第 XXXIII 部

M Root DNS サーバの運用

第 33 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指定される。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呼ばれるが、この名前解決を担当するシステムが DNS — Domain Name System — である。

DNS では、名前空間は Zone と呼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が行われる。木構造の頂点である Root Zone の解決を行う DNS サーバは、特に Root DNS サーバと呼ばれており、DNS の名前解決にとって非常に重要である。特に DNS での UDP を用いた場合のメッセージ長の制約から、多数の Root DNS サーバを設定することはできない。DNS ではキャッシュを多用することによって効率を改善するとともに、Root DNS サーバ等の上位ドメインに対応する zone を担当するサーバへの問合わせを減らすような努力がなされているが、Root DNS サーバが重要な存在であることには変わりはない。

Root DNS サーバは現在 `A.ROOT-SERVERS.NET`〜`M.ROOT-SERVERS.NET` という 13 システムで運用が行われている。このうち、`M.ROOT-SERVERS.NET` は、1997 年 8 月から WIDE Project によって運用されている。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一つであるため、障害等によるサービス中断を最低限に押さえる必要がある。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、ルータ Cisco4700M を一台と二台のサーバ（PentiumPro 200 MHz）で構成され、NSPIXP-2 に対して FDDI で接続されていた。サーバは primary/backup で運用され、primary サーバがダウンすると 1 分以内に backup に問合わせが届くようになり、サービスの中断を最小限度に保つことができるようになっていた。

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 FastEthernet に更新した。この構成では、図 2.1 に示すように二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 二台を経て、Pentium-III 1 GHz および Pentium-III 700 MHz を各一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

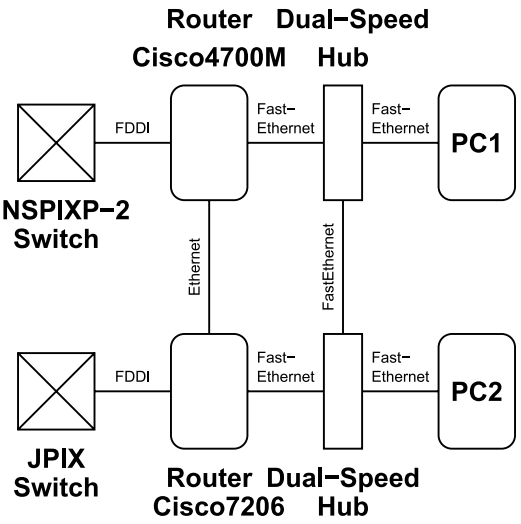


図 2.1.　単一故障点がない構成

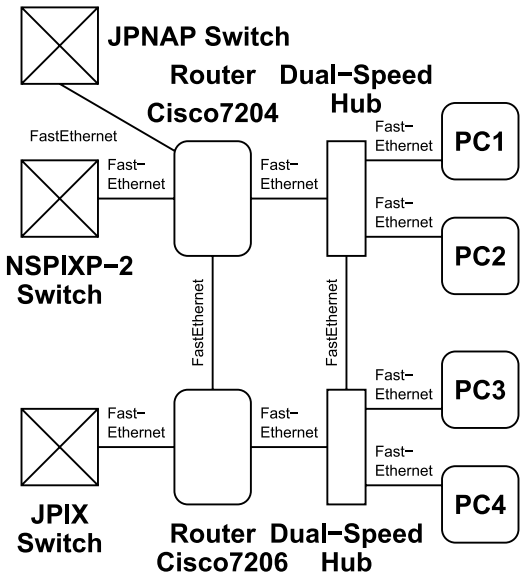


図 2.2.　2002 年から構成

四台（さらにバックアップ一台）に増強され、図 2.2 のような構成で運用された。

現在は後述の様に、以下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。

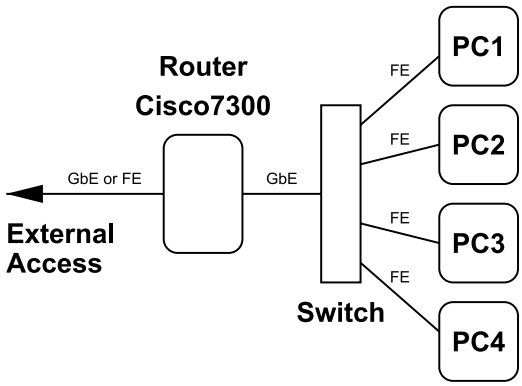


図 2.3.　Anycast 用基本構成

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。ルータは一台であるものの、NSPIXP-3 を始め、JPNAP/Osaka および JPIX/Osaka にそれぞれ接続されている。

