大規模 IaaS システムにおける予見可能な通信性能サービス

下國　治　　今井　祐二　　湯原　雅信
株式会社　富士通研究所
{osamus, imai.yuji, yuhara}@jp.fujitsu.com

筆者らはミッションクリティカルシステムを IaaS で動作させることを目的として、予見可能な通信性能サービスの実現方式を検討している。インターネットの実績をふまえ、サービスの実現にはキャパシティプランニングによる解決を図るべきだと考えている。キャパシティプランニングにおける資源増強プロセスでは、ネットワーク技術の進展によって、IaaS システムの帯域は柔軟に増強が可能になった。一方、帯域増強判断プロセスには課題が残っている。非定常時に必要な帯域は定常時に観測する使用帯域からは判断できないため、各 VM が出力可能な最大帯域を積算し、必要な帯域を過剰に予測してしまう課題である。筆者らは IaaS 提供者がサービス対象システムのネットワーク構成と VM の位置情報を把握しているという特徴を利用することにより、過剰な予測を抑制できる例を示した。

A Predictable Network Performance Service for Large Scale IaaS System

Osamu SHIMOKUNI　　Yuji IMAI　　Masanobu YUHARA
FUJITSU LABORATORIES LTD.

With the aim to operate mission-critical systems in IaaS, we have considered an implementation method of a predictable network performance services. Based on practices of the Internet, we should solve by capacity planning to realize the service. On the resource reinforcement process in capacity planning, bandwidth of IaaS system has enhanced flexibility by progress of network technology. On the other hand, the challenge remains in decision process to increase bandwidth. Because the bandwidth required for the failure state can't be determined from observing the typical state, it is a problem to accumulate the maximum possible output bandwidth of each VMs, thereby predicts the necessary bandwidth in excess. By the information of the VM location and the network configuration of the tenant system, we show an example which can suppress excessive prediction.

1. はじめに

1.1. IaaS 内での通信性能保証サービス

ある一定時間内に一連の処理が常に終了することが期待されるミッションクリティカルシステムを IaaS（Infrastructure as a Service）に構築することは困難であった。その理由は、リソースが複数の利用者に共有されているため、システム性能設計に必要な基礎データである CPU やストレージ、ネットワークの性能が IaaS では大幅

に変動するからである。例えば Wang[1]らは、Amazon EC2 内の通信で RTT の平均値が 100μ秒~1m 秒に分布しているのに対し、最悪値は 5m 秒~100m 秒に分布していることを報告している。

しかし、IaaS が提供する短いリードタイムでの資源確保可能性や柔軟なシステム拡張縮小性、ハードウェアの保守管理業務代行といった利点をミッションクリティカルシステムへ適用可能とするために、IaaS システムで CPU キャッピングやストレージの性能を保証するサービスが提供されはじめている。例えばストレージでは、Amazon EBS はプロビジョンド IOPS ボリュームサービスを 2012 年 8 月に開始している[2]。

ネットワークも同様に、通信性能の保証サービスを実現しようとする研究が行われている[3,4,5,6]。これらの研究は通信帯域確保のためのプロトコル提案や、ネットワークリンクの帯域を管理して通信経路の最適配置を探索するもので、Intserv や RSVP 等の、従来のインターネット QoS 技術研究の延長に位置づけることができる。

これらインターネット QoS 技術は 20 年以上の歴史と、多くの優れた論文やデモの実績にも関わらず、現実のインターネットでは普及しなかった。しかし、QoS 技術が必須と言われたアプリケーション、例えば動画像配信、音声通話、遠隔 TV 会議は、QoS 技術なしだがそれなりに実用的な品質で、現在多くの人々が低コストで利用している。これらのアプリケーションがインターネットで実用になっているのは、充分な通信資源が用意されているからである。インターネットサービスプロバイダやインターネットエクスチェンジなどがトラフィックを日々観測し、需要予測を基に回線と設備を増強するキャパシティプランニングという作業を適時実施して、通信品質を確保している。

インターネットでの経緯を踏まえると、IaaS システムで通信性能を保証するために必要なものはキャパシティプランニングという工学的プロセスだと我々は考える。

従来のインターネット QoS 研究では、与えられた通信帯域の使用を細かく制御することで通信性能を保証するアプローチが取られたが、インターネットでは管理組織数や、接続ノード数に対するスケーラビリティの課題があり、実際には運用が難しかった[7]。

IaaS システムのデータセンタ（以下 DC）内ネットワークもまた、数万テナント、数十万 VM が接続する巨大スケールのネットワークであり、使用帯域を細かく制御するアプローチはインターネットと同様のスケーラビリティに対する問題が発生すると予想できる。しかし、IaaS システムの DC 内ネットワークは予め設計しておけば増速のコストはインターネットに比較してはるかに低い。適時増速を行なって供給帯域を確保し、ネットワークトポロジを考慮せずに帯域量を管理するアプローチを取ることにより、スケーラビリティを上げることができると考えている。

