Vol.2012-IOT-18 No.4
2012/6/28

IP aliasと経路制御を用いた複製サーバ冗長化構成

大隅 淑弘[1,a)]　山井 成良[1]　藤原 崇起[1]　岡山 聖彦[1]　河野 圭太[1]　稗田 隆[1]

概要：我々の社会活動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が重要な課題となっている．計算機システムの冗長化構成では，HA クラスタなどによる構成方法が一般的であるが，分散配置したサーバには適していない．また，通信のボトルネックや単一障害点になりやすいなど，物理的な構成上の制約がある．そこで，本論文では，IP alias と経路情報によって anycast を行い，分散配置した複製サーバを冗長化する構成方法を提案する．この構成方法では，仮想 IP アドレスの経路制御によって冗長化を行うことにより，フェイルオーバ構成でも負荷分散構成でも動作することができる．また，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というメリットもある．提案方法に基づき試作システムで運用した結果，実用的に動作することが確認された．

Redundant Configuration of Replica Servers by IP Alias and IP Routing

Yoshihiro Ohsumi[1,a)]　Nariyoshi Yamai[1]　Takaoki Fujiwara[1]　Kiyohiko Okayama[1]
Keita Kawano[1]　Takashi Hieda[1]

Abstract: Recently, most of social activities depend on a information system, and the service continuity of the information system has become very important. For redundant configuration of the computer system, High-Availability cluster system is generally used. However this system is not suitable for the distributed replica servers since it may become a bottleneck of communication and "Single Point of Failure", and there are also restrictions on the physical configuration. In this paper, we propose a configuration method which makes the distributed replica servers redundant by means of anycast using IP alias and IP routing. In this method, since the redundancy function is performed by IP routing of the identical virtual IP address, it can be configured both failover configuration and load-balancing configuration. This method also have a merit that does not require an equipment such as a load balancer. According to a field testing, we confirmed the proposed system worked effective and practical.

1. はじめに

近年，我々の社会活動は様々な情報システムによって運営されており，これらのシステムによるサービスなくしては成り立たない時代になった．このような社会において情報システムがひとたび停止すると，社会活動に与える影響は大きく，その損害は甚大なものとなることは少なくない．このため，情報システムを止めない工夫が続けられているが，その一つに冗長化があげられる．冗長化の最も基本的な考え方は，予備機の用意や多重化であるが，これに基づいた構成方法においても，多くの技術が開発されている．ネットワークにサービスを提供しているサーバでは，DNS(Domain Name System)[1] やメールサーバのように，あるサービスに対して複数のサーバを定義し，負荷分散をしたり，障害時には接続先が自動的に切り替わったりすることができるものがあるが，サービスによってはそのような冗長化が利用できないものも多い．その場合，従来からの一般的な冗長化の構成方法では，多重化したサーバをデータセンタや計算機室に設置し，障害や負荷によってそれぞれのサーバに処理を振り分ける方式が基本となっている．これは，サーバがある程度集中して設置されている場

[1] 岡山大学情報統括センター
Center of Information Technology and Management, Okayama University
[a)] oosumi@cc.okayama-u.ac.jp

合に適した冗長化の方法であり，最近のサーバシステムの主流になっているクラウドにおいても同様の構成方法となる．しかしながら，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，企業や大学などでいくつかの拠点やキャンパスがあり，同一の機能を持つサーバを分散配置している場合などである．また，災害対策としてサーバを分散配置する必要のある場合などもある．

従来からの冗長化構成方法では，離れて設置されたサーバ間で冗長化を行うものはあまり例がなく，幅広く実用的に利用できる構成方法の開発が必要である．そこで，本研究ではネットワークに着目し，サーバのネットワークインタフェースに IP alias を用いることで anycast[2] を構成し，分散して配置されたサーバを冗長化する構成方法を提案する．この構成方法では，経路情報を変更することにより，フェイルオーバ構成でも負荷分散構成でも動作することができる．また，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が必要ないという利点もある．

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試作したシステムの実装と動作試験の結果について述べる．

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

従来からの冗長化構成方法では，デュアルシステム，デュプレックスシステム，フォールトトレラントシステムなどの構成方法が知られている．また，簡易な方法では,DNS による Round Robin[3] などもある。近年では，情報通信機器の高機能化やクラウドの普及などにより，実装方法として，HA（High Availability）クラスタ構成による，フェイルオーバクラスタ，負荷分散クラスタなどがよく利用されている．フェイルオーバクラスタでは，クラスタリングしたサーバ間で死活監視を行うことにより，障害時にはフェイルオーバしてサービスを継続する．負荷分散クラスタでは，一般にはロードバランサを用いる．ロードバランサの仮想 IP アドレスへのサービス要求を NAT(Network Address Translation)[4] により実サーバに振り分け，実サーバからのレスポンスもロードバランサで NAT 変換してクライアントに送る NAT 方式や，ロードバランサで宛先パケットを実サーバの MAC アドレスに変換して中継し，実サーバからのレスポンスは直接クライアントに送る DSR(Direct Server Return) 方式 [5] がある．

