# 第 XXXVII 部

# M Root DNS サーバの運用

# 第 37 部
# M Root DNS サーバの運用

## 第 1 章　はじめに

インターネット上の資源は、木構造の名前空間であるドメイン名によって命名される。与えられたドメイン名から、IP アドレスなどの名前に対応した種々の情報を得る操作は名前の解決とよばれ、この名前解決を担当するシステムが DNS である。DNS では、名前空間は Zone とよばれる連続した部分空間に分割して管理が行われており、分散的なアルゴリズムによって名前の解決も行われる。木構造の頂点である Root に対応した Zone の解決を行う DNS サーバは、特に Root DNS サーバとよばれているが、DNS の名前の解決はキャッシュを多用してその効率改善を図っているものの、基本的には名前の解決は Root からスタートする。

DNS の問い合わせに TCP を用いることも可能であるが、サーバ側での状態保持が必要であることや、TCP セッションの確立までに余計な RTT が必要であることから、極力 UDP を用いて問い合わせを行うことが推奨されている。UDP ではメッセージのフラグメント化を避けるため、IP や UDP ヘッダを除いたメッセージ長は 512 byte に制限されている。Root DNS サーバの一覧を問い合わせる QTYPE=NS QNAME=“.” という問い合わせの応答が単一メッセージに収まる必要があるため、Root DNS サーバの台数にも上限があり、現在は 13 台で運用が行われている。

この 13 台の Root DNS サーバのうち、M とよばれるサーバは、1997 年 8 月から WIDE Project によって運用が行われている。Root DNS サーバはインターネットにおける分散が制限されている資源の 1 つであるため、障害等によるサービス中断を最低限に抑える必要がある。そのため、M Root DNS サーバは、1997 年の運用開始時から、サーバの冗長構成を導入し、主サーバの障害時には副サーバが自動的にサーバ機能を提供するような運用を行っている。

## 第 2 章　構成

運用開始時には、M Root DNS サーバは、1 台のルータ Cisco4700M と 2 台のサーバ（PentiumPro 200 MHz）で構成され、NSPIXP-2 に対して FDDI で接続されていた。その後、1998 年にサービスを開始した商用 IX である JPIX から、接続およびルータ貸与の申し出があり、これを機にサーバシステム内部のネットワークを Ethernet から Fast Ethernet に更新した。この構成では、図 2.1 に示すように 2 台のルータが異なった IX に接続されており、単一故障点がない構成になっている。サーバも Pentium-II 450 MHz 2 台を経て、Pentium-III 1 GHz および Pentium-III 700 MHz を各 1 台という構成に更新された。

2001 年からは、第三の IX である JPNAP からポートおよびアクセス回線の提供を受け、また 2002 年 6 月からはサーバを Athlon XP-1900 を用いたもの 4 台（さらにバックアップ 1 台）に増強され、図 2.2 のような構成で運用された。