本稿では IaaS を構成する DC ネットワークにおける予見可能な通信性能サービスについて述べる。このサービスは厳密な意味での帯域保証をするのではなく、帯域の充分な余裕と統計多重によってある程度の幅で性能を保証しようという意味で、「予見可能な」通信性能サービスと呼ぶことにする。今後本稿での「帯域保証」はこの緩い意味での帯域保証とする。

1.2. 通信性能保証サービスと帯域

ミッションクリティカルシステムを IaaS で動作させるための通信への要件として、通信帯域と共に、通信遅延、パケットロスが一定の範囲内に常に収まることがあげられる。

Cisco の報告書[8]で示されているように、充分な帯域があれば通信遅延はある一定値以下に抑えられることが知られている。また、DC 内では輻輳が発生しなければパケットロスも機器故障以外には定常的には発生しない。

これらのことから、通信帯域の保証を行うことによって、通信遅延、パケットロスの要件も満たすことが可能であると我々は考えている。今後この論文では、通信性能保証サービスは通信帯域保証サービスを意味することとする。

1.3. 課題

キャパシティプランニングは一般的に図 1 のようなプロセスで構成される。キャパシティプランニングのプロセスに従って具体的に帯域保証を行うには次のような項目

が課題となる。

1. 資源増強計画と資源増強のプロセスを可能とする、帯域を増設できるネットワーク構造
2. 利用状況観測と需要予測のプロセスにおいて、帯域保証サービスとして必要な帯域の考え方と計算方法

本稿では、1 に対し 3.4 節で現状技術を報告する。2 に対しては、4.1 節で必要な帯域の考え方について議論し、4.3 節で必要な帯域を、より低く予測できる例を示す。

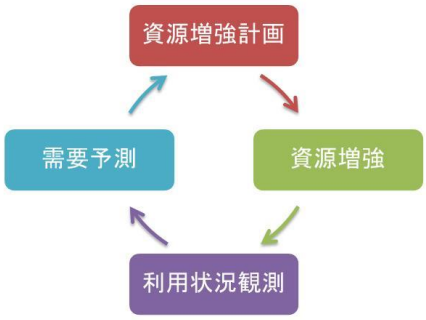

図 1. キャパシティプランニングプロセス

2. 関連研究

Oktopus[5]は仮想的な利用者システムに、VOC と呼ぶ仮想スイッチを導入し、仮想的なリンク帯域使用量に従って仮想スイッチをツリー状の物理スイッチに配置する。

我々はトポロジ情報を利用者から隠蔽することを指向しているが、Oktopus では IaaS の実際のシステムが 2 段のスイッチで構成されている時、利用者システムも 2 段の仮想スイッチを設定して、物理スイッチへの配置を探索する。モデルが詳細なので、1 つの VM の追加や移動によって複数のリンクにおいて必要な帯域の再計算が行われ、影響の発生した VM の再配置が発生する。配置探索にはグリーディ算法を用い、準最適な配置を決定する。

GateKeeper[6]は仮想的な一つのスイッチングハブに VM を接続し、ポート毎に帯域を保証するモデルを取っている。また、通信帯域が空いている場合には、契約帯域以上の通信も可能である。

受信帯域を超えると、送信側 VM の送信レートを下げるよう受信側 VM が接続しているインタフェースがフィードバック信号を送信する。制御は VM の端点のみで行い、コア網や ToR での帯域は充分にあると仮定している。

Seawall[4]は IaaS での帯域飽和型の DoS 防御という明確な目的をもって提案されている。

VM 間で利用者の仮想網を作るトンネルにフロー制御機構を入れ、（VM ではなく）物理マシンが過剰なトラフィックを受信した際に、各テナントのトンネルに帯域を分配する。この論文でも制御は VM の端点のみで行い、コア網や ToR の帯域は充分あることを仮定している。

3. 対象とする IaaS システム

3.1. IaaS システムのスケール

定量的な議論を進めるため、2020 年前後まで稼働する一つの DC 内の IaaS システムを前提として、スケール目標を以下のように仮定した。

表 1 スケール目標

	項目	量	備考
サーバ	VM	500,000	目標として仮定
	物理マシン	10,000	50VM / 物理マシン
ネットワーク	利用者システム	50,000	10VM / 利用者システム
	仮想 L2 セグメント	200,000	4 セグメント / 利用者システム
設備	ラック	350	物理マシン 32 台 / ラック その他ネットワーク、ストレージ等 40 ラック

数値は厳密ではなく、オーダを議論するためのものである。ここでは VM のサービスクラスとして最小クラスを仮定し、1 台の物理マシンに 50VM 搭載することとした。実際には、サービスクラスによって一つの VM は CPU を数分の 1 コア~数コア使用するので、これらの値は数倍の幅を持つ。

