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呼ばれ、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 サーバと呼ばれているが、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 のプロトコルとしては 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、メッセージ長は IP や UDP ヘッダを除いて 512 byte に制限されており、このため、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 と呼ばれるサーバは、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、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 [342]。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 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 3 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が行われている。

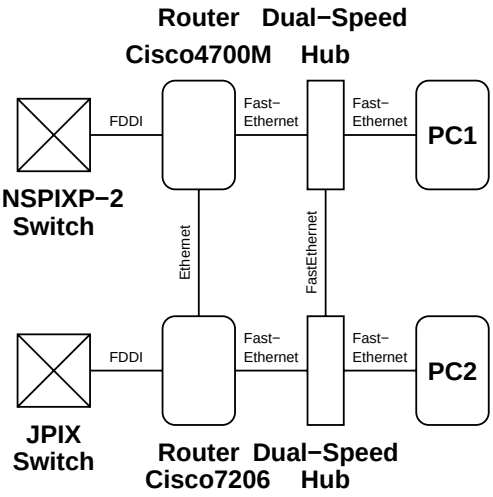

図 2.1.　単一故障点がない構成

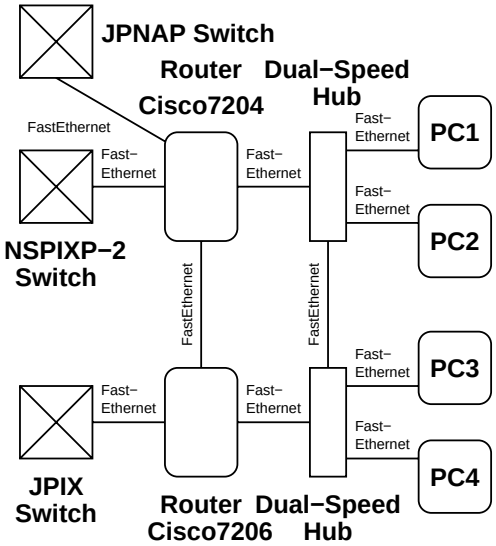

図 2.2.　現在の構成

第 3 章　トラフィック

運用開始時からの 1 日平均のトラフィックは図 3.1 に示す通り、運用開始時は 600 qps 程度であったが、2000 年から 2002 年にかけてほぼ線形に増加した。しかし、2003 年に入るとトラフィックの増加は収まりほぼ一定になっている。

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、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”[97] 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、議論が続けられてきた。M Root DNS サーバでは、2001 年 9 月に、図 2.1 において、NSPIXP-2（および JPNAP）から届いた問合わせは 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 Anycast であるということができる。これを “Anycast 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。

図 2.2 に示した構成で、当初は PC1 と PC3 が、JPNAP および NSPIXP-2 からの、あるいは JPIX からの問合わせを処理し、PC2 および PC4 はそれぞれの主サーバに対するバックアップサーバとして機能していた。2003 年 8 月に、PC2 および PC4 も PC1 と PC3 と同時にサービスを提供するように設

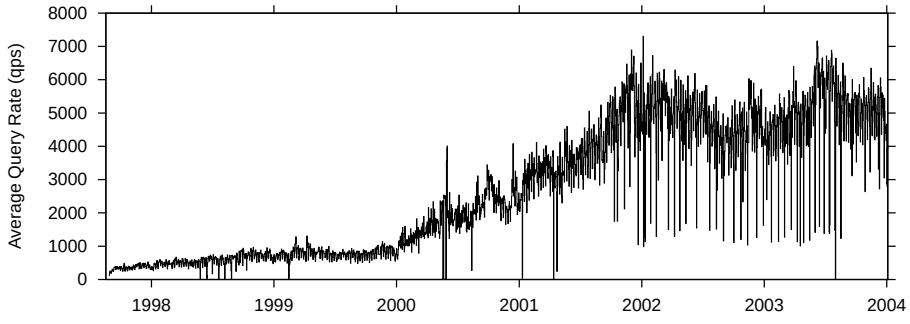

図 3.1.　bind が報告した問合わせの推移

定を変更した。つまり、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。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、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。

