

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: 23 July 2026

J. Tantsura
D. Sharp
V. Venkatraman
K. Muppalla
Nvidia
M. Rzehak
CoreWeave
A. Jouhari
Oracle
19 January 2026

BGP Unreachability Information SAFI
draft-tantsura-idr-unreachability-safi-01

Abstract

This document defines a new BGP Subsequent Address Family Identifier (SAFI) called "Unreachability Information" that allows the propagation of prefix unreachability information through BGP without affecting the installation or removal of routes in the Routing Information Base (RIB) or Forwarding Information Base (FIB). This mechanism enables network operators to share information about unreachable prefixes for monitoring, debugging, and coordination purposes while maintaining complete separation from the active routing plane.

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 23 July 2026.

Copyright Notice

Copyright (c) 2026 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Language	4
1.2. Terminology	4
2. Protocol Extensions	4
2.1. Unreachability Information SAFI	4
2.2. Capability Advertisement	5
2.3. BGP Path Attributes for Carrying Unreachability Information	6
3. NLRI Format	6
3.1. Approach to deal with multiple Reporters	7
3.2. IPv4 Unreachability NLRI (AFI=1, SAFI=TBD1)	7
3.3. IPv6 Unreachability NLRI (AFI=2, SAFI=TBD1)	7
3.4. Reporter TLV Format	8
3.5. Sub-TLV Format	9
3.5.1. Sub-TLV Type 1: Unreachability Reason Code	9
3.5.2. Sub-TLV Type 2: Timestamp	10
3.6. Encoding Examples	10
3.6.1. Single Reporter	10
3.6.2. Multiple Reporters	11
4. Operation	12
4.1. NLRI Processing and Aggregation	12
4.1.1. Aggregation Procedures	12
4.2. Withdrawal Procedures	13
4.2.1. Individual Reporter Withdrawal	13
4.2.2. Complete NLRI Withdrawal	14
4.2.3. Stale Reporter Detection	14
4.3. Path Selection for Aggregation	14
4.4. Interaction with Route Reflection	15
4.5. Communities and Attributes	15
4.6. Interaction with Graceful Restart	16
4.6.1. Graceful Restart Capability	16
4.6.2. Restarting Speaker Behavior	16
4.6.3. Receiving Speaker Behavior	17
4.6.4. Route Reflector Considerations	18
4.6.5. Implementation Recommendations	18
4.7. Preventing State Explosion	18
5. Deployment Considerations	19

5.1. Incremental Deployment	19
5.2. Use Cases	19
5.3. Operational Recommendations	20
6. Security Considerations	20
7. IANA Considerations	21
8. Acknowledgements	22
9. References	22
9.1. Normative References	22
9.2. Informative References	23
Appendix A. Implementation Considerations	23
Appendix B. Detailed Examples	24
B.1. Complete UPDATE Message Example	24
B.2. Aggregation Example	24
B.3. Withdrawal Example	25
Appendix C. Comparison with ADD-PATH Approach	26
C.1. Architectural Differences	26
C.2. Advantages of Nested TLV Approach	27
C.3. Disadvantages of Nested TLV Approach	27
C.4. When to Use Each Approach	27
Authors' Addresses	28

1. Introduction

BGP-4 [RFC4271] withdrawals are only propagated for prefixes that have been previously announced. This behavior, while preventing certain attack vectors, limits the ability of operators to share information about prefix unreachability for prefixes that were never announced in the first place.

There are several use cases where propagating unreachability information without affecting routing decisions would be valuable:

- * Debugging and troubleshooting routing issues across administrative domains
- * Sharing information about DDoS targets without null-routing traffic
- * Coordinating information about potentially hijacked prefixes
- * Monitoring and anomaly detection systems that need visibility into negative routing events
- * Providing telemetry about routing system health without affecting production traffic
- * Correlating unreachability reports from multiple network vantage points

This document defines a new SAFI that creates a parallel information plane for unreachability data, allowing BGP speakers to share this information while maintaining complete separation from the routing plane.

1.1. Requirements Language

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.

1.2. Terminology

UI-RIB: Unreachability Information RIB

NLRI: Network Layer Reachability Information

AFI: Address Family Identifier

SAFI: Subsequent Address Family Identifier

TLV: Type-Length-Value

Reporter TLV: A nested TLV structure containing information about one reporting BGP speaker and its associated unreachability details

Aggregation: The process of combining multiple Reporter TLVs from different paths into a single NLRI

Advertising Speaker: The BGP speaker that sends the UPDATE message containing the unreachability information

Reporting Speaker: The BGP speaker that originally generated the unreachability information (identified within a Reporter TLV)

2. Protocol Extensions

2.1. Unreachability Information SAFI

This document defines a new SAFI:

- * Value: TBD1 (to be assigned by IANA)

- * Name: Unreachability Information (UNREACH)

- * Applicable to AFI: 1 (IPv4) and 2 (IPv6)

2.2. Capability Advertisement

A BGP speaker that wishes to exchange Unreachability Information MUST advertise the corresponding AFI/SAFI capability as defined in [RFC5492].

