大規模データセンターのネットワーク仮想化

Network Virtualization for Large-Scale Data Centers

● 安藤達宏　● 下國　治　● 麻野克仁

あらまし

昨今，企業の大規模災害に向けた事業継続性(BCP)の観点からクラウドへの移行が進み，その結果データセンター事業者のシステムの大規模化が進んでいる。クラウドサービスにおいて，IaaS(Infrastructure as a Service)を例に取ると，大規模化が進むにつれ，既存技術での仮想ネットワークや物理ネットワークのネットワーク構築が困難となる。データセンターにおける仮想ネットワークは一般的にVLAN(Virtual LAN)により構築されるが，VLAN IDのビット長から4094までしか構築できないため，この制限の拡張が課題となっている。また物理ネットワークではSTP(Spanning Tree Protocol)によるネットワーク回線利用の非効率性やネットワーク設計の複雑性が課題となっている。

本稿では，大規模データセンターの条件，技術的な課題を整理する。また上記の仮想ネットワーク数の拡張，STPに代わる物理ネットワークの構築など，様々な課題を解決するための技術や各技術の比較や特徴について述べる。

Abstract

Enterprises are turning to cloud services as a business continuity plan (BCP) for large disasters, and as a result the systems used by operators of cloud data centers (DCs) are becoming larger these days. In the case of infrastructure as a service (IaaS), as cloud DCs become much larger so it becomes more difficult to configure virtual and physical networks with the existing technologies. Virtual networks are typically configured by VLANs in DCs; however, no more than 4,094 VLANs can be constructed because of the bit length limit of VLAN ID. Therefore, there is a need to have a technology to extend this limit. Moreover, with physical networks, the inefficiency of network interfaces and the complexity of network construction by STP have been problems. This paper describes the requirements that a large-scale DC must meet and the technical problems to overcome. Moreover, it gives a comparison of and the features of technology for solving the above problems, such as technology to extend the above-mentioned number of virtual networks and construct physical networks to replace STP. This paper also covers the future trends that we expect to see in DCs.

ま え が き

昨今，大規模災害に向けた事業継続性（BCP）の観点から，データ保管やWebサービスといった企業の情報処理やサービスの高可用化が求められている。情報システムの高可用化には高度なシステム設計と運用が必要なことからクラウドへの移行が進み，データセンター事業者のシステムの大規模化が進んでいる。

クラウドサービスは提供するサービスによりいくつかに分類され，サービス種別にはSaaS（Software as a Service）やPaaS（Platform as a Service），IaaS（Infrastructure as a Service）などが存在する。

特にIaaSでは，多数のテナント（ユーザ）が同一の物理システムを共同で利用するため，一つのシステムが大規模化する傾向にある。近年のサーバ性能の向上により，1台の物理サーバ上で稼働する仮想マシン（VM）数は20以上に増加している。ネットワークに関しても，サーバのネットワークカード（NIC）は10 G Ethernetが今後の主流となるため，更なる大容量化や高集積化が必要となることが予想される。

本稿では，データセンターが大規模化した場合のネットワークにおける課題，ネットワーク仮想化の大規模化方式，物理ネットワークの構築方法について述べる。

大規模データセンターにおける課題と要件

現状のパブリックIaaSは，ホストサーバが数百台，VMが数千台の規模で運用されている。[(1)] 本稿で今後の大規模データセンターについての課題を議論するために，2015年～ 2020年の大規模データセンターの規模を**表-1**のように仮定する。

現在のデータセンターが表-1のように大規模化した場合，以下の技術的課題や要件が挙げられる。

(1) クラウドOSと連携したサーバ・ストレージ・ネットワークの統合管理
(2) ライブマイグレーションと連動したネットワーク設定変更の自動化
(3) 4094を超えるテナント分離（仮想ネットワーク）
(4) トラフィック集中回避
(5) レイヤ2網のデータセンター間への遠隔延伸

更にビジネス的な観点から見ると，以下の課題が挙げられる。

(6) ベンダロックインの回避（キャリアは事業継続の観点から機器の調達を1社のベンダに依存することを避け，複数のベンダがサポートする方式を採用する）
(7) スモールスタート，スケーラビリティ

上記の課題を解決し，大規模データセンターでIaaSを構築することで，よりエンドユーザの利便性向上，顧客のOPEX（Operating Expense）/CAPEX（Capital Expenditure）削減が見込まれる。本稿では，上記の技術的課題（1）～（5）を解決する技術および方法について述べる。

