Last Version:draft-bates-bgp4-nlri-orig-verif-00.txt
Tracker Entry
Date:31-Aug-1999
Disposition:removed

Network Working Group
Internet Draft
Expiration Date: July 1998

Tony Bates
cisco Systems
Randy Bush
RGnet
Tony Li
Juniper Networks
Yakov Rekhter
cisco Systems

# DNS-based NLRI origin AS verification in BGP

draft-bates-bgp4-nlri-orig-verif-00.txt

## 1. Status of this Memo



## 2. Abstract

This document describes how a BGP speaker may verify that the Network Layer Reachability Information (NLRI) of a prefix received from a peer is consistent with the allocation of IP address space as determined by the Internet Registry system. These verification procedures rely on the DNS to provide a repository of information about address space allocation provided by the Internet Registry system.

Note that this is not a repository of announceable prefixes, but rather of allocation of delegated address space.

## 3. Motivations

IP address space is allocated by the Internet Registry system [RFC2050]. To provide Internet-wide IP connectivity it is imperative that the information provided by the Internet routing system be consistent with IP address allocation. Specifically, the consistency requirement implies that an organization should inject a route for a particular IP address prefix into inter-domain routing only if the address space covered by the prefix was allocated to the organization via the Internet Address Registry system. To provide adequate fault isolation and robust Internet-wide routing in the presence of either misconfiguration or malicious attacks on routing, procedures/mechanisms which allow operators to enforce this consistency requirement are essential.

This document describes procedures and mechanisms that will allow operators to determine the correct origin AS of a prefix advertised into interdomain routing. While other procedures and mechanisms are also necessary to provide reasonably secure routing, such procedures and mechanisms are beyond the scope of this document.

This document does not address the related, but orthogonal, issue of determining the authenticated identity of the routing domain advertising a given prefix.

## 4. Overview of Operations

We propose that information about IP address space allocation provided by the Internet Registry system be maintained within the DNS [DNS] under the bgp.in-addr.arpa domain. This domain is to be further subdivided into additional sub-domains to reflect the allocation structure associated with IP address space. Within this domain, and all of its subdomains, each node in the DNS tree (i.e., each fully qualified domain name (FQDN)) represents one or more IP address prefixes allocated by the Internet Registry system.

This document uses Autonomous System numbers (ASs) to identify entities to which address space has been allocated.

For each address prefix, the DNS may contain either (a) the AS to which the address space described by the prefix has been allocated, or (b) NS delegation to name servers authoritative for the zone identified by the address prefix, or (c) no information about the prefix.

When a BGP speaker receives a route from an external neighbor, the speaker uses the information provided by the DNS to verify that the prefix was legitimately allocated to the AS/organization that injected the route into the inter-domain routing system. If the speaker determines that the NLRI of the route wasn't legitimately allocated to the organization, the speaker rejects the route.

## 5. Extensions to the DNS

The mechanism described in this document requires one new Resource Record (RR):

Autonomous System RR (AS):

This RR consists of two fields, 'AS' and 'Prefix Length'. The AS field contains an AS number, encoded as a two octet unsigned integer (0 through 65535), with 0 and 65535 having special meanings. The Prefix Length field contains an octet encoding the length of the address prefix associated with the node named by the RR (0 through 32).

## 6. Procedures for populating the DNS with the Address Allocation information

A node under the bgp.in-addr.arpa domain may contain either (a) a set of NS RRs that specify name servers authoritative for the zone covered by the address prefix associated with the node, or (b) a CNAME RR, or (c) a set of one or more AS RRs, where each AS RR specifies the AS to which the address prefix has been allocated via the Internet Registry procedures and the length of the allocated address prefix.

Discussion: An alternative to constructing this information under the in-addr.arpa domain would be to pick up some other domain (e.g.,ipv4.nlri.ietf.org). Comments and suggestions are welcome.

CNAME RRs are used for allocations which occur on non-octet boundaries as described in draft-ietf-dnsind-classless-inaddr-04.txt.

The AS field of an AS RR may contain the special value zero (0). This value indicates that the DNS may not contain all the information about allocations out of the address prefix as defined by a combination of the node and the Prefix Length field of the AS RR. In other words, the allocation status of this space is not known. This is distinct from the case where the space is not allocated or it is known to be allocated to some particular AS.

The AS field of the AS RR may also contain the special value 65535. This value indicates that the address prefix associated with the address space has not been allocated by the Internet Registries. An AS RR with an AS value of 65535 can also be used to prevent authentication of certain prefixes.

While 0/0 is not allocated address space per se, as some routing domains use default announcement, default should be allowed in practice. Hence we propose 0/0 be considered unauthenticated (AS of zero) and all truely unallocated space be specifically so marked (AS 65535).

## 7. Conventions for encoding address allocations in DNS names

Syntactically, a DNS name is a series of text 'labels', separated by the '.' character. Within the bgp.in-addr.arpa domain, a label that is a decimal number is used to represent an octet within a prefix. To indicate partial octets, we use the notation <value>/<length> where the <value> contains the value of the last significant octet in the prefix and the <length> is the prefix length. Thus, the prefix 10.1.128/20 may be encoded in a DNS name as 128/20.1.10.bgp.in-addr.arpa.

