

BGP Edge Metadata Path Applicability to CATS

draft-dunbar-cats-5g-metadata-applicability-00

Linda Dunbar: ldunbar@futurewei.com

Cheng Li: c.l@huawei.com

IETF 125

Shen Zhen

Edge Metadata Path Attribute

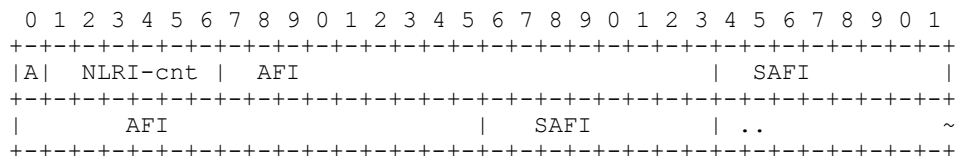
- An optional **Non-Transitive** BGP Path attribute to carry metrics and metadata about the edge services attached
 - If recognized, it can be used for local decision-making but must not advertise further.
 - If not recognize, it must silently discard it while still processing and possibly propagating the rest of the UPDATE (the route itself can still be propagated, just without that attribute).
- Only subset of prefixes' BGP advertisement include the Edge Metadata Path Attribute
- Local configuration dictates which prefix has Metadata Path Attribute attached.

```
BGP UPDATE Message
-----
Withdrawn Routes Length: 0
Total Path Attribute Length: <calculated>

Path Attributes:
- ORIGIN (Type Code 1): IGP
- AS_PATH (Type Code 2): { 65001 }
- NEXT_HOP (Type Code 3): 192.0.2.1
- MP_REACH_NLRI (Type Code 14):
  AFI: IPv4
  SAFI: Unicast
  Next Hop: 192.0.2.1
  NLRI: 198.51.100.0/24
- Metadata Path Attribute (Type Code 42):
  Flags: Optional, Non-Transitive
  Length: <variable>
  Sub-TLVs:
```

Limited Domain for the Metadata Distribution

- **BGP OPEN capability (Code 78) negotiates metadata support**
 - a one-octet Capability Code ((Value 78, assigned by IANA),
 - a one-octet Capability length, and
 - a variable-length Capability Value.



A Flag(1 bit):

- Set to 1 indicates that the Metadata attribute can be attached to any AFI/SAFI.
 - Set to 0 indicates that the Metadata attribute is restricted to specific AFI/SAFI pairs listed
- **Attribute Escape Prevention Mechanisms:**
 - NO-ADVERTISE Community
 - AS-Scope Sub-TLV
 - Route Filtering and Policies

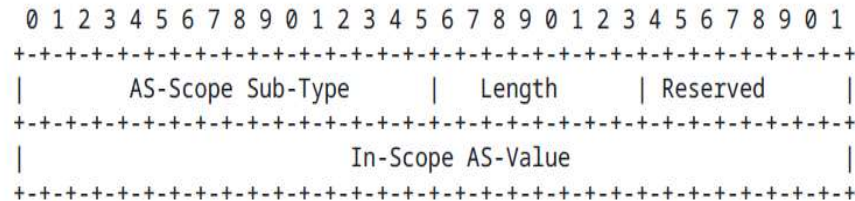
Limited Domain Provisioning

Typically localized with a well-defined topology. It may consist of edge sites (Edge DC) and local IP networks. E.g. connects 5G UPFs to those services hosted in the edge sites.

- ❑ Provisioning in a Limited Domain:
 - Routers in the domain are provisioned with policies that ensure traffic is routed based on both traditional BGP attributes and service-specific metadata. These policies must be consistent across the domain to ensure predictable behavior.
 - When the limited domain spans multiple ASes or administrative boundaries, inter-AS policies must be coordinated to ensure that routing behavior remains consistent.

Edge Metadata Propagation Scope

- To constrain the propagation:
 - **Non-transitive** when sending the BGP UPDATE message to its corresponding RR, so that routers that don't recognize them will drop those updates.
 - The RR can append the NO-ADVERTISE well-known community to the BGP UPDATE message with the Edge Metadata Path Attribute when forwarding it to the ingress routers.
 - signals to the ingress nodes that the associated route's Metadata Path Attribute SHOULD NOT be further advertised beyond their scope.
- AS-Scope SubTLV:
 - To prevent the Metadata Path Attribute from being leaked to unintended Autonomous Systems (ASes).



