# 仮想ルータを使ったスイッチレス・サーバクラスタリングのデザインパターン

松本直人[†1]

現在私たちを取り巻くサーバ環境は Ethernet では 10Gbit/sec から 56Gbit/sec に達し InfiniBand では 100Gbit/sec にまで広帯域化が進んでいる．しかしサーバに導入する NIC(Network Interface Card)や HCA(Host Channel Adapter)の低価格化は進んだが，広帯域ネットワークスイッチは依然として高価でありシステム導入の障害となっている．本稿では，物理サーバの仮想マシンマネージャ上に仮想スイッチおよび仮想ルータを導入することでサーバ間をリング状に直結し，広帯域ネットワークスイッチを購入せずコストを抑えたサーバクラスタリングを行うネットワーク設計手法について考察する．

# The design patterns of switch-less server clustering using virtual router

NAOTO MATSUMOTO[†1]

The server interconnect and network switch bandwidth is increasing day by day on the market. But the high bandwidth network switch cost is most important factor for small size server clustering. This paper is introduce to analysis the design pattern of the small size switch-less server clustering using virtual routers without high bandwidth network switch devices for system administrators.

## 1. はじめに

今日，サーバを取り巻くネットワーク環境はネットワークインターフェイスカードの広帯域化と低価格化が進んでいる．しかし広帯域ネットワークスイッチは依然として高価であり，小規模なサーバクラスタリングにおいてコスト面での障害となっている．(図 1)

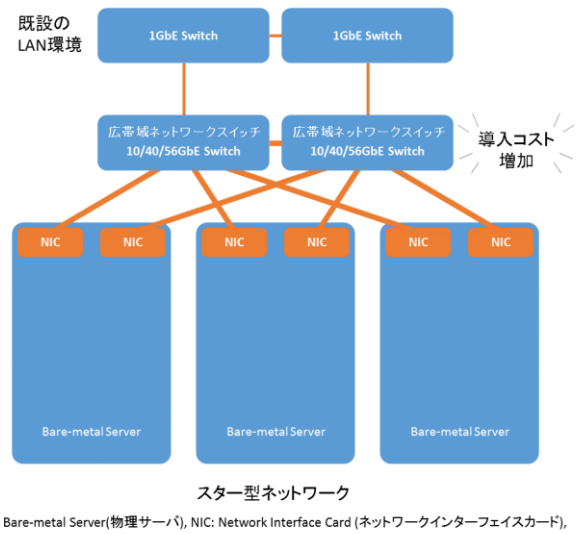



図 1 一般的なサーバクラスタリング
Figure 1 The diagram of typical server clustering.

本稿では VMware vSphere ESXi，Microsoft Hyper-V などの仮想マシンマネージャ上に仮想スイッチおよび仮想ルータを導入することで，高価な広帯域ネットワークスイッチを購入することなく，小規模なサーバクラスタリングの広帯域化を実現するネットワーク設計手法について考察する．(図 2)

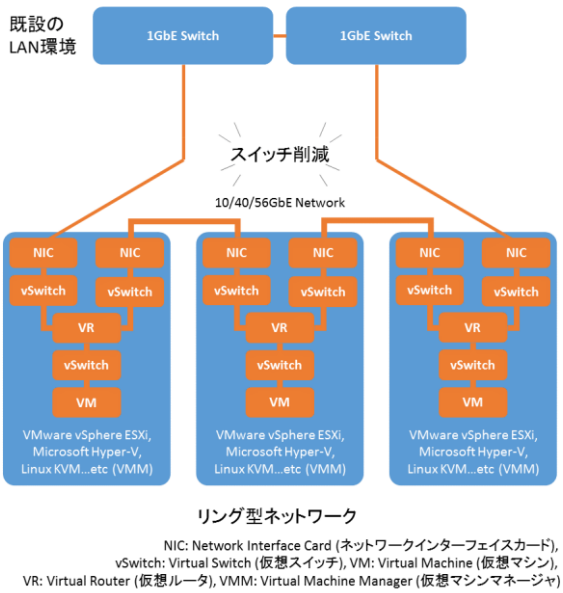



図 2 スイッチレス・サーバクラスタリング
Figure 2 The diagram of switch-less server clustering.

