Network Working Group
Request for Comments: 3258
Category: Informational

T. Hardie
Nominum, Inc.

April 2002

Distributing Authoritative Name Servers via Shared Unicast Addresses

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Copyright Notice

Copyright (C) The Internet Society (2002). All Rights Reserved.

Abstract

This memo describes a set of practices intended to enable an authoritative name server operator to provide access to a single named server in multiple locations. The primary motivation for the development and deployment of these practices is to increase the distribution of Domain Name System (DNS) servers to previously under-served areas of the network topology and to reduce the latency for DNS query responses in those areas.

1. Introduction

This memo describes a set of practices intended to enable an authoritative name server operator to provide access to a single named server in multiple locations. The primary motivation for the development and deployment of these practices is to increase the distribution of DNS servers to previously under-served areas of the network topology and to reduce the latency for DNS query responses in those areas. This document presumes a one-to-one mapping between named authoritative servers and administrative entities (operators). This document contains no guidelines or recommendations for caching name servers. The shared unicast system described here is specific to IPv4; applicability to IPv6 is an area for further study. It should also be noted that the system described here is related to that described in [ANYCAST], but it does not require dedicated address space, routing changes, or the other elements of a full anycast infrastructure which that document describes.

2. Architecture

2.1 Server Requirements

Operators of authoritative name servers may wish to refer to [SECONDARY] and [ROOT] for general guidance on appropriate practice for authoritative name servers. In addition to proper configuration as a standard authoritative name server, each of the hosts participating in a shared-unicast system should be configured with two network interfaces. These interfaces may be either two physical interfaces or one physical interface mapped to two logical interfaces. One of the network interfaces should use the IPv4 shared unicast address associated with the authoritative name server. The other interface, referred to as the administrative interface below, should use a distinct IPv4 address specific to that host. The host should respond to DNS queries only on the shared-unicast interface. In order to provide the most consistent set of responses from the mesh of anycast hosts, it is good practice to limit responses on that interface to zones for which the host is authoritative.

2.2 Zone file delivery

In order to minimize the risk of man-in-the-middle attacks, zone files should be delivered to the administrative interface of the servers participating in the mesh. Secure file transfer methods and strong authentication should be used for all transfers. If the hosts in the mesh make their zones available for zone transfer, the administrative interfaces should be used for those transfers as well, in order to avoid the problems with potential routing changes for TCP traffic noted in section 2.5 below.

2.3 Synchronization

Authoritative name servers may be loosely or tightly synchronized, depending on the practices set by the operating organization. As noted below in section 4.1.2, lack of synchronization among servers using the same shared unicast address could create problems for some users of this service. In order to minimize that risk, switch-overs from one data set to another data set should be coordinated as much as possible. The use of synchronized clocks on the participating hosts and set times for switch-overs provides a basic level of coordination. A more complete coordination process would involve:

a) receipt of zones at a distribution host
b) confirmation of the integrity of zones received
c) distribution of the zones to all of the servers in the mesh
d) confirmation of the integrity of the zones at each server

e) coordination of the switchover times for the servers in the mesh
f) institution of a failure process to ensure that servers that did not receive correct data or could not switchover to the new data ceased to respond to incoming queries until the problem could be resolved.

Depending on the size of the mesh, the distribution host may also be a participant; for authoritative servers, it may also be the host on which zones are generated.

This document presumes that the usual DNS failover methods are the only ones used to ensure reachability of the data for clients. It does not advise that the routes be withdrawn in the case of failure; it advises instead that the DNS process shutdown so that servers on other addresses are queried. This recommendation reflects a choice between performance and operational complexity. While it would be possible to have some process withdraw the route for a specific server instance when it is not available, there is considerable operational complexity involved in ensuring that this occurs reliably. Given the existing DNS failover methods, the marginal improvement in performance will not be sufficient to justify the additional complexity for most uses.

2.4 Server Placement

Though the geographic diversity of server placement helps reduce the effects of service disruptions due to local problems, it is diversity of placement in the network topology which is the driving force behind these distribution practices. Server placement should emphasize that diversity. Ideally, servers should be placed topologically near the points at which the operator exchanges routes and traffic with other networks.

