# Root ネームサーバの運用について

加藤 朗
東京大学大型計算機センター
kato@wide.ad.jp

民田 雅人
松井証券
minmin@wide.ad.jp

西野 大
インターネット総合研究所
dai@iri.co.jp

1997 年夏から我が国において、インターネットのドメインシステムの最上位のゾーン “.” のサービスを専門に行うルート DNS サーバが稼働している。インターネットにおいては、階層的に定義されたドメイン名からアドレスなどの情報を得ることができる DNS は、一般ユーザが意識的に利用することは稀であるが、現実のサービス提供に関して重要な役割を負っている。本稿では、DNS の検索サービスを安定に提供するための構成や運用に関して述べる。

# Operation of a Root DNS Server

Akira Kato
The University of Tokyo

Masato Minda
The Matsui Securities

Dai Nishino
Internet Research Institute

One of the root DNS servers has been operational since the summer in 1997. The server is dedicated to serve the top level DNS zone “.” only. In the Internet, DNS plays an important role to resolve hierchically defined domain names into IP addresses and other associated information, and the root DNS servers are extremely important to provide stable Internet services. In this paper, the configuration and the operation of the root DNS server to offer its service stably is described.

## 1 はじめに

ドメインネームシステム（DNS）[1] は、階層的に定義された名前空間から各種の情報を得ることができるシステムであり、分散的に設置された DNS サーバに対して問い合わせを送ることによって名前の解決が開始する [2]。DNS は階層的に定義された名前空間を構成しているため、適当なサーバに関する情報が無い場合には、名前木の頂点から探索を行えばよい。この頂点で管理されている情報を Root Zone と呼び、DNS 的には “.” で表現される。

この Root Zone に対するサーバである Root DNS サーバ の内の一つ `M.root-servers.net` が我が国で稼働を開始してから一年余りが経過した。本論文では、その運用に関する背景やハードウェア構成、ソフトウェア構成について紹介し、その運用の考察を行う。さらに現在計画中の拡張案についても紹介する。

## 2 経緯

Root DNS サーバは、実質的に全世界のインターネットの運用の一つのキーになっている。従って、Root DNS サーバを安定に運用し、定常的な問い合わせに対応できることは、インターネットの利用に対しては非常に重要である。DNS の問い合わせには UDP を用いており、IP での fragmentation を避けるため、メッセージ長は 512byte 以下と規定されている。そのため、ゾーン “.” の記述に含めることができるネームサーバの数は 13 個が上限となっている。そのうち従来から 9 個がサービスを提供してきた[1]。

[1] `A.root-servers.net` ～ `I.root-servers.net`

また、Root DNS サーバの重要性に関して、その運用基準に関するメモも作成され[3]、主要なIX にルータを介して接続することなどが規定されている。

従来の9個のサーバは、一つがStockholmに置かれていた他は全て U.S. にあり、世界的なインターネットの発展に対応できなくなってきた。そのため、残り4台の枠のうち2台を新たに設置することが1996年12月のIEPGで議論され、一台をヨーロッパ方面に対するサービスの強化を考えてロンドンのIXであるLINXに、もう一台をアジア太平洋地域に対して東京のNSPIXP-2に設置することが提案された。IEPGでの指名を受けて、WIDE Projectでは機材の調達を開始するとともに、関連ISPにRoot DNSサーバへのトランジットの提供の依頼を始めた。その結果、1997年4月には運用の準備が整った。

当初、`J.root-servers.net` がIANAから提案されていたが、"J" はInterNICにおいてgTLDサーバの試験運用に用いられていたため、"M" が用いられることになった。IANAからの通知に対して直ちに設定作業を行い、1997年8月22日付けで正式に運用が開始されたことがアナウンスされた。

## 3 ハードウェア構成

従来から稼働しているサーバは、"." のみならず ".com" や ".edu" などの古くからあるTLDも同時にサービスしているが、サーバの負荷などの問題でRFC2010では "." のみを提供することを基本にし、余裕がある場合には別の計算機でTLDのサービスを提供しても良いことになった。そのため、資源の確保の問題から、NSPIXP-2に設置するサーバは "." のみを提供することになった。

InterNICの運用経験では、PC UNIXを用いる方式は経済的なインパクトは少ないが、障害発生などによる人的コストが却って高くついてしまうことが知られていた。経済的問題と安定な運用を考えて、WIDE Projectでは以下の様な方針でシステムを設計することにした：

