

Networking Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: December 18, 2013

L. Ginsberg  
S. Previdi  
Y. Yang  
Cisco Systems  
June 16, 2013

IS-IS Flooding Scope LSPs  
draft-ginsberg-isis-fs-lsp-01.txt

## Abstract

Intermediate System To Intermediate System (IS-IS) provides efficient and reliable flooding of information to its peers. However the current flooding scopes are limited to either area wide scope or domain wide scope. There are existing use cases where support of other flooding scopes are desirable. This document defines new Protocol Data Units (PDUs) which provide support for new flooding scopes as well as additional space for advertising information targeted for the currently supported flooding scopes.

The protocol extensions defined in this document are not backwards compatible with existing implementations and so must be deployed with care.

## Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

## 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 <http://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 18, 2013.

## Copyright Notice

Copyright (c) 2013 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 (<http://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 Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

## Table of Contents

1. Introduction . . . . .	4
2. Definition of New PDUs . . . . .	5
2.1. Flooding Scoped LSP Format . . . . .	5
2.2. Flooding Scoped CSNP Format . . . . .	8
2.3. Flooding Scope PSNP Format . . . . .	9
3. Flooding Scope Update Process Operation . . . . .	11
3.1. Scope Types . . . . .	11
3.2. Operation on Point-to-Point Circuits . . . . .	11
3.3. Operation on Broadcast Circuits . . . . .	12
3.4. Use of Authentication . . . . .	12
3.5. Priority Flooding . . . . .	12
4. Deployment Considerations . . . . .	12
5. Graceful Restart Interactions . . . . .	13
6. Multi-instance Interactions . . . . .	13
7. Circuit Scoped Flooding . . . . .	13
8. Extending LSP Set Capacity . . . . .	14
9. Domain Scoped Flooding . . . . .	14
10. Announcing Support for Flooding Scopes . . . . .	15
11. IANA Considerations . . . . .	16
12. Security Considerations . . . . .	17
13. Acknowledgements . . . . .	17
14. References . . . . .	17
14.1. Normative References . . . . .	17
14.2. Informational References . . . . .	18
Authors' Addresses . . . . .	18

## 1. Introduction

The Update Process as defined by [IS-IS] provides reliable and efficient flooding of information to all routers in a given flooding scope. Currently the protocol supports two flooding scopes and associated Protocol Data Units (PDUs). Level 1 (L1) Link State PDUs (LSPs) are flooded to all routers in an area. Level 2 (L2) LSPs are flooded to all routers in the Level 2 sub-domain. The basic operation of the Update Process can be applied to any subset of the routers in a given topology so long as that topology is not partitioned. It is therefore possible to introduce new PDUs in support of other flooding scopes and utilize the same Update Process machinery to provide the same reliability and efficiency which the Update Process currently provides for L1 and L2 scopes. This document defines these new PDUs and the modified Update Process rules which are to be used in supporting new flooding scopes.

New deployment cases have introduced the need for reliable and efficient circuit scoped flooding. For example, Appointed Forwarder information as defined in [RFC6326] needs to be flooded reliably and efficiently to all RBridges on a broadcast circuit. Currently, only Intermediate System to Intermediate System Hellos (IIHs) have the matching scope - but IIHs are unreliable i.e. individual IIHs may be lost without affecting correct operation of the protocol. To provide reliability in cases where the set of information to be flooded exceeds the carrying capacity of a single PDU requires sending the information periodically even when no changes in the content have occurred. When the information content is large this is inefficient and still does not provide a guarantee of reliability. This document defines circuit scoped flooding in order to provide a solution for such cases.

Another existing limitation of [IS-IS] is the carrying capacity of an LSP set. It has been noted in [RFC5311] that the set of LSPs that may be originated by a system at each level is limited to 256 LSPs and the maximum size of each LSP is limited by the minimum Maximum Transmission Unit (MTU) of any link used to flood LSPs. [RFC5311] has defined a backwards compatible protocol extension which can be used to overcome this limitation if needed. While the [RFC5311] solution is viable, in order to be interoperable with routers which do not support the extension it imposes some restrictions on what can/cannot be advertised in the Extended LSPs and requires allocation of multiple unique system IDs to a given router. A more flexible and less constraining solution is possible if interoperability with legacy routers is not a requirement. As the introduction of new PDUs required to support new flooding scopes is by definition not interoperable with legacy routers, it is possible to simultaneously introduce an alternative solution to the limited LSP set carrying

capacity as part of the extensions defined in this document. This capability is also defined in this document.

The PDU type field in the common header for all IS-IS PDUs is a 5 bit field. The possible PDU types supported by the protocol are therefore limited to a maximum of 32. In order to minimize the need to introduce additional PDU types in the future, the new PDUs introduced in this document are defined so as to allow multiple flooding scopes to be associated with the same PDU type. This means if new flooding scopes are required in the future the same PDU type can be used.

## 2. Definition of New PDUs

In support of new flooding scopes the following new PDUs are required:

- o Flooding Scoped LSPs (FS-LSPs)
- o Flooding Scoped Complete Sequence Number PDUs (FS-CSNPs)
- o Flooding Scoped Partial Sequence Number PDUs (FS-PSNPs)

Each of these PDUs is intentionally defined with a header as similar in format as possible to the corresponding PDU types currently defined in [IS-IS]. Although it might have been possible to eliminate or redefine PDU header fields in a new way the existing formats are retained in order to allow maximum reuse of existing PDU processing logic in an implementation.

Note that in the case of all FS PDUs, the Maximum Area Addresses field in the header of the corresponding standard PDU has been replaced with a Scope field. The maximum area addresses checks specified in [IS-IS] are therefore not performed on FS PDUs.

### 2.1. Flooding Scoped LSP Format

An FS-LSP has the following format:

	No. of octets
+-----+	
Intradomain Routeing	1
Protocol Discriminator	
+-----+	
Length Indicator	1
+-----+	
Version/Protocol ID	1

Extension		
+-----+		
ID Length		1
+-----+		
R R R  PDU Type		1
+-----+		
Version		1
+-----+		
Reserved		1
+-----+		
P  Scope		1
+-----+		
PDU Length		2
+-----+		
Remaining Lifetime		2
+-----+		
FS LSP ID		ID Length + 2
+-----+		
Sequence Number		4
+-----+		
Checksum		2
+-----+		
Reserved LSPDBOL IS Type		1
+-----+		
: Variable Length Fields :		Variable
+-----+		

Intradomain Routeing Protocol Discriminator - 0x83  
(as defined in [IS-IS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1

ID Length - As defined in [IS-IS]

PDU Type - 10 (Subject to assignment by IANA) Format as  
defined in [IS-IS]

Version - 1

Reserved - transmitted as zero, ignored on receipt

Scope - Bits 1-7 define the flooding scope.

The value 0 is reserved

and MUST NOT be used. Received FS-LSPs with a scope of 0 MUST  
be ignored.

P - Bit 8 - Priority Bit. If set to 1 this LSP SHOULD be flooded

at high priority.

PDU Length - Entire Length of this PDU, in octets, including the header.

Remaining Lifetime - Number of seconds before this FS-LSP is considered expired.

FS LSP ID - the system ID of the source of the FS-LSP. One of the following two formats is used:

#### FS LSP ID Standard Format

+-----+	
Source ID	ID Length
+-----+	
Pseudonode ID	1
+-----+	
FS LSP Number	1
+-----+	

#### FS LSP ID Extended Format

+-----+	
Source ID	ID Length
+-----+	
Extended FS LSP Number	2
+-----+	

Which format is used is specific to the Scope and MUST be defined when the specific flooding scope is defined.

Sequence Number - sequence number of this FS-LSP

Checksum - Checksum of contents of FS-LSP from Source ID to end. Checksum is computed as defined in [IS-IS].

#### Reserved/LSPDBOL/IS Type

Bits 4-8 are reserved, which means they are transmitted as 0 and ignored on receipt.

LSPDBOL - Bit 3 - A value of 0 indicates no FS-LSP Database Overload and a value of 1 indicates that the FS-LSP Database is overloaded. The overload condition is specific to FS-LSPs with the scope specified in the scope field.

IS Type - Bits 1 and 2. The type of Intermediate System as defined

in [IS-IS].

Variable Length Fields which are allowed in an FS-LSP are specific to the defined scope.

## 2.2. Flooding Scoped CSNP Format

An FS-CSNP has the following format:

	No. of octets
-----+   Intradomain Routeing   Protocol Discriminator   +-----+	1
Length Indicator   +-----+	1
Version/Protocol ID   Extension   +-----+	1
ID Length   +-----+	1
R R R  PDU Type   +-----+	1
Version   +-----+	1
Reserved   +-----+	1
R  Scope   +-----+	1
PDU Length   +-----+	2
Source ID   +-----+	ID Length + 1
Start FS-LSP ID   +-----+	ID Length + 2
End FS-LSP ID   +-----+	ID Length + 2
: Variable Length Fields : +-----+	Variable

Intradomain Routeing Protocol Discriminator - 0x83  
(as defined in [IS-IS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1



ID Length - As defined in [IS-IS]

PDU Type - 11 (Subject to assignment by IANA) Format as defined in [IS-IS]

Version - 1

Reserved - transmitted as zero, ignored on receipt

Scope - Bits 1-7 define the flooding scope.

The value 0 is reserved

and MUST NOT be used. Received FS-CSNPs with a scope of 0 MUST be ignored.

Bit 8 is Reserved which means it is transmitted as 0 and ignored on receipt.

PDU Length - Entire Length of this PDU, in octets, including the header.

Source ID - the system ID of the Intermediate System (with zero Circuit ID) generating this Sequence Numbers PDU

Start FS-LSP ID - The FS-LSP ID of the first FS-LSP with the specified scope in the range covered by this FS-CSNP.

End FS-LSP ID - The FS-LSP ID of the last FS-LSP with the specified scope in the range covered by this FS-CSNP.

Variable Length Fields which are allowed in an FS-CSNP are limited to those TLVs which are supported by standard CSNP.

### 2.3. Flooding Scope PSNP Format

An FS-PSNP has the following format:

	No. of octets
+-----+	
Intradomain Routeing	1
Protocol Discriminator	
+-----+	
Length Indicator	1
+-----+	
Version/Protocol ID	1
Extension	
+-----+	
ID Length	1
+-----+	

R R R  PDU Type		1
+-----+		
Version		1
+-----+		
Reserved		1
+-----+		
U  Scope		1
+-----+		
PDU Length		2
+-----+		
Source ID		ID Length + 1
+-----+		
: Variable Length Fields :		Variable
+-----+		

Intradomain Routeing Protocol Discriminator - 0x83  
(as defined in [IS-IS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1

ID Length - As defined in [IS-IS]

PDU Type - 12 (Subject to assignment by IANA) Format  
as defined in [IS-IS]

Version - 1

Reserved - transmitted as zero, ignored on receipt

Scope - Bits 1-7 define the flooding scope.

The value 0 is reserved

and MUST NOT be used. Received FS-PSNPs with a scope of 0 MUST  
be ignored.

U - Bit 8 - A value of 0 indicates that the specified  
flooding scope is supported. A value of 1 indicates  
that the specified flooding scope is unsupported. When  
U = 1, variable length fields other than authentication  
MUST NOT be included in the PDU.

PDU Length - Entire Length of this PDU, in octets, including  
the header.

Source ID - the system ID of the Intermediate System  
(with zero Circuit ID) generating this Sequence Numbers PDU

Variable Length Fields which are allowed in an FS-PSNP are

limited to those TLVs which are supported by standard PSNPs.

### 3. Flooding Scope Update Process Operation

The Update Process as defined in [IS-IS] maintains a Link State Database (LSDB) for each level supported. Each level specific LSDB contains the full set of LSPs generated by all routers operating in that level specific scope. The introduction of FS-LSPs creates additional LSDBs (FS-LSDBs) for each additional scope supported. The set of FS-LSPs in each FS-LSDB consists of all FS-LSPs generated by all routers operating in that scope. There is therefore an additional instance of the Update Process for each supported flooding scope.

Operation of the scope specific Update Process follows the Update Process specification in [IS-IS]. The circuit(s) on which FS-LSPs are flooded are limited to those circuits which are participating in the given scope. Similarly the sending/receiving of FS-CSNPs and FS-PSNPs is limited to the circuits participating in the given scope.

Consistent support of a given flooding scope on a circuit by all routers operating on that circuit is required.

#### 3.1. Scope Types

A flooding scope may be limited to a single circuit (circuit scope). Circuit scopes may be further limited by level (L1 circuit scope/L2 circuit scope).

A flooding scope may be limited to all circuits enabled for L1 routing (area scope).

A flooding scope may be limited to all circuits enabled for L2 routing (L2 sub-domain scope).

