draft-ginsberg-lsr-isis-flooding-scale

Les Ginsberg, Cisco
Peter Psenak , Cisco
Marek Karasek, Cisco
Acee Lindem, Cisco
Tony Przygienda, Juniper

Significant Changes

Example algorithm removed

Algorithm is local matter – does not impact interoperability

Replaced with Guidelines => Requirements any solution must meet

No intent/requirement to standardize an algorithm

Algorithm Guidelines

- Flooding burst durations are not long-lived
  - 2000 LSPs/300 per sec is ~7 seconds
- Receiver performance may be affected by transient conditions
- Faster recovery requires minimizing retransmissions =>
  - Response time in small number of seconds (< 5)
- Aggressive slowdown / Less aggressive speedup
- Must work with enhanced nodes and legacy nodes
  - Receiver may ack quickly or slowly
  - Flooding optimizations? (Parallel link suppression, dynamic flooding)
  - Receiver may/may not implement optimized packet priority

POC Algorithm Overview

Tracks rate of transmissions vs rate of acknowledgments over a multi-second history

Configurable Parameters

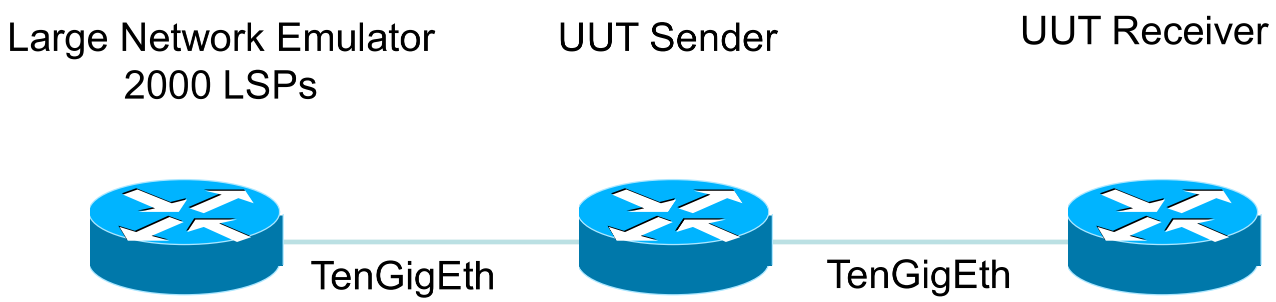
- Maximum LSP Transmission Rate (per node) (LSPTxMax)
- Receiver ACK Delay in ms (per neighbor)

Incorporates expected ACK delay

Agnostic to reason for delay (Tx loss, receiver input queue loss, punt path performance, CPU contention, …)

LSPTxRate is the current active flooding rate/second

Setup and Test procedure



Test procedure:

- Reset UUT Receiver and measure time to download 2000 LSPs from UUT Sender over P2P interface

LSP Transmission Bursts

Simplest scheme is to send one LSP every 1/LSPTxMax ms

At higher transmission rates it is unrealistic to expect scheduling at this resolution

Burst size is adjusted based on LSPTxRate with the expectation that:

- Transmitter will only be scheduled a limited number of times/second
- Some scheduling delays may occur – therefore we may need to “catch up”

We refer to this as “Optimized”

Baseline flooding

Test Procedure

- 2000 LSPs
- Receiver Unlimited
- No Tx adjustment

Flooding	Start/End LSPTxRate	LSPTxMax	Time [ms]	Retrans
Base	33/33	NA	66528	0
Base	333/333	NA	7432	0
Optimized	300/300	300	6324	0
Optimized	1000/1000	1000	1768	0
Optimized	2000/2000	2000	1092	0
Optimized	5000/5000	5000	832	0

