Network Working Group
Internet-Draft
Expires: July 5, 2006

J. Abley
Afilias Canada
K. Lindqvist
Netnod Internet Exchange
January 2006

Operation of Anycast Services
draft-ietf-grow-anycast-04

Status of this Memo

Copyright Notice

Abstract

As the Internet has grown, and as systems and networked services within enterprises have become more pervasive, many services with high availability requirements have emerged. These requirements have increased the demands on the reliability of the infrastructure on which those services rely.

Various techniques have been employed to increase the availability of

services deployed on the Internet. This document presents commentary and recommendations for distribution of services using anycast.

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 [RFC1546], [ISC-TN-2003-1] and [ISC-TN-2004-1].

The techniques and considerations described in this document apply to services reachable over both IPv4 and IPv6.

Anycast has in recent years become increasingly popular for adding redundancy to DNS servers to complement the redundancy which the DNS architecture itself already provides. Several root DNS 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 limitations on the nature of the service being distributed, including transaction longevity, transaction state held on servers and data synchronisation capabilities.

Although anycast is conceptually simple, its implementation introduces some pitfalls for operation of services. For example, monitoring the availability of the service becomes more difficult; the observed availability changes according to the location of the

client within the network, and the population of clients using individual anycast nodes is neither static, nor reliably 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 the Border Gateway Protocol (BGP) [RFC4271]. Many of the issues for monitoring and data synchronisation are common to both, but deployment issues differ substantially.

2. Terminology

Service Address: an IP address associated with a particular service (e.g. the destination address used by DNS resolvers to reach a particular authority server).

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. An Anycast Node might be as simple as a single host participating in a routing system with adjacent routers, or it might include a number of hosts connected in some more elaborate fashion; in either case, to the routing system across which the service is being anycast, each Anycast Node presents a unique path to the Service Address. The entire anycast system for the service consists of two or more separate Anycast Nodes.

Catchment: in physical geography, an area drained by a river, also known as a drainage basin. By analogy, as used in this document,

the topological region of a network within which packets directed at an anycast address are routed to one particular node.

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 a Service Address available to a routing system at Anycast Nodes in two or more discrete locations. The service provided by each node is generally consistent regardless of the particular node chosen by the routing system to handle a particular request (although some services may benefit from deliberate differences in the behaviours of individual nodes, in order to facilitate locality-specific behaviour; see Section 4.6).

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 decides which node is 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 achieved, 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) [RFC4271] 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 localising damage to single anycast nodes;

3. Constraint of distributed denial of service attacks or flash crowds to local regions around anycast nodes. Anycast distribution of a service provides the opportunity for traffic to be handled closer to its source, perhaps using high-performance peering links rather than oversubscribed, paid transit circuits;

4. To provide additional information to help identify the location of traffic sources in the case of attack (or query) traffic which incorporates spoofed source addresses. This information is derived from the property of anycast service distribution that the selection of the Anycast Node used to service a particular query may be related to the topological source of the request.

5. Improvement of query response time, by reducing the network distance between client and server with the provision of a local Anycast Node. The extent to which query response time is improved depends on the way that nodes are selected for the clients by the routing system. Topological nearness within the routing system does not, in general, correlate to round-trip performance across a network; in some cases response times may see no reduction, and may increase.

6. To reduce 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 nameserver authoritative for the parent zone.

4. Design

4.1. Protocol Suitability

When a service is anycast between two or more nodes, the routing system 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 and 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 substantially 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 (e.g. DNS transactions over UDP transport). Other services involve far longer-lived transactions (e.g. bulk file downloads and audio-visual media streaming).

Services may be anycast within very predictable routing systems, which can remain stable for long periods of time (e.g. anycast within a well-managed and topologically-simple IGP, where node selection changes only occur as a response to node failures). Other deployments have far less predictable characteristics (see Section 4.4.7).

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. In some cases, for new protocols, it may be practical to split large transactions into an initialisation phase which is handled by anycast servers, and a sustained phase which is provided by non-anycast servers, perhaps chosen during the initialisation phase.