現在は後述のように、図 2.3 に示すような基本構成をユニットとした Anycast を用いている。

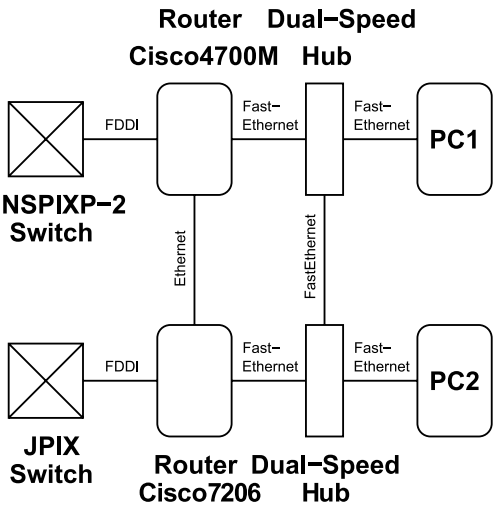


**図 2.1.**　単一故障点がない構成

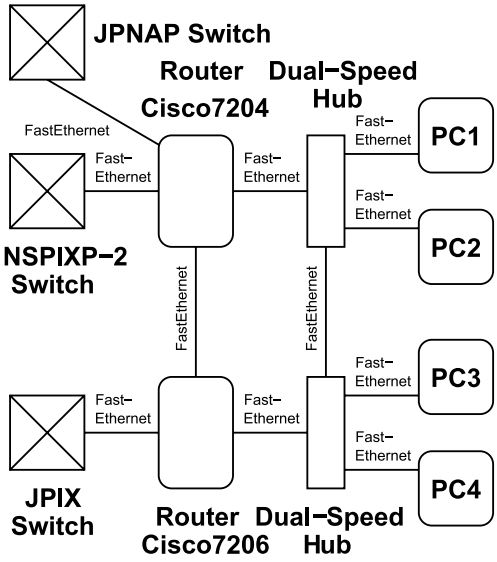


**図 2.2.**　2002 年からの構成

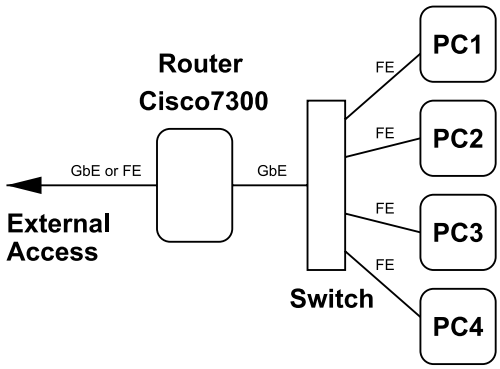


**図 2.3.**　Anycast 用基本構成

## 第 3 章　Backup サーバ

M Root DNS サーバは東京で運用されているが、東京で大災害等が発生した場合、サービス提供が不可能になる事態が想定される。そのため、2002 年 5 月、大阪にバックアップサーバの設置を行った。ルータは 1 台であるものの、NSPIXP-3 をはじめ、JPNAP/Osaka および JPIX/Osaka にそれぞれ接続されている。

当初は、誤動作を防ぐため、経路の広告をしないようにルータを設定しておき、東京での大災害発生時に手動でルータの設定を変更するようにしていた。しかし、2003 年夏の東京の電力危機によって、大規模な停電が発生することが懸念された。M Root DNS サーバは、商用電源の停電時でも、バッテリおよび発電機による電源のバックアップがなされているため、運用およびサービス提供には問題は発生しない。しかし、電源の切り替え時や発電機による運用中の不測の事故の発生を皆無にすることはできないため、大阪でのバックアップサーバは、サービスアドレスに対する経路広報を常時行うことにした。ただし、通常は東京の主サーバを優先するため、大阪のバックアップサーバは AS 番号を数回 prepend した経路情報を BGP で広告している。

## 第 4 章　Anycast

Root DNS サーバは 13 台と限られた存在であるため、インターネット上に普く分布させることはできない。そこで、同じデータを供給するサーバを複数インターネット上に設置し、それぞれのサーバは同一サービスアドレスでサービスを提供するようにする。このサービスアドレスを含む経路情報を BGP でアナウンスすることにより、BGP の経路選択ポリシに依存するものの、1 つのアドレスで複数台のサーバを運用することができる。この運用方法は RFC3258 “Distributing Authoritative Name Servers via Shared Unicast Addresses”[119] で定義されており、一般的には BGP Anycast とよばれている。

この Anycast に関しては、RFC が出版されたのは 2002 年 4 月であるが、最初の Internet Draft が IETF の DNSOP WG に提案されたのは 1999 年 10 月であり、その間議論が続けられてきた。