The Capability Code for Multiprotocol Extensions is 1. The Capability Value field contains:

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|           AFI           |   Reserved   |   SAFI       |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Where:

- * AFI = 1 (IPv4) or 2 (IPv6)
- * Reserved = 0 (MUST be set to 0 on transmit, ignored on receive)
- * SAFI = TBD1 (Unreachability Information)

Additionally, this document defines a new capability:

- * Capability Code: TBD2 (to be assigned by IANA)
- * Capability Name: Enhanced Unreachability Information
- * Capability Value: 1 octet flags field

```

    0 1 2 3 4 5 6 7
+-----+-----+
|A|   Reserved   |
+-----+-----+

```

Where:

- * A bit: If set, speaker supports Aggregation of multiple Reporter TLVs
- * Reserved: MUST be zero on transmit, ignored on receive

The A bit indicates support for the aggregation procedures in Section 4.1.1. A speaker that advertises A=1 SHALL combine Reporter TLVs from multiple paths into a single NLRI. A speaker that

advertises A=0 MUST NOT perform aggregation and therefore only advertises the Reporter TLVs present on its best path. A peer receiving A=0 MUST NOT send aggregated NLRI's on that session.

2.3. BGP Path Attributes for Carrying Unreachability Information

Unreachability Information NLRI is carried in BGP UPDATE messages using the Multiprotocol Extensions for BGP-4 as defined in [RFC4760]. Specifically:

- * MP_REACH_NLRI (Path Attribute Type Code 14): Used to announce unreachability information for one or more prefixes. The presence of a prefix in MP_REACH_NLRI with AFI/SAFI indicating Unreachability Information signifies that the prefix is being reported as unreachable by one or more reporting speakers.
- * MP_UNREACH_NLRI (Path Attribute Type Code 15): Used to withdraw previously announced unreachability information. The presence of a prefix in MP_UNREACH_NLRI with AFI/SAFI indicating Unreachability Information signifies that the unreachability condition for that prefix has been cleared by all previously reporting speakers.

The AFI field in both attributes MUST be set to 1 (IPv4) or 2 (IPv6), and the SAFI field MUST be set to TBD1 (Unreachability Information).

The NLRI field within MP_REACH_NLRI contains the Unreachability Information NLRI as described in Section 3. The NLRI field within MP_UNREACH_NLRI contains only the prefix (Prefix Length and Prefix) without TLVs.

Standard BGP path attributes (AS_PATH, ORIGIN, NEXT_HOP via MP_REACH_NLRI, etc.) apply as defined in [RFC4760] and [RFC4271]. These attributes represent the path taken by the UPDATE message itself, not the paths of individual reporters (which are preserved in Reporter TLVs).

3. NLRI Format

The NLRI is uniquely identified by the combination of Prefix Length and Prefix. Reporter TLVs are NOT part of the NLRI key but provide information about each reporting speaker. The presence of an Unreachability Information NLRI for a prefix signifies that one or more speakers report the prefix as unreachable. The withdrawal of such an NLRI indicates that all reporters have cleared their unreachability reports for that prefix.

3.1. Approach to deal with multiple Reporters

When multiple BGP speakers report unreachability for the same prefix, implementers have several options:

1. Single Reporter: Do nothing and allow the Reporter Identifier of the best path to be used as the only Reporter. This is the simplest approach but loses information from other reporters.
2. Nested TLV Aggregation (Recommended): Implement the nested TLV aggregation approach described in this specification to preserve all reporter perspectives in a single NLRI. This provides the most comprehensive view while maintaining a single BGP path per prefix.
3. BGP ADD-PATH: Use BGP ADD-PATH [RFC7911] to maintain multiple paths, each carrying its own Reporter TLV. This preserves full BGP path attributes per reporter but requires ADD-PATH support.

This specification focuses on the nested TLV aggregation approach as the preferred mechanism, providing detailed procedures and encodings for this method throughout the remainder of this document.

