Root DNSの運用について

加藤　朗

東京大学情報基盤センター

インターネットはIPデータグラムの搬送を行うネットワークであるが，ユーザは32bitのIPアドレスではなくドメイン名によって相手ホストの識別を行う．そのため，ドメイン名からIPアドレスを知るシステム－DNS－は必要不可欠である．本稿では，DNSの要であるRoot DNSサーバの1つを安定に運用するための構成に関して述べる．

はじめに

DNS（ドメインネームシステム）[1]は，階層的に定義された名前空間から，指定したノードに対応するIPアドレスをはじめとする各種の情報を得るシステムである．DNSでは，分散的に配置されたDNSサーバに対して問合せを送ることによって，名前の解決を行う．このDNSの木の頂点で管理されている情報をRoot Zoneと呼び，DNSでは“.”で表現される．Root Zoneに対するDNSサーバはRoot DNSサーバと呼ばれており，インターネットの運用の要の1つである．このRoot DNSサーバの1つM.root-servers.netが筆者の手元で稼働を開始してから2年が経過しようとしている．本稿では，その背景，ハードウェアおよびソフトウェア構成について紹介する．

DNSについて

インターネットはIPデータグラムの交換を行うネットワークである．宛先の指定は，最終的にはIPデータグラムヘッダ中のIPアドレスによって行われる．現在のインターネットは32bitのIPアドレスを使用しているが，より広範囲への拡張を目指した128bitのアドレスを用いるIPv6の開発も盛んに行われている．インターネットに使用する際，ユーザがIPアドレスを直接指定することは好ましくない．シンボリックな表現で宛先を指定し，何らかの機構でIPアドレスに変換することが必要になる．

インターネットではドメインと呼ばれる階層的に定義された名前空間を用いている．この名前空間は木構造になっており，その根となるノードは特に“.”で表現され，Rootとも呼ばれている．Rootにはいくつかの子ノードが定義されており，ISO-3166に指定される英字2文字の国コード（我が国にはJPが割り当てられている）や，従来からアメリカ合衆国を中心に使われてきたCOMやEDUなどが含まれている．名前空間は定義しただけでは何の意味もない．実際にユーザがドメイン名を用いて通信の相手を指定した場合，そのIPアドレスなどの情報が得られ，実際にIPデータグラムを構成できなければならない．この名前空間から対応する情報を検索するシステムはネームサーバと呼ばれるが，ドメイン名の解決を

行うサーバはDNS－Domain Name System－サーバと呼ばれている．

ユーザがアプリケーションプログラムなどに対してドメイン名で宛先を指定した場合，通信に先立っての解決を行う．これは，あらかじめ設定された近所のDNSサーバにIPデータグラムを送り，名前の解決を依頼する[2]．DNSサーバは，対応する情報が自分で管理しているデータベース中に含まれていれば，その情報をただちに返答する．DNSサーバにない情報に関する問合せを受けた場合には，問合せを解決できそうな他のDNSサーバに問い合わせる．適当なサーバがない場合には，Root DNSサーバに問合せを行い，順次木構造を辿ることによって，名前を解決できるサーバに到達することができる．

経緯

Root DNSサーバは，実質的に全世界に広がっているインターネットが有効に稼働するための要の1つである．そのため，Root DNSサーバを安定に運用し，世界中のDNSサーバからの問合せに対応することは，インターネットの利用に対しては非常に重要である．

DNSの問合せにはUDPを主に用い，IPでのfragmentationを避け，またreassemblyバッファがオーバフローしないようにするため，メッセージ長は512byte以下と規定されている☆1．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数の上限は13である．そのうち従来から9つがサービスを提供してきた☆2．また，Root DNSサーバの重要性を鑑み，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☆3．

従来から稼働している9つのサーバは，1つがStockholmに置かれていた他はすべてアメリカ合衆国に設置されており，世界的なインターネットの発展に十分対応できなくなってきた．そのため，残り4つの枠のうち，2つを新たに設置することがIEPG（Internet Engineering Planning Group）で議論されてきた．1996年12月のIEPGにおいて，1台（“L”）をヨーロッパのアクティビティを考慮してロンドンのIXであるLINXに，もう1台をアジア太平洋地域を考慮し東京の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（Internet Assigned Numbers Authority）から提案されていたが，“J”はInterNICにおいてgTLD☆4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ただちに設定作業を行い，1997年8月22日付けで運用開始がIANAから正式にアナウンスされた．

