DNS Anycast Operation of .JP

ICANN ccNSO @ Vancouver
30 Nov. 2005
Shinta Sato <shinta @ jprs.co.jp>
JPRS
Agenda

• Background
• Motivations
• .JP Anycast Overview
• Anycast management
Background

- IP anycast is...
  - A technology to share a single IP address in multiple servers
    - IGP anycast for inside AS
    - BGP anycast for outside AS
  - DNS service is one of the effective thing to introduce IP anycast
    - 1 packet udp transaction for both query and response
      (Response packet may fragment in EDNS0, but still no problem)
    - very short tcp session
  - IP anycast technology is now being deployed in authoritative name servers
    - Root servers (C, F, I, J, K, M)
    - Some TLD servers (.JP, .MX, .DE etc.)
BGP Anycast Overview

Incoming query packets from clients in AS5 will go to the nearest node #1 of AS1 via AS2.

Incoming query packets from clients in AS6 will go to the nearest node #2 of AS1 via AS3.

Both routers announce the same shared unicast IP address by BGP connection for a service address.
IGP Anycast Overview

A shared unicast IP address is assigned for the service, and is announced by the IGP.

Anycast node #1

AS 1

AS 2

AS 3

Incoming query packets from the clients will go to the nearest node after it gets in AS1.

A shared unicast IP address is assigned for the service, and is announced by the IGP.

AS 4

AS 5
Motivations

• Common motivations for using DNS anycast are,
  – Localize the DoS attack damages
  – Provide nameservers all over the world
  – IPv6 deployment
  – Simple maintenance and recovery
Localize the DoS attack damages

- IP Anycast can localize the DoS attack damages to the single node.
  - Other nodes will not be affected from the DoS attack
  - Only the nearest nodes from the DoS attacker will be damaged
  - In the DDoS case, if the attackers are gathering in the similar network, affects will be localized too.
Provide nameservers all over the world

• Placing more nameservers is one of the solutions to increase the stability of the DNS
• IP anycast can help to plan the placement of secondary servers
  – Adding a new anycast node improves the accessibility of the users
  – Users access only the nearest node
IPv6 deployment

• Adding IPv6 glue data in the higher level zone decrease the limit number of NS in less than 13
  – Number of NS is limited by the DNS response packet size of 512 octets
  – Serving AAAA (IPv6) information in the glue record require more data size in the additional section than A (IPv4) only
Simple maintenance and recovery

• IGP anycast can simplify server maintenance
  – Operator can stop individual server without outage of the service

• BGP anycast can simplify maintenance of the whole site
  – Operator can shutdown the BGP peer without outage of the service
  – Useful in the case of network troubles

• Able to rebuild the DNS node without thinking of other infrastructures placed in the same network
The current situation of .JP

- **JP DNS servers:**
  - 5 NSes
    - \{a,b,d,e,f\}.dns.jp
    - c.dns.jphas retired in Mar. 2005
  - Operated by 5 different organizations, with responsibility of JPRS
    - All organizations own their networks by their own AS numbers
  - Hold numbers of zones
    - .JP ccTLD zones (1 TLD and 63 SLDs)
      - 769,445 domains (1 Nov. 2005)
    - Also serve 339 of in-addr.arpa zones for JPNIC (NIR)
Introducing IP anycast servers to .JP

• Severe crisis of the power outage in Tokyo (2003)
  – JP DNS operators tried to move some of the servers out of Tokyo
    • Using IP address of their main network prevent us to change the location without changing the IP address at that time
    • This was the potential problem, which prevent us to recover the DNS without thinking of other infrastructures placed in the same network, even in the severe network trouble
  – JP DNS could not add more NSes
    • JP DNS operators were thinking of the deployment of IPv6 at that time
    • 4 IPv6 servers out of 6 NSes is the limit

Fortunately, the power outage did not happen
Introducing IP anycast servers to .JP (2)

• JP DNS took the following solution
  – Keep the number of NS in 6
  – Move to PI (Provider Independent) addresses and new ASNs if possible
  – Add more servers using IP anycast technology
    • Now we have servers in Tokyo, Osaka and US
Technical details of a.dns.jp
Concerns of IP Anycast management

- **IP address issues**
  - Anycast need PI address or unused /24 address block
    - ccTLD can have PI address blocks for their nameservers
  - Unicast address still needed for each anycast nodes
    - To update the zone data, to maintain the servers
  - At least 1 NS should remain in unicast (RFC 3258)

- **Budget issues**
  - IP anycast requires transit and / or IX connectivities for each nodes
  - Maybe expensive for individual service
    - This network serves only 1 IP address to the public

- **Measurement issues**
  - It is hard to know all the servers are up in anycast address
    - Checking unicast address is not enough
    - Multiple measuring address required
Nameserver configurations

• Multiple addresses are needed in a server
  – One for IP anycast service
  – One (or more) unicast address(es) for maintenance and zone update

• Not so much difference from unicast servers
  – in BIND9, following options should be considered to make zone updates to work
    • query-source
    • transfer-source
    • notify-source
Consideration points

• Local nodes and global nodes
  – Local nodes are for IX connections
    • No-export option in BGP peers
  – Global nodes are for transit connections
  – 2 global nodes and several local nodes may be good
  – Some trouble may occur by uRPF (unicast Reverse Path Forwarding)
    • Some ISPs use uRPF technology for very intelligent network filtering
Example of IP Anycast effect

- DoS like queries in Osaka node did not harm any in Tokyo node

Osaka node

Tokyo node
BCPs

- Some BCP activities exist
  - Distributing Authoritative Name Servers via Shared Unicast Addresses
    - RFC 3258
  - Operations of Anycast Services
    - draft-ietf-grow-anycast-02.txt
  - BGP Anycast Node for Authoritative Name Server Requirements
    - draft-morishita-dnsop-anycast-node-requirements-01.txt
Appendix: NS maximum number estimation

- DNS protocol has limitation in UDP response packet size
- More NSs make .JP DNS more reliable
  - Name compression
- Estimation for .JP (dns.jp)
- “preferred-glue a” and / or EDNS0 may moderate the limitation

<table>
<thead>
<tr>
<th>NS</th>
<th>AAAA</th>
<th>A</th>
<th>Add.</th>
<th>Judge</th>
<th>NS</th>
<th>AAAA</th>
<th>A</th>
<th>Addi.</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>AAAA x3, A x3</td>
<td>Nice</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>AAAA x4, A x3</td>
<td>OK</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>AAAA x3, A x4</td>
<td>Nice</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>AAAA x4, A x2</td>
<td>OK</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>5</td>
<td>AAAA x3, A x4</td>
<td>OK</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>AAAA x4, A x1</td>
<td>OK</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>AAAA x3, A x3</td>
<td>OK</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>AAAA &lt;4, A x0</td>
<td>NG</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>7</td>
<td>AAAA x3, A x2</td>
<td>OK</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>AAAA x5, A x1</td>
<td>OK</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>8</td>
<td>AAAA x3, A x1</td>
<td>OK</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>AAAA x5, A x0</td>
<td>Bad</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>9</td>
<td>AAAA x3, A x0</td>
<td>Bad</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>AAAA &lt;5, A x0</td>
<td>NG</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>10</td>
<td>AAAA &lt;3, A x0</td>
<td>NG</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>AAAA &lt;6, A x0</td>
<td>NG</td>
</tr>
</tbody>
</table>
Questions?

http://jprs.jp/

http://日本レジストリサービス.jp/