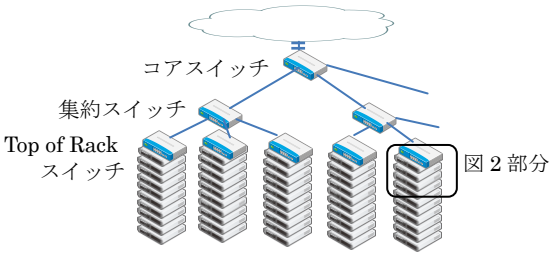

図 1. ネットワーク構成例

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．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)を利用しネットワークを構成することが多い．また，同一の利用者が複数の L2 セグメントを使用してシステム構築を行う場合がある．この場合も同様に VLAN を用いて実現されることが多い．

このようなネットワーク構成において，新規契約やスケールアウトを目的とした VM の増設要求があると，特定のルールに従って物理サーバ上に VM が割り当てられる．また，解約やスケールインのための削除要求があると，VM が取り除かれる．

† 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.

3．課題

利用者による VM の増減が繰り返されると，それぞれの物理サーバが提供できる CPU 資源やメモリ資源は断片化される。VM に提供できる資源が断片化されると，クラウドプロバイダーが持つ資産を効率的に活用できなくなり問題である。この解決方法として VM の効率的な配置方法を考える。既存の検討として、単一の物理サーバ資源を，収容した VM に対して等分(fair)に割り当てる方法については Ongaro[2]らが実装評価も含め研究を行っている．本研究は複数の物理サーバとネットワークに資源の対象を広げている点で異なる．また，本研究分野の製品として vmware 社の DRS(Distributed Resource Scheduler)[3]がある．CPU 及びメモリの使用率についてルールを設定し，VM の再配置を自動化できるがトラフィック交流は考慮することができない．

このとき、CPU 資源やメモリ資源だけに着目し、配置方法を決定していると別の問題が発生する。同じ利用者の VM，即ち通信が発生する VM 同士が異なる物理サーバに収容される可能性がある．アプリケーションによっては，トラフィックが異なる物理サーバ間を往復することになる．また，異なる利用者の VM，即ち異なるトラフィック特性を持った VM が同じ物理サーバに同居する場合がある．こういった状況では最悪の場合，物理サーバの NIC(Network Information Card)の帯域が不足しボトルネックになる場合があり[1]，高品質なクラウドサービスを提供できない．

図 2 で具体例を示す．典型的な 3 層構造の Web システムをクラウド環境で実現している．VM で提供されたロードバランサ(LB)，アプリケーションサーバ(App)と Web サーバ(Web)が異なる物理サーバに収容されている．また LB と同じ物理サーバに他利用者が使う動画配信サーバ(Streaming)が収容されている．

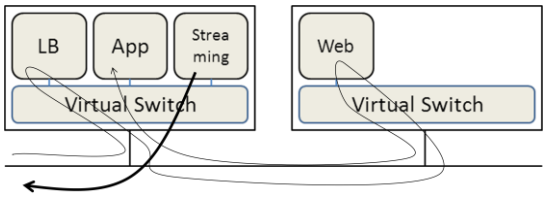

図 2. NIC でボトルネックが発生する可能性がある例

LB がリクエストを受け付け，Web にリクエストを送出する．Web は App にリクエストを要求する．レスポンスは逆を辿る．最悪の場合，リクエストだけで物理サーバにある NIC を 3 回経由する．現在，クラウド環境で使われる物理サーバの NIC は，コスト面から GbE が中心であり実測値の上限は 800～900Mbps である．リクエストが物理サーバ間を往復することで使用できる帯域は減少する．Streaming が

使う帯域を考慮すると利用可能帯域はさらに下がる．NIC の性能を上げて対処することは可能であるが，問題の本質は VM 間トラフィック交流を考慮せず VM の割り当てを決めているところにある．このため，トラフィックを考慮した VM の効率的な配置方法を課題とする．

４．提案するアルゴリズム

VM 間トラフィック交流に着目し，トラフィック量の大きい VM のペアを優先的に同じ物理サーバに割り当てて行くことで課題を解決するアルゴリズムを提案する．

N 個の VM が存在するとする．

b_{ij} = Traffic volume from v_i to v_j，$t_{ij} = \left(max(b_{ij}), v_\mathrm{i}, v_j\right)$

・基本アルゴリズム

入力: トラフィックの集合T = $\{t_{jj}\}$ (i = 1, … , N, j = i, … , N)

出力: アサインされた物理サーバの集合 S_{new}

手順 0: VM 間のトラフィック交流 T について各要素t_{ij}の$max(b_{ij})$の降順でソートしておく．

手順 1: 手順 0 でソートされた順に，t_{jj}各々のトラフィック t_{ij} = (max $(b_{ij}), v_i, v_j$)について a), b) の条件で評価し，2 つの VM, v_i とv_j を新しい物理サーバ集合 S_{new} のいずれかに割り当てる(図 3).

a) v_i が新しい物理サーバに割り当てられていない場合，未使用の物理サーバ $s \in S_{new}$に割り当てる．

b) v_j が新しい物理サーバに割り当てられていない場合，v_i が割り当てられている物理サーバ s_jに割り当てる．不可能な場合は，未使用の物理サーバ $\tilde{s} \in S_{new}$ に割り当てる．