従来からのRoot DNSサーバが“.”ゾーン以外に，COM，EDUなどのgTLDゾーンのサービスも同時に行っているのに対し，新規に設置された“L”Root DNSサーバおよび“M”Root DNSサーバの2つは“.”ゾーンのみのサービスを行っている．これはRFC2010の指定によるもので，Root DNSサーバを安定に運用するため，ホストに対する負荷および他のRoot DNSサーバがダウンした場合の負荷の集中に対するマージンを確保するという意味がある．また，この2つのサーバはそれぞれ主要なIXにルータを介して直接接続され，多くのISPと直接DNSトラフィックのやりとりが可能になっている．

またパフォーマンス上の理由から，Root DNSサーバは問合せに対する最終的な回答を返答する，いわゆる“recursive”サービスはサポートしていない．したがって，ほとんどの場合，各サーバは，Root DNSサーバから得られた情報を元に，他のDNSサーバへの問合せを継続することになる．また，同様な理由により，ゾーン転送も禁止されている．

ハードウェア構成

InterNICの運用経験によると，PCを用いる方式はワークステーションによる方式に比べて，ハードウェアは安価であるが，障害対応時などの人的コストがかかることが知られていた．経済的問題と安定運用という2つの目標を考慮し，WIDE Projectでは以下の方針でシステムを設計した：

- ハードウェアは性能価格比を考えてPCとするが，サーバ全体を2重化し，自動で切り替えを行うことによって，高い可用性を確保する．
- ルータを介してNSPIXP-2のGigaSwitchにDual Homeで接続することによって，片方のGiga Switchの障害時にもサービスの提供を行う．

M Root DNSサーバの構成は図-1のようになっている．PCは1997年4月からCPUにPen-

☆1 512byteより長いメッセージを要求できるようにするなどの拡張－EDNS0－が提案されている．

☆2 A.root-servers.net ～ I.root-servers.net

☆3 RFC2010の改訂案が現在IETF (Internet Enginnering Task Force) DNSOP WGで議論されている．

☆4 generic Top Level Domain

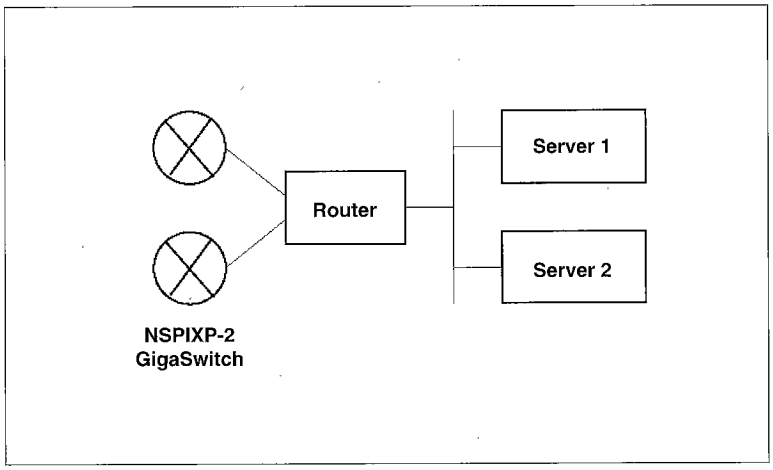

図-1　サーバの構成

tiumPro200MHzを使用したものを2台使用してきた．最近は障害が発生した場合に代替品の入手が困難になってきたことから，1999年3月にPentium-2 450MHzに交換している．

2重化の切り替え

サーバを単に2重化しただけでは，障害発生時にオペレータによる切り替えが必要になる．24時間365日のサービス提供を行うためには，切り替えを自動で行う機構が必要になる．比較的容易に実装できる切り替え方式として，経路制御による方式を用いることにした．

各サーバには，共通のサービス用アドレスをループバックインタフェースに割り振り，このサービス用アドレスをDNSサーバのアドレスとして公開している．その結果，DNSサーバに対する問合せは，サーバの物理アドレスではなく，このサービス用アドレスに対して到着する．

そこで，経路制御プロトコルによってこのサービス用アドレスをルータに対して広告する．サーバに障害が発生した場合，経路情報にタイムアウトが発生するため，ルータはサービス用アドレスに向かうパケットをバックアップサーバ側に切り替えることができる．

