Internet Draft
Category: Proposed Standard
Expires: November 2008

L. Donnerhacke
IKS GmbH
W. Wijngaards
NLnet Labs
May 5, 2008

DNSSEC protected routing announcements for BGP
draft-donnerhacke-sidr-bgp-verification-dnssec-04

Status of this Memo

Abstract

This document describes an infrastructure for real time verification of routes reveived via BGP4. Some DNS query types are introduced to check the origin of a prefix and validity of the AS path. The crypto part can be offloaded from the routing engine by sending a DNS query and checking the AD bit in the DNS response. The proposal depends on the DNS scalability and caching mechanisms as well as PKI introduced by DNSSEC.

Table of Contents

Nomenclature

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 [RFC2119].

The process of checking an DNS record set to match the DNSSEC key hierarchy is called "validation" in this document.

The process of checking an BGP route for origin and path consistency is called "verification" in this document.

An unordered collection of Autonomous System (AS) numbers is called "AS number set" in this document.

The DNS resource record representing an AS number set is called "ASSET" in this document. "ASSET RR" means the whole DNS resource record, while "ASSET RDATA" names the payload section of the ASSET RR.

The AS-SET object in the Internet Routing Databases (IIRB) is called "AS database set" in this document.

The aggregated AS number set stored in the BGP path information is called "aggregated AS path set" in this document.

1. Introduction

BGP hijacking is a serious problem in the current internet. In an ideal world those cases can't happen at all, because honest operators apply filters on their BGP4 [RFC4271] peerings in order to catch fat-fingered misconfigurations. The filters can automatically derived from existing, well maintained routing databases. A look at actual routing tables suffices for a reality check.

This document proposes a real time verification method of received BGP announcements for routers: An efficient, automatic, and external filter. The described infrastructure allows the filtering of bogus announcements even after some steps of transit.

All the routing resource meta information is simplified and mapped into a DNS hierarchy. The allocation and assignment chains for AS and IP numbers from the IANA via RIR and LIRs to the routing entities are reflected by the appropriate DNS delegation chain [iananum].

At the routing entity level (i.e. the ISP or customer) the delegated prefix is mapped to the AS number set, which injects the route into the DFZ. Futhermore the peering state is modeled as a two way announcement at this level.

Because of DNSSEC [RFC4033] all those delegations and announcements can be validated. When querying, the router can do the DNSSEC validation itself or delegate it to the next validating resolver. A validated response contains a special bit (Authenticated Data) assuming the trustworthiness of the link between the resolver and the router. So the router can work with validated data without performing expensive cryptographic operations and difficult lookup algorithms.

Some special issues arise from the interaction of building the routing table while requiring a working interconnection for verification, and from verification and other operational errors.

2. DNS Mapping

The mapping is designed to ease the route verification process. All verification steps should be performed in a building a simple DNS query and looking for a single value in the validated DNS response set. Furthermore the whole process should be easy to debug.

A new zone BGP.ARPA is introduced to hold the routing resources. For AS number mapping, the zone AS.BGP.ARPA is used. IPv4 prefixes are mapped into IPV4.BGP.ARPA and IPv6 prefixes are mapped into IPV6.BGP.ARPA.

2.1. The ASSET Resource Record

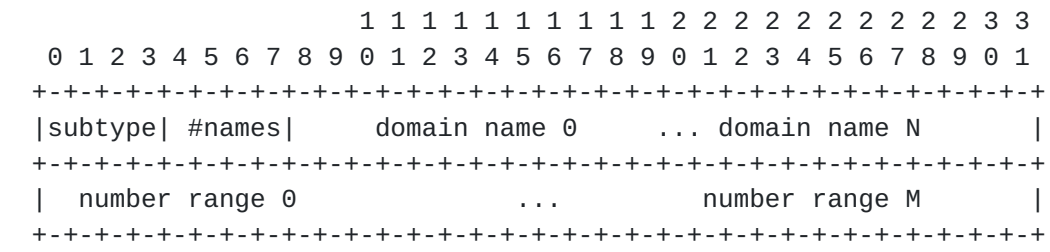
The ASSET RR contains a AS number set in a compact format. ASSET RRs can be point to multiple other ASSET RRs. Merging those referenced ASSET RRs allows to include AS database sets (in form of ASSET RRs) and to implement really huge AS number sets (as smaller ASSET RRs).

The type value for the ASSET RR is TBD (decimal).

The ASSET RR is defined for class IN.

