IDR
Internet Draft
Intended status: Standards Track
Expires: 09 December 2025

C. Lin
New H3C Technologies
H. Yao
China Mobile

June 7, 2025

Distribute Service Metric by BGP
draft-lin-idr-distribute-service-metric-05

Abstract

When calculating the path selection for service traffic, it is important to consider not only network metrics, but also the impact of service Metric. Therefore, it is necessary to transmit service Metric information from the service site to the user access site, in order to facilitate path selection for service traffic at the access router.

This document describes an approach for using the BGP Control Plane to steer traffic based on a set of metrics that reflect the underlying network conditions and other service-specific state collected from available service locations.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 09, 2025.

Copyright Notice

Table of Contents

1. Introduction

In scenarios such as edge services and Computing-Aware Traffic Steering (CATS) services, service instances are deployed across multiple geographically distributed sites to achieve better response times.

When selecting a path for service traffic, it is important to consider not only network metrics but also the operational status of the service, which includes CPU utilization, service queue length, memory usage, and other factors. These operational statuses of the service are abstracted as service metrics, allowing service requests to be directed to the optimal service instance based on both network metrics and service metrics.

Due to the rapid changes in service operational status, it is necessary for the service site to frequently send update messages regarding its operational status to the user side. Typically, the update frequency ranges from 1 to 5 minutes.

In scenarios with a large number of services, frequent updates of service metrics for each service instance can consume a significant amount of network bandwidth.

Since BGP has rich routing strategies that can adapt to the diversity and variability of services, and has a route reflector (RR) function that can reduce the session connections for service metric distribution, this document chooses BGP as the control plane of the networks that support service metric distribution. This document describes a service metric distribution framework based on BGP, which is designed to support the automatic discovery, start notification, and updating of service metrics.

1.1. Conventions and Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

This document uses the following terms from [draft-ietf-cats-framework]:

Computing-Aware Traffic Steering (CATS)

CATS Service ID (CS-ID)

CATS Instance Selector ID (CIS-ID)

Client

Service

Ingress CATS-Forwarder

Egress CATS-Forwarder

Service site

Additionally, this document uses the following terms from [RFC4364]:

Route Target (RT) Attribute

Route Distinguisher (RD)

Virtual Private Networks (VPN)

This document introduces the following terms:

Service Metric Routing: Routing based on Service Metric.

Discoverer Egress CATS-Forwarder which connected to the service site.

Originator Ingress CATS-Forwarder which connected to the client.

1.2. Gap

The process of Service Metric routing involves Egress CATS-Forwarder collecting service metric information and notifying it to Ingress CATS-Forwarder. When Ingress CATS-Forwarder needs to forward service traffic, it selects the optimal path for forwarding based on the network metric and service metric information.

Due to the frequent changes in service metrics, the Egress CATS-Forwarder needs to periodically notify the Ingress CATS-Forwarder of updates to the service metrics.

In the current implementation, BGP uses existing address families to distribute service metric routes. Consequently, if there are nodes that do not need to consider Service Metric Routing, additional filtering methods are required to avoid potential impacts on the efficiency of other route processing. Additionally, the CS-ID identifying the service does not necessarily have to be an anycast address. The existing address families are not sufficient to support future flexible expansion. Moreover, injecting periodically changing metric attributes into the existing address families may affect their stability. It is also essential to prevent attribute leaks to avoid security risks.

Alternatively, there is a current proposal to use a new BGP address family to propagate Service Metric Routing. The advantage of using a separate BGP address family is that routers not involved in service traffic processing are unaffected by Service Metric routing, as they do not pay attention to Service Metric routes. And the new BGP address family is highly extensible and can flexibly design packet formats, making packet packaging more efficient and reducing network load.

Furthermore, when the Ingress CATS-Forwarder router has not yet received service traffic, periodic updates of Service Metric routing

are unnecessary. This document presents a filtering notification mechanism for Service Metric routing, ensuring that notification to the corresponding Service Metric routing is only required when handling the respective service traffic. In this scenario, for environments supporting multiple services simultaneously, the Ingress CATS-Forwarder router only needs to focus on the Service Metric routing related to the services it handles. This approach significantly reduces the burden on the Ingress CATS-Forwarder.

2. Solution

To minimize the impact on existing routing information and to enhance security and scalability, service metrics are transmitted via an independent new address family.

2.1. Overview

For Service Metric routers, each service needs to be mapped to a service ID to differentiate between different services, called CS-ID. The CS-ID can be an IPv4 address, an IPv6 address, or more abstractly, an integer.

To differentiate between different service sites for the same service, each service site is assigned a service instance ID, called CIS-ID.

