The ISP Column
An occasional column on things Internet

October 2005

4-Byte AS Numbers

Geoff Huston

In the previous articles on the topic of AS numbers we've looked at the aspects of the use of AS numbers in the Internet's inter-domain routing space, and then looked at the range of projections that are derived from the existing AS number consumption data. The most likely consumption trend that is derived from this data is a model of an increasing AS number consumption rate in the coming years.

This model has a best fit projection that forecasts exhaustion of the unallocated pool of AS numbers in late 2010. In working backward from this date to the necessary steps to facilitate the associated industry transition, it would appear that the time to commence the transition is in the coming months rather than in the coming years. To recap the steps involved in this transition, we appear to be looking at a sequence of actions that includes:

- the completion of the relevant protocol standards for a larger AS number field in BGP,
- the production of code in available implementations of BGP that support this protocol standard,
- various forms of testing of this code, both in terms of its correct operation and interoperability, and in terms of the correctness and viability of the relevant transition steps,
- developing the necessary infrastructural support system to manage the distribution of this new number pool, and
- a process of deployment of this protocol so that the deployment of larger AS numbers can commence well before the point at which the existing AS number pool is exhausted.

One of the very first steps is to look at what is being proposed to address this forecast exhaustion is standardize the use of a larger AS number pool within the BGP protocol, and also understand the implications of an associated transition plan.

In this article we'll look at the current proposal for a larger AS number pool.

> As of July 2005 the document defining this proposal is an IETF Internet Draft. The most recent version of the draft at the time of writing of this article was draft-ietf-idr-as4bytes-10.txt.

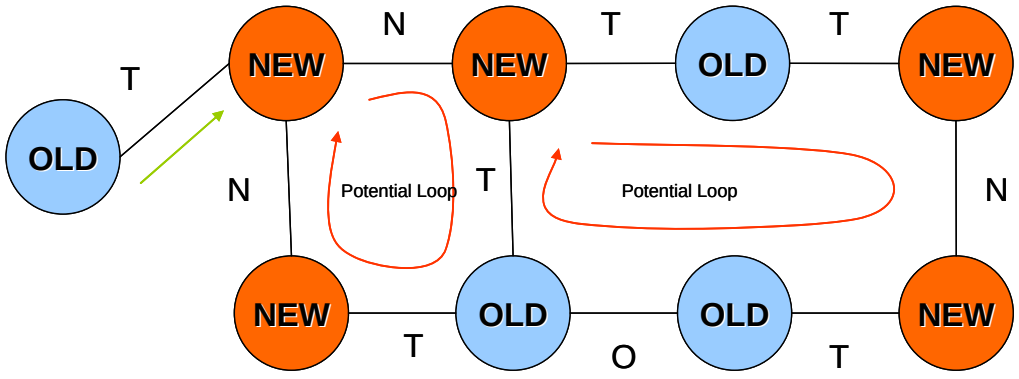
The approach proposed in this work-in-progress document is to expand the size of the AS number pool space from 16 bits to 32 bits. In number terms this expands the number space from a pool of

65,536 numbers to 4,294,967,296 billion numbers. In terms of the current use of AS numbers, the current scaling properties of the BGP routing protocol, and the use of AS's in the context of inter-domain routing, a pool of 4.4 billion numbers would easily encompass a network environment of significantly greater levels of domains, and inter-domain interconnection density. Such a pool size would exceed some current guesses of the scaling capabilities of the BGP protocol by up to a further two orders of magnitude.

Its also proposed to preserve the first block of 4-Byte AS Numbers to align with the allocations of the 2-Byte numbers.

> Let's use a new form of terminology here for 4-Byte AS number values, where the first 65,536 AS numbers are numbers use the form "0:0" through to "0:65535". The second set of 65536 numbers would be written as 1:0 through to 1:65535, and so on. So we'll be using a number format of <upper16 bits>:<lower 16 bits>.