Additional scopes may be defined which include all circuits enabled for either L1 or L2 routing (domain-wide scope).

#### 3.2. Operation on Point-to-Point Circuits

When a new adjacency is formed, synchronization of all FS-LSDBs supported on that circuit is required. Therefore FS-CSNPs for all supported scopes MUST be sent when a new adjacency reaches the UP state. Send Receive Message (SRM) bit MUST be set for all FS-LSPs associated with the scopes supported on that circuit. Receipt of an FS-PSNP with the U bit equal to 1 indicates that the neighbor does

not support that scope (although it does support FS PDUs). This MUST cause SRM bit to be cleared for all FS-LSPs with the matching scope which are currently marked for flooding on that circuit.

### 3.3. Operation on Broadcast Circuits

FS PDUs are sent to the same destination address(es) as standard PDUs for the given protocol instance. For specification of the defined destination addresses consult [IS-IS], [IEEE802.1], [RFC6822], and [RFC6325].

The Designated Intermediate System (DIS) for a broadcast circuit has the responsibility to generate periodic scope specific FS-CSNPs for all supported scopes. A scope specific DIS is NOT elected as all routers on a circuit MUST support a consistent set of flooding scopes.

It is possible that a scope may be defined which is not level specific. In such a case the DIS for each level enabled on a broadcast circuit MUST independently send FS PDUs for that scope to the appropriate level specific destination address. This may result in redundant flooding of FS-LSPs for that scope.

### 3.4. Use of Authentication

Authentication TLVs MAY be included in FS PDUs. When authentication is in use, the scope is first used to select the authentication configuration that is applicable. The authentication check is then performed as normal. Although scope specific authentication MAY be used, sharing of authentication among multiple scopes and/or with the standard LSP/CSNP/PSNP PDUs is considered sufficient.

### 3.5. Priority Flooding

When the FS LSP ID Extended Format is used the set of LSPs generated by an IS may be quite large. It may be useful to identify those LSPs in the set which contain information of higher priority. Such LSPs will have the P bit set to 1 in the Scope field in the LSP header. Such LSPs SHOULD be flooded at a higher priority than LSPs with the P bit set to 0. This is a suggested behavior on the part of the originator of the LSP. When an LSP is purged the original state of the P bit MUST be preserved.

## 4. Deployment Considerations

Introduction of new PDU types is incompatible with legacy implementations. Legacy implementations do not support the FS

specific Update process(es) and therefore flooding of the FS-LSPs throughout the defined scope is unreliable when not all routers in the defined scope support FS PDUs. Further, legacy implementations will likely treat the reception of an FS PDUs as an error. Even when all routers in a given scope support FS PDUs, if not all routers in the flooding domain for a given scope support that scope flooding of the FS-LSPs may be compromised. Therefore all routers in the flooding domain for a given scope SHOULD support both FS PDUs and the specified scope before use of that scope can be enabled.

The U bit in FS-PSNPs provides a means to suppress retransmissions of unsupported scopes. Routers which support FS PDUs SHOULD support the sending of PSNPs with the U bit equal to 1 when an FS-LSP is received with a scope which is unsupported. Routers which support FS PDUs SHOULD trigger management notifications when FS PDUs are received for unsupported scopes and when PSNPs with the U bit equal to 1 are received.

## 5. Graceful Restart Interactions

[RFC5306] defines protocol extensions in support of graceful restart of a routing instance. Synchronization of all supported FS-LSDBs is required in order for database synchronization to be complete. This involves the use of additional T2 timers. Receipt of a PSNP with the U bit equal to 1 will cause FS-LSDB synchronization with that neighbor to be considered complete for that scope. See [RFC5306] for further details.

## 6. Multi-instance Interactions

In cases where FS-PDUs are associated with a non-zero instance the use of IID-TLVs in FS-PDUs follows the rules for use in LSPs, CSNPs, PSNPs as defined in [RFC6822].

## 7. Circuit Scoped Flooding

This document defines two circuit scoped flooding identifiers:

- o Level 1 circuit scope (L1CS)
- o Level 2 circuit scope (L2CS)

FS-LSPs with the scope field set to one of these values contain information specific to the circuit on which they are flooded. When received, such FS-LSPs MUST NOT be flooded on any other circuit. The

FS LSP ID Extended format is used in these PDUs. The FS-LSDB associated with circuit scoped FS-LSPs consists of the set of FS-LSPs which both have matching circuit scope and are transmitted (locally generated) or received on a specific circuit.

The set of TLVs which may be included in such FS-LSPs is specific to the given use case and is outside the scope of this document.

## 8. Extending LSP Set Capacity

The need for additional space in the set of LSPs generated by a single IS has been articulated in [RFC5311]. When legacy interoperability is not a requirement, the use of FS-LSPs meets that need without requiring the assignment of alias system-ids to a single IS. Two flooding scopes are defined for this purpose:

- o Level 1 Scoped FS-LSPs (L1-FS-LSP)
- o Level 2 Scoped FS-LSPs (L2-FS-LSP)

The FS LSP ID Extended format is used in these PDUs. This provides 64K of additional LSPs which may be generated by a single system at each level.

Lx-FS-LSPs are used by the level specific Decision Process (defined in [IS-IS]) in the same manner as standard LSPs (i.e. as additional information sourced by the same IS) subject to the following restrictions:

- o A valid version of LSP #0 from the same IS at the corresponding Level MUST be present in the LSDB in order for the FS-LSP set to be usable
- o Information in an Lx-FS-LSP (e.g. IS-Neighbor information) which supports using the originating IS as a transit node MUST NOT be used when the Overload bit is set in LSP #0
- o Existing TLVs which are restricted to LSP #0 MUST NOT appear in Lx-FS-LSPs.

There are no further restrictions as to what TLVs may be advertised in FS-LSPs.

## 9. Domain Scoped Flooding

Existing support for flooding information domain wide (i.e. to L1

routers in all areas as well as to routers in the Level 2 sub-domain) requires the use of leaking procedures between levels. For further details see [RFC4971]. This is sufficient when the data being flooded domain-wide consists of individual TLVs. If it is desired to retain the identity of the originating IS for the complete contents of a PDU, then support for flooding the unchanged PDU is desirable. This document therefore defines a domain-wide flooding scope. FS-LSPs with this scope MUST be flooded on all circuits regardless of what level(s) are supported on that circuit.

The FS LSP ID Extended format is used in these PDUs.

Use of information in FS-LSPs for a given scope depends on determining the reachability to the IS originating the FS-LSP. This presents challenges for FS-LSPs with domain-scopes because no single IS has the full view of the topology across all areas. It is therefore necessary for the originator of domain scoped FS-LSPs to advertise an identifier which will allow an IS who receives such an FS-LSP to determine whether the source of the FS-LSP is currently reachable. The identifier required depends on what "address-families" are being advertised.

When IS-IS is deployed in support of Layer 3 routing for IPv4 and/or IPv6 then FS-LSP #0 with domain-wide scope MUST include at least one of the following TLVs:

- o IPv4 Traffic Engineering Router ID (TLV 134)
- o IPv6 Traffic Engineering Router ID (TLV 140)

When IS-IS is deployed in support of Layer 2 routing, current standards (e.g. [RFC6325]) only support a single area. Therefore domain-wide scope is not yet applicable. When the Layer 2 standards are updated to include multi-area support the identifiers which can be used to support inter-area reachability will be defined - at which point the use of domain-wide scope for Layer 2 can be fully defined.

## 10. Announcing Support for Flooding Scopes

Announcements of support for flooding scope may be useful in validating that full support has been deployed and/or in isolating the reasons for incomplete flooding of FS-LSPs for a given scope.

ISs supporting FS-PDUs MAY announce supported scopes in IIH PDUs. To do so a new TLV is defined.

## Scoped Flooding Support

Type: 243 (suggested - to be assigned by IANA)

Length: 1 - 127

Value

	No of octets
+-----+  R  Supported Scope	1
+-----+ : :	
+-----+  R  Supported Scope	1
+-----+	

A list of the circuit scopes supported on this circuit and other non-circuit flooding scopes supported.

R bit MUST be 0 and is ignored on receipt.

In a Point-Point IIH L1, L2 and domain-wide scopes MAY be advertised.

In Level 1 LAN IIHs L1 and domain-wide scopes MAY be advertised.

In Level 2 LAN IIHs L2 and domain-wide scopes MAY be advertised.

Information in this TLV MUST NOT be considered in adjacency formation.

Whether information in this TLV is used to determine when FS-LSPs associated with a locally supported scope are flooded is an implementation choice.

## 11. IANA Considerations

This document requires the definition of three new PDU types that need to be reflected in the ISIS PDU registry. Values below are suggested values subject to assignment by IANA.

Value	Description
10	FS-LSP
11	FS-CSNP
12	FS-PSNP

This document requires that a new IANA registry be created to control the assignment of scope identifiers in FS-PDUs. The registration procedure is "Expert Review" as defined in [RFC5226]. Suggested registry name is "LSP Flooding Scoped Identifier Registry". A scope identifier is a number from 1-127 inclusive. The following scope



identifiers are defined by this document. Values are suggested values subject to assignment by IANA.

Value	Description	FS LSP ID Format
1	Level 1 Circuit Flooding Scope	Extended
2	Level 2 Circuit Flooding Scope	Extended
3	Level 1 Flooding Scope	Extended
4	Level 2 Flooding Scope	Extended
5	Domain-wide Flooding Scope	Extended

This document requires the definition of a new IS-IS TLV to be reflected in the "IS-IS TLV Codepoints" registry:

Type	Description	IIH	LSP	SNP	Purge
243	Circuit Scoped Flooding Support	Y	N	N	N

## 12. Security Considerations

Security concerns for IS-IS are addressed in [IS-IS], [RFC5304], and [RFC5310].

The new PDUs introduced are subject to the same security issues associated with their standard LSP/CSNP/PSNP counterparts. To the extent that additional PDUs represent additional load for routers in the network this increases the opportunity for denial of service attacks.

## 13. Acknowledgements

The authors wish to thank Ayan Banerjee, Donald Eastlake, and Mike Shand for their comments.

## 14. References

### 14.1. Normative References

- [IEEEaq] "Standard for Local and metropolitan area networks: Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks - Amendment 20: Shortest Path Bridging", IEEE Std 802.1aq-2012, 29 June 2012.", 2012.

- [IS-IS] "Intermediate system to Intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4971] Vasseur, JP., Shen, N., and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, October 2008.
- [RFC5306] Shand, M. and L. Ginsberg, "Restart Signaling for IS-IS", RFC 5306, October 2008.
- [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.
- [RFC6822] Previdi, S., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, December 2012.

#### 14.2. Informational References

- [RFC5311] McPherson, D., Ginsberg, L., Previdi, S., and M. Shand, "Simplified Extension of Link State PDU (LSP) Space for IS-IS", RFC 5311, February 2009.
- [RFC6325] Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC6326] Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 6326, July 2011.

Authors' Addresses

Les Ginsberg  
Cisco Systems  
510 McCarthy Blvd.  
Milpitas, CA 95035  
USA

Email: ginsberg@cisco.com

Stefano Previdi  
Cisco Systems  
Via Del Serafico 200  
Rome 0144  
Italy

Email: sprevidi@cisco.com

Yi Yang  
Cisco Systems  
7100-9 Kit Creek Road  
Research Triangle Park, North Carolina 27709-4987  
USA

Email: yiya@cisco.com



Networking Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 14, 2014

L. Ginsberg  
S. Mirtorabi  
S. Previdi  
A. Roy  
Cisco Systems  
July 13, 2013

IS-IS Support for Unidirectional Links  
draft-ginsberg-isis-udl-01.txt

Abstract

This document defines support for the operation of IS-IS over Unidirectional Links without the use of tunnels or encapsulation of IS-IS Protocol Data Units. Adjacency establishment when the return path from the router at the receive end of a unidirectional link to the router at the transmit end of the unidirectional link is via another unidirectional link is supported. The extensions defined here are backwards compatible - only the routers directly connected to a unidirectional link need to be upgraded.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 <http://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 January 14, 2014.

