第 XXXVI 部

M Root DNS サーバの運用

第 36 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（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、図 1.1 に示すような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、とくに 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はかっているものの、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が 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。図 1.2 に QTYPE=NS QNAME=“.” という問い合わせの結果を示す。これが 512 byte 以内に収まる必要がある。

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 プロジェクト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、障害など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。

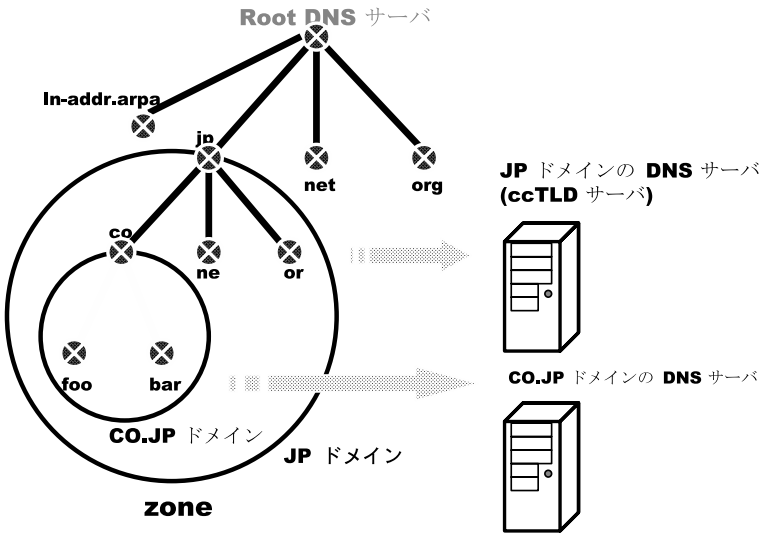

図 1.1.　DNS のデータ空間

```
; <<>> DiG 9.2.4 <<>> ns .
;; global options:  printcmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 63858
;; flags: qr rd ra; QUERY: 1, ANSWER: 13, AUTHORITY: 0, ADDITIONAL: 13

;; QUESTION SECTION:
;.                              IN      NS

;; ANSWER SECTION:
.                       164672  IN      NS      F.ROOT-SERVERS.NET.
.                       164672  IN      NS      G.ROOT-SERVERS.NET.
.                       164672  IN      NS      H.ROOT-SERVERS.NET.
.                       164672  IN      NS      I.ROOT-SERVERS.NET.
.                       164672  IN      NS      J.ROOT-SERVERS.NET.
.                       164672  IN      NS      K.ROOT-SERVERS.NET.
.                       164672  IN      NS      L.ROOT-SERVERS.NET.
.                       164672  IN      NS      M.ROOT-SERVERS.NET.
.                       164672  IN      NS      A.ROOT-SERVERS.NET.
.                       164672  IN      NS      B.ROOT-SERVERS.NET.
.                       164672  IN      NS      C.ROOT-SERVERS.NET.
.                       164672  IN      NS      D.ROOT-SERVERS.NET.
.                       164672  IN      NS      E.ROOT-SERVERS.NET.

;; ADDITIONAL SECTION:
A.ROOT-SERVERS.NET.     139419  IN      A       198.41.0.4
B.ROOT-SERVERS.NET.     72916   IN      A       192.228.79.201
C.ROOT-SERVERS.NET.     139419  IN      A       192.33.4.12
D.ROOT-SERVERS.NET.     139419  IN      A       128.8.10.90
E.ROOT-SERVERS.NET.     146384  IN      A       192.203.230.10
F.ROOT-SERVERS.NET.     164649  IN      A       192.5.5.241
G.ROOT-SERVERS.NET.     148810  IN      A       192.112.36.4
H.ROOT-SERVERS.NET.     147636  IN      A       128.63.2.53
I.ROOT-SERVERS.NET.     139419  IN      A       192.36.148.17
J.ROOT-SERVERS.NET.     169613  IN      A       192.58.128.30
K.ROOT-SERVERS.NET.     164362  IN      A       193.0.14.129
L.ROOT-SERVERS.NET.     139419  IN      A       198.32.64.12
M.ROOT-SERVERS.NET.     139419  IN      A       202.12.27.33

;; Query time: 4 msec
;; SERVER: 133.11.124.164#53(133.11.124.164)
;; WHEN: Thu Feb 23 14:21:16 2006
;; MSG SIZE  rcvd: 436
```

