広域分散配置 Web サーバにおける最適サーバ探索システムの検討

荻野 司　　松田 和宏　　須藤 一顕　　針山 欣之　　向阪 正彦
FastNet, Inc.　インターネット事業部

概要

Web サーバの負荷を分散させるために、サーバクラスターを構成しアクセスを分散させる方法や、地理的、ネットワーク的に分散したミラーサーバを配置することが一般的に行われている。しかし、時々刻々変化するサーバ、ネットワーク状態に応じて、クライアントを最適なサーバに導くことは、種々の提案がなされているものの、決定的な解決方法が見いだされていない。

本稿では、広域分散配置されたＷｅｂサーバ群において、動的に変化するサーバ、ネットワーク状態を計測する手段の提案を、また、その計測手段を用いて真に最適なサーバを検出する新たな方式の提案を、さらに、アクセスクライアントを検出した最適なサーバに導くための最適サーバ探索システムの提案をする。

本方式では、経路情報（ＢＧＰ：Border Gateway Protocol）の AS path(Autonomous System)をネットワークの論理的な距離計測手段判断子として用いる。また、各種サーバ、ネットワーク情報計測ツールを用いた結果と併せて、最適な Web サーバを決定するものである。本稿では、日米各々に実証実験用 Web サーバサイトを構築、実際のインターネット上においてプロトタイプシステムをインプリメント、性能評価を実施した結果についても併せて報告する。

Study of an Efficient Server Selection Method for Widely Distributed Web Server Network

Tsukasa Ogino　Kazuhiro Matsuda　Kazuaki Sudo　Yoshiyuki Hariyama　Masahiko Kousaka
FastNet, Inc.

Abstract:

In order to disperse the load on a Web server, generally the server cluster is configured to distribute access requests, or mirror servers are distributed geographically or situated on different networks. However, although there are several proposals for leading clients to the most efficient server according to the constantly changing server and network condition, as yet no definitive solution has been proposed.

In this document, we propose a measurement method for dynamically changing server and network enviroment, a new selecting method to find the most efficient server based on the measurement method, and an efficient server selection system for leading the access clients to the most efficient server among the distributed Web server network.

Under this method, we use AS (Autonomous System) path routing information of BGP (Border Gateway Protocol) as the factors for evaluating the logical distance of the network. We also try to determine the most efficient Web server by using various server/network information measurement tools. Experimental Web server sites have been set up in both Japan and the USA, and a prototype system was implemented on the Internet and its performance was evaluated; results of the experiments and evaluation are included in this report.

1. Introduction

インターネットが急速に普及するなかで、手軽に情報を配信・閲覧する道具として、WorldWideWeb（以降 Web と称する）を用いた情報配信システムが急速に拡大している。最近では、全世界をアクセス可能対象とした Web サーバサイトも出現し、その重要性は一段と高まってきている。このような全世界的に情報配信を行う Web サーバサイトは、全世界から集中してアクセスをされるために、多量なアクセスに対して十分な準備をする必要がある。このような対策として、Web サーバの負荷分散が重要な検討課題として着目されている。Web サーバの負荷を分散させるためには、サイト内に複数の Web サーバでサーバクラスターを構成する事により、集中するアクセスを複数の Web サーバに分散させる方法や、地理的、ネットワーク的に異なる他サイトに、同様のコンテンツを持つミラーサーバを配置することによって、地域的、ネットワーク的にアクセスを分散させるといった方法が一般的に行われており、アクセスするクライアントに対して、最適なサーバに導くための、種々の提案がなされている。しかし、後述のような理由で、このような一般的な手法では、集中するアクセス負荷を的確に分散する事は非常に困難である。それは、以下の理由からである。Web サーバへのアクセスは、地域的、ネットワーク的に均一ではない。また、アクセス時間においてもアクセス分布は変化する。例えば、Web サーバへのアクセスは、地域的、ネットワーク的に均一ではなく、むしろかなり偏ったアクセス分布を示す。全世界的な広域の web サーバがそれで、その地域のローカル時間にアクセス分布は大きく影響される。アクセスされる時間においても、ビジネスタイムと深夜とで、そのアクセス分布は大きく異なる。

本稿では、広域分散配置されたＷｅｂサーバ群において、動的に変化するサーバ、ネットワーク状態を計測する手段の提案を、また、その計測手段を用いて真に最適なサーバを検出する新たな方式の提案を、さらに、アクセスクライアントを検出した最適なサーバに導くための最適サーバ探索システムの提案をする。