Copyright Notice

Copyright (c) 2013 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 (<http://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 Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

## Table of Contents

1. Introduction . . . . .	4
2. Encoding Extensions . . . . .	4
2.1. UDL LSPs and the UDL-TLV . . . . .	4
2.2. UDL Intermediate System Neighbors sub-TLV . . . . .	5
2.2.1. UDL Point-to-Point Intermediate System Neighbor Sub-TLV . . . . .	5
2.2.2. UDL LAN Intermediate System Neighbor Sub-TLV . . . . .	6
2.3. UDL LSP Range sub-TLV . . . . .	7
2.4. UDL LSP Entry sub-TLV . . . . .	7
2.5. UDL Manual Area Addresses sub-TLV . . . . .	8
3. Adjacency Establishment . . . . .	9
3.1. Adjacency Establishment in Point-to-Point Mode . . . . .	9
3.2. Adjacency Establishment in Broadcast Mode . . . . .	10
3.3. UDL link metric configuration . . . . .	10
4. Adjacency Maintenance . . . . .	11
4.1. Adjacency Maintenance by IS-T . . . . .	11
4.2. Adjacency Maintenance by IS-R . . . . .	12
4.3. Use of BFD . . . . .	12
5. Operation of the Update Process on a UDL . . . . .	13
6. Support for UDL on the Return Path . . . . .	14
7. IANA Considerations . . . . .	15
8. Security Considerations . . . . .	15
9. Acknowledgements . . . . .	15
10. References . . . . .	16
10.1. Normative References . . . . .	16
10.2. Informational References . . . . .	16
Authors' Addresses . . . . .	17

## 1. Introduction

Operation of IS-IS depends upon two-way connectivity. Adjacencies are formed by exchanging hellos on a link, flooding of the link state database is made reliable by exchanges between neighbors on a link, etc. However, there are deployments where operation of the protocol is desired over links which are unidirectional i.e., one end of the link can only send Protocol Data Units (PDUs) and one end of the link can only receive PDUs. Traditional methods of supporting Unidirectional Links (UDLs) have involved establishing a tunnel from the Intermediate System (IS) at the receive end of the UDL to the IS at the transmit end of the UDL, encapsulating/decapsulating the IS-IS PDUs as they enter/exit the tunnel, and associating the PDUs received via the tunnel with the UDL at the transmit end. This typically requires static configuration and may introduce Maximum Transmission Unit (MTU) issues due to the required encapsulation.

This specification defines extensions to the protocol which support correct and reliable operation of IS-IS over UDLs without the need for tunnels or any form of encapsulation.

## 2. Encoding Extensions

Although the IS at the transmit end of a UDL link (IS-T) can send IS-IS PDUs normally on the link, the IS at the receive end of a UDL link (IS-R) requires assistance from other ISs in the network to pass the information it would normally send directly to IS-T. The Update Process as defined in [IS-IS] allows information generated by one IS in the network to be reliably flooded to all other ISs in the network using Link State PDUs (LSPs). The extensions defined here utilize LSPs to allow IS-R to send information normally sent in hellos (IIHs) or sequence number PDUs (SNPs) to IS-T in LSPs. As LSPs are flooded to all ISs in an area/sub-domain, care is taken to minimize the LSP churn necessary to support adjacency establishment and maintenance between IS-T and IS-R.

### 2.1. UDL LSPs and the UDL-TLV

Routers on the receive end of a UDL MUST reserve at least one LSP (for each level supported on the UDL) to advertise the UDL information described below. Such LSPs are referred to as UDL-LSPs although the only distinction between a UDL-LSP and other LSPs is in the TLV information which is present in such an LSP. LSP #0 MUST NOT be used to send UDL information. UDL-LSPs have the following special characteristics:



1. The only TLV which may be advertised in UDL-LSPs is the UDL TLV described below and (optionally) an Authentication TLV and/or Purge Originator Identification TLV [RFC6232] . This requirement is enforced by the originator of the UDL-LSP but is not checked by receiving systems i.e., other TLVs which are included in a UDL-LSP are processed normally. The reason for the restriction is to minimize the number of LSPs which have UDL information content.
2. Routers on the transmit side of a UDL flood UDL-LSPs regardless of the existence of an adjacency in the UP state on that circuit. Flooding of UDL-LSPs on circuits other than a UDL is as specified in [IS-IS] i.e., no special handling.

A new TLV is defined in which UDL specific information appears. All information in a UDL-TLV is encoded in sub-TLVs. UDL sub-TLVs are formatted as specified in [RFC5305]. The format of the UDL-TLV is therefore:

	No. of octets
+-----+	
Type (11)	1
(To be assigned by IANA)	
+-----+	
Length	1
+-----+	
Sub-TLVs	3 - 255
:	:
+-----+	

## 2.2. UDL Intermediate System Neighbors sub-TLV

UDL links may operate in Point-to-Point mode or in broadcast mode (assuming the subnetwork is a broadcast subnetwork). There are therefore two types of Intermediate System Neighbors sub-TLVs defined. A UDL-TLV MUST NOT contain more than one Intermediate System Neighbors sub-TLV. If multiple Intermediate System Neighbors sub-TLVs appear in a UDL-TLV all information in that UDL-TLV MUST be ignored.

### 2.2.1. UDL Point-to-Point Intermediate System Neighbor Sub-TLV

The UDL Point-to-Point Intermediate System Neighbor Sub-TLV describes an adjacency on a UDL which is operating in Point-to-Point mode i.e. either a Point-to-Point subnetwork or a LAN subnetwork operating in Point-to-Point mode as described in [RFC5309]. The information

encoded follows the format for the Point-to-Point Three-Way Adjacency TLV as defined in [RFC5303] but may also include the local LAN address when the underlying subnetwork is a LAN.

No. of octets	
Type (240)   (To be assigned by IANA)	1
Length (9 + ID Length)   to (15 + ID Length)	1
Adjacency 3-way state	1
Extended Local Circuit ID	4
Neighbor System ID	ID Length
Neighbor Extended Local   Circuit ID	4
Local LAN Address	6

#### 2.2.2. UDL LAN Intermediate System Neighbor Sub-TLV

The UDL LAN Intermediate System Neighbor sub-TLV describes an adjacency on a UDL operating in broadcast mode on a LAN subnetwork.

No. of octets	
Type (6)   (To be assigned by IANA)	1
Length (7 + ID Length)	1
Neighbor LAN ID	ID Length + 1
Local LAN Address	6

### 2.3. UDL LSP Range sub-TLV

The content of this sub-TLV describes a range of LSPs for which the originating router requires an update. A UDL Intermediate System Neighbor sub-TLV MUST be included in any UDL-TLV where the UDL LSP Range sub-TLV is included. This is necessary so that only the specified neighbor processes the LSP range mentioned in the sub-TLV.

	No. of octets
+-----+	
Type (8)	1
(To be assigned by IANA)	
+-----+	
Length (ID Length + 2)* 2	1
+-----+	
Start LSP ID	ID Length + 2
+-----+	
End LSP ID	ID Length + 2
+-----+	

### 2.4. UDL LSP Entry sub-TLV

The content of this sub-TLV describes LSPs for which the originating router requires an update. A UDL Intermediate System Neighbor sub-TLV MUST be included in any UDL-TLV where the UDL LSP Entry sub-TLV is included. This is necessary so that only the specified neighbor processes the LSP entries mentioned in the sub-TLV.

	No. of octets
+-----+	
Type (9)	1
(To be assigned by IANA)	
+-----+	
Length (10 + ID Length)*N	1
+-----+	
: LSP Entries	:
+-----+	

Each LSP Entry has the following format:

+-----+	
Remaining Lifetime	2
+-----+	
LSP ID	ID Length + 2
+-----+	
LSP Sequence Number	4
+-----+	
Checksum	2
+-----+	

## 2.5. UDL Manual Area Addresses sub-TLV

This sub-TLV specifies the set of manualAreaAddresses of the originating system. No other sub-TLVs are allowed in a UDL-TLV which has this sub-TLV. Any other sub-TLVs in such a UDL-TLV are ignored on receipt.

	No. of octets
+-----+	
Type (1)	1
(To be assigned by IANA)	
+-----+	
Length	1
+-----+	
: Area Address(es)	:
+-----+	

Each Area Address has the following format:

+-----+	
Address Length	1
+-----+	
Area Address	Address Length
+-----+	

### 3. Adjacency Establishment

An adjacency over a UDL link may be established over a link operating in Point-to-Point mode (including a LAN subnetwork configured to operate in Point-to-Point mode) or a link operating in broadcast mode. Operation in either mode is identical except for some differences in the manner of adjacency establishment as specified in the following sub-sections.

IS-T utilizes the set of manualAreaAddresses advertised by IS-R in a UDL Manual Area Address sub-TLV in combination with the UDL Intermediate System Neighbor sub-TLV(s) to IS-T advertised by IS-R to determine the level(s) associated with any adjacency to IS-R.

#### 3.1. Adjacency Establishment in Point-to-Point Mode

Adjacency establishment makes use of Three Way Handshake as defined in [RFC5303] when operating in Point-to-Point mode. When operating over a LAN subnetwork, the use of point-to-point operation over LAN as defined in [RFC5309] is also used.

IS-T initiates adjacency establishment by sending Point-to-Point IIHs over the UDL as normal i.e., including Three-Way Handshake TLV. Note that the local circuit ID specified by IS-T need only be unique among the set of Point-to-Point UDL links supported by IS-T on which IS-T is at the transmit end.

Upon receipt of a Point-to-Point IIH IS-R creates an adjacency in the INIT state with IS-T and advertises the existence of the adjacency in its UDL-LSP(s) utilizing the UDL Point-to-Point Intermediate System Neighbor sub-TLV. The Local LAN address is included if the link is a LAN subnetwork operating in Point-to-Point mode. UDL-LSPs of the appropriate level(s) are generated according to the type of the adjacency with IS-T.

When IS-T receives the UDL-LSP(s) generated by IS-R containing the UDL Point-to-Point Intermediate System Neighbor sub-TLV it validates the 3 way information and, if valid, transitions its adjacency to UP state. In subsequent Point-to-Point IIHs IS-T includes IS-R's circuit ID information as indicated in the UDL Point-to-Point IS Neighbor sub-TLV in its 3 way handshake TLV. A complete set of CSNPs is sent to IS-R for the level(s) appropriate for the type of adjacency. LSPs which are updated as a result of the existence of the adjacency to IS-R are sent to IS-R, but IS-T does NOT propagate its full LSP Database. This is done to minimize the amount of redundant flooding.

IS-R uses normal adjacency bring up rules based on the 3 way

handshake information it receives in Point-to-Point IIHs from IS-T and advertises its IS neighbor to IS-T in the usual manner i.e. in an LSP other than a UDL-LSP. Following transition of the adjacency to IS-T to the UP state IS-R MAY request IS-T to flood its complete LSP Database by sending an LSP Range sub-TLV to IS-T in a UDL-LSP.

### 3.2. Adjacency Establishment in Broadcast Mode

IS-T initiates adjacency establishment by sending LAN IIHs of the appropriate level(s) over the UDL as normal. IS-T specifies itself in the LAN ID field of the IIH, including a non-zero circuit ID. Note that the local circuit ID specified by IS-T need only be unique among the set of LAN UDL links supported by IS-T on which IS-T is at the transmit end. This is because pseudo-node LSPs will never be generated for a UDL. Operation in broadcast mode supports a UDL with a single IS-T and multiple IS-Rs.

Upon receipt of a LAN IIH PDU IS-R creates an adjacency in the INIT state with IS-T and advertises the existence of the adjacency in its UDL-LSP(s) utilizing the UDL LAN Intermediate System Neighbor sub-TLV. UDL-LSPs of the appropriate level(s) are generated according to the levels supported by IS-R and IS-T.

When IS-T receives the UDL-LSP(s) generated by IS-R containing the UDL LAN Intermediate System Neighbor sub-TLV(s) it validates the LANID and, if valid, transitions its adjacency to UP state. In subsequent LAN IIH PDUs, IS-T includes IS-R's LAN Address as indicated in the UDL LAN IS Neighbor info. A complete set of CSNPs for the appropriate level is sent over the circuit. LSPs which are updated as a result of the existence of the adjacency to IS-R are sent to IS-R, but IS-T does NOT propagate its full LSP Database. This is done to minimize the amount of redundant flooding.

IS-R uses normal adjacency bring up rules based on the IS Neighbor LAN Address information it receives in LAN IIH PDUs from IS-T and advertises its IS neighbor to IS-T in an LSP other than a UDL-LSP. Note that there is no pseudo-node on a UDL LAN circuit - therefore both IS-T and IS-R MUST advertise an IS Neighbor TLV to each other, not to a pseudo-node. This is identical to what is done on a Point-to-Point subnetwork. Following transition of the adjacency to IS-T to the UP state IS-R MAY request IS-T to flood its complete LSP Database by sending an LSP Range sub-TLV to IS-T in a UDL-LSP.