2.2 従来システムの問題点

従来の多重化や HA クラスタ構成などでは，多くの場合運用管理などの理由から，冗長化されたサーバをデータセンタ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．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．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．また，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．

このように従来からの実装方法では，物理的な構成上の制約があったり，冗長化するサーバが遠隔拠点に分散して設置されていたりする場合には適していない．

3. IP alias と経路制御による冗長化構成

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場合があり，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，IP alias と経路制御による冗長化構成を行う．

この方法では，冗長化する各サーバで共通の IP アドレス (以下，仮想 IP アドレスとする) を設定し，各サーバが仮想 IP アドレスの経路情報を広告することにより anycast を構成する．クライアントはネットワーク的に最も近いサーバに接続することで冗長化を行う．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，複数の拠点がルータで接続されており，いくつかの拠点にサーバが設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，組織のネットワーク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．

なお，本提案方法は anycast の応用的な構成方法である．anycast は負荷分散や冗長化のために利用される通信方式である．ネットワーク上の様々な場所から同じ IP アドレス範囲を同時に広告することで，その範囲のアドレスを宛先とするパケットは、そのアドレスを広告した最も近い場所にルーティングされる．anycast は一般には，DNS のルートサーバなどのように，インターネット上の様々な場所か

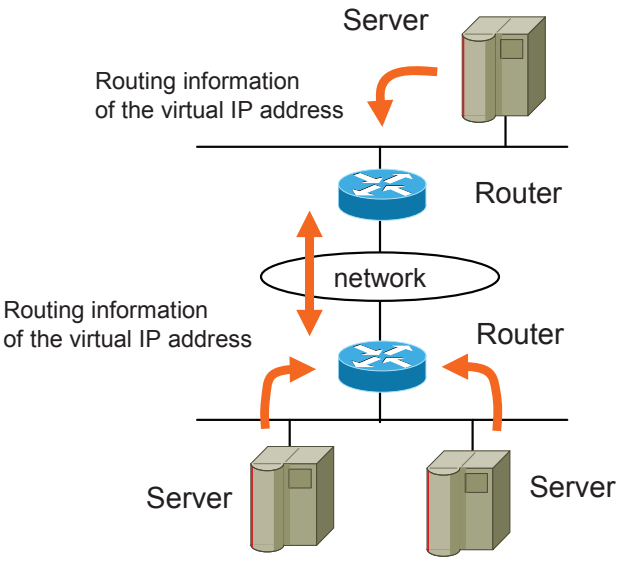

図 1 システムの構成例
Fig. 1 System configuration example.

らの接続を冗長化するような，比較的大規模な範囲に利用されている．しかしながら，本論文の前提条件にあるような組織内での冗長化を想定したものはあまり例がない．

本提案方法は 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一切必要なく，また，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という利点がある．さらに，負荷分散構成では DDoS 攻撃 (Distributed Denial of Service attack) の効果を低減させるなどの anycast の利点も有している．

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

物理的な構成要件としては，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サーバが接続されていることであるが，構成方法について以下に説明をする．

まず，IP alias で各サーバのループバックインタフェースに，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 アドレスを経路情報として，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．ループバックインタフェースを用いるのは，仮想 IP アドレスを経路情報に載せるためであり，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ときの対策を行う場合や，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 台のサーバを冗長構成する例を**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．また，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．

(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．

3.4 サーバの死活監視機能

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．ひとつは，一般的によく用いられるように，別に運用する監視用のシステムで死活監視を行い，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方法である．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を検討する必要がある．

もうひとつは，別の監視用システムを利用することなく，サーバ自身が死活監視を行う方法である．この場合には，さらに，自分自身の死活監視を行い，障害時には自分への経路情報の優先度を下げる方法と，他のサーバの死活監視を行い，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は，サーバの台数が増えると正常なサーバ間で負荷分散がうまく行われない危険性があり，また応答を返さないサーバから誤って

優先度の高い経路情報を広告すると全体のサービスが停止する可能性もあるため注意が必要である．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．

また，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 の場合には，通常は 30 秒間隔で経路情報が広告されるが，経路情報の広告が停止してさらに 3 分間が経過しないと経路情報は更新されない．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的に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. 試作システムの実装

