GROW
Internet-Draft
Expires: April 22, 2005

K. Lindqvist
Netnod Internet Exchange
J. Abley
ISC

October 22, 2004

Operation of Anycast Services
draft-kurtis-anycast-bcp-00.txt

Status of this Memo

Copyright Notice

Abstract

As the Internet has grown, many services with high availability requirements have emerged. The requirements of these services have increased the demands on the reliability of the infrastructure on which those services rely.

Many techniques have been employed to increase the availability of

services deployed on the Internet. This document presents operational experience of wide-scale service distribution using anycast, and proposes a series of recommendations for others using this approach.

Table of Contents

1. Introduction

To distribute a service using anycast, the service is first associated with a stable set of IP addresses, and reachability to those addresses is advertised in a routing system from multiple, independent service nodes. Various techniques for anycast deployment of services are discussed in RFC 1546 [4], ISC-TN-2003-1 [12] and ISC-TN-2004-1 [13].

Anycast has in recent years become increasingly popular for adding redundancy to DNS servers. Several root server operators have distributed their servers widely around the Internet, and both resolver and authority servers are commonly distributed within the networks of service providers. Anycast distribution has been used by commercial DNS authority server operators for several years. The use of anycast is not limited to the DNS, although the use of anycast imposes some additional requirements on the nature of the service being distributed, including transaction longevity, transaction state held on servers and data synchronization capabilities.

Although anycast is conceptually simple, its implementation introduces some pitfalls for operation of the service. For example, monitoring the availability of the service becomes more difficult; the observed availability changes according to the source of the query, and the client catchment of individual anycast nodes is not static, nor especially deterministic.

This document will describe the use of anycast for both local scope distribution of services using an Interior Gateway Protocol (IGP) and global distribution using BGP [5]. Many of the issues for monitoring and data synchronization are common to both, but deployment issues differ substantially.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

Service Address: an IP address associated with a particular service (e.g. the address of a nameserver).

Anycast: the practice of making a particular Service Address
available in multiple, discrete, autonomous locations, such that
datagrams sent are routed to one of several available locations.
Anycast Node: an internally-connected collection of hosts and routers
which together provide service for an anycast service address.

Local-Scope Anycast: reachability information for the anycast service address is propagated through a routing system in such a way that a particular anycast node is only visible to a subset of the whole routing system.

Local Node: an Anycast Node providing service using a Local-Scope Anycast address.

Global-Scope Anycast: reachability information for the anycast service address is propagated through a routing system in such a way that a particular anycast node is potentially visible to the whole routing system.

Global Node: an Anycast Node providing service using a Global-Scope Anycast address.

3. Anycast Service Distribution

3.1 General Description

Anycast is the name given to the practice of making one or more Service Addresses available to a routing system at Anycast Nodes in two or more discrete locations. The service provided by each node is necessarily consistent regardless of the particular node chosen by the routing system to handle a particular request.

For services distributed using anycast, there is no inherent requirement for referrals to other servers or name-based service distribution ("round-robin DNS"), although those techniques could be combined with anycast service distribution if an application required it. The routing system makes the decision of the node to be used for each request, based on the topological design of the routing system and the point in the network at which the request originates.

The Anycast Node chosen to service a particular query can be influenced by the traffic engineering capabilities of the routing protocols which make up the routing system. The degree of influence available to the operator of the node depends on the scale of the routing system within which the Service Address is anycast.

Load-balancing between Anycast Nodes is typically difficult to achieve (load distribution between nodes is generally unbalanced in terms of request and traffic load). Distribution of load between nodes for the purposes of reliability, and coarse-grained distribution of load for the purposes of making popular services

scalable can often be accommodated, however.

The scale of the routing system through which a service is anycast can vary from a small Interior Gateway Protocol (IGP) connecting a small handful of components, to the Border Gateway Protocol (BGP) [5] connecting the global Internet, depending on the nature of the

service distribution that is required.