3.2.　利用者のネットワーク帯域のモデル

サービス利用者に提示するネットワーク帯域のモデルとして、我々は GateKeeper と同様に、利用者のひとつの L2 ネットワークセグメントを仮想的な一つのスイッチングハブとして扱うモデルを採用する（図 2)。この仮想的なスイッチングハブの各ポートに VM やルータを接続する。帯域のモデルとして、ハブの内部は輻輳が起きない、充分な帯域があると仮定する。

その一方、全ポートが同一の 1Gbps や 100Mbps の速度指定がされており、送信と同様に受信も、通信は各ポートで速度制限される。IaaS 提供者はこのモデルの基で可能な通信帯域を利用者に保証する。

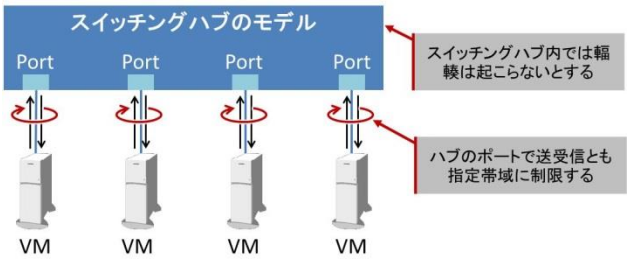

図 2 利用者のネットワーク帯域のモデル

3.3.　帯域保証サービスとベストエフォートとの混在

我々は、帯域保証サービス利用者のシステムと、ベストエフォートの帯域を利用する利用者のシステムを一つの IaaS ネットワークに混在させることを検討している。その理由は、帯域保証サービス利用者が予約したが使用していない帯域を、ベストエフォートサービス利用者が利用することにより、帯域の有効活用を図ることが可能だからである。

混在させた 2 つのクラスのトラフィックを分離し、操作するために、QoS 技術として確立している Diffserv を使用する。クラス分けは、帯域保証サービスとベストエフォートの 2 クラスだけとし、パケットスケジューリングでは厳格な優先制御と Expedited Forwarding を採用する。

実現方法は以下のとおりである。テナント分離方式として各テナントの L2 フレームを IP トンネリングする nvo3[9]を採用する。各 VM が接続しているトンネルの端点において、nvo3 トンネルの Outer Ethernet ヘッダに対しては IEEE802.1Q を、IP ヘッダに対しては DSCP を指定する（図 3)。

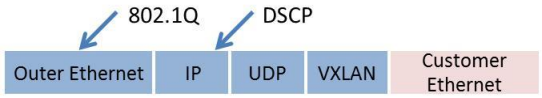

図 3 Diffserv の設定場所例: VXLAN

3.4.　帯域を増減可能なアンダーレイネットワーク

対象としている IaaS システムは数千~数万物理マシンの巨大システムなので、一つのシステムの設計からサービス終了までは 5 年から 10 年の長期にわたる。しかし、設計時に 5 年後や 10 年後に必要となるマシン数や帯域は予測不能である。また予測できたとしても、システム構築当初は充分な量の利用者がいないので、投資は非効率であるし、技術進歩に伴った機器のコストダウンを考慮すると、IaaS システムは小規模な状態から提供をはじめ、適時柔軟に拡張できることが必要となる。

IaaS システムの物理的なネットワークは、VM のマイグレーションを目的として、2010 年前後まではフラットな Ethernet のツリー構造を取っていた。ところが巨大なツリー構造ではルート部分のトラフィックがボトルネックとなることから、Leaf & Spine と呼ばれる、Folded Clos ネットワークが提案された[10,11]。この構成はラック間のコアスイッチを適時追加することによって、コアスイッチで構成されるコア網の帯域を増加することが可能になっている（図 4)。インターネットなど外部との接続は、ラックの ToR の代わりにルータを設置し、コア網に接続する。

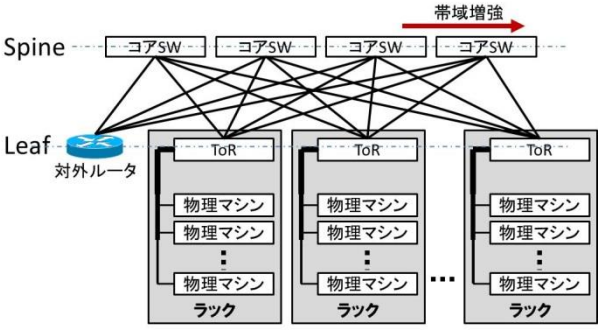

図 4. Leaf & Spine 構成

Leaf となる ToR（Top of Rack スイッチ）では、物理マシンからの総帯域よりも、コア網への総帯域が小さくなり、輻輳が発生することを許容して設計している。これはオーバサブスクリプションと呼ばれている。一方 Spine となるコアスイッチでは、多段化した場合でも前後の段の間で帯域の増減は無いように設計し、ToR から入力したパケットは原則としてロス無しにコア網を通過するようにする。