まず次章では，課題（1），（2）に関して，サーバ・ストレージ・ネットワークを一元的に管理するIaaS基盤の構築について述べる。

IaaS基盤におけるシステム構築

データセンターはクラウド化の動きが活発であり，アーキテクチャの変化が最も激しい部分であ

表-1　2015年～ 2020年の大規模データセンター諸元

大項目	中項目	基準値
サーバ	VM数	20 000以上
	物理サーバ台数	1024以上
ストレージ	VMあたりのストレージサイズ	75 Gバイト以上
	総ストレージサイズ	500 Tバイト以上
	ストレージ台数	2台×2以上
ネットワーク	テナント数	2000以上
	仮想ネットワーク数（VLAN/VXLANなど）	8000以上
	インターネットとの接続箇所	2か所以上
ファシリティ	データセンター拠点数	2か所以上

る。従来，サーバ・ストレージとネットワークは別々の部門が個々に管理していた。しかし，サーバ内のVM間で通信するための仮想スイッチはホストOS内に実装されるため，サーバ上のVMと仮想スイッチを一括管理した方が利便性が高い。また，サーバ上のVMを動的に別のサーバへ移動させるライブマイグレーションも頻繁に行われるため，VMの移動に合わせ，ネットワークも設定変更を自動的に対応させる必要がある。そのため，サーバ・ストレージ・ネットワークを統合的に管理するクラウド運用管理ソフト（クラウドOS）の普及が進むと思われる。

特にOpenStackやCloudStackといったオープンソースのクラウドOSの普及が進み，複数の仮想化基盤ソフトウェアやハードウェアがサポートされるようになってきた。この結果，従来ベンダロックインされた（同一ベンダ機器で構築されていた）システムが，様々なベンダ機器で構築できるようになってきている。

クラウドシステムのアーキテクチャを**図-1**に示す。

次章では，課題（3），（5）に関して，既存技術の課題や課題を解決する技術について述べる。

大規模データセンターを構成するネットワーク仮想化

IaaS基盤では，**図-2**に示すように，サーバ仮想化同様，ネットワークにおいてもテナントごと（サービス利用部門および会社）のネットワークを独立して構築し，同一の物理ネットワークを使用可能とする技術，すなわちネットワーク仮想化が必要である。

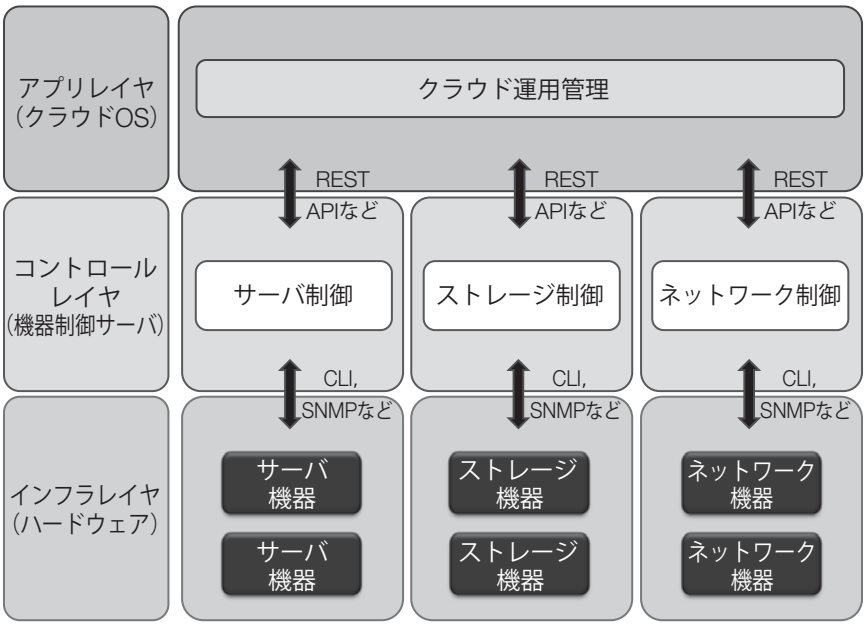
現在でも，ネットワークの仮想化はVLAN（Virtual LAN）により可能だが，分離できる仮想ネットワーク数（ID数）に4094という制限があり，大規模データセンターには対応できなくなってきている。各テナントのシステム構成に依存するが，各テナントでVLANを1個ずつ消費するのではなく，平均的に4 ～ 5個消費するため，一つのVLANエリアで収容できるテナント数は800程度に制限される。