What is the inventory of issues that need to specifically addressed here? Obviously there is a need for some changes to the routing protocol, and this change needs to be able to accommodate a number of situations. Firstly the assumption is that each domain (or Autonomous System will indeed undertake a "flag day" and transition all its BGP speakers to support 4-byte AS Numbers in a coordinated fashion. But beyond this per-domain transition an ordered inter-domain transition will be unrealistic to expect. More reasonable on the expectation scale is the piecemeal transition of domains, where individual domains will shift to supporting 4-byte AS's in their own time. Domains that are currently using 2-byte AS's may have less reason to undergo an early transition to 4-byte AS support, while those domains who are assigned a non-mappable 4-byte AS number will find that they have to support 4-byte AS numbers from the outset. The following diagram hopefully illustrates most of the potential issues associated with this kind of piecemeal transition. Her "OLD" refers to 2-byte AS number domains and "NEW" refers to 4-byte AS number domains.



T – Transitional BGP session NEW / OLD boundary
O – OLD BGP peering session
N – NEW BGP peering session

Figure 1 – BGP Transition Cases

Figure 1 describes an original prefix injection from an OLD BGP originating AS and a number of transitions from OLD to NEW BGP AS's and NEW to OLD BGP AS's, as well as OLD and NEW BGP peerings. The diagram also contains a number of potential loops, encompassing both OLD and NEW transits. Any proposed solution should be able to loop detect in this scenario without having to alter the behaviour of the OLD BGP speakers.

Changes to the BGP protocol

BGP has two major parts within its protocol: opening up a BGP conversation with a peer BGP speaker, and then transfer of protocol objects that describe reachability of address prefixes and associated attributes of these address prefixes. Both parts include AS Number components, and in considering changes to the current protocol, both parts of the protocol require some change. The message objects that need to be considered here are therefore the BGP OPEN message and the BGP UPDATE message.

The changes to the BGP protocol create a "NEW" BGP implementation that is capable of supporting a 4-byte AS number environment. The essential task of the changes is to define mechanisms that all NEW BGP speakers to speak to each other and pass all AS number values in 4 byte fields. However the Internet is way too large to set up a "flag day" at which point the entire collection of BGP speakers will undertake a switch from "OLD" BGP to NEW BGP. Accordingly, its also necessary to define protocol interactions in NEW BGP where the transition in the Internet will be gradual and essentially uncoordinated. NEW BGP speakers will have to set up sessions with OLD BGP speakers, and of course OLD BGP speakers will also be peering with other OLD BGP speakers. The information associated with 4-Byte AS paths must be passed across sections of the network that normally support only 2-Byte AS Paths. In other words 4-Byte AS information needs to be passed to OLD BGP speakers and between OLD BGP speakers.

Opening a BGP session

BGP carries its own AS number in the "My Autonomous System" field of the BGP OPEN message.

The proposed approach is to initiate a NEW BGP session in a mode that is compatible with the OLD BGP protocol, and also inform the remote peer of its capability to conduct a NEW BGP conversation if the remote peer is also a NEW BGP speaker. This capability advertisement is part of OLD BGP, and OLD BGP speakers who open a peer session with a NEW BGP speaker will simply ignore the NEW capability and operate in OLD mode. A NEW BGP peer will respond positively to the NEW capability, and the BGP session can then operate in NEW mode.

The BGP OPEN message includes a fixed length 2-byte "My AS field" (as shown in Figure 2) as well as potentially containing a capability query as part of the Optional Parameters section. In order to ensure that NEW and OLD speakers can communicate, then this 2 byte My AS field needs to be preserved in NEW BGP even when the Optional Parameters section encompasses the capability to undertake a NEW peering session. This may appear contradictory in the first instance, as the OPEN message then contains both a 2-byte Autonomous System number and a 4-byte AS Capabilities Query.

The mechanism proposed for the OPEN Message varies according to whether the NEW speaker is using a mappable AS number drawn from the original pool (i.e. with a My AS number in the range 0:0 through to 0:65535), or its using a number drawn from a higher-numbered 4-byte number

block. In the first case the OPEN message would use the 2-byte mapped value in the My AS field (dropping out the zero-valued high order 16 bits of the AS value), while in the second case the BGP speaker would use for My AS a special 2-byte value that is reserved for this purpose (AS 23456). In both cases the Optional Parameter section would include a capability code to indicate that the local BGP speaker can support 4-byte AS Numbers (Capability Code 65).