図 5 は一般的に用いられている構成例で、10Gbps のリンクを持つ物理マシンを 32 台 ToR に接続しているが、コア網へは 4 つのスイッチに 10Gbps のリンクを 2 本ずつ接続しており、オーバサブスクリプション率は 4:1 となる。

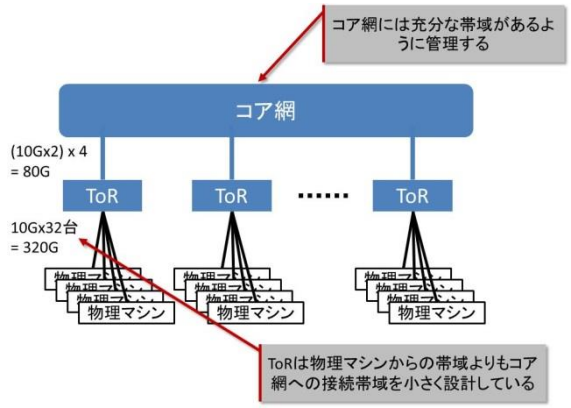

図 5. Leaf & Spine での帯域モデル

このモデルは 3.2 節で述べた利用者のネットワーク帯域のモデルと、コア網とスイッチングハブは共に帯域が充分にあり、ToR とポートは共に帯域制限があるという、共通の特徴をもっている。

帯域保証サービスを実現する場合、コア網へ利用者モデルの各スイッチングハブを重畳して配置すると考える（図 6)。利用者へ提示したモデルにおいて、仮想的なスイッチングハブ内で輻輳が発生しないようにするためには、コア網の帯域量に関するキャパシティプランニングを実施することが必要となる。

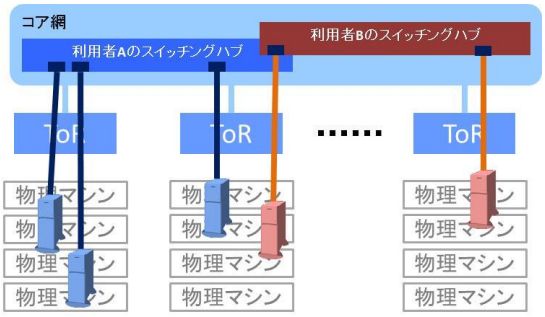

図 6. Leaf & Spine の帯域モデルに、利用者システムを重畳した模式図

一方、利用者モデルと Leaf & Spine モデルでは、帯域制限の位置では相違点が発生する。利用者モデルのポートの帯域制限は実装では VM の仮想インタフェースでかかる一方、Leaf & Spine モデルは ToR で帯域制限がかかる。この相違点のため、VM の仮想インタフェースから ToR のコア網へのアップリンク間では輻輳しないように、VM 配置の可否を判断するアドミッション制御を実施する必要がある。

大規模な DC 内ネットワークの実装方式として、Lapukhov[12,13]らは Leaf & Spine 構成を IP ネットワークとして構築し、BGP を使用して経路制御を実現することを提案している。BGP を使う利点の一つとして ECMP（Equal Cost Multi Path）によって、複数のラック間接続ネットワークの間で負荷分散が可能になることがある。ECMP と、多数の利用者トラフィックによる大数の法則により、コア網の帯域が 4.1 節で説明する単純化した通信量のモデルで近似して扱えるようになる。

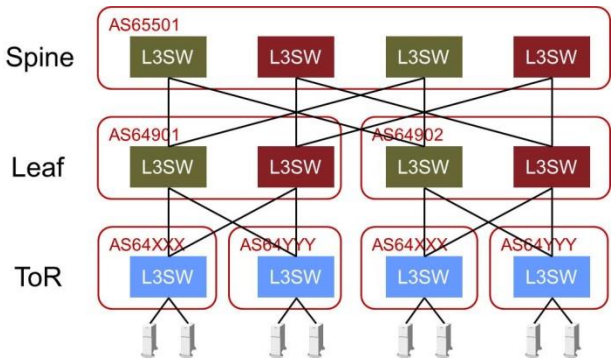

図 7　BGP を利用した 3 段 Leaf & Spine 構成

4. 必要な帯域の計算

4.1. 帯域保証サービスに用意すべき帯域

3.4 節ではキャパシティプランニングに必要な、柔軟な帯域増強が DC 内ネットワークで可能になったことを述べた。この節では帯域増強判断をするために、用意すべき帯域について述べる。

