Ptomaine
Internet-Draft
Expires: November 11, 2003

G. Huston
Telstra
May 13, 2003

NOPEER community for BGP route scope control
draft-ietf-ptomaine-nopeer-03.txt

Status of this Memo

Copyright Notice

Abstract

This document describes the use of a scope control BGP community. This well-known advisory transitive community allows an origin AS to specify the extent to which a specific route should be externally propagated. In particular this community, NOPEER, allows an origin AS to specify that a route with this attribute need not be advertised across bilateral peer connections.

1. Introduction

BGP today has a limited number of commonly defined mechanisms that allow a route to be propagated across some subset of the routing system. The NOEXPORT community allows a BGP speaker to specify that redistribution should extend only to the neighbouring AS. Providers

commonly define a number of communities that allow their neighbours to specify how advertised routes should be re-advertised. Current operational practice is that such communities are defined on as AS by AS basis, and while they allow an AS to influence the re-advertisement behaviour of routes passed from a neighbouring AS, they do not allow this scope definition ability to be passed in a transitive fashion to a remote AS.

Advertisement scope specification is of most use in specifying the boundary conditions of route propagation. The specification can take on a number of forms, including as AS transit hop count, a set of target ASs, the presence of a particular route object, or a particular characteristic of the inter-AS connection.

There are a number of motivations for controlling the scope of advertisement of route prefixes, including support of limited transit services where advertisements are restricted to certain transit providers, and various forms of selective transit in a multi-homed environment.

This memo does not attempt to address all such motivations of scope control, and addresses in particular the situation of both multi-homing and traffic engineering. The commonly adopted operational technique is that the originating AS advertises an encompassing aggregate route to all multi-home neighbours, and also selectively advertises a collection of more specific routes. This implements a form of destination-based traffic engineering with some level of fail over protection. The more specific routes typically cease to lever any useful traffic engineering outcome beyond a certain radius of redistribution, and a means of advising that such routes need not to be distributed beyond such a point is of some value in moderating one of the factors of continued route table growth.

Analysis of the BGP routing tables reveals a significant use of the technique of advertising more specific prefixes in addition to advertising a covering aggregate. In an effort to ameliorate some of the effects of this practice, in terms of overall growth of the BGP routing tables in the Internet and the associated burden of global propagation of dynamic changes in the reachability of such more specific address prefixes, this memo describes the use of a transitive BGP route attribute that allows more specific route tables entries to be discarded from the BGP tables under appropriate conditions. Specifically, this attribute, NOPEER, allows a remote AS not to advertise a route object to a neighbour AS when the two AS's are interconnected under the conditions of some form of sender keep all arrangement, as distinct from some form of provider / customer arrangement.

2. NOPEER Attribute

This memo defines the use a new well-known bgp transitive community, NOPEER.

The semantics of this attribute is to allow an AS to interpret the presence of this community as an advisory qualification to readvertisement of a route prefix, permitting an AS not to readvertise the route prefix to all external bilateral peer neighbour AS's. It is consistent with these semantics that an AS may filter received prefixes that are received across a peering session that the receiver regards as a bilateral peer sessions.

3. Motivation

The size of the BGP routing table has been increasing at an accelerating rate since late 1998. At the time of publication of this memo the BGP forwarding table contains over 118,000 entries, and the three year growth rate of this table shows a trend rate which can be correlated to a compound growth rate of no less than 10% per year [2].

One of the aspects of the current BGP routing table is the widespread use of the technique of advertising both an aggregate and a number of more specific address prefixes. For example, the table may contain a routing entry for the prefix 10.0.0.0/23 and also contain entries for the prefixes 10.0.0.0/24 and 10.0.1.0/24. In this example the specific routes fully cover the aggregate announcement. Sparse coverage of aggregates with more specifics is also observed, where, for example, routing entries for 10.0.0.0/8 and 10.0.1.0/24 both exist in the routing table. In total, these more specific route entries occupy some 51% of the routing table, so that more than one half of the routing table does not add additional address reachability information into the routing system, but instead is used to impose a finer level of detail on existing reachability information.

There are a number of motivations for having both an aggregate route and a number of more specific routes in the routing table, including various forms of multi-homed configurations, where there is a requirement to specify a different reachability policy for a part of the advertised address space.