The side effect is that in the OLD BGP domains AS 23456 may appear to be connected to the 2-byte BGP realm in many different locations, and advertising a collection of different address prefixes in different locations. From the OLD BGP realm this does not present a protocol problem, although, as always, there is the potential here that this repeated use of AS 23456 as a 4-byte AS substitution token may create a somewhat confusing BGP-view of the Internet from the perspective of the OLD BGP world!

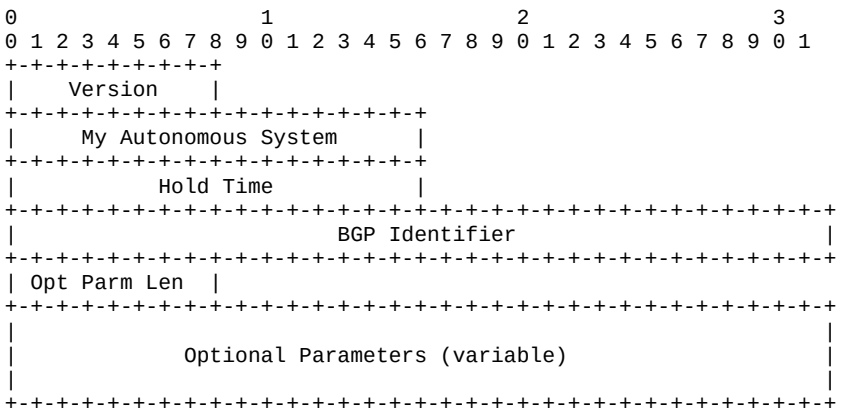


Figure 2- BGP Open Protocol Message – From "draft-ietf-idr-bgp4-26.txt"

The capability exchange uses a protocol described in RFC3392. The NEW BGP speaker adds an optional capability field to the OPEN message. The 4-byte AS capability code 65 carries as its capability value the local 4-byte local AS number value. For a NEW peer this capability value is to be interpreted as the actual AS of the remote side, on the basis that the My AS field in the body of the OPEN is either a truncation of the local 4-byte AS value (in the case of mappable 4-byte AS values), or the special value of AS 23456.

One response from the remote BGP speaker is to accept the capabilities announcement with a comparable OPEN message, in which case the remote side is also a NEW BGP speaker, and the session may proceed using 4-byte AS values.

If the session is being opened with an OLD BGP peer, the OLD BGP peer may respond with a NOTIFICATION message indicating that the 4-byte capability is an Unsupported Optional Capability parameter. In response to this unsupported notification the NEW BGP speaker will re-establish the connection by resending the OPEN message, and this time drop the 4-byte capability option from the message. The NEW BGP speaker will then assume that it is peering with an OLD BGP peer.

The "Unsupported" response to a capabilities parameter was not included in the original specification. Older versions of BGP allowed a BGP speaker to optionally send a NOTIFICATION message and terminate the peer session. If the NEW BGP speaker sees a session termination in response to its OPEN message it may need to re-open the TCP session, and this time omitting the

4-byte capability advertisement in the initial BGP OPEN message. Once again, the NEW BGP speaker will then assume that it is peering with an OLD BGP peer.

In general, however, a BGP implementation should not send a NOTIFICATION when a capability parameter is unrecognized because the Capabilities Optional Parameter is still optional. With such general implementations, the OLD speaker would just pick up the 2-byte AS (23456) in the OPEN received form the NEW speaker. As the OLD speaker does not advertise the 4-byte AS Capability in its OPEN, the NEW speaker has to use the 2-byte AS it advertised in the OPEN (that is, the AS_TRAN - 23456) for peering. A NOTIFICATION is not involved in this scenario.

The BGP UPDATE Message

For a NEW BGP session (4-byte peering with 4-byte) the changes to the protocol are the use of 4-byte AS numbers in the AS_PATH attribute of UPDATE messages. All 2-byte AS values are padded with a zero high order 16 bits. If the AGGREGATOR attribute is used it is similarly carried as a 4-byte value. So in the 4-byte peering, all 2-byte information is carried in mapped 4-byte AS numbers.

