RL-004

地理的に分散したサーバ間のフェイルオーバ・フェイルバックを可能にする複製サーバ冗長化構成*

Redundant Configuration of Replication Servers for Failover and Failback Among Servers in Different Locations

大隅 淑弘† 山井 成良† 岡山 聖彦† 河野 圭太† 藤原 崇起†
Yoshihiro Ohsumi Nariyoshi Yamai Kiyohiko Okayama Keita Kawano Takaoki Fujiwara

1. はじめに

我々の社会は多くの情報システムに依存しており，情報システムのサービス継続性は重要な課題となっている．我々の社会活動の基盤となっている情報システム，例えば銀行の預金管理システム，公共における住民情報の管理システム，列車の運行管理システム等々はいかなる場合も正しく動作しなければならない．このため，システムの信頼性向上を目指した開発が行われている．従来から用いられている一般的な方法が，ロードバランサや仮想化などによるサーバの冗長化であり，フェイルオーバシステムや負荷分散システムなどがある．これらは，サーバがある程度集中して設置されている場合に適した冗長化の方法である．しかし，場合によっては離れた場所にあるサーバを冗長化したいことがある．例えば，大きな組織でいくつかの拠点があり，サーバが分散して配置されているような場合や事業継続計画 (Business Continuity Plan)のためにサーバを分散配置する必要のある場合などである．

また，ネットワーク上のサービスでは，DHCP(Dynamic Host Configuration Protocol) [2] サーバや MTA(Message Transfer Agent)のように，あるサービスに対して複数のサーバを定義し，サーバの障害時には接続先が自動的に切り替わることができるものがある．このようなサービスでは，サーバを分散配置することもできる．しかし，サービスによっては，主系のサーバが障害でダウンし，待機系に切り替わった後で主系のサービスが復旧しても，クライアントからのアクセスが主系に戻らないものがある．これは冗長化を負荷分散の目的で運用している場合には問題となる．

一方で DNS ルートサーバの運用に用いられている IP anycast[3]による冗長化構成がある．IP anycast では，分散配置されたサーバを冗長化することができる他にもいくつかの利点を有している．

そこで本論文では，IP anycast に着目し，組織内において分散配置された複製サーバを冗長化する構成方法を提案する．IP anycast を用いることにより，サーバがダウンし，その後に復旧したときも，クライアントのアクセスが元のサーバに戻るため，負荷分散構成を回復させることができる．また，サーバの経路情報を操作することにより，フェイルオーバ構成でも負荷分散構成でも動作可能である．さらに，OS の基本機能で実装することが可能なため，ロードバランサなどの新たな装置や機器は必要ない．

以下，2 章では，従来の冗長化構成方法とその問題点について述べる．次に 3 章では，提案するサーバの冗長化構成について述べ，4 章では提案方法に基づいて構成した LDAP サーバと今後の実装計画について述べる．

2. 従来の冗長化構成方法と問題点

2.1 従来の冗長化構成方法

従来，ハードウェアによる冗長化構成方法には，一般にアクティブ・スタンバイ構成，Fault Tolerant(FT)構成，ロードバランサ構成などが利用されている．アクティブ・スタンバイ構成では，運用系と待機系によりサーバの多重化を行う．FT 構成では，システムを複数のノードで動作させ，障害発生時には別のノードが即座に処理を引き継ぐことにより，サービスの中断を最小限に抑える多重化を行う．ロードバランサ構成では，ロードバランサが各複製サーバを監視し，クライアントからのリクエストを最適なサーバに振り分けることで，多重化や負荷分散を行う．近年では，仮想化サーバによる High Availability(HA)構成，FT 構成が多重化として一般的になっている．

近年では仮想化サーバによるライブマイグレーションも利用することができる．ライブマイグレーションでは，稼働している OS やソフトウェアを停止させずに，別の計算機に移動させることができるため，比較的柔軟にフェイルオーバとしての冗長化を構成することができる．