図 1.2.　Root DNS サーバに対する Root ゾーン問い合わせ結果

第 2 章　構成

1997 年の運用開始時には、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[1]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[2]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 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[3]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。

現在は図 2.3 に示す基本構成を 1 つの Anycast ユニットとして、各 IX への接続を行っている。

1　http://nspixp.wide.ad.jp/
2　http://www.jpix.ad.jp
3　http://www.mfeed.ad.jp/

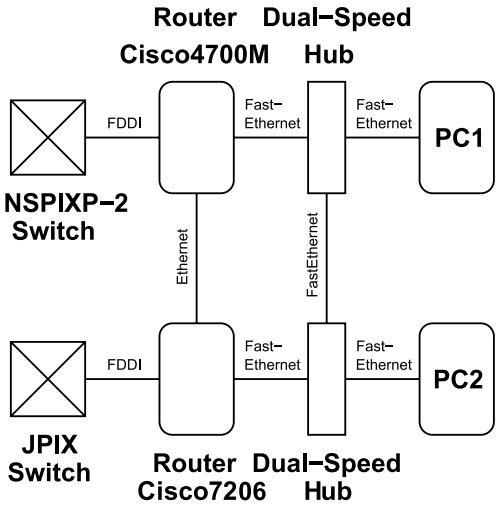

図 2.1. 単一故障点がない構成

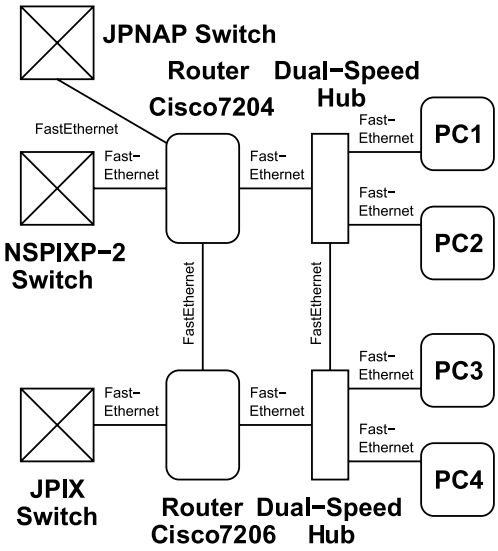

図 2.2. 2002 年からの構成

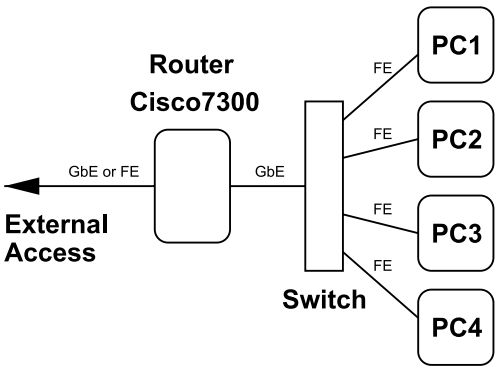

図 2.3. Anycast 用基本構成

第 3 章 Backup サーバ

M Root DNS サーバの中心となる運用拠点は東京であるが、東京で大災害などが発生した場合、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、大規模な停電によるサービス障害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、2003 年夏より、大阪でのバックアップサーバにて、サービスアドレスの経路広報を常時行うことにした。ただし、通常は東京の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 にて広告している。

第 4 章 Anycast

Root DNS サーバは 13 台と限られた存在であるため、容易に拠点を増加させることはできない。そこで、Anycast と呼ばれる技術を用いて、サービス拠点の増設を行っている。Anycast では、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 IP アドレスでサービスを提供するよう構成される。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”[103] 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 に示すように、従来は全ての問い合わせを 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はかられている。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。これにより、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（Distributed Denial of Service）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 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）でのサービス拠点設置を行い、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。また、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 50〜100 qps 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 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 に transit を提供してもらっている。現在は `NO_EXPORT` をつけて経路広告を行っている。しかし、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 peering 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、4000 qps 程度の問い合わせがある。