経路制御プロトコルとしては，BGPやOSPFを用いることもできるが，ルータにおけるタイムアウト値の設定の自由度が高いことから，RIP[4]を使用することにした☆5．

経路制御を用いた切り替えを行う場合，既存の経路制御プログラムを用いると，サーバのハードウェア障害時には経路情報の更新が途絶える．そのため，ルータにおいて対応する経路情報にタイムアウトが発生し，これによってバックアップ側の切り替えを行う．M Root DNSサーバでは，サーバプロセスのダウンにも対応するため，RIPパケットを送信するプログラムを別途作成した．このプログラムは，定期的にサーバプロセスの活動の有無をチェックし，プロセスダウンを検出した場合，ただちにmetric=16の経路を送出し，現在の経路が無効であることを通知する．この機構によって，タイムアウトを待たずに短時間でバックアップ側に切り替えを行うことができる．

M Root DNSサーバでは，通常は30秒に1回あるRIPアナウンスを10秒に1回にし，標準では180秒であるholddownするまでのタイマを40秒に変更して運用している．また，プライマリサーバ側からの経路がholddownした後でバックアップ側からの経路の広告があった場合，ただちに経路を切り替えるようにしている．毎回のRIPアナウンスに先立って，ネームサーバプロセスの存在をチェックするため，ネームサーバプロセスの異常などの場合には20秒以内に，またプライマリサーバにハードウェア的な障害が発生し，ダウンした場合でも50秒以内に切り替えを行うことができる．

サーバを2重化できることは，DNSサーバプログラムやOSのバージョンアップを頻繁に行わなければならない場合に特に有効である．プライマリ側の経路制御プロトコルにシグナルを送ることによって，サービスをバックアップ側に切り替えておき，作業終了後，再びシグナルを送ることによってプライマリ側へとサービスを復旧することができ，サービスの中断を発生させずに作業を行うことができる．

DNSの問合せは原則としてUDPを用いることになっている．そのため，両方のサーバに対してルータで問合せを分散させ，全体の性能を上げることも検討された．しかし，全体の0.9%程

☆5 ルータとして使用しているCiscoのソフトウェアでは，RIPのプロトコルの規定にかかわらず，タイマ値を変更することができ，この機能を利用している．

☆6 通常はIPデータグラムの宛先アドレスのみによって経路制御を行うが，それ以外のフィールド，たとえばプロトコルフィールドや送信元アドレスなどを参照して経路を決定すること．

度の問合せはTCPを利用しており，単純にルータで振り分けを行った場合には，TCPによる問合せに対応することはできない．ポリシールーティング機構☆6を利用することも検討されたが，ルータに対する負荷を考慮して見送られた．

運用

M Root DNSサーバへの問合せは，図-2に示すように，当初は1日平均で毎秒200件程度であったが，その後は徐々に増加し，最近では1日平均で毎秒710件程度に増加している．

最近のM Root DNSサーバへの問合せを分類すると，図-3に示すように，IPv4アドレスに関する問合せを示すAが65.2%と全体の2/3近くを占めており，次いでいわゆる逆引きに関連するPTR，電子メールのメールホストを示すMX，ネームサーバ情報であるNSの順になっている．IPv6アドレスに関する問合せAAAAは，現在のところ0.046%にとどまっている．

M Root DNSサーバは我が国で運用されているが，そのサービスを提供している範囲がどのように分布しているかは，Root DNSの運用を考える上で興味ある情報である．1999年7月6日に，00:00-24:00 UTCの24時間分の問合せパケットのヘッダ情報を収集したところ，29万台を超えるDNSサーバから，合計約6,700万回の問合せがあった．これをBGPの経路表（60,644経路，121,190パス）と照合すると，問合せ元は4,714のAS☆7に分散していることが分かった．ASを国別に集計すると，109カ国からの問合せがあり，その比率を図示すると，図-4のようになった．アメリカ合衆国からの問合せが全体の42%を占めており，次いで韓国の9.1%，我が国の8.7%となっている．図示できなかった国をISO-3166の国別コードで頻度順に並べると，FR，DK，NO，IT，ZA，NZ，IE，SG，ES，RU，PH，BE，SE，CH，HK，FI，IN，MY，NL，AR，AT，TH，TR，ILとなる．

また，この日における問合せの時刻変動を，図-5に示す．これは1分単位で平均化したものをプロットしたものであるが，午前2時から3時にピークがあるほかはあまり変動はなく，24時間絶え間なく問合せが到着していることが分かる．

