マルチメディア通信と分散処理 78－4
（1996．9．27）

NATによるWWWサーバの負荷分散機構の実装

井上 博之[1]　　山口 英
{h-inoue,suguru}@is.aist-nara.ac.jp
奈良先端科学技術大学院大学

概要

インターネット上の主要なサービスの1つであるWWWサーバは、そのアクセスの集中による過負荷が問題になっている。本稿では、サービスを分散させる機構として LB-NAT(Load Balancing Network Address Translator)と呼ぶ負荷分散機構の実装を提案する。NAT技術を用いることでサーバプール機構を実現し、複数のホストすなわち複数のIPアドレスからなるWWWサーバ群を、1つのIPアドレスを持つホストに見せかける。これにより、WWWクライアントからのアクセスは、その単一のIPアドレスによって行われ、現在の負荷が最小であるWWWサーバに振り分けることで負荷分散を実現する。本実装ではWWWクライアントからのアクセスは完全に透過的である。また、見かけ上のWWWサーバの耐故障性を向上することも可能となる。

Implementation of Load Balancing of WWW Server using NAT

Hiroyuki Inoue[2]　　Suguru Yamaguchi
{h-inoue,suguru}@is.aist-nara.ac.jp
Nara Institute of Science and Technology, Ikoma, Nara, JAPAN

Abstract

One of major issues on the Internet services, especially on the WWW service, is "server overload" caused by excessive concentration of accesses from their clients. We purpose the load balancing mechanism called LB-NAT (Load Balancing Network Address Translator) in order to improve service scalability on the Internet. In the LB-NAT which uses the NAT technique as a "connection switch", we can advertise a single IP address for the pool of WWW servers, and an LB-NAT server dispatches requests to a WWW server with the minimum load. For WWW client, this LB-NAT mechanism provides transparent accesses to a pool of WWW servers. Furthermore, it can also add the fault-tolerance capability to the operation of WWW servers.

---

[1] 住友電気工業㈱から留学中

[2] He also works for Sumitomo Electric Industries, Ltd.

1 はじめに

インターネットの代表的な情報サーバであるWWW(World Wide Web)が広く普及してきた一方で、WWWサーバへのアクセスの集中によるサーバの反応時間の増大が問題になってきている。インターネット上で人気のあるWWWサーバでは1秒あたり100から500ものアクセス[1]の処理が要求される。そのため、WWWサーバとして大型計算機並みのシステムを用意することや、特殊なネームサーバを使って負荷が最低であるWWWサーバのIPアドレスを返すことでWWWサーバの負荷分散を行う研究などが行われている[2]。また、接続相手の種類によってサーバが提供するサービスを切り替えるような、サービススイッチ機構にNATを利用した研究もある[3]。

本稿では、WWWサーバへのアクセスを分散させるための機構として、NAT(Network Address Translation/Translator)を使った実装を提案する。複数のホストの集合から成るWWWサーバを、NATによって1つのIPアドレスを持つホストに見せかける。WWWクライアントからのアクセスはその単一のIPアドレスによって行われ、NATによって現在の負荷が最小であるWWWサーバに振り分けることで負荷分散を実現する。以下では、負荷分散を行うNATのことをLB-NAT(Load Balancing NAT)と呼ぶ。

2 NAT

NAT[4]は組織内のプライベートIPアドレス[5]をインターネットで通用するグローバルIPアドレスに変換するために提案された。図1にNATの典型的な運用環境を示す。インターネットからNATの上半分まではInterNICから割り当てられたグローバルIPアドレスを使用し、NATの下半分から組織内のネットワークはRFC1579[5]で規定されたローカルIPアドレスを使用する。組織内ネットワークのあるクライアントからインターネット上のあるサーバにアクセスがあったとき、NATが管理している複数のグローバルIPアドレスから1つをクライアントのために割り当て、両者のアドレスを双方向にマッピングすることでアクセスを実現している。アクセスはクライアントから見て完全に透過的である。