第 5 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 6 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、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。

2004 年 2 月時点での Root DNS サーバの設置状況を表 6.1 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 Communica-

表 6.1.　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)		
G	Vienna, VA			
H	Aberdeen, MD			
I	Stockholm (SE)	Helsinki (FI)	Milan (IT)	London (UK)
J	Dulles, VA	Mountain View, CA	Sterling, VA	Seattle, WA
	Amsterdam (NL)	Atlanta, GA	Los Angeles, CA	Miami, FL
	Stockholm (SE)	London (UK)		
K	London (UK)	Amsterdam (NL)	Frankfurt (DE)	
L	Los Angeles, CA			
M	Tokyo (JP)			

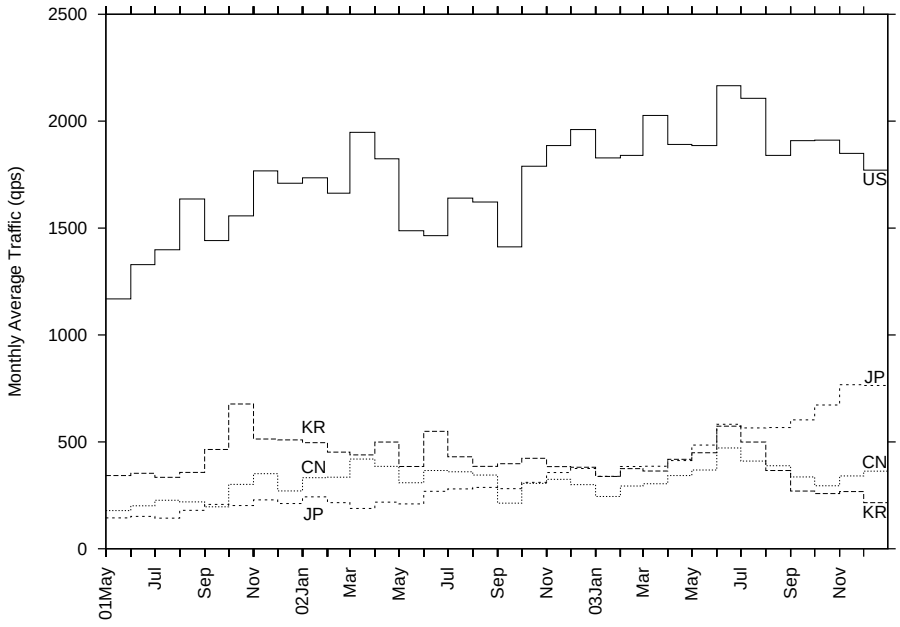

図 6.1.　M Root DNS サーバへの問合わせ数の推移

tions 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。

M Root DNS サーバは、現在は東京のみで運用が行われているが、Seoul (KR) および Paris (FR) での Anycast サーバの運用の準備が行われており、Seoul に関しては KINX — Korea Internet Neutral Exchange —の、また Paris に関しては Telehouse Paris および France Telecom、Renater の協力により 2004 年 3 月の運用開始を予定している。また M Root DNS サーバに到達する問合わせの 40%以上が U.S. からのものであるため、Palo Alto に Anycast サーバを設置する計画もある。

ところで、Anycast サーバ、特に F Root DNS サーバのアジア太平洋地区における精力的な設置に対する影響であるが、M Root DNS サーバに送信した問合わせを発信元別に分類した場合 [149, 341]、U.S.、韓国、日本、中国の上位 4 ヵ国の問合わせの推移を図 6.1 に示す。F Root DNS サーバの Seoul (KR) での運用は 2003 年 8 月に始まり、また Beijin (CN) は同 10 月に始まったことを念頭に図 6.1 をみると、韓国からの問合わせはピークのおよそ半分に減少しており、Anycast の効果を知ることができる。中国はそれほど減少していないが、これは中国では規制によって電気通信事業者の免許がない場合には IX に直接接続することが許されていないため、Anycast サーバのサービス領域が限定されていることが想定される。

第 7 章　まとめ

M Root DNS サーバは、6 年以上に渡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。今後は、Seoul や Paris 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。