分散システム運用技術　9－7
（1998．5．15）

DNS を用いた広域負荷分散の実装

馬場 始三　　山口 英

倉敷芸術科学大学　　奈良先端科学技術大学院大学

概要

WWW をはじめとするインターネットの情報サービスが普及するにつれてサービス利用者数が大幅に増加し、特定のサーバやコンテンツへのアクセス集中により、サーバレスポンスの低下やネットワークの輻輳といった問題が顕在化している。本論文では、既存の負荷分散技術について紹介した後、災害時に利用することを目的に開発が進められている広域分散アプリケーション IAA システムが利用する、DNS を用いた広域負荷分散手法について述べる。また、第3回インターネット災害訓練において実験した結果、DNS の重み付けラウンドロビン処理による静的なアクセス負荷分散が実現できたことを報告する。

A DNS based implementation on widely load balancing mechanism

Tomomitsu BABA　　Suguru YAMAGUCHI

Kurashiki University of Science and the Arts　　Nara Institute of Science and Technology

ABSTRACT

As the Internet is increasingly being viewed as providing information services, the number of users have become exponentially increased. Due to the increased WWW traffic simultaneously, it is essential to improve the Web servers' performance and to avoid the network congestion.

In this paper we show the related works regarding load balancing mechanism, and we describe our DNS based implementation to address the widely load balancing mechanism. Furthermore, we report the result of the weighted ROUND ROBIN approach in the 3rd Internet Disaster Support Drill.

1 はじめに

災害時に役立つ情報に代表されるような大多数の利用者から注目される情報を取り扱うインターネット上のアプリケーションにとり、サービスのスケーラビリティを備えることは重要な課題である。注目度の高い情報を提供する World Wide Web や電子メールサービス等では、サーバの処理能力の上限や情報が伝送されるネットワークの帯域幅、ならびに伝送遅延などがスケーラビリティを妨げる主要な要因として考えられる。このようなスケーラビリティの問題を解決する手段として広域ネットワークにおける負荷分散技術があり、大別するとサーバ複製技術、サーバ選択技術、サーバリダイレクト技術やキャッシュ技術がこれまでにいくつか提案、実装されている。

このような広域負荷分散技術は日常時に十分に役立つ技術である一方で、災害時においても重要な役割を担うと考えることができる。災害規模は予測できないものの、負荷分散技術は災害時に役立つアプリケーションに対して利用者数に対するスケーラビリティを提供する。また一方で、被災場所の予測は困難ゆえに、等価なサービス内容を持つサーバ群を地理的に分散したいという要求に応える頑健なシステムの設計技術として役立つ。

本論文では、災害時に必須となる広域負荷分散技術について、まず既存手法を紹介する。次に、1998年1月に開催された第3回インターネット災害訓練における、DNS(Domain Name System)[1] にもとづく広域負荷分散の実装について報告する。

2 負荷分散手法

2.1 サーバ複製技術

DNS や NIS[1] には、同一情報を複数サーバから提供するためにサーバ間で情報を同期する機構が組み込まれている。データを更新管理できるのはプライマリサーバまたはマスターサーバと呼ばれるサーバのみである。また、Sybase の replication サーバ [2] では、分散データベースの処理レベルで情報の正しさを保証しつつ、各 replication サーバ上で情報更新が可能となっている。また replication サーバ群のネットワーク監視や管理を簡易化している特長をもつ。

[1]Network Information Service

2.2 サーバ選択技術

DNS のラウンドロビン処理応答によるサーバ選択が広く一般的に利用されている。サーバを提供するサイトのホスト情報を記述した DNS ゾーン情報の中において、提供されるサーバのサービスホスト名に複数 IP アドレスを割り当てる。この結果、クライアントからサービスホスト名の名前解決要求が DNS へ送られた時に、IP アドレスのリストをラウンドロビン処理することにより、異なる IP アドレス情報をクライアントへ DNS が送り返す。

