IETF DNSOP Working Group
Internet-Draft
Expires: September 28, 2006

Y. Morishita
S. Sato
T. Matsuura
JPRS

March 27, 2006

BGP Anycast Node Requirements for Authoritative Name Servers
draft-morishita-dnsop-anycast-node-requirements-03.txt

Status of this Memo

Copyright Notice

Abstract

IP anycast [1] is a technology to share one IP address for Internet services with multiple server nodes. It is now being deployed for improving service reliability, scalability, and stability.

Especially, "Global-Scope" IP anycast is now being deployed for authoritative name servers, typically for root servers.

RFC 3258 [2] describes a set of practices to apply IP anycast technology for authoritative name servers. And "Operation of Anycast Services" Internet-Draft [3] (hereafter, called "Anycast BCP") describes a series of recommendations and considerations for distribution of services using IP anycast.

On the other hand, operators of authoritative name servers can also refer to RFC 2182 [4] and 2870 [5] for general guidances on appropriate practices for authoritative name servers.

This memo describes the details of requirements and preconditions for making "Global-Scope" IP anycast nodes for authoritative name servers, with the conformance of the practices in RFC 2182, 2870, 3258 and Anycast BCP.

And this memo also describes our findings and experiences for making IP anycast nodes for us. Authors hope that it is useful for DNS operators that they will walk on the same way in the future, especially for TLD operators.

1. Introduction

By applying the IP anycast technology to DNS, name server operators can increase the number of authoritative name server nodes, and distribute them in topologically and geographically diverse locations, without violating the DNS protocol limitations [6] [7]. If IP anycast is appropreately operated for DNS server nodes, it improves the robustness against Denial-of-Service Attack and Distributed Denial-of-Service Attack and the reliability of DNS service. And it improves the DNS total response by decreasing RTT for authoritative name server, and distributes authoritative name servers' load, too.

However, IP anycast needs more careful operation for achieving the original goals and improvements. In the IP anycast technology, IP address does not specify the individual (real) end-point for the Internet communications any longer. Which means that the real communication peer can not be specified by the destination IP address only. It means some new pitfalls and risks the DNS service, for example, monitoring the availability of the service becomes more difficult, and data of all DNS server nodes need syncing.

For achieving the original goals of IP anycast, improving the robustness and the reliability of service. So by introduction of IP anycast, The reliability of the entire system must not decrease.

Especially, DNS is one of an important infrastructures of the

Internet, so, introducing "Global-Scope" IP anycast in critical DNS authoritative servers should be carefully.

This memo describes the details of requirements and preconditions for making "Global-Scope" IP anycast nodes for authoritative name servers, with the conformance of the practices in RFC 2182, 2870, 3258 and Anycast BCP.

And this memo also describes our findings and experiences for making IP anycast nodes for us. Authors hope that it is useful for DNS operators that they will walk on the same way in the future, especially for TLD operators.

In this memo, authors focus on "BGP anycast" for making "Global-scope" IP anycast nodes. It is the currently most popular technology for making IP anycast nodes and it is already used for making some root and TLD DNS server nodes. Authors think a part of the basic point of view can be applied to "Local-scope" IP anycast, too.

2. BGP Anycast and DNS Service

BGP anycast is a part of IP anycasting technology. It uses a shared IP address block and a shared AS number for BGP anycast nodes, and their nodes are placed in the Internet. Reachability of each nodes are served by BGP routing protocol [8].

Each anycast nodes propagate the routing information of the shared IP address block and AS number by BGP. Each BGP routers in the Internet choose 'nearest' node by BGP's best route selection algorithm. That is, the accesses to the shared IP address block will be distributed to the each anycast nodes, depending on the clients' locations and network topologies.

BGP anycast can control each anycast nodes by configuring as 'local node' or 'global node' using BGP's routing framework. Concretely, using the 'NO_EXPORT' [9] or 'NOPEER' [11] community, the 'local node' operators can limit distributing the routing information of anycast node only for the directly peering sites. Therefore, the 'local node' can localize the access to anycast node from directly peering sites. On the other hand, the 'global node' operators apply the normal BGP anycast for its node.