4.1 システムの冗長化構成

前章で述べた提案方法に基づき，我々は試作システムの実装を行った．SMTP[6] サーバとして，CentOS5[7] 上で Postfix[8] を運用し，岡山大学の津島キャンパスと鹿田キャンパスにそれぞれ設置した．なお，両キャンパスの L3-SW は，アラクサラネットワークス社の AX6708S[9] 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．津島キャンパスの SMTP サーバ (以下，SMTP サーバ A とする) では，自分の IP アドレス (eth0=150.46.a.b/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.E.F/32) を設定した．鹿田キャンパスの SMTP サーバ (以下，SMTP サーバ B とする) でも，自分の 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フェースに仮想 IP アドレス (lo:0=150.46.E.F/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[10] 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[11] を使用しているため，quagga では ripd を使用した．設定した ripd.conf を以下に示す．

- SMTP サーバ A

```
hostname  smtp-a
password  PASSWORD
router rip
 network 150.46.0.0/16
 redistribute connected metric MA
```

- SMTP サーバ B

```
hostname  smtp-b
password  PASSWORD
router rip
 network 150.46.0.0/16
 redistribute connected metric MB
```

但し，MA，MB は定数

以下にフェイルオーバ構成で運用する場合と，負荷分散構成で運用する場合の構成手順を示す．

4.1.1 フェイルオーバ構成

ここでは SMTP サーバ B をスタンバイとするため，MA=1，MB=3 とする．これにより，津島キャンパスでの仮想 IP アドレスのメトリックが，SMTP サーバ A では 2 に，SMTP サーバ B では 5 になる．また，鹿田キャンパスでの仮想 IP アドレスのメトリックは，SMTP サーバ A では 3 に，SMTP サーバ B では 4 になり，通常状態では，両キャンパスのクライアントからは，SMTP サーバ A の経路が選択される．ここで，例えば SMTP サーバ A で障害が発生した場合には，経路情報が書き換わることで，両キャンパスのクライアントは SMTP サーバ B に接続される．この様子を**図 2** に示す．SMTP サーバ B で障害が発生した場合にも同様の振る舞いにより，クライアントは SMTP サーバ A に接続されるようになる．

4.1.2 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，津島キャンパスでは SMTP サーバ A のメトリックが 2 に，SMTP サーバ B のメトリックが 3 になり，津島キャンパスのクライアントからは，SMTP サーバ A の経路が選択される．同様に，鹿田キャンパスでは，SMTP サーバ A のメトリックが 3 に，SMTP サーバ B のメトリックが 2 になり，鹿田キャンパスのクライアントからは，SMTP サーバ B の経路が選択される．ここで，SMTP サーバ A で障害が発生した場合には，経路情報が書き換わることで，両キャンパスのクライアントは SMTP サーバ B に接続される．この様子を**図 3** に示す．SMTP サーバ B で障害が発生した場合にも同様の振る舞いにより，クライアントは SMTP サーバ A に接続されるようになる．

4.2 サーバの死活監視

試作システムでは，まず，サーバ自身が自分の死活監視を行い，障害時には経路情報を変更して自分への優先度を下げる構成方法で実装した．各サーバが自分の postfix プロセスについて死活監視を行い，障害を検知すると，直ちに自分の仮想 IP アドレスについて，メトリックを 16 に変更する．RIP ではメトリックが 15 を越えたものは，到達不能とみなされるため，障害が発生したサーバの経路情報がフラッシュされ，クライアントからの接続は相手サーバに向けられる．

次に，各サーバが相手のサーバの死活監視を行い，障害時には経路情報を変更して自分への優先度を上げる構成方法で実装した．この場合には，まず，4.1 節で示した MA，

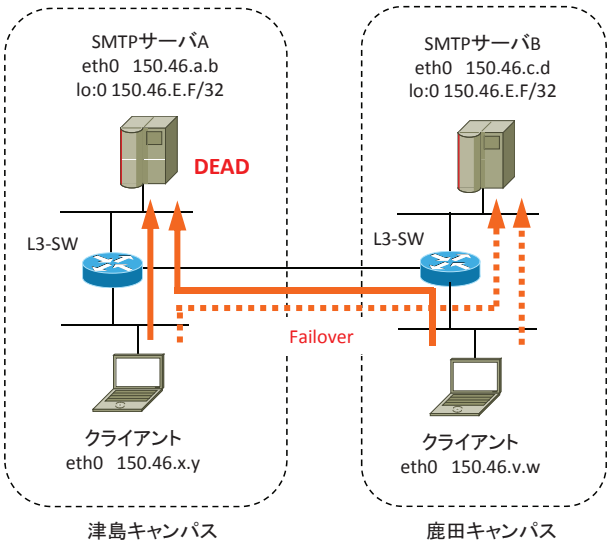

図 2 フェイルオーバ構成
Fig. 2 System configuration for failover.