†1 さくらインターネット(株)
SAKURA Internet, Inc.

ⓒ2014 Information Processing Society of Japan

## 2. スイッチレス・サーバクラスタリング

### 2.1 コスト削減効果とデメリット

本稿では仮想マシンマネージャ上に仮想スイッチおよび仮想ルータを導入することで，高価な広帯域ネットワークスイッチを購入することなく，小規模なサーバクラスタリングの広帯域化を実現するサーバクラスタリングのネットワーク設計手法を「スイッチレス・サーバクラスタリング」と表記する．

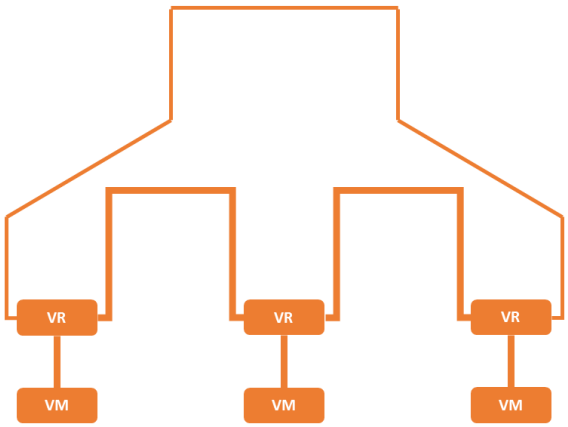
この手法では，高価な広帯域スイッチに代わり物理サーバの仮想マシンマネージャ上に仮想スイッチおよび仮想ルータを導入することで実現されるコスト削減効果が大きく期待される．

しかし，スイッチレス・サーバクラスタリングで構成されるリング型ネットワークには，一つの物理サーバ上で起こる障害は広帯域スイッチと同等のネットワーク耐障害性を備えるが，同時に複数の物理サーバで起こる複合障害には対処できないというデメリットを持つ．

このためスイッチレス・サーバクラスタリング導入の要求分析には十分な検討と注意が必要である．

### 2.2 ネットワーク設計手法の概念理解

図 3 は，スイッチレス・サーバクラスタリングのネットワーク設計手法の概念理解を深めるために図 2 を IP(Internet Protocol)層に着目して抽象化したものである．この図により一見複雑に入り組んだサーバクラスタ構成も，仮想ルータが単純なリング状に接続しているだけであることが分かる．




VM: Virtual Machine (仮想マシン), VR: Virtual Router (仮想ルータ)

図 3 スイッチレス・サーバクラスタリング抽象化モデル
Figure 3 Switch-less server clustering: Abstract model.

### 2.3 構成要素と利用技術

スイッチレス・サーバクラスタリングでは，サーバクラスタ内の耐障害性を高めるため OSPF(Open Shortest Path First)[1]プロトコルによる経路制御と VRRP(Virtual Router Redundancy Protocol)[2]による仮想ルータの冗長化が行われている．