● 適用可能な技術

その課題を解決するために，以下のような技術が考えられている。

（1）VLANエリア接続方式

VLANで構築可能な規模のIaaSシステムを複数個連結して大規模なシステムを構築する技術であ



REST API：REpresentational State Transfer Application Programming Interface
CLI：　Command Line Interface
SNMP：　Simple Network Management Protocol

図-1　クラウドシステムのアーキテクチャ

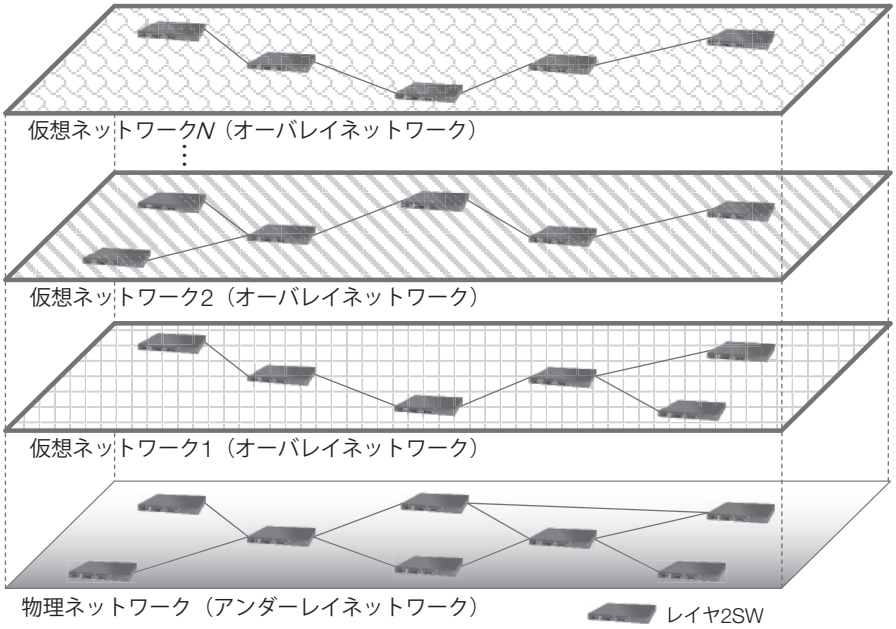


図-2　ネットワークの仮想化

る。あるテナントに対するVLAN IDを，一つのVLANエリア内に制限し，ほかのエリアで該当するIDは別のテナントが使用できるようにする。複数のVLANエリアにまたがった仮想システムを構築するために，エリア間にVLAN ID変換スイッチを設置する。VLANエリアを相互接続することで，4096個×エリア数の単位で仮想ネットワーク数を拡張する技術である。また，データセンター間の仮想網延伸は，VPLS（Virtual Private LAN Service）などの技術により実現可能である。

（2）VM間ラベルスイッチング方式

各テナントを構成するVM間の通信を，従来のVLANの代わりに，ラベルスイッチング（MPLS）技術を用いて制御する技術である。データセンター内の物理ネットワークをMPLSで構築して，サーバ間をテナントごと，リンクごとにラベルを用意し，パケットを転送する。MPLSのShimヘッダのラベルは20ビットあるため，100万以上の仮想ネットワークを構築することが可能となる。データセンター間の仮想網延伸もMPLSにより実現可能である。

（3）L2 over L3トンネリング方式

各テナントを構成するVM間の通信を，従来のVLANの代わりに，レイヤ3トンネリング技術により制御する技術である。理論上構築可能な仮想ネットワーク数はトンネル識別子のビット長で制限され，例えばVXLAN[(2)]の場合は24ビットあるため1600万以上の仮想ネットワークが構築可能である。またデータセンター間の仮想網延伸はレイヤ3トンネルの到達性が保証されれば良いため，既存のレイヤ3技術により実現可能である。

●　各ネットワーク仮想化方式の特徴

上述の三つのネットワーク仮想化方式の特徴を説明する（**表-2**）。なお，各ネットワーク仮想化方式には長所・短所があるため，顧客要件や求めるネットワーク特性を見極め，どの方式が良いかを検討することが重要である。