M Root DNS サーバでは、図 2.1 において、従来はすべての問い合わせを PC1 で処理し、PC1 がダウンした際には PC2 がバックアップする、という運用を行ってきた。2001 年 9 月にその運用方式を変更し、NSPIXP-2（および JPNAP）から届いた問い合わせは PC1 で、JPIX から届いた問い合わせは PC2 で処理を行うようにした。これは、地理的な分散はないものの、PC1/PC2 がインターネットのトポロジ的に異なった場所に接続されていることになり、限定された形式の Anycast であるということができ

る。これを “Anycsat in a Rack” とよんでいる。この構成では、両方のサーバがサービスに参加しており、全体としてのサーバの能力の向上が図られている。また、片方のサーバが停止した場合には、サーバ全体としての能力は低下するが、他方のサーバがサービスを提供することにより、継続的なサービスの提供を可能にしている。

2002年6月からは、図2.2に示した構成で、JPNAPおよびNSPIXP-2経由で到着した問い合わせはPC1あるいはPC2のいずれかで、JPIX経由で届いた問い合わせはPC3あるいはPC4のいずれかで処理されるようにした。このようにすることにより、負荷にばらつきがあるものの、4台のサーバでサービスが提供されることになり、DDoS攻撃などに対する耐久力を増すことができた。しかしながら、地理的には全体が1本のラックに収まっており、Anycastのもう1つの利点である、各顧客からサーバへのRTTの減少は実現されていなかった。

M Root DNSサーバでは、2004年に入り、Seoul（KR）およびParis（FR）での設置を行い、運用準備を進めてきた。このうち、Seoulに関しては、韓国で唯一のLayer-2 IXであるKINX—Korea Internet Neutral Exchange—のご協力を得て、2004年7月21日より運用を開始した。経路広告にBGPの`NO_EXPORT`属性を添付するいわゆるlocal anycastとして運用を行っているが、学術系のネットワークの収容を目的としてNCA—National Computerization Agency—が運用しているLayer-3 IXであるKIXでは、`NO_EXPORT`を外して学術系ネットワークに対して経路の広報を行っている。しかし、韓国での主要二大ISPであるKTおよびDaemonへの接続性がないため、現在、Seoulで処理されている問い合わせは毎秒50〜100程度と大きくない。

一方、ParisはTelehouse Europe、Renater、France Telecom、および Open Transit の協力を得て、Telehouse Voltaire にて 2004 年 9 月 1 日より運用を開始した。ここでは 2 つの独立な IX である Renater が運用する SFINX と France Telecom が運用する PARIX に接続しているほか、10 月からは TISCALI が独立に transit を提供して頂いている。現在は `NO_EXPORT` をつけて経路広告を行っているが、ヨーロッパ全域にサービスを提供しているtransit ISP とも多く peer しているため、ヨーロッパ全体をカバーしているわけではないが、そのサービスエリアはフランスに留まっていない。そのため、毎秒 4000 程度の問い合わせがある。

図 4.1 に 2002 年 6 月以降の、M-Root サーバに届いたすべての問い合わせの増加を示す。2004 年 9 月からの問い合わせ数が増加しているのは、主にParis で運用されている Anycast サーバへの問い合わせが、peer の増加や TISCALI からの transit の提供などの原因で増加しているためである。このうち、