ソフトウェアによる冗長化構成方法としては，DNS ラウンドロビン[4]やサービス自身がフェイルオーバの機能を持つものがある．DNS ラウンドロビンは，クライアントの DNS クエリに対して，DNS サーバが複数のサーバから IP アドレスを順に回答することで主に負荷分散を行う．サービスが冗長化の機能を実装しているものについて，DHCP サーバはネットワーク上に複数のサーバを設置することで多重化を行う．また，MTA では MX レコードを複数登録しておくことで，主システムの障害時には，自動的に待機系にフェイルオーバする．

また，IP anycast を用いた冗長化構成方法がある．DNS ルートサーバでは，UDP パケットの制限による問題を回避しながらも多数のサーバで負荷分散，冗長化を行うため，IP anycast によって運用されている．IP anycast では，通常，DNS ルートサーバに見られるように，ルータによる BGP[5]の経路情報制御によって，世界規模あるいは国や地方等の大きな範囲で実装されるのが一般的である．

2.2 従来システムの問題点

従来のハードウェアによる冗長化構成方法では，仮想化による構成を含めて，多くの場合冗長化されたサーバをデ

---

* 本論文は文献[1]を発展させたものである．

† 岡山大学情報統括センター Center for Information Technology and Management, Okayama University

ータセンターや組織の計算機室などに集中的に設置して運用することを前提にしている．このため，離れた場所にあるサーバを冗長化したい場合には利用できない．遠隔クラスタなどの製品もあるが，利用目的によって特殊な機能を実装している．また，このようなシステムでは，死活監視を行ったり，フェイルオーバさせたりするためのサーバや機器を動作させる場合には，そこが単一障害点（以下，SPOF(Single Point of Failure)）になる可能性がある．さらに，ロードバランサはクライアントと実サーバの間に設置して通信を中継するため，往復あるいは片道の通信がロードバランサを経由することになり，通信のボトルネックや SPOF の問題がある．

また，仮想化システムによるライブマイグレーションを利用するためには，仮想 OS を移動する物理サーバ間でハードウェアやソフトウェアの細かな要件をクリアする必要がある．

一方，ソフトウェアによる冗長化構成方法では，ハードウェアによる冗長化構成の問題は回避できるが，障害によってフェイルオーバした後でシステムが復旧し，フェイルバックしても，クライアントのリクエスト先が元のサーバに戻らないものがある．例えば Dovecot[6]で LDAP[7]認証を利用する場合，LDAP サーバとして複数のサーバを設定することができるため，通常アクセスしているサーバで障害が発生しても，他の LDAP サーバにフェイルオーバしてサービスは継続される．しかし，障害の起きた LDAP サーバが復活しても，Dovecot は元の LDAP サーバにフェイルバックしない．これではサーバの冗長化を負荷分散の目的にしている場合には問題となる．

このように従来からの実装方法では，システムの構成上の問題，冗長化サーバの分散配置の制約，フェイルバックなどにおいて問題がある．

3. IP anycast を用いた複製サーバの冗長化構成

3.1 冗長化構成の概要

冗長化するサービスにおいては，サーバを遠隔拠点に分散して設置する必要があり，障害が発生した場合には自動的にフェイルオーバしてサービスを継続するが，サーバ間でセッション管理や処理の継続，データ同期などを保証する必要のないものがある．そこで，本論文では，このような利用条件を想定し，分散配置されたサーバにも適用できる冗長化方法として， 組織内のネットワークで IP anycast を用いることにより，複製サーバの冗長化を行う．

この方法では，多重化したどこかのサーバで障害が発生して他のサーバにフェイルオーバした後に，そのサーバが復活したときには，クライアントのアクセス要求が元のサーバに戻ることで負荷分散が回復する．各サーバは通信できる範囲のネットワーク上のどこにあってもよく，各サーバで広告する経路情報をコントロールすることにより，フェイルオーバ構成でも，負荷分散構成でも動作が可能である．但し，負荷分散構成は各サーバが異なるルータに接続しており，ルータ間で経路情報が交換されるとメトリックが加算されるような場合に適用できる．すなわち，組織内の複数の拠点や範囲がルータで接続されており，異なるルータにサーバが配置されているような場合である．また，サーバが組織外へもサービスをしている場合には，組織がマルチホーミングの環境であり，各エッジルータから各サーバまでのメトリックが異なる場合には負荷分散構成も可能であるが，その他の場合には，エッジルータからのメトリックに差が出ないため，フェイルオーバによる構成となる．本提案方法では次の利点を有している．