3.2. IPv4 Unreachability NLRI (AFI=1, SAFI=TBD1)

```

0          1          2          3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Prefix Length |           IPv4 Prefix (variable)           |
+-----+-----+-----+-----+-----+-----+-----+-----+
|           Reporter TLVs (variable)           |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

* Prefix Length: 1 octet (0-32 bits)

* IPv4 Prefix: Variable, encoded as in [RFC4271] Section 4.3, and carried in MP_REACH_NLRI per [RFC4760]

* Reporter TLVs: One or more Reporter TLVs (Section 3.4)

3.3. IPv6 Unreachability NLRI (AFI=2, SAFI=TBD1)

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Prefix Length |               IPv6 Prefix (variable)               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Reporter TLVs (variable)               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

* Prefix Length: 1 octet (0-128 bits)

* IPv6 Prefix: Variable, encoded in MP_REACH_NLRI as defined in [RFC4760]

* Reporter TLVs: One or more Reporter TLVs (Section 3.4)

Example IPv6 prefix: 2001:db8::/32.

3.4. Reporter TLV Format

The Reporter TLV is a container that encapsulates information about a single reporting BGP speaker and its associated unreachability details. Multiple Reporter TLVs MAY appear in a single NLRI to represent reports from different speakers.

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|   Type   |               Length               |
+-----+-----+-----+-----+-----+-----+-----+
|               Reporter Identifier (4 octets)               |
+-----+-----+-----+-----+-----+-----+-----+
|               Reporter AS Number (4 octets)               |
+-----+-----+-----+-----+-----+-----+-----+
|               Sub-TLVs (variable)               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Reporter TLV Fields:

Type: 1 octet. Value: 1 (Reporter)

Length: 2 octets. Total length of the Reporter Identifier, Reporter AS Number, and Sub-TLVs fields (minimum 8 octets)

Reporter Identifier: 4 octets. BGP Identifier (Router ID) of the reporting speaker in network byte order

Reporter AS Number: 4 octets. 4-octet AS number of the reporting

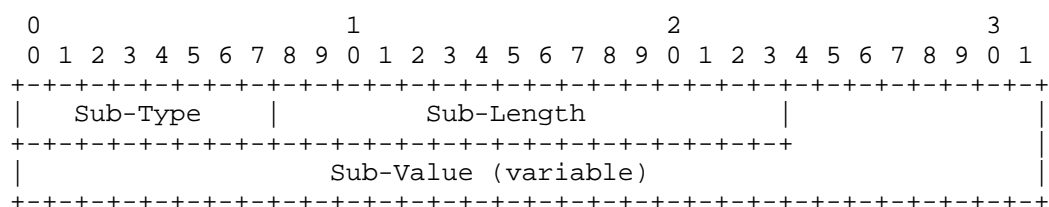
speaker in network byte order. If the AS number is less than 65536, the upper 2 octets are set to 0

Sub-TLVs: Variable length. Contains one or more Sub-TLVs providing details about this reporter's unreachability observation

The combination of Reporter Identifier and Reporter AS Number uniquely identifies the reporting speaker. Multiple Reporter TLVs with the same Reporter Identifier and AS Number MUST NOT appear in the same NLRI. If such duplication occurs, only the first occurrence SHOULD be processed.

3.5. Sub-TLV Format

Sub-TLVs appear within Reporter TLVs and provide specific details about the unreachability observation by that reporter.



Defined Sub-TLV Types:

Implementations MUST ignore unknown Sub-TLV types to allow for future extensibility. Multiple Sub-TLVs of the same type SHOULD NOT appear within a single Reporter TLV; if present, only the first occurrence SHOULD be processed.

3.5.1. Sub-TLV Type 1: Unreachability Reason Code

- * Sub-Type: 1
- * Sub-Length: 2 octets
- * Sub-Value: 2-octet reason code in network byte order

Defined Reason Codes:

- * 0: Unspecified
- * 1: Policy Blocked
- * 2: Security Filtered

- * 3: RPKI Invalid
- * 4: No Export Policy
- * 5: Martian Address
- * 6: Bogon Prefix
- * 7: Route Dampening
- * 8: Local Administrative Action
- * 9: Local Link Down
- * 10-64535: Reserved
- * 64536-65535: Reserved for Private Use

3.5.2. Sub-TLV Type 2: Timestamp

- * Sub-Type: 2
- * Sub-Length: 8 octets
- * Sub-Value: Unix timestamp (seconds since epoch) in network byte order, indicates when the unreachability event occurred or was detected by this reporter

3.6. Encoding Examples

3.6.1. Single Reporter

Example: 192.0.2.0/24 unreachable, reported by AS 65001, Router ID 198.51.100.1, Reason: RPKI Invalid

Prefix Length: 24 (0x18)
Prefix: 192.0.2.0 (0xC0000200)

Reporter TLV:

Type: 1
Length: 24 (0x0018)
Reporter Identifier: 198.51.100.1 (0xC6336401)
Reporter AS: 65001 (0x0000FDE9)

Sub-TLV (Reason):

Sub-Type: 1
Sub-Length: 2 (0x0002)
Sub-Value: 3 (0x0003) [RPKI Invalid]

Sub-TLV (Timestamp):

Sub-Type: 2
Sub-Length: 8 (0x0008)
Sub-Value: 1733789400 (0x0000000067596958)

Hexadecimal encoding:

18 C0 00 02 01 00 18 C6 33 64 01 00 00 FD E9 01
00 02 00 03 02 00 08 00 00 00 00 67 59 69 58

3.6.2. Multiple Reporters

Example: 192.0.2.0/24 unreachable, reported by two speakers

Prefix Length: 24
Prefix: 192.0.2.0

Reporter TLV #1:

Type: 1
