Vadim Antonov
Sprint
March 1995

BGP AS Path Metrics

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard. Distribution of this memo is unlimited.

This document is an Internet Draft. Internet Drafts are working documents of the Internet Engineering Task Force (IETF), its Areas, and its Working Groups. Note that other groups may also distribute working documents as Internet Drafts.

Internet Drafts are draft documents valid for a maximum of six months. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "work in progress".

Abstract

In this memo, the extended (per-hop) BGP metrics are proposed to allow network operators to implement routing policies which are common cases in the practice but require complicated ad-hoc configurations given the virtual absence of globally significant metrics in BGP-4.

BGP is an inter-autonomous system routing protocol designed for TCP/IP internets. The current version is BGP-4 documented in RFC 1654, versions 1-3 are documented in RFCs 1105, 1163 and 1267 accordingly.

This document is a supplement to RFC 1654, and describes the new features in terms of differences from BGP-4.

1. Current State

BGP version 4 has three distinct metrics:

(1) LOCAL_PREF, providing a way for local operator to override the standard best path discrimination mechanism; local preferences are not distributed outside of autonomous systems;

(2) MULTI_EXIT_DISC, which can be used as tie-breaking mechanism to select between otherwise identical paths; multi-exit discriminators are meaningful on per-peer basis, and are not propagated between AS-es which are not directly connected;

(3) Length of AS-path (number of components in an AS-path).

Although the length of AS-path is not explicitly stated to be a global metric it in fact is, and used as such. Since there is no way to attach administrative distances to the AS-path length two crude workarounds were devised, first being insertion of physical routers to change the length of AS-path, and second is AS-path modification knobs in some BGP implementations to achieve the same effect (i.e. to make an AS path to appear longer) without adding "real" autonomous systems in between.

Although it solved the original problem (i.e. to provide globally-significant administrative distances) to some extent, the BGP path selection process remains baroque and confusing, particularly in case if "iBGP Hack" is used to propagate the exterior routing information across a transit autonomous system. The default best path selection algorithm is (each following step is used if previous steps resulted in ties):

- first, local preferences are compared, and the path with higher local preference is selected;

- exterior BGP routes are preferred to interior;

- routes with shorter AS path are preferred;

- routes with smaller MED are preferred;

- routes with smaller origin code are preferred;

- routes with smaller interior metric are preferred;

- routes from router with a largest ID are preferred.

Needless to say that as a result there are routers in the Internet with routing policies no human being is able to understand completely.

A much worse effect is that even trivial backup arrangements like:

```
     B --------- C
    /             \
   A = X = Y = Z = D
```

(where `-` is the slow path, `=` is the fast path, and `A` wants all other AS-es to prefer the path through `X` even when there are shorter as-paths available) require extensive co-ordination between operators of all AS-es in the path, to override the default path selection. The task would be much easier if `A` could simply announce path through `B` with a much higher metric. This can be done (sort of) by inserting pseudo-ASes but with real-life spaghetti networks it is real hard to make sure the length of the path will be sufficient.

However, there is a clear consensus that a single metric is not sufficient for systems with no central administration (that was the reason why there is no such metric in BGP in the first place).

2. Per-Hop Metrics

The proposed solution is to provide a vector of metrics (one per element of AS path) $\{m_i\}$ instead of a single metric, and allow network administrators to specify the vector of weights $\{w_i\}$ to calculate the single metric as a weighted sum

$$m = \sum_{i=1}^{n} w_i m_i .$$

where n is the number of components in AS path of the route.

A per-exit local preference, generated by local border gateways m_0 and the appropriately scaled IGP metric can be added to the summary path metric.

The rationale for using weighted sum instead of plain metric is to provide users of routing information with some degree of control on how to interpret the metrics generated by sources or transit autonomous systems they cannot control. At the same time, the original metric information remains unmodified. It is a valid argument that this approach provides too much flexibility, allowing creation of totally insane routing policies; however, it also provides the equally strong defense.

The BGP-4 multi-exit discriminators are often generated from the interior metrics, allowing to request the remote AS to modify its behavior in response to the local routing conditions (it is also very handy for establishing nearly-symmetrical routing in case of autonomous system connected at several geographically diverse interconnection points). To simulate that behavior, the metrics in the as-path could be derived from the appropriately scaled interior metrics. However, the "dynamic" (or IGP-derived) metrics tend to generate more routing updates than static administrative metrics. In the marginal case, all exterior metrics are derived from BGP metrics to the effect that the entire Internet has one common IGP which obviously provides for nearly-optimal routing but not less obviously does not scale.