（1）VLANエリア接続方式

既存のレイヤ2技術による接続のため，レイヤ2スイッチでネットワークを構成することにより，ルータでネットワークを構築するほかの方式と比較して，低遅延のネットワークを構築することが可能である。

（2）VM間ラベルスイッチング方式

スイッチ機器によるネットワークトポロジーがTree構造など，ある程度制約が出てしまうVLANエリア接続方式と比較して，MPLS技術でネットワークを構成するため，自由なトポロジーでネットワークを構築することが可能である。

表-2　仮想ネットワーク構築方式の比較

	VLANエリア接続方式	VM間ラベルスイッチング方式	L2 over L3トンネリング方式
長所	・従来のVLAN使用により，新技術不要。既存ノウハウが活用可能 ・レイヤ2ベースの低遅延なネットワークを構築可能	・レイヤ3ベースのネットワーク構築のため，ネットワーク設計の自由度が高い	・仮想スイッチ以外，既存機器流用し，低価格で構築可能 ・物理ネットワークを仮想ネットワークとは独立に構築可能
短所	・DC間はVPLSなどで接続要 ・最新の負荷分散技術（TRILL）導入には同一ベンダ装置での構築が必要	・高価なネットワーク装置が必要 ・サーバ内の仮想スイッチでMPLSをサポート要	・トンネルヘッダのオーバヘッドによる性能劣化 ・ジャンボフレーム対応要 ・DC間は双方向IPマルチキャスト対応要
用途	・低遅延性重視	・ネットワーク設計の自由度重視	・価格・既存ネットワークとの親和性重視

（3）L2 over L3トンネリング方式

各方式の中で，既存のネットワークとの親和性が高い技術である。トンネルの終端はサーバ内の仮想スイッチまたは外付けのハードウェアスイッチで行うため，中継するネットワーク機器は従来のレイヤ2またはレイヤ3ネットワーク機器で対応可能であり，既存のネットワーク機器を新規に置き換える必要がないため，導入に当たってのハードルが低いと思われる。ただし，サーバでトンネルを終端する場合，サーバ内の仮想スイッチでの性能劣化による遅延増大やスループット低下が懸念される。またL2 over L3トンネリング方式は，現在，IETFのNVO3（Network Virtualization Overlay 3）WGで課題設定やフレームワークを規定しようとしている。トンネリング方式に関しても，VXLAN,[2] NVGRE,[3] STT[4] など多数あり，IETFで標準化を議論中のため，今後の動向に注目する必要がある。

次章では，仮想ネットワーク（オーバレイネットワーク）を支える物理ネットワーク（アンダーレイネットワーク）基盤，および課題（4）のトラフィック集中回避について述べる。

仮想ネットワークを支える物理ネットワーク基盤

仮想ネットワークを支える物理ネットワークに関しても様々な構築方法があり，どれを選択するかが重要である。サーバにおいてVM多重度増大により，サーバ間のトラフィック増大が今後懸念されるため，物理ネットワークは大容量化に対応できる必要がある。

通常，データセンターのネットワーク構成はラックごとに存在するToR（Top-of-Rack）スイッチ，ToRスイッチを収容する集約スイッチの2段構成，または集約スイッチを相互に接続するコアスイッチを含めた3段構成を取る。従来のSTP（Spanning Tree Protocol）を用いた構成ではブロッキングポートによる回線の無駄が発生するため，ループを防止し効率的なネットワーク構築が重要となる。

現在，レイヤ2およびレイヤ3ごとにそれぞれ以下の方法が考えられる。物理網の構築技術および方法を**図-3**に示す。

（1）ファブリックスイッチ

複数のスイッチを束ね，一つの仮想的なスイッチとして制御する技術である。ファブリックスイッチはSTPの代わりに，IS-IS（Intermediate System to Intermediate System）によるルーティングプロトコルを持つ技術｛SPB（Shortest Path Bridging），TRILL（TRansparent Interconnection of Lots of Links）｝をEthernetに適用し，転送先に到達できる経路テーブルを基に転送することで，ループを防止し，ネットワークを有効活用する。また，各経路のコスト値を制御し，複数の経路に負荷分散することで，トラフィックの集中回避が可能である。