図 4 は，五台の仮想ルータを用いた小規模なスイッチレス・サーバクラスタリングの構成手順を示したものである．

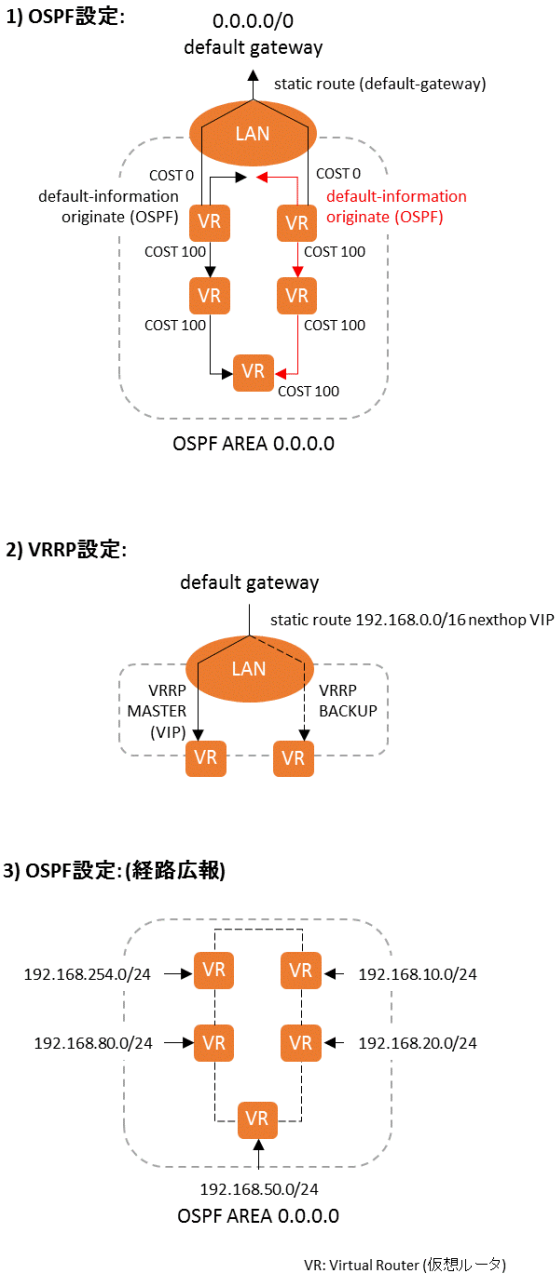



図 4 スイッチレス・サーバクラスタリングの構成例
Figure 4 Switch-less server clustering: design pattern.

既存の LAN 環境に隣接する二台の仮想ルータには，静的なデフォルトルートが上位のルータに設定されており，その他の仮想ルータには OSPF プロトコルを通じてデフォルトルートが伝搬する．

この二台の仮想ルータは VRRP により冗長化されており，上位のルータからサーバクラスタ内に存在するすべての経路に対して，VRRP で管理される仮想 IP アドレス(VIP: Virtual IP Address)を宛先として静的経路が設定されている．また，すべての仮想ルータは OSPF プロトコルを通じて，仮想ルータが管理する下位ネットワークの経路広報を行っている．

これらすべての構成要素によってスイッチレス・サーバクラスタリングは実現されている．

ⓒ2014 Information Processing Society of Japan

既存の LAN 環境と仮想ルータはリング状のネットワークを形成し，あらかじめ設定された OSPF コスト・パラメータや VRRP 設定などの条件に従い，一つの仮想ルータに障害が発生しても隣接する仮想ルータを経由してデータ通信が継続するようになっている．

OSPF コスト・パラメータは設定する仮想ルータから見て，既存の LAN 環境にホップ数が近いインターフェイスに 0 が設定され，迂回路となるインターフェイスには 100 が設定されている．

この際の例外として，設定する仮想ルータから見て，両方のインターフェイスから既存の LAN 環境へのホップ数が同数となった場合，VRRP MASTER として設定される仮想ルータに近いインターフェイスが 0 に設定され，迂回路となるインターフェイスには 100 が設定される．

## 3. ネットワーク冗長化の動作検証

### 3.1 検証環境

本稿の検証環境では，一台の物理サーバ上に仮想マシンマネージャとして VMware vSphere ESXi 5.1，仮想ルータとして Brocade Vyatta 5400 vRouter 6.6R5[3]を用意し，複数台の仮想ルータと仮想スイッチによる疑似的なスイッチレス・サーバクラスタリング環境で動作検証を行った．

### 3.2 正常系の動作確認

サーバクラスタ内のすべての仮想ルータは，あらかじめ設定された OSPF コスト・パラメータと VRRP 設定に従い経路選択が行われる．(図 5)