BGP Edge Metadata Path Attribute Mapping to CATS

- The Site Preference Index → CATS L2 policy metric:
E.g., a) high computing power; b) high bandwidth; c) high storage capacity.
- Site Physical Availability Index → "Shared Resource" [ll-idr-cats-bgp-extension]:
Indicates the percentage of impact on a group of routes associated with a common physical characteristic, e.g., a pod, a row of server racks, a floor, or an entire DC.
E.g., a power outage to a pod -> Site Physical Availability Index to be 0% for all the routes in the pod.
 - Site Index Associated to Routes
 - Standalone Site Availability Index
- Service Delay Prediction → Service Processing Delay (CATS L0 or L1):
0-100, with 0 indicating that the service delay is negligible and 100 indicating that the site has the most significant delay compared to all other sites for the same service.
- Service-Oriented Capability → Compute/Resource Capability (CATS L1/L2):
Available resources for a specific service in each deployment environment.
- Raw Measurement → Traffic Load / Utilization (CATS L0-L1/L2)

Metadata in Route Selection

- **Deployment-Specific Attribute Selection:**
 - Each deployment may choose to consider only a subset of the available metadata attributes
- **Influence on BGP Decision Process:**
 - The Metadata selectively impacts only a subset of client routes explicitly configured to consider service-specific attributes.
 - Different from traditional BGP metrics [RFC 4271] which can affect many routes sharing the same next hop.
- **Handling Degraded Metrics (with policy configured):**
 - If a service-specific metric indicates a degraded service quality (e.g., the Service Delay Prediction Index exceeds a threshold), the ingress router may deprioritize that route, even if traditional BGP attributes suggest it is optimal.
 - When the Capacity Availability Index or Service Delay Prediction Index for a service instance degrades beyond a configured threshold, the ingress router will treat that route as ineligible for traffic steering, similar to how BGP handles route withdrawals.
- **Fallback to Non-Metadata Routes:**
 - If no suitable routes with the required metadata are available, the BGP decision process defaults to traditional attribute evaluation [RFC 4271].

Enhanced Route Selection Procedure

- **Scenario 1: Metadata Value Higher Priority than Traditional BGP attributes**

Prefer Highest Local Preference → Prefer Highest Metadata Value → Traditional BGP Tie-breaking Steps (AS_PATH Length → Origin Type → MED) → Continue with Remaining BGP Tie-breaking (eBGP over iBGP → lowest IGP metrics → prefer oldest route → lowest route ID)

- **Scenario 2: Metadata Value at the Same Level as Traditional BGP Metrics**

Traditional BGP Steps (Local Preference → AS_PATH Length → Origin Type → MED) → Prefer Highest Metadata Value → Continue with Remaining BGP Tie-breaking

- **Scenario 3: Traditional BGP Metrics Higher Priority than Metadata ~~Composite-Value~~**

Traditional BGP Steps [RFC4271] → Metadata Value as Final Tie-Breaker

Gaps and Extensions Needed for CATS

- **Per-SCI Differentiation:**
 - Add an Instance-ID field or tie individual metric Sub-TLVs to distinct NLRIs so ingress routers can distinguish metrics for different service instances behind an anycast prefix.
- **Explicit Resource-Type Identification (for L0->L1 transitions):**
 - CATS defines specific raw metrics (CPU load, GPU usage, memory pressure. Adding optional resource-type enumerations would allow Raw Measurement or Capability TLVs to expose more structured CATS L0 primitives when needed.
- **Metric Lifetime / Validity Interval:**
 - CATS defines explicit metric-freshness semantics. The 5G Edge Metadata mechanism can be extended with: TTL or "valid until" timestamps, change-threshold indicators, or explicit "stale" flags.
- **Scope Identifier for Shared Resource Metrics:**
 - Allow SPAI-like metrics to reference a named resource pool (e.g., "DC-Room-A", "GPU-Cluster-2") when a site contains multiple independent shared resources.
- **Support for More Dynamic Metric Classes:**
 - For highly dynamic CATS L0 metrics, support for sub-second updates is generally not appropriate in BGP. Should mandate local smoothing/aggregation before exporting updates into the control plane.

Questions?