1. ハードウェアは性能価格比を考えてPCとするが、安定運用を考えて二重化する。
2. ルータを介してNSPIXP-2のGigaSwitchにDASで接続する。

そのため、図1のような構成になった。

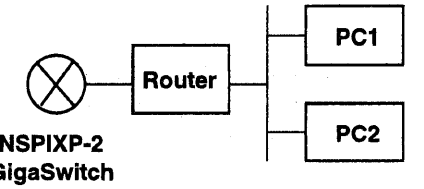


図1: サーバの構成

ルータはメモリ 64MB を実装したCisco4700Mを用い、PCは1997年4月現在でほぼ最速であったPentiumPro 200MHz CPUに64MBのメモリ、2GBのSCSIハードディスクという構成で現在も稼働している。各PCにはEthernetインターフェースを2枚実装し、片方をサービス用、もう片方をメンテナンス用とした。

## 4 経路制御による自動切替え

サーバを単に二重化しただけでは、片方のサーバに障害が発生した場合、オペレータが切替えを行う必要がある。このために、24時間体制で監視を行うのは現実的ではないため、障害時には自動的に切替えを行う必要がある。自動切替えの方法としては幾つかの方法が考えられるが、比較的手軽に実装できる方法として、経路制御を利用する方法を用いることにした。

各サーバには共通のサービス用のアドレスをEthernetインターフェースのアドレスとは別途割り振る。各サーバはサービスが可能な場合には、サービス用アドレスを経路制御プロトコルを通じて広告する。ルータは、経路制御情報を参考に、問い合わせを適当なサーバに送り、サービスを提供する。

Cisco IOSではRIPやIGRPなどの距離ベクトル型の経路制御プロトコルに関しては、各種のタイマ値を変更することができる機能があ

る。そこで経路制御プロトコルに RIP を使用して、

1. サービス用のアドレスは PC の loopback インタフェースに割り振る。
2. ネームサーバとしては bind 8.1.1[2]を用いる。
3. ネームサーバプロセスが生きている場合に限り、1 のアドレスを RIP で広告する。この際の metric は設定によって変更できるようにするが、主サーバからは 1 で、副サーバからは 2 で広告する。

という要領で運用することにした。

標準の RIP では、30 秒に一回更新メッセージを送り、180 秒メッセージが来ない場合には経路に問題があると判断し、更に 120 秒メッセージが来ない場合には経路を消去する、というパラメータが規定されている。サーバ障害の際の切替えに 180 秒掛かるというのでは安定なサービスを提供しているとは言い難いため、パラメータを変更し、10 秒に一回更新メッセージを送ることにした。それによって、40 秒メッセージが来ない場合には経路に問題があると判定し、この場合直ちに別な経路があればそれを採用することにした[3]。

この場合には、

- 主サーバがクラッシュした場合や主サーバのネットワークに障害が発生した場合には、40 秒以内に副サーバ側に経路が切り替わる。
- 主サーバの RIP プログラムが異常を検出した場合や、オペレータが RIP プログラムに signal を送った場合には、10 秒以内に副サーバ側に経路が切り替わる。

という応答性を得ることができる。

ルータは NSPIXP-2 のほとんどの ISP と BGP セッションを設定し、それぞれの ISP 独自の経路およびその顧客の経路の供給を受けている。また国際リンクを運用している幾つかの ISP は、トランジットサービスを提供して頂いている。1998 年 10 月中旬現在で約 106,000 Path・約 53500 経路がルータで管理されている。また、64MB のメインメモリは約 25MB の未使用領域があり、当面メモリ問題は発生しないことが予想される。

なお、二つのサーバから同じ metric でサービスアドレスを広告し、ルータでの経路キャッシュ機能を off にすると、問い合わせを両方のサーバに分散することができる。運用開始前のテストでは、この機能が正しく動作することは確認されている。ところが、TCP を利用した問い合わせに対して、そのセッションはどちらか片方のサーバに固定しないと正常に通信できないという問題が発生する。そのため、TCP による問い合わせはほとんど無いにも関わらず[4]、実際の運用には供されていない。

## 5 lorip プログラム

PC 側では、サービス用のアドレスに対応するホスト経路を RIP でルータに対して送らなければならない。しかし、タイミングを RIP の標準値ではなく 10 秒毎にしなければならないこと、ネームサーバプロセスに異常がある場合にはホスト経路の広告を通常の metric ではなく metric=16 で送出しなければならないことなどから、従来のプログラムではなく独自に RIP 送信用プログラム `lorip` を作成した。