また、日米各々に実証実験用 Web サーバサイトを構築、実際のインターネット上においてプロトタイプシステムをインプリメント、性能評価を実施した結果についても併せて報告する。

2. 最適サーバ探索システム

2.1 最適サーバ決定方式の概要

広域分散されたＷｅｂサーバ群から、最適なＷｅｂサーバを決定するにあたり、本提案の方式では、以下の事項を目標として定義した。

1) 同一サイト内において負荷を均一に分散する。
2) 複数に分散配置されたサイト間においても負荷を均一に分散する。
3) Ｗｅｂサーバに別途過度の負荷を要求しない。
4) クライアントからのレスポンスには、高速で対応をする。

上記目的を達成するため、本方式では、以下の情報を収集、分析し、その結果から総合的に最適サーバを決定するシステムを構築した。

- アクセスクライアントとｗｅｂサーバとのネットワーク的距離を経路情報より取得
- 各々のＷｅｂサーバから、クライアントまでのネットワーク状態を計測
- サイト内ネットワーク情報計測
- 各々Ｗｅｂサーバ状態を計測

本方式では、以上の計測結果に基づいて最適なＷｅｂサーバを決定する。

2.2 プロトタイプシステム構成

本システムは４つの機能で構成されている（図 1）。

- Network Status Probe (NS-P)
- Network Status Server (NS Server)
- NS-Agent
- Route Server

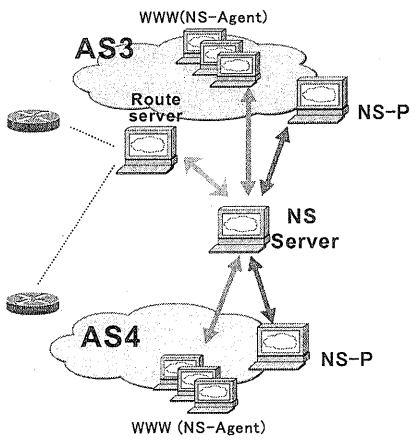

図1 システム構成

2.3 システム機能

(1) Network Status Probe

Network Status Probe（以下 NS-P）は、次の３つの探査機能を持つ。

・ネットワーク情報探査機能
・サイト内ネットワーク情報探査機能
・サーバ状況探査機能

探査機能の実装は、UNIX コマンドと同等な機能を直接組み込んだ(表１-表３)。

表１　ネットワーク情報探査機能一覧

計測項目	測定単位	同等コマンド
RTT	ms	ping
パケットロス	%	ping
スループット	bit/s	(独自)
ルータホップ数	段数	traceroute
ＡＳホップ数	段数	(RS)

表２　サイト内ネットワーク情報探査機能一覧

計測項目	測定単位	同等コマンド
送受信トラフィック	packets	netstat n
パケットエラー数	packets	netstat n
コリジョン発生数	回数	netstat n
ルータトラフィック	bps	snmp
ルータ廃棄パケット数	packets	snmp

表３　サーバ状況探査機能一覧

計測項目	測定単位	同等コマンド
TCP コネクション数	本数	netstat
ディスク負荷	転送回数/s	iostat n
CPU アイドル値	%	iostat n
ロードアベレージ	プロセス数	uptime

(2) Network Status Server

Network Status Server（以下 NS サーバ）は、本システム全体の中枢的な位置にあるシステムであり５つの機能を持っている。

①ネットワーク情報取得機能では、NS-P が探査したネットワーク情報を収集する。②最適サイト決定機能では、Route server(以下 RS)や NS-P から取得した計測データから最適サイトを決定する。③最適サーバ決定機能では、各サイトの NS-P から現在クライアントのアクセスを受けさせるのに最適な状態にあるサーバ（負荷の最も軽いサーバ）を取得しかつ決定する。④データ集約機能では、クライアントの情報を集約し管理している。⑤サイト管理機能では、各サイトの負荷状態を把握している。

(3) NS-Agent

NS-Agent は、Web サーバの Apache に「モジュール」として実装している。モジュール[3]は、Apache の機能を自由に変更できるように提供された仕組みである。NS-Agent は３つの処理を実行する。①クライアントからのアクセスを Apache から受け取る、②最適サーバを知るため NS サーバへ最適サーバ情報を問い合わせる、③HTTP レスポンスでクライアントへ最適サーバを知らせると同時に最適サーバへリダイレクトさせる指示を出す。このとき、HTTP レスポンスコードは 302 (Moved Temporarily[4]) を設定する。このコードを受信したクライアントの Web ブラウザは、指定されたＩＰアドレスのサーバにリダイレクトする。