When CS-ID is used as an IPv4 or IPv6 address, it corresponds to the Anycast mode. The advantage of using Anycast mode is that it can leverage the existing routing and forwarding infrastructure. However, the drawback is that it can impact non-Service Metric routing, as all routers have to process Anycast routes. Therefore, we consider adopting a more general approach, which is to use a universal CS-ID instead of IPv4/IPv6 addresses.

The processing flow of this new approach is shown in the figure 1. The reachability information of the service site location is published to the ingress device through the BGP existing address family or other routing protocols to ensure the reachability of the service site location. The egress device sends the service metric information to the ingress device through the BGP new address family. The service metric information carries the service location information, and the outbound interface information of service is iterated through the service location information to form the service forwarding information, which is used to guide the forwarding of service traffic.

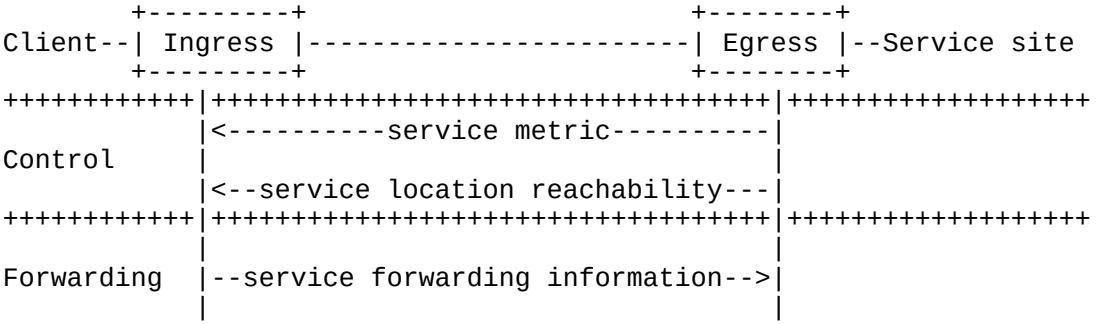


Figure 1: New approach Processing Flow

The mapping from service characteristics to CS-ID needs to be announced by the egress router. The Ingress CATS-Forwarder stores the mapping relationship and maps the received service traffic to the corresponding service CS-ID according to the mapping relationship. Service characteristics could include protocol type, service port number, TOS type, and so on. How the service characteristics is defined, and how the mapping relationship is published, are out of the scope of this document.

When CS-ID is used as an Anycast address, no service characteristics are required since the destination address of the service request message is the Anycast address of CS-ID.

The function of automatically discovering and announcing the mapping of service characteristics to CS-ID by the egress router can be abstracted into a module: Discoverer.

To facilitate the filtering of Service Metric routes by nodes that do not concern Service Metric routing, considering the characteristic of frequent updates in Service Metric routing, this document defines a new BGP Address-Family called BGP Service Metric (SM) AFI, and two new BGP Sub-Address-Family (BGP SM SAFI and BGP VPN SM SAFI), which leverages the characteristic of frequent updates in Service Metric routing.

The Ingress CATS-Forwarder receives the service mapping announcement sent by the Discoverer and saves the corresponding service mapping. In order to further reduce the bandwidth consumed by Service Metric routes, a dynamic filtering notification mechanism is introduced. If it needs to pay attention to the service metric information, it sends a Start-Notification for service metrics to the Discoverer. Here, we abstract a new module called Originator.

The Discoverer first sends a service automatic discovery route to notify the Ingress CATS-Forwarder about the existence of Service Metric routes. If the Ingress CATS-Forwarder needs to obtain the service metric information, it acts as an Originator and sends a start notification for service metrics to the Discoverer. On the contrary, if the Ingress CATS-Forwarder hasn't received any traffic related to the service yet, it doesn't need to pay attention to the service metrics at the moment.

Subsequently, The Discoverer receives the service metric start notification message sent by the Originator, records the start notification status, and sends service metric updates to the Originator.

In general, the Originator only needs to send start notification routes to request service metric information when it receives service requests related to the specific service. However, for simplification purposes, the Originator can also choose not to use the dynamic filtering notification mechanism and directly send start notification routes to request service metric information upon receiving automatic discovery routes.

Sending a start notification message without any service traffic can improve the response speed when the service traffic is first received. However, the downside is that it increases the load on the Ingress CATS-Forwarder. The specific usage scenario needs to be assessed based on whether priority is given to the response speed to service requests or to reducing the load on the Ingress CATS-Forwarder. This can also be determined based on the characteristics of each service. For example, for services with higher real-time requirements, immediate notification can be adopted, while other services can use on-demand notification.