10 秒毎にネームサーバプロセスの動作をチェックし、サーバが稼働していると思われる場合に loopback インターフェースに追加されたアドレスに対するホスト経路を、起動時にパラメータで指定された metric で送出する。このネームサーバプロセスのチェックは、

- `/var/run/named.pid` ファイルのチェック
- そのファイルのプロセス番号のプロセスに対して、無害なシグナル (`SIGINFO` を使用) を送ることができるかどうかを確認

という手順で行っている。

ネームサーバソフトウェアの更新を行う際には、まず副サーバ側のネームサーバプロセスが

[2]運用開始当時、現在は bind 8.1.2 を使用

[3]Cisco のパラメータでは、`timers basic 10 40 0 40` となる。

[4]ほとんど Root DNS サーバでは、Zone 転送は禁止している。

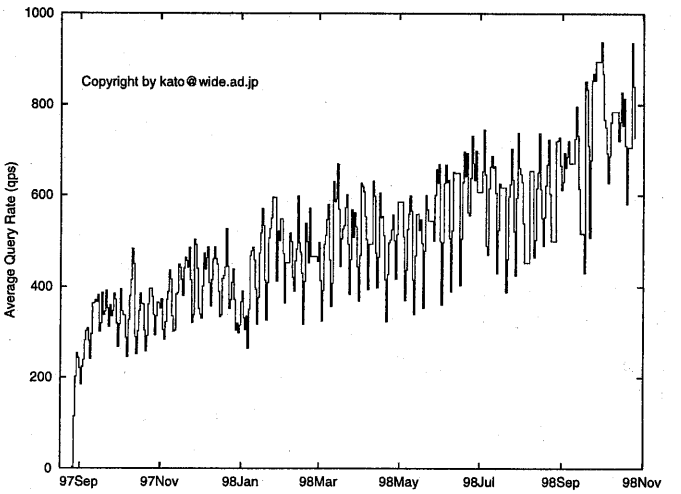


図 2: サーバへの問い合わせ頻度の推移

稼働していることを確認し、主サーバの `lorip` プログラムに `INT` シグナルを送る。すると、指定した metric の代わりに metric=16 で経路を送るので、ルータは直ちに副サーバ側に問い合わせを送るようになる。必要な処置を行った後、`lorip` プログラムを起動することによって、ルータは主サーバ側に問い合わせを送るようになる。

## 6 運用および考察

RFC2010 に規定されるガイドラインによれば、Root DNS サーバの障害に関しては 24 時間 365 日の対応が必要になる。サーバの二重化によって障害発生時に直ちに対応を取る必要はないが、副サーバに障害が発生する前に主サーバの障害回復措置を取らなければならない。そのため、運用は一人では不可能で、NSPIXP-2 に接続している ISP のオペレータの内の有志に協力を依頼した。

1997 年 8 月の運用開始から現在に至るまでの 14 ヵ月、ハードウェア的な障害は発生していない。ソフトウェア的には、`bind 8.1.2` に内在するメモリ管理問題によって、サーバソフトウェアがダウンする現象が数回発生している。この場合、自動的に副サーバ上の `bind` プロセスによってサービスが継続しており、重大な事態には至っていない。ソフトウェアの事故によって、二重化が正しく動作していることがときどき確認されている。

図 2 に、運用開始から現在に至るまでの "`M`" サーバへの問い合わせ頻度の推移を示す。当初は 200qps 程度であったが IANA のアナウンスによって各ネームサーバの設定の変更が行われると問い合わせも増加し、400qps 程度になった。その後、徐々に問い合わせが増加し、現在は 800qps を越えるようになってきた。また、図 3 に、最近のネットワークの負荷の状況を示す。Inbound のトラフィックはネームサーバへの問い合わせであり、0.5Mbps 程度であるが、その応答のトラフィック（outbound）は 2~3Mbps に達している。

ネームサーバの問い合わせの種類は表 1 に示す通りアドレスに対する問い合わせが 1/3 近くを占めている。なお IPv6 アドレスに対する問い合わせ "`AAAA`" は、現在 0.0025% に留まっている。

## 7 拡張

現在の "`M`" サーバの CPU の load average は 0.2~0.3 程度であるため、直ちに CPU の更新は必要ではない。PentiumPro 200MHz の二倍程度の性能を有する Pentium-II 450MHz が容