This document deliberately avoids prescribing rules as to which protocols or services are suitable for distribution by anycast; to attempt to do so would be presumptuous.

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 DNS recursive resolver service might be distributed within an ISP's network, one Anycast Node per site.

- A root DNS server service might be distributed throughout the Internet; Anycast Nodes could be 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 deliberate DNS incoherence (causing DNS queries to yield different answers depending on where the queries originate).

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

An IGP will generally have no inherent restriction on the length of

prefix that can be introduced to it. In this case there is no need to construct a covering prefix for particular Service Addresses; host

routes corresponding to the Service Address can instead be introduced to the routing system. See Section 4.4.2 for more discussion of the requirement for a covering prefix.

IGPs often feature little or no aggregation of routes, partly due to algorithmic complexities in supporting aggregation. There is little motivation for aggregation in many networks' IGPs in many cases, since the amount of routing information carried in the IGP is small enough that scaling concerns in routers do not arise. For discussion of aggregation risks in other routing systems, see Section 4.4.8.

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 [ISC-TN-2004-1].

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;

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 essential 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 server which provide the service being distributed, which are configured to advertise and withdraw the route advertisement in conjunction with the availability (and health) of the software on the host which processes service

requests. An example of such an arrangement for a DNS service is included in [ISC-TN-2004-1].

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 in the case where the service is down at one Anycast Node is to route requests to a different Anycast Node where the service is working normally. This approach is discussed in Section 4.8.

Rapid advertisement/withdrawal oscillations can cause operational problems, and nodes should be configured such that rapid oscillations are avoided (e.g. by implementing a minimum delay following a withdrawal before the service can be re-advertised). See Section 4.4.4 for a discussion of route oscillations in BGP.

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 the route will propagate 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 discarded by commonly-deployed import policies. For IPv4 Service Addresses, this is often a 24-bit prefix, but there are other well-documented examples of IPv4 import polices which filter on Regional Internet Registry (RIR) allocation boundaries, and hence some experimentation may be prudent. Corresponding import policies for IPv6 prefixes also exist. See Section 4.5 for more discussion of

IPv6 Service Addresses and corresponding anycast routes.

The propagation of a single route per service has some associated scaling issues which are discussed in Section 4.4.8.

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 signalling availability of individual services is discussed in Section 4.4.1 and Section 4.8.

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 different request

packets associated with a single transaction might be delivered to more than one node. Services provided over TCP [RFC0793] necessarily involve transactions with multiple request packets, due to the TCP setup handshake.

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.

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 an example of the use of hash-based ECMP selection in anycast service distribution, see [ISC-TN-2004-1].

Other ECMP selection algorithms are commonly available, including those in which packets from the same flow are not guaranteed to be routed towards the same destination. ECMP algorithms which select a route on a per-packet basis rather than per-flow are commonly referred to as performing "Per Packet Load Balancing" (PPLB).

With respect to anycast service distribution, some uses of PPLB may cause different packets from a single multi-packet transaction sent by a client to be delivered to different anycast nodes, effectively making the anycast service unavailable. Whether this affects specific anycast services will depend on how and where anycast nodes are deployed within the routing system, and on where the PPLB is being performed:

1. PPLB across multiple, parallel links between the same pair of routers should cause no node selection problems;

2. PPLB across diverse paths within a single autonomous system (AS), where the paths converge to a single exit as they leave the AS, should cause no node selection problems;

3. PPLB across links to different neighbour ASes where the neighbour ASes have selected different nodes for a particular anycast destination will, in general, cause request packets to be distributed across multiple anycast nodes. This will have the effect that the anycast service is unavailable to clients downstream of the router performing PPLB.

The uses of PPLB which have the potential to interact badly with anycast service distribution can also cause persistent packet reordering. A network path that persistently reorders segments will degrade the performance of traffic carried by TCP [Allman2000]. TCP, according to several documented measurements, accounts for the bulk of traffic carried on the Internet ([McCreary2000], [Fomenkov2004]). Consequently, in many cases it is reasonable to consider networks making such use of PPLB to be pathological.