Therefore, the distribution of the IGP-derived metrics should be limited; but associated administrative metrics should be preserved. To achieve that the current proposal assumes that each per-hop metric is really a product of the static part s, and the dynamic part d, with attached indicator of scope of the dynamic part (an unsigned integer number). The scope indicators should be set by the border gateways adding elements to the AS paths as the length of AS path plus the number of additional AS-hops during which the dynamic indicators should be preserved. When the length of AS path reaches the value of the scope indicator the dynamic part of the metric *expires* and its value is assumed to be 1.0.

The dynamic part of a metric biases the static administrative value and is generally derived from the interior metric(s) using some scaling function. Generally, the dynamic part should have the value close to 1.0 for transit routes, so expiration of dynamic part won't result in a drastic change in the composite metric value.

The resulting formula for the scalar (composite) metric is:

$$m = (c + \varepsilon)M + w_0 m_0 + \sum_{i=1}^{n} w_i d_i s_i$$

where M is the local (interior) metric, c is its scaling co-efficient and ε is a very small number, allowing tie-breaking but not allowing interior metric to change the result of comparison of the exterior metric.

To ensure consistency, all gateways in AS should have the same vector of weights. The default value of w_i is 1.0 and default value of c is 0.0.

To provide further flexibility implementations can make selection of the weight vector to be conditional on AS path matching a pre-defined pattern. The sets of weight vectors and patterns should be consistent in all routers in the autonomous system.

3. Changes In Border Gateway Protocol

The changes only affect some path attributes. Two new attributes are added to carry information on static and dynamic components of the path metric.

3.1. MULTI_EXIT_DISC

BGP Path Metrics makes the MULTI_EXIT_DISC attribute obsolete. Gateways generating path metrics should *not* generate this attribute.

If a path learned from a neighbor AS has MULTI_EXIT_DISC attribute, that value attribute should be preserved when the path is announced to the other local (interior) systems.

If the presence of the multi-exit discriminator with at least one path to the destination, the constants c and ε in the formulae for the scalar metric are assumed to be zero, and the multi-exit discriminator is used to resolve ties. If MEDs are identical, then the interior metric is used to pick the best route.

If one or more path has multi-exit discriminator, and other paths do not, paths without MED are assumed to have MED equal to zero.

3.2. LOCAL_PREF

The local preference value is now interpreted as m_0 for the purpose of calculation of the local metric. The four-octet integer value of the attribute is interpreted as fixed-point unsigned number with 16 bits in fractional, and 16 bits in integer part.

If LOCAL_PREF attribute is not present, it is assumed to have value zero.

3.3. AS_PATH_METRICS

This is the new optional transitive attribute with Type Code *X* (to be determined, code 8 is suggested) which contains the ordered sequence of 4 or 6-octet metric elements. This attribute is used to carry the vector of AS path metric information as described above.

Metric elements have the following format:

```
 0             7 8            15
+---------------+
| Dynamic Scope |
+---------------+
|     Index     |
+---------------+---------------+
|          Static Part          |
+---------------+---------------+
|    Dynamic Part (optional)    |
+---------------+---------------+
```

Dynamic Scope

This 1-octet unsigned field indicates the maximal length of AS path at which the dynamic part of the metric should still be taken into account. The length of AS path is defined as the total number of 2-octet AS numbers in the AS_SEQUENCE components of AS_PATH attribute plus the total number of AS_SET components. If *Dynamic Scope* has the value zero, the last two octets (the *Dynamic Part*) of the metric element are omitted.

Index

This 1-octet unsigned field points to the component of AS path associated with the metric element (a component of AS path is a single element of AS_SEQUENCE or the entire AS_SET). The index of the oldest component in AS path is zero, and every newer element has index increased by one. The maximal value of *Index* should not be greater than AS path length minus one.

To simplify the processing of AS path metric information elements of the AS_PATH_METRICS attribute must be ordered by their *Indices* in ascending order. There should be no metric elements with identical values of *Index* within the same attribute.