The specific processing steps are as follows:

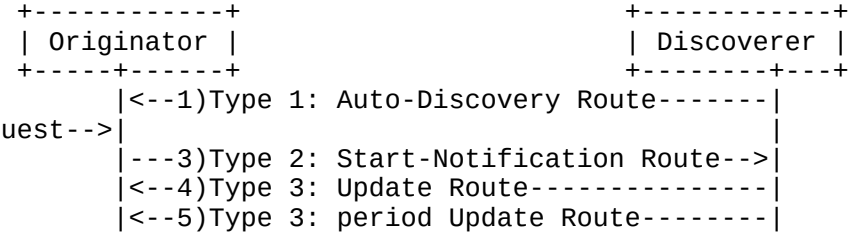


Figure 2: BGP Service Metric Route Process

1) The Discoverer gathers service information and sends an Automatic Discovery Route to the Originator to indicate the existence of a

service. The specific format of Automatic Discovery routes is shown in Section 3.1. If the Originator chooses not to use the On-Demand filtering notification mechanism, it skips the 2) step and proceeds directly to the 3) step upon receiving Automatic Discovery routes.

2) When the Originator receives a service request, it checks if it matches the characteristics of service specified in the previously received Automatic Discovery routes. If there is a match, it associates the request with the corresponding service type, as 3).

3) Originator sends a Start Notification Route to obtain the service metric information. The format of the Start Notification message is shown in Section 3.2.

4) Upon receiving the Start Notification route, the Discoverer sends an Update Route to notify the service metric information. The Update Route format is as shown in Section 3.4. In the case of multiple discoverers, the Originator needs to send Start Notification messages to all discoverers, and after receiving the Update Route from each Discoverer site, it selects the optimal route to guide the Service Metric route forwarding.

5) Thereafter, the Discoverer periodically sends Update Routes to update the service metric information when it changes.

2.2. Discovery and Notification for Service Metric

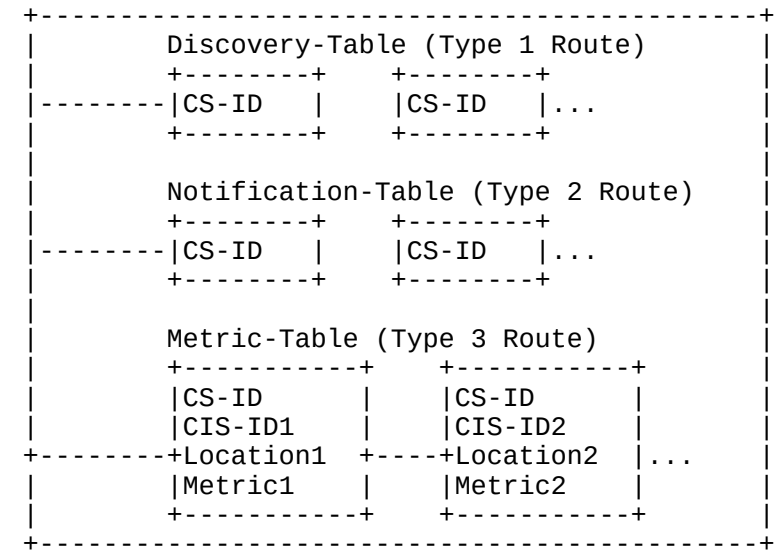


Figure 3: Service Metric Table for Public Network

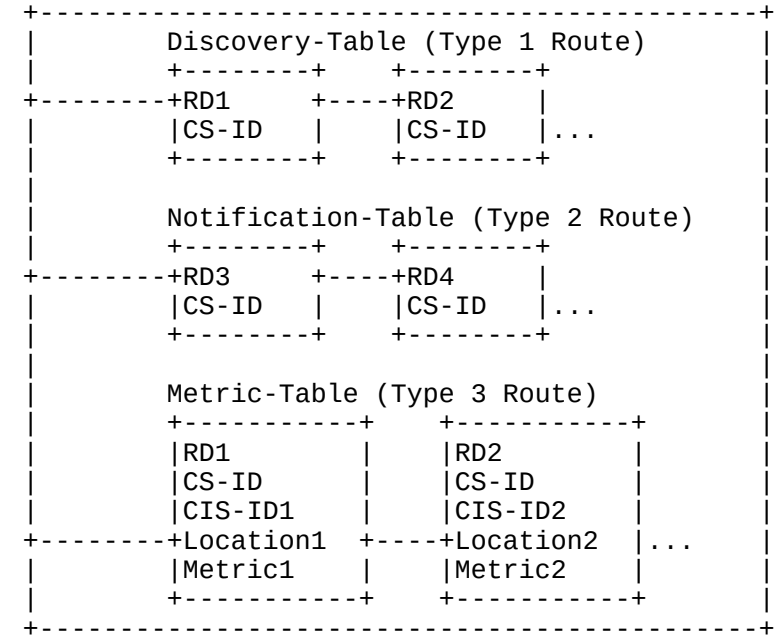


Figure 4: Service Metric Table for VPN Network

For each service type, maintain a Service Metric Table that records the CS-ID for each service. As shown in Figures 3 and 4, The Service Metric Table consists of a Discovery Table, a Notification Table, and a Metric Table. For VPN network, the service table contains RD, while for public network, it does not contain RD.