（2）Staticレイヤ3ネットワーク

ToRスイッチからのトラフィックを集約する集約スイッチを多数台並べ，パケットのNext-Hopアドレスに集約スイッチのIPアドレスを指定することで，どの集約スイッチにどの位のトラフィックを流すかを明示的に指定するネットワーク構築方法である。この方法による負荷分散でトラフィック集中回避が可能となる。

（3）MPLSネットワーク

従来のIP技術およびMPLS技術を用いてネット

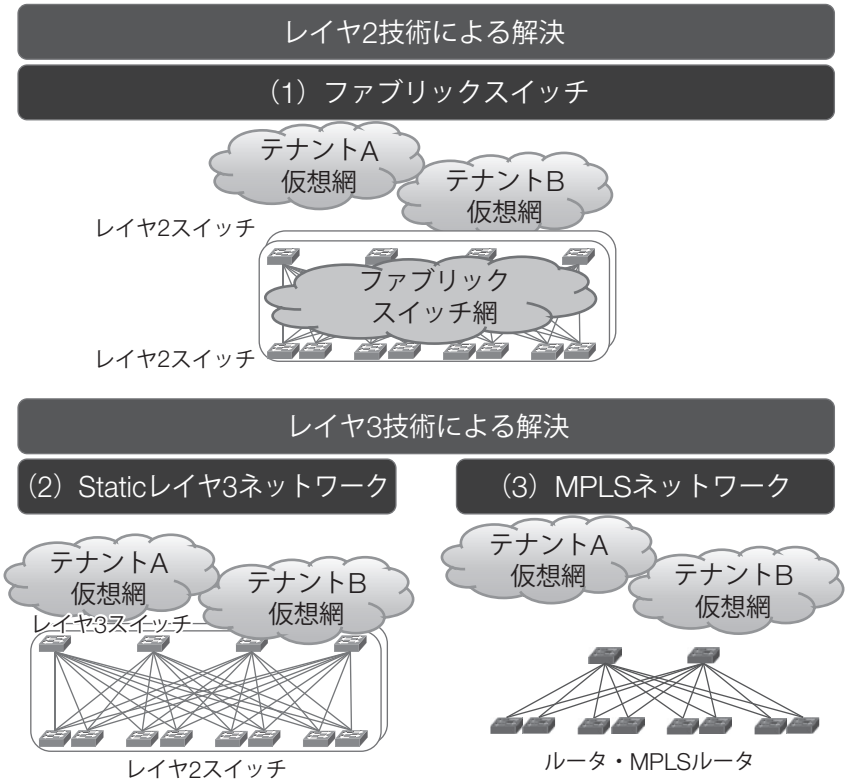


図-3　物理網の種類

ワークを構築することにより，既存のルーティングプロトコルでループを防止し，ネットワークを有効活用する技術である。また，MPLS-TE（Traffic Engineering）により明示的に複数の経路に負荷分散することで，トラフィック集中回避が可能である。

前章の仮想ネットワーク構築において，VLANエリア接続方式はレイヤ2ベースの技術であるため（1）のファブリックスイッチ，VM間ラベルスイッチング方式は（3）のMPLSネットワーク，L2 over L3トンネリング方式では物理ネットワークは仮想ネットワークと独立に構築可能なため，（1）～（3）のどの方式とも組み合わせることが可能である。

仮想ネットワークと物理ネットワークの構築方式の関係を**図-4**に示す。

ファブリックスイッチはレイヤ2スイッチで構成することにより，ほかの方式と比較して，低遅延なネットワークを構成することが可能である。

Staticレイヤ3ネットワークでは既存レイヤ2スイッチとレイヤ3スイッチの組合せで構築可能なため，比較的に安価でかつ安定した運用を期待することが可能である。

MPLS ネットワークは，ほかの方式と比較して，MPLSでネットワークを構築するため，ネットワーク設計の自由度が高く，実際の構築例も多いため，ネットワークの運用リスクが低い。

各方式の長所・短所を**表-3**で整理する。

ネットワーク仮想化方式と同様，物理ネットワークに関しても，顧客要件や重視する特性に応じてネットワーク構築方式を検討することが重要となる。

次章ではIaaS構築に当たり，これまでに述べた仮想ネットワークおよび物理ネットワークの一元管理について述べる。

物理および仮想ネットワークの一元管理

従来，ネットワーク機器は，装置1台ごとにコマンドにより設定を行っていた。またベンダごとに制御コマンドラインが異なるため，ネットワークの設定作業や変更作業は従来かなり時間やコストのかかるものとなっていた。そのため，制御サーバからネットワークの機器を一元的に制御できるようにすることが期待されている。制御するため

