BGP情報を用いたアプリケーションレベルのマルチホームを実現するシステムの構築

八代一浩[1], 大西康雄[1], 二戸麻砂彦[1], 中俣賢司[2],
笹本正樹[3], 岡裕人[4], 林英輔[5]

山梨県立女子短期大学[1], 山梨大学[2], ニスカ株式会社[3],
(株) ウインテックコミュニケーションズ[4], 麗澤大学[5]

概要

2002年度山梨県立女子短期大学インターネット接続システムの更新に伴い，インターネット接続システムをサーバ系とアクセス系の2つに分離を行った．サーバ系のシステムはISP(Internet ServiceProvider)のハウジングサービスを利用し，アクセス系システムはSINET[9]に接続を行った．そして，この両システムを地域IX(Internet eXchange)を介して高速に接続した．この環境において，http(Hyper Text Transfer Protocol)を停止させないシステムとして，BGP情報とトランスペアレント（透過的）プロキシーサーバを用いたシステムを構築した．このシステムを利用することによって，利用者が意識せずにアプリケーションレベルで二重化することができるシステムを構築することができる．本稿では，システムの設計および設計を検証するための実験について報告する．

A multihomed system in application layer with BGP information.

Kazuhiro YATSUSHIRO[1], Yasuo Ohnishi[1], Masahiko Nito[1], Satoshi Nakamata[2], Masaki Sasamoto[3], Hirohito Oka[4], Eisuke Hayashi[5]

Yamanashi Women's Junior Collge[1], Yamanashi Univ.[2], NISCA Corp.[3], Wingtechnology Communications Inc.[4], Reitaku Univ.[5]

Abstract

Yamanashi Women's Junior College Internet connecting system have updated in 2002. The system has divided two sub systems: (1) Server system locates on ISP housing service. (2) Access system connects to SINET. These systems are connecting with high-speed network over a regional IX.

We designed the new http backup system, which use transparent proxy servers over the Internet conecting system. BGP information is used to change the direction of http stream.

In this paper, we introduce the system and evaluate it with experiments.

1 はじめに

山梨県立女子短期大学は国文科，幼児教育科，生活科学科，国際教養科からなる学生数400名，教職員数60名の文科系短期大学である．インターネット接続システム[2]は学内ネットワークシステム(KAINS[1])とインターネットを接続するためのシステムである．

2002年度におけるインターネット接続システムの更新では，全体をサーバ系システムとアクセス系システムの2つに分離した．そして両システムを地域IXを経由した高速ネットワークを使って接続する「分散型ネットワークシステム」を構築した．サーバ系システムはISP内に配置することにより，24時間365日停止しない環境が作れる．また，サーバの運用をISPに依頼することにより，セキュリティの管理についても専任のスタッフに任せることができる．他方でアクセス系システムはサーバがないために，セキュリティについては最小限の努力で運用が行える．しかも，間にFireWallシステムやNATシステムを入れる必要がないので，End to Endでインターネットを利用することが可能なシステムである．

このシステムはサーバ系のデフォルトルートはハウジングを行うISPの経路になり，アクセス系はSINET経由になる．つまり，インターネットに対して2つの出口を持つことになる．この点に注目し，本研究ではアクセス系の出口が停止しても，サーバ系の出口からインターネットにアクセスできるシステムの構築を行う．システムで扱うプロトコルはhttpを対象とし，アプリケーションレベルでのマルチホームを実現するシステムを構築した．

本稿では，2002年度に更新したシステムの概要について説明する．次に，アプリケーションレベルのマルチホームを実現するシステムの設計について説明する．そして，このシステムを検証する目的で行った実験および評価について説明をする．

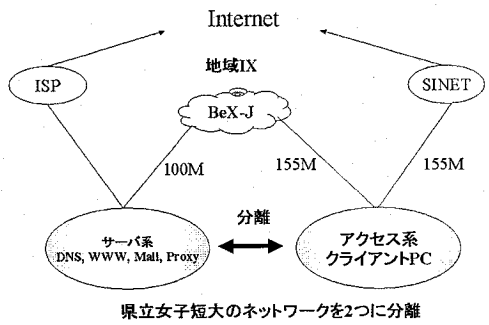

図1: インターネット接続システム

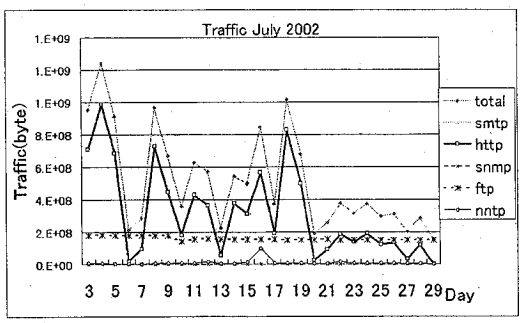

図2: FireWallを流れるトラフィック

2 インターネット接続システム

2.1 システム概要

2002年度に更新したインターネット接続システムを図1に示す．サーバ系とアクセス系のネットワークを分離し，両者の間を地域IX(BeX-J)[3]を介して高速に接続している．