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 "dampened", as described in [RFC2439].

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.

Some implementations of flap dampening penalise oscillating advertisements based on the observed AS_PATH, and not on Network Layer Reachability Information (NLRI; see [RFC4271]). For this reason, network instability which leads to route flapping from a single anycast node will not generally cause advertisements from other nodes (which have different AS_PATH attributes) to be dampened by these implementations.

To limit the opportunity of such implementations to penalise advertisements originating from different Anycast Nodes in response to oscillations from just a single node, care should be taken to arrange that the AS_PATH attributes on routes from different nodes are as diverse as possible. For example, Anycast Nodes should use the same origin AS for their advertisements, but might have different upstream ASes.

Where different implementations of flap dampening are prevalent, individual nodes' instability may result in stable nodes becoming unavailable. In mitigation, the following measures may be useful:

1. Judicious deployment of Local Nodes in combination with especially stable Global Nodes (with high inter-AS path splay, redundant hardware, power, etc.) may help limit oscillation problems to the Local Nodes' limited regions of influence;

2. Aggressive flap-dampening of the service prefix close to the origin (e.g. within an Anycast Node, or in adjacent ASes of each Anycast Node) may also help reduce the opportunity of remote ASes to see oscillations at all.

4.4.5. Reverse Path Forwarding Checks

Reverse Path Forwarding (RPF) checks, first described in [RFC2267], are commonly deployed as part of ingress interface packet filters on routers in the Internet in order to deny packets whose source addresses are spoofed (see also RFC 2827 [RFC2827]). Deployed

implementations of RPF make several modes of operation available (e.g. "loose" and "strict").

Some modes of RPF 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. This issue is discussed in [RFC3704].

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 each anycast node is treated as a multi-homed network, and that the corresponding recommendations in [RFC3704] with respect to RPF checks are heeded.

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 [RFC1997] or NOPEER [RFC3765], or by arranging with peers to apply a conventional

"peering" import policy instead of a "transit" import policy, or some suitable combination of measures.

Advertising reachability to Service Addresses from Local Nodes should ideally be made using a routing policy that require presence of explicit attributes for propagation, rather than relying on implicit (default) policy. Inadvertent propagation of a route beyond its intended horizon can result in capacity problems for Local Nodes which might degrade service performance network-wide.

4.4.7. Other Peoples' Networks

When Anycast services are deployed across networks operated by others, their reachability is dependent on routing policies and topology changes (planned and unplanned) which are unpredictable and sometimes difficult to identify. Since the routing system may include networks operated by multiple, unrelated organisations, the possibility of unforeseen interactions resulting from the combinations of unrelated changes also exists.

The stability and predictability of such a routing system should be taken into consideration when assessing the suitability of anycast as a distribution strategy for particular services and protocols (see also Section 4.1).

By way of mitigation, routing policies used by Anycast Nodes across such routing systems 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 from many points in order to avoid unpleasant surprises (see Section 5.1).

4.4.8. Aggregation Risks

The propagation of a single route for each anycast service does not scale well for routing systems in which the load of routing information which must be carried is a concern, and where there are potentially many services to distribute. For example, an autonomous system which provides services to the Internet with N Service Addresses covered by a single exported route, would need to advertise (N+1) routes if each of those services were to be distributed using anycast.

The common practice of applying minimum prefix-length filters in

import policies on the Internet (see Section 4.4.2) means that for a route covering a Service Address to be usefully propagated the prefix length must be substantially less than that required to advertise just the host route. Widespread advertisement of short prefixes for individual services hence also has a negative impact on address conservation.

Both of these issues can be mitigated to some extent by the use of a single covering prefix to accommodate multiple Service Addresses, as described in Section 4.8. This implies a de-coupling of the route advertisement from individual service availability (see Section 4.4.1), however, with attendant risks to the stability of the service as a whole (see Section 4.7).