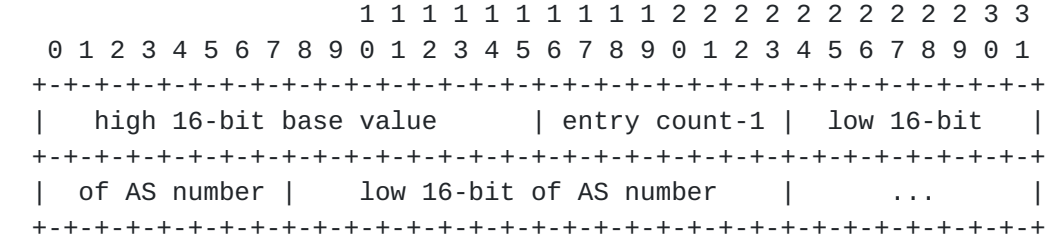
2.1.1. ASSET RDATA wire format

The ASSET RDATA is the concatenation of a single octet with subtype and name nibbles. The name nibble is bits 4-7, and indicates how many names will follow, zero or more names of referenced ASSET RRs. After the names are zero or more number ranges up to the end of RDATA. The subtype and name count are unsigned integers in network order.



A sender MUST NOT use DNS name compression for the names. This allows the ASSET RR to be handled by older software [RFC3597].

An number range is encoded using an unsigned 16-bit base value in network byte order, a single octet range length which is an unsigned integer with the number of entries - 1 and up to 256 entries of 16-bit offset values. Each range encodes 32-bit AS numbers by combining the offset as lower 16-bit with the base as higher 16-bit.



Encoder software MAY reorder AS numbers for efficent encoding. Encoders MAY issue a warning, if the encoded RDATA exceeds 1000 bytes. Encoders SHOULD reject the input, if the encoded RDATA exceeds 3500 bytes. Encoders MUST reject the input, if the encoded RDATA exceeds 55 kbytes.

The following values are defined for the subtype value:

0 Union of the referenced ASSET RRs and embedded number ranges

This ASSET RR corresponds to the union set of all AS number sets corresponding to the embedded number ranges and all the AS number sets corresponding to the named ASSET RRs recursively. While processing ASSET references, the querier MUST provide loop protection. ASSET references are likely to become circular. An ASSET RR with no reference names and no number ranges is allowed and corresponds to the empty set.

1 Set of all possible AS numbers

This ASSET RR provides a catch all for any possible AS number. If this resource record is found while recursively processing subtype 0 records, the whole recursion process can be aborted resulting in

the largest possible AS number set. This RR does not contain any referenced names nor any number ranges. So the RDATA wire format of this subtype consists of the single octet "16" (decimal).

2 Transition marker

This ASSET RR is used while setting up the global infrastructure to mark "to be done" points. Initially this RR has the same semantics as the subtype 1 RR. In the medium term, the semantics of this RR will be changed to generate warnings or errors. In the long term this RR will vanish. This RR does not contain any referenced names nor any number ranges. So the RDATA wire format of this subtype consists of the single octet "32" (decimal).

3-15 Reserved

2.1.2. ASSET RDATA representation format

Subtype 0 resource records are represented by a space seperated sequence of domain names (of the referenced ASSETs) followed a space seperated sequence of AS numbers in the asdot format [as4byte]. Name representations without a trailing dot are abbrivated names in the current $ORIGIN of the zone file. The first term matching the asdot format, i.e. consisting of digits and an optional single dot only, terminates the domain name sequence and starts the AS number sequence.

Subtype 1 resource records are represented by the case insensitive term "any".

Subtype 2 resource records are represented by the case insensitive term "transition".

Ambigous domains names SHOULD not be abbrivated.

2.1.3. Fallback to TXT

To ease deployment, ASSET RR can be implemented as TXT records, containing the representation format of the ASSET RR as RDATA. This allows to provide DNS mapped data in the BGP.ARPA zone without running ASSET aware DNSSEC tools or DNS servers.

Routing devices MUST first query for and understand the ASSET RR. Only if the final response contains an authenticated denial of existance (NSEC) record proving the existance of a TXT record for exactly the queried name, the routing device MUST ask for the TXT record. The TXT record is not queried for in other circumstances. So a minimal amount of queries is sent.

This fallback procedure will be declared obsolete in the medium term.

2.2. Prefix origin

To query the origin AS number set for a prefix, the prefix is transformed similar to reverse lookups and the DNS is queried for ASSET RRs. The DNS response results in a (possibly empty) AS number set.