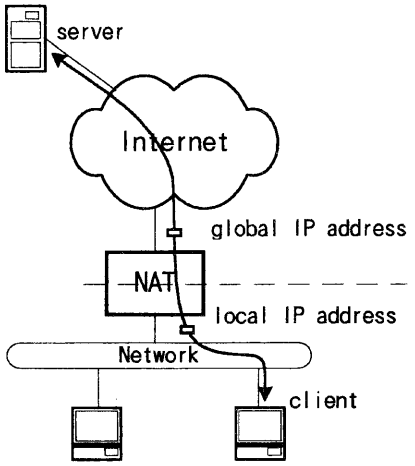

図 1: NATの動作

このNAT機能を搭載した商品としてのルータは1995年春にnti社[3]から発売され[6]、その後多くのメーカが同様の機能を提供するルータを発売している。

3 モデル

3.1 構成要素

負荷分散機構のシステムモデルは図2のようになる。構成要素として以下のものを定義する。

[3] 現在はCisco社に買収されている

- サーバ

  WWWのサービスを提供する計算機で、N台のホストから成る(N≧1)。その提供するWWWコンテンツは同一であるとする。
- クライアント

  WWWのサービスを受ける計算機で、M台のホストから成る(M≧1)。
- LB-NAT

  サーバ間で負荷分散を行うクライアントからの要求をサーバに割り振る。ネットワーク1とネットワーク2の間のルータとして扱われる。
- ネットワーク1

  サーバ群とLB-NATを接続するためのネットワークで、高速なLAN(Local Area Network)から成る。
- ネットワーク2

  クライアント群とLB-NATを接続するためのネットワークで、広域に分散したインターネットである。
- セッション

  クライアントからサーバに対して行われているWWWデータへの1つのアクセスで、HTTP(Hypertext Transport Protocol)[7]におけるconnectionのことである。

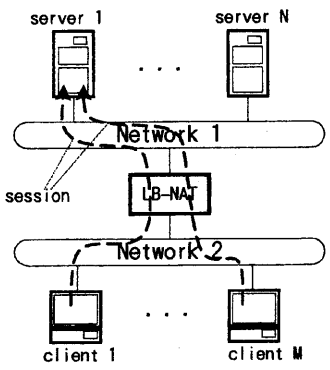

図 2: システムモデル

3.2 サーバの負荷

あるサーバ $server_i$ の負荷を表す指標として次のものを定義する。

- サーバの実行待ち行列にあるプロセス数 L_i
- サーバが処理中のセッションの数 S_i

4 実装

LB-NATにおけるアドレス変換処理とサーバ間の負荷分散処理の実装について説明する。

4.1 アドレス変換処理

サーバ群 $server_j$ は、WWWサービスを提供する1つのIPアドレス S_A とTCPポート S_P の組$[S_A,S_P]$として、ネットワーク2側から認識されるものとする。LB-NATにおける、アドレス変換処理は次のようになる。

1. ネットワーク2から受け取ったパケットを調べ、あるクライアント C_i から$[S_A,S_P]$へのセッション接続要求を検出する。セッション接続要求はTCP(Transport Control Protocol)のSYN flagで知ることができる。
2. 後述のアルゴリズムで適当なサーバ S_j を選び、$[[C_{iA},C_{iP}],[S_{jA},S_{jP}]]$の組合わせを記憶する。この4つの情報の組合わせは C_i と S_j を対応づけるもので、アドレス変換におけるセッション・タプルと呼ぶ。
3. ネットワーク2から受け取ったパケットを調べ、クライアント$[C_{iA},C_{iP}]$から$[S_A,S_P]$に宛てたパケットを検出すると、その行き先アドレスとポート番号を$[S_{jA},S_{jP}]$に置換してネットワーク1に転送する。

1. ネットワーク 1 から受け取ったパケットを調べ、サーバ[S_{jA},S_{jP}]からクライアント[C_{iA},C_{iP}]に宛てたパケットを検出すると、その送り元アドレスとポート番号を[S_A,S_P]に置換してネットワーク 2 に転送する。
2. クライアント[C_{iA},C_{iP}]またはサーバ[S_{jA},S_{jP}]からセッション終了要求のパケットを受け取った時(TCP では FIN flag で知ることができる)は、3,4 と同様の処理を行った後に、セッション・タプルの状態を half-close とする。さらに、逆向きのセッション終了要求のパケットを受け取った時点で、セッション・タプルを消去する。