3.2 Goals

A service may be anycast for a variety of reasons. A number of common objectives are:

1. Coarse ("unbalanced") distribution of load across nodes, to allow infrastructure to scale to increased numbers of queries and to accommodate transient query peaks;
2. Mitigation of non-distributed denial of service attacks by localizing damage to single anycast nodes;
3. Constraint of distributed denial of service attacks or flash crowds to local regions around anycast nodes (perhaps restricting query traffic to local peering links, rather than paid transit circuits);
4. Triangulation of traffic sources, in the case of attack (or query) traffic which incorporates spoofed source addresses;
5. Improvement of query response time, by reducing the network RTT between client and server with the provision of a local Anycast Node.
6. Reduction of a list of servers to a single, distributed address. For example, a large number of authoritative nameservers for a zone may be deployed using a small set of anycast service addresses; this approach can increase the accessibility of zone data in the DNS without increasing the size of a referral response from a parent nameserver.

4. Design

4.1 Protocol Suitability

When a service is anycast between two or more nodes, the routing system effectively makes the node selection decision on behalf of a client. Since it is usually a requirement that a single client-server interaction is carried out between a client the same server node for the duration of the transaction, it follows that the routing system's node selection decision ought to be stable for an order of magnitude longer than the expected transaction time, if the service is to be provided reliably.

Some services have very short transaction times, and may even be carried out using a single packet request and a single packet reply in some cases (the DNS is an example of this). Other services involve far longer-lived transactions (e.g. bulk file downloads and audio-visual media streaming).

Some anycast deployments have very predictable routing systems, which

can remain stable for long periods of time (e.g. anycast within an IGP, where node selection changes only occur as a response to node failures). Other deployments have far less predictable characteristics (e.g. a densely-deployed array of nodes across the global Internet).

The stability of the routing system together with the transaction time of the service should be carefully compared when deciding whether a service is suitable for distribution using anycast.

4.2 Node Placement

Decisions as to where Anycast Nodes should be placed will depend to a large extent on the goals of the service distribution. For example:

- A recursive resolver service might be distributed within an ISP's network, one Anycast Node per PoP.
- A root server service might be distributed throughout the Internet with nodes located in regions with poor external connectivity, to ensure that the DNS functions adequately within the region during times of external network failure.
- An FTP mirror service might include local nodes located at exchange points, so that ISPs connected to that exchange point could download bulk data more cheaply than if they had to use expensive transit circuits.

In general node placement decisions should be made with consideration of likely traffic requirements, the potential for flash crowds or denial-of-service traffic, the stability of the local routing system and the failure modes with respect to node failure, or local routing system failure.

4.3 Routing Systems

4.3.1 Anycast within an IGP

There are several common motivations for the distribution of a Service Address within the scope of an IGP:

1. to improve service response times, by hosting a service close to

other users of the network;
2. to improve service reliability by providing automatic fail-over to backup nodes; and
3. to keep service traffic local, to avoid congesting wide-area links.

In each case the decisions as to where and how services are provisioned can be made by network engineers without requiring such

operational complexities as regional variances in the configuration of client computers, or DNS tricks which respond differently to requests from clients in different locations.

When a service is anycast within an IGP the routing system is typically under the control of the same organization who is providing the service, and hence the relationship between service transaction characteristics and network stability are likely to be relatively well-understood. This technique is consequently applicable to a larger number of applications than Internet-wide anycast service distribution (see Section 4.1).

By reducing the scope of the IGP to just the hosts providing service (together with one or more gateway routers) this technique can be applied to the construction of server clusters. This application is discussed in some detail in [13].

4.3.2 Anycast within the Global Internet

Service Addresses may be anycast within the global Internet routing system in order to distribute services across the entire network. The principal differences between this application and the IGP-scope distribution discussed in Section 4.3.1 are that:

1. the routing system is, in general, controlled by other people; and
2. the routing protocol concerned (BGP), and commonly-accepted practices in its deployment, impose some additional constraints (see Section 4.4).