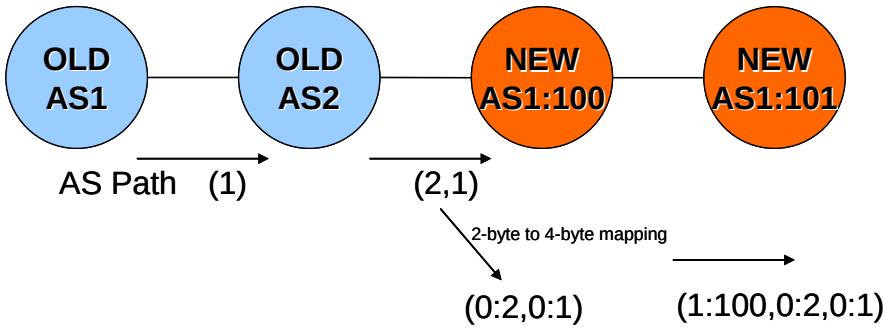


Figure 3 – OLD to NEW BGP AS Path Mapping

In this way AS Path length is preserved without change when translating 2-byte AS information into the 4-byte domain.

The next case is where an OLD BGP peers with a NEW BGP. We've already seen the simple case where the information is coming from a 2-byte path and there is no additional 4-byte information, and in this case the 2-byte values are simply mapped into 4-byte values. What about the reverse case where 4-byte information is being passed back into the 2-byte world?

There are two parts to this case: firstly creating an equivalent 2-byte AS Path and secondly packing up the 4-byte AS Path information in such a way that it transits across the 2-byte domain in such a way that it can be reassembled in any subsequent transition into a 4-byte domain. In the first case the equivalent path information is constructed by either stripping off the high order 2-bytes off the AS value, as long as this part is all zeros. Where this is not possible the transition AS number, 23456, is substituted in its place. This is indicated in Figure 4.

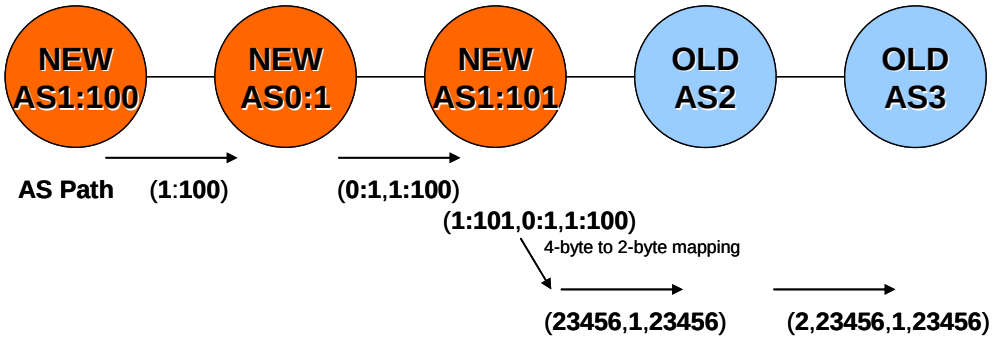


Figure 4 – NEW to OLD BGP AS Path Mapping

In this way the AS Path length metric is preserved, and the prevention of count-to-infinity loops in the 2-byte domain is avoided.

The second part to this case is packaging up the 4-byte path into the OLD BGP session in such a way that it can be unpacked at any subsequent boundary into a 4-byte realm. Here the proposal calls for new transitive community attributes to be supported for OLD BGP. These attributes are defined as transitive attributes, and should be passed through the OLD BGP peering sessions without alteration. It should be noted that this is not a protocol change per se, but it does require the explicit support within OLD BGP implementations of this attribute as a transitive community. The proposed mechanism is an extended community attribute called "NEW_AS_PATH". When a NEW BGP speaker is speaking to an OLD BGP, the NEW BGP prepends its own AS value to the AS PATH and copies this information into the NEW_AS_PATH. It then translates the 4-byte AS Path into a 2-byte equivalent AS Path. The translation is straightforward, in that where the 4-byte AS has all zeros in the high order 2 bytes, the translation truncates the AS value to a 2-byte value, and where the high order 2-bytes are non-zero the translation substitutes the reserved 2-byte value AS 23456 in its place.