Length: 24
Reporter Identifier: 198.51.100.1
Reporter AS: 65001
Sub-TLV (Reason): 3 (RPKI Invalid)
Sub-TLV (Timestamp): 1733789400

Reporter TLV #2:

Type: 1
Length: 24
Reporter Identifier: 198.51.100.2
Reporter AS: 65002
Sub-TLV (Reason): 1 (Policy Blocked)
Sub-TLV (Timestamp): 1733789410

Total NLRI length: 4 + 27 + 27 = 58 octets

4. Operation

4.1. NLRI Processing and Aggregation

When a BGP speaker receives an UPDATE message with Unreachability Information SAFI:

1. It MUST NOT install or remove any routes in the Loc-RIB based on this information
2. It MUST maintain a separate Unreachability Information RIB (UI-RIB) for this SAFI
3. It SHOULD apply standard BGP path selection to UI-RIB entries for consistency
4. It MAY propagate the information according to standard BGP rules, local policy, and aggregation procedures defined in Section 4.1.1
5. It MUST NOT mix Unreachability Information NLRI with other SAFIs in the same UPDATE message

4.1.1. Aggregation Procedures

When multiple UPDATE messages arrive advertising unreachability for the same prefix from different neighbors, a BGP speaker supporting aggregation (A bit set in capability) SHOULD combine the Reporter TLVs according to the following procedure:

1. Perform standard BGP path selection on the received updates. The "best path" is determined based on standard BGP decision process, considering only standard BGP attributes (AS_PATH length, LOCAL_PREF, etc.), NOT the content of Reporter TLVs.
2. Extract all Reporter TLVs from the best path.
3. For each non-selected path that would be feasible (not filtered by policy):
 - * Extract all Reporter TLVs from that path
 - * For each Reporter TLV, check if a Reporter with the same Reporter Identifier and Reporter AS already exists in the aggregated set
 - * If not present, add the Reporter TLV to the aggregated set

- * If already present, compare timestamps (if present) and keep the more recent report, or keep the existing entry if timestamps are equal or absent

4. Create a new NLRI containing all unique Reporter TLVs
5. Advertise this aggregated NLRI to appropriate neighbors, with BGP path attributes taken from the best path

The maximum number of Reporter TLVs that can be aggregated in a single NLRI is limited by the maximum BGP UPDATE message size (4096 octets). Implementations SHOULD limit the number of Reporter TLVs to prevent NLRI size from becoming unwieldy. A RECOMMENDED maximum is 50 Reporter TLVs per prefix, which allows for comprehensive multi-vantage-point monitoring while maintaining reasonable message sizes.

If the maximum is reached and a new reporter must be added, implementations SHOULD remove the oldest Reporter TLV (based on Timestamp Sub-TLV if present), unless this is the reporter of the best path. In that case the second oldest reporter should be removed.

4.2. Withdrawal Procedures

Withdrawal of unreachability information operates at two levels:

4.2.1. Individual Reporter Withdrawal

When a BGP speaker determines that a specific reporter no longer considers a prefix unreachable (e.g., receives an UPDATE from that reporter's AS that doesn't include the unreachability NLRI, or local policy determines the report is stale), it SHOULD:

1. Remove the corresponding Reporter TLV from the NLRI
2. If other Reporter TLVs remain, re-advertise the NLRI with the remaining Reporter TLVs
3. If no Reporter TLVs remain, withdraw the entire NLRI as described in Section 4.2.2

To facilitate individual reporter withdrawal, implementations MUST track the source of each Reporter TLV (which BGP neighbor or local process it came from).

4.2.2. Complete NLRI Withdrawal

A BGP speaker **MUST** withdraw an Unreachability Information NLRI (send the prefix in MP_UNREACH_NLRI) when:

- * All Reporter TLVs have been removed
- * The prefix is explicitly withdrawn by all upstream sources
- * Local policy dictates the information should no longer be propagated

The MP_UNREACH_NLRI contains only the prefix (Prefix Length and Prefix) without any TLVs.

4.2.3. Stale Reporter Detection

Implementations **SHOULD** implement aging mechanisms to remove stale Reporter TLVs:

- * If a Timestamp Sub-TLV is present and indicates the report is older than a configurable threshold (RECOMMENDED default: 24 hours), the Reporter TLV **MAY** be removed
- * If the BGP session to the neighbor that provided a Reporter TLV goes down, implementations **SHOULD** mark associated Reporter TLVs as potentially stale and **MAY** remove them after a grace period

4.3. Path Selection for Aggregation