IPv4 prefixes are queried in the same way as classless IN-ADDR.ARPA reverse delegation [RFC2317], but in IPV4.BGP.ARPA instead of IN-ADDR.ARPA. The least specific label MUST contain the netmask of the prefix.

IPv6 prefixes are queried in the same way as IP6.ARPA reverse delegation [RFC3596], but in IPV6.BGP.ARPA instead of IP6.ARPA. If a delegated misses the nibble boundary, the same technique MUST be used as for IPv4. The least specific label MUST contain the netmask of the prefix.

Prefixes which MUST NOT appear in global routing tables do not get an entry in the delegation hierarchy. I.e. IPV6.BGP.ARPA should not contain an entry for F. For locally distributed prefixes, the local resolver SHOULD provide more specific zones and trust anchors for those prefixes. This way exception handling in the routing devices is minimized: They simply ask for the data they have to verify.

During rollout of this proposal, a transition period is necessary to allow the AS operators to set up the necessary zones and get the delegations. During the transition, the RIRs SHOULD derive the AS data from the [irdb] or MAY add the "transition" ASSET subtype for the allocated prefixes.

Please note, that for multicast routing the destination addresses are not distributed via BGP4, but only the source addresses. So the multicast group addresses from 224.0.0.0/4 and FF00::/8 are never looked up and will not be delegated in BGP.ARPA.

Example:

```
$ORIGIN 192/20.17.217.IPV4.BGP.ARPA.
@ ASSET 15725     ; CNAME delegation not necessary

$ORIGIN 8.D.B.4.1.0.0.2.IPV6.BGP.ARPA.
8/32 ASSET 15725  ; Delegation will have a CNAME for it
```

2.3. AS Peering

Peering between two AS is fourfold: Sending and accepting on each site of a peering session. Futhermore peering policies depend on the

address family of the prefix [RFC4012].

To query the peering policy of AS A in regard to AS B, both AS numbers are put together with the protocol and the peering direction, and the DNS is queried for ASSET records. The DNS response results in a (possibly empty) AS number set.

Local use of private AS numbers SHOULD be announced by adding specific zones and trust anchors at the local resolver. This way exception handling in the routing devices is minimized: Routing devices handle private numbers in the same way as ordinary assigned AS numbers.

During rollout of this proposal, a transition period is necessary to allow the AS operators to set up the necessary zones and get the delegations. During the transition, the RIRs SHOULD derive the AS data from the [irdb] or MAY insert the "transition" subtype of ASSET.

To ease the delegation of AS numbers ranges to a RIR and in order to keep the zone size small for efficent DNSSEC operation, the combining of the two AS numbers for a peering from AS A to AS B is processed in the following way: The 32-bit AS number of A is written as <high order 16-bit value in decimal>.<low order 16-bit value in decimal as five dot seperated digits>, then the order of the labels is reversed, and AS.BGP.ARPA appended. The resulting zone SHOULD be under the control of the AS operators. The asdot format of AS B followed by the peering direction ("import" or "export") and the protocol family is prepended to this zone apex.

Conversion example:

```
AS15725 ->  0.15725 -> 5.2.7.5.1.0
AS3.10  ->  3.00010 -> 0.1.0.0.0.3
AS12.34 -> 12.00034 -> 4.3.0.0.0.12
```

Peering information example:

```
$ORIGIN multicast.ipv4.5.2.7.5.1.0.AS.BGP.ARPA.
3.3.export ASSET 15725   ; AS15725 exports to AS3.3 only itself
3.3.import ASSET 3.3     ; AS15725 imports from AS3.3 only 3.3
15725.export ASSET ANY   ; AS15725 may prepend
15725.import ASSET ANY   ; AS15725 may prepend

$ORIGIN ipv4.3.0.0.0.0.3.AS.BGP.ARPA.
5539.import.unicast ASSET ANY
5539.export.unicast ASSET 3.3
6695.import.unicast ASSET as-decix.5.9.6.6.0.0.AS.BGP.ARPA.
6695.export.unicast ASSET 3.3
15725.import.multicast ASSET 3.3
15725.export.multicast ASSET ANY
```

```
$ORIGIN 5.9.6.6.0.0.AS.BGP.ARPA.
as-decix ASSET local as-hosteurope.3.7.7.0.2.0.AS.BGP.ARPA. ...
local    ASSET 12510 12989 20899 25286 31334 31529 41039 42416

$ORIGIN 3.7.7.0.2.0.AS.BGP.ARPA.
as-hosteurope ASSET 20773
```

2.4. Delegation hierarchy

Currently IPv4 addresses are allocated to the RIRs as /8. The delegation at IPV4.BGP.ARPA follows this and delegate the zones to RIR's name servers. This mimics the delegation from IANA to the RIRs in IN-ADDR.ARPA.

