

大規模なスイッチレス・サーバクラスタリング構築運用の考察

松本直人^{†1}

現在サーバを取り巻くネットワーク環境は広帯域化が進み、対応するネットワークアダプタの低価格化も進んでいる。しかし広帯域スイッチは高価でありシステム導入のひとつの障害となっている。本稿では、ハイパバイザ上に仮想ルータを導入しサーバ間を大規模に相互接続することで、広帯域スイッチの導入コストを大幅に抑えるシステム試作について考察する。

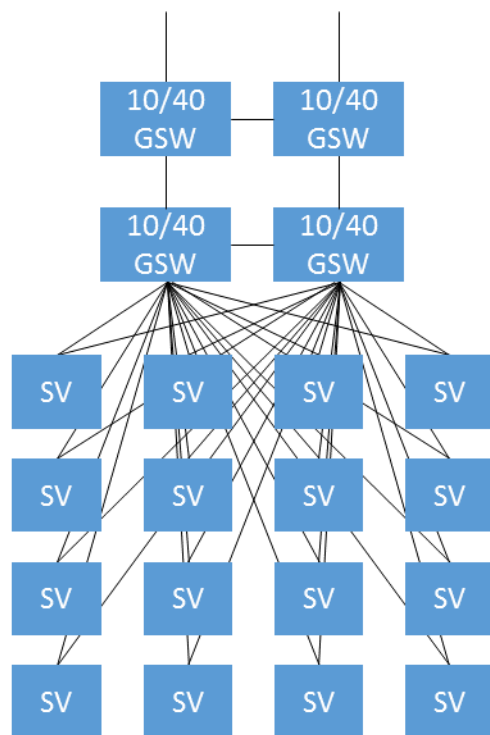
How to build and manage a large scale switchless server clustering

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 how to build and manage a large scale switchless server clustering using virtual routers without high bandwidth network switch devices for system administrators.

1. はじめに

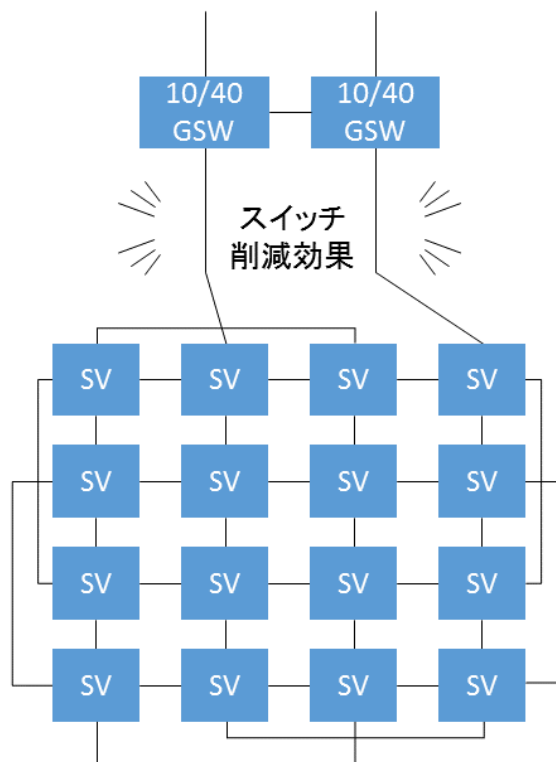
今日、サーバを取り巻くネットワーク環境はネットワークインターフェイスカードの広帯域化と低価格化が進んでいる。しかし広帯域ネットワークスイッチは依然として高価であり、大規模なサーバ構成を組む場合においてコスト面の大きな負担となっている。(図1)



10/40GSW: 10/40 Gigabit Ethernet Switch, SV: Server

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, Linux KVM などの仮想マシンマネージャ(ハイパバイザ)上に仮想スイッチおよび仮想ルータを導入しサーバ間を大規模に相互接続することで、広帯域スイッチの導入コストを大幅に抑えるシステム試作について考察する。(図2)



10/40GSW: 10/40 Gigabit Ethernet Switch, SV: Server

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.