The path selection for Unreachability Information SAFI follows standard BGP best path selection (RFC 4271 Section 9.1) with the following clarifications:

- * Weight/Local Preference: Apply normally based on local policy.
- * AS_PATH Length: Shorter AS_PATH is preferred. This represents the path the UPDATE message took.
- * ORIGIN: IGP preferred over EGP over INCOMPLETE.
- * MED: Apply if comparing paths from the same neighboring AS.
- * BGP Identifier: Use for tie-breaking.

The content of Reporter TLVs (number of reporters, reason codes, etc.) MUST NOT influence path selection. Path selection determines which UPDATE's BGP attributes are used for propagation, while aggregation combines Reporter TLVs from multiple paths.

4.4. Interaction with Route Reflection

Route Reflectors process Unreachability Information SAFI like any other AFI/SAFI combination:

- * Apply standard route reflection rules
- * ORIGINATOR_ID and CLUSTER_LIST attributes apply normally to the UPDATE message, not to individual reporters
- * Route Reflectors SHOULD support aggregation to combine reports from multiple clients
- * When reflecting to clients, include all aggregated Reporter TLVs

The distinction between the ORIGINATOR_ID BGP attribute and the Reporter Identifier field in Reporter TLVs is important:

- * ORIGINATOR_ID identifies the originator of the BGP UPDATE message for loop prevention
- * Reporter Identifier identifies the speaker that observed and reported the unreachability condition
- * These MAY be different in aggregated scenarios

4.5. Communities and Attributes

Standard BGP communities and attributes apply to the UPDATE message:

- * NO_EXPORT, NO_ADVERTISE, and NO_EXPORT_SUBCONFED work as defined
- * Large Communities [RFC8092] MAY be used for policy control of aggregation behavior
- * AS_PATH is constructed normally for the UPDATE message path
- * ORIGIN SHOULD be set to IGP for locally generated information

4.6. Interaction with Graceful Restart

BGP Graceful Restart (GR) as defined in [RFC4724] applies to the Unreachability Information SAFI. This section describes how UI-RIB is handled during restart scenarios.

4.6.1. Graceful Restart Capability

A BGP speaker that supports Graceful Restart for Unreachability Information SAFI MUST include the AFI/SAFI pair (AFI=1 or 2, SAFI=TBD1) in the Graceful Restart Capability advertisement as defined in RFC 4724.

The "Forwarding State" (F) bit for this SAFI:

- * Since Unreachability Information does not affect the forwarding plane (Loc-RIB or FIB), there is no forwarding state to preserve.
- * The F bit SHOULD be set to 0 in the Graceful Restart Capability for this AFI/SAFI combination.
- * Receiving speakers MUST NOT interpret the F bit for this SAFI as indicating preservation of forwarding state. The F bit, if set, has no defined meaning for this SAFI and MUST be ignored.
- * Implementations MAY use the F bit in future extensions to signal UI-RIB preservation capabilities, but such usage is outside the scope of this document.

4.6.2. Restarting Speaker Behavior

When a BGP speaker restarts and has negotiated GR for the Unreachability Information SAFI:

1. The speaker SHOULD set the F bit to 0 in the Graceful Restart Capability for this AFI/SAFI pair, as there is no forwarding state associated with this SAFI.
2. If the speaker preserved its UI-RIB across the restart, it SHOULD re-advertise all retained UI-RIB entries to its peers as soon as possible after restart, but MAY do so gradually to avoid overwhelming the network.
3. If the speaker did NOT preserve its UI-RIB across the restart, it SHOULD rebuild the UI-RIB from local information sources before re-advertising entries.

4. The speaker **MUST** send an End-of-RIB (EoR) marker for this SAFI after completing the re-advertisement of UI-RIB entries. The EoR marker is an UPDATE message with no NLRI and the MP_UNREACH_NLRI attribute containing the AFI/SAFI pair but no withdrawn routes.

4.6.3. Receiving Speaker Behavior

When a BGP speaker detects that a peer has restarted with GR capability for Unreachability Information SAFI:

1. The speaker **MUST** mark all UI-RIB entries learned from the restarting peer as stale. Stale marking occurs regardless of the F bit value, since the F bit has no defined semantics for this SAFI.
2. Stale entries **MUST NOT** be immediately withdrawn. They **MUST** be retained for the duration of the Restart Time advertised in the peer's GR Capability, or until the End-of-RIB marker is received, whichever comes first.
3. During the Restart Time period:
 - * Stale entries **MAY** be used for monitoring and correlation purposes
