LSR Working Group
Internet-Draft
Intended status: Informational
Expires: August 15, 2019

S. Hares
Huawei

February 11, 2019

IPRAN Grid-Ring IGP convergence problems
draft-hares-lsr-grid-ring-convergence-00.txt

Abstract

This draft describes problems with IGP convergence time in some IPRAN networks that use a physical topology of grid backbones that connect rings of routers. Part of these IPRAN network topologies exist in data centers with sufficient power and interconnections, but some network equipment sits in remote sites impacted by power loss. In some geographic areas these remote sites may be subject to rolling blackouts. These rolling power blackouts could cause multiple simultaneous node and link failures. In these remote networks with blackouts, it is often critical that the IPRAN phone network re-converge quickly.

The IGP running in these networks may run in a single level of the IGP. This document seeks to briefly describe these problems to determine if the emerging IGP technologies (flexible algorithms, dynamic flooding, layers of hierarchy in IGPs) can be applied to help reduce convergence times. It also seeks to determine if the improvements of these algorithms or the IP-Fast re-route algorithms are thwarted by the failure of multiple link and nodes.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 15, 2019.

Table of Contents

1. Introduction

This draft describes problems with IGP convergence time in some IPRAN networks. The physical topologies of these IPRAN networks combine a grid backbone topology with a ring topology to support phone networks (see figure 1). Routers are attached to the rings that route traffic from the IPRAN devices (see figure 2). Each of the rings is attached to two grid nodes in order to provide redundancy. All of the routers in the IPRAN ring-grid network topology run a single IGP (IS-IS).

Some current deployments attach 10-30 routers per ring with a 20 by 20 grid of routers. In these deployments, a grid of 400 routers supports between 10,000 - 15,000 routers on the IPRAN rings.

Convergence of the IGP after a single link failure on one ring router is over 1 second for these topologies. The desired convergence time for a single link failure is less than 200 ms for phone networks.

Initial convergence of the full network may take on the order of minutes.

Part of these IPRAN network topologies exist in data centers with sufficient power and interconnections, but some network equipment sits in remote sites impacted by power loss. In some geographic regions, these remote sites may be subject to rolling blackouts. These rolling power blackouts could cause multiple simultaneous link or node failures. In these remote networks with blackouts, it is often critical that the IPRAN network converge quickly to restore what mobile phone service it can. Keeping isolated portions of the network working may be critical to keep some phone service working. Converging the isolated portions back into the network when repairs are made also causes further disruptions.

Due to the topologies of the IPRAN network, this document examines how the flooding of IGP informations causes the longer IGP convergence times for single links. The potential multiple simultaneous link and node failures mean that the assumptions in most IGP and fast IP-Route algorithms do not apply.

This document seeks to briefly describe these problems to determine if the following emerging IGP technologies an be applied to solve the convergence problem:

flexible algorithms [I-D.ietf-lsr-flex-algo],

dynamic flooding [I-D.li-lsr-dynamic-flooding],

Level 1 abstraction for ISIS [I-D.li-area-abstraction]

hierarchical IS-IS [I-D.li-hierarchical-isis]

2. IPRAN Topologies

A bit of background on the IPRan sizes.

Grid topologies can be any size of square topologies. Figure 1 shows a 3 router by 3 router topologies (3x3) with 9 nodes). Other sizes could be 10 routers by 10 routers (10X10) with 100 nodes, 15 routers by 15 routers (15X15) with 225 routers, or 50 nodes by 50 nodes (20X20) with 400 routers. A grid with network topology of a 100x100 grid would have 10,000 gird-routers (grid only and ring-grid). Suppose that for every two grid nodes, 3 rings would be attached and

on each ring there are 50 nodes. This topology would result in 750,000 ring routers plus 10,000 grid routers. The size of this topology rivals data center sizes, but the IPRAN network does not have the infrastructure advantages of the data center.