(4) Route Server

RS は zebra[5] を本システム用に変更して使用した。具体的には、各サイトの Gate way(GW)ルータアドレスを登録できるようにしたのと、クライアントの IP アドレスを受信できるようにした。zebra はクライアントの IP アドレスから、クライアントと各サイトのＧＷルータ間のＡＳホップ数を算出し、各サイトのＧＷルータとＡＳホップ数のリストをコマンド発行元へレスポンスする。

2.4 アルゴリズム

プロトタイプでは最適サーバを２ステップのアルゴリズムで決定している。最初のステップは収集したネットワーク情報から最適サイトを決定する。次に最適サイト内で最も負荷の少ないサーバを探査し最適サーバに決定する。

2.4.1 最適サイト決定

最適サイト決定には、クライアントのアクセスが初回なのかそうでないかによって次の２つの方法をとる。

・ＡＳホップ数で決定する（初回アクセスのとき）
・収集してあるネットワーク情報から決定する（再アクセスのとき）

(1) ＡＳホップ数で決定

NS サーバは、ＡＳホップ数を取得するため、クライアントの IP アドレスをセットしたＡＳホップ数取得コマンドをＲＳへ発行し、レスポンスとして

ＡＳホップ数のリストを得る。リストの中で最小のAS ホップ数になるサイトを最適サイトに決定する。

(2) ネットワーク情報から決定

ここでは、ネットワーク状態値とサイト状態値を使って最適サイトを算出する。まず、ネットワーク情報からネットワーク状態値を算出する。次に、サイト内ネットワーク情報からサイト状態値を算出する。算出されたネットワーク状態値と、サイト状態値を加算した結果がサイトの最適サイト判定値である。

ネットワーク状態値計算式を式(1)に示す。

ネットワーク状態値
=ASL・A＋RTT・B＋RN・C＋PL・D＋TP・E　(1)

- ASL：クライアントまでのＡＳホップ数
- RTT：クライアントまでのRTT 計測値
- RN　：クライアントまでのルータホップ数
- PL　：パケットロス率
- TP　：クライアントまでのスループット
- A-E：重み係数

サイト内ネットワーク状態値式を式（2）に示す。

サイト内ネットワーク状態値
＝ CS・F＋PS・G＋ES・H＋RTR・I＋RTE・J　(2)

- CS　：サイト内で発生したコリジョン数
- PS　：サイト内で発生したパケット数
- ES　：サイト内で発生したパケットエラー数
- RTR：GWルータトラフィック
- RTE：GWルータ廃棄パケット数
- F-J：重み係数

2.4.2 最適サーバ決定

サーバ負荷をリアルアイムに計測し、負荷の少ないサーバを最適サーバに決定する。計測した負荷からサーバ評価値を算出し、各サーバの評価値を比較する。サーバ評価値式を式（3）に示す。

サーバ評価値
＝ LINK・K＋IO・L＋IDLE・M＋CPU・N　(3)

- LINK：TCP コネクション確立数
- IO　：ディスク負荷
- IDEL：CPU アイドル状態
- CPU　：ロードアベレージ
- K-N：重み係数

3.性能評価

本章では、作成したプロトタイプシステムを使って実施した予備実験及び、システムテストの結果について述べる。

3.1 ネットワーク情報の評価

(1) 予備実験の目的および概要

実際のインターネット上のサイトを使ったテストによって、「最適サイト決定」のための「ネットワーク状態値」がどの程度有効性（的中率）を持っているのか調査することを目的とした。本実験では、データ転送時間をネットワーク的な距離として考え、転送時間が短いほど、ネットワーク的には近いと考える。データ転送時間は、Web サーバからクライアントへのデータ転送に要した時間で、クライアント<->Web サーバ（サイト）間のコネクション接続時間を計測して求める。このとき同時にクライアント<->Web サーバ間のネットワーク状態を計測しデータ転送時間とネットワーク状態の相関を調べる。