IPv4 addresses are allocated to the LIRs in various sizes. Delegation of the allocate is done by the RIR in classless manner. Futhermore the classless prefixes at this level up to the next classful boundary have to be delegated to the LIR, too. The use of CNAME for classless delegations and DNAME for smaller prefixes is REQUIRED.

Example:

```
$ORIGIN 17.217.IPV4.BGP.ARPA.
192/20 NS avalon.iks-jena.de.
$GENERATE 192-207/8 $/21 CNAME $/21.192/20
$GENERATE 192-207/4 $/22 CNAME $/22.192/20
$GENERATE 192-207/2 $/23 CNAME $/23.192/20
$GENERATE 192-207/1 $/24 CNAME $/24.192/20
$GENERATE 192-207   $    DNAME $.192/20
```

If the AS operators announces the full allocate, the LIR adds the ASSET RR to the delegated zone. If the AS operators deaggregate the allocate and/or permit assignments to be seperatly announced, the LIR adds further ASSET records or set up delegations to the AS operators.

IPv6 address delegation mimics the delegation in IP6.ARPA. Please note the similarity to IPv4 if an allocate or assignment miss the nibble boundary. Futhermore the classless prefixes at this level up to the next classful boundary have to be delegated to the LIR, too. The use of CNAME for classless delegations and DNAME for smaller prefixes is REQUIRED.

Example:

```
$ORIGIN 0.1.0.0.2.IPV6.BGP.ARPA.
8/22 NS ns.ripe.net.
$GENERATE 8-15/2 ${0,0,x}/23 CNAME ${0,0,x}/23.8/22
$GENERATE 8-15/1 ${0,0,x}/24 CNAME ${0,0,x}/24.8/22
$GENERATE 8-15   ${0,0,x}    DNAME ${0,0,x}.8/22
```

AS number allocations from IANA to the RIRs are done in large blocks. IANA has to delegate every zone for which the RIR might be responsible, but not more. Additional zones MAY be introduced using DNAME to delegate single AS numbers via RIRs, if the RIR can't maintain the LIRs data directly in the IANA zone (sometimes the IANA delegation can be directly to the LIR).

RIRs assign single AS numbers to the LIRs and delegate the appropriate zone.

AS database sets are a common tool in the Internet Routing Registy [irdb] and maintained by a AS operators. AS operators SHOULD provide their common AS database sets of the routing registry directly as ASSET RR in their associated AS.BGP.ARPA zone. Other AS operators are encouraged to refer to those ASSET records instead of generating the own ASSET RR using a database toolset. Referencing provides much smaller zone files and "automatic" update of changes. On the other hand generating the whole AS number set directly from the database provides a locally cached and therefore more stable version of the peering information.

2.5. Private numbers

The delegation described in the previous section can't cover usage of private addresses or AS numbers. Private numbers are not delegated, but only reserved by IANA. Instead of officially marking reserved ranges to hand over the control to local router configuration, the reserved ranges are simply not delegated at all.

If private addresses or numbers are in use, the DNS operators of this environment SHOULD set up local zones in BGP.ARPA, sign them and locally distribute the trust anchors. This way the verification process for routers stays simple. The zones SHOULD be shared between between involved AS to avoid duplication of configuration data.

Configured local zones for private space MUST NOT be redistributed in the official BGP.ARPA tree. DNS operators need to make sure, that those zones are not visible in unrelated AS. The authoritive name servers serving local zones in BGP.ARPA SHOULD be kept seperate from the authoritive name servers visible to the public. When using local zones in BGP.ARPA, the recursive, validating resolver used for router equipment SHOULD be kept seperate from the DNS resolvers for customers.

2.6. Route and AS path aggregation

A not uncommon BGP setup is to aggregate several more specific routes to a larger prefix. The aggregated prefix is injected into the

global routing table by the aggregating AS. Optionally the AS path can contain a aggregated AS path set, in order to prevent the aggregated route to be propagated back.

For the purpose of verifying the origin of a prefix, the whole aggregation process as well as the aggregated AS path set can be ignored. So aggregated AS path sets MUST be stripped from the AS path before verification. The aggregating AS is considered as the origin of the aggregated prefix.