When Discoverer establishing a new BGP neighbor, the Type 1 automatic discovery routes is advertised to the neighbor to notify the associated CS-ID of the service.

When Originator receives discovery routes, it maintains a service discovery table based on the CS-ID.

If local on-demand filtering notification is required, the Originator only sends start notification routes to the Discoverer to request service metric information when it receives a local service request. Otherwise, it directly sends start notification routes to request service metric information.

Upon receiving Type 2 start notification routes from Originator, Discoverer sends Type 3 updated routes to the Originator to update the service metric information, and the Originator of this service is recorded for future use in sending updated routes based on this information.

When the service metric information changes afterwards, Discoverer sends Type 3 updated routes to the Originator based on the Notification-Table.

The service metric information is stored as Service Metric Tables and published via Type 3 routes. During publication, it is only sent to originators. Start Notification information is stored in the Notification Table.

To avoid frequent updates of service metric information, the updated routes are sent based on the minimum refresh time.

3. BGP Service Metric AFI and SAFI

In order to carry out the transmission of service metric information between different routers, this document defines a new BGP Service Metric Address Family Identifier (BGP SM AFI) and two new BGP Service Metric Subsequent Address Family Identifier (BGP SM SAFI and BGP VPN SM SAFI). Both SM SAFI and VPN SM SAFI SHOULD be applied to BGP SM AFI.

BGP SM SAFI can be enabled on transport devices in a provider network (underlay) to complete service metric transport across the

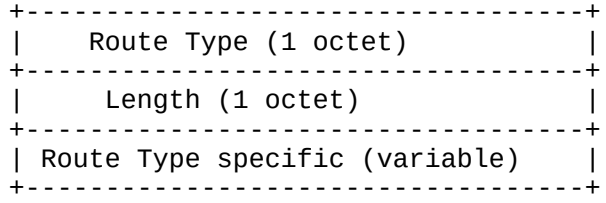
provider network. The multi-domain transport network may comprise of multiple BGP ASs as well as multiple IGP domains within a single BGP AS. BGP SM SAFI can also be enabled within a VRF on a PE router towards a peering CE router, and on devices within a customer network. BGP VPN SM SAFI is used for the distribution of service metric routes from different customers received on a PE router across the provider network, maintaining the separation of the customer address spaces that may overlap.

This document also defines an extensible NLRI model for both SAFIs that allow multiple NLRI types to be defined for different use cases. Each type of NLRI contains key and TLV based non-key fields for efficient encoding of different per-prefix information. The specific format information of the NLRI will be described in section 3.

4. BGP Service Metric Routes

This document defines a new BGP Network Layer Reachability Information (NLRI) called the Service Metric NLRI.

The format of the Service Metric NLRI is as follows:



The Route Type field defines the encoding of the rest of the Service Metric NLRI (Route Type specific Service Metric NLRI).

The Length field indicates the length in octets of the Route Type specific field of the Service Metric NLRI.

This document defines the following Route Types:

+ 1 - Service Metric Automatic Discovery route

+ 2 - Service Metric Start Notification route

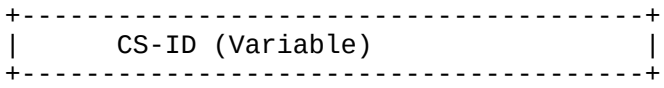
+ 3 - Service Metric Update route

The detailed encoding and procedures for these route types are described in subsequent sections.

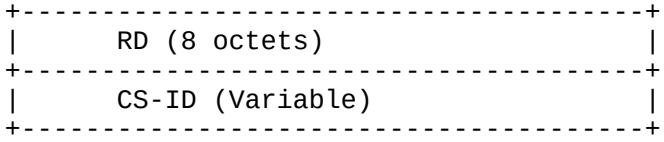
The Service Metric NLRI is carried in BGP [RFC4271] using BGP Multiprotocol Extensions [RFC4760] with an AFI of TBD and two SAFIs of Service Metric (To be assigned by IANA). The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute contains the Service Metric NLRI (encoded as specified above). Because the Service Metric Route does not use the next hop address of the MP_REACH_NLRI attribute, the length of Next Hop Network Address is set to 0, which helps to package service metric routes, that is, encapsulate multiple NLRIs in the same BGP update message to reduce network overhead.