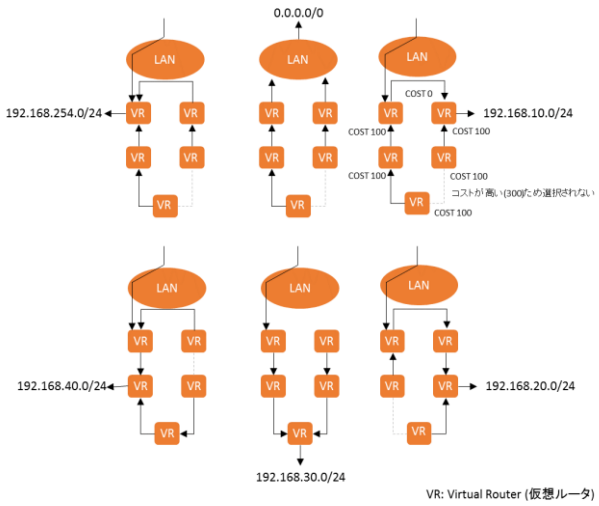



図 5 スイッチレス・サーバクラスタリングの動作検証
Figure 5 Switch-less server clustering: route select.

この際，経路によっては 192.168.10.0/24 のように隣接する仮想ルータが目的とするネットワークまでのホップ数として近くとも，OSPF コスト・パラメータの設定に従い経路選択が行われる．

### 3.3 異常系の動作確認

図 6 は，既存の LAN 環境への基幹経路，もしくは仮想ルータ間の中間経路で起こる障害を想定した検証結果である．

基幹経路の障害では，VRRP MASTER として設定された仮想ルータが停止した際，VRRP BACKUP として設定された仮想ルータが昇格し経路制御が引き継がれる．

この際，OSPF 設定の情報更新タイミグに合わせて VRRP MASTER から広報される OSPF default-information originate が消え，VRRP BACKUP からの OSPF default-information originate がデフォルトルート情報となる．

これにより静的なデフォルトルート設定を持たないすべての仮想ルータのデフォルトルートは VRRP BACKUP へ切り替わる．

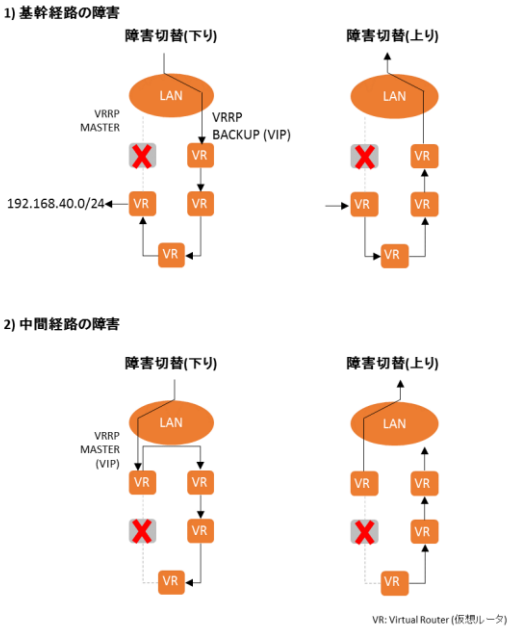



図 6 仮想ルータの障害切替
Figure 6 Virtual router failover.

中間経路の障害では，VRRP MASTER として設定された仮想ルータに近い経路の仮想ルータが停止した際，VRRP BACKUP を経由する迂回経路が選択される．

この際，OSPF 設定の情報更新タイミグに合わせて VRRP MASTER から広報される OSPF default-information originate の OSPF コスト値が変化し，VRRP BACKUP の OSPF default-information originate がデフォルトルート情報となる．この際，VRRP MASTER および VRRP BACKUP にはネットワーク冗長化としての変化が起きないため，データ通信は上りと下りで非対称経路となる．

## 4. ネットワーク設計の考察

### 4.1 ネットワーク帯域制御

スイッチレス・サーバクラスタリングでは，すべての仮想ルータが一本の輪のように接続され，ネットワーク帯域を共有する．

このため，すべての仮想ルータはネットワークインターフェイスカードの帯域上限に達しないよう，仮想ルータ単位で帯域制御する必要がある．(図 7)