In this memo, authors focus on 'global node' as main target, but authors believe it can be applied as 'local node' too.

When one of BGP anycast nodes goes down, routing informations will be automatically recalculated. The datagrams to the anycast node are

automatically rerouted to other anycast nodes. Thus, BGP anycast can provide redundancy for the Internet services.

Current BGP anycast is hard to apply for longer-lived transaction service, because the instability of the dynamic routing protocol is harmful for them. But most of the DNS queries and answers are based on a single UDP packet, so, DNS is considered to be one of typical service in which BGP anycast can be applied.

In this memo, authors focus the critical DNS infrastructure, especially root and TLD DNS servers. So, authors only focus a case of "a single IP address for DNS service covered by a single exported route". It means advertisement and withdrawal of a single covering prefix can be tightly coupled to the availability of the single associated service, as described "4.8.1. Multiple Covering Prefixes" of Anycast BCP.

In this situation, DNS service providers need an exclusive IP address block which is a provider independent CIDR block and exclusive an AS number for making each anycast nodes.

3. Requirements and Preconditions for Critical DNS Server Nodes

As described before, IP anycast is one of the effective ways for improving the robustness and the reliability of service. And in recent critical authoritative name servers, especially for large TLDs, ability to handle more data than before, more frequent data updating, and higher reliability are required.

So, when BGP anycast technology is applied to their critical servers, carefulness must be necessary for their operators. Authors expect that the requirements and preconditions which described by this memo would be useful.

In this section, this memo describes requirements and preconditions for making BGP anycast nodes for authoritave name servers in the following two points of view, the Internet access service, and data center.

3.1. Choosing the Internet Access Service

For making BGP anycast nodes distributed in the wide area, it is important to make network environment with geographical and network topological diversity.

In case of making such network environment, each anycast nodes should have Internet connectivity from different Internet access service

provider (hereafter, called ISP) for ensuring network diversity.

And in case of ensuring the BGP connectivity, the owner of the authoritative name servers must consider the following preconditions and requirements to choose the Internet access services.

3.1.1. Reliability of the Backbone Network

When making a critical authoritative name server, higher reliability for ISP's network itself is needed. For implementing this, it is desirable for ISP itself to have owned and managed its backbone network.

In general an ISP, which owns and manages the backbone network itself, is expected to have stronger responsibility for network stability. Then it is expectable that the stability of a network is higher. Of course, it is not absolute requirement, but it will surely be one of the important elements.

3.1.2. Connectivity of Outside Area

In case of critical authoritative name servers, such as served for root and/or TLD zones, there are many accesses from not only its country and its local area but also outside area. Thus, connectivity for them must be needed. In the same reason of Section 3.1.1, it is desirable that an ISP which owns and manages the stable connectivity of all areas.

3.1.3. Peering

When ensuring highly reliable Internet connectivity, it is an important element for ensuring the diversity of Internet routes including multiple alternative paths. Moreover, providing DNS service to multiple ISP networks efficiently, it is desirable for the ISP to have many BGP peers with other ISPs, and they are much stable.

3.1.4. Connectivity for Provider Independent CIDR Block and AS Number

When making a set of BGP anycast node for critical authritative name servers, a provider independent CIDR block and an AS number must be prepare in advance, and they must be used for DNS service at each anycast nodes. It is also needed for making the multihomed connectivity.

In this case, the ISP must support propagating CIDR block and AS number for anycasting service to the Internet widely, and the ISP must provide connectivity for them from the Internet. Concretely, the ISP must provide "transit" service.

3.1.5. Connectivity for Administration and Data Synchronization

As in RFC 3258, an Internet connectivity which is different for IP anycasting must be needed for anycast node administration. And as in Anycast BCP, the synchronisation of data between anycast nodes will involve transactions between non-anycast addresses.

3.1.6. Connectivity for IPv6

RFC 3513 [10] prohibited host-based anycast in IPv6. But new version of RFC 4291 [12] removes this limitation and it RFC 3513 is generally obsoleted.