- 通信のボトルネックや SPOF の問題を回避することができる
- フェイルバックするとクライアントからのアクセスは元のサーバに戻る
- OS の基本機能によって動作するため，ロードバランサなどの新たな装置や機器の導入が必要ない
- 負荷分散構成では IP anycast の特徴である平均応答時間の短縮，DoS (Denial of Service) 攻撃の局所化，DDoS (Distributed Denial of Service) 攻撃の効果低減を期待できる

なお，IP anycast の特徴として，接続条件によりパケットが到達する機器が異なること，送信元が同じでもパケットが到達する機器が異なることが挙げられるため，特に TCP による通信の場合は，ACK パケットの到達やパケットの分割を考慮して実装を行う必要がある．

前述のとおり一般に IP anycast は DNS ルートサーバのように，BGP を用いて大規模な範囲で利用されているが，本論文では組織内のネットワークに適用するものであり，他にはあまり例がない．

以下では，提案方法の詳細について述べる．

3.2 システムの構成

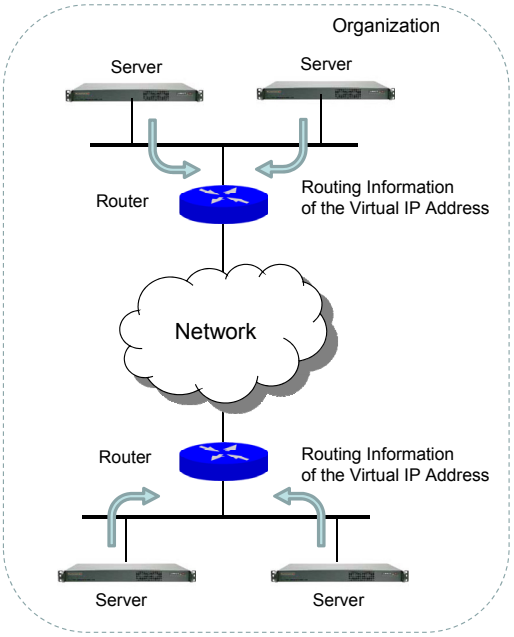

図 1　システムの構成例

構成要件としては，冗長化する複製サーバが 1 台以上のルータによって構成されているネットワークに接続されていることである．

まず，IP alias で各サーバのループバックインターフェースに，ネットマスクが 32 ビットの仮想 IP アドレスを設定する．各サーバでは，この仮想 IP アドレスを IP anycast アドレスとして，ネットワーク上に広告するためのルーティ

ングデーモンを動作させる．ネットマスクを 32 ビットにすることにより，経路情報の最長一致によってサーバが組織のどこにあっても接続が可能となる．この構成だけでも動作するが，目的のサービスだけが停止した場合の対策や，システムダウンなどにおいてサービスの停止時間を短縮する場合には，死活監視機能が必要である．この機能については，3.4 節で説明をする．3 台のサーバで冗長構成する例を図 1 に示す．

3.3 提案方法の動作手順

本提案方法の動作手順を以下に説明する．

(1) 各サーバでルーティングデーモンを動作させ，仮想 IP アドレスの経路情報をネットワーク上に広告する．このとき目的とする冗長化の構成方法によって，デフォルトのメトリックを決定して設定しておく．各サーバ間でフェイルオーバ構成をする場合には，クライアントから見て，スタンバイのサーバのメトリックが，アクティブなサーバのそれよりも大きくなるように設定する．また，負荷分散構成をする場合には，負荷分散をする範囲のクライアントから見て，最寄りのサーバのメトリックが他のサーバのそれよりも小さくなるように設定する．

(2) クライアントは，接続先サーバの IP アドレスとして，仮想 IP アドレスを指定する．

(3) クライアントがサーバに接続しようとすると，ルータ上の経路情報からメトリックの最も小さいサーバへの経路が選択されて接続される．

(4) いずれかのサーバで障害が発生した場合には，そのサーバの仮想 IP アドレスの経路情報がルータからフラッシュされることにより，他のサーバに自動的にフェイルオーバする．あるいは，3.4 節で述べるサーバの死活監視機能によって，経路情報の優先度が下げられて，他のサーバにフェイルオーバする．