### 3.3. UDL link metric configuration

What metrics are configured on a UDL depend upon the intended use of the UDL. If the UDL is to be used for unicast forwarding, then IS-T should be configured with the value appropriate to its intended

preference in the network topology and IS-R should be configured with maximum link metric ( $2^{24} - 1$ ) as defined in [RFC5305] (assuming wide metrics are in use). If the UDL is to be used for building a multicast Reverse Path Forwarding tree, then IS-R should be configured with the value appropriate to its intended preference in the network topology and IS-T should be configured with maximum link metric ( $2^{24} - 1$ ). If the link is to be used for both unicast forwarding and multicast, then it is necessary to have two different metric configurations and perform two different SPF calculations. This may be achieved through the use of multi-topology extensions as defined in [RFC5120]. Note that the configured link metrics have no bearing on adjacency establishment - they only affect the building of a Shortest Path Tree (SPT).

#### 4. Adjacency Maintenance

This section defines how adjacencies are maintained once established. Adjacency maintenance is defined without the need to send periodic UDL-LSP updates as this would be a significant burden on the entire network.

##### 4.1. Adjacency Maintenance by IS-T

IS-T sends IIH PDUs as normal on a UDL. As IS-R does NOT send IIH PDUs to IS-T, IS-T maintains the adjacency to IS-R so long as all of the following conditions are TRUE:

- o IS-T has a valid UDL-LSP from IS-R which includes Point-to-Point UDL IS Neighbor information or LAN UDL IS Neighbor information (as appropriate) regarding the adjacency IS-R has with IS-T on the UDL.
- o IS-T can calculate a return path rooted at IS-R to IS-T which does not traverse the UDL on which the adjacency is associated

When either of the above conditions becomes FALSE, IS-T brings down its adjacency to IS-R. Note that the return path calculation is only required when a topology change occurs in the network. It therefore need only be done in conjunction with a normal event driven SPF calculation.

NOTE: Immediately after the adjacency to IS-R has come up, if the only available return path traverses a UDL link on which the adjacency is still in the process of coming UP, the return path check will fail. This is possible because we bypass normal flooding rules to allow the UDL-LSP to be flooded even when the adjacency is not UP on a UDL link (as described later in this document). If IS-T

immediately brings the adjacency to IS-R down in this case, a circular dependency condition arises. To avoid this, if the return path check fails immediately after the adjacency comes up, a timer  $T_p$  is started. The timer is cancelled when a return path check succeeds. If the timer expires, IS-T brings down the adjacency to IS-R. A recommended value for the timer  $T_p$  is a small multiple (e.g., "twice") of the estimated time necessary to propagate LSPs across the entire domain.

Although it is unorthodox to bring up an adjacency without confirmed two way connectivity, the extension is well grounded because the receipt of IS-R's UDL-LSP by IS-T is indicative of the existence of a return path even though it cannot yet be confirmed by examination of the LSP database. This unconfirmed two way connectivity is a condition which we do not want to persist indefinitely - hence the use of timer  $T_p$ .

#### 4.2. Adjacency Maintenance by IS-R

IS-R maintains its adjacency with IS-T based on receipt of IIHs from IS-T as normal. So long as IS-T follows the rules for adjacency maintenance described in the previous section this is sufficient.

Further protection against pathological behavior on the part of IS-T (e.g., failure to perform the return path calculation after a topology change) MAY be implemented by IS-R. When IS-R receives a CSNP from IS-T which contains an SNP entry identifying an LSP which is not in IS-R's Link State Database (LSDB) a timer  $T_f$  is started for each such LSP. This includes entries which are older than, newer than, or non-existent in IS-R's LSDB. The timer  $T_f$  is cancelled if:

- o The associated LSP is received by IS-R on any circuit by normal operation of the Update process or
- o A subsequent set of CSNPs received from IS-T does not include the LSP entry

If any timer  $T_f$  expires IS-R brings down the adjacency with IS-T.

In the absence of pathological behavior by IS-T the  $T_f$  extension is not required. Its use is therefore optional.

#### 4.3. Use of BFD

A multi-hop BFD session [RFC5883] MAY be established between IS-T and IS-R. This can be used to provide fast failure detection. If used, this would also make the calculation by IS-T of a return path from IS-R to IS-T optional.



## 5. Operation of the Update Process on a UDL

For purposes of LSP propagation IS-T views the UDL as if it were a broadcast subnetwork where IS-T is the Designated Intermediate System (DIS). This is true regardless of the mode of operation of the circuit (point-to-point or broadcast). Therefore, IS-T propagates new LSPs on the UDL as they arrive but after sending an LSP on the UDL the SRM flag for that LSP is cleared i.e. no acknowledgement for the LSP is required or expected. IS-T also sends periodic CSNPs on the UDL.

IS-R cannot propagate LSPs to IS-T on the UDL. IS-R also cannot acknowledge LSPs received from IS-T on the UDL. In this respect IS-R operates on the UDL in a manner identical to a non-DIS on a broadcast circuit. If an LSP entry in a CSNP received from IS-T identifies an LSP which is "newer than" an LSP in IS-R's LSDB, IS-R MAY request the LSP from IS-T by sending a UDL-LSP with an LSP entry as described above. Since IS-R's UDL-LSP(s) will be propagated throughout the network even though the information is only of use to IS-Ts, it is recommended that some small delay occur between the receipt of a CSNP from IS-T and the generation of a UDL-LSP with an updated LSP entry by IS-R so as to allow for the possible receipt of the LSP either from IS-T or on another link.

If the number of LSP entries to be requested exceeds the space available in the UDL TLV associated with the adjacency to IS-T, IS-R MUST NOT generate multiple UDL TLVs associated with the same adjacency. Instead it should maintain the state of SSN flags appropriately for the LSP entries that require updates and send additional LSP entries (if necessary) in a subsequent UDL-LSP after the previously requested updates arrive.

Use of the LSP Range sub-TLV by IS-R allows more efficient encoding of a request for multiple LSPs. This could be especially useful following an adjacency UP event on a UDL. As described in Section 3, IS-T does NOT propagate its full LSP database following transition of an adjacency to IS-R to the UP state. This is consistent with IS-T operating in the role of DIS on a broadcast circuit. If IS-R has neighbors on other circuits it is possible that it will have received LSPs from other neighbors. In such a case flooding of the full LSP database by IS-T would be redundant. It is therefore left to the discretion of IS-R to request those portions of the LSP database which are not current. This is consistent with IS-R operating as a non-DIS on a broadcast circuit.

On receipt of a UDL-LSP generated by IS-R, IS-T checks the neighbor information in each UDL-TLV. If the information matches an existing adjacency that IS-T has with IS-R then IS-T sets SRM flag on the UDL

for any LSPs in its LSDB which are "newer" than the corresponding entries IS-R sent in LSP Entry sub-TLVs in UDL TLVs. SRM flags are also set on the UDL for LSPs which fall in the ranges specified in LSP Range sub-TLVs in UDL TLVs. UDL-TLVs associated with adjacencies to routers other than IS-T are ignored by IS-T.

## 6. Support for UDL on the Return Path

If all return paths from IS-R to IS-T traverse a UDL, then in order to bring up the adjacency between IS-T and IS-R at least one of the adjacencies on a return path UDL must already be UP. This is required because IS-T relies on receiving the UDL-LSP(s) generated by IS-R in order to bring up its adjacency. In order to overcome a circular dependency in the case where multiple pairs of UDL neighbors are trying to bring up an adjacency at the same time, an extension to LSP propagation rules is required.

When a new UDL-LSP is received by any IS which has one or more active UDLs on which it is operating as an IS-T, the set of neighbors other than the local system which are advertised in UDL-TLVs in the received UDL-LSP is extracted - call this UDL-LSP-ISN-SET. A return path from the originating IS-R to each neighbor in the UDL-LSP-ISN-SET is calculated. If there is no return path to one or more neighbors in this set periodic propagation of that UDL-LSP on all UDLs on which the local system acts as IS-T is initiated regardless of the state of an adjacency on that UDL. Periodic transmission of that UDL-LSP continues until a return path to all neighbors in the UDL-LSP-ISN-SET exists. This calculation is redone whenever the UDL-LSP is updated and when a topology change in the network occurs as a result of updates to the LSDB. Note that periodic retransmission is only done on UDLs on which the local system acts as IS-T.

If the network is partitioned the lack of a return path from a given IS-R to a given IS-T may persist. It is therefore recommended that the periodic retransmission employ an exponential backoff timer such that when the partition persists the periodic retransmission period is long enough so as to not represent a significant burden. It is recommended that the periodic retransmission be initially set to the locally configured CSNP interval. Note that periodic retransmission is only performed on UDL links and if an IS-R has previously received the same UDL-LSP it will silently ignore the retransmission since the UDL-LSP will already be in its LSDB. Unnecessary reflooding of the retransmitted UDL-LSP beyond the UDL does not occur.

IS-R MUST accept and propagate UDL-LSPs received on a UDL even when there is no adjacency in the UP state on the UDL circuit. Flooding of UDL-LSPs by IS-R uses normal flooding rules. LSPs received by

IS-R on the UDL which do NOT include UDL TLVs are discarded unless the adjacency is UP (normal processing).

This extension allows establishment of an adjacency on a UDL even when the return path transits another UDL which is also in the process of bringing up an adjacency. The periodic nature of the flooding is meant to compensate for the unreliability of the flooding. After the adjacency is UP, IS-R can request LSPs from IS-T by putting LSP entries into UDL-LSPs - but that ability is not available until the adjacency is UP.

## 7. IANA Considerations

This document requires the definition of a new IS-IS TLV to be reflected in the "IS-IS TLV Codepoints" registry:

Type	Description	IIH	LSP	SNP	Purge
----	-----	---	---	---	----
11	Unidirectional Link Information	N	Y	N	Y

This document requires that a new IANA registry be created to control the assignment of sub-TLV code points to be advertised within a Unidirectional Link Information TLV. The registration procedure is "Expert Review" as defined in [RFC5226]. The following sub-TLVs are defined by this document. Values are suggested values subject to assignment by IANA.

Value	Description
-----	-----
1	Manual Area Addresses
6	LAN IS Neighbor
9	LSP Entry
240	Point-to-Point IS Neighbor

## 8. Security Considerations

Security concerns for IS-IS are addressed in [IS-IS], [RFC5304], and [RFC5310].

## 9. Acknowledgements

The idea of supporting IS-IS on UDLs without using tunnels or encapsulation was originally introduced in the US patent "Support of

unidirectional link in IS-IS without IP encapsulation and in presence of unidirectional return path" (patent number: 7,957,380), by Sina Mirtorabi, Abhay Kumar Roy, Lester Ginsberg.

## 10. References

### 10.1. Normative References

- [IS-IS] "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5303] Katz, D., Saluja, R., and D. Eastlake, "Three-Way Handshake for IS-IS Point-to-Point Adjacencies", RFC 5303, October 2008.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

### 10.2. Informational References

- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, October 2008.
- [RFC5309] Shen, N. and A. Zinin, "Point-to-Point Operation over LAN in Link State Routing Protocols", RFC 5309, October 2008.
- [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.
- [RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, June 2010.

[RFC6232] Wei, F., Qin, Y., Li, Z., Li, T., and J. Dong, "Purge  
Originator Identification TLV for IS-IS", RFC 6232,  
May 2011.

Authors' Addresses

Les Ginsberg  
Cisco Systems  
510 McCarthy Blvd.  
Milpitas, CA 95035  
USA

Email: ginsberg@cisco.com

Sina Mirtorabi  
Cisco Systems  
3800 Zankar Road  
San Jose, CA 95134  
USA

Email: smirtora@cisco.com

Stefano Previdi  
Cisco Systems  
Via Del Serafico 200  
Rome 0144  
Italy

Email: sprevidi@cisco.com

Abhay Roy  
Cisco Systems  
560 McCarthy Blvd.  
Milpitas, CA 95135  
USA

Email: akr@cisco.com



Networking Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: December 19, 2013

S. Previdi, Ed.  
Cisco Systems, Inc.  
S. Giacalone  
Thomson Reuters  
D. Ward  
Cisco Systems, Inc.  
J. Drake  
A. Atlas  
Juniper Networks  
C. Filsfils  
Cisco Systems, Inc.  
June 17, 2013

IS-IS Traffic Engineering (TE) Metric Extensions  
draft-ietf-isis-te-metric-extensions-00

Abstract

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance criteria (e.g. latency) are becoming as critical to data path selection as other metrics.

This document describes extensions to IS-IS TE [RFC5305] such that network performance information can be distributed and collected in a scalable fashion. The information distributed using ISIS TE Metric Extensions can then be used to make path selection decisions based on network performance.

Note that this document only covers the mechanisms with which network performance information is distributed. The mechanisms for measuring network performance or acting on that information, once distributed, are outside the scope of this document.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

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 <http://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 19, 2013.

#### Copyright Notice

Copyright (c) 2013 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 (<http://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 Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.