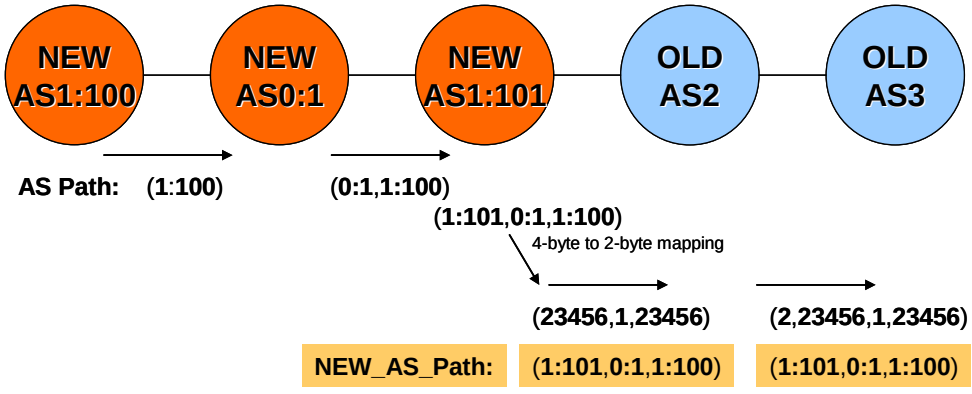


Figure 5 – NEW to OLD BGP AS Path Mapping

The transit across the OLD BGP domains leaves the NEW_AS_PATH untouched, and prepends 2-byte AS values to the AS_PATH.

The next transition is one from the OLD to the NEW domain, as shown by a continuation of the previous example

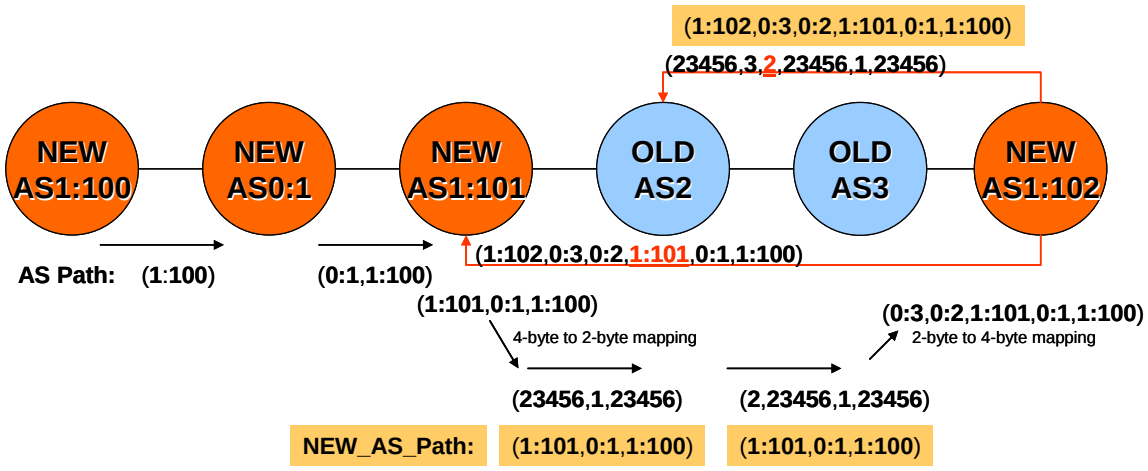


Figure 6 – NEW-> OLD-> NEW Transition with Potential Routing Loops

The Figure 6 there is a further OLD to NEW transition. In this case the NEW BGP speaker takes the AS Path as presented by the OLD BGP speaker and converts the 2-byte values to 4-byte values by adding 2-bytes of zero padding to each entry, and then overwrites the trailing entries with the values specified by the NEW_AS_PATH attribute. The net result is that the 4-byte path that entered the 2-byte sequence is prepended with the 2-byte transit AS sequence. The NEW_AS_PATH is then removed, leaving an intact 4-byte path as the AS_PATH attribute.