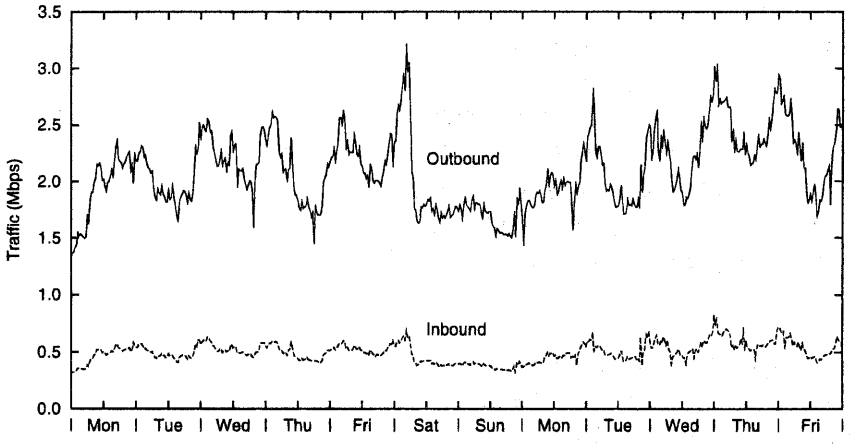


図 3: ネットワークトラフィックの状況

表 1: 問い合わせの種類

| 種類 | 割合 (%) |
|---|---|
| A | 63.03 |
| PTR | 24.14 |
| MX | 9.57 |
| ANY | 2.06 |
| NS | 0.49 |
| TXT | 0.22 |
| CNAME | 0.19 |

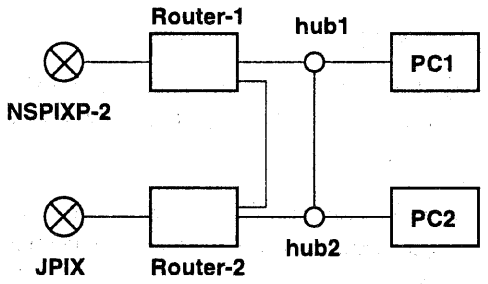


図 4: 完全二重化構成

易に入手できるようになってきたため、1999 年春頃に CPU の更新を計画したい。

サーバの設置場所は第一種電気通信事業者のハウジングであり、電源については発電設備を含めて非常に高い信頼性があると考えられるが、現在の構成上の問題点として

- ルータが一台のみであり、ルータ障害時や保守時のサービス提供に問題がある。
- ルータとサーバを接続する Ethernet Hub に障害が発生すると、サービス提供が不可能になる。

などの単一障害には耐えられない構成になっていることが挙げられる。

この対策としては、ルータを二重化することが考えられる。幸い、我が国の商用 IX である JPIX が近隣に存在するため、それとの接続も行うことによってルータのみならず、単一 IX の障害に対しても耐故障性を得ることができる。この場合、図 4 に示すような構成をとることにより、電源や空調以外の構成要素に関して、単一故障が発生してもサービスの提供を継続することができる。

このハードウェア構成の場合、Router-1 は PC1 に、Router-2 は PC2 に問い合わせを送るようにし、また PC1 は Router-1 に、PC2 は Router-2 に対して default 経路を設定する。さらに Router-1 と Router-2 を接続し、Router-1

が PC1 からの経路を失ったとしても Router-2 経由で PC2 を利用して、サービスを継続できる。正常時には NSPIXP-2 方面からのネームサーバへの問い合わせは PC1 が、JPIX 方面からの問い合わせには PC2 が対応するため、負荷分散も図ることができる。現在はハードウェア的には図 4 の構成になっており、ソフトウェア的な調整の確認作業中である。

## 謝辞

アジア太平洋地域用サーバの設置場所として NSPIXP-2 を推薦して頂いた IEPG 諸氏及びその決定を頂いた IANA に感謝します。ルータを貸与して頂いている日本シスコシステムズ（株）および日本インターネットエクスチェンジ（株）および "M" サーバの運用チームの諸氏に感謝します。

## 参考文献

[1] P. V. Mockapetris. Domain Names – Concepts and Facilities. RFC 1034, November 1987.

[2] P. V. Mockapetris. Domain Names – Implementation and Specification. RFC 1035, November 1987.

[3] B. Manning and P. Vixie. Operational Criteria for Root Name Servers. RFC2010, October 1996.