## Table of Contents

1. Introduction . . . . .	4
2. TE Metric Extensions to IS-IS . . . . .	5
3. Interface and Neighbor Addresses . . . . .	6
4. Sub TLV Details . . . . .	6
4.1. Unidirectional Link Delay Sub-TLV . . . . .	7
4.2. Min/Max Unidirectional Link Delay Sub-TLV . . . . .	7
4.3. Unidirectional Delay Variation Sub-TLV . . . . .	9
4.4. Unidirectional Link Loss Sub-TLV . . . . .	9
4.5. Unidirectional Residual Bandwidth Sub-TLV . . . . .	10
4.6. Unidirectional Available Bandwidth Sub-TLV . . . . .	11
5. Announcement Thresholds and Filters . . . . .	12
6. Announcement Suppression . . . . .	13
7. Network Stability and Announcement Periodicity . . . . .	14
8. Enabling and Disabling Sub-TLVs . . . . .	14
9. Static Metric Override . . . . .	14
10. Compatibility . . . . .	14
11. Security Considerations . . . . .	14
12. IANA Considerations . . . . .	15
13. Acknowledgements . . . . .	15
14. References . . . . .	15
14.1. Normative References . . . . .	15
14.2. Informative References . . . . .	16
Authors' Addresses . . . . .	16

## 1. Introduction

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance information (e.g. latency) is becoming as critical to data path selection as other metrics.

In these networks, extremely large amounts of money rest on the ability to access market data in "real time" and to predictably make trades faster than the competition. Because of this, using metrics such as hop count or cost as routing metrics is becoming only tangentially important. Rather, it would be beneficial to be able to make path selection decisions based on performance data (such as latency) in a cost-effective and scalable way.

This document describes extensions to IS-IS Extended Reachability TLV defined in [RFC5305] (hereafter called "IS-IS TE Metric Extensions"), that can be used to distribute network performance information (such as link delay, delay variation, packet loss, residual bandwidth, and available bandwidth).

The data distributed by the TE Metric Extensions proposed in this document is meant to be used as part of the operation of the routing protocol (e.g. by replacing cost with latency or considering bandwidth as well as cost), by enhancing Constrained-SPF (CSPF), or for other uses such as supplementing the data used by an ALTO server [I-D.ietf-alto-protocol]. With respect to CSPF, the data distributed by ISIS TE Metric Extensions can be used to setup, fail over, and fail back data paths using protocols such as RSVP-TE [RFC3209]; [I-D.atlas-mpls-te-express-path] describes some methods for using this information to compute Label Switched Paths (LSPs) at the LSP ingress.

Note that the mechanisms described in this document only disseminate performance information. The methods for initially gathering that performance information, such as [RFC6375], or acting on it once it is distributed are outside the scope of this document. Example mechanisms to measure latency, delay variation, and loss in an MPLS network are given in [RFC6374]. While this document does not specify how the performance information should be obtained, the measurement of delay SHOULD NOT vary significantly based upon the offered traffic load. Thus, queuing delays SHOULD NOT be included in the delay measurement. For links, such as Forwarding Adjacencies, care must be taken that measurement of the associated delay avoids significant queuing delay; that could be accomplished in a variety of ways, including either by measuring with a traffic class that experiences minimal queuing or by summing the measured link delays of the components of the link's path.

## 2. TE Metric Extensions to IS-IS

This document proposes new IS-IS TE sub-TLVs that can be announced in ISIS Extended Reachability TLV (TLV-22) to distribute network performance information. The extensions in this document build on the ones provided in IS-IS TE [RFC5305] and GMPLS [RFC4203].

IS-IS Extended Reachability TLV 22 (defined in [RFC5305]), Inter-AS reachability information TLV 141 (defined in [RFC5316]) and MT-ISN TLV 222 (defined in [RFC5120]) have nested sub-TLVs which permit the TLVs to be readily extended. This document proposes several additional sub-TLVs:

Type	Value
-----	
TBA	Unidirectional Link Delay
TBA	Low/High Unidirectional Link Delay
TBA	Unidirectional Delay Variation
TBA	Unidirectional Packet Loss
TBA	Unidirectional Residual Bandwidth
TBA	Unidirectional Available Bandwidth

As can be seen in the list above, the sub-TLVs described in this document carry different types of network performance information. The new sub-TLVs include a bit called the Anomalous (or "A") bit. When the A bit is clear (or when the sub-TLV does not include an A bit), the sub-TLV describes steady state link performance. This information could conceivably be used to construct a steady state performance topology for initial tunnel path computation, or to verify alternative failover paths.

When network performance violates configurable link-local thresholds a sub-TLV with the A bit set is advertised. These sub-TLVs could be used by the receiving node to determine whether to fail traffic to a backup path, or whether to calculate an entirely new path. From an MPLS perspective, the intent of the A bit is to permit LSP ingress nodes to:

- A) Determine whether the link referenced in the sub-TLV affects any of the LSPs for which it is ingress. If there are, then:
- B) Determine whether those LSPs still meet end-to-end performance objectives. If not, then:
- C) The node could then conceivably move affected traffic to a pre-established protection LSP or establish a new LSP and place the traffic in it.

If link performance then improves beyond a configurable minimum value (reuse threshold), that sub-TLV can be re-advertised with the Anomalous bit cleared. In this case, a receiving node can conceivably do whatever re-optimization (or fallback) it wishes to do (including nothing).

Note that when a sub-TLV does not include the A bit, that sub-TLV cannot be used for failover purposes. The A bit was intentionally omitted from some sub-TLVs to help mitigate oscillations. See Section 5 for more information.

Consistent with existing IS-IS TE specifications [RFC5305], the bandwidth advertisements defined in this draft MUST be encoded as IEEE floating point values. The delay and delay variation advertisements defined in this draft MUST be encoded as integer values. Delay values MUST be quantified in units of microseconds, packet loss MUST be quantified as a percentage of packets sent, and bandwidth MUST be sent as bytes per second. All values (except residual bandwidth) MUST be calculated as rolling averages where the averaging period MUST be a configurable period of time. See Section 5 for more information.

### 3. Interface and Neighbor Addresses

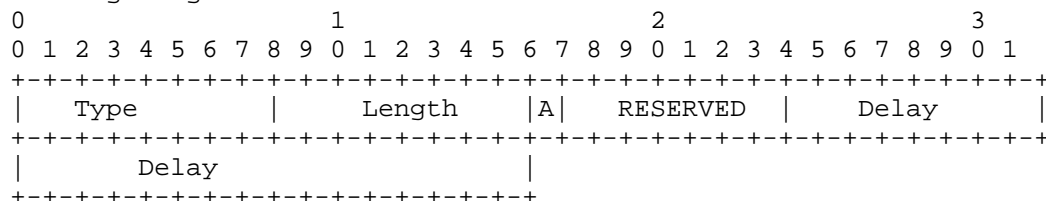
The use of TE Metric Extensions SubTLVs is not confined to the TE context. In other words, IS-IS TE Metric Extensions SubTLVs defined in this document can also be used for computing paths in the absence of a TE subsystem.

However, as for the TE case, Interface Address and Neighbor Address SubTLVs (IPv4 or IPv6) MUST be present. The encoding is defined in [RFC5305] for IPv4 and in [RFC6119] for IPv6.

### 4. Sub TLV Details

#### 4.1. Unidirectional Link Delay Sub-TLV

This sub-TLV advertises the average link delay between two directly connected IS-IS neighbors. The delay advertised by this sub-TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:



where:

Figure 1

Type: TBA

Length: 4

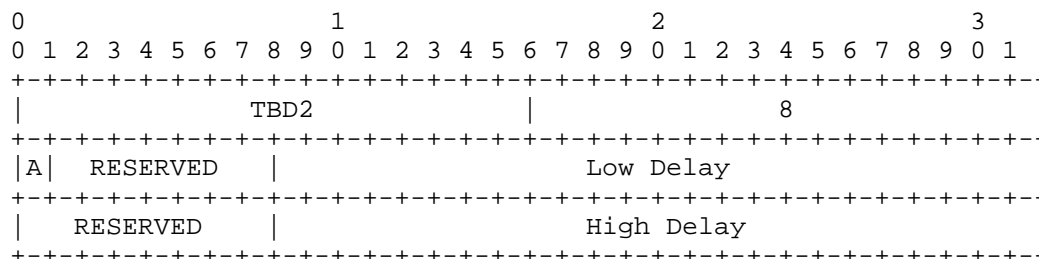
A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link performance.

RESERVED. This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Delay. This 24-bit field carries the average link delay over a configurable interval in micro-seconds, encoded as an integer value. When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger. If there is no value to send (unmeasured and not statically specified), then the sub-TLV should not be sent or be withdrawn.

#### 4.2. Min/Max Unidirectional Link Delay Sub-TLV

This sub-TLV advertises the minimum and maximum delay values between two directly connected IS-IS neighbors. The delay advertised by this sub-TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:



where:

Figure 2

Type: TBA

Length: 8

A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link performance.

RESERVED. This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Low Delay. This 24-bit field carries minimum measured link delay value (in microseconds) over a configurable interval, encoded as an integer value.

High Delay. This 24-bit field carries the maximum measured link delay value (in microseconds) over a configurable interval, encoded as an integer value.

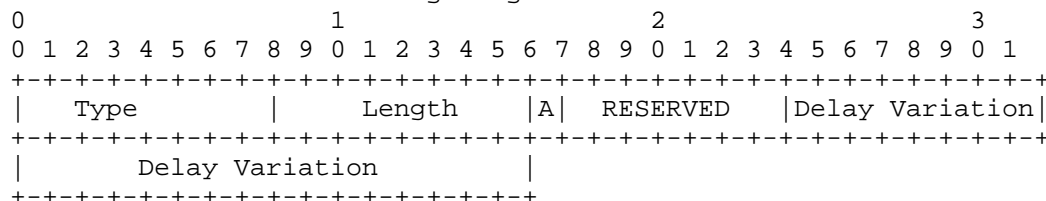
Implementations MAY also permit the configuration of a static (non dynamic) offset value (in microseconds) to be added to the measured delay value, to facilitate the communication of operator specific delay constraints.

It is possible for the high delay and low delay to be the same value.

When the delay value (Low or High) is set to maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger.

#### 4.3. Unidirectional Delay Variation Sub-TLV

This sub-TLV advertises the average link delay variation between two directly connected IS-IS neighbors. The delay variation advertised by this sub-TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:



where:

Figure 3

Type: TBA.

Length: 4.

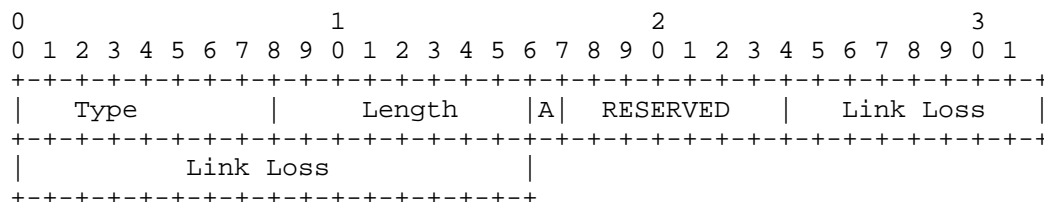
A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link performance.

RESERVED. This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Delay Variation. This 24-bit field carries the average link delay variation over a configurable interval in micro-seconds, encoded as an integer value. When set to 0, it has not been measured. When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger.

#### 4.4. Unidirectional Link Loss Sub-TLV

This sub-TLV advertises the loss (as a packet percentage) between two directly connected IS-IS neighbors. The link loss advertised by this sub-TLV MUST be the packet loss from the local neighbor to the remote one (i.e. the forward path loss). The format of this sub-TLV is shown in the following diagram:



This sub-TLV has a type of TBD3.  
The length is 4.

where:

Type: TBA.

Length: 4.

A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link performance.

A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link performance.

RESERVED. This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

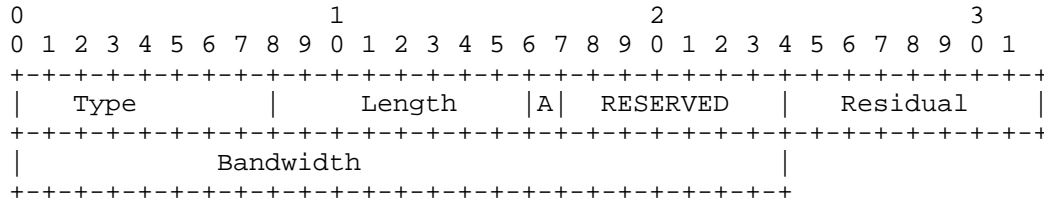
Link Loss. This 24-bit field carries link packet loss as a percentage of the total traffic sent over a configurable interval. The basic unit is 0.000003%, where  $(2^{24} - 2)$  is 50.331642%. This value is the highest packet loss percentage that can be expressed (the assumption being that precision is more important on high speed links than the ability to advertise loss rates greater than this, and that high speed links with over 50% loss are unusable). Therefore, measured values that are larger than the field maximum SHOULD be encoded as the maximum value. When set to a value of all 1s ( $2^{24} - 1$ ), the link packet loss has not been measured.