ネットワークが輻輳状態にならないことを厳密に保証するには、ネットワークに接続している全マシンが同時に最大帯域で通信しても充分な帯域を用意しなければならない（図 8 中 a）。しかし実際には、全マシンが同時に最大帯域で通信することは極めて稀で、しかもその帯域を用意するには巨大なコストがかかるので、資源削減の面からより少ない資源量を用意することが一般に行われている。用意する帯域値（図中 b）は通常状態で使用している帯域（図中 c）を観測し、今後のピークやトレンドを予測して決定する。

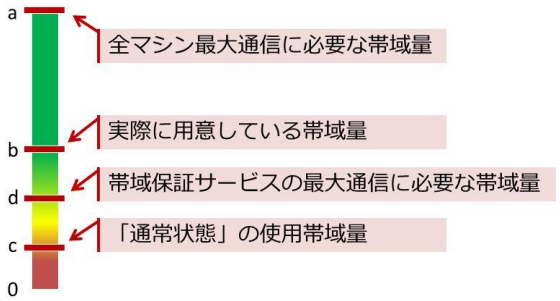

図 8　必要な帯域量の模式図

帯域保証サービスを提供する IaaS システムでは、IaaS 提供者によって示された利用者のネットワーク帯域のモデルに基づいて、帯域保証サービスの全 VM が出力可能な全帯域分は確保しておく必要がある（図中 d）。この値に経験から得られる安全係数を掛け、ベストエフォートの帯域分を加えて、実際に用意するべき帯域（図中 b）を決定する。

一方で、ミッションクリティカルシステムとはいえ、VM がすべて同時に最大出力をする確率はやはり低いので、定常的に流量計測して統計的に安全な帯域を用意するという考え方もある。

しかし、これは採用できないと我々は考えている。その理由は、過去の障害事例[14]にあるように、異常時に利用者システムが復旧のために大量の通信を実施する可能性があり、定常時のトラフィックだけを基にした統計からの予測では帯域保証ができないからである。そして、利用者にとって帯域保証はこのような異常時こそ必要と我々は考えているからである。

ただし、「利用者のネットワーク帯域のモデルに基いて、出力可能な全帯域」d の予測を、正確により低く見積もることが可能であれば、より低い帯域量を用意するだけですみ、より安いコストでサービスを提供できる。

以降の節では、d を低く見積もるための一方式を紹介する。

4.2. 単純な最大通信帯域見積もり方式

4.1 節で述べたように、IaaS 提供者は利用者の VM の動作に対する情報が無いことから、帯域保証サービスを適用している VM すべてのトラフィックがコア網へ送出されることを仮定しなければならない（今後この計算方式を方式 α と呼ぶ）。

IaaS システムには複数の利用者システムがあり、利用者システムはそれぞれ複数の仮想 L2 ネットワークセグメントを持っているとする。ある利用者の仮想 L2 ネットワークセグメントに接続する VM の個数を n, その仮想網への保証帯域を w とするとき、帯域保証サービスの最大通信の帯域は以下の式で示される。

$$\sum_{\text{各利用者}} \sum_{\text{各仮想網}} n \cdot w$$

また、3.4 節でのべたように、物理マシンおよび ToR はアドミッション制御が必要となる。配下に存在する帯域保証サービスの VM の総最大通信用量が実回線の帯域を
を超えると、サービスを受理しないようにする。

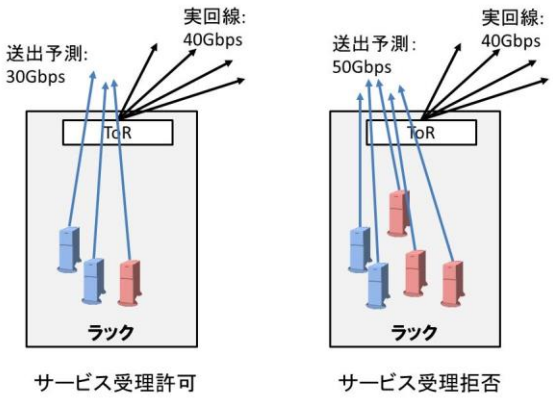

図 9. ToR におけるアドミッション制御

利用者の仮想 L2 ネットワークセグメントへの保証帯域を w, その仮想 L2 ネットワークセグメントの該当ラックの搭載数を k とするとき、アドミッション制御における該当ラックの総帯域は

$$\sum_{\text{各利用者}} \sum_{\text{各仮想網}} k \cdot w$$

となる。

4.3.　VM の配置と TCP 通信を考慮した計算

前節では、IaaS 提供者が VM の動作に関する情報を一切所有していないとして、計算を実施した。しかし、他の情報を用いて、より少なく必要帯域を見積もることが可能である。

IaaS 提供者は、利用者システムの仮想ネットワークとそれに接続している VM、VM の物理マシン配置場所を把握している。また、一般のシステムでは大部分の通信が TCP であることが知られており、Murray[15]らは、企業ネットワークにおいて、帯域比で 92%が TCP であると報告している。