以上のアドレスを変換を行うことで、クライアント Ci からはサーバ[S_A,S_P]にアクセスしているかのように見えて、実際はサーバ Sj にアクセスするという仕組みが実現できる。このアクセスは完全に透過的であり、LB-NAT の存在はクライアントからは見えない。

4.2 LB-NAT の構成

LB-NAT は図 3 に示した各部から構成される。

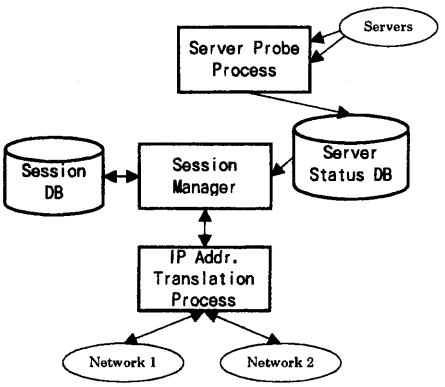

図 3: LB-NAT の構成

表1: サーバ負荷データベースの内容

名前	内容
last_update	情報の最終更新時刻
server_addr	サーバの IP アドレス
reachable	サーバと LB-NAT の間でネットワークが到達可能かどうかの情報
reachable_last	上の情報の最終変化時刻
serviceable	サーバの WWW サービスが稼動中かどうかの情報
serviceable_last	上の情報の最終変化時刻
load	サーバの実行待ち行列にあるプロセス数
load_last	上の情報の最終変化時刻
nsession	サーバが処理中のセッションの数
nsession_last	上の情報の最終変化時刻

- サーバ負荷収集部
  全サーバの負荷情報を収集し、データベースに格納する。サーバ状態データベースのサーバ毎に持つ情報は表 1 に示すようになる。
- セッション管理部
  適切なサーバを選択し、セッションを管理する。また、サーバ自身あるいはそのサービスの停止を検出し、回復処理を行う。
- IP アドレス変換処理部
  クライアントとサーバ間の IP パケットのアドレス変換処理を、セッション・タプルの情報を基に行う。

4.3 サーバの選択

負荷が最小であるサーバを選択するには、(1)サーバの実行待ち行列にあるプロセス数 L_i が最小のものを選ぶ、(2)サーバが処理中のセッションの数 S_i が最小のものを選ぶ、のいずれかで行う。L_i は Unix システムにおいては Load Average という情報として容易に得ることができる。S_i は LB-NAT のセッション・タプル

を検索することで得ることができる。

また、この実装において、収集したサーバの負荷は即時性のある情報ではないため、同程度の負荷のサーバ間でラウンドロビンでセッションを振り分ける処理も必要になると考えられる。

4.4 サーバ故障時の回復

LB-NAT は負荷が最小のサーバを選ぶと同時に、サービスが停止していると考えられるサーバを選ばないことも同様に可能である。すなわち、あるサーバが故障した時にそれを検出することにより、システム全体としての対故障性を向上させることができる。

サーバの故障を検出するには次のような方法を用い、いずれも選択すべきサーバから除外する。

- サーバの負荷収集時に検出する
  サーバの負荷を収集する際に応答のないサーバはダウンしているものとみなす。
- セッションの開始時に検出する
  LB-NAT で発生する ICMP(Internet Control Message Protocol) Host Unreachable または Port Unreachable メッセージを検出する。
- セッションの途中で検出する
  通信中にサーバがダウンしても途中の経路では原理的に検出できない。このセッションはタイムアウトして通信エラーになってしまう。