4.1. The Service Metric Automatic Discovery NLRI

For BGP SM SAFI, A Service Metric Automatic Discovery route type specific Service Metric NLRI consists of the following:



For BGP VPN SM SAFI, A Service Metric Automatic Discovery route type specific Service Metric NLRI consists of the following:



The Discoverer utilizes Service Metric Automatic Discovery messages to publish service characteristics and their associated CS-ID. Originator that are interested in this service can need to obtain the service metric information of this service.

CS-ID: includes a 1-byte type field and a variable-length field.

Type: 1 Byte, indicates the type of CS-ID.

1 The CS-ID type is a 4-byte unsigned integer, and also contains 1-byte address family identifier (AFI = IPv4 or IPv6). AFI indicates the address family to which the service belongs.

2: The CS-ID type is an IPv4 Anycast address (4-byte), which indicates that the service belongs to the IPv4

address family.

3: The CS-ID type is an IPv6 Anycast address (16-byte),

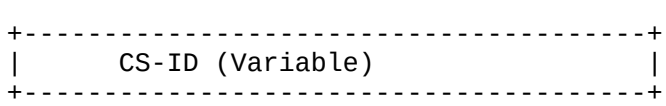
which indicates that the service belongs to the IPv6

address family.

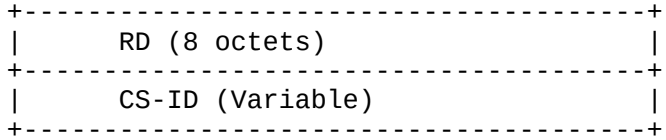
For the purpose of BGP route key processing, only CS-ID is considered to be part of the prefix in the NLRI.

4.2. The Service Metric Start Notification NLRI

For BGP SM SAFI, A Service Metric Start Notification route type specific Service Metric NLRI consists of the following:



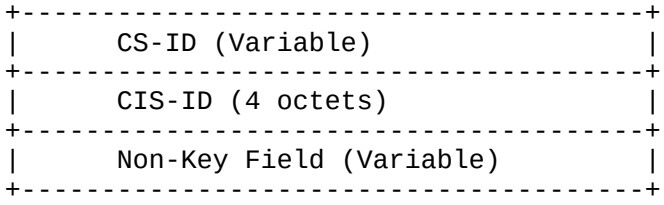
For BGP VPN SM SAFI, A Service Metric Start Notification route type specific Service Metric NLRI consists of the following:



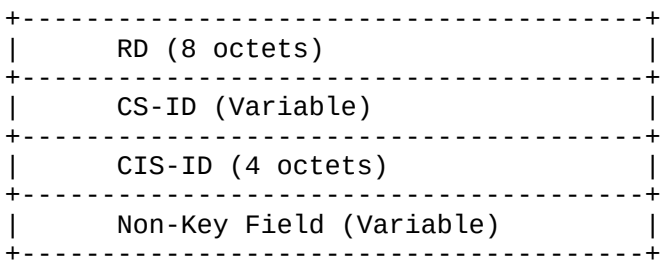
There are two instances in which a Discoverer sends an Automatic Discovery message. The first is when on-demand filtering notification is supported and local service metric information for a requested service is not available. In this scenario, the Originator sends a Start Notification message to Discoverer based on the automatic discovery message in order to request the corresponding service metric information from the Discoverer. The second instance is when on-demand filtering notification is not supported. In this scenario, upon receiving an automatic discovery message from the Discoverer, the Originator immediately sends a Start Notification message to request the corresponding service metric information.

4.3. The Service Metric Update NLRI

For BGP SM SAFI, A Service Metric Update route type specific Service Metric NLRI consists of the following:



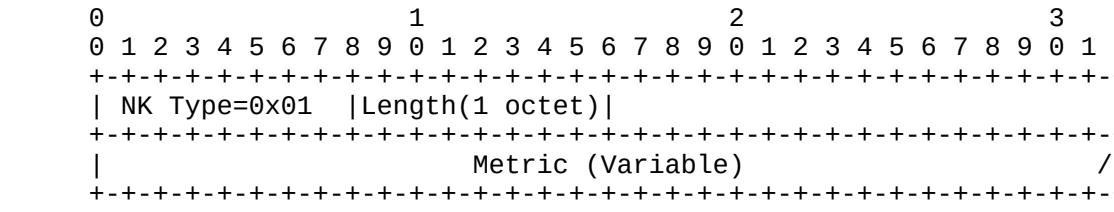
For BGP VPN SM SAFI, A Service Metric Update route type specific Service Metric NLRI consists of the following:



For the purpose of BGP route key processing, only CS-ID and CIS-ID are considered to be part of the prefix in the NLRI. CIS-ID can be an address or an integer. To simplify the implementation, this article defines CIS-ID as a 4-byte integer. The Non-Key Field is to be treated as a route attribute as opposed to being part of the route. The Non-Key Field is a series of TLV format contents.