(2) 計測方法

クライアント<->サイト間におけるデータ転送時間は、tcpdump コマンドで計測し、ネットワーク状態は作成した NS サーバ、NS-P を使って計測する。

実験の手順（図２）は、1) Ｗｅｂページのトップページに画像（イメージファイル）を２つ埋め込む。画像サイズは2,525 バイトである。2) この画像の実体を、片方はＵＳサイトにおき、もう片方はＪＰＮサイトにおく(トップページではリンクする)。3) クライアントがトップページにアクセスすると、画像ファイルはリンクになっているため、クライアントのブラウザはＵＳサイトとＪＰＮサイトへ画像ファイルを取得にいく。4)ここでクライアントと２つのサイト間のコネクション接続時間（図３）を計測する。その結果コネクション接続時間の短い方が、データ転送時間が短かくネットワーク的に近いということになる[6]。

今回、NSS、NS-P によって計測（図４）したネットワーク状態は、クライアントまでのＲＴＴ、スループット、ルータのホップ数、AS パス数である。

コネクション接続時間の計測は、JPN サイト、

USA サイトのそれぞれの Web サーバ上で tcpdump コマンドを実行し、Web アクセスに関するログを収集して行う。

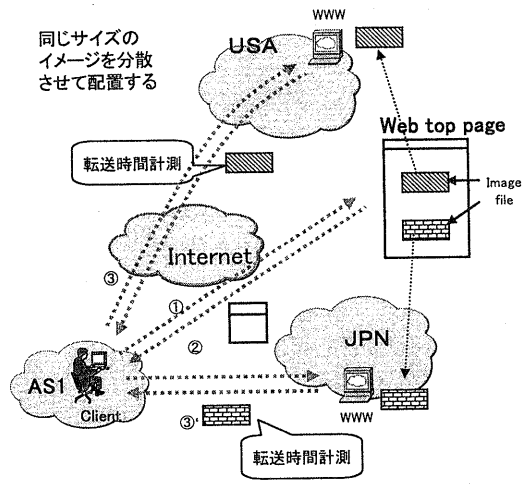

図2　同じサイズのイメージを分散させて配置する

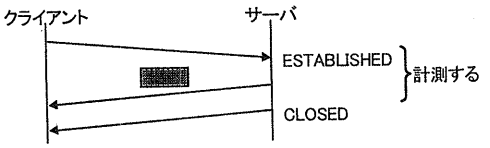

図3　コネクション接続時間を計測する

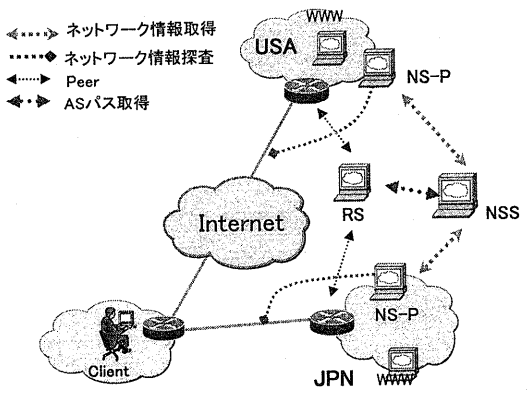

図4　ネットワーク情報探査

今回の予備実験にあたって留意した点は次のとおり。

・広範囲な地域からアクセスされる Web サーバサイトを選択
・日本、米国ともに同一のマシン仕様
・埋め込む２つの画像サイズは同じ大きさ
・時間帯、曜日のネットワークパターンを調査できるように任意の１週間を連続してデータを取得
・各サーバの内部ＬＡＮにおける影響が無い

(3) データ評価の条件

NSS で取得したネットワーク情報の各データとコネクションの接続時間とを比較し、NSS の判断が正しいか判定する。判定は、ネットワーク情報が JPN ＜ USA のとき、コネクション接続時間が JPN＜ USA なら判断は正しい（的中）とし、逆にコネクション接続時間が JPN≧USA になっていた場合、判断は間違いとする（TP は判断が逆）。