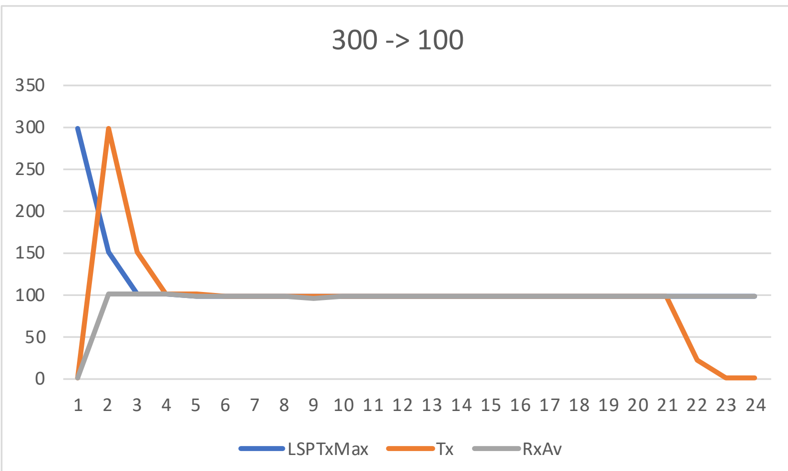
SlowingDown

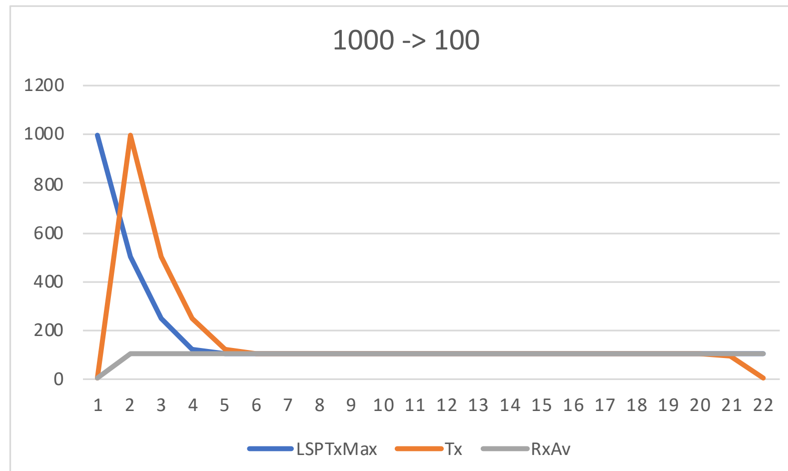
Test procedure:

- 2000 LSPs
- Receiver’s capability changed to 100 LSP/sec
- Sender detects lagging acks and adjusts rate

Flooding	Start/End LSPTxRate	LSPTxMax	Time [ms]	Retrans
Base	33/33	NA	66268	0
Base	333/333	NA	20988	2439 (122%)
Optimized	300/100	300	20076	257 (13%)
Optimized	1000/100	1000	19456	1475 (74%)

SlowingDown





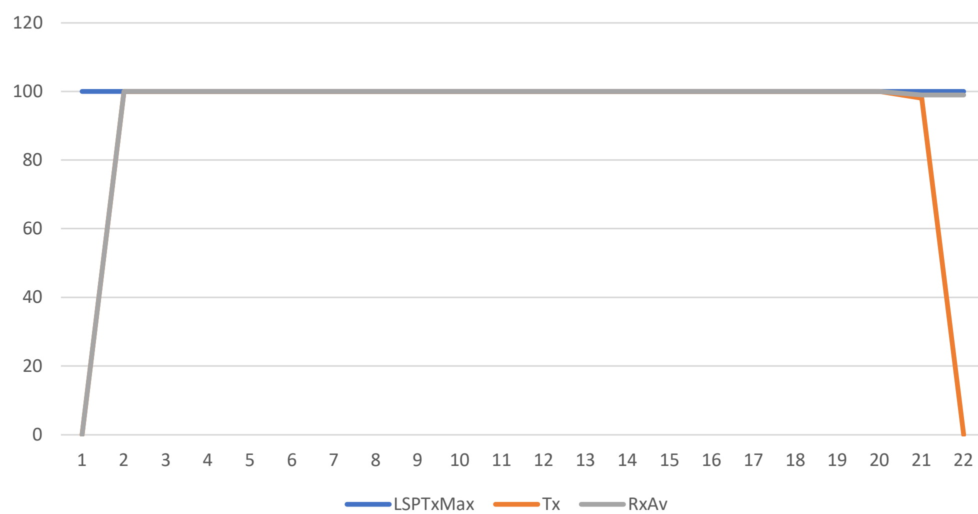
Steady state, receiver affected

Test procedure:

- 2000 LSPs
- Receiver's capability is 100 LSP/sec
- Sender no adjustment required

Flooding	Start/End LSPTxRate	LSPTxMax	Time [ms]	Retrans
Base	33/33	NA	66268	0
Base	333/333	NA	20988	2439 (122%)
Optimized	100/100	300	19444	0
Optimized	100/100	1000	19488	0