NTT の DyMS [3] では、アクセスナビゲーションサーバがクライアントの HTTP アクセス要求に対して、分散先サーバ名を記述した再宛先指示ファイルをクライアントへ返送することにより、トラヒックの分散を考慮しつつ分散先サーバへの再アクセス方式を実現している。

BPROBE / CPROBE [4] はクライアントサーバ間のネットワークパスに対する帯域幅と RTT を調べるツールである。このツールを利用することにより、クライアント側から制御する動的なサーバ選択が実現可能である。

anycast [5],[6] はネットワーク層の anycast ではネットワーク層の情報しか得られないことを欠点としてとらえ、アプリケーション層における anycast を検討したサーバ選択手法であるが、anycast 時にサーバの選択尺度となる情報の収集方法が課題である。

2.3 サーバリダイレクト技術

Cisco LocalDirector [7] と DistributedDirector [8] は、サーバの追加削除をサービス利用者に隠蔽する技術である。複雑に構成されたサーバ群を単一の IP アドレスや URL 情報として利用者へ提供する。LocalDirector は LAN に配置されたサーバ群へのアクセス負荷分散をはかる目的で利用される。一方、DistributedDirector は WAN に複数配置されたサーバを対象に、DNS のキャッシュサーバ動作モードか、または HTTP セッションのリダイレクトをおこなう動作モードのいずれかでドメ

イン単位で利用される。DistributedDirector は相互に DRP (Director Response Protocol) を利用して協調して動作するための情報交換をおこなっている。

NAT [9] 技術を用いることでサーバプール機構を実現し、複数の IP アドレスから構成されるアプリケーションサーバ群を単一の IP アドレスを持つホストとして扱うことが可能である。また、サーバ群の中で最小負荷のサーバへクライアントからのアクセスを動的に振り分ける機能を組み込むことができる [10]。

WWW サーバによるリダイレクト機能も存在する。Apache や NCSA サーバに代表される WWW サーバでは、*Redirect* というサーバ設定を利用して、ある URL に対する HTTP アクセスを別のサーバを参照するようにクライアントへ情報を渡すことが可能である。この場合、*Redirect* を設定したサーバを一度経由して別のサーバへアクセスすることになるため、最初のサーバではリダイレクト先の URL 情報を含むサーバレスポンスをクライアントへ返す HTTP 処理が必要となる。

2.4 キャッシュ技術

キャッシュ技術を積極的に利用するものとして、WWW キャッシュサーバと DNS キャッシュサーバを挙げる。WWW キャッシュサーバは WWW オブジェクト単位で情報をキャッシュする。キャッシュ有効時間はキャッシュ情報が個々に持つことは稀であり、キャッシュサーバごとに設定されている。キャッシュ有効時間は、キャッシュ情報中に含まれる Last-Modified:*date* と Date:*date* ヘッダ情報の二つの時間差をもとにサーバごとの運用ポリシーを反映した上で計算される。

DNS キャッシュサーバは、ゾーン情報単位または各資源レコード単位の TTL 値にもとづいて、キャッシュサーバ上にキャッシュされた情報の有効時間を管理する。現在の BIND の実装ではゾーン情報単位で指定する TTL 値は各資源レコードの TTL 初期値として利用可能となっており、資源レコードへ明示的に指定した TTL 値が優先される。キャッシュサーバが提供する情報は、参照するクライアントの都合に合わせてキャッシュ制御できない。

3 DNS を用いた広域負荷分散

インターネット災害訓練 [11] では、DNS を用いた負荷分散手法を用いて IAA システム（生存者情報データベースシステム）[12] へ集中する大量のアクセス数を分散をはかってきた。本論文では、特に第 3 回訓練時の広域負荷分散手法である、重み付けラウンドロビン方式に焦点をあててこの手法を説明する。

3.1 インターネット災害訓練と IAA システム

インターネット災害訓練[2] は、災害時に役立つインターネットアプリケーションやネットワーク運用技術の検証の場として、一般のインターネット利用者の訓練参加のもと、1996 年 1 月から 1998 年 4 月現在までに 3 回実施されている。同訓練では、寸断された幹線経路の迂回経路構築や、災害時の利用を目指したアプリケーションである IAA システム等が訓練プログラムとして運用されてきた。IAA システムは WAN 上の複数箇所へ設置された IAA クラスタと呼ばれる機能単位から構成され、IAA クラスタはシステム独自の情報同期機構から同一のコンテンツを個々に保有する。