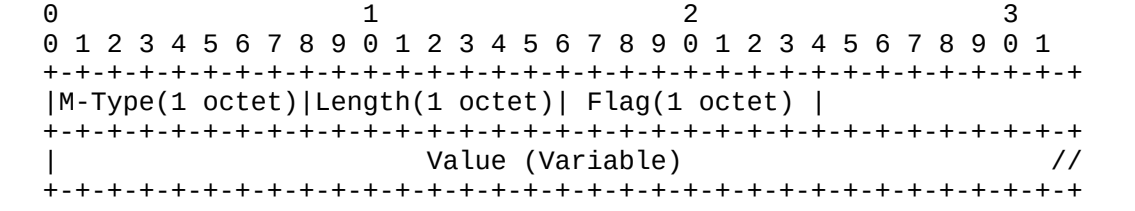
Non-Key (NK) Field: includes a 1-byte type field, a 1-byte length field and a variable-length value field.

Non-Key (NK) Type = 0x01: represents metric information, which is encoded as follows:



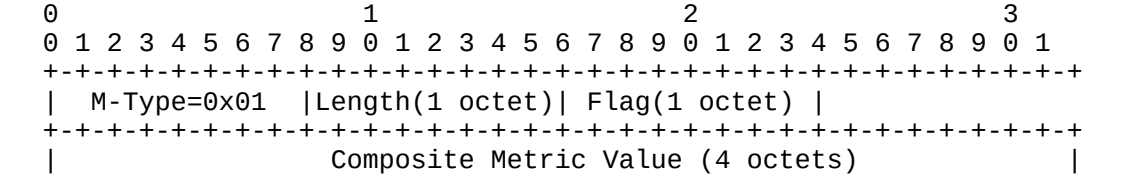
Where,

* Metric field: is defined here to be a set of elements encoded as "M-Type/Length/Flag/Value" (i.e., a set of TLVs). Each such TLV is encoded as shown blow:

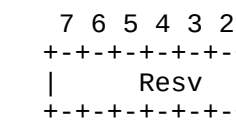


Where, Flag must be set to 0 if it is not used, and M-Type defines three commonly used types, as follows:

* M-Type = 0x01: represents a composite metric information, which is usually used to represent the execution capability of the service when a service has no special application requirements. The composite metric is an average of the raw metrics of the service, but the specific calculation method is beyond the scope of this article. This composite metric type is encoded as follows:



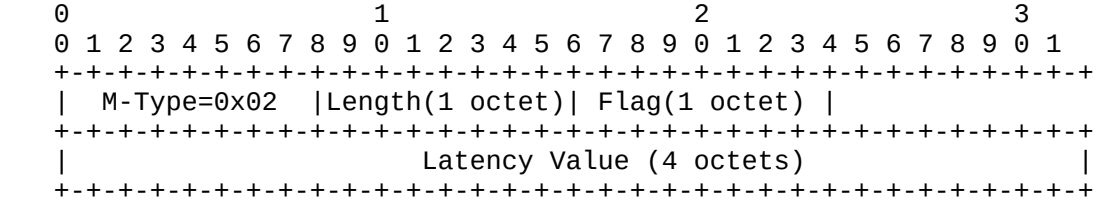
Where, The Flag carry additional information about the composite metric, the Flag is encoded as follows:



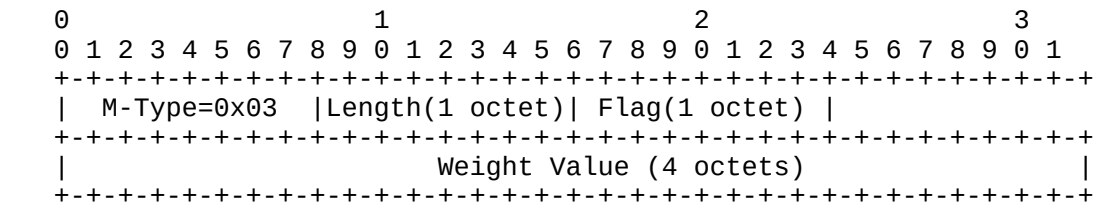
Bit R: represents composite metrics belonging to resource types. 1 indicates resource type, 0 indicates non-resource type.

Resv: Reserved for future use. MUST be set to zero.