(5) 障害が発生したサーバが修復され，サービスが復旧すると，クライアントのアクセス先は障害発生前のサーバにフェイルバックする．

以下に，2 台のルータと 2 台の Server がある場合を例にして動作の説明をする．

3.3.1 フェイルオーバ構成

ここでは Server B をスタンバイとするため，Server A，Server B のデフォルトのメトリック値をそれぞれ MA=1，MB=3 とする．これにより，Client A から見ると，各 Server の仮想 IP アドレスのメトリックが Server A では 2 に，Server B では 5 になる．また，Client B から見ると，Server A では 3 に，Server B では 4 になり，通常状態では，両サイトのクライアントは，Server A に接続される．ここで，例えば Server A で障害が発生した場合には，Server A の経路情報がフラッシュされることにより，両サイトのクライアントは自動的に Server B に接続される． Server B で障害が発生した場合にも同様の振る舞いにより，両サイトのクライアントは Server A に接続される．この様子を図 2 に示す．

3.3.2 負荷分散構成

2 台のサーバを負荷分散構成とするため，MA=MB=1 とする．仮想 IP アドレスについて，Client A では Server A のメトリックが 2 に，Server B のメトリックが 3 になるため，サイト A の Client は Server A に接続される．同様に，サイト B の Client B は Server B に接続される．Server A で障害が発生した場合には，Server A の経路情報がフラッシュされることにより，Client A は Server B に接続される．Server B で障害が発生した場合も同様に，Client A も Client B も Server A に接続される．この様子を図 3 に示す．

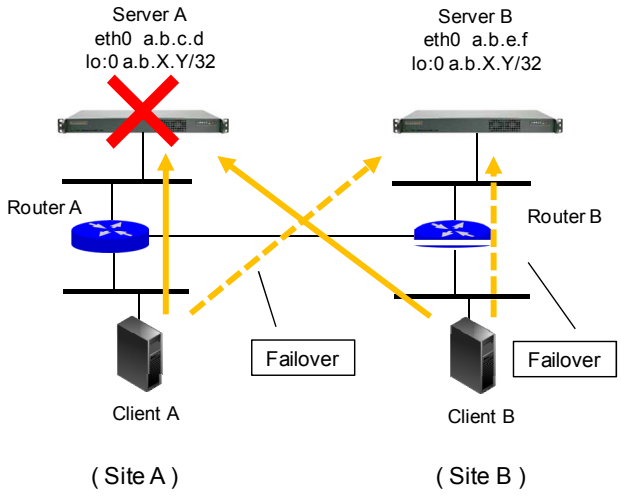

図 2　フェイルオーバ構成の動作

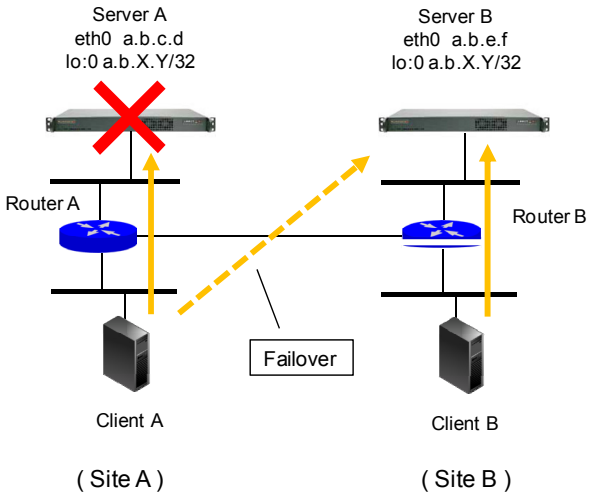

図 3　負荷分散構成の動作

3.4 サーバの死活監視

本提案による冗長化構成では，サーバあるいはルーティングデーモンが停止したときには，経路情報が書き換わってバックアップとなるサーバに自動的にフェイルオーバするが，サービスのプロセスだけが停止した場合には，自動的にフェイルオーバしない．このため，障害を検知して直ちに経路情報を更新し，フェイルオーバさせるための機能を運用する．これにはいくつかの方法が考えられる．