IAA システムが同一コンテンツを持つ複数の IAA クラスタから構成される理由として、地理的にみて一箇所への情報の集中を避けることと、システムへ寄せられる大量のアクセス負荷を複数の IAA クラスタへ分散させる狙いがある。この WAN 上へ配置された IAA クラスタ間の負荷分散の手法として、IAA システムでは DNS の資源レコードのラウンドロビン機能により、システムへのアクセス負荷を IAA クラスタへ分散している。

3.2 重み付けラウンドロビン方式

1996 年 1 月に実施された第 1 回訓練では、均等な処理性能を持つ 4 箇所の IAA クラスタに対して、アクセス要求をラウンドロビン方式によりほぼ均等に分散させることができた。一方、1998 年 1 月に 2 日間にわたって実施された第 3 回訓練では、表 1 に示す通り、不均一な性能を持つ 3 箇所の IAA クラスタ内のサーバに対して適切に負荷

[2] http://www.iaa.wide.ad.jp/

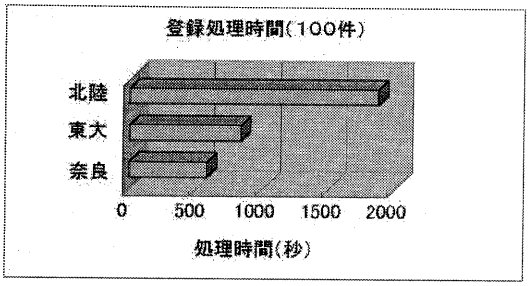

図 1: 100 件単位の登録処理時間

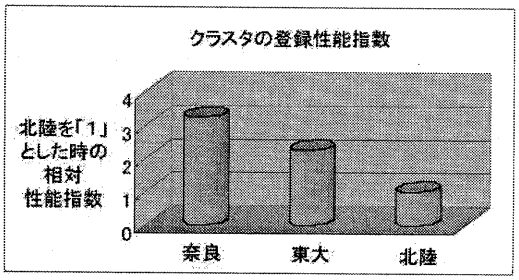

図 2: 相対的な登録性能指数

分散を行なう必要があった。そこで第 3 回訓練の IAA システムでは、不均一さをサーバの性能差から定量化することにより、重み付けラウンドロビン方式によりクラスタの処理性能に応じる形で負荷分散させることを考えた。

まず、生存者情報の登録シミュレータを作成し、一定量の登録要求を事前にサーバへ自動的に流し込むことでサーバ処理能力を定量化したところ、図 1 の通りとなった。したがって、もっとも処理数の少ない北陸に設置した IAA クラスタの性能を「1」として相対的な性能指数を求め、図 2 の結果を得た。この結果から IAA クラスタ間の負荷分散比率を表 1 の最下列に示す値へ定めた。

次に、IAA システムが利用者に公開する WWW サーバ名を DNS で名前解決する際に、3 箇所の IAA クラスタが提供する 3 台のサーバの IP アドレスを用いて重み付けラウンドロビンされるように DNS の A レコードを定義した（図 3 参照）。このように、同一サーバへ複数 IP アドレスを割り当てるため、サーバが稼働する BSDI BSD/OS

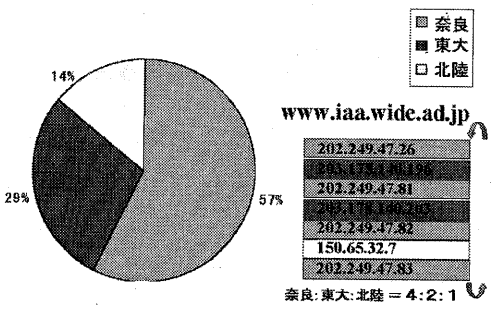

図 3: 広域負荷分散の実装

3.1 の ifconfig コマンドを用いて、エイリアス機能により複数アドレスを必要数だけ割り当てた。この時、奈良の IAA クラスタではサーバを接続するネットワークセグメントの IP アドレス空間が枯渇したため、新たに別サブネットの経路情報を同サーバからアナウンスすることによって、必要アドレス数を確保した。