In addition, for this proposal to work, the hierarchy of address

allocations must be explicitly encoded in the name through the addition of one or more labels. This also implies that no labels may be removed as part of the allocation of portions of a prefix.

If a prefix is allocated on a non-octet boundary, then the allocating domain constructs the name by first adding the labels for the additional full octets, if any, in reversed order to the leftmost position in the name. Then, the label for the partial octet is added as the leftmost position in the name. This name is given an NS RR. As always, normal DNS syntax applies and the resulting name need not be fully qualified.

For non-octet allocations, the NS record is necessary but not sufficient. In addition, a number of CNAME RRs must be added. Recall that the partial octet specifies a number of significant bits in the least significant octet in the prefix. One CNAME RR must be created for each possible value of the remaining bits. The name that the CNAME RR points to (i.e., the name on the right hand side) is constructed by using the value of the least significant octet concatenated with the fully qualified name used for the NS RR. These RRs allow lookups for longer prefixes to redirect through the correct allocation.

A prefix can be extracted from a DNS name constructed using the above conventions by using the labels that represent full octets and the leftmost label (if any) that represents a partial octet. These labels are then reversed in the normal in-addr.arpa manner.

This particular naming scheme is a suggested convention, and alternate semantically equivalent conventions are also perfectly acceptable.

## 8. An Example

The following example illustrates how the DNS might be populated with address allocation information.

```
; the root
$ORIGIN     bgp.in-addr.arpa.       ;well-known root zone
@           SOA     (...)           ;presume ns etc. for zone
0           AS      0 0             ;default not allocated but Ok
1           NS      ns0.bbn.com.    ;allocate 1/8 to bbn
205         NS      ns0.arin.net.   ;allocate 205/8 to arin

; ns0.bbn.com  -  bbn's server
$ORIGIN     1.bgp.in-addr.arpa.
@           SOA     (...)           ;presume ns etc. for zone
            AS      1 8             ;claim allocation for 1/8

; ns0.arin.net  -  arin's server
$ORIGIN     205.bgp.in-addr.arpa.
@           SOA     (...)           ;presume ns etc. for zone
            AS      65535 8         ;205/8 is delegated in parts
0           NS      ns0.digex.net.  ;delegating 205.0/16
1           NS      ns0.verio.net.  ;delegating 205.1/16

; ns0.digex.net  -  digex's server
$ORIGIN     0.205.bgp.in-addr.arpa.
@           SOA     (...)           ;presume ns etc. for zone
```

```
            AS      2548 16         ;claim allocation for 205.0/16

; ns0.verio.net  -  verio's server
$ORIGIN     1.205.bgp.in-addr.arpa.
@           SOA     (...)           ;presume ns etc. for zone
            AS      2914 16         ;205.1/16 is allocated to AS 2914
0           AS      777  24         ;205.1.0/24 is allocated to AS 777
1           AS      2914 24         ;205.1.1/24 is allocated to AS 2914
; delegate 205.1.2/23 using the classless in-addr hack
2           CNAME   2.2/23.1.205.bgp.in-addr.arpa.
3           CNAME   3.2/23.1.205.bgp.in-addr.arpa.
2/23        NS      ns.cust.com.    ;delegate 205.1.2/23, or
                                    ;205.1.2/24 and 205.1.3/24
                                    ;to customer server
4           AS      42 22           ;205.1.4/22 is allocated to AS 42
            AS      0  22           ;also allocated elsewhere
8           AS      666 21          ;205.1.8/21 is allocated to AS 666
; delegate 205.1.16/22 using the classless in-addr hack
16          CNAME   16.16/22.1.205.bgp.in-addr.arpa.
17          CNAME   17.16/22.1.205.bgp.in-addr.arpa.
18          CNAME   18.16/22.1.205.bgp.in-addr.arpa.
19          CNAME   19.16/22.1.205.bgp.in-addr.arpa.
16/22       NS      ns.cust.net.    ;delegate 205.1.16/22 and longer
                                    ;to customer

; ns.cust.com - 2/23 server
$ORIGIN     2/23.1.205.bgp.in-addr.arpa.
@           SOA     (...)           ;presume ns etc. for zone
            AS      4242 23         ;AS 4242 claims 205.1.2/23

; ns.cust.net - 16/22 server
$ORIGIN     16/22.1.205.bgp.in-addr.arpa.
@           SOA     (...)           ;presume ns etc. for zone
16          AS      222  23         ;AS 222 claims 205.1.16/23
18          NS      ns.c1.cust.net. ;delegate 205.1.18/24
                                    ;to a customer's campus
```

## 9. Procedures for verifying BGP routing information

Given a prefix, a lookup in the bgp.in-addr.arpa domain is done by padding the least significant side of the prefix with zeros to an octet boundary and then reversing the octets, as is normally done within the bgp.in-addr.arpa domain. A normal DNS lookup on the resulting name may involve multiple CNAME records, eventually resulting in a FQDN.