In general, the scaling problems described here prevent anycast from being a useful, general approach for service distribution on the global Internet. It remains, however, a useful technique for distributing a limited number of Internet-critical services, as well as in smaller networks where the aggregation concerns discussed here do not apply.

4.5. Addressing Considerations

Service Addresses should be unique within the routing system that connects all Anycast Nodes to all possible clients of the service. Service Addresses must also be chosen so that corresponding routes will be allowed to propagate within that routing system.

For an IPv4-numbered service deployed across the Internet, for example, an address might be chosen from a block where the minimum RIR allocation size is 24 bits, and reachability to that address might be provided by originating the covering 24-bit prefix.

For an IPv4-numbered service deployed within a private network, a locally-unused [RFC1918] address might be chosen, and reachability to that address might be signalled using a (32-bit) host route.

For IPv6-numbered services, Anycast Addresses are not scoped differently from unicast addresses. As such the guidelines presented for IPv4 with respect to address suitability follow for IPv6. Note that historical prohibitions on anycast distribution of services over IPv6 have been removed from the IPv6 addressing specification in [RFC4291].

4.6. Data Synchronisation

Although some services have been deployed in localised form (such

that clients from particular regions are presented with regionally-

relevant content) many services have the property that responses to client requests should be consistent, regardless of where the request originates. For a service distributed using anycast, that implies that different Anycast Nodes must operate in a consistent manner and, where that consistent behaviour is based on a data set, that the data concerned be synchronised between nodes.

The mechanism by which data is synchronised depends on the nature of the service; examples are zone transfers for authoritative DNS servers and rsync for FTP archives. In general, the synchronisation of data between Anycast Nodes will involve transactions between non-anycast addresses.

Data synchronisation across public networks should be carried out with appropriate authentication and encryption.

4.7. Node Autonomy

For an Anycast deployment whose goals include improved reliability through redundancy, it is important to minimise 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.

Co-dependencies 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 implementations of operating system, server software, routing protocol software, etc., such that a defect which appears in a single component does not affect the whole system.

It should be noted that these approaches to increase node autonomy are, to varying degrees, contrary to the practical goals of making a deployed service straightforward to operate. A service which is over-complex is more likely to suffer from operator error than a

service which is more straightforward to run. Careful consideration should be given to all of these aspects so that an appropriate balance may be found.

4.8. Multi-Service Nodes

For a service distributed across a routing system where covering prefixes are required to announce reachability to a single Service Address (see Section 4.4.2), special consideration is required in the case where multiple services need to be distributed across a single set of nodes. This results from the requirement to signal availability of individual services to the routing system so that requests for service are not received by nodes which are not able to process them (see Section 4.4.1).

Several approaches are described in the following sections.

4.8.1. Multiple Covering Prefixes

Each Service Address is chosen such that only one Service Address is covered by each advertised prefix. Advertisement and withdrawal of a single covering prefix can be tightly coupled to the availability of the single associated service.

This is the most straightforward approach. However, since it makes very poor utilisation of globally-unique addresses, it is only suitable for use for a small number of critical, infrastructural services such as root DNS servers. General Internet-wide deployment of services using this approach will not scale.

4.8.2. Pessimistic Withdrawal

Multiple Service Addresses are chosen such that they are covered by a

single prefix. Advertisement and withdrawal of the single covering prefix is coupled to the availability of all associated services; if any individual service becomes unavailable, the covering prefix is withdrawn.

The coupling between service availability and advertisement of the covering prefix is complicated by the requirement that all Service Addresses must be available -- the announcement needs to be triggered by the presence of all component routes, and not just a single covered route.

The fact that a single malfunctioning service causes all deployed services in a node to be taken off-line may make this approach unsuitable for many applications.

4.8.3. Intra-Node Interior Connectivity

Multiple Service Addresses are chosen such that they are covered by a single prefix. Advertisement and withdrawal of the single covering prefix is coupled to the availability of any one service. Nodes have interior connectivity, e.g. using tunnels, and host routes for service addresses are distributed using an IGP which extends to include routers at all nodes.