 - * Implementations **MAY** mark stale entries distinctly in display and APIs
 - * Stale entries **SHOULD NOT** be propagated to other peers unless explicitly configured to do so
4. Upon receiving the End-of-RIB marker from the restarting peer:
 - * All stale entries that were not refreshed **MUST** be removed from the UI-RIB
 - * Reporter TLVs from the restarted peer that were part of aggregated NLRIs **MUST** be removed if not refreshed
 - * If removal of Reporter TLVs leaves other Reporter TLVs for the same prefix, the NLRI **SHOULD** be re-advertised with the remaining Reporter TLVs
5. If the Restart Time expires before receiving the End-of-RIB marker, all stale entries **MUST** be removed immediately.

4.6.4. Route Reflector Considerations

Route Reflectors implementing Graceful Restart for this SAFI:

- * MUST properly handle stale marking of UI-RIB entries from restarting clients
- * SHOULD NOT reflect stale entries to other clients unless configured with a specific policy to do so
- * MUST correctly manage ORIGINATOR_ID and CLUSTER_LIST for entries that transition through stale and refresh phases
- * SHOULD send End-of-RIB markers to clients after the RR itself completes restart processing

4.6.5. Implementation Recommendations

- * Restart Time for this SAFI SHOULD be configurable independently from other AFI/SAFI combinations, with a RECOMMENDED default of 120 seconds.
- * Implementations SHOULD provide configuration options to:
 - Enable/disable preservation of UI-RIB across restarts
 - Control whether stale entries are propagated during GR
 - Set the Restart Time for this SAFI
 - Configure actions when End-of-RIB is not received in time
- * Implementations SHOULD log GR events for this SAFI to aid in debugging, including:
 - Detection of peer restart
 - Stale marking of entries
 - Receipt of End-of-RIB marker
 - Removal of stale entries

4.7. Preventing State Explosion

To prevent unbounded growth of the UI-RIB:

1. Implementations SHOULD limit the number of Reporter TLVs per prefix (RECOMMENDED maximum: 50)
2. Implementations SHOULD implement rate limiting for accepting new unreachability information
3. Default maximum UI-RIB size SHOULD be configurable with a reasonable default (e.g., 100,000 prefixes)
4. Implementations SHOULD implement memory limits for total Reporter TLV storage

5. Deployment Considerations

5.1. Incremental Deployment

The Unreachability Information SAFI can be deployed incrementally:

- * Speakers that don't support it simply don't negotiate the capability
- * Mixed environments work correctly with normal BGP capability negotiation
- * Can be enabled on specific sessions for testing
- * Aggregation support (A bit in capability) is OPTIONAL; speakers without it can still propagate single-reporter NLRIs
- * Speakers MAY use BGP dynamic capabilities to enable or disable this SAFI without resetting the BGP session

5.2. Use Cases

Example deployment scenarios:

Inter-AS Debugging: Enable between cooperating ASes for troubleshooting. Aggregation provides comprehensive view of why different ASes find a prefix unreachable.

Route Collectors: Deploy on route collector sessions for enhanced telemetry. Collectors can aggregate reports from multiple feeders to provide consolidated unreachability view.

DDoS Coordination: Share attack target information without null-routing. Multiple reports from different locations confirm attack patterns.

Security Monitoring: Track suspicious unreachability patterns. Correlation of reports from multiple vantage points aids in distinguishing localized issues from widespread problems.

RPKI Validation Monitoring: Track RPKI validation failures across different ASes. Aggregation shows consensus or disagreement on RPKI status.

5.3. Operational Recommendations

- * Enable aggregation on route collectors and monitoring systems to maximize visibility
- * Configure reasonable Reporter TLV limits based on expected number of reporters
- * Use Timestamp Sub-TLVs to facilitate debugging of temporal aspects of unreachability
- * Monitor UI-RIB size and Reporter TLV counts for capacity planning

6. Security Considerations

The Unreachability Information SAFI introduces new security considerations:

1. Information Leakage: Unreachability information might reveal network topology or operational issues. Reporter TLVs explicitly identify reporting ASes and routers. Operators SHOULD carefully consider peering policies for this SAFI.
2. State Exhaustion: Malicious peers could attempt to exhaust memory by advertising large numbers of unreachable prefixes or including excessive Reporter TLVs. Implementations SHOULD enforce limits as described in Section 4.7.