This ensures that the resultant BGP environment can detect loops in both the NEW 4-byte and OLD 2-byte realms.

Further extending this example, we can construct a potential loop in the 4-byte world by adding a path back to AS 1:101. The restoration of the original 4-byte AS Path at the OLD-to-NEW transition ensures that the potential loop is discarded even when the loop needs to traverse one or more 2-byte OLD BGP AS's. A similar form of loop can be constructed for a 2-byte OLD BGP AS, that traverses a 4-byte NEW BGP AS. Again the transition mapping ensures that the potential routing loop is detected by BGP.

The ability to perform AS Path Prepending is also unaltered in this mixed NEW and OLD BGP environment, The AS simply prepends its local AS value to the AS_PATH as normal. In the case of prepending on a a NEW-to-OLD boundary the prepended AS Path is mapped into the NEW_AS_Path attribute as described above.

The earlier article of use of AS numbers in routing also noted the less common use of AS PATH poisoning, where the prepending uses a different AS number value in order to ensure that the particular advertisement is not learned by a remote AS. For NEW BGP speakers there is no change to this capability. For OLD BGP speakers the AS Path Poisoning can only be directed towards 2-byte AS's, as the OLD BGP speaker has no knowledge of the structure or content of the NEW AS PATH attribute. From the perspective of the OLD BGP speaker, the NEW_AS_PATH attribute is an opaque data block.

The same translation technique applies to the AGGREGATOR attribute. In a NEW-to-OLD transition the AGGREGATOR may be a mappable AS number, in which case the value is truncated to 2-bytes and no further action is required. Otherwise the 4-byte AGGREGATOR value is rewritten to the NEW_AGGREGATOR attribute and the transition 2-byte value, AS 2356 is placed into the AGGREGATOR attribute. On an OLD-to-NEW transition the NEW_AGGREGATOR attribute is copied

back int the AGGREGATOR attribute, if defined, otherwise the AGGREGATOR is padded out with leading zeros.

The general approach adopted for transition is preserve AS Path length information across the OLD and NEW BGP boundaries, while recognising that some 4-byte AS information cannot be cleanly mapped into a 2-byte AS Path. In order to preserve 4-byte information, which is necessary to prevent loop formation for 4-byte AS's, the 4-byte information is preserved across OLD transit paths and restored upon rentry into NEW BGP realms.

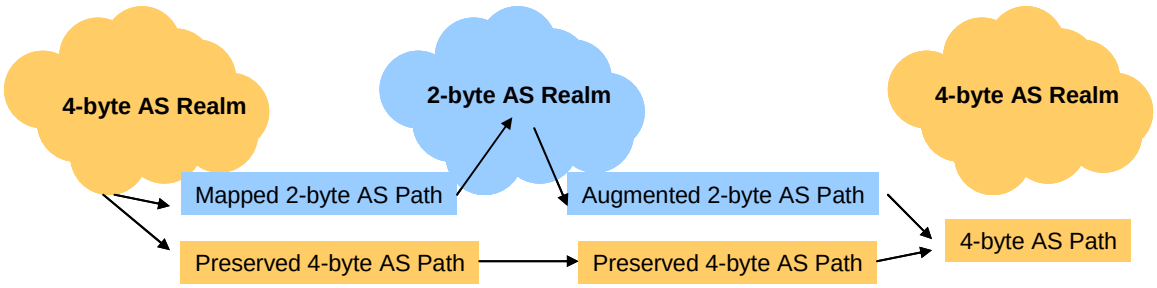


Figure 7 – 2-byte and 4-byte AS Realms

BGP Communities

BGP communities require some additional consideration. If the high order 16 bits of the community attribute are neither all zeros or all ones, then it is assumed to contain a 2-byte AS value. Where it is necessary to specify a 4-byte AS number in the community attribute it is necessary to turn to the extended community attribute to support this [This extended communities feature is documented in the Internet draft: draft-ramachandra-bgp-ext-communities-10.txt]

Transition

Transition in this environment is relatively straightforward. NEW BGP speakers can be deployed within the network in a piecemeal fashion without any major concerns. The size of BGP UPDATE messages is slightly longer due to the extended length of the AS PATH attribute in NEW BGP and the NEW_AS_PATH attribute that has been added in the OLD BGP environment, but it should not prove to be a major factor.