In the event that a service is unavailable at one node, but available at other nodes, a request may be routed over the interior network from the receiving node towards some other node for processing.

In the event that some local services in a node are down and the node is disconnected from other nodes, continued advertisement of the covering prefix might cause requests to become black-holed.

This approach allows reasonable address utilisation of the netblock covered by the announced prefix, at the expense of reduced autonomy of individual nodes; the IGP in which all nodes participate can be viewed as a single point of failure.

4.9. Node Identification by Clients

From time to time, all clients of deployed services experience problems, and those problems require diagnosis. A service

distributed using anycast imposes an additional variable on the diagnostic process over a simple, unicast service -- the particular anycast node which is handling a client's request.

In some cases, common network-level diagnostic tools such as traceroute may be sufficient to identify the node being used by a client. However, the use of such tools may be beyond the abilities of users at the client side of a transaction, and in any case network conditions at the time of the problem may change by the time such tools are exercised.

Troubleshooting problems with anycast services is greatly facilitated if mechanisms to determine the identity of a node are designed in to the protocol. Examples of such mechanisms include the NSID option in DNS [I-D.ietf-dnsext-nsid] and the common inclusion of hostname information in SMTP servers' initial greeting at session initiation [RFC2821].

Provision of such in-band mechanisms for node identification is strongly recommended for services to be distributed using anycast.

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. The RIPE NCC DNSMON service [1] is an example of such monitoring for the DNS.

Monitoring the routing system (from a variety of places, in the case

of routing systems where perspective is relevant) 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 the RIPE NCC Routing Information Service [2] and the University of Oregon Route Views Project [3].

Monitoring the health of the component devices in an Anycast deployment of a service (hosts, routers, etc.) is straightforward, and can be achieved using the same tools and techniques commonly used to manage other network-connected infrastructure, without the additional complexity involved in monitoring Anycast service addresses.

6. Security Considerations

6.1. Denial-of-Service Attack Mitigation

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

1. An Anycast Node can act as a sink for attack traffic originated within its sphere of influence, preventing nodes elsewhere from having to deal with that traffic;

2. The task of dealing with attack traffic whose sources are widely

distributed is itself distributed across all the nodes which contribute to the service. Since the problem of sorting between legitimate and attack traffic is distributed, this may lead to better scaling properties than a service which is not distributed.

6.2. Service Compromise

The distribution of a service across several (or many) autonomous nodes imposes increased monitoring as well as an increased systems administration burden on the operator of the service which might reduce the effectiveness of host and router security.

The potential benefit of being able to take compromised servers off-line without compromising the service can only be realised if there are working procedures to do so quickly and reliably.

6.3. Service Hijacking

It is possible that an unauthorised party might advertise routes corresponding to anycast Service Addresses across a network, and by doing so capture legitimate request traffic or process requests in a manner which compromises the service (or both). A rogue Anycast Node might be difficult to detect by clients or by the operator of the service.

The risk of service hijacking by manipulation of the routing system exists regardless of whether a service is distributed using anycast. However, the fact that legitimate Anycast Nodes are observable in the routing system may make it more difficult to detect rogue nodes.

Many protocols which incorporate authentication or integrity protection provide those features in a robust fashion, e.g. using periodic re-authentication within a single session, or integrity protection at either the channel (e.g. [RFC2845], [RFC2487]) or message (e.g. [RFC4033], [RFC2311]) levels. Protocols which are

less robust may be more vulnerable to session hijacking. Given the greater opportunity for undetected session hijack with anycast services, the use of robust protocols is recommended for anycast services that require authentication or integrity protection.

7. Protocol Considerations

This document does not impose any protocol considerations.

8. IANA Considerations

This document requests no action from IANA.