4.4 Routing Considerations

4.4.1 Signalling Service Availability

When a routing system is provided with reachability information for a Service Address from an individual node, packets addressed to that Service Address will start to arrive at the node. Since it is desirable for the node to be ready to accept requests before they start to arrive, a coupling between the routing information and the availability of the service at a particular node is desirable.

Where a routing advertisement from a node corresponds to a single Service Address, this coupling might be such that availability of the service triggers the route advertisement, and non-availability of the service triggers a route withdrawal. This can be achieved using routing protocol implementations on the same servers which provide the service being distributed.

Where a routing advertisement from a node corresponds to two or more Service Addresses, it may not be appropriate to trigger a route withdrawal due to the non-availability of a single service. Another approach is to tunnel requests from nodes that cannot handle individual services to other nodes that can, perhaps using an IGP which extends over tunnels between nodes, in which servers participate. Circumstances which might lead to multiple Service Addresses being covered by a single route are discussed in Section 4.4.2.

4.4.2 Covering Prefix

In some routing systems (e.g. the BGP-based routing system of the global Internet) it is not possible, in general, to propagate a host route with confidence that availability of the route will be signaled throughout the network. This is a consequence of operational policy, and not a protocol restriction.

In such cases it is necessary to propagate a route which covers the Service Address, and which has a sufficiently short prefix that it will not be caught by commonly-deployed import policies. In many cases this will be a 24-bit prefix, but there are other well-documented examples of import polices which filter on RIR allocation boundaries, and hence some experimentation may be prudent.

Where multiple Service Addresses are covered by the same covering route, there is no longer a tight coupling between the advertisement of that route and the individual services associated with the covered host routes. The resulting impact on signaling availability of individual services is discussed in Section 4.4.1.

4.4.3 Equal-Cost Paths

Some routing systems support equal-cost paths to the same destination. Where multiple, equal-cost paths exist and lead to different anycast nodes, there is a risk that request packets associated with a single transaction might be delivered to more than one node. Services provided over TCP necessarily involve transactions with multiple request packets, due to the TCP setup handshake.

Equal cost paths are commonly supported in IGPs. Multi-node selection for a single transaction can be avoided in most cases by careful consideration of IGP link metrics, or by applying equal-cost multi-path (ECMP) selection algorithms which cause a single node to be selected for a single multi-packet transaction. For a description of hash-based ECMP selection, see [13].

For services which are distributed across the global Internet using BGP, equal-cost paths are normally not a consideration: BGP's exit selection algorithm usually selects a single, consistent exit for a single destination regardless of whether multiple candidate paths exist. Implementations of BGP exist that support multi-path exit selection, however, and corner cases where dual selected exits route to different nodes are possible. Analysis of the likely incidence of such corner cases for particular distributions of Anycast Nodes are recommended for services which involve multi-packet transactions.

4.4.4 Route Dampening

Frequent advertisements and withdrawals of individual prefixes in BGP are known as flaps. Rapid flapping can lead to CPU exhaustion on routers quite remote from the source of the instability, and for this reason rapid route oscillations are frequently "damped", as described in [9].

A dampened path will be suppressed by routers for an interval which increases according to the frequency of the observed oscillation; a suppressed path will not propagate. Hence a single router can prevent the propagation of a flapping prefix to the rest of an autonomous system, affording other routers in the network protection from the instability.

Common implementations of flap dampening penalizes oscillating advertisements based on the observed AS_PATH, and not on the NLRI. For this reason, network instability which leads to route flapping from a single anycast node ought not to cause advertisements from other nodes (which have different AS_PATH attributes) to be dampened.

As dampening works on advertisements with the same AS_PATH attribute, care should be taken so that the AS_PATH is as diverse as possible for the anycasted nodes. The Anycasted nodes should have the same origin AS for their advertisements, but they should have different upstream AS:es for each node. If the upstream AS is the same at all locations, there is a risk that the upstream AS will peer with the AS:es at multiple locations and could therefor propagate the same AS_PATH, but for different Anycast nodes. This could render the service for multiple Anycast nodes unavailable due to dampening caused by only one of them.