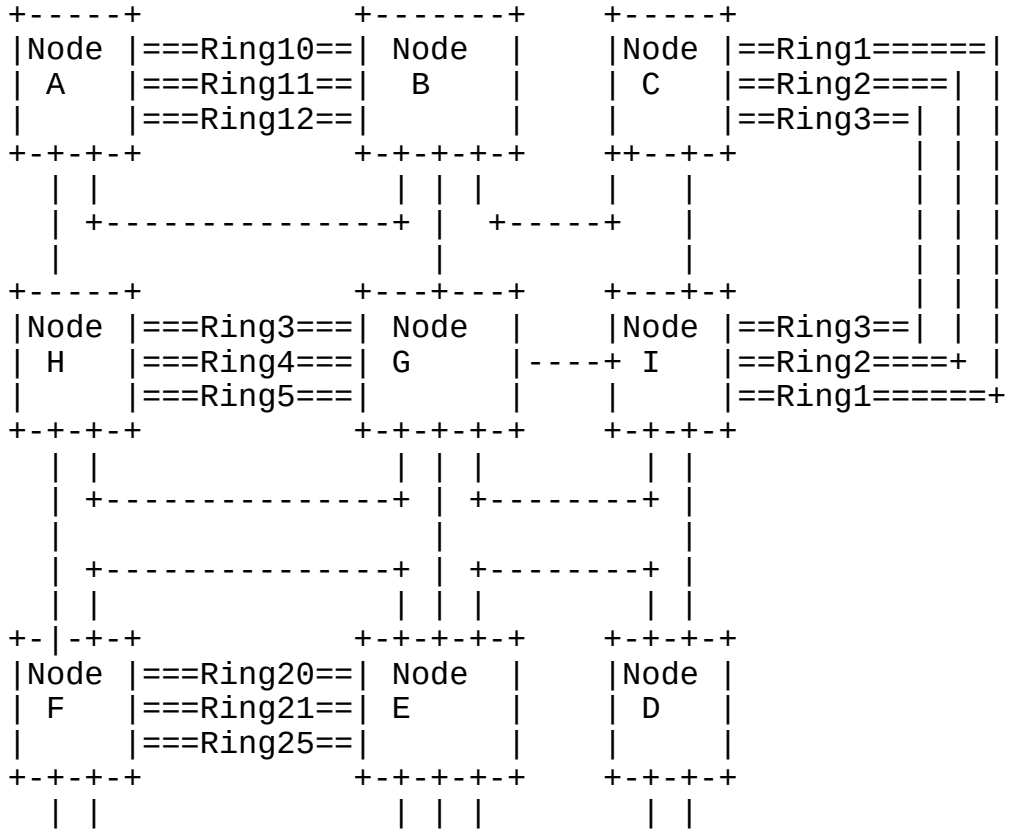

Figure 1

Figure 1: Example IPRAN Grid-Ring Topology

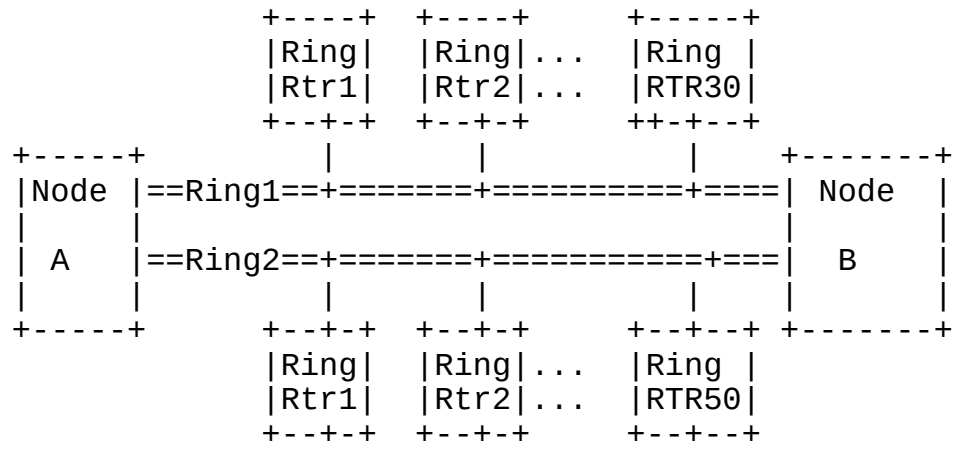

Figure 2

Figure 2: Example IPRAN Ring Topology

One characteristics of a grid is that a basic 3X3 square can be overlaid on most grids. Figure 3 shows a 10 by 10 grid with 3 by 3. Notice that the grid squares overlaid on column 10 and row 10 form partial squares (see GS4, GS8, GS12, GS13, GS14, GS15, and GS16).

If additional connections were made most of column 10 could form a single Grid (GS4, GS8, and GS12), and most of row 10 could form a single grid (GS13, GS14, and GS15). Alternatively, with a single connection, GS16 could merge with GS15 to form a partial grid of 4 nodes.