#### 4.5. Unidirectional Residual Bandwidth Sub-TLV

This TLV advertises the residual bandwidth between two directly connected IS-IS neighbors. The residual bandwidth advertised by this sub-TLV MUST be the residual bandwidth from the system originating



the LSA to its neighbor.



where:

Type: TBA.

Length: 4.

A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link performance.

RESERVED. This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

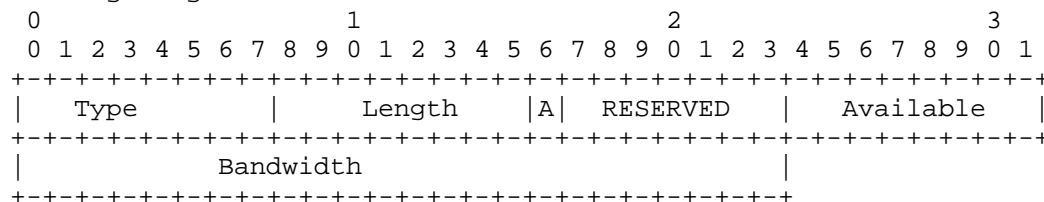
Residual Bandwidth. This field carries the residual bandwidth on a link, forwarding adjacency [RFC4206], or bundled link in IEEE floating point format with units of bytes per second. For a link or forwarding adjacency, residual bandwidth is defined to be Maximum Bandwidth [RFC3630] minus the bandwidth currently allocated to RSVP-TE LSPs. For a bundled link, residual bandwidth is defined to be the sum of the component link residual bandwidths.

The calculation of Residual Bandwidth is different than that of Unreserved Bandwidth [RFC3630]. Residual Bandwidth subtracts tunnel reservations from Maximum Bandwidth (i.e. the link capacity) [RFC3630] and provides an aggregated remainder across QoS classes. Unreserved Bandwidth [RFC3630], on the other hand, is subtracted from the Maximum Reservable Bandwidth (the bandwidth that can theoretically be reserved) [RFC3630] and provides per-QoS-class remainders. Residual Bandwidth and Unreserved Bandwidth [RFC3630] can be used concurrently, and each has a separate use case (e.g. the former can be used for applications like Weighted ECMP while the latter can be used for call admission control).

#### 4.6. Unidirectional Available Bandwidth Sub-TLV

This Sub-TLV advertises the available bandwidth between two directly connected IS-IS neighbors. The available bandwidth advertised by

this sub-TLV MUST be the available bandwidth from the system originating this Sub-TLV. The format of this Sub-TLV is shown in the following diagram:



where:

Figure 4

Type: TBA.

Length: 4.

A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link performance.

RESERVED. This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Available Bandwidth. This field carries the available bandwidth on a link, forwarding adjacency, or bundled link in IEEE floating point format with units of bytes per second. For a link or forwarding adjacency, available bandwidth is defined to be residual bandwidth minus the measured bandwidth used for the actual forwarding of non-RSVP-TE LSP packets. For a bundled link, available bandwidth is defined to be the sum of the component link available bandwidths minus the measured bandwidth used for the actual forwarding of non-RSVP-TE Label Switched Paths packets. For a bundled link, available bandwidth is defined to be the sum of the component link available bandwidths.

## 5. Announcement Thresholds and Filters

The values advertised in all sub-TLVs (except Low/High delay and residual bandwidth) MUST represent an average over a period or be obtained by a filter that is reasonably representative of an average. For example, a rolling average is one such filter.

Low or High delay MAY be the lowest and/or highest measured value over a measurement interval or MAY make use of a filter, or other technique to obtain a reasonable representation of a low and high value representative of the interval with compensation for outliers.

The measurement interval, any filter coefficients, and any advertisement intervals MUST be configurable per sub-TLV.

In addition to the measurement intervals governing re-advertisement, implementations SHOULD provide per sub-TLV configurable accelerated advertisement thresholds, such that:

1. If the measured parameter falls outside a configured upper bound for all but the low delay metric (or lower bound for low-delay metric only) and the advertised sub-TLV is not already outside that bound or,
2. If the difference between the last advertised value and current measured value exceed a configured threshold then,
3. The advertisement is made immediately.
4. For sub-TLVs which include an A-bit (except low/high delay), an additional threshold SHOULD be included corresponding to the threshold for which the performance is considered anomalous (and sub-TLVs with the A-bit are sent). The A-bit is cleared when the sub-TLV's performance has been below (or re-crosses) this threshold for an advertisement interval(s) to permit fail back.

To prevent oscillations, only the high threshold or the low threshold (but not both) may be used to trigger any given sub-TLV that supports both.

Additionally, once outside of the bounds of the threshold, any readvertisement of a measurement within the bounds would remain governed solely by the measurement interval for that sub-TLV.

## 6. Announcement Suppression

When link performance values change by small amounts that fall under thresholds that would cause the announcement of a sub-TLV, implementations SHOULD suppress sub-TLV readvertisement and/or lengthen the period within which they are refreshed.

Only the accelerated advertisement threshold mechanism may shorten the re-advertisement interval. All suppression and re-advertisement interval backoff timer features SHOULD be configurable.

## 7. Network Stability and Announcement Periodicity

Section 5 and Section 6 provide configurable mechanisms to bound the number of re-advertisements. Instability might occur in very large networks if measurement intervals are set low enough to overwhelm the processing of flooded information at some of the routers in the topology. Therefore care SHOULD be taken in setting these values.

Additionally, the default measurement interval for all sub-TLVs SHOULD be 30 seconds.

Announcements MUST also be able to be throttled using configurable inter-update throttle timers. The minimum announcement periodicity is 1 announcement per second. The default value SHOULD be set to 120 seconds.

Implementations SHOULD NOT permit the inter-update timer to be lower than the measurement interval.

Furthermore, it is RECOMMENDED that any underlying performance measurement mechanisms not include any significant buffer delay, any significant buffer induced delay variation, or any significant loss due to buffer overflow or due to active queue management.

## 8. Enabling and Disabling Sub-TLVs

Implementations MUST make it possible to individually enable or disable each sub-TLV based on configuration.

## 9. Static Metric Override

Implementations SHOULD permit the static configuration and/or manual override of dynamic measurements data on a per sub-TLV, per metric basis in order to simplify migrations and to mitigate scenarios where measurements are not possible across an entire network.

## 10. Compatibility

As per [RFC5305], unrecognized Sub-TLVs should be silently ignored

## 11. Security Considerations

This document does not introduce security issues beyond those discussed in [RFC3630] and [RFC5329].

## 12. IANA Considerations

IANA maintains the registry for the sub-TLVs. IS-IS TE Metric Extensions will require one new type code per sub-TLV defined in this document.

## 13. Acknowledgements

The authors would like to recognize Ayman Soliman, Nabil Bitar, David McDysan, Les Ginsberg, Edward Crabbe and Don Fedyk for their contributions.

The authors also recognize Curtis Villamizar for significant comments and direct content collaboration.

## 14. References

### 14.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in

Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

- [RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, September 2008.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

#### 14.2. Informative References

- [I-D.atlas-mpls-te-express-path]  
Atlas, A., Drake, J., Giacalone, S., Ward, D., Previdi, S., and C. Filsfils, "Performance-based Path Selection for Explicitly Routed LSPs", draft-atlas-mpls-te-express-path-02 (work in progress), February 2013.
- [I-D.ietf-alto-protocol]  
Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol", draft-ietf-alto-protocol-16 (work in progress), May 2013.
- [RFC6375] Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.

#### Authors' Addresses

Stefano Previdi (editor)  
Cisco Systems, Inc.  
Via Del Serafico 200  
Rome 00191  
IT

Email: sprevidi@cisco.com

Spencer Giacalone  
Thomson Reuters  
195 Broadway  
New York, NY 10007  
USA

Email: [Spencer.giacalone@thomsonreuters.com](mailto:Spencer.giacalone@thomsonreuters.com)

Dave Ward  
Cisco Systems, Inc.  
3700 Cisco Way  
SAN JOSE, CA 95134  
US

Email: [wardd@cisco.com](mailto:wardd@cisco.com)

John Drake  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
USA

Email: [jdrake@juniper.net](mailto:jdrake@juniper.net)

Alia Atlas  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
USA

Email: [akatlas@juniper.net](mailto:akatlas@juniper.net)

Clarence Filsfils  
Cisco Systems, Inc.  
Brussels  
Belgium

Email: [cfilsfil@cisco.com](mailto:cfilsfil@cisco.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 4, 2014

C. Lin  
H. Zhang  
HangZhou H3C Co. Limited  
V. Manral  
Hewlett-Packard Co.  
July 5, 2013

Simplified Extension of interface Space for IS-IS  
draft-lz-isis-relax-interfaces-limit-00

Abstract

This document describes a simplified method for extending the interface space beyond the 255 interfaces limit. The proposed mechanism does not require any changes to the IS-IS protocol.

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 <http://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 19, 2013.

Copyright Notice

Copyright (c) 2012 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 (<http://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 Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.



## Table of Contents

1. Introduction . . . . .	3
2. Requirements Language . . . . .	3
3. Definition of Commonly Used Terms . . . . .	3
4. Proposed Solution . . . . .	4
5. Security Considerations . . . . .	5
6. IANA Considerations . . . . .	5
7. Acknowledgements . . . . .	5
8. References . . . . .	5
8.1. Normative References . . . . .	5
8.2. Informative References . . . . .	5
Authors' Addresses . . . . .	5

## 1. Introduction

The IS-IS specification has an implicit limit of 255 interfaces, as constrained by the eight-bit Circuit ID field carried in various packets. Moderately clever implementers have realized that the only true constraint is that of 255 LAN interfaces, and for that matter only 255 LAN interfaces for which a system is the Designated IS. This is because the only place that the circuit ID is advertised in LSPs is in the pseudo-node LSP ID.

Implementers have treated the point-to-point circuit ID number space as being independent from that of the LAN interfaces, since these circuit IDs appear only in IIH PDUs and are only used for detection of a change in identity at the other end of a link. More than 255 point-to-point interfaces have been supported by sending the same circuit ID on multiple interfaces. See [RFC5303].

However, that solution suffers from restrictions required to maintain interoperability with systems that do not support the extensions.

This document defines extensions that allow a system to exceed the 255 interfaces limit and do so in a way that has no interoperability issues with systems that do not support the extension.

## 2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

## 3. Definition of Commonly Used Terms

This section provides definitions for terms that are used throughout the text. The terminology is consistent with that used in RFC 5311.

**Originating IS:** A physical IS running the IS-IS protocol. As this document describes a method that allows a single physical IS to run additional interfaces in name of multiple extend ISs, the Originating IS represents the single physical IS.

**Normal system-id:** The system-id of an Originating IS as defined by [IS-IS].

**Additional system-id:** A system-id other than the "Normal system-id", that is assigned by the network administrator to an Extend-IS in order to extend the interface range. The Additional system-id, like the Normal system-id, must be unique throughout the routing area (Level-1) or domain (Level-2), and must be different with the Additional system-id used to extend LSPs in RFC5311.

**Extending IS:** The system, identified by an Additional system-id, for the interfaces beyond 255 to enabled with.

**Local System:** A physical IS running the IS-IS protocol, including Originating IS and Extending ISs.



#### 4. Proposed Solution

The extension proposed to IS-IS to relax the 255 interfaces limit, Extending IS, same defined as Virtual IS in RFC 5311, is introduced to be only used for extending the interfaces. Circuit index is allocated based on IS, one extended IS 255 more interfaces. Here's the diagrams:

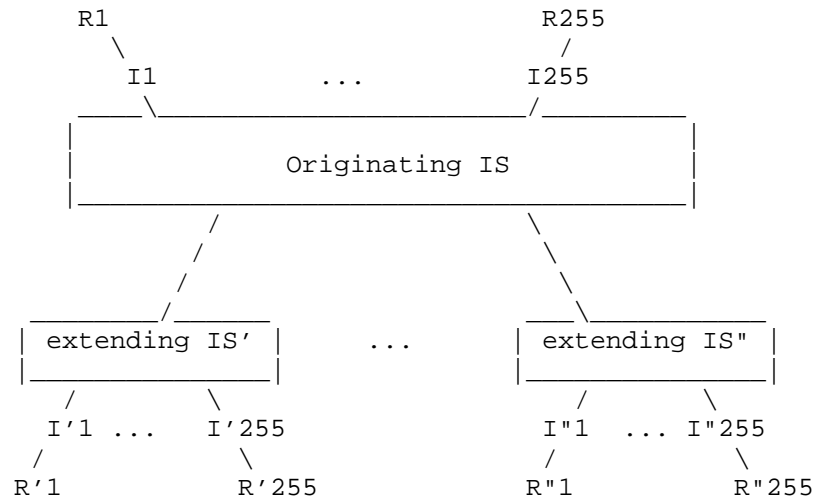


Figure 1: Extend interface space by extending IS

where Rx are remote routers, Ix are interface to remote routers.