3.4.1 死活監視用に別システムを運用する

別に運用する監視用のシステムでフェイルスチェックを行い，障害が発生したときには，ルータやサーバに指示を出してフェイルオーバする．この方法では，ネットワーク

やサーバの状態を総合的に管理できる利点があるが，SPOF になる可能性があり，監視用システムの冗長化なども検討する必要がある．

3.4.2 冗長化するサーバ自身が死活監視を行う

別の監視用システムを利用することなく，サーバ自身がフェイルスチェックを行う．この場合には，さらに，自分自身のフェイルスチェックを行い，障害時には自分への経路情報の優先度を下げる方法と，他のサーバのフェイルスチェックを行い，障害時には自分への経路情報の優先度を上げる方法がある．但し，自分の優先度を上げる方法では，特に多重度が大きい場合には負荷が集中する場合のあること，サーバの動作に異常がある場合にはブラックホールとなって通信障害を起こすなどの理由から注意が必要である．これらの方法は，前述のような総合的な管理機能が必要でない場合に有効であり，SPOF や通信経路のボトルネックの問題を回避することができると同時に，シンプルな冗長システムとして運用することが可能である．

なお，サーバあるいはルーティングデーモンが停止した場合のフェイルオーバまでの時間は，ルータのルーティングプロトコルのパラメータに基づいている．RIP[8]の場合には，Update=30 秒，Invalid=180 秒，Hold down=180 秒，Flush=240 秒である．サーバやルーティングデーモンが停止した場合に直ちに経路情報を変更し，フェイルオーバ時間を短縮させる場合にも，死活監視によって経路情報を強制変更する機能を運用する．ルーティングプロトコルのパラメータを変更することも考えられるが，ネットワーク全体に影響が及ぶため，注意が必要である．

4. システムの実装

4.1 LDAP サーバの冗長化

岡山大学では統合認証システムのサービスを行っており，津島キャンパスと鹿田キャンパスに 2 台の LDAP サーバを運用している．ここでは，津島キャンパスのものを LDAP Server A，鹿田キャンパスのものを LDAP Server B とする．2 台の LDAP サーバはレプリケーションによって同じ情報を有する複製サーバとなっている．これらの LDAP サーバは本学の教職員，学生，来訪者の一時アカウントなど全てのユーザのアカウント情報を有しており，ロケーションフリーネットワークシステム[9]の 2 台の Radius[10]サーバ，教務システムの 2 台のサーバ及び 16 台の PC 端末から認証サーバとして利用されている．LDAP クライアントであるこれらのサーバや PC 端末では，運用しているシステム構成により，LDAP サーバとして設定できるホストが 1 台に限られる問題があった．このため，通常では各 LDAP クライアントは LDAP Server A を参照しているが，その障害時には各 LDAP クライアントの設定を手作業で変更して LDAP Server B を参照させる必要があった．

そこで，2 台の LDAP サーバについて，IP anycast を構成し，フェイルオーバとしてシステムを実装した．通常では，津島キャンパスの LDAP クライアントは LDAP Server A を，鹿田キャンパスの LDAP クライアントは LDAP Server B にアクセスしているが，どちらかの LDAP サーバが障害の場合，例えば LDAP Server A で障害が発生すると，津島キャンパスの LDAP クライアントでは LDAP Server B にアクセス先が自動的に変更される．なお，LDAP サーバは，NEC 社の Enterprise Directory Server[11]（以下，EDS とする）を用いており，OS は RedHat Enterprise Linux 5 [12]である．また，津島キャンパス及び鹿田キャンパスの L3-SW は，Alaxala 社の AX6708S[13]であり，ダークファイバによって 10GbE×2 回線で接続されている． LDAP サーバの冗長化構成を図 4 に示す．