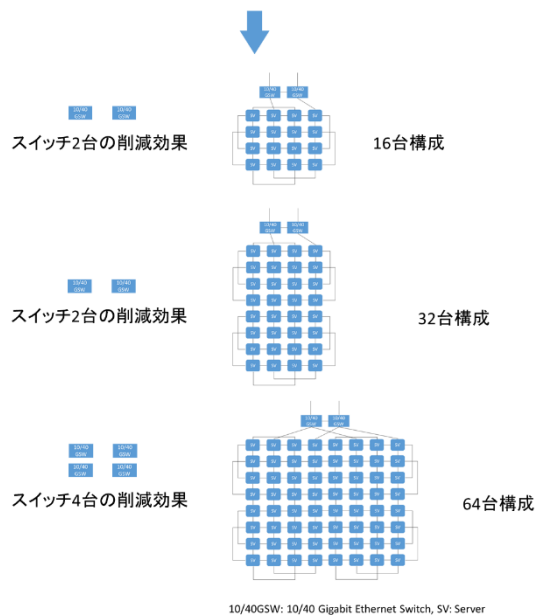
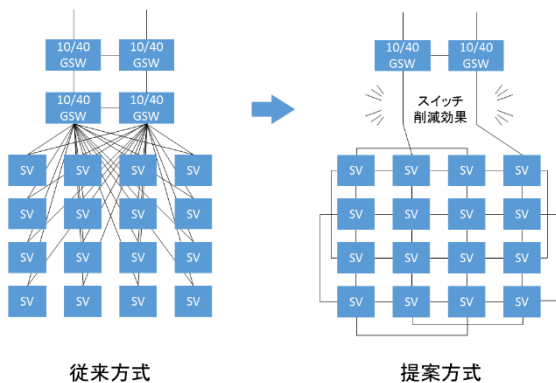
^{†1} さくらインターネット(株)
SAKURA Internet, Inc.

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(ハイパバイザ)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、大規模な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(ハイパバイザ)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。(図3)



台数が増えれば増えるだけスイッチ台数が減るコスト削減効果

図3 サーバクラスタリングのコストメリット
Figure 3 Switch-less server clustering: Cost merit

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ネットワーク内部には、外部接続に帯域を共有するためボトルネックが生じる点と、クラスタ内部への通信に複数サーバを経由するため伝送遅延が生じる点にデメリットがある。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4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を示したものである。サーバクラスタ内は経路制御と耐障害性を高めるため OSPF (Open Shortest Path First) version 3 [1] プロトコルによって構成されている。この際、サーバ間の相互接続には IPv6 Link-local Address および Unique Local IPv6 Unicast Addresses [3] が用いられ、ネットワーク設計および設定の簡素化が行われている。

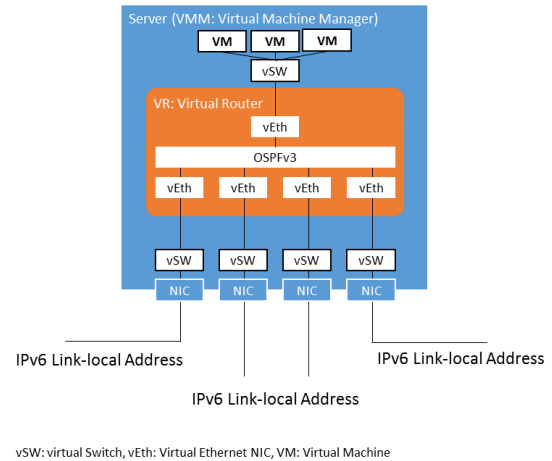


図4 スイッチレス・サーバクラスタリング抽象化モデル
Figure 4 Switch-less server clustering: Abstract model.

またサーバクラスタ外部との接続には経路制御とトラフィック制御の自由度を高めるため、BGP (Border Gateway Protocol 4) [2] が用いられている。この際、BGP で広報される経路はサーバクラスタ内の OSPF 経路に再配布されことで、ネットワーク中に点在する他のサーバクラスタとの相互接続を実現している。