3. False Information: Peers could advertise false unreachability information or spoof Reporter TLVs. This SAFI SHOULD only be enabled with trusted peers. Consider validating Reporter Identifiers and AS Numbers against known valid values.
4. Prefix Hijacking: The SAFI could be used to claim prefixes are unreachable when they're not. Since this doesn't affect routing, the impact is limited to monitoring systems.

5. Reporter Impersonation: A malicious speaker could include Reporter TLVs claiming to represent other ASes. Implementations SHOULD validate that Reporter AS Numbers in TLVs are consistent with the AS_PATH of UPDATES that introduced them.
6. Aggregation Amplification: A malicious speaker could send many UPDATES with different Reporter TLVs for the same prefix, causing downstream aggregating speakers to accumulate and propagate large numbers of Reporter TLVs. Rate limiting and Reporter TLV limits mitigate this.

Operators SHOULD:

- * Use BGP TCP-AO [RFC5925] or MD5 for session protection
- * Implement prefix filtering for unreachability information
- * Monitor UI-RIB size and Reporter TLV counts
- * Enable this SAFI only with explicitly trusted peers
- * Validate Reporter AS Numbers against expected values
- * Configure appropriate Reporter TLV limits per prefix
- * Implement rate limiting on incoming unreachability UPDATES

7. IANA Considerations

IANA is requested to assign:

1. A new SAFI value (TBD1) for "Unreachability Information" applicable to AFI 1 and AFI 2 in the "Subsequent Address Family Identifiers (SAFI) Parameters" registry.
2. A new BGP Capability Code (TBD2) for "Enhanced Unreachability Information" in the "Capability Codes" registry.
3. Create a new registry called "BGP Unreachability Information Reporter TLV Types" under the "Border Gateway Protocol (BGP) Parameters" registry page. The registration procedure is "Standards Action". Initial value:
 - * Type 1: Reporter TLV (this document)

4. Create a new registry called "BGP Unreachability Information Sub-TLV Types" under the "Border Gateway Protocol (BGP) Parameters" registry page. The registration procedure is "RFC Required". Initial values:
 - * Sub-Type 1: Unreachability Reason Code (this document)
 - * Sub-Type 2: Timestamp (this document)
5. Create a new registry called "BGP Unreachability Reason Codes" under the "Border Gateway Protocol (BGP) Parameters" registry page. The registration procedure is "RFC Required". Initial values defined in Section 3.5.

8. Acknowledgements

The authors would like to thank the IDR working group for their valuable feedback and suggestions on this proposal.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <<https://www.rfc-editor.org/info/rfc4724>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.

- [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>>.

9.2. Informative References

- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <<https://www.rfc-editor.org/info/rfc7854>>.
- [RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <<https://www.rfc-editor.org/info/rfc7911>>.
- [RFC8092] Heitz, J., Ed., Snijders, J., Ed., Patel, K., Bagdonas, I., and N. Hilliard, "BGP Large Communities Attribute", RFC 8092, DOI 10.17487/RFC8092, February 2017, <<https://www.rfc-editor.org/info/rfc8092>>.

Appendix A. Implementation Considerations

Implementers should consider the following aspects when implementing the Unreachability Information SAFI with nested Reporter TLVs:

1. UI-RIB should store both the NLRI and associated Reporter TLVs with tracking of which neighbor provided each Reporter TLV
2. Show commands should clearly display all reporters for a given prefix, including Reporter ID, AS, Reason, and Timestamp
3. MIB/YANG models need structures to represent multiple reporters per prefix
4. BMP [RFC7854] should be extended to convey Reporter TLV information in monitoring messages
5. Efficient data structures for Reporter TLV storage and lookup are important for performance
6. Consider implementing a local database to track reporter history for forensic analysis

Appendix B. Detailed Examples

B.1. Complete UPDATE Message Example

An UPDATE message advertising that 192.0.2.0/24 is unreachable, reported by AS 65001 (Router 198.51.100.1) due to RPKI validation failure:

BGP UPDATE Message:

Withdrawn Routes Length: 0

Total Path Attribute Length: (calculated)

Path Attributes:

MP_REACH_NLRI (Type 14, Flags 0x90):

AFI: 1 (IPv4)

SAFI: TBD1 (Unreachability Information)

Next Hop Length: 0

Reserved: 0

NLRI:

Prefix Length: 24

Prefix: 192.0.2.0

Reporter TLV:

Type: 1