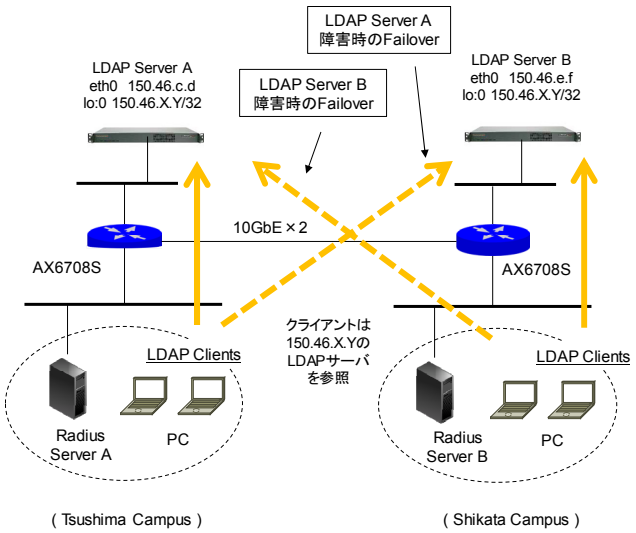

図 4　LDAP サーバの冗長構成

具体的な構成としては，まず，津島キャンパスの LDAP Server A では，ホストの IP アドレス(eth0=150.46.c.d/24)の他に，IP alias でループバックインターフェースに仮想 IP アドレス(lo:0=150.46.X.Y/32) を設定した．鹿田キャンパスの LDAP Server B では，ホストの IP アドレス(eth0=150.46.e.f/24)の他に，ループバックインターフェースに仮想 IP アドレス(lo:0=150.46.X.Y/32) を設定した．また，各サーバではルーティングデーモンとして，quagga-0.98[14]を使用した．岡山大学では，支線部のルーティングプロトコルには，RIP version2 を使用しているため，quagga では ripd を使用した．現在の構成では，2 台の LDAP サーバのデフォルトのメトリック値を 3 にしている．これは，目的の冗長化が負荷分散であること，死活監視が 3.4.2 節の自分自身のフェイルスチェックを行い，障害時には自分への経路情報の優先度を下げる方法であること，また，今後の状況によっては死活監視を他のサーバのフェイルスチェックを行い，障害時には自分への優先度を上げる方法に変更することを可能にするためである．

設定した ripd.conf を以下に示す．

- LDAP Server A

```
hostname ldap-a
password PASSWORD
router rip
 network 150.46.0.0/16
 redistribute connected metric 3
```

- LDAP Server B

```
hostname ldap-b
password PASSWORD
router rip
```

```
network 150.46.0.0/16
redistribute connected metric 3
```

また，死活監視の方法は，3.4.2 節に示した“自分自身のフェイルスチェックを行い，障害時には自分への経路情報の優先度を下げる方法”を用いた．

4.2 WEB 認証サーバの冗長化

岡山大学ではキャンパス情報ネットワークにおいて，ロケーションフリーネットワークシステムのサービスを行っている．このサービスは教職員，学生を含む大学の全構成員にそれぞれの所属に応じた VLAN を割り当て，学内のどこで端末を接続しても割り当てられたネットワークに接続できるサービスである．2013 年 4 月現在，岡山大学の構成員と学外からの一時的な来訪者を含めて 2 万人程度に VLAN を割り当てており，その大部分がロケーションフリーネットワークによって学内 LAN を利用している．このネットワークでは，各建物にある約 320 台のフロアスイッチが，WEB 認証，MAC アドレス認証，IEEE802.1x 認証機能を有しているが，ほとんどの利用者は WEB 認証によって認証され，所定の VLAN を割り当てられてネットワークを利用している．

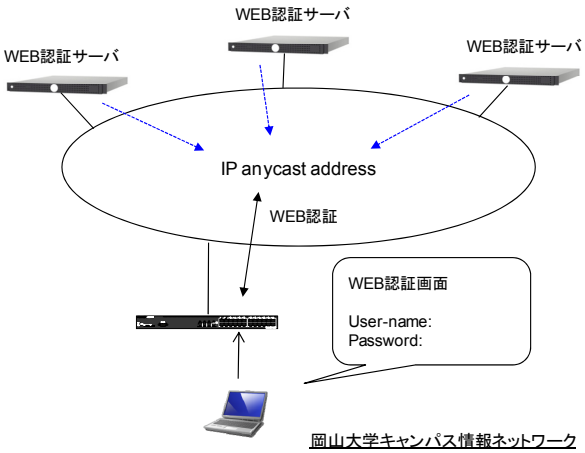

図 5　WEB 認証サーバの冗長化（計画中）

この WEB 認証では，認証のための URI を各フロアスイッチに設定しているが，それに複数の WEB サーバを指定できない．そこで，本論文による構成方法を適用することで，実際には複数の WEB サーバを運用しておき，フロアスイッチには IP anycast による URI を設定することで冗長化することを計画している．