2.2 利用状況

本システムの利用状況を調査するために，FireWallシステムを流れるトラフィックを2002年7月3日から7月29日まで観測を行った．観測にはtcpdump[10]を用い，対外セグメントを構成するswitching hubでFireWallシステムに接続したポートを計測した．結果を図2に示す．

totalとhttpがほぼ連動しており，httpのバケットが全体に大きく影響している．ftpが毎日ほぼ同じトラフィックを運んでいるが，これは深夜に外部セグメントに配置してあるサーバのバックアップをftpを利用して行うためである．

トラフィックの内訳を図3に示す．これらの状況から，本学の利用状況はhttp, ftpが利用の99%を占めていることがわかる．すなわち，利用者へのサービスという観点からはhttpの運用が最も重要であり，停止させないことが重要な課題である．

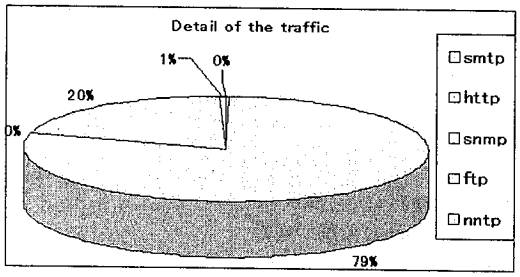

図 3: トラフィックの内訳

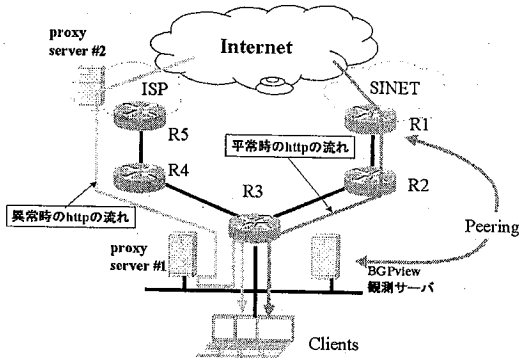

図 4: システムの概要

3 アプリケーションレベルでのマルチホーム化

本研究では，障害時にアプリケーションレベルで迂回路を形成することを目的としているが，これはマルチホームの特殊なケースと考えることができる．一般的には，マルチホーム化を行う手法として，以下の 4 つの手法が知られている．

1. AS 番号を取得してネットワークレイヤでマルチホームを実現する方法 [8]
2. NAT(Network Address Translation) を利用する方法 [6][5]
3. 負荷分散装置を利用する方法
4. ALG(Application Level Gateway) を利用する方法 [4]

1. の方法は技術的には最も有効な方法であるが，現実的には，エンドユーザがマルチホームを行うためのアドレス空間を入手することが不可能であり，実現ができない．2. の方法では，NAT を行うために，アクセス系システム（本学に設置しているシステム）の出口付近に ISP のアドレスを割り振る必要がある．しかしながら，ISP と本学のシステムの間には public な地域 IX があるため，この状況を作ることが困難である．3. の方法は 2. と同じ理由で困難であることに加えて，小さな組織においては，機器が高価である．

このような理由から我々は 4. の ALG を利用する手法を用いて設計および実装を行う．ALG へのトラフィック誘導には透過的な手法を用いる．これはデフォルトルータにおいて，組織内から組織外への http アクセスをすべてプロキシーサーバに行わせる方法である．この手法自体は，主にキャッシュサーバを利用して，バックボーントラフィックの要求を軽減させる目的，もしくは，キャッシュされた内容を共有することを目的として使われている．本システムでは，この手法を地域 IX を介したバックアップシステムとして機能させる．

4 システムの設計

4.1 システムの概要

本システムでは，ALG を用いたシステムを利用し，インターネットへの接続に障害が検知された場合に，アプリケーションレベルで迂回路を形成し，利用者にサービスを提供できるシステムの構築を行う．図 4 にシステムの概要を示す．

R1 は SINET が所有するルータ，R2,R3,R4 は山梨県立女子短大の所有するルータ，R5 は ISP が所有するルータである．平常時に http は SINET を経由している．この経路の接続性を観測するためのサーバを本学内に配置する（観測サーバ）．観測サーバは SINET のルータ (R1) と Multihop BGP を用いてピアリングする．観測サーバによって，経路の異常が観測されると，観測サーバは R3 の設定を変更し，平常時には利用されていない Proxy Server #1 を透過的プロキシーサーバとして機能させるように