3. Verification

A router receives routes in a given address family consisting of a prefix and a AS path via BGP4. The router has to verify, if the incoming route is allowed or not.

The router has to check the following criteria:
- is the originating AS allowed to inject the route?
- do all the AS in the path peer as claimed?
- does the recorded path fullfill the peering policies?

3.1. Verification algorithm

To check the origin, the router queries for the prefix as described in 2.2. If the last AS in the path, which is not part of an aggregated AS path set, is in the AS number set of the DNS response, the origin is verified. If the prefix can't be found, the check fails.

To check the peering policies, for each pair of sequenced AS in the path a query as described in 2.3. is performed. Aggregated AS path sets are ignored. The policy of the sending AS MUST contain all AS numbers of the path tail including the sending AS number for the address family and for the direction "export". The policy of the receiving AS MUST contain all AS numbers of the path tail including the sending AS number for the address family and for the direction "import". If an AS can't be found, the check fails.

The router SHOULD NOT check the recursive peering policy for duplicate AS numbers, which are the result of prepending. AS operators SHOULD add a self peering entry, if they use prepending.

If all checks succeed, the route is accepted.

If the check fails, the processing for this route MUST be delayed and retried. This is necessary, because BGP4 does announce a route only once during a peering session. If the problem with the DNS disappears, the route will not be reannounced in the BGP4 session, but MUST be accepted now.

Routers MAY record the TTL of the responses and assign the route the minimum of all TTLs to regularly reverify the route. Routers MUST NOT drop the route solely because the TTL times out.

3.2. Offloading crypto

Routers are not designed for DNS processing and should not do it. DNSSEC offers a validating resolver and a Authenticated Data bit in the response header. Routers SHOULD ask a validating resolver and rely on the AD bit in the response [RFC4033].

Using this approach, PKI processing, caching, and debugging is handed over to specialized software and admins.

3.3. Zone slaving

Normally name servers of new AS can't be reached, because the new route to the prefix of the AS can't be verified until the route to the nameserver is active.

That's why all zones in BGP.ARPA MUST have secondaries in other AS. The RIRs are urged to provide public secondaries for their LIRs and their routing customers.

To avoid a net split after a hypothetical major outage, running secondaries of other zones, especially of those of the peering AS, is RECOMMENDED. Name server operators in BGP.ARPA SHOULD allow zone transfers to everyone [RFC1034].

3.4. Utilizing peer's cache

Querying each record from the authoritive name servers for every recursive resolver would cause a storm of queries from the whole internet if a prefix is injected or flaps. Such a query storm is similar to a DDoS and should be avoided.

Any received prefix comes from a peer router which should have verified the prefix before sending. So the peer's router knows it's local resolver which in turn may have cached all the necessary data to validate the prefix.

Routing devices SHOULD add the peer's router name as NS for BGP.ARPA in the authority section, and the peer's router address as A or AAAA for the router name to the additional section of it's own queries to it's own validating resolver. The name for the NS and A/AAAA entry is not important, it only connects the NS RR and the A/AAAA RR. The qname of the NS RR can be considered as the maximum scope of allowed DNS queries.

The resolver SHOULD ask the mentioned address first for all necessary recursive queries regarding this query. It MUST NOT add the router address into the cache as a valid nameserver for the zone BGP.ARPA. If the peer's resolver denies access or is unreachable, the resolver MUST NOT query the peer's resolver for a reasonable time. If the necessary data can not be obtained from the peer's resolver, the resolver MUST start the normal DNS resolving algorithm. Sending DNS queries to a different host is a security risk, so resolvers SHOULD permit this redirection only for known sources (their own routers) and MAY limit this feature to zones under BGP.ARPA.

The peer's router SHOULD forward the queries to it's local resolver. It is NOT RECOMMENDED for the router to provide this service for everyone, so the routing device SHOULD permit DNS forwarding only for sources of the peering AS and MAY use it's BGP routing table for this purpose.

The peer's resolver SHOULD respond using it's cache data as a regular recursor providing forwarding service. The resolver MUST take care not to serve information for private zones, this can also be accomplished by having two resolvers, one for the router, one for outside queries.

3.5. Bootstrapping

There are two strategies to handle the startup of AS routing.

3.5.1. Delaying verficiation

Routers SHOULD postphone all the checkings but accept all the routes as long as the routing table stays below to a configurable value. This behaviour allows a cold start after disasterous problems: The verification is postphoned until DNS becomes useable.