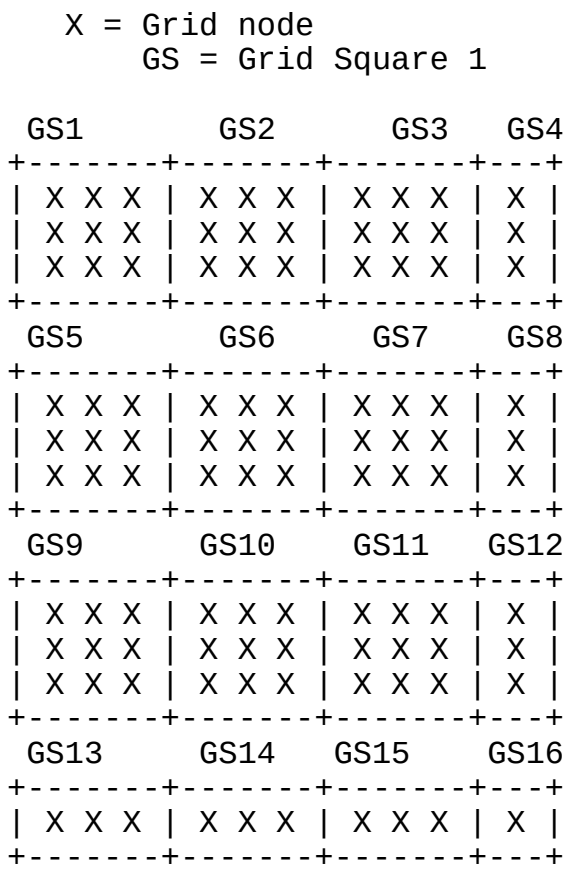

Figure 3

Figure 3: Overlaying Grid Squares on IPRAN Grid

The grid topology is currently one flat IGP. However, logical grid squares could form Level 1 areas within the IGP. If one desired to create an L1 Area abstraction such as defined [I-D.li-area-abstraction], then the grid-square areas could be created as L1 areas and connected by 1-3 links to adjacent areas. Figure 4 shows a logical topology for grid squares 1-8 from figure 2.

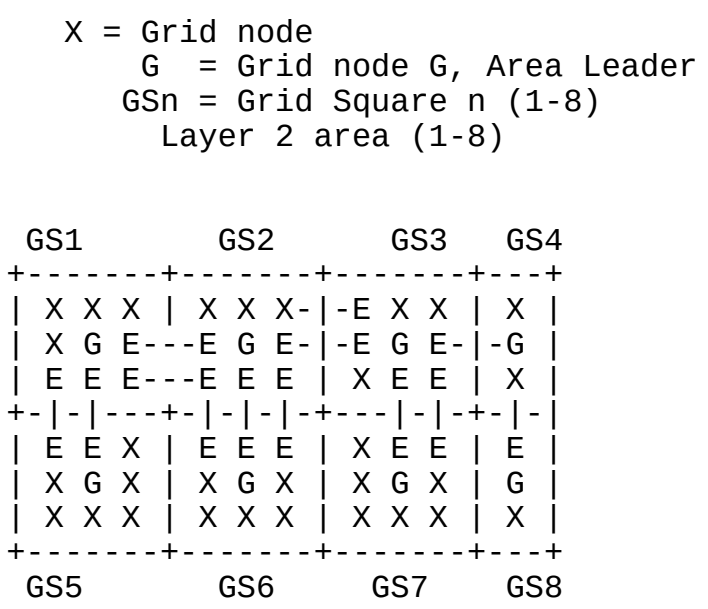

Figure 4

Figure 4: Grid Squares Area Leaders and Area Edge Nodes

3. Definitions

This section provides definitions for nodes within the IPRAN routing infrastructure:

ring router: a routing device only attach to a ring in an IPRAN topology which routes end-system information

ring-grid router routing device attached to ring and the grid topology

grid router: a routing device which is only attached to the IPRAN Grid network

pseudo-node for grid area: a pseudo-node which summarizes for an IGP a grid area at one level for a higher level.

3.1. 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].

4. Problem detection using theoretical IGP Convergence

Theoretical "best" convergence times for a single link failure on ring depths of 30 nodes suggests the flooding time is a major component for the flat IGP. Estimates of theoretical best convergence times may be based on set of equations shown in figure 5. These equations show how network convergence is the maximum time for the information on a link change (down (failure) or up) to spread to all routers in the network. The change travels along a pathway of routers from the change to any particular router. Therefore, convergence is really topology dependent on the convergence time in each router and the pathways.

The theoretical convergence equations in figure 5 include updating the RIB/FIB (Trib) and forwarding elements (Tdd). Some IGPS may forward IGP traffic after calculating the SPF (Tspf)and updating the RIB/FIB, but before updating the FIB line cards (Tdd). In this case, these factors would be zero in the equation.