It is possible that other implementations of flap dampening may become prevalent in the future, causing individual nodes' instability to result in stable nodes becoming unavailable. Judicious deployment of Local Nodes in combination with especially stable Global Nodes may help mitigate such problems, should they ever arise.

4.4.5 Reverse Path Forwarding Checks

Reverse Path Forwarding (RPF) checks, first described in [8], are commonly deployed as part of ingress interface packet filters on routers in the global Internet in order to deny packets whose source addresses are spoofed (see also [10]). Deployed implementations of RPF make available two modes of operation: a loose mode, and a strict mode.

Strict-mode RPF checks can cause non-spoofed packets to be denied when they originate from multi-homed site, since selected paths might legitimately not correspond with the ingress interface of non-spoofed packets from the multi-homed site. A collection of anycast nodes deployed across the Internet is largely indistinguishable from a distributed, multi-homed site to the routing system, and hence this risk also exists for anycast nodes, even if individual nodes are not multi-homed.

Care should be taken to ensure that strict-mode RPF is not enabled in peer networks connecting to anycast nodes.

4.4.6 Propagation Scope

In the context of Anycast service distribution across the global Internet, Global Nodes are those which are capable of providing service to clients anywhere in the network; reachability information for the service is propagated globally, without restriction, by advertising the routes covering the Service Addresses for global transit to one or more providers.

More than one Global Node can exist for a single service (and indeed this is often the case, for reasons of redundancy and load-sharing).

In contrast, it is sometimes desirable to deploy an Anycast Node which only provides services to a local catchment of autonomous systems, and which is deliberately not available to the entire Internet; such nodes are referred to in this document as Local Nodes. An example of circumstances in which a Local Node may be appropriate are nodes designed to serve a region with rich internal connectivity but unreliable, congested or expensive access to the rest of the Internet.

Local Nodes advertise covering routes for Service Addresses in such a way that their propagation is restricted. This might be done using well-known community string attributes such as NO_EXPORT [6] or NOPEER [11], or by arranging with peers to apply a conventional "peering" import policy instead of a "transit" import policy, or some suitable combination of measures.

4.4.7 Other Peoples' Networks

When Anycast services are deployed across networks operated by others, their reachability is dependent on routing polices and topology changes (planned and unplanned) which are unpredictable and sometimes difficult to identify. Consequently, routing policies used by Anycast Nodes should be conservative, individual nodes' internal and external/connecting infrastructure should be scaled to support loads far in excess of the average, and the service should be monitored proactively (Section 5.1) from many points in order to avoid unpleasant surprises.

4.5 Data Synchronisation

As a client contacting a anycasted service will expect all possible servers to serve the same data, the Anycast service needs to assure data consistency across all Anycast Nodes. This includes periodic updating of all data, and verification of a successful transfer of data.

How data is synchronized depends on the service being Anycasted. The methods used could for example be a zone transfer for an authoritative set of DNS-servers, rsync for a FTP archive or no synchronization needed for a DNS resolver service. In the DNS examples, synchronization comes with the service and the associated protocol. For other services, this will be an external mechanism to the protocol. In both cases, the synchronization needs to be run from a local IP address that is not the service address. The data transfer should be authenticated in order to prevent spoofing of the data on the Anycasted nodes and the data should be verified.

Verification can be done with for example TSIG for DNS, or for example a MD5 hash[2] for verification of other data. The method might vary but should verify that all data was transfered, and that the data is correct and not manipulated.

Authentication of the data source can be based either on the protocol in use, as is the case with TSIG for DNS, or some other external mechanism. For example a IP tunnel protected by authentication and encryption as described in [7].

4.6 Node Autonomy

For an Anycast deployment whose goals include improved reliability through redundancy, it is important to minimize the opportunity for a single defect to compromise many (or all) nodes, or for the failure of one node to provide a cascading failure bringing down additional successive nodes until the service as a whole is defeated.