3.5.2. Utilizing peer's resolver

While bootstrapping, foreign AS will need security information to accept routes originating from an AS. This can be accomplished by putting master authority DNS servers for the AS AS.BGP.ARPA zone, the AS prefixes in IPV4.BGP.ARPA and IPV6.BGP.ARPA inside the AS and reachable by the forwarding resolver. Far away AS can then query their neighboring routers, which will forward the query to their resolver, which will ask routers that are closer, and so on, towards the authority server.

A resolver performing such router forwarding MUST be able get the address from its router for the resolver in a neighboring AS that is closer to a destination AS or prefix. The router consults its

routing tables to determine the AS neighbor closer for a prefix. For unrouted prefixes, the router has no answer, because it does not know a closer AS, or the resolver address for the closer AS.

The router is queried for the neighboring resolver address with a query of type NEIGHBOR_NS, and name in AS.BGP.ARPA, IPV4.BGP.ARPA, IPV6.BGP.ARPA. The reply contains an NS for BGP.ARPA and addresses for the remote server that handles forwarded router queries in the neighboring AS. This NS MUST NOT be stored in the validator cache as a nameserver for BGP.ARPA. Query RR type NEIGHBOR_NS has type code TBD3 (decimal).

To be able to validate the DNSSEC chain of trust while the root, IANA, RIR and other servers are unreachable during bootstrapping, the DNSSEC chain of struct information MUST be stored. The AS stores such information in CHAINOFTRUST RRs at the zone apex for its AS.BGP.ARPA, IPV4.BGP.ARPA and IPV6.BGP.ARPA zones. The information was inserted at the last zone sign for the zone, so may be out of date regarding current information served by parent zones, but the information MUST be verifiable using the current trust anchors.

The CHAINOFTRUST RR has type code TBD2 (decimal) and is class independent. Its wire format consists of a 16-bit value type code and an uncompressed original domain name, and the remainder up to rdata length is the original rdata and presented in base64. The RR type is used to wrap DNSSEC chain of trust data so that it can be stored at the authority servers of the AS without conflicting with data from other AS. It is RFC3597 compliant. The data can be copied from the parent authority servers verbatim. The CHAINOFTRUST RRset must also be signed by the ZSK as usual. An example:

```
$ORIGIN 3.0.0.0.3.as.bgp.arpa.
@ CHAINOFTRUST DNSKEY . <base64 data of RR>
  CHAINOFTRUST RRSIG . <DNSKEY data>
  CHAINOFTRUST DS arpa. <data>
  CHAINOFTRUST RRSIG . <DS data>
  CHAINOFTRUST DNSKEY arpa. <data>
  CHAINOFTRUST RRSIG arpa. <DNSKEY data>
  CHAINOFTRUST DS bgp.arpa. <data>
  CHAINOFTRUST RRSIG arpa. <DS data>
  CHAINOFTRUST DNSKEY bgp.arpa. <data>
  CHAINOFTRUST RRSIG bgp.arpa. <DNSKEY data>
  CHAINOFTRUST DS as.bgp.arpa. <data>
  CHAINOFTRUST RRSIG bgp.arpa. <DS data>
  CHAINOFTRUST DNSKEY as.bgp.arpa. <data>
  CHAINOFTRUST RRSIG as.bgp.arpa. <DNSKEY data>
  CHAINOFTRUST DS 3.as.bgp.arpa. <data>
  CHAINOFTRUST RRSIG as.bgp.arpa. <DS data>
```

```
CHAINOFTRUST DNSKEY 3.as.bgp.arpa. <data>
CHAINOFTRUST RRSIG 3.as.bgp.arpa. <DNSKEY data>
CHAINOFTRUST DS 3.0.0.0.3.as.bgp.arpa. <data>
CHAINOFTRUST RRSIG 3.bgp.arpa. <DS data>
CHAINOFTRUST DNSKEY 3.0.0.0.0.3.as.bgp.arpa. <data>
CHAINOFTRUST RRSIG 3.0.0.0.3.as.bgp.arpa. <DNSKEY data>
RRSIG 3.0.0.0.3.as.bgp.arpa. <CHAINOFTRUST data>
```

The CHAINOFTRUST type can thus become fairly large, and will probably require TCP failover when queried for. Storing a CHAINOFTRUST with original type CHAINOFSTRUCT can be used to refer a validator to more CHAINOFTRUST RRs which can be found at the name pointed to by the domain name stored.

4. Related work

The idea is not new. Directly after the specification of DNSSEC, the provided infrastructure was applied for verifying BGP announcements. Prefix originating verification was proposed by [bates] and discussed by [liauth]. AS mapping to DNS was proposed by [eastlake].