4 訓練結果と考察

1998 年 1 月の第 3 回インターネット災害訓練の 2 日間という訓練期間を通じて得られた WWW サーバへのアクセス数の経時変化を図 4 に示す。

また同訓練の広域負荷分散結果を図 5 ならびに 図 6 に示す。このデータは各 IAA クラスタの WWW サーバアクセスログから算出した。

訓練過程において、奈良の IAA クラスタが 17 日の午後 3 時から午後 4 時までの 1 時間近く、動作不安定に陥ったことを除き、他の IAA クラスタは継続してサービスを提供した。訓練期間の 2 日間を通してまとめると、奈良の IAA クラスタの停止時間も含めたアクセス分布の場合、2 日間を通して ± 3 %の幅で分布が目標値に近付いた。奈良の IAA クラスタの停止時間を除いたアクセス分布の場合は、さらに 2 日間を通して ± 2 %の幅で分布が目標値に近づくことができた。

一方、1 時間単位で実際のアクセス数の分布がどのように推移しているのかをまとめてみた（図 7 参照）。処理率の分布は 17 日の 15 時台に奈良の IAA クラスタがサービスを一時的に停止して

表 1: IAA クラスタ間の広域負荷分散比率

IAA クラスタ	奈良	東京	北陸
サーバ OS	BSDI BSD/OS 3.1		
サーバ CPU	Pentium Pro/200 MHz	Pentium/200 MHz	Pentium/166 MHz
サーバ メモリ	128MB	128MB	64MB
負荷分散比率	4	2	1

いたことを示している。また、18 日の早朝に東大の IAA クラスタの処理率が落ち込んでいるが、この時間帯はサンプル数が少ないことと、開発者が特定の IAA クラスタへ直接アクセスする行為によって、分布がひずんでいると考えることができる。以上より、ネームサーバの重み付けラウンドロビンによる広域負荷分散は目標をほぼ達成したといえる。

5 今後の課題

今回は静的な負荷分散比率を導入したが、訓練時に発生したようにサーバの障害時や新たな IAA クラスタの組み込みに対応するためには、動的に負荷分散比率を制御できることが望ましい。特に、IAA システムを構成する IAA クラスタの障害発生時に、サーバの切り離しやサーバの追加を手動で行なうのではなく、障害の自動検知機構とともにサーバ単位のプラグイン、プラグアウトが行なえることが望ましい。

一方、分散比率の決定尺度としてサーバの登録性能を利用したが、サーバ間で OS やデータベースが異なる場合は、ネットワーク処理能力や検索性能といった多角的に複数の尺度を検討する必要がある。

6 おわりに

本論文では既存の負荷分散技法について概括したあと、インターネット災害訓練における IAA システムにおいて利用された、DNS を用いた広域負荷分散手法について述べた。この手法を用いることにより、ほぼ想定した比率でアクセス数を複数の複製サーバへ分散させることができたことを示した。

DNS のラウンドロビンを用いた単一手法による負荷分散には限界があることから、今後は先に述べた課題を踏まえた上で、他の負荷分散技術と組み合わせる形で頑健性を兼ね備えたスケーラブルな広域負荷分散手法についてさらなる検討を行ないたい。

謝辞

本研究において、貴重な御意見を与えてくださった WIDE プロジェクトの皆様、また、インターネット災害訓練に参加いただいた皆様、災害訓練を開催するにあたりご協力いただいた多くの組織の方々に感謝いたします。

参考文献

[1] P.V.Mockapetris and K.J.Dunlap, editors. *Development of the Domain Name System.* ACM SIGCOMM'88, 1988.

[2] Sybase, Inc. SYBASE Replication Server, 1997. Hypertext document. Available electronically at http://www.sybase.com/products/datamove/repserver_wpaper.html.