ⓒ2014 Information Processing Society of Japan

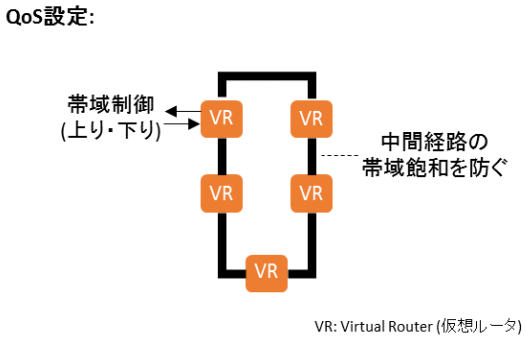



図 7 スイッチレス・サーバクラスタの帯域制御
Figure 7 Switch-less server clustering: QoS config

この際，既存の LAN 環境に隣接する仮想ルータのようにネットワーク帯域が細い場合，状況により広帯域のネットワークインターフェイスカードを追加する必要がある．(図 8)

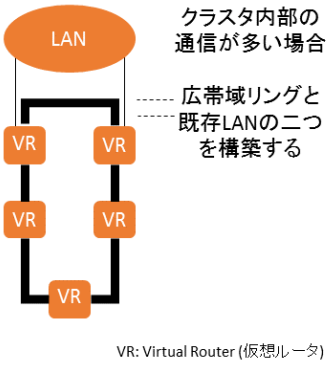



図 8 スイッチレス・サーバクラスタリングの帯域設計
Figure 8 Switch-less server clustering: dual-net.

### 4.2 パケット処理性能

サーバクラスタ内で閉じたデータ通信量が大きい場合，仮想ルータや仮想マシン，仮想マシンマネージャの MTU(Maximum Transmission Unit)サイズを変更する必要がある．(図 9)

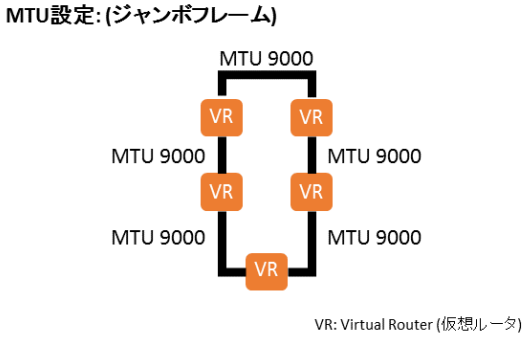



図 9 仮想マシンマネージャ・仮想ルータの MTU サイズ変更
Figure 9 MTU size up config on VMM and VR.

これは仮想ルータや仮想マシン，仮想マシンマネージャにおけるパケット転送処理が強く CPU に依存するためであり，大量のパケット転送処理が仮想マシンや仮想ルータで行われた場合，物理サーバの CPU を大きく占有しかねないためである．

このため，仮想ルータや仮想マシン，仮想マシンマネージャにあらかじめ設定されている一つのパケットあたりの送受信可能な MTU サイズを，標準的な 1,500 バイトから 9,000 バイトなどに変更する必要がある．

図 10 は，標準的な Linux OS における 10Gbit/sec Ethernet のパケット処理性能(Packet/sec)を表したものである．赤線は理論上のハードウェア処理性能を表し，青い縦棒はパケット生成装置を用いた検証結果を表している．

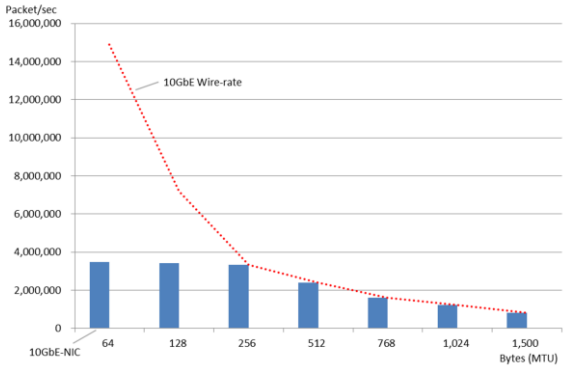



図 10 10Gbit/sec Ethernet におけるパケット処理限界
Figure 10 10Gbit/sec Ethernet: Packet/sec limit

この検証結果からも，パケット転送をハードウェア処理する広帯域ネットワークスイッチに比べ，Linux OS をベースとしてソフトウェア処理される仮想ルータでは，MTU サイズが 256 バイト以下の比較的小さい場合，性能差に開きが出てくることが確認できる．