[bates] prefered to define the new record type AS in order to keep the current semantics of TXT. This proposal initially prefered TXT.

Filling the testbed with real world data reveals AS database sets with more than 20000 AS numbers after deaggregation. Using TXT records, the record set exceeds 100 kbyte and all limites for DNS packets. Such record sets can't be retrieved. Mr. Wijngaards developed the ASSET type with bitfields and name chaining. Following the responsibility principle, chaining was extended to multiple references.

Multiple encoding variants of ASSET where tried with real world data: Decimal encoding as TXT, binary encoding of 32-bit numbers, binary encoding if 16-bit numbers within a high 16-bit window, and NSEC like bitmaps within a 32-bit base window. Bitmap encoding is more efficent if RDATA exceeds about 700 bytes. In all other cases the 16-bit encoding as described in 2.1.1 is more compact.

[eastlake] define the AS mapping to DNS using the asplain notation combined with a length indicator of the significat digits. With the introduction of four-byte AS numbers [RFC4893], IANA chooses to allocate a whole <high 16bit> to the a single RIR only, which suggests asdot usage. Futhermore fixed sized formats are easier to handle in embedded devices.

The current proposal chooses to expand the <low 16bit> to five decimal digits and and append the whole <high 16bit> as a single decimal number. This decision does only scale, as long as the number of allocated <high 16bit> keeps small.

The alternate approach of coding the AS number in hex as in IP6.ARPA offers the possibility to follow the IANA allocation policy more closely (allocation step is 0x100). Tests show, that currently 3455 delegations based on decimal number vom IANA to RIRs are necessary, but only 3440 based on hexadecimal numbers. Only if lookup would be done on binary numbers, the number of delegations would drop to 70. In order to ease debugging, this proposal chooses to stick on decimal numbers.

The actual work of the SIDR WG focuses on automatic generation and validation of filters [sidrwg]. AS Path checks are not yet developed.

[wijngaards] is very similar to the current proposal, so the results where merged.

There are other proposals, i.e. a redesign of the BGP4 protocol to include cryptographic authentication of the path and origin [bartels].

5. Test environment

A testbed was build to test implementations and verify assumptions based on this recommendation. The data in the testbed is derived on snapshots of the Internet Routing Registy [irdb] with focus of the RIPE region.

The primary NS for the testbed of BGP.ARPA is IANA.BGP.IKS-JENA.DE. If you run secondaries, Lutz is happy to add them as name servers for the test zones. If you like to get a delegation to maintain your own part in the testbed, please contact Lutz Donnerhacke.

IANA and RIRs are especially encouraged to maintain their own area of responsibility. This way the testbed would be more accurate and the communication channels between the participating parties could be covered.

To gain experience with DNSSEC signed domains up to the root, Lutz Donnerhacke runs a signed root [iksroot], which is expanded to cover the BGP.ARPA testbed. You MUST NOT consider this environment as a permanent resource. It will vanish as soon as the root gets signed [rootsign].

6. Security Considerations

All zones in BGP.ARPA MUST be signed. Local infrastructure between the routers and the validating recursive resolvers SHOULD be secured against data modification or spoofing attacks.

Operational errors in DNSSEC or DNS handling will cause routing problems. Operational errors at RIR or IANA will cause larger shutdowns of global routing. These errors may be mitigated if the CHAINOFTRUST types are queried, and contain data from before the error.

Injecting or flapping routes may cause a storm of DNS queries from routers of the whole internet. Such a request storm is similar to a DDoS attack. Be prepared. Have secondaries. Don't flap.

7. IANA Considerations

IANA should gracefully add the BGP.ARPA zone and maintain the delegations to the RIRs.

IANA should sign the all the zones from the RIR delegation point down to the root. IANA should maintain the resigning and key rollover procedures for those zones.

IANA should set up a Delegate Signer (i.e. manual) update protocol for the delegation points to allow the RIRs to change their keys.

IANA should maintain a registry of ASSET subtype numbers. Those numbers should be updated by IETF consensus.

IANA should assign RR type codes for ASSET, CHAINOFTRUST and NEIGHBOR_NS.