有効データに関しては以下の条件を取り入れた。

・AS パス：　等しい場合は除く
・RT（ルータホップ数）、ＲＴＴ、TP（スループット）：両サイトからクライアントまで到達した場合のみ有効にする。

実験データの比較は小数点第3位を切り捨てて行う。実験は２回行った。

(4) 結果

結果は表４のようになった。

表４　ネットワーク情報の正解率

１回目）５日間：データ総計　２１００８件

	平均正解率	サンプル件数
ＡＳ	５３．４％	１４６０５
ＲＴ	４８．９％	５７２７
ＲＴＴ	７１．１％	６７４３
ＴＰ	６７．８％	６０２８

２回目）８日間：データ総計　３９４７４件

	平均正解率	サンプル件数
ＡＳ	５２．２％	２７３６５
ＲＴ	４８．９％	１０２５０
ＲＴＴ	７１．５％	１２６６６
ＴＰ	６６．２％	９９５８

経路情報（AS パス数）による最短経路は、約５０％の割合で最適な経路を選択。一方、ＲＴＴでは、約７０％の割合で最適な経路を選択した。この結果からうまく両者のデータを利用すれば、最適な経路を短時間で取得できる。なお、日、週においても動的にネットワーク変化している事が確かめられた。

3.2 システムテスト

今回作成したプロトタイプでは、ネットワーク状態値とサイト状態値の算出において利用するパラメータはASホップ数、ルータホップ数だけ有効にした。現時点でのパラメータの計算によって得られた最適サーバへのリダイレクト動作を確認できた。

4. 考察

4.1 パラメータについて

今回の予備実験ではネットワーク情報のパラメータ4つ（クライアントまでのＲＴＴ、スループット、ルータのホップ数、ＡＳパス数）について有効性を検証した。2回実施した実験において、1回目、2回目の平均正解率をパラメータ毎に比較すると、ほぼ同じ（2%以内）結果であることからデータの信頼性は高いと考える。一番低い正解率であったのは、ルータのホップ数であった。一般的にルータのホップ数が多いとルータ内でのオーバヘッドが影響し、ネットワーク的な距離は遠くなると考えられるが、今回の結果では、現在のインターネット（ＷＡＮ）上において、ネットワーク（＝回線）そのものの混雑による遅延の方がルータ内のオーバヘッドより上回っているため、ルータ内のオーバヘッドはそれほど大きく影響しないと推測できる。

4.2 アルゴリズムについて

最適サーバを決定するための手順としてプロトタイプでは2ステップのアルゴリズムで決定した。最初のステップで最適サイトを決定し、次のステップで最適サーバを決定した。その結果、当初の想定通りの動作を確認できた。今後は、計測パラメータをさらに吟味し、より精度の高い決定ができるように、システムの調整が必要である。

4.3 システムについて

最適サーバへの振り分けをHTTPリダイレクトによって実現した。HTTPリダイレクトはオーバヘッドが大きく、負荷がそれほど高くないシステムでは有効に機能するが、負荷の高いシステムではオーバヘッドがボトルネックになる可能性がある。

5.今後の課題とまとめ

本論文では、広域に分散配置されたＷｅｂサーバにおいて、最適サーバを探索するためのシステムについて検討した。実際のインターネット上にサイトを構築（日本と米国に計測用のＷｅｂサーバサイトを構築）し、サーバとクライアント間のデータ転送時間および各ネットワーク状態値を計測し、これらの相関から最適なＷｅｂサーバ探索の手法を検討した。

今後の課題としては、

- 予備実験で調査できなかったサイト内ネットワーク情報など、残りのパラメータの有効性について調査
- 調査した結果からネットワーク情報とサイト内ネットワーク情報の重みのバランス（パラメータの調整）を再度検討

などがある。

また、今後はこれら残りのパラメータの調査を進めるとともに、パラメータの精度を上げ実用性を高めていく予定である。

謝辞 プロトタイプシステムの開発にあたりRS(zebra)の改良に協力していただいた（株）デジタル・マジック・ラボの石黒邦宏氏に感謝いたします。

参考文献

1) ttcp - Enger, R., Reynolds, J., FYI on a Network Management Tool Catalog, RFC1470(1993)

2) netperf - http://www.netperf.org/netperf/NetperfPage.html

3) Apache modules - http://modules.apache.org/

4) Berners-Lee, T., Fielding, R. and Frystyk, H., Hypertext Transfer Protocol -- HTTP/1.0, RFC1945(1996).

5) GNU Zebra - http://www.zebra.org/

6) Yutaka Nakamura, Ken-ichi Chinen, Suguru Yamaguchi, Hideki Sunahara:
"Proposal of WWW Server Behavior Observation Method by Packet Monitoring", In Proceedings of the Internet Conference 1998, December 1998.

7) W.Richard Stevens, TCP/IP Illustrated, Volume1:The Protocols. Addison-Wesley Publishing Company 1995

8) W.Richard Stevens, UNIX Network Programming, Prentice Hall, 1990