する．Proxy Server #1 は ISP 内に配置されている Proxy Server #2 と連携し，http の流れを ISP 経由となるように機能する．これら一連の動きにおいて，クライアントには何ら設定の変更は必要ない．

4.2 インターネット接続断検知アルゴリズム

一般的にインターネットへの接続口が 1 つしかない場合には，デフォルト経路を静的に設定し運用を行う．この方法だと，インターネットへの接続性を確認することは非常に困難である．一方，経路情報プロトコルを利用していると，インターネットへの接続断があると，経路情報が極端に少なくなり，接続断を検知することができる．現在のインターネットでは ISP 同士の経路情報交換には BGP4[7] が用いられている．そこで，本研究においても BGP4 を用いて上位の ISP (本研究の場合は SINET) と経路情報を交換することを行う．ここで注意するのは，AS 番号を用いて複数の ISP と接続を行うマルチホーム化とは違い，上位の ISP とのみピアリングを行う点である．つまり，上位 ISP とは private AS を用いた運用を行うことができる．また，観測の目的でピアリングを行っているため，BGP の情報は経路制御には影響を与えない．

BGP4 はコネクション指向の経路情報交換プロトコルである．一度セッションが作られ，最初の経路情報が交換されると，後は定期的な keepalive メッセージの交換や，経路情報が変更された際に update メッセージの交換等が行われる．本研究では，切断を関知するための情報として，2 つの状況を想定する．一つは，ピアリングを行っているルータ間のリンクがダウンする場合であり，もう一つはピアリングを行っているルータの上位ルータに障害が発生している場合である．一つ目の状況では，keepalive メッセージが届かなくなり，holdtime が expire して切断を関知できる．二つ目の状況では，大量の withdraw を伴う update メッセージを受け取ることによって切断を関知できる．

障害回復時には，リンクダウンによる切断の場合には BGP のセッションが再び established されることにより関知できる．また，上位ルータの障害の場合は大量の update メッセージが届くことによって関知できる．

障害検出のアルゴリズムを以下にまとめる．

切断の検知

```
if (keepalive が届かない)
      peer との間が切断
else if (大量の withdraw メッセージが届いた)
      上位ルータで障害
else
      問題なし
```

復旧

```
while(障害)
{
    if (peer との間が切断した)
       established 状態になったら復旧
    else if (上位ルータで障害)
       大量の update があったら復旧
}
```

5 実験

5.1 実験環境

本システムを評価するため図 5 に示すシステムで実験を行った．実験では，経路障害を検知してから，実際に http の経路が変更されるまでの時間を計測し，どの程度の時間で http の経路が変更されるかを測定した．bgpview の設定で，keepalive の間隔を 30 秒とし，holdtime を 60 秒とした．

5.2 実装

実装は以下の通りである．

- R1: cisco 2514
- R2: zebra 0.93a (Vine Linux 2.6)
- 観測サーバ: bgpview alpha 0.34 (FreeBSD 4.7 Release)
- Proxy Server: squid 2.4.STABLE7-0vl1 (Vine Linux 2.6)

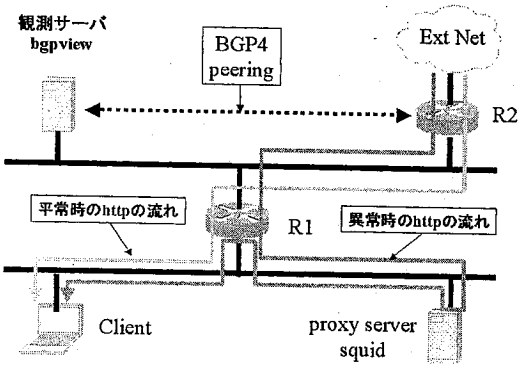

図 5: 実験環境

また，以下にハードウェア環境を示す．

- R2: Pentium3 500MHz 128MB 100BaseTX
- 観測サーバ: Pentium Pro 200MHz 64MB 100BaseTX
- Proxy Server: Pentium3 500MHz 256MB 100BaseTX

観測サーバは R2 とピアリングを行い，接続性を観測している．接続断が生じた際には，観測サーバからシェルスクリプトを用い，expect を利用して，R1 に telnet 接続をし，tftp サーバに保存されている，経路が切断されたときに用いる設定ファイルをダウンロードし，R1 の running-conf を変更することによって経路の変更をさせる．この実験では，接続断の発生を，R2 上で実行されている bgpd プロセスを終了させることによって再現した．なお，この方法では，bgpd を終了させた瞬間に，bgpview は接続断を発見する．また，実行時間をはかるために，シェルスクリプト開始時に time コマンドを用いることによって計測した．