Codependencies are avoided by making each node as autonomous and self-sufficient as possible. The degree to which nodes can survive failure elsewhere depends on the nature of the service being delivered, but for services which accommodate disconnected operation (e.g. the timed propagation of changes between master and slave servers in the DNS) a high degree of autonomy can be achieved.

The possibility of cascading failure due to load can also be reduced by the deployment of both Global and Local Nodes for a single service, since the effective fail-over path of traffic is, in general, from Local Node to Global Node; traffic that might sink one Local Node is unlikely to sink all Local Nodes, except in the most degenerate cases.

The chance of cascading failure due to a software defect in an operating system or server can be reduced in many cases by deploying nodes running different software implementations.

5. Service Management

5.1 Monitoring

Monitoring a service which is distributed is more complex than monitoring a non-distributed service, since the observed accuracy and availability of the service is, in general, different when viewed from clients attached to different parts of the network. When a problem is identified, it is also not always obvious which node served the request, and hence which node is malfunctioning.

It is recommended that distributed services are monitored from probes distributed representatively across the routing system, and, where possible, the identity of the node answering individual requests is recorded along with performance and availability statistics.

Monitoring the routing system (from a variety of places, in the case of routing systems where perspective counts) can also provide useful diagnostics for troubleshooting service availability. This can be achieved using dedicated probes, or public route measurement facilities on the Internet such as RIPE's Routing Information Service [14] and the University of Oregon Route

Views Project [15].

5.2 Self-Healing Nodes

As is described in having the Anycast Node avoid black-holing traffic in the event of a failure on the software or subsystem providing the service should be avoided. As described, this can be done with withdrawing the announcement of the prefix corresponding to

the service address, or the covering route. However, the nodes could also try and handle the failure in a number of ways. This can be with as also previously described tunneling to other instances of the Anycasted service, and using a IGP over the tunnels, route incoming client queries to the other destination. The Anycasted node could also contain separate systems for trying to restart the service in question, and if successful again re-announce the service prefix.

6. Security Considerations

This document describes mechanisms for deploying services on the Internet which can be used to mitigate vulnerability to attack.

The distribution of a service across several (or many) autonomous nodes imposes an increased monitoring load on the operator of the service, and which also imposes an additional systems administration load on the service operator which might reduce the effectiveness of host and router security. It is recommended that these factors be taken into account when assessing the risks and benefits of distributing services using anycast.

7. Protocol Considerations

This document does not impose any protocol considerations.

8. IANA Considerations

This document requests no action from IANA.

9 References

[1] Oran, D., "OSI IS-IS Intra-domain Routing Protocol", RFC 1142, February 1990.

[2] Rivest, R., "The MD5 Message-Digest Algorithm", RFC 1321, April 1992.

[3] Moy, J., "OSPF Version 2", RFC 1247, July 1991.

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

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

[6] Chandrasekeran, R., Traina, P. and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[7] Kent, S. and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

[8] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", RFC 2267, January 1998.

[9] Villamizar, C., Chandra, R. and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

[10] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

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

[12] Abley, J., "Hierarchical Anycast for Global Service Distribution", March 2003, <http://www.isc.org/pubs/tn/isc-tn-2003-1.html>.

[13] Abley, J., "A Software Approach to Distributing Requests for DNS Service using GNU Zebra, ISC BIND 9 and FreeBSD", March 2004, <http://www.isc.org/pubs/tn/isc-tn-2004-1.html>.

[14] <http://ris.ripe.net>

[15] <http://www.route-views.org>

Authors' Addresses

Kurt Erik Lindqvist
Netnod Internet Exchange
Bellmansgatan 30
118 47 Stockholm
Sweden

EMail: kurtis@kurtis.pp.se
URI: http://www.netnod.se/

Joe Abley
Internet Systems Consortium, Inc.
950 Charter Street
Redwood City, CA 94063
USA

Phone: +1 650 423 1317
EMail: jabley@isc.org
URI: http://www.isc.org/

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment

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