

A Root DNS Server

Akira Kato



WIDE Project
kato@wide.ad.jp

Brief Overview of M-Root

- ☆ **Assumes basic knowledge on DNS**
 - Dr. Tatsuya Jinmei has introduced in Nov 19, 2004

What's Root Servers?

- ☆ **Start point of the DNS name resolution process**
 - Root servers serves the top node of name tree
 - " "
- ☆ **All DNS servers offering caching service**
 - Need to know all Root Servers
 - Names of each server
 - IP address of each server
 - Otherwise bootstrap problem happens
 - No DNS names are resolvable!
- ☆ **Root servers only provide information of**
 - Root servers
 - TLD servers
 - in-addr.arpa servers

When queries are sent to a Root Server?

- ☆ **At bootstrap (priming in bind terminology)**
 - DNS (bind) queries up-to-date list of Root Servers
 - Equivalent to the following command
 - `% dig @[a-m].root-servers.net . ns`
 - Their IP addresses are described in "root.cache" file
 - It might be obsoluted
 - Old Root still responds correctly or no response
 - Select the destination at random
- ☆ **When got a query with an uncached TLD**
 - TLDs of Valid queries are likely in the cache
 - Invalid TLD are likely subjects for Root query
 - Queries with Typo in the TLD portion
 - `.local/.domain/...`

Brief History

- ☆ **1987**
 - 7 servers with 10 IP addresses, all in U.S.
- ☆ **1990**
 - 8 servers with 11 IP addresses, one in SE
- ☆ **1995**
 - 9 servers with 9 IP addresses
 - Renamed to [A-I].root-servers.net
 - to allow maximum compression
- ☆ **1997 Jan**
 - J/K introduced at NSI (InterNIC)
- ☆ **1997 May**
 - L/M introduced at ISI (IANA)
 - K moved to London operated by RIPE/NCC
- ☆ **1997 Aug**
 - M moved to Tokyo operated by WIDE

Number of Root Servers

- ☆ **DNS transport depends on UDP**
 - UDP doesn't need circuit setup
 - Prompt response
 - Message size is limited to 512bytes
 - UDP/IP headers are extra
- ☆ **Priming response must fit into 512byte**
 - $SZ = 33 + 15*NS + 16*A$
 - With 13 NS, 13 A, response size is 436bytes
 - 15 is the absolute maximum number
 - 76bytes remaining with 13 Roots
 - This is a valuable space for AAAA records
 - $SZ = 33 + 15*NS + 16*A + 28*AAAA$

To serve better

- ☆ **Number of Root servers is limited**
 - Too small to install in every country
- ☆ **The powerful tool is anycasting**
 - Install the server in multiple locations
 - Sharing the same service IP address
 - Provide the same service in all servers
 - Advertise the prefix via BGP
 - Let routing system to choose the destination
 - It is likely to choose the nearest server

Anycasting

- ☆ **RFC3258**
 - All of the instances are managed by single entity
 - Point of contact is identical among all instances
- ☆ **Currently RFC3258 style anycasting is deployed**
 - at C/F/I/J/K/M
 - 84 instances from A-M are working now
- ☆ **See <http://www.root-servers.org/>**
 - For M-Root, see <http://m.root-servers.org/>

Which instance serves you?

☆ To identify, execute the following unix command

```
% dig @m.root-servers.net hostname.bind chaos txt
```

...

```
:: ANSWER SECTION:  
HOSTNAME.BIND.      0    CH   TXT   "M-d1"
```

- The string "M-d1" identifies the server instance
 - if not clearly states where it is located

Server Selection

☆ Which server to use if multiple candidates exist?

- TinyDNS in DJBDNS chooses one at random
- **BIND** chooses one with smallest RTT
 - RTT is decreased slightly for other candidates
 - Eventually be tried to know up-to-date RTT
 - Prefers the nearest server
- **Note that all servers are used, anyway**
 - But the frequency may vary very much

Benefit of Anycasting

- ☆ **In many cases, it provides smaller RTT**
 - Not always a case, unfortunately
 - But BIND prefers other letters in such a case
- ☆ **Enhance the total server capacity**
 - Multiple servers are working in parallel
- ☆ **Limit the holizon of DDoS attack**
 - Only a fraction of the instances get affected
 - Nearest ones of DDoS sources
 - Other instances work as usual

Issue of Anycasting

- ☆ **Anycasting may breaks in TCP**
 - A TCP session involes multiple packets
- ☆ **Some packets may delivered to other instances**
 - Yielding the TCP session breaks
 - When the routing topology changes
 - This rarely happens in short-live TCP sessions
 - When ISP performs per-packet load-balancing
 - TCP performance get suffered due to re-order
 - This is highly discouranged in general

Future enhancements

☆ IPv6 transport support

- July 2004 root zone includes AAAA glues for TLD servers
- No AAAA for Root server is defined
 - B/F/H/K/M is ready to respond in IPv6
 - But they are totally useless
 - No AAAA is defined in root-servers.net zone
- The way to include more than 2 AAAA is undefined
 - Need to evaluate the method carefully

☆ DNSSEC

- It is highly required for additional security in DNS
- Packet size is also an issue
- BIND9.3 supports DNSSEC based on latest spec
 - Need to evaluate its performance as well