9. Acknowledgements

The authors gratefully acknowledge the contributions from various participants of the grow working group, and in particular Geoff Huston, Pekka Savola, Danny McPherson, Ben Black and Alan Barrett.

This work was supported by the US National Science Foundation (research grant SCI-0427144) and DNS-OARC.

10. References

10.1. Normative References

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

[RFC2827] 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.

[RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing

Architecture", RFC 4291, February 2006.

10.2. Informative References

[Allman2000]
Allman, M. and E. Blanton, "On Making TCP More Robust to Packet Reordering", January 2000, <http://www.icir.org/mallman/papers/tcp-reorder-ccr.ps>.

[Fomenkov2004]
Fomenkov, M., Keys, K., Moore, D., and k. claffy, "Longitudinal Study of Internet Traffic from 1999-2003", January 2004, <http://www.caida.org/outreach/papers/2003/nlanr/nlanr_overview.pdf>.

[I-D.ietf-dnsext-nsid]
Austein, R., "DNS Name Server Identifier Option (NSID)", draft-ietf-dnsext-nsid-02 (work in progress), June 2006.

[ISC-TN-2003-1]

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

[ISC-TN-2004-1]
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>.

[McCreary2000]
McCreary, S. and k. claffy, "Trends in Wide Area IP Traffic Patterns: A View from Ames Internet Exchange", September 2000, <http://www.caida.org/outreach/papers/2000/AIX0005/AIX0005.pdf>.

[RFC1546] Partridge, C., Mendez, T., and W. Milliken, "Host Anycasting Service", RFC 1546, November 1993.

[RFC2267] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source

Address Spoofing", RFC 2267, January 1998.

[RFC2311] Dusse, S., Hoffman, P., Ramsdell, B., Lundblade, L., and L. Repka, "S/MIME Version 2 Message Specification", RFC 2311, March 1998.

[RFC2487] Hoffman, P., "SMTP Service Extension for Secure SMTP over TLS", RFC 2487, January 1999.

[RFC2821] Klensin, J., "Simple Mail Transfer Protocol", RFC 2821, April 2001.

[RFC2845] Vixie, P., Gudmundsson, O., Eastlake, D., and B. Wellington, "Secret Key Transaction Authentication for DNS (TSIG)", RFC 2845, May 2000.

[RFC3765] Huston, G., "NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control", RFC 3765, April 2004.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.

URIs

[1] <http://dnsmon.ripe.net/>

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

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

Appendix A. Change History

This section should be removed before publication.

Intended category: BCP.

draft-kurtis-anycast-bcp-00: Initial draft. Discussed at IETF 61 in

the grow meeting and adopted as a working group document shortly afterwards.

draft-ietf-grow-anycast-00: Missing and empty sections completed; some structural reorganisation; general wordsmithing. Document discussed at IETF 62.

draft-ietf-grow-anycast-01: This appendix added; acknowledgements section added; commentary on RFC3513 prohibition of anycast on hosts removed; minor sentence re-casting and related jiggery-pokery. This revision published for discussion at IETF 63.

draft-ietf-grow-anycast-02: Normative reference to draft-ietf-ipv6-addr-arch-v4" added (in the RFC editor's queue at the time of writing; reference should be updated to an RFC number when available). Added commentary on per-packet load balancing.

draft-ietf-grow-anycast-03: Editorial changes and language clean-up at the request of the IESG.

draftt-ietf-grow-anycast-04: Replaced reference to RFC1771 with a reference to RFC4271. Replaced reference to draft-ietf-ipv6-addr-arch-v4 with a reference to RFC 4291. Changed author address for Abley. Wordsmithing in response to Gen-ART review by Sharon Chrisholm and Secdir review by Rob Austein. Added Section 4.9 at the suggestion of Rob Austein.

Authors' Addresses

Joe Abley
Afilias Canada, Corp.
204 - 4141 Yonge Street
Toronto, ON M2P 2A8
Canada

Phone: +1 416 673 4176
Email: jabley@ca.afilias.info
URI: http://afilias.info/

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

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

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment

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