Static Part

This 2-octet field contains the representation of the static part s_i of the metric as an unsigned fixed-point number with 8 bits in integer and 8 bits in fractional parts.

Dynamic Part

This 2-octet optional field contains the representation of the dynamic part d_i of the metric as an unsigned fixed-point number with 8 bits in integer and 8 bits in fractional parts.

This field is omitted if *Dynamic Scope* is zero, thus reducing the size of the metric element down to 4 octets.

When a BGP speaker propagates a route which it has learned from another speaker's UPDATE message, it should modify the route's AS_PATH_MERTICS attribute accordingly to the following procedure:

(1) Scan all metric elements to find elements with non-zero *Dynamic Scope* which is less or equal the length of AS path *prior* to the addition of the speaker's own AS. If the value of *Static Part* of such element equals 1.0 (0x100), the element can be removed completely, otherwise it is truncated to 4 octets.

A BGP speaker can restrict the propagation of dynamic metric information by decreasing the *Dynamic Scope* fields accordingly to the administrative restrictions.

(2) If the given BGP speaker announces a route to another BGP speaker in its own autonomous system, the following step should be skipped.

(3) If the speaker wants to add the metric with a dynamic or static part not equal 1.0 (0x100), it should append the new metric element to the AS_PATH_METRICS attribute (if there was no attribute in the route, the attribute with the single element should be created).

If the dynamic part is equal 1.0, a 4-byte metric element should be created (*Dynamic Scope* in this case should be zero), otherwise 6-byte element is created with *Dynamic Scope* equal to the length of AS path *prior* to the addition of the speaker's own AS plus the administratively configured number of AS hops the dynamic metric should survive.

The *Index* field of the new element should be set to be equal to the length of AS path *prior* to the addition of the speaker's own AS.

(4) If the resulting value of AS_PATH_METRIC attribute has zero length, the attribute can be removed completely.

When a BGP speaker originates the route the same steps as above should be taken, as in case if AS_PATH_METRICS attribute is absent.

If more specific routes $\{r\}$ are suppressed by a BGP speaker performing proxy-aggregation, the value of AS_PATH_METRICS attribute should be replaced with a single element with static part of the metric equal to

$$s_1 = \min_r \sum_{i=1}^{n_r} s_{r,i}.$$

The dynamic part of such aggregate's metric should be set to 1.0.

4. Scalar Metric Computation

All computations related to finding the scalar metric (except the computation of $(c + \varepsilon)M$ which depends on the nature of the interior metric) should be performed on 4-octet unsigned fixed-point numbers, with 16 bits in fractional and 16 bits in integer part. Conversion of 2-octet fixed-point numbers to 4-octet fixed-point numbers should be performed by an operation equivalent to prepending and appending all-zero octets.

It is recommended to keep weight vectors $\{w_i\}$ as sequences of 2-octet fixed-point numbers. The representation of weight vectors shall allow generation of variable-length vectors from the fixed-length prefix, suffix and a single infix weight value repeated the required number of times. If the AS path is short enough, the prefix values (i.e. weights for the closest components of AS path) take precedence over the overlapping suffix values, and the infix value is not used. If AS path is shorter than prefix the prefix is truncated, so only the weights for the closest components are used.

References

(1) Lougheed, K.; Rekhter, Y. *Border Gateway Protocol (BGP)*, RFC 1105, 1989 June;

(2) Lougheed, K.; Rekhter, Y. *Border Gateway Protocol (BGP)*, RFC 1163, 1990 June;

(3) Rekhter, Y.,ed., *BGP protocol analysis*, RFC 1265, 1991 October;

(4) Rekhter, Y.,ed., Experience with the BGP protocol, RFC 1266, 1991 October;

(5) Lougheed, K.; Rekhter, Y., *A Border Gateway Protocol 3 (BGP-3)*, RFC 1267, 1991 October;

(6) Rekhter, Y.; Li, T.,eds., *A Border Gateway Protocol 4 (BGP-4)*, RFC 1654, 1994 July;

(7) Rekhter, Y.; Gross, P.,eds., *Application of the Border Gateway Protocol in the Internet*, RFC 1655, 1994 July;

(8) Traina, P., *BGP-4 Protocol Document Roadmap and Implementation Experience*, RFC 1656, 1994 July.