Length: 24

Reporter ID: 198.51.100.1

Reporter AS: 65001

Sub-TLV (Reason):

Sub-Type: 1

Sub-Length: 2

Reason: 3 (RPKI Invalid)

Sub-TLV (Timestamp):

Sub-Type: 2

Sub-Length: 8

Value: 1733789400

AS_PATH (Type 2):

Segment Type: AS_SEQUENCE

Segment Length: 1

AS: 65001

ORIGIN (Type 1):

Value: IGP (0)

B.2. Aggregation Example

Router R1 receives two UPDATES for 192.0.2.0/24:

UPDATE 1 (from Neighbor N1, AS 65100):
AS_PATH: 65100
Reporter TLV:
Reporter ID: 198.51.100.1, AS: 65001
Reason: RPKI Invalid (3)
Timestamp: 1733789400

UPDATE 2 (from Neighbor N2, AS 65200):
AS_PATH: 65200
Reporter TLV:
Reporter ID: 198.51.100.2, AS: 65002
Reason: Policy Blocked (1)
Timestamp: 1733789410

R1 Path Selection:
- Compare AS_PATH length: both length 1
- Compare by BGP ID: UPDATE 1 wins

R1 Aggregation:
- Extract Reporter TLV from UPDATE 1 (best path)
- Extract Reporter TLV from UPDATE 2 (add to aggregated set)
- Result: NLRI with 2 Reporter TLVs

R1 Advertisement to downstream:
AS_PATH: 65100 (from best path)
NLRI for 192.0.2.0/24:
Reporter TLV #1:
Reporter ID: 198.51.100.1, AS: 65001
Reason: 3, Timestamp: 1733789400
Reporter TLV #2:
Reporter ID: 198.51.100.2, AS: 65002
Reason: 1, Timestamp: 1733789410

B.3. Withdrawal Example

Scenario: Reporter 198.51.100.1 (AS 65001) clears its unreachability report, but Reporter 198.51.100.2 (AS 65002) maintains its report.

Initial State on Router R1:

UI-RIB Entry for 192.0.2.0/24:

Reporter TLV #1: 198.51.100.1/AS65001 (from Neighbor N1)

Reporter TLV #2: 198.51.100.2/AS65002 (from Neighbor N2)

Event: N1 sends MP_UNREACH_NLRI for 192.0.2.0/24

R1 Processing:

1. Identify that withdrawal came from N1
2. Find Reporter TLVs associated with N1
3. Remove Reporter TLV for 198.51.100.1/AS65001
4. Check remaining Reporter TLVs
5. Reporter TLV #2 still present
6. Re-advertise NLRI with remaining Reporter TLV

R1 Advertisement to downstream:

MP_REACH_NLRI for 192.0.2.0/24:

Reporter TLV #2:

Reporter ID: 198.51.100.2, AS: 65002

Reason: 1, Timestamp: 1733789410

Later Event: N2 also sends MP_UNREACH_NLRI for 192.0.2.0/24

R1 Processing:

1. Remove Reporter TLV for 198.51.100.2/AS65002
2. No Reporter TLVs remain
3. Send MP_UNREACH_NLRI for 192.0.2.0/24

R1 Advertisement to downstream:

MP_UNREACH_NLRI:

AFI: 1, SAFI: TBD1

Withdrawn Route: 192.0.2.0/24

Appendix C. Comparison with ADD-PATH Approach

This nested TLV approach differs from using BGP ADD-PATH [RFC7911] in several fundamental ways:

C.1. Architectural Differences

ADD-PATH Approach: Maintains multiple complete BGP paths for the same prefix, each with full set of BGP attributes (AS_PATH, communities, etc.). Each reporter's information travels as a separate BGP UPDATE path.

Nested TLV Approach: Aggregates multiple reporter perspectives into a single BGP path. Only one set of BGP attributes (representing the advertising path), but multiple Reporter TLVs within the NLRI.

C.2. Advantages of Nested TLV Approach

- * No ADD-PATH capability negotiation required - works with all BGP implementations
- * More compact representation - single UPDATE can carry reports from many speakers
- * Explicit aggregation model designed for consolidating multiple perspectives
- * Lower BGP state - one path entry instead of multiple
- * Easier correlation - all reports for a prefix in one NLRI

C.3. Disadvantages of Nested TLV Approach

- * More complex specification and implementation
- * BGP attribute ambiguity - AS_PATH represents the path of the UPDATE message, not the original paths reporters observed
- * Withdrawal complexity - requires tracking which neighbor provided each Reporter TLV
- * NLRI size can grow large with many reporters
- * Non-standard pattern in BGP protocol design
- * Path selection doesn't consider reporter content
- * New capability bit and processing logic required

C.4. When to Use Each Approach

Use ADD-PATH when:

- * Full BGP path information per reporter is important
- * Independent lifecycle management is critical
- * ADD-PATH is already widely deployed in the network
- * Standard BGP mechanisms are preferred

Use Nested TLV when:

- * ADD-PATH support is limited or unavailable
- * Compact aggregation is desired

- * Monitoring systems want single consolidated view
- * Lower BGP state is important

Authors' Addresses

Jeff Tantsura
Nvidia
Email: jefftant.ietf@gmail.com

Donald Sharp
Nvidia
Email: sharpd@nvidia.com

Vivek Venkatraman
Nvidia
Email: vivek@nvidia.com

Karthikeya Venkat Muppalla
Nvidia
Email: kmuppalla@nvidia.com

Maciej Rzehak
CoreWeave
Email: mrzehak@coreweave.com

Abderrahman Jouhari
Oracle
Email: jouharii@gmail.com