When interface space is exceeded, for example I'x and I"x as illustrated in Figure 1 is enabled in the name of extending IS, which means R'x is peered with extending IS' and R"x is peered with extending IS", NOT with Originating IS. The Originating IS MUST specify extending ISS as a neighbor, with metric set to zero. Extending ISSs MUST specify the Originating IS as a neighbor with metric set to zero. The adjacency between Originating IS and Extending ISSs SHOULD be considered as point-to-point.

Hello packet sending

Additional system-id is used for the Hello Packets sending on The interface which is running in Extending IS.

LSP FLOOD

When a new LSP has been received, it must be flooded out some set of the local system's interfaces including Original IS's interfaces and all Extending ISSs's interfaces. Also, a self-originated LSP must be flooded out all the local system's interfaces including Original IS's interfaces and all Extending ISSs's interfaces.

Route Calculation

In local system, LSP Database including all Extending ISSs's LSP should be used in route calculation. All Extending ISSs's interfaces

should be used in nexthop calculation.

## 5. Security Considerations

This document raises no new security issues for IS-IS. IS-IS security may be used to secure the IS-IS messages discussed here. See [RFC5304].

## 6. IANA Considerations

This document has no IANA actions.

## 7. Acknowledgements

## 8. References

### 8.1. Normative References

[IS-IS] ISO, "Intermediate system to Intermediate system routing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:2002, Second Edition.

[RFC5303] D. Katz, R. Saluja, D. Eastlake 3rd,  
"Three-Way Handshake for IS-IS Point-to-Point Adjacencies."

[RFC5311] Hermelin, A., Previdi, S., and M. Shand,  
"Simplified Extension of Link State PDU (LSP) Space for IS-IS", RFC 5311, February 2009.

### 8.2. Informative References

[draft-ietf-isis-wg-255adj-02.txt] T. Przygienda, Maintaining more than 255 circuits in IS-IS

## Authors' Addresses

Changwang Lin  
Oriental Electronic Bld., 2 Chuangye Road,  
Shang-Di Information Industry Base, Hai-Dian District  
Beijing  
P.R.China

Email: linchangwang.04414@h3c.com

Haifeng Zhang  
Hangzhou H3C Co. Limited  
310 Liuhe Road, Zhijiang Science Park  
Hangzhou  
P.R. China

Email: zhanghf@h3c.com

Manral Vishwas  
Hewlett-Packard Co.

USA

Email: vishwas.manral@hp.com

Lin & Zhang & vishwas Expires January 4, 2014

[Page 5]





Network Working Group  
Internet-Draft  
Obsoletes: 1142 (if approved)  
Intended status: Informational  
Expires: January 8, 2014

M. Shand  
  
L. Ginsberg  
Cisco Systems  
July 07, 2013

Reclassification of RFC 1142 to Historic  
draft-shand-rfc1142-to-historic-02

Abstract

This memo reclassifies RFC 1142, OSI IS-IS Intra-domain Routing Protocol, to Historic status. This memo also obsoletes RFC 1142.

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 <http://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 January 8, 2014.

Copyright Notice

Copyright (c) 2013 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 (<http://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 Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## 1. Introduction

IS-IS is the "OSI Intermediate system to Intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", otherwise known as ISO/IEC 10589. It has been extended for use with IP by RFC 1195[RFC1195] and subsequently enhanced by many other RFCs.

RFC 1142[RFC1142] was a republication of ISO DP 10589 originally provided as a service to the Internet community. However, ISO DP 10589 was an ISO "Draft Proposal" which differed in a considerable number of significant respects from the final standardised version published as ISO/IEC 10589[ISO10589-First-Edition], and subsequently revised as ISO/IEC 10589 second edition[ISO10589-Second-Edition]. It has been an ongoing source of confusion when RFC 1142 has been unwittingly quoted or referenced in place of ISO/IEC 10589 itself.

All references to IS-IS should be to ISO/IEC 10589:2002, Second Edition and RFC 1142 is only of historic interest.

## 2. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

## 3. Security Considerations

Reclassifying RFC 1142 has no security considerations.

## 4. References

### 4.1. Normative References

[ISO10589-First-Edition]  
International Organization for Standardization,  
"Intermediate system to Intermediate system intra-domain  
routing information exchange protocol for use in  
conjunction with the protocol for providing the  
connectionless-mode Network Service (ISO 8473)", ISO/  
IEC 10589:1992, First Edition, Nov 1992.

[ISO10589-Second-Edition]

International Organization for Standardization,  
"Intermediate system to Intermediate system intra-domain  
routing information exchange protocol for use in  
conjunction with the protocol for providing the  
connectionless-mode Network Service (ISO 8473)", ISO/  
IEC 10589:2002, Second Edition, Nov 2002.

[RFC1142] Oran, D., "OSI IS-IS Intra-domain Routing Protocol",  
RFC 1142, February 1990.

#### 4.2. Informative References

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and  
dual environments", RFC 1195, December 1990.

#### Authors' Addresses

Mike Shand

Email: [imc.shand@gmail.com](mailto:imc.shand@gmail.com)

Les Ginsberg  
Cisco Systems  
510 McCarthy Blvd.  
Milpitas, CA 95035  
USA

Email: [ginsberg@cisco.com](mailto:ginsberg@cisco.com)



Network Working Group  
Internet Draft  
Category: Standard Track

L. Yong  
W. Hao  
D. Eastlake  
Huawei

Expires: January 2014

July 8, 2013

ISIS Protocol Extension For Building Distribution Trees  
draft-yong-isis-ext-4-distribution-tree-00

Abstract

This document proposes an IS-IS protocol extension for automatically building bi-directional distribution trees to transport multi-destination traffic in an IP network.

Status of this document

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

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 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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on January 8, 2014.

## Copyright Notice

Copyright (c) 2013 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 (<http://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.

## Table of Contents

1. Introduction.....	3
1.1. Conventions used in this document.....	4
2. IS-IS Protocol Extension.....	4
2.1. RTADDR sub-TLV.....	4
2.2. RTADDRV6 sub-TLV.....	6
2.3. The Group Address Sub-TLV.....	7
3. Procedures.....	8
3.1. Distribution Tree Computation.....	8
3.2. Parent Selection.....	8
3.3. Parallel Local Link Selection.....	9
3.4. Tree Selection for a Group.....	10
3.5. Pruning a Distribution Tree for a Group.....	10
3.6. RPF Mechanism.....	10
3.7. Forwarding Using a Pruned Distribution Tree.....	10
3.8. Local Forwarding at Edge Router.....	11
3.9. Distribution Tree across different IGP Levels.....	12
4. Backward Compatibility.....	12
5. Security Considerations.....	12
6. IANA Considerations.....	12
7. Acknowledgements.....	12
8. References.....	12
8.1. Normative References.....	12
8.2. Informative References.....	13

## 1. Introduction

The computer virtualization and cloud applications motivate the DC network virtualization technology [NVO3FRWK]. This technology decouples the end-points networking from the DC physical infrastructure network in terms of address space and configuration [NVO3FRWK].

DC network virtualization solutions are necessary to carry all types of traffic in today's DC physical networks including multi-destination traffic. It is also desirable to use IP network as the DC underlying network for the overlay virtual networks [NVO3FRWK].

IP network technology does not yet support multi-destination traffic forwarding. A variant of Protocol Independent Multicast (PIM) solutions [RFC4601] [RFC5015] are designed to carry IP multicast traffic over IP networks. However the PIM solutions use their own hello protocol and hop-to-hop Join/Leave message so each router does not have global information about the receivers; in the PIM solution, the data packets could be forwarded unnecessarily to the Rendezvous Point (RP), and then get dropped there when no receiver at all or the sender and receivers for a multicast group are on the same branch towards the RP, which consumes network resources. Furthermore PIM solutions maintain a lot of soft-state, have intensive CPU utilization, and have additional convergence time besides IGP's under a failure condition.

Although the PIM protocol is mature and has been deployed in IP networks, applying PIM to the IP network that supports the Network Virtualization can be an extreme challenge [MCASTISS]. For example, VXLAN [VXLAN] solutions requires multicast support in the underlying network to simulate overlay L2 broadcast capability, where every edge node in an overlay virtual network (VN) is a multicast source and receiver. An overlay VN topology may be sparse and dynamic compared to the underlying IP network topology. Also large number of overlay VNs may exist in a DC, which PIM solutions can't scale to.

This document uses extensions to the IS-IS protocol to build a distribution tree for multi-destination traffic transport in an IP network. A router uses Router Capability message to announce the tree root address and the multicast groups associated to the tree. With this information, routers in the IGP can compute rooted distribution trees by using the link state information, i.e. LSDB, and shortest path algorithm. Edge routers include information in their LSPs to announce their multicast group-memberships. Routers perform distribution tree pruning for each multicast group based on

router's group membership announcement. A router forwards the multi-destination traffic along the pruned tree.

In this solution, edge routers use IGMP query messages to inform the attached hosts and the hosts use IGMP report message to response with their interested multicast group(s). The edge routers announce interested multicast groups in their LSPs so they are flooded to whole network.

The benefits of this solution are 1) protocol convergence: use single protocol for both unicast and multicast traffic transport and get the same convergence time for unicast and multicast traffic. 2) multi-destination transport simplification: rely on the LSDB for computing a distribution tree and not run PIM hello protocol. 3) forwarding efficiency: no need to always forward the traffic to the RP; 4) better scalability: no need to maintain heavy PIM soft states. TRILL [RFC6325] has used IS-IS protocol for both single destination and multi-destination packet transport, which proves the protocol capability for doing both.