TCP の動作と、3.2 節で提示した帯域通信モデルでの VM 受信帯域の制限により、VM のコア網への出力帯域を方式 α より低く見積もる例（今後この計算方式を方式 β と呼ぶ）を以下に示す。

10 個の VM が、それぞれ 1Gbps の帯域保証ポートを持つ利用者仮想ネットワークに接続しており、3 つのラック A, B, C にそれぞれ、VM が 6 個、3 個、1 個配置されているとする。今、ラック A に注目する。方式 α では 6×1Gbps＝6Gbps 分の通信を ToR からコア網へ出す可能性があると予測する。しかし、ラック B, C には合計 4VM しか存在しておらず、最大受信帯域は 4Gbps である。TCP を考慮すると、たとえ一時的にコア網へ 4Gbps を超えるパケットが流れたとしても、ラック B, C の VM の受信ポートでロストして、送信側は TCP の輻輳回避アルゴリズムによって帯域を制限する。したがって、ラック A からの送出予測帯域は 6Gbps ではなく、4Gbps と見積もれる。

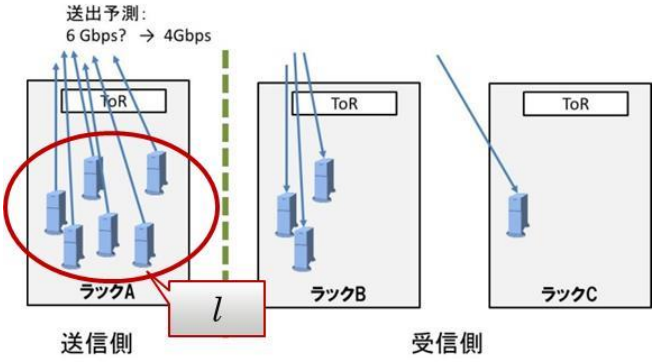

図 10. 受信性能を考慮した送出予測

例を定式化すると、前節と同様に利用者仮想ネットワークに接続する VM の個数を n, その仮想網への保証帯域を w, その仮想網の該当ラックの搭載数を k とするとき、総帯域は以下の式で示される。

$$\sum_{\text{各利用者}} \sum_{\text{各仮想網}} \sum_{\text{各ラック}} \min(k, n-k) \cdot w$$

ここで、$1 \le k \le n$から、min を場合分けして、

$$\min(k, n-k) \cdot w = \begin{cases} (n-k) \cdot w, & if\ k \ge \dfrac{n}{2} \\ k \cdot w, & otherwise \end{cases}$$

さらに、$k \geqq n/2$ となるラックは高々一つであることがわかる。これは、一つのラックが条件を満たすと、他のすべての VM 総数は $n/2$ 未満となるので、他のラックは条件を満たさないからである。そこで仮想網ごとの VM における一つの最大ラック搭載数を l とすると、仮想網ごとの送出予測帯域は

$$\sum_{\text{各ラック}} \min(k, n-k)\cdot \mathrm{w}$$

$$= (n-l)\cdot w + \begin{cases} (n-\mathrm{l})\cdot w, & if\ l > \frac{n}{2} \\ l\cdot w, & otherwise \end{cases}$$

（最大ラック以外　＋　最大ラック部分）

$$= \begin{cases} (n-l)\cdot 2w, & if\ l > \frac{n}{2} \\ n\cdot w, & otherwise \end{cases}$$

となる。

$k \geqq n/2$ となるラックは高々一つであることから、VM 数の増加や移動に伴う影響は、移動などが発生した該当のラックと、VM 最大搭載ラックに限られる。よって、仮想ネットワーク毎に VM 最大搭載ラックの情報を記憶しておけば、この値は差分を計算するだけで更新が可能になる。

この計算方式はすべての通信が TCP であると仮定していた。UDP や他の輻輳制御をしないプロトコルの場合は、求めた値より多くの帯域をコア網に送出してしまう。ただし、出力帯域はポートの保証帯域で制限されていること、一般の企業システムでは非 TCP 通信は帯域比で 8%以下なので無視出来るとした。

4.4.　方式 β の効果の確認

方式 β の計算式によれば一つの利用者システムを一つのラックに封じ込めるとコア網への送出通信量をゼロにできた。これは実際の状況と合致し、VM 配置設計制御[16]によってコア網への送出通信量の削減を図ることが可能である。

しかし、実際の IaaS システムでは、ライセンスやサポートの制約条件から、特定の種類の VM は特定のラックに収容する場合があり、一つのラックに封じ込められない場合がある。

ここで、方式 β で予測を低く見積もれる効果を調査するため、ある大規模 IaaS からサンプリングしたデータを用いて、4.3 節における l の平均値および標準偏差を求めた（図 11）。