Steady state, receiver affected



SpeedingUp

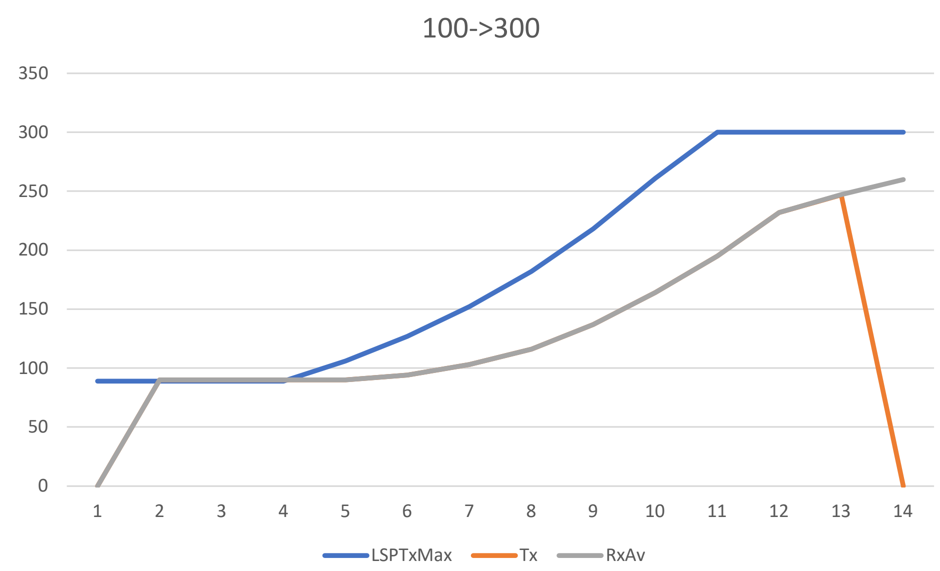
Test procedure:

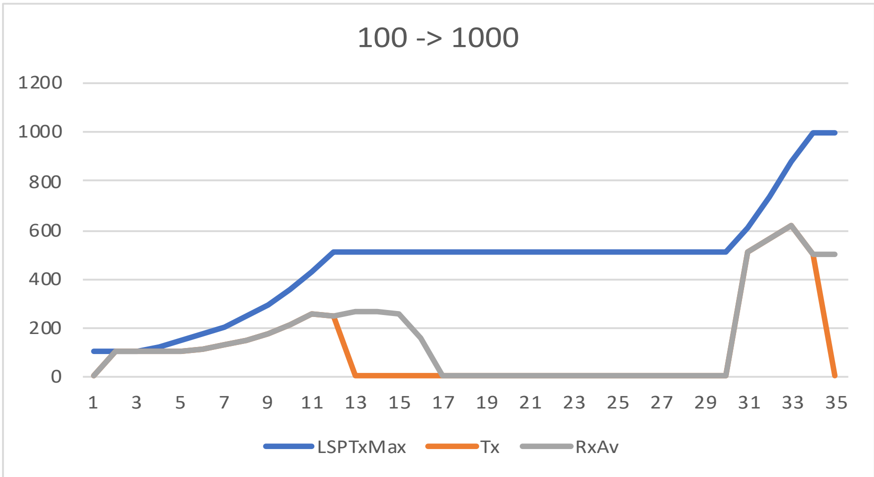
- 2000 LSPs
- Receiver's capability changed from 100 LSP/sec to no limit
- Sender has to detect change and adjust rate

Flooding	Start/End LSPTxRate	LSPTxMax	Time [ms]	Retrans
Base	33/33	NA	66528	0
Base	333/333	NA	7432	0
Optimized	100/300	300	11388	0
Optimized	100/1000	1000	10200/3072**	0

** Multiple burst needed

SpeedingUp


100->300
350
300
250
200
150
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
LSPTxMax
Tx
RxAv


100 -> 1000
1200
1000
800
600
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
LSPTxMax
Tx
RxAv

Some Discussion Points

Issues w static controls

Determination of state at the receiver

Signaling in real time

Comparisons to TCP

Issues with Static Controls

Receive Performance may be impacted by:

- Number of neighbors
- # of nodes in the network
- Flooding optimizations supported (mesh groups, parallel neighbor suppression, dynamic flooding) by each neighbor
- Other protocols (BGP, BFD, OAM, link PM)
- Link bandwidth
- Hardware speed/memory
- SRLG deployment
- …

How are all of these variables accounted for if a static value is used?

Determination of State at the Receiver

Platform implementations are not all alike. Some combination of:

- Policed Input queue
  - Per Line card (not per interface)
  - May be per protocol or combine multiple protocols (“all routing”)
- Punt queue
  - May be per input queue or combine many input queues
- Control plane input queue (multiple line cards)
- IS-IS Input queue (IIHs, LSPs, SNPs) Multiple interfaces/Line Cards
  - Distribution to specific Instance
  - Separation of PDU types (Prioritization)

Receiver based detection does not account for Tx drops/corruption

Every stage has queue limits, interaction with other activities, CPU

Signaling In Real Time

During a burst both transmitter and receiver are busy

Nodes act as both transmitter and receiver simultaneously

Hellos and SNPs are unreliable – may be dropped

Signaling delays will increase likelihood of retransmissions

Comparisons to TCP

TCP	IS-IS
Byte Stream	Packet Based
Data from a single source	Data from multiple sources
Ordered delivery	Unordered delivery
Single independent data stream	Multiple interface streams
Resources managed by control plane	Resources dependent on dataplane