So, now there are no protocol limitations for using IP anycasting for IPv6 network by using the same technology as IPv4. But authors think that more experiences are needed for deployment of IPv6 anycasting.

That is, the anycast node owner should ensure IPv6 connectivity.

3.2. Choosing the Location

For choosing the BGP anycast node locations for critical authoritative name servers, RFC 2182 and 2870 can be refered for useful guidance on appropriate practices for them. By referencing them, when choosing the location for BGP anycast node for critical authoritative name servers, the owner of them must consider the following preconditions and requirements.

3.2.1. Providing Higher Security Level

To realize higher defense performance to physical destruction and/or the intrusion from the outside, the location must provide higher security level.

3.2.2. Providing Higher Redundancy of Electrical Power Supply

DNS service requires high continuity and stability, the location must provide higher redundancy of electrical power supply and urgent power supply equipment for emergency.

3.2.3. Providing Higher Tolerance Against Disasters

For the same reason of Section 3.2.2, the location must provide higher tolerance against disasters, for example fire, earthquake and other disasters.

3.2.4. Providing the Diversity of Locations

For ensuring tolerance and redundancy, the diversity of locations is needed. Concretely, even if a fatal disaster occurred at one location, the continuity of critical DNS service must be ensured.

4. Cost and Operational Issue

In the technical point of view, BGP anycast nodes can be made in numbers of locations. But it is not realistic to prepare them more than necessity. In general, to satisfy the preconditions and requirements which is previously described, BGP anycast node needs high cost, including financial, human and operational resources.

In the current condition, this cost is mandatory for making BGP anycast node. Especially, to guarantee the quality of service, for example SLA (Service Level Agreement), needs higher cost than normal Internet connectivity. This is one of big burden for operating BGP anycast node. The authors believe that this is one of in the future issue for deploying IP anycasting.

Furthermore, for administrating remote anycast nodes smoothly, many human recources are needed, including local and remote technical staffs and operators. When making BGP anycast node for critical DNS service, the owner of authoritative name servers must consider about this issue.

5. Measurement Issue

To evaluate the practical effect of IP anycast, for example, to verify whether the selections of the IP anycast nodes are appropriate or not, objective measurement is very important. When making BGP anycast mesh in the wide area, the measurement must also be carried out in the wide area.

In such case, there is an effective guideline defined by ICANN, called "CNNP test" [13]. This guideline is useful for making critical DNS server nodes.

But the cost of the measurement is very high, and it is hard to make and maintain for many measurement points/probes.

The continuity is one of an important points for measurement. And operators should verify that the continuity of DNS service is ensured by measurement. RIPE NCC's DNSMON service [14] is one of typical notable project.

6. Our Findings through Making Oversea Anycast nodes

In this section, this memo describes our findings for making oversea IP anycast nodes for our DNS server. At this point, these server nodes are still for reserach and development, but when they were made, we did some useful experiences and got some useful findings. Authors hope that it is useful for DNS operators that they will walk on the same way in the future, especially for TLD operators.

6.1. Running Cost is More Dominant

Running cost is more dominant than initial cost. Human resources and traveling expense for troubleshooting and trouble recovery. Especially for oversea site, sometimes higher the wall of language exists. For instance, a simply replacement sometimes may reduce the total cost from fixing by using remote hands.

6.2. Difference of Custom Practices

Difference of custom of each country is sometimes imporatant issue for practical server node operation. Some data center requires damage insurance contract. But it is hard to have contract with foreign customer. As a result, we had to contact and negotiate a lot of insurance companies.

6.3. Others to Remind

During our operations, we encountered some unexpected trouble. In this section, this memo describes the term of them.

6.3.1. Thermal Overheat

Thermal overheat is occurred due to cabinet placement and ventilation. So, we had to order rack rotation work and additional cabinet fans.

6.3.2. Broken Hardware Replacement

Due to our contract of agent, broken hardware replacement must be "parts-based", not a whole of server hardware. So we had to need additional work for fixing server at oversea place. We cannot order the remote hands because the work is so complex and difficult.