図 4.1 に 2002 年 6 月から 2005 年 2 月までに、M Root DNS サーバに届いた問い合わせの総量を示す。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、peering 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、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[4] になっている。

また、図 4.2 に、2005 年 2 月から 2006 年 2 月までに M-Root DNS サーバに届いた問い合わせの総量を示す。

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。

4　ハードウェアの増強によりラック一本に収まらなくなった。

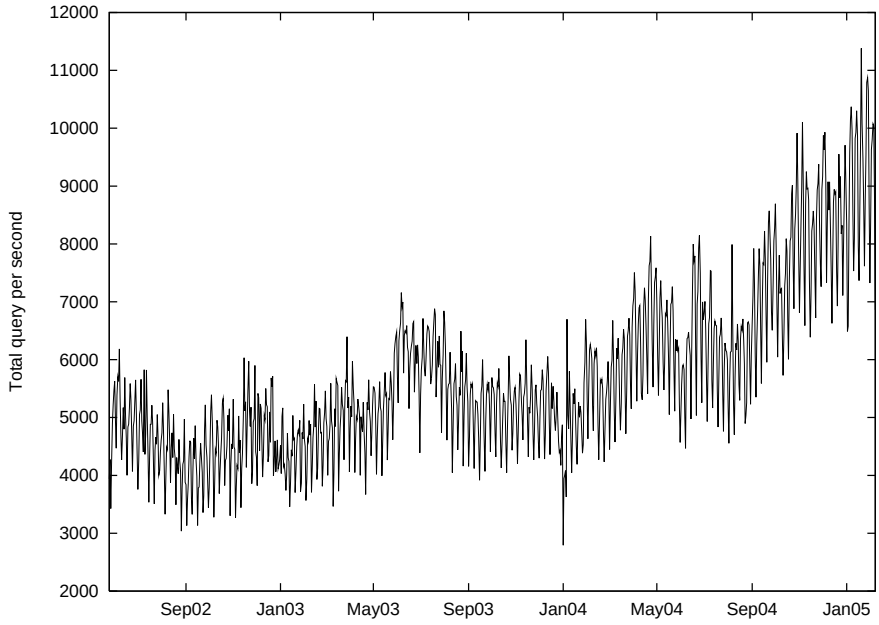

図 4.1. M-Root 全体の問い合わせ数の推移

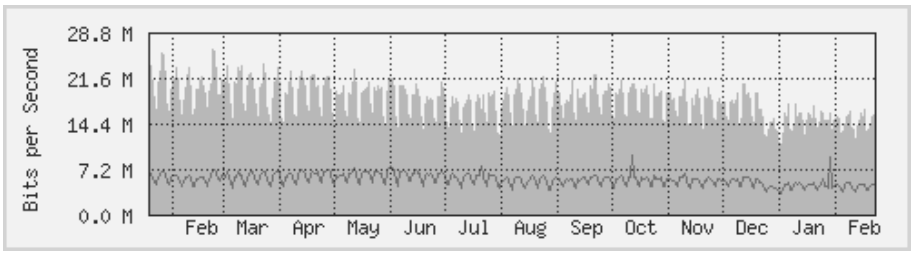

図 4.2. 2005 年度における M Root DNS 問い合わせ数の推移

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 サーバの設置をはかっている。特に、ISC が運用している F Root DNS サーバや、Autonomica が運用している I Root DNS サーバでは、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各サーバの中心運用拠点であり、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。

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 de Chile (CL)	Dhaka (BD)	Karachi (PK)
	Torino (IT)			
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)	Palo Alto, CA	Jakarta (ID)	Wellington (NZ)
	Johannesburg (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
	Stockholm (SE)	London (UK)	Tokyo (JP)	Seoul (KR)
	Singapore (SG)	Sydney (AU)		
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)	

第 6 章　まとめ

M Root DNS サーバは、8 年半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（RFC[2010][171]）ことを考えると、運用面に大きなメリットをもたらした。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。