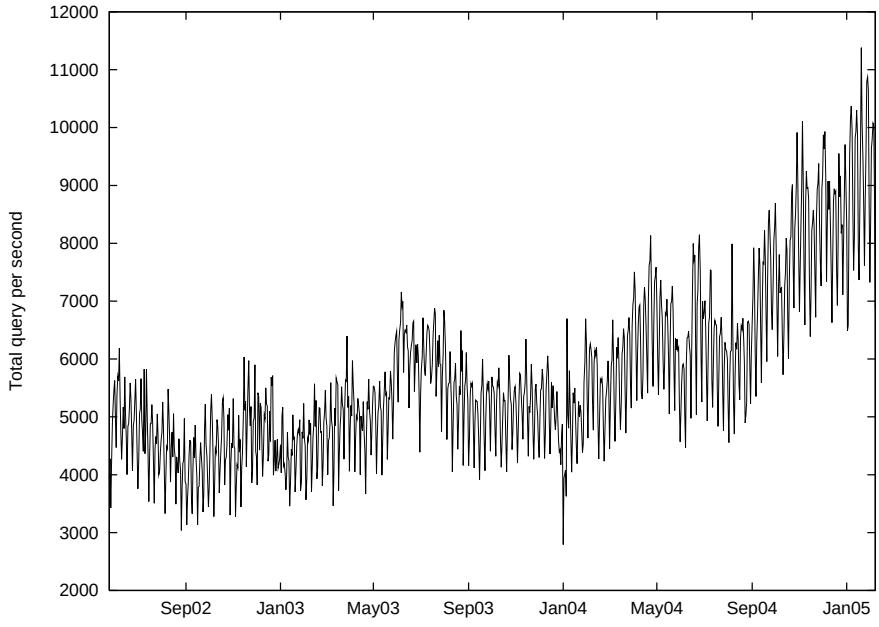


**図 4.1.** M-Root 全体の問い合わせ数の推移

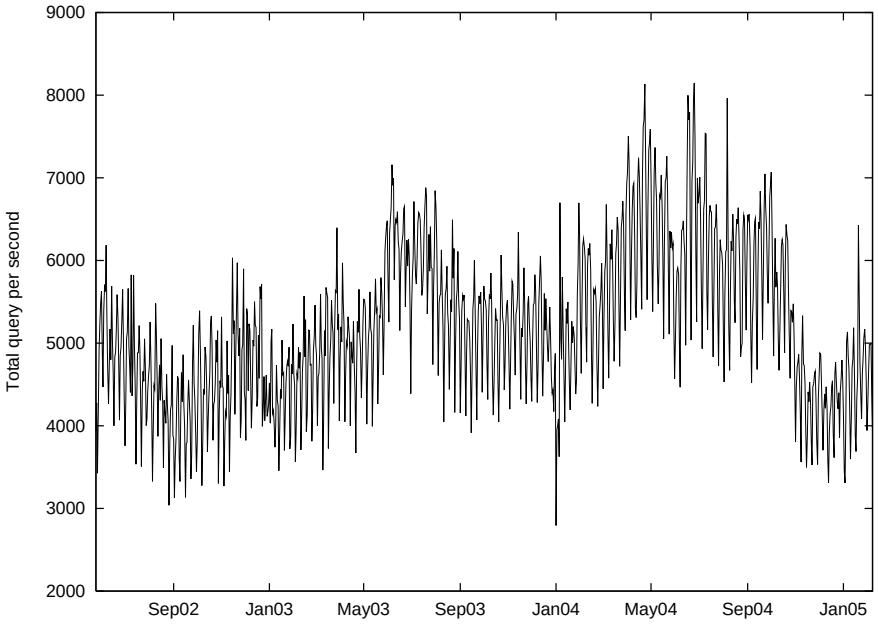


**図 4.2.**　東京のサーバでの問い合わせ数の推移

東京でのサーバは、DIX-IE、JPIX、JPNAP それぞれに独立したシステムが運用を担当する Anycast in a cage[1]になっているが、これらを合計した問い合わせ数の推移は図 4.2 の通りであり、Paris でのサーバ運用開始時に半減したが、その後やや増加している。

M-Root では、US からの問い合わせが多いことを考慮し、San Francisco でのサービス開始に向けて現在準備中である。MAE-LA および LAIIX は WIDE 経由で、その他は、PAIX/Palo Alto 経由でのサービスを予定しているが、U.S. でも大手の ISP は peering に前向きではないところも多く、調整が必要である。

## 第 5 章　他の Root DNS サーバ

2002 年 10 月 22 日早朝（日本時間）に発生した 13 台の Root DNS サーバをターゲットにした DDoS 攻撃をきっかけに、いくつかの Root DNS サーバでは、Anycast サーバの設置を図っている。とくに、ISC が運用している F Root DNS サーバでは、APNIC などとの協調により、精力的に Anycast サーバの設置を行っている。