BGP loop prevention appears to be adequately addressed in all commonly encountered situations, and there appear to be no other significant transition considerations.

There does appear to be one precondition for the use of 4-byte AS numbers, and that is for a routing domain to actually be numbered with a non-mappable 4-byte AS number, all the BGP speakers in the domain should be NEW BGP speakers. Aside from that consideration there do not appear to be any further constraints associated with this transition.

Comments

It is certainly a confronting task to contemplate an environment when we would exhaust a 4-byte AS number space, but i suppose that same consideration was in the minds of the original BGP protocol designers when they opted to use 2-byte AS numbers. Of course a 32 bit number pool is not double the pool size of a 16 bit number – its 65,536 times larger. That does appear to lead one to believe that this time it will be far more challenging to exhaust this number pool.

It should be considered that this approach appears to offer a path of minimal disruption and minimal change in terms of operational configuration, storage, message size and processing overheads for BGP. Nothing much has changed here except the range of the number space, and some ancillary considerations relating to transitional arrangements.

Other labelling spaces remain possibilities, and could well use the same transitional approach. There is no significance whatsoever in the AS number apart from its uniqueness, and any other form of namespace would function equally well in terms of its role in BGP. One could use domain names, URIs, fixed length hashes of public keys, the public keys themselves, or even IPv6 addresses as a distinguishing AS identifier. There is no direct requirement for summarization of AS number ranges within the protocol use, and there is no direct requirement within the protocol to continue to use number identifiers, and no direct requirement to stick with values that are encoded in a find length field.

However, such approaches would add to the size of BGP UPDATE messages, increase the storage requirements, and, perhaps marginally, increase processing overheads for BGP. The more complex the identity space (and here I'm thinking about the use of public keys as AS identifiers) the more complex the basic task of BGP configuration and the higher the possibility of mistakes. Borrowing from another namespace, such as domain names, or derived URIs has the associated issue that the uniqueness of the space is derived from the inherent stability and uniqueness of the name space upon which the identifiers are derived. It's a definite possibility that at times this trust is misplaced.

Numbers are certainly simple neutral identifiers. The decision to simply extend the number pool space appears to be another instance of a design trade-off between the size of the number space and the additional BGP overheads and ease of transition. In sticking with numeric labels this approach represents minimal change to the installed base of BGP speakers, and there is no requirement for an existing routing domain using a 2-byte AS number and OLD BGP to make any changes to its routing environment at all.

The 4-byte transition appears to offer flexibility, orderly transition and minimal disruptions to existing operational practices.

However, it should be remembered that we are running out of the 2-byte AS number pool, and an industry of this size needs to have long periods of advance warning of change in order to be able to integrate such changes in operational cycles of testing and deployment. The first steps that need to happen are the completion of this approach in the form of an IETF Standard and the production of NEW BGP implementations from the existing BGP implementations. All all we have to do is get on with the job!

Acknowledgement

Thanks to Enke Chen, one of the authors of the 4-Byte AS working document, for some clarification regarding the OPEN behaviour between OLD and NEW BGP implementations

Disclaimer

If were are any views expressed in this article, they do not necessarily represent the views or positions of the Asia Pacific Network Information Centre, nor those of the Internet Society.

About the Author

GEOFF HUSTON holds a B.Sc. and a M.Sc. from the Australian National University. He has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. He is author of a number of Internet-related books, and is currently the Senior Internet Researcher at APNIC, the Regional Internet Registry serving the Asia Pacific region. He has served as a Trustee of the Internet Society, and also as a member of the Internet Architecture Board of the Internet Engineering Task Force.