draft-ginsberg-lsr-isis-flooding-scale

Les Ginsberg, Cisco
Peter Psenak , Cisco
Acee Lindem, Cisco

Historical Behavior

(From ISO 10589)

minimumBroadcastLSPTransmissionInterval - the minimum interval between PDU arrivals which can be processed by the slowest Intermediate System on the LAN.

The default value was defined as 33 milliseconds.

NOTE: It was permitted to send multiple LSPs "back-to-back" as a burst, but this was limited to 10 LSPs in a one second period.

This has been broadly interpreted to apply to P2P interfaces as well – even though this was not the intent of ISO 10589

Section 12.1.2.4.3 states:

On point-to-point links the peak rate of arrival is limited only by the speed of the data link and the other traffic flowing on that link.

Network Scale

As the number of nodes in the network increases, the number of LSPs in the LSPDB increases.

As the number of neighbors a given node has increases, the number of interfaces on which the LSPDB needs to be flooded increases.

If the number of LSPs in the LSPDB == 1000, it would take 30+ seconds simply to flood the LSPDB to a single neighbor at 33 LSPs/second.

Convergence

Step 1: Detection

Link state changes, adjacency state changes

Step 2: Advertise

Update and Flood LSPs

Step 3: Run Decision Process

Calculate new SPT, best paths to destinations

Step 4: Update the local forwarding plane

Steps 2, 3, 4 need to be completed on all nodes in the network

Link State Flooding

ISO 10589

7.3.14.3 The Update Process scans the Link State Database for Link State PDUs with SRMflags set. When one is found, provided the timestamp lastSent indicates that it was propagated no more recently than minimumLSPTransmissionInterval, the IS shall

*a) transmit it on **all circuits** with SRMflags set,*

ISO 10589

7.3.15.5 Action on expiration of the minimumLSPTransmissionInterval

An IS shall perform the following action every minimumLSPTransmissionInterval with jitter applied as described in 10.1:

*- For **all point-to-point circuits C** (including non-DA DED circuits and virtual links) transmit all LSPs that have SRMflag set on circuit C*

Link State Flooding(2)

Convergence depends on all nodes in the network having the same LSPDB.

Update Process: Reliably floods on all supported interfaces

Goal of flooding is to achieve LSPDB convergence network-wide AS FAST AS POSSIBLE!!

Pacing of flooding exists to deal with:

- Fairness to other data and control traffic on the same interface
- Limitations on the processing rate of incoming control traffic

Flooding Rate is not intended to be a per interface parameter (despite many existing knobs)

Bandwidth Utilization

LSPs/Second	100 Mb Link1	1 Gb Link
100	1.2%	0.1%
500	6.1%	0.6%
1000	12.1%	1.2%

Per interface Flow Control

Temporary overload due to transient loads on neighbor
(Consistent overload is pathological and represents a fault)

LSPs which are to be flooded are marked per interface (SRM)

LSPs remain marked until acknowledged

Retransmit timer results in periodic retransmissions of unacknowledged LSPs

Based on the size of the “retransmission queue” sender knows neighbor has been unable to process the LSPs sent (and which ones)

This accounts for all reasons (drops before receive queueing, receiver overloaded CPU, etc.)

Example Flow Control Algo

MaxLSPTx = maximum # LSPs transmitted/second/interface

Umax = Maximum Unacknowledged LSP/Interface

Usafe = Safe level of Unacknowledged LSP/Interface

U(i) = # of unacknowledged LSPs previously transmitted/interface

LSPTx(intf) = max # LSPs transmitted/second for a specific interface

1)LSPTx(intf) = MaxLSPTx

2)U(i) >= Umax (this should be logged)

- **only retransmissions of unacknowledged LSPs are performed**
- **LSPTx(intf) = MaxLSPTx/2**

3)For each second U(i) >= Usafe

- **LSPTx(intf) = LSPTx(i)/2**

4)When U <= Usafe

- **LSPTx(intf) = MaxLSPTx**
- **new LSPs may be transmitted**

Packet Prioritization on Receive

PDU Types: (Hellos, LSPs, SNPs)

Hellos often prioritized to avoid adjacency flapping due to transient peak loads

SNPs serve to acknowledge LSP reception

SNPs should be prioritized over LSPs to minimize unnecessary LSP retransmission

Minimizing LSP Generation

Unit of flooding is an LSP numbered (0-255)

LSP contains TLVs which advertise neighbors, prefixes, etc.

By preserving the association of a given TLV with a specific LSP#, the number of LSPs which change as a result of a topology change is minimized. This reduces overall flooding.

A.00-00
- Neighbor 1
- Neighbor 2
- Neighbor 3
- Neighbor 4
- Neighbor 5

A.00-01
- Neighbor 6
- Neighbor 7
- Neighbor 8
- Neighbor 9
- Neighbor 10

A.00-02
- Neighbor 11

A.00-00
- Neighbor 1
- Neighbor 2
- Neighbor 4
- Neighbor 5

A.00-01
- Neighbor 6
- Neighbor 7
- Neighbor 8
- Neighbor 9
- Neighbor 10

A.00-02
- Neighbor 11

Method 1

A.00-00
- Neighbor 1
- Neighbor 2
- Neighbor 4
- Neighbor 5
- Neighbor 6

A.00-01
- Neighbor 7
- Neighbor 8
- Neighbor 9
- Neighbor 10
- Neighbor 11

A.00-02 (Purge)

Method 2

Redundant Flooding

In cases where a network is highly meshed, there can be a significant amount of redundant flooding. Nodes will receive multiple copies of each updated LSP.

There are defined mechanisms which can greatly reduce the redundant flooding. These include:

- Suppress flooding on parallel links to same neighbor
- Mesh Groups ([RFC2973])
- Dynamic Flooding ([I-D.ietf-lsr-dynamic-flooding])

Jumbo Frames

Max LSPBufferSize MUST be consistent on all nodes and must be <= MTU of all interfaces enabled for IS-IS

(IS-IS does not support fragmentation)

Default size of LSPs is 1492.

In networks where Jumbo frames are supported a larger maxLSPBufferSize can be used.

This may reduce the total number of LSPs in the LSPDB.

What should we do…

- Encourage vendors to increase flooding rate.
- Emphasize that flooding rate should NOT vary/interface
- Use Tx based flow control to dampen flooding rate when necessary.
- Prioritize SNP reception over LSP reception
- Minimize LSP Generation by preserving TLV/LSP association
- Reduce redundant flooding
- Use Jumbo Frames where possible