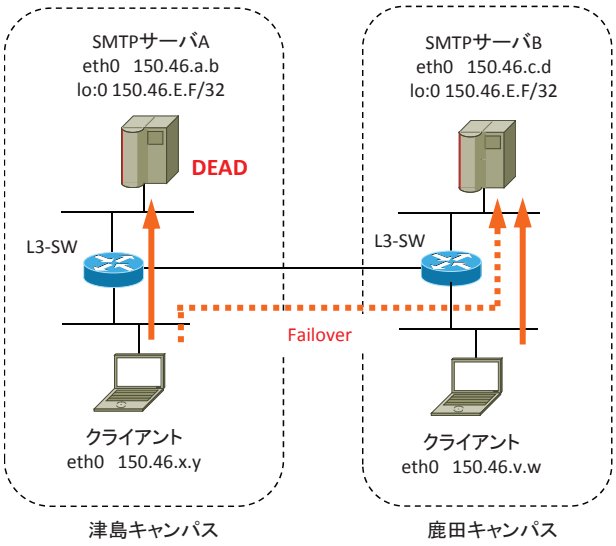

図 3 負荷分散構成
Fig. 3 System configuration for load balancing.

MB の値について，いずれも設定値に 2 を加算したものを適用しておくことで，両キャンパスにおいて各サーバの仮想 IP アドレスへのメトリックの最小値が 4 になるようにしておく．死活監視は相手サーバの実アドレスに対して実行する．相手サーバの障害を検知した場合には，自分の仮想 IP アドレスのメトリックを 1 にすることにより，両キャンパスでのメトリックが最大で 3 になり，クライアントからの接続が自分に対して行われるようになる．

試作システムでは，cron によって死活監視を動作させたため，1 分間隔での死活監視となる．

4.3 動作試験

クライアントを津島キャンパスに接続して以下の動作試験を行い，提案する冗長化構成が有効に動作することを確認した．

4.3.1 サーバが自分の死活監視を行う場合

フェイルオーバ構成では，まず，クライアントが SMTP サーバ A に接続されることを確認後，SMTP サーバ A について postfix を停止したところ，L3-SW から SMTP サーバ A の仮想 IP アドレスの経路情報がフラッシュされ，クライアントが SMTP サーバ B に接続されることを確認した．

次に，負荷分散構成では，まず，クライアントが SMTP サーバ A に接続されることを確認後，SMTP サーバ A について，postfix を停止したところ，同様にクライアントが SMTP サーバ B に接続されることを確認した．

4.3.2 サーバが相手の死活監視を行う場合

フェイルオーバ構成および負荷分散構成について，クライアントが SMTP サーバ A に接続されることを確認後，SMTP サーバ A について postfix を停止したところ，SMTP サーバ B の仮想 IP アドレスの経路情報が変更され，クライアントが SMTP サーバ B に接続されることを確認した．

4.3.3 サーバの停止や ripd が停止した場合

フェイルオーバ構成および負荷分散構成について，クライアントが SMTP サーバ A に接続されることを確認後，SMTP サーバ A についてサーバをシャットダウンしたり，ripd を停止したりしたところ，クライアントが SMTP サーバ B に接続されることを確認した．

5. まとめ

本論文では，分散して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ができることを示した．さらに，OS の基本機能によって動作することを示し，試作システムの実装および動作試験によってその有効性を確認した．

今後の課題としては，サーバの負荷状態によって動的に負荷分散することや，フェイルオーバしたときにサーバ間のセッション管理や処理の継続ができるように，本提案方法を拡張することがあげられる．

参考文献

[1] Mockapetris, P.: DOMAIN NAMES - CONCEPTS AND FACILITIES, RFC1034, IETF (1987).
[2] Partridge, C., Milliken, W.: Host Anycasting Service, RFC1546, IETF (1993).
[3] Brisco, T.: DNS Support for Load Balancing, RFC1794, IETF (1995).
[4] Egevang, K. and Francis, P.: The IP Network Address Translator (NAT), RFC1631, IETF (1994).
[5] Bourke, T.: Server Load Balancing, O'Reilly (2001)
[6] Klensin, J.: Simple Mail Transfer Protocol, RFC5321,

IETF (2008).
[7] The CentOS Project: CentOS-5 (online), available from ⟨https://www.centos.org/⟩ (accessed 2012-06-05).
[8] Wietse Venema: The Postfix Home Page (online), available from ⟨http://www.postfix.org/⟩ (accessed 2012-06-05).
[9] 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 2012-06-05).
[10] Quagga Routing Suite (online), available from ⟨http://www.nongnu.org/quagga/⟩ (accessed 2012-06-05).
[11] Malkin, G.: RIP Version 2, RFC2453, IETF (1998).