8. References

8.1. Normative References

[RFC1034] Mockapetris, P, "Domain Names - Concepts and Facilities", RfC 1034, November 1987

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC2119, March 1997

[RFC2317] Eidnes, H. and de Groot, G. and Vixie, P., "Classless IN-ADDR.ARPA delegation", RFC2317, March 1998

[RFC3596] Thomson, S. and Huitema, C. and Ksinant, V. and Souissi, M., "DNS Extensions to support IP version 6", RFC 3596, October 2003

[RFC3597] Gustafsson, A., "Handling of Unknown DNS Resource Record (RR) Types", RFC 3597, September 2003

[RFC4012] Blunk, L. and Damas, J. and Parent, F. and Robachevsky, A., "Routing Policy Specification Language next generation", RFC4012, March 2005

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

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

[RFC4893] Vohra, Q. and Chen, E., "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007

[as4byte] Michaelson, G. and Hustone, G., "Canonical Text Representation of Four-octet AS Numbers", Work in Progress: draft-michaelson-4byte-as-representation-05, December 2007

8.2. Informal References

[bates] Bates, T. and Bush, R. and Li, T. and Rekhter, Y., "DNS-based NLRI origin AS verification in BGP", Expired work in progress: draft-bates-bgp4-nlri-orig-verif-00, December 1997

[eastlake] Eastlake, D., "Mapping Autonomous Systems Number into the Domain Name System", Expired work in progress: draft-ietf-dnssec-as-map-05, July 1997

[liauth] Li, T., "Origin Authentication in BGP", Expired work in progress: http://www.academ.com/nanog/feb1998/origin.html, February 1998

[irdb] "The Internet Routing Registry: History and Purpose", http://www.ripe.net/db/irr.html

[iananum] "Number Resources", http://www.iana.org/numbers/

[sidrwg] "Secure Inter-Domain Routing", http://tools.ietf.org/wg/sidr/

[rootsign] "IANA (DEMO) DNSSEC Status",
https://ns.iana.org/dnssec/status.html

[youtube] RIPE NCC, "YouTube Hijacking: A RIPE NCC RIS case study",
http://www.ripe.net/news/study-youtube-hijacking.html,
February 2008

[bartels] Bartels, O., "Requirements for a new routing protocol",
Work in Progress: news:6msps3tgjug-
mvlkk1hcr26jpo8nrfhbmj0@4ax.com, March 2008

[wijngaards] Wijngaards, W., "Securing BGP using DNSSEC", unpub-
lished, April 2008

[iksroot] Donnerhacke, L., "Instructions for a signed root",
https://www.iks-jena.de/leistungen/keys.txt, December 2007

9. Changes history

This section will not appear in the final document. It does provide some convenience hints what changed between the document version. It is not complete nor normative.

Important differences from 02 to 03:
- IP delegation requires always a netmask for propper delegation

Important differences from 02 to 03:
- Wouter Wijngaards added as author
- ASSET RR added in favor of TXT RR
- Peering direction and address familiy moved from RDATA to NAME.
- DNAME for delegations are now REQUIRED instead of RECOMMENDED.
- IRDB AS-Set mappings added
- Bootstrapping seperated out as an extra section
- Utilizing peer's cache section added
- Testbed responsibility assigned to Lutz Donnerhacke
- Added DDoS risks
- Added subtype registry for IANA

Important differences from 01 to 02:
- Removed reserved handling in favor to local served DNS zones.
- Added aggregate handling.

Important differences from 00 to 01:
- Added handling of reserved address space using wildcards.
- Added handling of non routable address space using denial of existence.
- Added classification of multicast address space as non routable space.
- Added transition phase where information is copied from [irdb] or verification is explicitly turned off.
- Added recommendation to explicitly announce prepending as self peering.
- Raised recheck of delayed verifications from SHOULD to MUST.
- Added a section about related work and reasons for design decisions.

10. Acknowledgements

The proposal was developed with the help of Gert Doering and Oliver Bartels in a USENET News discussion about the YouTube hijacking in February 2008 [youtube].

Many thanks go to Tony Li for pointing out several historic documents, and his invaluable comments on the transition phase, reserved areas, and readvertisement of received prefixes.

Wouter Wijngaards independently developed a very similar proposal. Both proposals were merged. Mr. Wijngaards does a wonderful job in developing the DNS related parts.

Authors' Addresses

Lutz Donnerhacke
IKS GmbH
Leutragraben 1
07743 Jena
Germany
Phone: +49-3641-573561
EMail: lutz@iks-jena.de

Wouter Wijngaards
NLnet Labs
Kruislaan 419
Amsterdam 1098 VA
The Netherlands
Phone: +31-20-888-4551
EMail: wouter@nlnetlabs.nl

Full Copyright Statement

Disclamer

Intellectual Property