以下に expect を利用するためのシェルスクリプトを示す．

```
#!/usr/local/bin/expect

spawn telnet 133.23.140.***
expect -ex "Password:"
send "*****\r"
expect -ex "Router>"
send "en\r"
expect -ex "Password:"
send "*****\r"
expect -ex "Router#"
send "copy tftp running-conf\r"
expect -ex "]?"
send "133.23.140.***\r"
expect -ex "]?"
send "normal.conf\r"
expect -ex "]?"
send "\r"
expect -ex "Router#"
send "exit\r"

expect eof
```

5.3 実験結果

測定結果を示す．表 1 は，tftp サーバから設定ファイル (1138byte) をダウンロードし，running-conf を書き変える時間を 10 回測定した値とその平均を示している．左から順に，回数，expect の実行にかかった時間，tftp サーバからのダウンロードにかかった時間である．

5.4 考察

経路切断を検知してから，expect の実行を完了するまでの時間を見ると，平均 10.71 秒を示している．この結果より，充分に速い時間で経路先を変更できることを意味している．また，各実行結果は何度試してもあまり差が生じなかった．これは，実験に利用したネットワーク上には，他のパケットが流れておらず，かなり安定していたことが影響している．実際のシステムに組み込む場合には，利用者の http パケットなどが流れることも充分予想されるが，大量のパケットが流れない限りは，大幅な遅延はないと思われる．また，expect の実行を完了するまでの時

表 1: 測定結果

回数	expect(秒)	tftp(秒)
1	10.84	5.876
2	10.65	5.622
3	10.68	5.670
4	10.70	5.701
5	11.03	5.972
6	10.72	5.699
7	10.58	5.595
8	10.77	5.784
9	10.62	5.591
10	10.49	5.510
平均	10.71	5.702

間のうち，tftp での running-conf の転送にかかる時間の平均が 5.702 秒と，全体の結果に対する割合は約 53%とかなりの割合を占めている．このため，よりよい結果を得るためには，単純にサーバのパフォーマンスを上げる方法以外に，転送する running-conf ファイルのサイズを抑えることや，転送率の向上が効果的である．

6 おわりに

本稿では，BGP 情報を利用して経路障害を検知し，検知後ただちに，透過的プロキシーサーバを用いて，http の経路を変更するシステムの設計を行った．設計したモデルにしたがい，実験環境を構築し，システムの評価を行った．その結果，実験環境で，経路障害を検知してから約 10 秒後には経路変更を行うことができた．今後の課題として，

- 実システムへの導入
- より短い時間での経路の変更
- 積極的な http 経路選択システムの導入

などがあげられる．また，この実験では，expect を実行するためのシェルスクリプトにルータのパスワードを直接書き込んでしまっているため，セキュリティ面で若干の不安をかかえている点も課題である．

参考文献

[1] 八代一浩，大西康雄，二戸麻砂彦:"文化系短期大学における教育計算機環境の構築と運用"，分散システム／インターネット運用技術研究報告，情報処理学会， Vol.2001-DSM-21, No. 50, pp. 1-6(2001).

[2] 八代一浩，大西康雄，二戸麻砂彦，笹本正樹，岡裕人: "地域 IX を利用した分散型大学ネットワークの運営", 分散システム／インターネット運用技術研究報告，情報処理学会，Vol.2002-DSM-28, pp. 37-42(2002)

[3] 八代一浩, 林英輔:"MAN 技術を用いた地域商用 IX の構築"，情報処理学会論文誌 Vol.42, No.12, pp.2909-2915(2001)

[4] 中川郁夫，上谷一，鍋島公章，樋地正浩，今野幸典: "マルチホーム環境におけるアプリケーションルーティング技術の提案", 分散システム／インターネット運用技術研究報告，情報処理学会，Vol. 1998-DSM-12, pp. 37-42(1998)

[5] 岡山聖彦，山井成良，島本裕志，宮下卓也，岡本卓爾: "マルチホームネットワークにおける透過的な動的トラヒック分散"，情報処理学会論文誌 Vol.41, No.12, pp.3255-3264(2000)

[6] 梶田将司，結縁祥治: "NAT による準マルチホーム化技法"，情報処理学会論文誌 Vol.42, No.12, pp.2818-2825(2001)

[7] Y. Rekhter, T. Li: "A Border Gateway Protocol 4", RFC1771, Mar. (1995)

[8] S. Halabi: "Internet Routing Architectures Second Edition", CISCO Press, (2001)

[9] http://www.sinet.ad.jp/

[10] http://www.tcpdump.org/

謝辞

本研究は (株) 日本ネットワークサービスによる研究補助を受けている．(株) 日本ネットワークサービスに深く感謝いたします．