2005 年 2 月時点での Root DNS サーバの設置状況を表 5.1 に示す。各サーバの最初の都市が元々運用されていた都市であり、それ以降は Anycast によるものである。Anycast の運用形式も各サーバで異なっており、たとえば、`C` では Cogent Communications のバックボーンにおける IGP による Anycast を実施しているほか、`F` では、Palo Alto、CA と San Francisco、CA のサーバはグローバルな経路広告を行っているのに対し、その他の `F` サーバは原則として、経路情報に `NO_EXPORT` BGP Community を添付することによるローカルな Anycast サービスを提供している。

1　ハードウェアの増強によりラック 1 本に収まらなくなった.

**表 5.1.** Root DNS サーバの設置状況

| サーバ | 設置都市 | | | |
|---|---|---|---|---|
| A | Dulles, VA | | | |
| B | Marina Del Rey, CA | | | |
| C | Herndon, VA | Los Angeles, CA | New York, NY | Chicago, IL |
| D | College Park, MD | | | |
| E | Mountain View, CA | | | |
| F | Palo Alto, CA | San Francisco, CA | Ottawa (CA) | San Jose, CA |
| | New York, NY | Madrid (ES) | Hong Kong (HK) | Los Angeles, CA |
| | Rome (IT) | Auckland (NZ) | Sao Paulo (BR) | Beijing (CN) |
| | Seoul (KR) | Moscow (RU) | Taipei (TW) | Dubai (AE) |
| | Paris (FR) | Singapore (SG) | Brisbanse (AU) | Toronto (CA) |
| | Monterrey (MX) | Lisbon (PT) | Johanesburg (ZA) | Tel Aviv (IL) |
| | Jakarta (ID) | Munich (DE) | Osaka (JP) | Prague (CZ) |
| G | Vienna, VA | | | |
| H | Aberdeen, MD | | | |
| I | Stockholm (SE) | Helsinki (FI) | Milan (IT) | London (UK) |
| | Geneva (CH) | Amsterdam (NL) | Olso (NO) | Bangkok (TH) |
| | Hong Kong (HK) | Brussels (BE) | Frankfurt (DE) | Ankara (TR) |
| | Buchareset (RO) | Chicago, IL | Washington D.C. | Tokyo (JP) |
| | Kuala Lumpur (MY) | | | |
| J | Dulles, VA | Mountain View, CA | Sterling, VA | Seattle, WA |
| | Amsterdam (NL) | Atlanta, GA | Los Angeles, CA | Miami, FL |
| | Stockholm (SE) | London (UK) | Tokyo (JP) | Seoul (KR) |
| | Singapore (SG) | | | |
| K | London (UK) | Amsterdam (NL) | Frankfurt (DE) | Athens (GR) |
| | Doha (QA) | Milan (IT) | Reykjavik (IS) | Helsinki (FI) |
| | Geneva (CH) | Poznan (PL) | Budapest (HU) | |
| L | Los Angeles, CA | | | |
| M | Tokyo (JP) | Seoul (KR) | Paris (FR) | |

## 第 6 章　まとめ

M Root DNS サーバは、7 年半以上にわたり安定的にサービスを提供してきた。特に冗長構成の導入により、サービスの停止をともなわずにサーバやサーバソフトウェアの保守作業が可能になったことは、サービス停止をともなう保守作業は 72 時間前に他の Root DNS サーバオペレータに連絡することが要請されていることを考えると、運用面で大きなメリットがある。また、数多くの ISP や IX の協力により、サーバそのものの安定運用に留まらず、インターネットの広い範囲に対して安定なサービスを提供できたことも特筆すべきである。今後は、Seoul や Paris に加えて San Francisco での Anycast サービスの提供およびその評価を通じて、DNS の安定運用に貢献していきたい。