OSPF 経路から BGP 経路への再配布は、サーバクラスタ内部での経路異常および設定ミスを防ぐ観点から設計上行っていない。

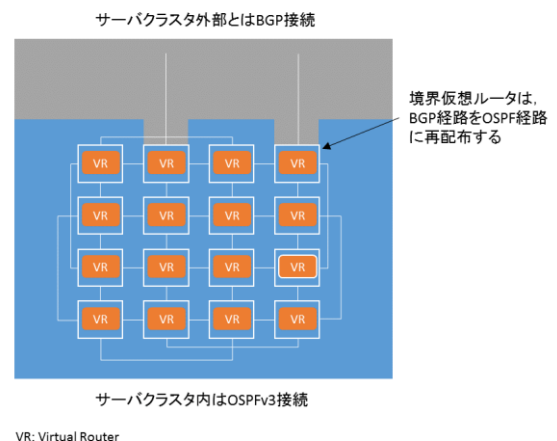
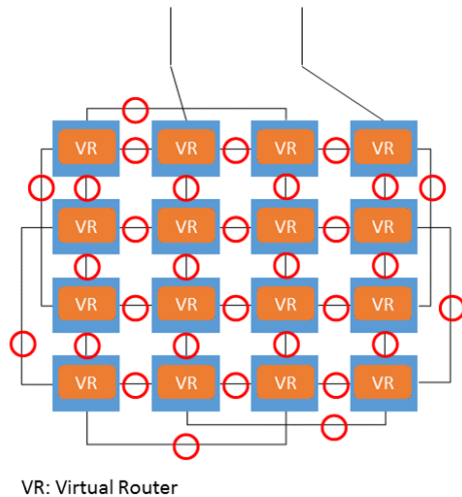


図4 スイッチレス・サーバクラスタリング経路設計
Figure 4 Switch-less server clustering: Routing model.

2.3 利用技術の比較

スイッチレス・サーバクラスタリングでは、ネットワーク設計および設定の簡素化を図るため OSPF(Open Shortest Path First) version 3[1]プロトコルと IPv6 Link-local Address が用いられている。スイッチレス・サーバクラスタリングは、その性質上サーバ間に多数のネットワーク設計および設定を必要とする。そのため、従来からの IPv4 および OSPFv2 を用いたネットワーク設計では、大規模であればあるだけ設定数が増大するという課題がある。(図 5)



○ IPv4およびOSPFv2でネットワーク設定を要する箇所

図 5 スイッチレス・サーバクラスタリング OSPFv2 設定
Figure 5 Switch-less server clustering: OSPFv2 config.

本稿提案方式では、この課題を解決するために OSPF(Open Shortest Path First) version 3[1]プロトコルと IPv6 Link-local Address を組み合わせることで、ネットワーク設計および設定の簡素化を図っている。(図 6)

作業 手順	IPv4 OSPFv2	IPv6 OSPFv3
物理サーバのNICポートと仮想ルータの仮想NICポートの対応を確認する。	○	—
仮想ルータの仮想NICポートに、IPアドレスを付与し疎通確認する。	○	—
OSPF設定を行い経路広報する。	○	○

<pre> router ospf6 interface eth0 area 0.0.0.0 interface eth1 area 0.0.0.0 interface eth2 area 0.0.0.0 interface eth3 area 0.0.0.0 router-id 10.0.0.16 area 0.0.0.0 range fd50:d026:7eee::/48 : </pre>	IPv6 OSPFv3設定例
--	----------------

図 6 スイッチレス・サーバクラスタリング OSPFv3 設定
Figure 6 Switch-less server clustering: OSPFv3 config.

3. 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 Microsoft Windows 8.1 Pro および Hyper-V 環境、仮想ルータとして CentOS 7(Linux 3.10.0-123.13.2.el7)および Quagga 0.99.22.4[4]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、OSPFv3 より自動構成された経路選択が行われる。(図 7)

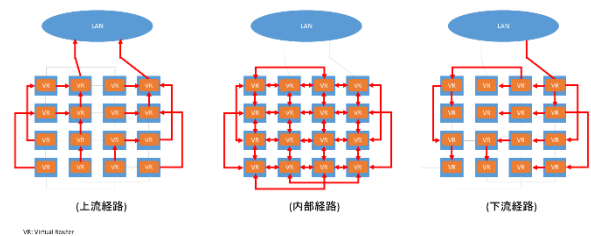


図 7 スイッチレス・サーバクラスタリングの動作検証
Figure 7 Switch-less server clustering: route select.