One of the observed common requirements in the multi-homed network configuration is that of undertaking some form of load balancing of incoming traffic across a number of external connections to a number of different neighbouring ASs. If, for example, an AS wishes to use a multi-homed configuration for routing-based load balancing and some

form of mutual fail over between the multiple access connections for incoming traffic, then one approach is for the AS to advertise the same aggregate address prefix to a number of its upstream transit providers, and then advertise a number of more specifics to individual upstream providers. In such a case all of the traffic destined to the more specific address prefixes will be received only over those connections where the more specific has been advertised. If the neighbour BGP peering session of the more specific advertisement fails, the more specific will cease to be announced and incoming traffic will then be passed to the originating network based on the path associated with the advertisement of the encompassing aggregate. In this situation the more specific routes are not automatically subsumed by the presence of the aggregate at any remote AS. Both the aggregate and the associated more specifics are redistributed across the entire external BGP routing domain. In many cases, particularly those associated with desire to undertake traffic engineering and service resilience, the more specific routes are redistributed well beyond the scope where there is any outcomes in terms of traffic differentiation.

To the extent that remote analysis of BGP tables can observe this form of configuration, the number of entries in the BGP forwarding table where more specific entries share a common origin AS with their immediately enclosing aggregates comprise some 20% of the total number of FIB entries. Using a slightly stricter criteria where the AS path of the more specific route matches the immediately enclosing aggregate, the number of more specific routes comprises some 14% of the number of FIB entries.

One protocol mechanism that could be useful in this context is to allow the originator of an advertisement to state some additional qualification on the redistribution of the advertisement, allowing a remote AS to suppress further redistribution under some originator-specified criteria.

The redistribution qualification condition can be specified either by enumeration or by classification. Enumeration would encompass the use of a well-known transitive extended community to specify a list of remote AS's where further redistribution is not advised. The weakness of this approach is that the originating AS would need to constantly revise this enumerated AS list to reflect the changes in inter-AS topology, as, otherwise, the more specific routes would leak beyond the intended redistribution scope. An approach of classification allows an originating AS to specify the conditions where further redistribution is not advised without having to refer to the particular AS's where a match to such conditions are anticipated.

The approach described here to specifying the redistribution boundary

condition is one based on the type of bilateral inter-AS peering. Where one AS can be considered as a customer, and the other AS can be considered as a contracted agent of the customer, or provider, then the relationship is one where the provider, as an agent of the customer, carries the routes and associated policy associated with the routes. Where neither AS can be considered as a customer of the other, then the relationship is one of bilateral peering, and neither AS can be considered as an agent of the other in redistributing policies associated with routes. This latter arrangement is commonly referred to as a "sender keep all peer" relationship, or "peering". This peer boundary can be regarded as a logical point where the redistribution of additional reachability policy imposed by the origin AS on a route is no longer an imposed requirement.

This approach allows an originator of a prefix to attach a commonly defined policy to a route prefix, indicate that a route should be re-advertised conditionally, based on the characteristics of the inter-AS connection.

4. IANA considerations

The IANA should register NOPEER as a well-known community, as defined in [1], as having global significance.

NOPEER (0xFFFFFF04)

This is an advisory qualification to readvertisement of a route prefix, permitting an AS not to readvertise the route prefix to all external bilateral peer neighbour AS's. It is consistent with these semantics that an AS may filter received prefixes that are received across a peering session that the receiver regards as a bilateral peer sessions

5. Security considerations

BGP is an instance of a relaying protocol, where route information is received, processed and forwarded. BGP contains no specific mechanisms to prevent the unauthorized modification of the information by a forwarding agent, allowing routing information to be modified, deleted or false information to be inserted without the knowledge of the originator of the routing information or any of the recipients.

The NOPEER community does not alter this overall situation concerning the integrity of BGP as a routing system.

Use of the NOPEER community has the capability to introduce additional attack mechanisms into BGP by allowing the potential for

man-in-the-middle, session-hijacking, or denial of service attacks for an address prefix range being launched by a remote AS.

Unauthorized addition of this community to a route prefix by a transit provider where there is no covering aggregate route prefix may cause a denial of service attack based on denial of reachability to the prefix. Even in the case that there is a covering aggregate, if the more specific route has a different origin AS than the aggregate, the addition of this community by a transit AS may cause a denial of service attack on the origin AS of the more specific prefix.

BGP is already vulnerable to a denial of service attack based on the injection of false routing information. It is possible to use this community to limit the redistribution of a false route entry such that its visibility can be limited and detection and rectification of the problem can be more difficult under the circumstances of limited redistribution.

References

Normative References:

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

INformative References:

[2] Huston, G., "Commentary on Inter-Domain Routing in the Internet", RFC 3221, December 2001.

Author's Address

Geoff Huston
Telstra

Intellectual Property Statement

Full Copyright Statement

Acknowledgement

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