[3] 中島伊佐美, 堀米英明, 松村龍太郎, 巳波弘佳. サーバ集中トラヒック分散技術. NTT 技術ジャーナル, pp. 27–29, November 1997.

[4] Robert L. Carter and Mark E. Crovella. Server Selection Using Dynamic Path Char-

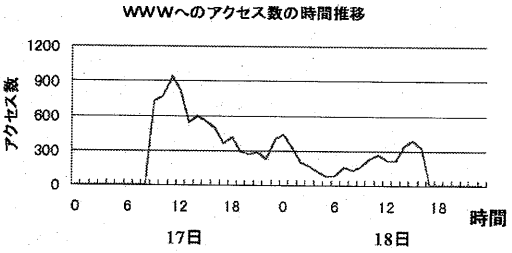

図 4: IAA システムに対するアクセス数の時間推移

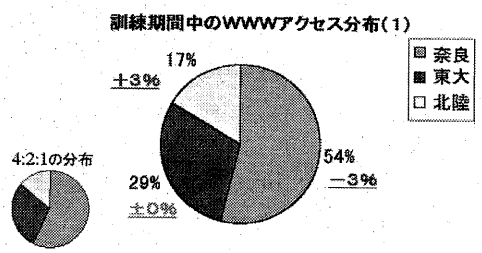

図 5: 分散結果 (1)

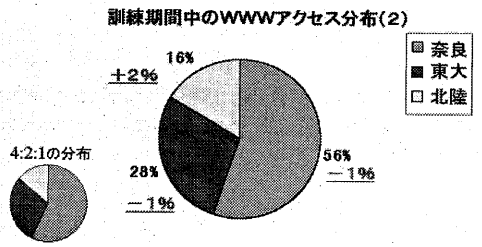

図 6: 分散結果 (2)

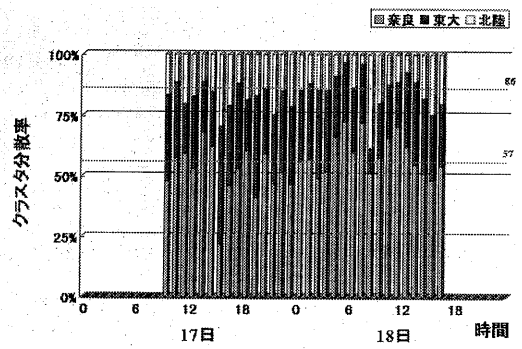

図 7: 1 時間単位の分散結果

acterization in Wide-Area Networks. In *Proc. IEEE INFOCOM'97*, April 1997.

[5] T.Mendez C.Partridge and W.Milliken. *Host anycasting service*, RFC 1546, November 1993.

[6] Samrat Bhattacharjee, Mostafa H. Ammar, Ellen W. Zegura, Viren Shah, and Zongming Fei. Application-Layer Anycasting. In *Proc. IEEE INFOCOM'97*, April 1997.

[7] Cisco Systems Inc. LocalDirector, 1996. Hypertext document. Available electronically at http://www.cisco.com/warp/public/751/lodir/lodir_wp.htm.

[8] Cisco Systems Inc. DistributedDirector, 1997. Hypertext document. Available electronically at http://www.cisco.com/warp/public/751/distdir/dd_wp.htm.

[9] P. Francis and K. Egevang. The IP Network Address Translator (Nat). RFC 1631, May 1994.

[10] 井上博之, 山口英. NAT による WWW サーバの負荷分散機構の実装. 情報処理学会 マルチメディア通信と分散処理研究会 資 料 96–DPS–78–4, September 1996.

[11] Yoichi Shinoda, Tomomitsu Baba, Nobuhiko Tada, Akira Kato, and Jun Murai. Experiences from the 1st Internet Disaster Support Drill. In *Proceedings of INET'96*, June 1996.

[12] 馬場始三, 篠田陽一. 第 1 回インターネット防災訓練における生存者情報データベースについて. インターネットコンファレンス'96 論文集 (pp.17-24), July 1996.