またこの際、サーバクラスタ内の隣接する仮想ルータは、もっとも近い回線を用いてトラフィック交換が行われる。

3.3 異常系の動作確認

図 8 は、複数個所のサーバが障害により停止した場合を想定した検証結果である。サーバクラスタ内部では利用可能な回線を用いた経路が選択されるようになり、それによってサーバクラスタ外部への接続性も引き継がれる。

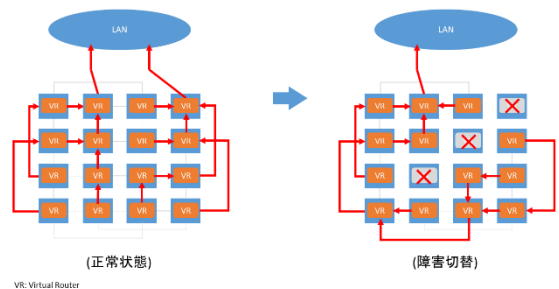


図 8 仮想ルータの障害切替
Figure 8 Virtual router failover.

この際、本稿の提案方式では IPv6 Link-local Address および Unique Local IPv6 Unicast Addresses[3]であるため、OSPF default-information originate などデフォルトルートの経路管理は行っていない点に注意が必要である。

4. ネットワーク設計の考察

4.1 サーバラック設計およびケーブリング

スイッチレス・サーバクラスターリングは、理論上のネットワーク設計において自由度は高い。しかしデータセンターの 19 インチ・サーバラックには物理的な制限があるためサーバ収容およびケーブリングに注意が必要である。

図 9 はスイッチレス・サーバクラスターリングの論理構成から物理構成への構成展開を助けるため概念を図示したものである。

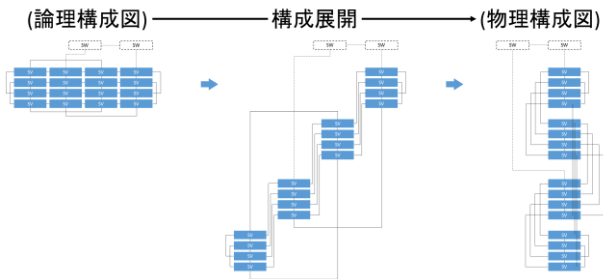
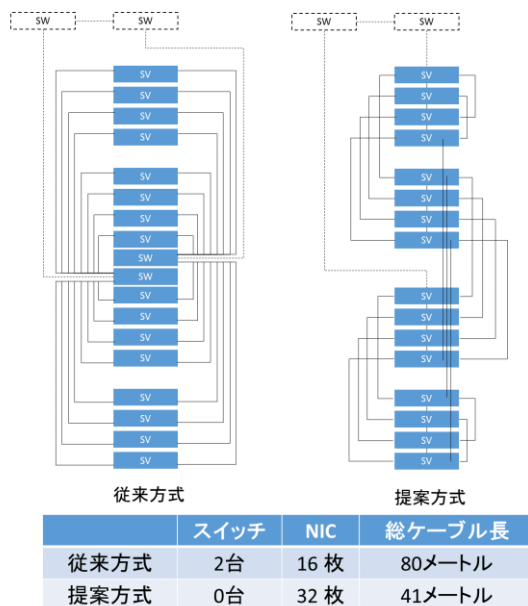


図 9 スイッチレス・サーバクラスターの 19' ラック設計
Figure 9 Switch-less server clustering: 19'RACK config

物理的に隣接するサーバ間は短い Ethernet ケーブルにより接続され、迂回経路をとる接続には長い Ethernet ケーブルが用いられる。スイッチレス・サーバクラスターリングの最小構成において本稿提案方式は従来方式に比べ Ethernet ケーブルの総ケーブル長も削減できる。(図 10)



「スイッチ 2 台の価格 > NIC 16 枚の価格」である時、提案方式が有効となる。

SW: 10/40 Gigabit Ethernet Switch, SV: Server

図 10 スイッチレス・サーバクラスターのケーブリング
Figure 10 Switch-less server clustering: Cabling

本稿の提案方式では、Twinax 方式など広帯域ネットワーク特有の太いケーブルに起因するサーバ排熱の妨げとなる

ケーブル余長を削減でき、それに伴うエアフローの改善も期待できる。

4.2 オーバレイネットワーク

本稿の提案方式では、サーバクラスター内に OSPF (Open Shortest Path First) version 3[1] プロトコルおよび IPv6 Link-local Address によりシステムが構成されるため、仮想マシンマネージャ (ハイパバイザ) 上の仮想マシンを Layer2 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する L2TPv3[5] などオーバレイネットワーク技術が必要となる。