#### 1.1. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

## 2. IS-IS Protocol Extension

### 2.1. RTADDR sub-TLV

This is the sub-TLV of Router Capability TLV. Each RTADDR sub-TLV contains a root IPv4 address and multicast group addresses that associate to the tree. A router may use multiple RTADDR sub-TLVs to announce multiple root addresses and associated multicast groups with each root. RTADDR sub-TLV format is below.



```

+---+---+---+---+---+
|Type=RTADDR      |                               (1 byte)
+---+---+---+---+---+
|   Length        |                               (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Root IPv4 Address                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| RESV  |          Topology ID  |          (2 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Tree Priority |                               (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|Num of Groups  |                               (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Group Address (1)                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Group Mask (1)                                |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
~                                                                    ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               GROUP Address (N)                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Group Mask (N)                                |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Where:

Type: sub-TLV of Router Capability for RTADDR (TBD)

Length: variable depending on the number of associated groups

Topology ID: This field carries a topology ID [RFC5120] or zero if topologies are not in use.

Root IP Address: IPv4 Address for a root

Tree Priority: high number means higher priority. Zero means no priority.

Num of Groups: the number of group addresses

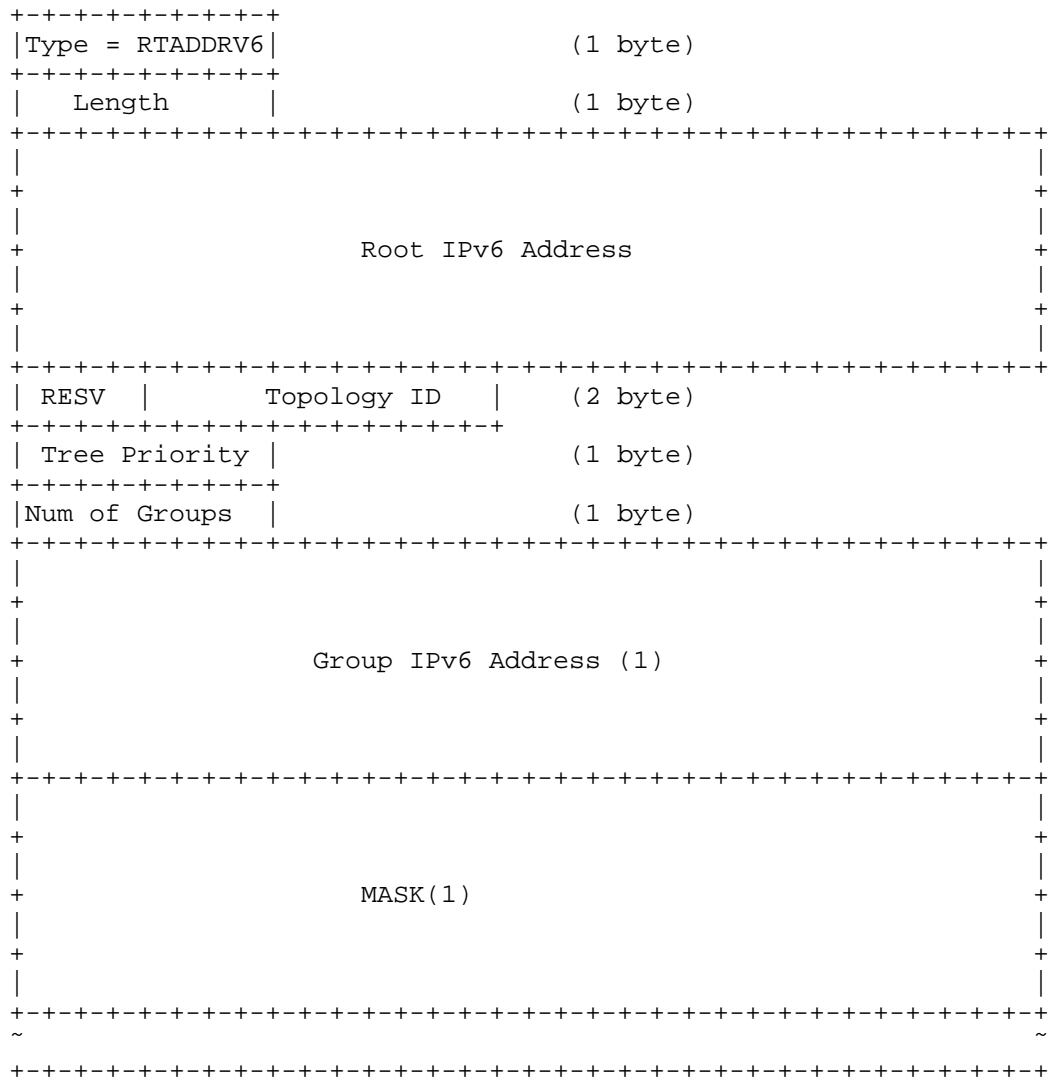
Group Address: IPv4 Address for the group

Group Mask: multicast group range

One router may be the root for multiple trees, each tree associates to a set of multicast groups. In this case, a router encodes multiple RTADDR sub-TLVs to announce root addresses, one for each root, in a router capability TLV. The group address/mask in different sub-TLVs can overlap. See section 3 for detail.

## 2.2. RTADDRV6 sub-TLV

This sub-TLV is used in IPv6 network. It has the same format and usage except that the addresses are in IPv6.



### 2.3. The Group Address Sub-TLV

The Group Address TLV and a set of Group Address sub-TLVs are defined in RFC6326-bis [RFC6326BIS]. The GIP-ADDR and GIPV6-ADDR sub-TLVs are used in this solution. An edge router uses the GIP-ADDR sub-TLV or GIPV6-ADDR to announce its interested multicast groups.

The GIP-ADDR sub-TLV applies to an IPv4 network and GIPV6-ADDR sub-TLV for IPv6 network.

When using a GIP-ADDR or GIPV6-ADDR sub-TLV, the field VLAN-ID MUST set to zero and be ignored. Other field usage remains the same as [RFC6326-BIS]

### 3. Procedures

When an operator selects a router as a distribution tree root, he/she configures the tree root address and associated multicast groups on the router. A tree root address can be an interface address or router loopback address. After the configuration, the router will include a RTADDR sub-TLV, inside a router capability TLV, where the tree root address and multicast groups are specified. If multiple trees are configured on the router, multiple RTADDR sub-TLVs are added in one router capability TLV to specify individual tree roots. For IPv4 network, RTADDR sub-TLV is used. For IPv6, RTADDRV6 sub-TLV is used. Note that the rest of document specifies the processes for an IPv4 network only and the processes for an IPv6 network is the same.

Operator may associate one multicast group to more than one tree for the redundancy purpose and use the tree priority to specify the primary tree preference. Section 3.2 describes the primary tree selection.

#### 3.1. Distribution Tree Computation

Upon receiving RTADDR sub-TLVs, routers track the tree roots and associated multicast groups. When the LSDB stabilizes, routers calculate all rooted trees according to the LSDB and shortest path algorithm.

One multicast group may associate to multiple trees. It is important that all the routers choose the same tree for a multicast group. Section 3.2 and 3.3 describes the tiebreaking rule for primary tree selection for a multicast group and parent selection in case of equal-cost to potential children.

#### 3.2. Parent Selection

It is important, when building a distribution tree, that all routers choose the same links for the tree. Therefore, when there are equal costs from a potential child node to possible parent nodes, all routers need to use the same tiebreakers. It is also desirable to

allow splitting of traffic on as many links as possible in such situations. TRILL [RFC6325] achieves this by defining multiple rooted trees and using the tiebreakers to enable these trees to choose different parents. This draft uses the same tiebreakers as TRILL [RFC6325].

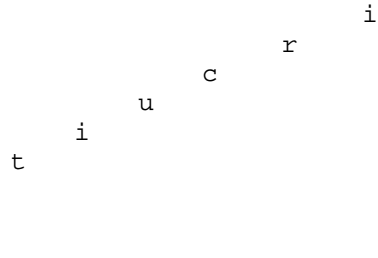
If there are  $k$  distribution trees in the network, when each router computes these trees, the  $k$  trees calculated are ordered and numbered from 0 to  $k-1$  in ascending order according to root IP addresses.

The tiebreaker rule is: When building the tree number  $j$ , remember all possible equal cost parents for router  $N$ . After calculating the entire "tree" (actually, directed graph), for each router  $N$ , if  $N$  has " $p$ " parents, then order the parents in ascending order according to the 7-octet IS-IS ID considered as an unsigned integer, and number them starting at zero. For tree  $j$ , choose  $N$ 's parent as choice  $j \bmod p$ .

### 3.3. Parallel Local Link Selection

If there are parallel links between two routers, say  $R1$  and  $R2$ , these parallel links would be visible to  $R1$  and  $R2$ , but not to other routers. If this bundle of parallel links is included in a tree, it is important for  $R1$  and  $R2$  to decide which link to use; if the  $R1$ - $R2$  link is the branch for multiple trees, it is desirable to split traffic over as many link as possible. However the local link selection for a tree irrelevant to other Routers. Therefore, the tiebreaking algorithm need not be visible to any Routers other than  $R1$  and  $R2$ .

When there are  $L$  parallel links between  $R1$  and  $R2$  and they both are on  $K$  trees.  $L$  links are ordered from 0 to  $L-1$  in ascending order of  $C$



Circuit ID as associated with the adjacency by the router with the highest System ID, and  $K$  trees are ordered from 0 to  $K-1$  in ascending order of root IP addresses. The tiebreaker rule is: for tree  $k$ , select the link as choice  $k \bmod L$ .

Note that if multiple distribution trees are configured in a network or on a router, better load balance among parallel links through the tie-breaking algorithm can be achieved. Otherwise, if there is only one tree is configured, then only one link in parallel links can be used for the corresponding distribution tree. However, calculating and maintaining many trees is resource consuming. Operators need to balance between two.



### 3.4. Tree Selection for a Group

Routers receive one or more possible multicast group-range-to-tree mappings. Each mapping specifies a range of multicast groups. It is possible that a group-range is associated with multiple trees that may have the same or different priority. When a multicast group-range associates with more than one tree, all routers has to select the same tree for the group-range. The tiebreaker rules specified in PIM [RFC4601] are used. They are:

- o Perform longest match on group-range to get a list of trees.
- o Select the tree with highest priority.
- o If only one tree with the highest priority, select the tree for the group-range.
- o If multiple trees are with the highest priority, use the PIM hash function to choose one. PIM hash function is described in section 4.1.1 in RFC4601 [RFC4601].

### 3.5. Pruning a Distribution Tree for a Group

Routers prune the distribution tree for each associated multicast group, i.e. eliminating branches that have no potential downstream receivers. Multi-destination packets SHOULD only be forwarded on branches that are not pruned. The assumption here is that a multicast source is also a multicast receiver but a multicast receiver may not be a multicast source.

Routers prune the trees based on the groups specified in GRADD-TLV from edge routers. Routers maintain a list of adjacency interfaces that are on the pruned tree for a multicast group. Among these interfaces, one interface may be toward the tree-root router and other are toward the egress routers.

### 3.6. RPF Mechanism

For the further study.

### 3.7. Forwarding Using a Pruned Distribution Tree

Forwarding a multi-destination packet follows the pruned tree for the group that the packet belongs to. It is done as follows.

- o The router receives a multi-destination packet with group IP address that does not associated with any tree, the packet MUST be dropped.
- o Else check if the link that the packet arrives on is one of the ports in the pruned distribution tree. If not, the packet MUST be dropped.
- o Else perform RPF checking (section 3.5). If it fails, the packet SHOULD be dropped.
- o Else the packet is forwarded onto all the adjacency interfaces in the list for the group except the interface where the packet receive.

### 3.8. Local Forwarding at Edge Router

Upon receiving a multi-destination packet, besides forwarding it along the pruned tree, an edge router may also need to forward the packet to the local hosts attached to it. This is referred to as local forwarding in this document.

The local group database is needed to keep track of the group membership of the router's directly attached network or host. Each entry in the local group database is a [group, network/host] pair, which indicates that the attached network has one or more hosts belonging to the multicast group. When receiving a multi-destination packet, the edge router forwards the packet to the network/host that match the [group, network/host] pair in the local group database.

The local group database is built through the operation of the IGMPv3 [RFC3376]. When an edge router becomes Designated Router on an attached network, say N1, it starts sending periodic IGMPv3 Host Membership Queries on the network. Hosts then respond with IGMPv3 Host Membership Reports, one for each multicast group to which they belong. Upon receiving a Host Membership Report for a multicast group A, the router updates its local group database by adding/refreshing the entry [Group A, N1]. If at a later time Reports for Group A cease to be heard on the network, the entry is then deleted from the local group database. The Designated Router further sends the LSP message with GRADDR sub-TLV to inform other routers about the group memberships in the local group database. A router MUST ignore Host Membership Reports received on those networks where the router has not been elected Designated Router.



### 3.9. Distribution Tree across different IGP Levels

Coming soon.

## 4. Backward Compatibility

If a router does not support the distribution tree function described in this document, distribution tree computation MUST NOT include this router. This may result the incomplete tree. Operator can build a tunnel between two routers, which allows a single rooted tree to be built. How to build the tunnel is outside scope of this document.

## 5. Security Considerations

Coming soon.

## 6. IANA Considerations

The document requires two new sub-TLVs, RTADDR and RTADDRV6 for the Router Capability TLV in IANA registry.

## 7. Acknowledgements

Authors like to thank Mike McBride and Linda Dunbar for their valuable inputs.

## 8. References

### 8.1. Normative References

[RFC3376] Cain B., etc, ''Internet Group Management Protocol, Version 3'', rfc4604, October 2002

[RFC4601] Fenner, B., etc, ''Protocol Independent multicast -  
- Sparse  
Mode (PIM-SM): Protocol Specification'', rfc4601, August 2006

[RFC5015] Handley, M., etc, ''Bidirectional Protocol Independent  
Multicast (BIDIR-PIM'', rfc5015, October 2007

[RFC6325] Perlman, R., et al, ''Routing Bridges (Rbridges): Base  
Protocol Specification'', RFC6325, July 2011

[RFC6326] Eastlake D, et al, ''      Transparent Interconnection of

Lots of Links (TRILL) Use of IS-IS'', RFC6326, July 2011

## 8.2. Informative References

[MCASTISS] Ghanvani, A., ''Multicast Issues in Networks Using NVO3'', draft-ghanwani-nvo3-mcast-issues-00, work in progress

[NVO3FRWK] Lasserre, M., ''Framework for DC Network Virtualization'', draft-ietf-nvo3-framework-02.txt, work in progress.

[RFC6326BIS] Eastlake, D., etc, ''Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS'', draft-ietf-isis-rfc6326bis-01, work in progress

## Authors' Addresses

Lucy Yong  
Huawei USA  
5340 Legacy Drive  
Plano, TX 75025 USA

Phone: 469-277-5837  
Email: lucy.yong@huawei.com

Weiguo Hao  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012  
China

Phone: +86-25-56623144  
Email: haoweiguo@huawei.com

Donald Eastlake  
Huawei  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
EMail: d3e3e3@gmail.com