このような比較的小さいパケット転送を頻繁に行うようなサーバクラスタを設計する場合，要求分析に十分な考慮が必要である．

### 4.3 オーバレイネットワーク

サーバクラスタ内の異なる仮想ルータ配下に存在する仮想マシンを Ethernet レベルで同一セグメント化するには，トンネリングとブリッジ機能を有した GRE[4]や L2TPv3[5]などオーバレイネットワーク技術が必要となる．

この際，データ通信の暗号化が要求仕様にあれば IPSec[6]などの適用も必要である．

図 11 は，サーバクラスタ内におけるオーバレイネットワークの動作と耐障害性を確認したものである．

これにより，仮想マシンを Ethernet レベルで同一セグメント化している仮想ルータ間にネットワーク接続性がある限り，オーバレイネットワークが維持されることが確認できる．

ⓒ2014 Information Processing Society of Japan

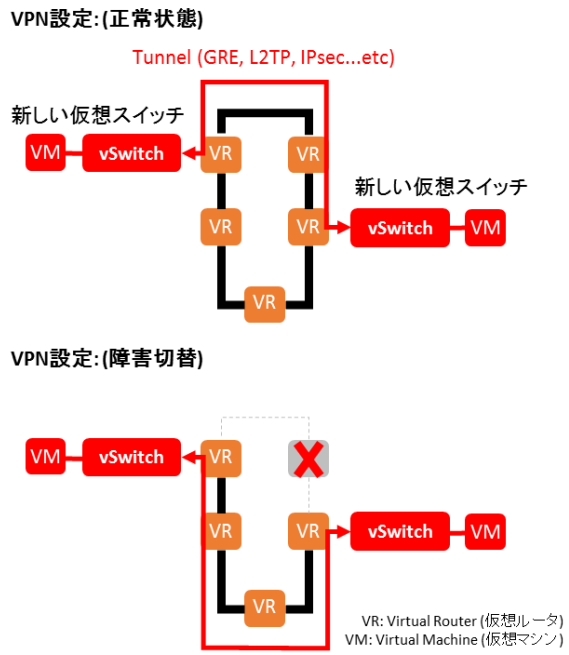



図 11 仮想ルータのトンネル技術における障害切替
Figure 11 Virtual router tunneling failover.

### 4.4 IP マルチキャスト

サーバクラスタ内の仮想マシンに IP マルチキャストサービスを提供するには，マルチキャストのドメイン間通信に用いられる MSDP(Multicast Source Discovery Protocol)[7]やマルチキャストルーティングプロトコルである PIM-SM (Protocol-Independent Multicast Sparse Mode) [8]と Anycast[9]を併用する必要がある．

図 12 は，サーバクラスタ内における IP マルチキャストの動作と耐障害性を確認したものである．

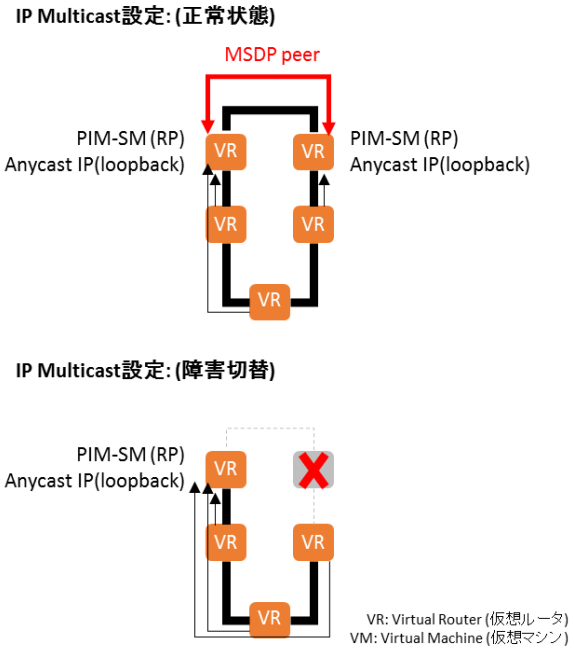



図 12 スイッチレス・サーバクラスタのマルチキャスト
Figure 12 Switch-less server clustering: Multicast config.