このため、仮想ルータには OSPF (Open Shortest Path First) version 3[1] プロトコルおよび L2TPv3[5] などオーバレイネットワーク機能を同時に機能させる性能を有している必要がある。(図 11)

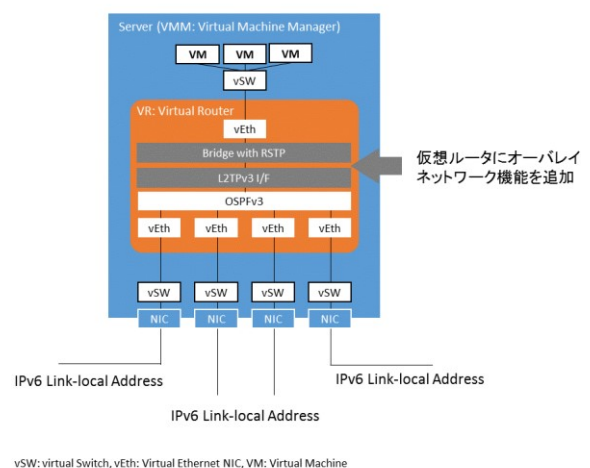


図 11 仮想ルータの L2TPv3 トンネル技術の適用
Figure 11 L2TPv3 Tunneling with the virtual router.

L2TPv3[5] などオーバレイネットワーク機能を仮想ルータにより実現することで、サーバクラスター内およびサーバクラスターを跨いだテナント毎の Layer2 over Layer3 ネットワーキングによる同一セグメント化が可能となる。(図 12)

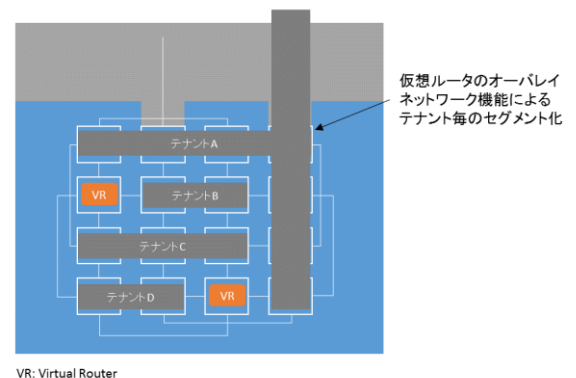


図 12 Layer2 over Layer 3 ネットワーキング
Figure 12 Layer2 over Layer 3 networking

4.3 パケット処理性能とフレームサイズ変更

サーバクラスタ内で L2TPv3[5] などオーバーレイネットワーク機能を仮想ルータにより実現するため、仮想ルータや仮想マシンマネージャ (ハイパバイザ) の MTU (Maximum Transmission Unit) サイズを変更する必要がある。

仮想ルータや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 バイトなどに適切なサイズへ変更すると良いだろう。

5. まとめ

本稿では仮想マシンマネージャ (ハイパバイザ) 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、大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メリットとして、大規模であればあるだけ広帯域スイッチの導入コストを削減でき、サーバ排熱の妨げとなる物理的なケーブル余長の削減にも効果が期待できる。

しかしデメリットとして、サーバクラスタ外部接続に帯域を共有するためボトルネックが生じる点と、クラスタ内部への通信に複数サーバを経由するため伝送遅延が生じる点があるため、導入の要求分析には十分な検討と注意が必要である。

本稿の提案方式が、今後も増大するサーバクラスタ内部のトラフィックを適切に制御できる一つの方式として活用されること期待する。

参考文献

- 1) Rob Coltun et al.: OSPF for IPv6. IETF, RF 5340
- 2) Yakov Rekhter et al.: A Border Gateway Protocol 4 (BGP-4). IETF, RFC4271
- 3) Robert M. Hinden et al.: Unique Local IPv6 Unicast Addresses. IETF, RFC 4193.
- 4) Quagga Routing Software Suite.: <http://www.nongnu.org/quagga/>
- 5) Jed, Lau. et al.: Layer Two Tunneling Protocol - Version 3 (L2TPv3). IETF, RFC3931.