7. Other Considerations and Issues

In this section, this memo describes other related considerations and issues for making BGP anycast nodes for critical authoritative name

servers. Authors will describe requirements and preconditions for these issues, but not yet issued.

- Selection of Node Locations
- Selection of Server Hardware
- Selection of Server Software
- Selection of Remote Maintenance Tools
- Effective Remote Maintenance
- Effective Measurement

8. Security Considerations

IP anycast mitigates Denial-of-Service attack effect and constrains Distributed Denial-of-Service attack in their local network. It is one of most important goals of IP anycast.

To keeping higher security level of each DNS server nodes is one of most important points of managing critical authrotative name servers. Of couese, it is the same as non-anycasted DNS service, but in IP anycasting environment, all of IP anycast nodes have the same IP address for authoritative DNS service, so, it means it is much important than single server because all of nodes should be applied the same higher security policy.

In IP anycasting environment, number of nodes increases compared with non-anycasting environment. It means the place where the safety of data must be guaranteed also increases. And practical secure data synchronization method between nodes must be required, typically data encryption.

9. IANA Considerations

This document requests no action from IANA.

10. Acknowledgements

Paul Vixie and Bill Manning reviewed a previous version of this document. Joe Abley reviewed a previous version of this document and provided detailed and useful comments.

Yoshiki Ishida, Tomoya Yoshida, George Michaelson and Peter Koch provided some useful comments and suggestions.

The authors acknowledge many helpful suggestions from the members of JPRS Research and Development Department and System administration

Department.

This memo is included in the results of the research activities funded by National Institute of Information and Communications Technology (NICT).

11. References

[1] Partridge, C., Mendez, T., and W. Milliken, "Host Anycasting Service", RFC 1546, November 1993.

[2] Hardie, T., "Distributing Authoritative Name Servers via Shared Unicast Addresses", RFC 3258, April 2002.

[3] Abley, J. and K. Lindqvist, "Operation of Anycast Services", draft-ietf-grow-anycast-03.txt (work in progress), January 2006.

[4] Elz, R., Bradner, S., and M. Patton, "Selection and Operation of Secondary DNS Servers", RFC 2182, July 1997.

[5] Bush, R., Karrenberg, D., Kosters, M., and R. Plzak, "Root Name Server Operational Requirements", RFC 2870, June 2000.

[6] Mockapetris, P., "DOMAIN NAMES - CONCEPTS AND FACILITIES", RFC 1034, November 1987.

[7] Mockapetris, P., "DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION", RFC 1035, November 1987.

[8] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[9] Chen, E. and J. Stewart, "A Framework for Inter-Domain Route Aggregation", RFC 2519, February 1999.

[10] Hinden, R. and S. Deering, "Internet Protocol Version 6 (IPv6) Addressing Architecture", RFC 3513, April 2003.

[11] Huston, G., "NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control", RFC 3765, April 2004.

[12] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[13] "ICANN Unsponsored TLD Agreement: Appendix D (.info)", May 2001.

[14] "RIPE-NCC/DNS Server Monitoring", <http://dnsmon.ripe.net/>.

Authors' Addresses

Yasuhiro Orange Morishita
Research and Development Department, Japan Registry Services Co.,Ltd.
Chiyoda First Bldg. East 13F, 3-8-1 Nishi-Kanda
Chiyoda-ku, Tokyo 101-0065
Japan

Email: yasuhiro@jprs.co.jp

Shinta Sato
System Administration Department, Japan Registry Services Co.,Ltd.
Chiyoda First Bldg. East 13F, 3-8-1 Nishi-Kanda
Chiyoda-ku, Tokyo 101-0065
Japan

Email: shinta@jprs.co.jp

Takayasu Matsuura
System Administration Department, Japan Registry Services Co.,Ltd.
Chiyoda First Bldg. East 13F, 3-8-1 Nishi-Kanda
Chiyoda-ku, Tokyo 101-0065
Japan

Email: matuura@jprs.co.jp

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.