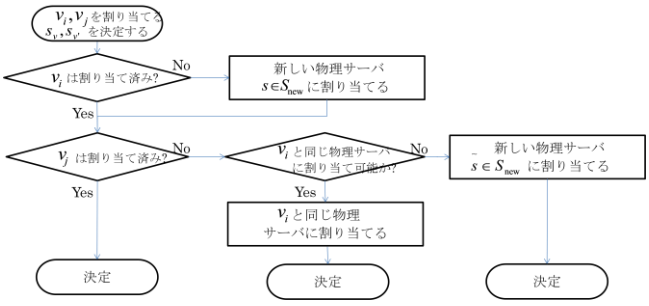

図 3. 基本アルゴリズム手順 1

ここで割り当て可能かどうかについては以下のようなアルゴリズムで判断する．

・割当判断アルゴリズム

与えられた VM が要求する CPU を c_{new}, メモリを m_{new}, 割当たときに物理サーバ外へ出るトラフィックを$\tilde{t}_{new}$とする. また, 割り当て対象となる物理サーバが提供できる CPU 資源をC ，メモリ資源をM，物理 NIC の処理能力を$\tilde{T}$，収容している仮想サーバが各々使用している CPU を c_i，メモリをm_i，物理サーバ外へのトラフィック量を$\tilde{t}_i$とする。

一定の資源(αおよびβ)を残し， c_{new}, m_{new} を C, M から提供できる場合、収容できると判断する．

$$\sum m_i + m_{new} < M - \alpha$$
$$\sum c_i + c_{new} < C - \beta$$
$$\sum \tilde{t}_i + \tilde{t}_{new} < \tilde{T} * \gamma$$

(α , β > 0 の場合は資源留保ありの収容判断，γ はトラフィック量の総和/uplink = oversubscription の係数)

αおよびβは定数や，収容を判断する VM のスケールアップ実績に応じた値を変数としてとる場合が考えられる。

５．実験

M＝2000 台の仮想サーバを仮定し，複数台の物理サーバ(10CPU，32GB メモリ)に割り振る．それぞれの仮想サーバが要求する CPU 量(最大 3CPU)，メモリ量(最大 10GB)をランダムに割り振る。トラフィック 1VM あたり最大 5 交流あるとし，交流一つあたり最大 1Gbps の範囲でランダムに割り振った。結果は，乱数発生の種を変えつつ 10 回行いその平均を示す.

上記の条件で(1)トラフィックを考慮せず VM が要求する CPU とメモリのみを満たせるよう物理サーバに順番に割り当てた場合，(2)提案アルゴリズム(内部留保なし; α = β =0, γ =2.4)，を比較した．結果は表 1 の通り．サーバ外通信は物理サーバの外に流れるトラフィック量の総和である．

	物理サーバ	サーバ外通信/台	サーバ外通信
(1)	483.2 台	7.03Gbps	3089.2Gbps
(2)	1592.7 台	1.54Gbps	2233.4Gbps

表 1. 実験結果の比較

提案アルゴリズムは 2000 台の VM を約 20%集約し, 物理 NIC のボトルネックになる可能性のある物理サーバ外通信は平均 1.54Gbps に抑えられた．本値は各 VM の最大値の合計で計算しており，GbE NIC を装備した物理サーバと仮定すると, 約 1.5:1 程度の oversubscription である. これに対し，従来の方法である(1)の場合，約 7:1 にもなるため，ボトルネックが発生しサービスレスポンスへの悪影響を及ぼす可能性がある．また物理サーバ外に抜けるトラフィック量については(1)の手法と比較し，約 27%削減できた．提案手法では，ネットワーク全体のトラフィックを削減できることも分かった．一方，提案アルゴリズムでは一つの VM しか収容していない物理サーバが 1068.6 台あることも分かった為，一層の物理サーバ台数削減が可能であると考えられる．

７．まとめ

本稿ではサーバ仮想化が利用者にもたらす特徴的な効果を述べ，それにより発生する物理サーバ資源の断片化という問題を挙げた。解決のため，VM の効率的な配置方法を提案した。このとき，VM 間のトラフィック交流を考慮したことが従来と異なる．実験評価を行った結果，CPU 及びメモリ資源のみを考慮したときに比べ，必要な物理サーバは増えるが，NIC がボトルネックになる可能性は少なくなった．また，必要な物理サーバについては，まだ削減可能であると考えられる．

今回は VM 間の片方向トラフィックを評価に使ったが，NIC は通常全二重で使われるため双方向のトラフィックを対象とすれば高効率な配置ができる可能性がある．また，得られた最適配置を実現するためにマイグレーションが必要となるが，VM を入れ替える順番や移動回数も考える必要がある．これらを考慮した，収容アルゴリズムへ拡張を図る予定である．

[1] “すべてわかる仮想化大全 2011”，日経 BP, 2011

[2] Diego Ongaro, Alan L. Cox and Scott Rixner, Scheduling I/O in Virtual Machine Monitors, In Proceedings of the ACM SIGPLAN/SIGOPS International Conference on Virtual Execution environments, 2008

[3] DRS の機能，仮想インフラストラクチャ内の仮想マシンのリソース，vmware 社，available at http://www.vmware.com/jp/products/vi/vc/drs_features.html