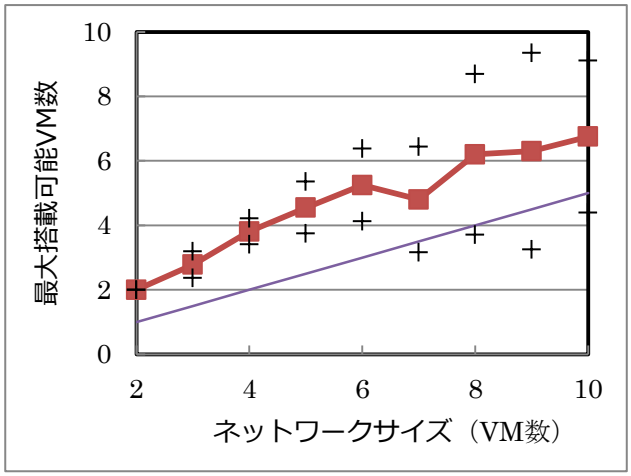

図 11. 同一ラックへの最大搭載可能 VM 数平均

解析できた有効なサンプル数は VM 数 774、L2 ネットワークセグメント数は 165 であった。VM は OS が 3 種類とライセンス付き DB の計 4 種類に分別し、それぞれを別のラックに搭載することと仮定した。ここで、一つの L2 ネットワークに接続している VM 数をその L2 ネットワークの「ネットワークサイズ」と呼ぶとすると、その個数比率は図 12 のようであった。

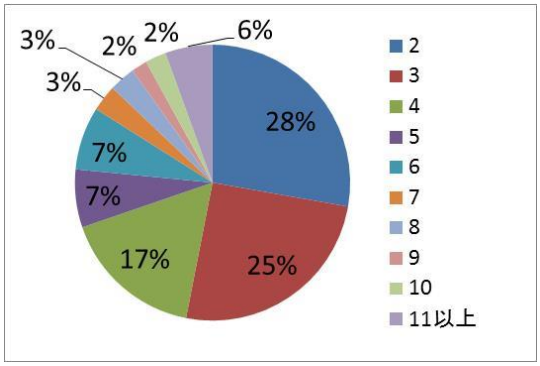

図 12. ネットワークサイズ比率

なお、この IaaS システムでは利用者システムにはもれなくソフトウェアルータが付いているので、最小ネットワークサイズは 2 である。ネットワークサイズが 10VM 以上はサンプルが少なかったので平均、分散や以下の評価では除外した。

図 11 において、各ネットワークにおける同一種 VM の最大数の平均値はそれぞれ全 VM 数の 1/2（図中マーカー無し直線）以上となっており、この計算法の適用による効

果が現れることを示している。図中のマーカ「+」は標準偏差（σ）であり、サイズが大きくなるにつればらつきが大きくなることが分かる。これは、サイズが大きなシステムになると、それぞれの役割毎に VM を割り振り、異なった OS やライセンスが必要なアプリケーションを使用する傾向が強くなるからである。またばらつきが小さい、接続 VM 数が 6 までのネットワークは、図 12 の比率を見ると 80%以上となっており効果が出やすい状況であることがわかる。

また、この計算法の効果を確認するため、サンプリングした利用者システムにそれぞれ 1Gbps の帯域保証サービスを適用するとして、方式α、方式βでコア網の最大通信帯域を計算した。

表 2　コア網の最大通信帯域

	方式α	方式β
総帯域'(Gbps)	590	134

サンプリングしたシステムでは、方式βを適用することにより、コア網への最大予想帯域が 134/590＝約 1/4 に削減できるという結果になった。

5. 議論

方式βによって、4.1 節で述べた、IaaS 提供者によって提示された帯域モデルの元で出力可能な全帯域の予測を、VM の配置情報と TCP 通信の特性を利用して、サンプリングしたデータでは 1/4 程度に小さくできることが判明した。

ここで、アドミッション制御から可能なサービス収容率を考える。3.1 節で述べた 1 ラックに、それぞれ 10Gbps の NIC を持つ 32 台の物理マシンが収容され、各 50VM が搭載されているとする。ToR はコア網へ 40Gbps のアップリンクが 2 系統、計 80Gbps 用意されているとする。このとき、ToR のオーバサブスクリプション率は 1:4 であり、IaaS システムとしては標準的な構成である。

また、Cisco の報告書[8]にある通り、帯域使用率 80%以下ならば、遅延、パケットロスが低く押さえられるとする。すると、帯域保証サービスとしては 1 ラック当たり 64Gbps 分の収容が可能と考える。

仮にすべての帯域保証サービス対象システムが 1Gbps の契約をしている場合、方式αで計算したアドミッション制御を行うと、このラックには帯域保証サービス対象 VM は 64 台が収容可能である。これはラック内 1600VM に対して 64/1600=4%となる。残りの 96%のベストエフォート利用者VMとサービス対象VMのネットワーク資源の使用比率を最大出力時で計算すると、この 4%の利用者が 16Gbps 分の帯域を使用するので、16/96:64/4 = 1:96 となり、ベストエフォート利用者の 96 倍の資源を使用することとなる。