2.5 Routing

The organization administering the mesh of servers sharing a unicast address must have an autonomous system number and speak BGP to its peers. To those peers, the organization announces a route to the network containing the shared-unicast address of the name server. The organization’s border routers must then deliver the traffic destined for the name server to the nearest instantiation. Routing to the administrative interfaces for the servers can use the normal routing methods for the administering organization.

One potential problem with using shared unicast addresses is that routers forwarding traffic to them may have more than one available route, and those routes may, in fact, reach different instances of

the shared unicast address. Applications like the DNS, whose communication typically consists of independent request-response messages each fitting in a single UDP packet present no problem. Other applications, in which multiple packets must reach the same endpoint (e.g., TCP) may fail or present unworkable performance characteristics in some circumstances. Split-destination failures may occur when a router does per-packet (or round-robin) load sharing, a topology change occurs that changes the relative metrics of two paths to the same anycast destination, etc.

Four things mitigate the severity of this problem. The first is that UDP is a fairly high proportion of the query traffic to name servers. The second is that the aim of this proposal is to diversify topological placement; for most users, this means that the coordination of placement will ensure that new instances of a name server will be at a significantly different cost metric from existing instances. Some set of users may end up in the middle, but that should be relatively rare. The third is that per packet load sharing is only one of the possible load sharing mechanisms, and other mechanisms are increasing in popularity.

Lastly, in the case where the traffic is TCP, per packet load sharing is used, and equal cost routes to different instances of a name server are available, any DNS implementation which measures the performance of servers to select a preferred server will quickly prefer a server for which this problem does not occur. For the DNS failover mechanisms to reliably avoid this problem, however, those using shared unicast distribution mechanisms must take care that all of the servers for a specific zone are not participants in the same shared-unicast mesh. To guard even against the case where multiple meshes have a set of users affected by per packet load sharing along equal cost routes, organizations implementing these practices should always provide at least one authoritative server which is not a participant in any shared unicast mesh. Those deploying shared-unicast meshes should note that any specific host may become unreachable to a client should a server fail, a path fail, or the route to that host be withdrawn. These error conditions are, however, not specific to shared-unicast distributions, but would occur for standard unicast hosts.

Since ICMP response packets might go to a different member of the mesh than that sending a packet, packets sent with a shared unicast source address should also avoid using path MTU discovery.

Appendix A. contains an ASCII diagram of an example of a simple implementation of this system. In it, the odd numbered routers deliver traffic to the shared-unicast interface network and filter traffic from the administrative network; the even numbered routers

deliver traffic to the administrative network and filter traffic from the shared-unicast network. These are depicted as separate routers for the ease this gives in explanation, but they could easily be separate interfaces on the same router. Similarly, a local NTP source is depicted for synchronization, but the level of synchronization needed would not require that source to be either local or a stratum one NTP server.

3. Administration

3.1 Points of Contact

A single point of contact for reporting problems is crucial to the correct administration of this system. If an external user of the system needs to report a problem related to the service, there must be no ambiguity about whom to contact. If internal monitoring does not indicate a problem, the contact may, of course, need to work with the external user to identify which server generated the error.

4. Security Considerations

As a core piece of Internet infrastructure, authoritative name servers are common targets of attack. The practices outlined here increase the risk of certain kinds of attacks and reduce the risk of others.

4.1 Increased Risks

4.1.1 Increase in physical servers

The architecture outlined in this document increases the number of physical servers, which could increase the possibility that a server mis-configuration will occur which allows for a security breach. In general, the entity administering a mesh should ensure that patches and security mechanisms applied to a single member of the mesh are appropriate for and applied to all of the members of a mesh. "Genetic diversity" (code from different code bases) can be a useful security measure in avoiding attacks based on vulnerabilities in a specific code base; in order to ensure consistency of responses from a single named server, however, that diversity should be applied to different shared-unicast meshes or between a mesh and a related unicast authoritative server.