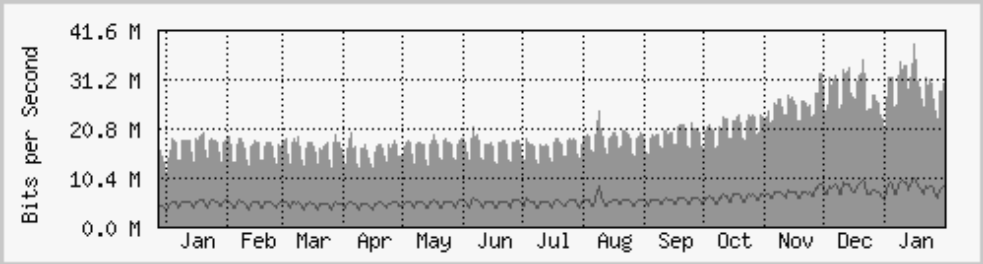
当初は、誤動作を防ぐため、経路の広告をしないようにルータを設定しておき、東京での大災害発生時に手動でルータの設定を変更するようにしていた。しかし、2003 年夏の東京の電力危機によって、大規模な停電が発生することが懸念された。M Root DNS サーバは、商用電源の停電時でも、バッテリおよび発電機による電源のバックアップがなされているため、運用およびサービス提供には問題は発生しない。しかし、電源の切り替え時や発電機による運用中の不測の事故の発生を皆無にすることはできないため、大阪でのバックアップサーバは、サービスアドレスに対する経路広報を常時行うことにした。ただし、通常は東京の主サーバを優先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

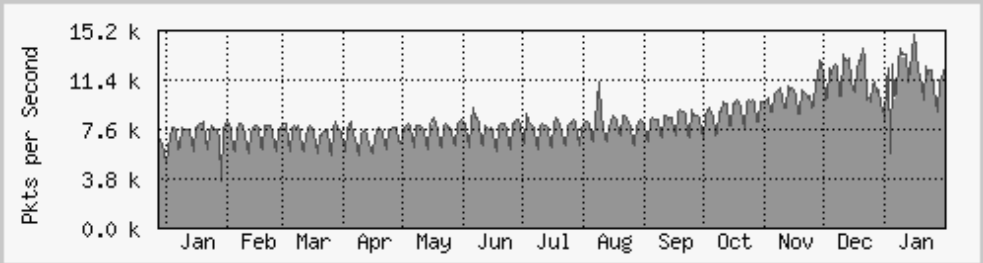
Root DNS サーバは 13 台と限られた存在であるため、インターネット上に普く分布させることはできない。そこで、同じデータを供給するサーバを複数インターネット上に設置し、それぞれのサーバは同一サービスアドレスでサービスを提供する様にする。このサービスアドレスを含む経路情報を BGP でアナウンスすることにより、BGP の経路選択ポリシに依存するものの、一つのアドレスで複数台のサーバを運用することができる。この運用方法は RFC3258 “Distributing Authoritative Name Servers via Shared Unicast Addresses”[83] で定義されており、一般的には BGP Anycast と呼ばれている。

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、従来は全ての問合わせを 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、という運用を行なってきた。2001 年 9 月にその運用方式を変更し、NSPIXP-2（および JPNAP）から届いた問合わせは