OSPF default-information originate を管理する二つの仮想ルータは，Anycast IP アドレスと PIM-SM によって IP マルチキャストドメインの幹(RP: rendezvous point)となる．

この際，二つの仮想ルータは MSDP により IP マルチキャスト・ソース情報も共有する．

幹となった仮想ルータはいずれか一方が停止しても，Anycast IP アドレスを宛先とした IP マルチキャスト通信が，一方の幹となる仮想ルータで管理されるため，IP マルチキャストドメインは維持される．

これにより，図 13 に抽象化して示されるような，ひとつの大きな IP マルチキャストドメインがサーバクラスタ内に形成される．

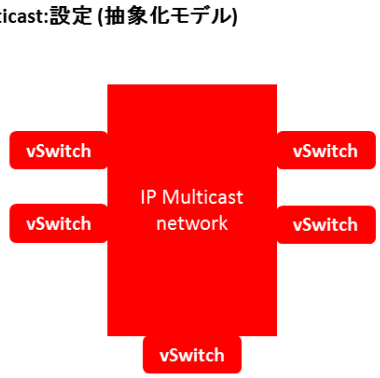



図 13 IP マルチキャストネットワークの抽象化モデル
Figure 13 IP Multicast network: Abstract model.

## 5. まとめ

本稿では VMware vSphere ESXi，Microsoft Hyper-V などの仮想マシンマネージャ上に仮想スイッチおよび仮想ルータを導入することで，高価な広帯域ネットワークスイッチを購入することなく，小規模なサーバクラスタリングの広帯域化を実現するネットワーク設計手法について考察した．

本稿で考察したスイッチレス・サーバクラスタリングは広帯域ネットワークスイッチの導入に比べ，複数の物理サーバで起こる複合障害に対処できない点や仮想ルータにおける一部パケット処理性能の低下，オーバレイネットワークの運用，IP マルチキャストのための Anycast の運用といった技術面での差異は大きい．(表 1)

ⓒ2014 Information Processing Society of Japan

表 1 スイッチレス・サーバクラスタリングの適応範囲
Table 1 Switch-less server clustering: Capability

| | 広帯域ネットワークスイッチ | スイッチレス・サーバクラスタリング |
|---|---|---|
| ネットワーク冗長化 | ○ | △*1 |
| ネットワーク帯域制御 | ○ | ○ |
| パケット処理性能 | ○ | △*2 |
| IPマルチキャスト | ○ | △*3 |
| VLAN | ○ | △*4 |
| システム導入コスト | 高い | 低い |

*1: 複数の物理サーバにおける複合障害には適さない
*2: 比較的小さいパケット処理には適さない
*3: Anycastなどの適用が必要
*4: オーバレイネットワーク技術(GRE, L2TPv3)などの適用が必要

しかし，設計・構築・運用するサーバクラスタリングの要求仕様によっては，高価な広帯域ネットワークスイッチを購入せずとも，ネットワークの広帯域化が行えるコスト削減効果は大きい.

本稿考察が，今後も発展し続けるサーバのネットワーク広帯域化に対して，管理者の技術的な選択肢を広げる一助となることを期待する.

## 参考文献

1) John, Moy.: OSPF Version 2. IETF, RFC2328.
2) Robert, Hinden.: Virtual Router Redundancy Protocol (VRRP). IETF, RFC3768.
3) Brocade Vyatta 5400 vRouter, http://www.brocade.com/
4) Dino, Farinacci. et al.: Generic Routing Encapsulation (GRE). IETF, RFC2784.
5) Jed, Lau. et al.: Layer Two Tunneling Protocol - Version 3 (L2TPv3). IETF, RFC3931.
6) Stephen, Kent. et al.: Security Architecture for the Internet Protocol. IETF, RFC4301.
7) Bill, Fenner and David, Mayer.: Multicast Source Discovery Protocol (MSDP). IETF, RFC3618.
8) Bill, Fenner. et al.: PIM-SM Protocol Independent Multicast - Sparse Mode (PIM-SM). IETF, RFC4601.
9) Joe, Abley. et al.: Operation of Anycast Services. IETF, RFC4786.

ⓒ2014 Information Processing Society of Japan