☆7 Autonomous Systemの略．インターネットの経路制御の単位で，ある程度の規模のプロバイダに対応していることが多い．16bitのAS番号で識別され，たとえば，WIDE InternetのAS番号は2500である．

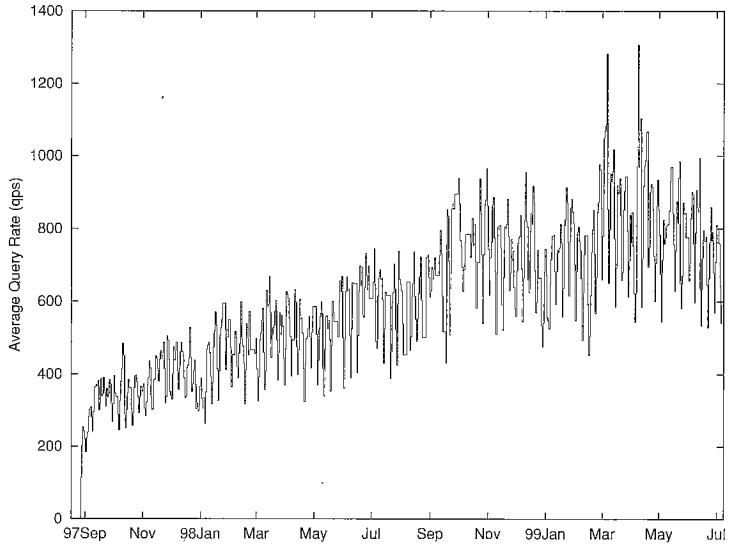

図-2　M Root DNSサーバの問合せ件数

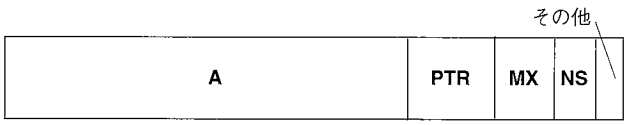

図-3　問合せ種別

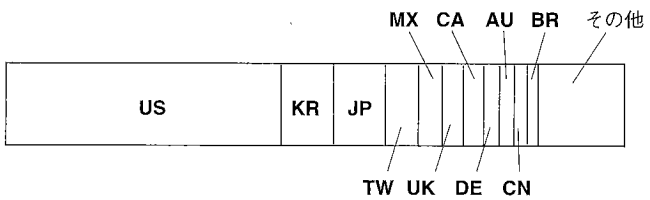

図-4　M Root DNSサーバの問合せ元の国別分布

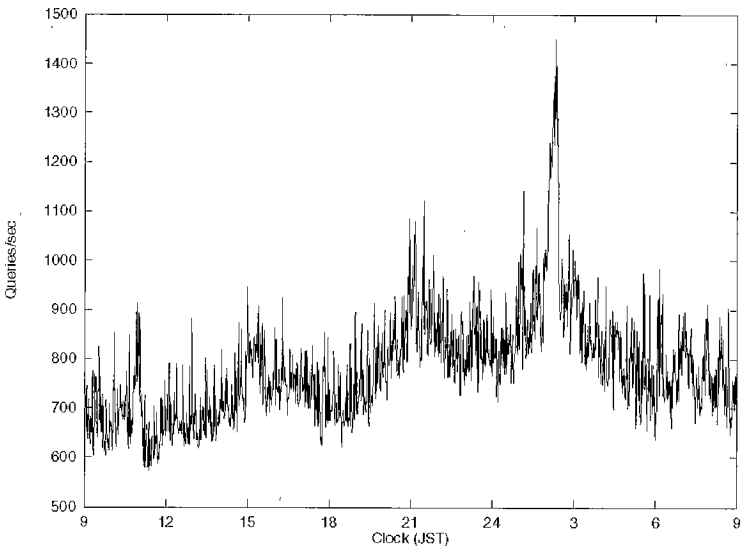

図-5　M DNSの問合せの時刻変動

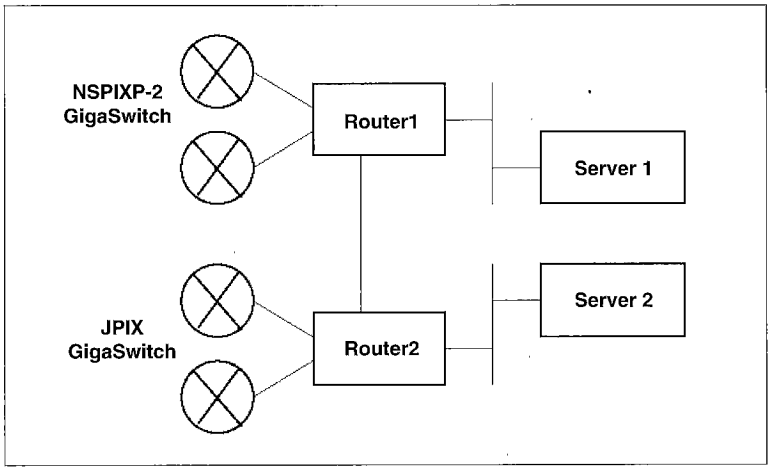

図-6　M DNS サーバの拡張

考察

サーバの構成を自動切り替えが可能な2重系とし，ネームサーバプロセスの異常終了時にも切り替えを行うようにしたため，稼働開始時から，システムとしては非常に高い可用性が得られている．幸い，現在までハードウェアによる障害は発生していない．

ネームサーバプログラムのあるバージョンにはメモリ管理に若干問題があり，1週間程度の運用でプロセスがダウンするという現象があった．この場合でも，自動的にバックアップ側のシステムに切り替わり，長時間のサービス断を避けることができた．また，このような構成をとっているため，サービスを停止せずにサーバソフトウェアやOSのバージョンアップを行うことができ，また，サーバのハードウェアの交換も実施することができた．

本システムは現在RIPを用いたサーバの切り替え機構を用いて運用しているが，他の経路制御プロトコルを用いても差し支えない．しかしOSPFは複雑であり，既存のプログラムを流用したとしても，安定度や，ネームサーバプロセス異常時に対応するための改良が必要になるという問題もある．BGPは，定常時にはKeepAliveメッセージを交換するだけでよいためオーバヘッドは小さいが，Root DNSシステムのような限定された環境では，RIPに比べて格段優れているとはいえない．

本システムはバッテリバックアップおよび自家発電機が装備されている環境で運用されており，またNSPIXP-2のFDDIスイッチにDual Homeで接続しているため，ルータ以外のハードウェアは2重化されていることになる．しかし，ルータは2重化されていないため，ルータのハードウェア障害時にはサービスが停止する．また，ルータのバージョンアップやソフトウェアの問題で再起動が必要になった場合，2分程度のサービス中断が発生するという問題がある．