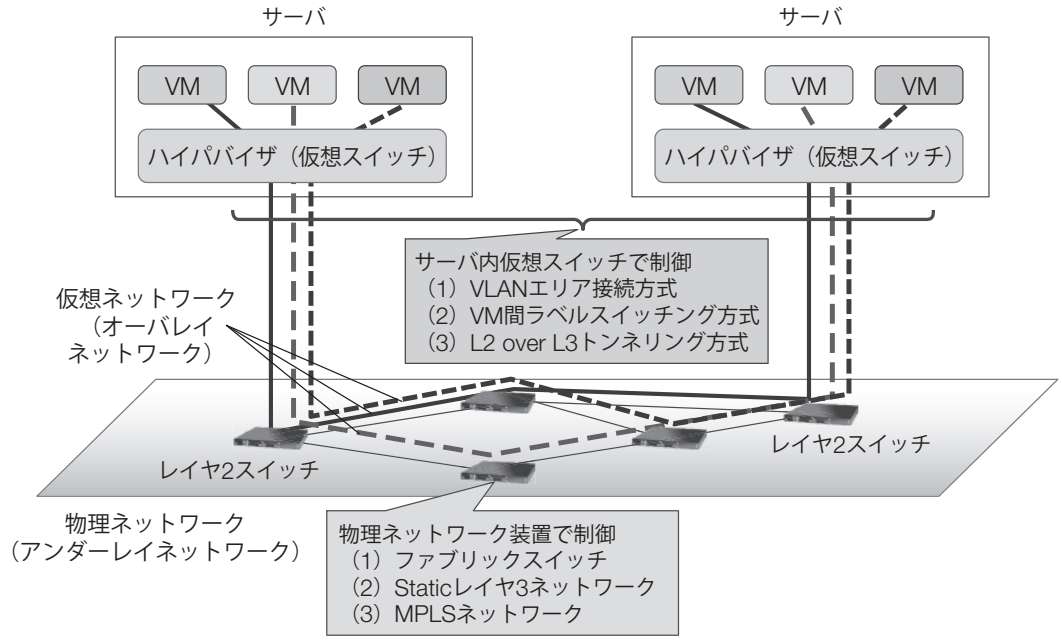


図-4　仮想ネットワーク構築方式と物理ネットワーク構築方式の関係

表-3　物理ネットワーク構築方式の比較

	ファブリックスイッチ	Staticレイヤ3ネットワーク	MPLSネットワーク
長所	・低遅延	・既存機器での構築により低価格	・ネットワーク設計の自由度が高い ・既存技術により低リスク
短所	・同一ベンダ製品のみで構成要 ・TRILLでは更に20バイトのオーバヘッドあり	・static設定により設定手順が煩雑	・ルータやレイヤ3スイッチにより，遅延時間が比較的大きい
用途	・価格・低遅延性重視	・既存ネットワークとの親和性重視	・ネットワーク設計の自由度や運用リスク低減重視