4.3 SSL-VPN 装置の認証サーバ冗長化

岡山大学では大学の全構成員を対象に学外から学内に接続する場合に SSL-VPN 接続サービスを行っている．このサービスでは，F5 社の FirePass4120[15]のアプライアンス装置を使用しているが，ユーザ認証に使用する RADIUS サーバの指定について，複数のサーバを設定はできるが，それらのサーバで障害が発生し，フェイルオーバした後にサーバが復旧してもフェイルバックしないと思われる．そこで，フェイルバックした時にも負荷分散構成を回復させるため，本論文による構成方法を適用し，複数の RADIUS サーバに IP anycast のアドレスを設定して冗長化することを計画している．

5. まとめ

本論文では，組織内において IP anycast を適用し，分散配置されたサーバを冗長化する構成方法を提案した．この方法ではメトリックを変更することで，フェイルオーバ構成でも，負荷分散構成でも運用が可能である．また，障害が発生したサーバが復活すると，クライアントのアクセスは障害発生前のサーバフェイルバックすることで負荷分散構成が回復することを示した．また，死活監視を各サーバ自身が実装することにより，通信経路のボトルネックや SPOF の問題を回避すること，OS の基本機能によって動作することを示した．

さらに，提案方法によって岡山大学の LDAP サーバを構成し，実際の運用からその有効性を確認した．

今後の課題としては，死活監視機能の精度を上げ，サービスがダウンした場合のフェイルオーバ時間を短縮すること，サーバ間でのセッション管理や処理の継続，データ同期などを実装すること，本論文による冗長化構成の適応範囲を広げることなどがあげられる．

参考文献

[1] 大隅淑弘 山井成良, 藤原崇起, 岡山聖彦, 河野圭太, 稗田隆: IP alias と経路制御を用いた複製サーバ冗長化構成, 情報処理学会研究報告. IOT, [インターネットと運用技術] 2012-IOT-18(4)
[2] Droms, R.: Dynamic Host Configuration Protocol, RFC 2131, IETF (1997).
[3] Partridge, C., Milliken, W.: Host Anycasting Service, RFC1546, IETF (1993).
[4] Brisco, T.: DNS Support for Load Balancing, RFC1794, IETF (1995).
[5] Rekhter, Y., Hares, S.: A Border Gateway Protocol 4 (BGP-4), RFC4271, IETF (2006).
[6] Dovecot (online), available from <http://dovecot.org/index.html> (accessed 2013-04-14).
[7] Wahl, M., Howes, T., Kille, S.:Lightweight Directory Access Protocol (v3), RFC2251, IETF (1997)
[8] Malkin, G.: RIP Version 2, RFC2453, IETF (1998).
[9] 大隅淑弘 岡山聖彦, 山井成良, 藤原崇起, 稗田隆: 電子ジャーナルの地理的なサイトライセンス契約条件に適応するロケーションフリーネットワークシステム, 平成 24 年度第 2 回 情報処理学会インターネットと運用技術研究会 (IOT).
[10] Rigney, C.,Willens, S., Rubens, A. and Simpson,W.: Remote Authentication Dial In User Service (RADIUS), RFC 2865, IETF (2000).
[11] 日本電気株式会社: EnterpriseDirectoryServer (online), available from <http://www.nec.co.jp/middle/WebSAM/products/secmaster/eds/> (accessed 2013-04-14).
[12] Red Hat, Inc.: Red Hat Enterprise Linux (online), available from <http://www.redhat.com/products/enterprise-linux/> (accessed 2013-04-14).
[13] アラクサラネットワークス株式会社: AX6700S (online), available from <http://www.alaxala.com/jp/products/AX6700S/index.html> (accessed 2013-04-14).
[14] Quagga Routing Suite (online), available from <http://www.nongnu.org/quagga/> (accessed 2013-04-15).
[15] F5 Networks, Inc.: FirePass (online), available from < http://www.f5networks.co.jp/product/firepass/spec.html> (accessed 2013-04-14).