Root DNSサーバは全部で13台が稼働しており，その1/3がダウンしたとしても，インターネット全体の運用に深刻な影響はないが，サービス中断はないに越したことはない．幸い，近くに商用のIXであるJPIXがあるため，図-6のようにルータを2台にし，NSPIXP-2とJPIX両方に接続している複数のISPにM Root DNSサーバに関するトラフィックの通過をお願いすることによって，ルータを含めた完全な2重化を行うことができる．

2台のルータはIBGPによって経路情報の交換を行い，直結しているサーバの障害を検出した場合には，問合せをルータ間リンクを利用して他方へ送る．この構成において，すべての隣接ASがNSPIXP-2とJPIXのどちらか片方の経路を選択するような経路制御を実施した場合，2台のサーバが稼働状態にあれば，負荷を分散して応答性を向上させることができる．この場合には，ポリシルーティングを導入することなく，TCPを用いた問合せにも対応することができる．現在，この拡張作業中であり，1999年秋からは現状と同じプライマリ・バックアップの運用を行う予定である．また，実験室および実機での評価を行った後で，負荷分散運用も考慮したい．

謝辞

アジア太平洋地域用サーバの設置場所としてNSPIXP-2を推薦していただいたIEPG諸氏およびその決定をいただいたIANAに感謝します．ルータを貸与していただ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，さらに運用に関して助言いただいたRootサーバ運用グループの諸氏に感謝します．

参考文献

1) Mockapetris, P.V.: Domain Names − Concepts and Facilities, RFC1034 (1987).
2) Mockapetris, P. V.: Domain Names − Implementation and Specication, RFC1035 (1987).
3) Manning, B. and Vixie, P.: Operational Criteria for Root Name Servers, RFC2010 (1996).
4) Hedrick, C. L.: Routing Information Protocol, RFC1058 (1988).

(平成11年8月3日受付)