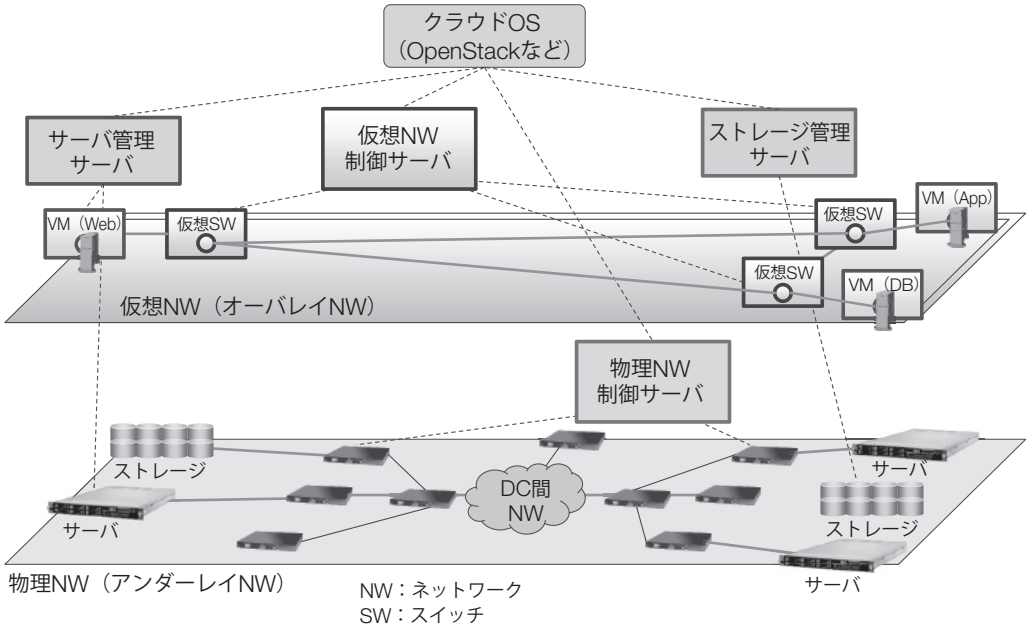


図-5　クラウドOSによるIaaSシステムの一元管理

のプロトコルとして，既存のCLI（Command Line Interface）やSNMP（Simple Network Management Protocol）がある。物理ネットワークおよびテナントごとの仮想ネットワークを制御サーバから制御することが必要である。一つの制御サーバで物理ネットワークと仮想ネットワークを集中制御してもよいし，それぞれ別々の制御サーバを設置し，制御を分担してもよい。

IaaS構築においては，ネットワークのみに限らず，サーバ・ストレージ・ネットワークを統合的に管理するために，クラウドOSと呼ばれるソフトウェアを導入することが今後の主流になると思われる。

クラウドOSとして，オープンソースのOpenStack，CloudStack，あるいはベンダ独自のクラウドOSが現在製品発表されている。オープンソースのクラウドOSはベンダロックインを回避することができ，価格や機能を考慮し，自由にベンダ機器を組み合わせることが可能であることがメリットとして挙げられる。またベンダ独自のクラウドOSでは，そのベンダ独自に拡張され，機能拡張や運用管理で利便性向上を図っているものが多いことがメリットである。

それぞれ顧客ごとの機能要件や性能要件により，適切なクラウドOSを選択していくことが重要である。

クラウドOSによるIaaSシステムの一元管理イメージを**図-5**に示す。

むすび

本稿では，大規模データセンターにおける課題を踏まえ，大規模化に対応可能な仮想ネットワークの実現方式について述べた。またその仮想ネットワークを支える物理ネットワークの構築方式についても述べた。

今後の展開として，データセンターを構成するサーバやストレージの高密度化も進み，10 Gbps以上の超高速インタフェースの使用が主流になってくると思われる。そのため，今以上にトラフィック設計が重要である。またVM間のトラフィックを局所化し，集約スイッチやコアスイッチに流入するトラフィックを抑制するため，VMの配置設計が非常に重要になってくる。

今後も大規模化の一途をたどるデータセンターの安定運用に向けたネットワーク技術の開発に取り組んでいく。

参考文献

(1) 大久保修一：「さくらのクラウド」におけるL2ネットワークの課題．2011年10月，MPLS Japan 2011.
http://mpls.jp/2011/presentations/20111024-sakura-mplsjp.pdf

(2) IETF：VXLAN：A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks draft-mahalingam-dutt-dcops-vxlan-01.
http://datatracker.ietf.org/doc/draft-mahalingam-dutt-dcops-vxlan/

(3) IETF：NVGRE：Network Virtualization using Generic Routing Encapsulation draft-sridharan-virtualization-nvgre-01.
https://datatracker.ietf.org/doc/draft-sridharan-virtualization-nvgre/

(4) IETF：A Stateless Transport Tunneling Protocol for Network Virtualization（STT）draft-davie-stt-01.
http://tools.ietf.org/html/draft-davie-stt-01

著者紹介

安藤達宏（あんどう　たつひろ）
ネットワークソリューション事業本部
サービスノード事業部 所属
現在，通信事業者向け新規ソリューション開発に従事。

下國　治（しもくに　おさむ）
クラウドコンピューティング研究センター 所属
現在，通信事業者およびデータセンター事業者のデータセンターネットワーク技術における研究開発に従事。

麻野克仁（あさの　かつひと）
ネットワークソリューション事業本部
サービスノード事業部 所属
現在，通信事業者向け新規ソリューション開発に従事。