定常的には、余剰帯域をベストエフォート利用者が使用可能であり、96 倍の課金をするわけではないが、この使用比率は、インターネット QoS が普及しなかった原因となっていた、ベストエフォートサービスに比較して同一帯域で 10 倍以上の高い課金を実施せねばならない一因である。

いっぽう方式βで計算し、4.4 節の例のようにコア網への帯域出力を 1/4 に小さくできるとすると、ラック内の 16%、すなわち 64 台の VM が帯域保証サービスの対象として収容可能になる。このとき、資源使用比は 16/84:64/16=1:21 となる。これも 21 倍の課金を実施するわけではないが、利用者にとって納得できる価格付けをするには、非定常時のモデルを更に深く考える必要があると思える。

今後の研究として、TCP 通信の特性以外にも、利用者への帯域保証のモデルや IaaS システムの条件、保証条件の緩和によって、さらに必要帯域を低く見積もることを検討する予定である。

IC2013 では、利用者にとって意味のある保証のモデルや現実的な保証条件について議論したい。

6. まとめ

本稿では、IaaS における通信性能保証サービスの実現に向け、通信帯域量を充分に確保することで、品質とシス

テムのスケーラビリティを確保することを考察した。

通信帯域を充分に管理するためにキャパシティプランニングのサイクルを示し、柔軟に通信帯域増強を可能とするネットワークとして Leaf & Spine モデルを示し、ネットワークトポロジの考慮なしに単純化したモデルで帯域を扱うことを紹介した。

一方、必要な帯域を予測するには、通信性能保証サービスの性質を考えると、通常状態の帯域観測ではなく、利用者に提示したモデルに基づく、利用者の出力可能な最大帯域を仮定しなければならないことを考察した。

この帯域を単純に計算すると、帯域保証サービスには大きな帯域を用意しなければならない。この予測を低く見積もる方法として、VM の配置と TCP 通信の特性を利用した方式を一例として示した。

我々は IC2013 での議論を元に、利用者にとって意味のあるサービスを保ちつつ、必要な帯域をさらに正確に低く求める方式を今後検討していく予定である。

参考文献

[1] G. Wang, T. Ng: The Impact of Virtualization on Network Performance of Amazon EC2 Data Center, In *Proc. of INFOCOM'10*, pp. 1163-1171.

[2] Amazon EBS Provisioned IOPS service: http://aws.amazon.com/jp/ebs/

[3] J.C. Mogul, L. Popa: What We Talk About When We Talk About Cloud Network Performance, *ACM Sigcomm CCReview*, Vol.42, No. 5, 2012 Oct.

[4] A. Shieh, et al.: performance isolation for cloud datacenter networks. In *Proc. HotCloud, 2010*.

[5] H. Ballani et al.: Towards predictable datacenter networks, In *Proc. SIGCOMM*, pages 242–253, 2011.

[6] H. Rodrigues et al.: Gatekeeper: supporting bandwidth guarantees for multi-tenant datacenter networks. In *Proc. WIOV, 2011*.

[7] B. Teitelbaum, S. Shalunov: Why Premium IP Service Has Not Deployed (and Probably Never Will), *Internet2 QoS Working Group Informational Document*, May 3, 2002

[8] Cisco: Best Practices in Core Network Capacity Planning, http://www.cisco.com/en/US/solutions/collateral/ns341/ns973/ns1225/white_paper_c11-728551.html

[9] Network Virtualization Overlays (nvo3): internet draft, http://datatracker.ietf.org/wg/nvo3/

[10] A. Greenberg, et al. VL2: A Scalable and Flexible Data Center Network, *CACM*, Vol. 54, NO. 3, March 2011.

[11] M. Al-Fares, et al.: A Scalable, Commodity Data Center Network Architecture. In *Proc. of SIGCOMM,* 2008.

[12] P. Lapukhov: Internet Draft: draft-lapukhov-bgp-routing-large-dc-05

[13] P. Lapukhov: Routing Design for Large Scale Data Centers: BGP is a better IGP!, *Nanog 55.*

[14] Amazon: Summary of the Amazon EC2 and Amazon RDS Service Disruption in the US East Region, http://aws.amazon.com/message/65648/

[15] D. Murray, T. Koziniec: The State of Enterprise Network Traffic in 2012, *18th Asia-Pacific Conference on Communications (APCC 2012).*

[16] 山島 他: 大規模データセンターにおける最適な VM 配置設計手法, *信学技報, vol. 111, no. 163, CPSY2011-19*, pp. 61-66, 2011 年 7 月.