4.1.2 Data synchronization problems

The level of systemic synchronization described above should be augmented by synchronization of the data present at each of the servers. While the DNS itself is a loosely coupled system, debugging

problems with data in specific zones would be far more difficult if two different servers sharing a single unicast address might return different responses to the same query. For example, if the data associated with www.example.com has changed and the administrators of the domain are testing for the changes at the example.com authoritative name servers, they should not need to check each instance of a named authoritative server. The use of NTP to provide a synchronized time for switch-over eliminates some aspects of this problem, but mechanisms to handle failure during the switchover are required. In particular, a server which cannot make the switchover must not roll-back to a previous version; it must cease to respond to queries so that other servers are queried.

4.1.3 Distribution risks

If the mechanism used to distribute zone files among the servers is not well secured, a man-in-the-middle attack could result in the injection of false information. Digital signatures will alleviate this risk, but encrypted transport and tight access lists are a necessary adjunct to them. Since zone files will be distributed to the administrative interfaces of meshed servers, the access control list for distribution of the zone files should include the administrative interface of the server or servers, rather than their shared unicast addresses.

4.2 Decreased Risks

The increase in number of physical servers reduces the likelihood that a denial-of-service attack will take out a significant portion of the DNS infrastructure. The increase in servers also reduces the effect of machine crashes, fiber cuts, and localized disasters by reducing the number of users dependent on a specific machine.

5. Acknowledgments

Masataka Ohta, Bill Manning, Randy Bush, Chris Yarnell, Ray Plzak, Mark Andrews, Robert Elz, Geoff Huston, Bill Norton, Akira Kato, Suzanne Woolf, Bernard Aboba, Casey Ajalat, and Gunnar Lindberg all provided input and commentary on this work. The editor wishes to remember in particular the contribution of the late Scott Tucker, whose extensive systems experience and plain common sense both contributed greatly to the editor’s own deployment experience and are missed by all who knew him.

6. References

[SECONDARY] Elz, R., Bush, R., Bradner, S. and M. Patton, "Selection and Operation of Secondary DNS Servers", BCP 16, RFC 2182, July 1997.

[ROOT] Bush, R., Karrenberg, D., Kosters, M. and R. Plzak, "Root Name Server Operational Requirements", BCP 40, RFC 2870, June 2000.

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

Appendix A.

```
        _________________
Peer 1-|                 |
Peer 2-|                 |
Peer 3-|     Switch      |
Transit|                 |  _________                 _________
etc    |                 |--|Router1|---|----|----------|Router2|---WAN-|
       |                 |  ---------   |    |          ---------       |
       |                 |              |    |                          |
       |                 |              |    |                          |
       ------------------            [NTP] [DNS]                        |
                                                                        |
                                                                        |
                                                                        |
                                                                        |
        _________________                                               |
Peer 1-|                 |                                               |
Peer 2-|                 |                                               |
Peer 3-|     Switch      |                                               |
Transit|                 |  _________                 _________          |
etc    |                 |--|Router3|---|----|----------|Router4|---WAN-|
       |                 |  ---------   |    |          ---------       |
       |                 |              |    |                          |
       |                 |              |    |                          |
       ------------------            [NTP] [DNS]                        |
                                                                        |
                                                                        |
                                                                        |
                                                                        |
        _________________                                               |
Peer 1-|                 |                                               |
Peer 2-|                 |                                               |
Peer 3-|     Switch      |                                               |
Transit|                 |  _________                 _________          |
etc    |                 |--|Router5|---|----|----------|Router6|---WAN-|
       |                 |  ---------   |    |          ---------       |
       |                 |              |    |                          |
       |                 |              |    |                          |
       ------------------            [NTP] [DNS]                        |
                                                                        |
                                                                        |
                                                                        |
```

```
                                                                  |
       __________________                                         |
Peer 1-|                |                                         |
Peer 2-|                |                                         |
Peer 3-|     Switch     |                                         |
Transit|                |  ________                 _________     |
etc    |                |--|Router7|---|----|----------|Router8|---WAN-|
       |                |  ---------   |    |          ---------
       |                |              |    |
       |                |              |    |
       ------------------            [NTP] [DNS]
```

7. Editor's Address

Ted Hardie
Nominum, Inc.
2385 Bay Road.
Redwood City, CA 94063

Phone: 1.650.381.6226
EMail: Ted.Hardie@nominum.com

8. Full Copyright Statement

Copyright (C) The Internet Society (2002). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Acknowledgement

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