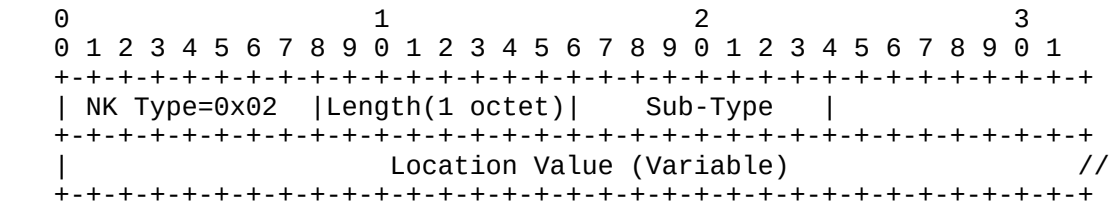
* M-Type = 0x02: represents a latency metric information, which is usually used as an indicator for selecting services when such services are sensitive to low latency. Calculation of latency in the network is beyond the scope of this document. This latency metric type is encoded as follows:



* M-Type = 0x03: represents a weight metric information, which is used as the weight value for selecting a service when the service needs to support load balancing deployment network. Again, the calculation of the weight is beyond the scope of this document. This weight metric type is encoded as follows:



Non-Key Type = 0x02: represents a Location information of service site, which is encoded as follows:



Where,
Sub-Type: 1 octet,

0x01: Location Value is IPv4 address (4 octets);

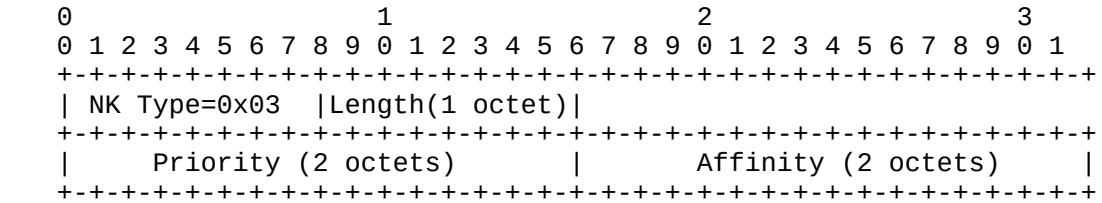
0x02: Location Value is IPv6 address (16 octets);

Location Value: The Location address of service site, the specific content depends on the Sub-Type.

The IPv4 or IPv6 Service site address must be advertised in other address family, such as in the EVPN address family. The routing path for service routes is forwarded through the actual path corresponding to the IPv4 or IPv6 Service site address. For example, when the IPv4 or IPv6 Service site address is forwarded through SR-

MPLS or SRv6, the service routes also inherit the corresponding forwarding path.

Non-Key Type = 0x03: represents a Priority information, which is encoded as follows:



Where,
Priority: 2 octets, Priority of the Service site.

Affinity: 2 octets, Affinity of the gateway where the Service site is located.

When the Discoverer receives a Service Metric start notification message for the first time, it sends an Update message to the Originator to notify the update of service metric information. Subsequently, if there are any changes in the service metric information, an Update message is sent to the originators to notify the update of service metric information. To avoid frequent updates of service metric, it is necessary to have a last update period to control the minimum interval for updating the service metric of a specified service.

5. Procedure

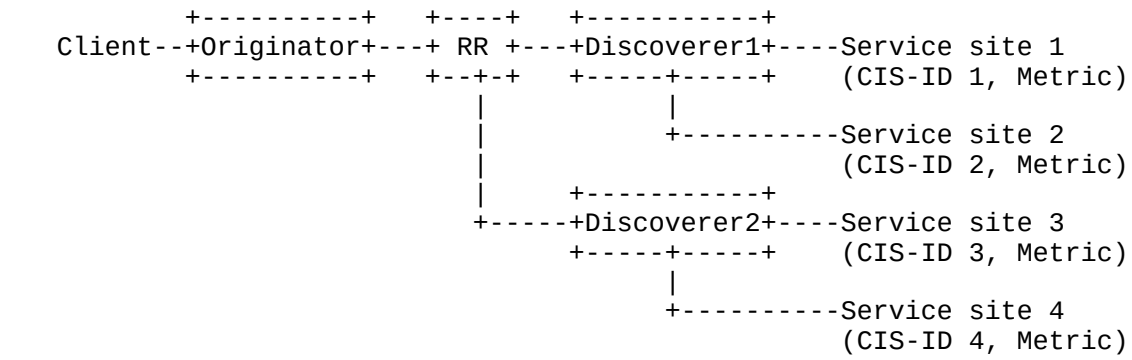


Figure 5: Service Metric Network Topology

The client is connected to the Originator, Service site 1 and Service site 2 are connected to Discoverer1, while Service site 3 and Service site 4 are connected to Discoverer2. The route reflector (RR) is used to receive and advertise service metric routes from all CATS-Forwarders which include Originator and Discoverer.