故障から回復したサーバの検出は、サーバの負荷収集時に同様に行い、サービスが再開されていると判断された場合は、再び選択対象として扱うようにする。

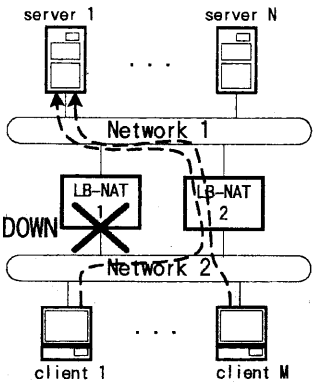

図 4: LB-NAT の冗長構成

4.5 LB-NAT の冗長構成

図 2 のシステムモデルでわかるように、LB-NAT 自身に異常が発生すると、サーバのサービスを継続することができなくなってしまう。しかしながら、LB-NAT はネットワーク 1 と 2 の間のルータとして見えており、LB-NAT を並列に複数接続することによって冗長構成をとることが可能である。

ただし、セッションを管理する必要上から全てのパケットはただ 1 つの LB-NAT しか経由することはできず、それがダウンした時に初めて別の LB-NAT が引き継ぐ形になる（この方式にあったルーティングプロトコル、例えば OSPF[8] を使用しなければならない）。よって、LB-NAT 間での負荷分散はできない。また、LB-NAT の切替わり時に既存のセッション情報は失われてしまうが、最善努力型のインターネットの性質上、特に対処は行わない。

4.6 実装環境

4.4BSD ベースの Unix システム(BSDI 社 BSD/OS2.1)上に LB-NAT を実装した。IP アドレス変換部とセッション管理部はカーネルのネットワーク層の一部として組込み、サーバ

負荷収集部はユーザ空間のプロセスとして組込んだ。

5 現状と今度の課題

現在、上記アルゴリズムに基づく LB-NAT を評価中である。負荷分散方式の妥当性、サーバ故障時の回復動作の効果、システム性能、特に LB-NAT のスループットの評価などを行っている。

また、今回の方式では、サーバの負荷としてクライアントの性能、サーバのネットワークデバイスやディスク装置のI/O量を考慮しておらず、この点についても実験を通じて明からにしていきたいと考えている。

6 まとめ

本研究では、WWW サーバの負荷を分散し、またサーバの故障時にも継続してサービスを提供する手法として、NAT を使った機構を提案した。HTTP の connection を 1 つのセッションとして扱い、セッション毎に[[Client IP addr., Client TCP Port], [Server IP addr., Server TCP Port]]からなるセッション・タプルを LB-NAT で管理し IP アドレスを変換することで、クライアントから見て完全に透過的なサーバへのアクセスを実現できる。タプル作成時にサーバの負荷量に応じてサーバを選択することで、サーバ間の負荷分散を実現する。また、サーバがダウンするなどの故障を検出した場合は、サーバ選択から除外することで、クライアントから見て継続的なサービスを提供することができる。LB-NAT 自身も並列に複数接続することで冗長構成をとることができる。

謝辞

本研究をすすめるにあたり、多大なご協力を頂いた奈良先端科学技術大学院大学情報ネットワーク講座の諸氏、および貴重なアドバイスを頂いた WIDE Project のメンバに感謝します。

参考文献

[1] The Standard Performance Evaluation Corporation. SPECweb96 Benchmark, http://www.specbench.org/osg/web/, July 1996.

[2] IBM Corporation. Load Balancing Among Internet Servers at UCLA, http://www.csc.ibm.com/advisor/provensolutions/pcid/75a6_602.html, August 1996.

[3] 多田信彦, 山口英, 山本平一. Service Switching 技術の研究, 情報処理学会研究報告 94-DPS-65, May 1994.

[4] K. Egevang and P. Francis. The IP Network Address Translator (NAT), RFC1631. May 1994.

[5] Y. Reckhter, B. Moskowitz, D. Karrenberg and G. de Groot. Address Allocation for Private Internets, RFC1597. March 1994.

[6] Andrew Foss. Primer: Perimeter Network Design with Private Internet Exchange Firewall, http://www.translation.com/. August 1995.

[7] T. Berners-Lee, R. Fielding and H. Frystyk. Hypertext Transfer Protocol - HTTP/1.0, RFC1945, May 1996.

[8] J. Moy. OSPF Version 2, RFC1583, March 1994.