We define that a DNS node, authenticated by DNSSEC and under the bgp.in-addr.arpa domain, 'matches' a particular prefix if (a) the result of a lookup on the prefix is the node, and (b) the node contains an AS RR with the value of the Prefix Length field less than or equal to the length of the prefix. We refer to any such AS RR as a "matching" AS RR.

If a BGP speaker performs a lookup on a prefix and cannot find a match, it first clears the least significant set bit in the least significant octet in the prefix and performs another lookup. If there is no set bit in the least significant octet, it then discards the least significant octet from the prefix and performs another lookup. The AS RRs that result from this lookup are compared to the original, unmodified prefix to determine if there is a match.

Using the example from the previous section, an address prefix 205.1.4/22 matches the node 4.1.205.bgp.in-addr.arpa. The matching AS RR is "AS 42 22". An address prefix 205.1/16 matches the node 1.205.bgp.in-addr.arpa with a matching AS RR of (AS 2914 16).

Further, an address prefix 205.1.0/18 matches the node 1.205.bgp.in-addr.arpa, with the matching AS RR as (AS 2914 16). Note that in this case, the first lookup fails and requires a second lookup. Similarly, the prefix 205.1.5/24 matches the node 4.1.205.bgp.in-addr.arpa with the matching AS RR as (AS 42 22).

The following assumes that a BGP speaker has sufficient routing information to have access into the DNS system.

A route may be marked "Authenticated", "Unauthenticated", or "Authentication Failed".

When a BGP speaker receives a route from an external peer, the speaker marks the route as "Unauthenticated", and then performs the following:

- the speaker checks the DNS for the presence of a node that "matches" the NLRI of the route.

- if there is a matching node with an AS RR where the value of the AS field is equal to the origin AS of the BGP AS_PATH attribute of the received prefix, the route is marked as "Authenticated."

- if there is a matching node with an AS RR where the value of the AS field is 65535, the route is marked as "Authentication Failed."

- if there is a matching node with an AS RR where the value of the AS field is is zero (0), the route is left as as "Unauthenticated."

- if there is a matching node with an AS RR where the value of the AS field is neither 0, 65535, nor the origin AS of the received prefix, the route is marked as "Authentication Failed."

- in all other cases the marking of the route is not modified, i.e. it remains "Unauthenticated."

The authentication status of a route has a time limit, maintained in the authentication status timer. If the origin of a route is Authenticated, then the timer is set to the Time To Live (TTL) of the matching AS RR. The timer for a route marked as "Unauthenticated" is set to RouteAuthenticationRetryTimer value (by default 24 hours). Note that the authentication status timer is not propagated in BGP Update messages.

When the timer expires, the route is marked as "Unauthenticated", and the BGP speaker performs the above procedures.

A BGP speaker MAY use and advertise to other BGP speakers a route that is marked as either Unauthenticated or Authenticated. As a matter of local policy the BGP speaker in its route selection policy MAY give preference to routes marked as Authenticated.

A BGP speaker MUST NOT use and/or advertise to other BGP speakers a

route that is marked as "Authentication Failed".

Since a BGP speaker may perform the above procedures asynchronously with route installation and advertisement, a speaker may advertise a route marked as "Unauthenticated", but then might later mark the route as "Authentication Failed". In this case the speaker MUST withdraw the route, and stop using it.

As a local matter a BGP speaker MAY "preload" as much of the DNS information as it wants. Doing this would allow the speaker to accelerate the marking of a newly received routes.

A BGP speaker, MAY (under control of its local configuration) exempt certain routes from the above verification procedures.

In addition to address allocations, the bgp.in-addr.arpa domain can be used to encode aggregated prefixes. As with other prefixes, the AS RR is used to indicate the origin of the aggregate. Insertion of information about the aggregate requires the cooperation of the entity controlling the appropriate point in the namespace.

## 10. Use of TXT RRs

Instead of introducing a new RR type, the AS RR, the scheme described in this document might use a TXT RR, where the information encoded in the TXT RR would be the same as in the AS RR (although the encoding will be different). One of the problems with using the TXT RRs is that it redefines the semantics of the TXT RR, which at least will be somewhat confusing. Further, if a TXT RR is used for initial deployment, there is a likelihood that no transition would ever be made to the AS RR.

## 11. Security Considerations

This entire document is about security considerations.

DNSSEC should be used in conjunction with the procedures described in this document to provide authentication for the DNS information.

## 12. Acknowledgments

The authors would like to acknowledge the contributions of the DNSSEC working group and the authors of draft-ietf-dnsind-classless-inaddr-04.txt for their contributions without which, this work would have been impossible. Additionally, the authors would like to thank Jerry Scharf for commenting on the work as it progressed.

## 13. References

[BGP-4]

[DNS]

[DNSSEC]

## 14. Author Information

Tony Bates
cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134
Email:  tbates@cisco.com

Randy Bush
RGnet, Inc.
5147 Crystal Springs
Bainbridge Island, WA 98110
E-mail: randy@psg.com

Tony Li
Juniper Networks, Inc.
385 Ravendale Dr.
Mountain View, CA 94043
E-mail: tli@juniper.com

Yakov Rekhter
cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134
Email:  yakov@cisco.com