The following is the process of Originator and Discoverer interacting with BGP for Service Metric routing:

1) Discoverer1 sends a Type 1 Automatic Discovery Route to announce the service attributes associated with CS-ID 1. The Originator receives the Automatic Discovery route and maintains the discovery information for this service, recording the association between CS-ID and service attributes. If the Originator Forwarder itself does not support on-demand filtering notification, it directly proceeds to the 3) step and immediately sends Type 2 start notification routes.

2) When the Originator Forwarder receives a service request from the Client for the first time, it associates the request with the CS-ID based on the service attributes and the maintained discovery information. It then proceeds to the 3): sending Type 2 start notification routes to all the recorded Discoverers of this service.

3) The Originator Forwarder sends a Type 2 start notification route to all recorded Discoverers of this service based on the CS-ID, in order to request the metric information for this service.

4) When the Discoverer1 receives the Type 2 start notification routes, it sends the metric information for this service to originators by using Type 3 Update routes.

5) When the Originator Forwarder receives Type 3 Update route, Originator uses the Service site address carried by Type 3 Update route to iterate out the outbound interface of the Overlay network to guide the forwarding of Client service messages.

6) Discoverer2 establishes a BGP neighbor ship with Originator and sends Type 1 automatic discovery routes to notice service characteristics, associating them with CS-ID 1.

7) When Originator receives new discovery routes and if there are already other discoverers and service metric tables, it sends start notification routes to the new discoverer, requesting new service metric.

8) When there is a change in the service metric, the Discoverer1 or Discoverer2 sends Type 3 update routes to all originators based on the Type 2 start notification routes. Update routes are sent only to the originators, which helps in reducing network load.

9) When there is no service traffic for a long period of time, the service metric table is aged out, and Originator sends withdrawal of Type 2 start notification routes to all discoverers.

Discoverer advertises the Type 1 Automatic Discovery Route for BGP VPN SM SAFI in the form below:

```
    Type 1 Automatic Discovery Route UPDATE

        NLRI: AFI=TBD and SAFI=TBD
          Prefix: RD, CS-ID 1
          Next Hop Length: 0
        Attributes:
          Extended Community RT: RT Attribute for Service site
Location VPN
```

Figure 6: Type 1 Automatic Discovery Route Update Message

Originator advertises the Type 2 Start Notification Route for BGP VPN SM SAFI in the form below:

```
    Type 2 Start Notification Route UPDATE

        NLRI: AFI=TBD and SAFI=TBD
          Prefix: RD, CS-ID 1
          Next Hop Length: 0
        Attributes:
          Extended Community RT: RT Attribute for Client Location VPN
```

Figure 7: Type 2 Start Notification Route Update Message

Discoverer advertises the Type 3 Update Route for BGP VPN SM SAFI in the form below:

```
    Type 3 Update Route UPDATE

        NLRI: AFI=TBD and SAFI=TBD
          Prefix: RD, CS-ID 1, CIS-ID 1
          Non-Key Field: Metric (Composite Metric), Location (IPv4
Service site address), Priority (Priority=1, Affinity=2)
          Next Hop Length: 0
        Attributes:
```

Extended Community RT: RT Attribute for Service site Location VPN

Figure 8: Type 3 Update Route Update Message

6. Security Considerations

TBD.

7. IANA Considerations

7.1. Service Metric AFI and SAFI

This document requests a code point for Service Metric AFI (SM AFI) and SAFIs (SM SAFI and VPN SM SAFI) from the registry of Address Family Numbers and Subsequent Address Family Numbers.

7.2. Service Metric Route Types Registry

This document requests creation of a new registry for Service Metric Automatic Discovery, Service Metric Start Notification, and Service Metric Update. And IANA is requested to assign a new registry for "Non-Key Field" of Service Metric Update NLRI.

Also IANA is requested to assign a new registry for Metric type of Metric information in Service Metric Update NLRI.

8. Acknowledgements

The authors would like to thank Susan Hares for their comments to this document.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[I-D. draft-ietf-cats-framework] C. Li.,Z. Du.,M. Boucadair., L. M. Contreras.,J. Drake., " A Framework for Computing-Aware Traffic Steering (CATS)", draft-ietf-cats-framework-07(work in progress), April 2025.

Authors' Addresses

Changwang Lin
New H3C Technologies
Beijing
China
Email: linchangwang.04414@h3c.com

Huijuan Yao
China Mobile
No.32 XuanWuMen West Street
Beijing
100053
China
Email: yaohuijuan@chinamobile.com