(a) トラフィックの推移



(b) パケット数の推移

図 4.1. M-Root 全体の問合わせ数の推移

PC1 で、JPIX から届いた問合わせは 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” と呼んでいる。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 経由で到着した問合わせは PC1 あるいは PC2 のいずれかで、JPIX 経由で届いた問合わせは 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、負荷にはばらつきはあるものの、四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、地理的には全体が一本のラックに収まっており、Anycast のもう一つの利点である各顧客からサーバへの RTT を減少することができること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）での設置を行ない、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 として運用を行なっているが、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なっている。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 2004 年 9 月 1 日より運用を開始した。ここでは二つの独立な IX である Renater が運用する SFINX と France Telecom が運用する PARIX に接続している他、10 月からは TISCALI が独立に transit を提供して頂いている。現在は多くの ISP に対して `NO_EXPORT` をつけて経路広告を行なっているが、幾つかの ISP に対しては `NO_EXPORT` なしに経路広告をしている。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、そのサービスエリアはフランスに留まっていない。このため、毎秒 4000 程度の問合わせがある。

San Francisco は WIDE San Francisco NOC に設置されており、WIDE とは別な FastEthernet で

PAIX/Palo Alto に接続されている。WIDE の Los Angeles での upstream である AS701 からのトラフィックは東京に送るのではなく San Francisco で処理されている。また、アメリカ合衆国の研究教育ネットワークである Abilene とは IPv6 による PAIX 上の peer をしているが、2006 年夏に IPv4 での peer を追加した。これによって、アメリカ合衆国の主な大学からの M-Root DNS Server への問合わせは TransPAC 等を経由して東京で処理されるのではなく、San Francisco で処理されるようになり、RTT の改善に貢献している。

図 4.1 に M-Root 全体に対するトラフィックの最近の推移を示す。2006 年の前半までには大きな変化は起っていないが、10 月以降トラフィックが増加している。これは、特に β 版を含む Windows Vista の IPv6 サポートによる影響ではないかと考えられている。

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

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)
	Amsterdam (NL)	Barcelona (ES)	Nairobi (KE)	Chennai (IN)
	London (UK)	Santiago (CL)	Dhaka (BD)	Karachi (PK)
	Torino (IT)	Chicago, IL	Buenos Aires (AR)	Caracas (VE)
G	Colombus, OH			
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)	Palo Alto, CA	Jakarta (ID)	Wellington (NZ)
	Johanesburg (ZA)	Perth (AU)	San Francisco, CA	New York, NY
	Singapore (SG)	Miami, FL	Ashburn, VA	Mumbai (IN)
	Beijing (CN)			
J	Dulles, VA	Mountain View, CA	Sterling, VA	Seattle, WA
	Amsterdam (NL)	Atlanta, GA	Los Angeles, CA	Miami, FL
	Sunnyvale, CA	Stockholm (SE)	London (UK)	Dublin (IR)
	Tokyo (JP)	Seoul (KR)	Singapore (SG)	Sydney (AU)
	Sao Paolo (BR)	Brasilia (BR)	Toronto (CA)	Montreal (CA)
K	London (UK)	Amsterdam (NL)	Frankfurt (DE)	Athens (GR)
	Doha (QA)	Milan (IT)	Reykjavik (IS)	Helsinki (FI)
	Geneva (CH)	Poznan (PL)	Budapest (HU)	Abu Dhabi (AE)
	Tokyo (JP)	Brisbane (AU)	Miami, FL	Delhi (IN)
	Novosibirsk (RU)			
L	Los Angeles, CA			
M	Tokyo (JP)	Seoul (KR)	Paris (FR)	San Francisco, CA

攻撃をきっかけに、幾つかのRoot DNSサーバでは、Anycastサーバの設置を図っている。特に、ISCが運用しているF Root DNSサーバでは、APNIC等との協調により、精力的にAnycastサーバの設置を行っている。

2007年1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、例えば、`C`ではCogent Communicationsのバックボーンにおける IGPによるAnycastを実施している他、`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。

第6章　まとめ

M Root DNSサーバは、9年半以上に渡り安定的にサービスを提供してきた。特に多階層の冗長構成の導入により、サービスの停止を伴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、サービス停止を伴う保守作業は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。M-Root DNS Serverは JPRSと共同で管理運用が行われているが、今後とも各方面と協力の上、より安定なサービス提供に努めたい。