If several factors are zero or a constant, then the convergence may be determined by one element in the equation that dominates the convergence per node.

```
            CT-Node = Td + To + Tf + Tspf + Trib + Tdd

                CT-Node = Node convergence time
                Td = link failure detection time
                 (or link up detection time)
                To = time to originate LSP
                     describing the new topology

                Tf = Time to flood the change
                     from this node to other nodes
                         that must perform a flood update

                Tspf = Time for shortest path calculation

                Trib = Time to update the RIB and FIB

                Tdd = time to distribute the FIB to line cards

                CT-path(i) = sum [CT-Node(j), .. CT-Node-(n))
                                        where i = path through network
                                              j = nodes on path (1..n)

                CTnetwork = maximum (CT-path(i))
                            where i = 0..n paths
                      Figure 5
```

Figure 5: Convergence equations

[My first experience with an equation like this was Cengiz Alaettinoglu research in IGP around 2000 at NANOG. (Please let me know if you have a good scholarly reference or presentation reference for these equations).]

4.1. Equation applied to Data Center IGP Convergence

Some early SPF implementations were slow with large IGP topologies. In this case, IGP's SPF calculations dominates the convergence time for all nodes. Thus the Tspf dominates the time for each network path and the entire networks convergence time. One might summarize the convergence as:

CT-network = (Tspf + constant) * maximum path-length

The maximum path length is often called the network depth. The network depth of a full mesh network is 1. The network depth of a dense mesh fat tree in a data center with 3 levels (top of rack, aggregate, spine) is 3. If Tspf dominates the calculation then:

CT-network = (Tspf + constant) * 3

Centralized algorithms might improve convergence time if Tspf is the main factor. Rather than using routers with typically low calculation power, centralized devices could be optimized for the calculation. If the difference in network depth of sending the information end-to-end on any network path and sending it to the centralized processor and back is minimal, then centralized processing may be more effective.

If flooding (Tf) dominates the per node convergence, the equation is:

CT-network = (Tf + constant) * 3

Many of the authors of the IGP flooding enhancements to reduce the data flooded understand that the flooding depends on the maximum pathway length for pathways in the IGP graph. (see 802.1aq [I-D.allan-lsr-flooding-algorithm], Li et al. [I-D.li-lsr-dynamic-flooding], Shen, Ginsberg, and Thyamagundalu [I-D.shen-isis-spine-leaf-ext]). Others mention creating a sub-graph of the entire topology to reduce the flooding traffic and reduce convergence time (Chen et al. [I-D.cc-ospf-flooding-reduction]).

Some of the IGP flooding reductions are identifying and limiting the number of global pathways without mentioning their concern for length. (see Chunduri and Eckert [I-D.ce-lsr-ppr-graph]).

The point behind this is that each algorithm has a set of goals. Those goals may impact other things that impact convergence. Some questions one can ask are:

- Does the algorithm seek to reduced data flooded and stored?
- Does the algorithm seek to reduce convergence time?
- If the algorithm tries to both reduce the data flooded and stored, what trade-offs did the algorithm make?
- what is the impact of the topology?

If one looks to adapt the algorithms developed for the dense interconnections of the 3 tier data center to the IPRAN Grid-ring network structure, these questions are important.

4.2. Flooding Problem on the Rings

Putting 30 or 50 ring routers on a ring may help operational costs. Within a city the higher density of rings may allow more cells for the phone. In the rural networks, it may allow the cells to be deployed over a larger physical area.

Every router one puts on a ring increases the network depth of the path through a fully operational ring or a partitioned ring that is still connected to the network. The network depth of a ring is

network depth = (n-ring-nodes + n-grid-ring)/2

where

n-ring-nodes = 30 to 50 nodes

n-grid-nodes = 2 nodes

A partitioned ring may have the full network depth if the link between a grid-router and the ring router attached to it fails.

This flooding time is only for the on-ring path. For a network path that involves the link failure of a ring router link the pathway is:

```
network depth = depth(failed-ring) +
                depth(grid) +
                depth(remote-ring)

depth(failed-ring)= network depth of ring with
                    failed link.

depth(grid) = network depth of pathway
              through Grid

depth(remote-ring) = network depth of pathway
                     through remote ring
```

Figure 6

Figure 6: Convergence equations

The worse case IGP convergence time combines the worse case for each of these network depths.

4.3. Flooding problem on the grid

The network depth of grid topologies grows as the size of the grid grows from 3X3 to 10X10 to 100X100. The network depth of the best case pathway through the grid is a single hop as it is on the same ring-grid router. The worse case path is the one from x1 to X2 in figure 7. A network pathway that goes from x1 to X2 by using routers in the following grid squares: pathway of GS2, GS3, GS4, GS8, GS12 could take 19 hops.

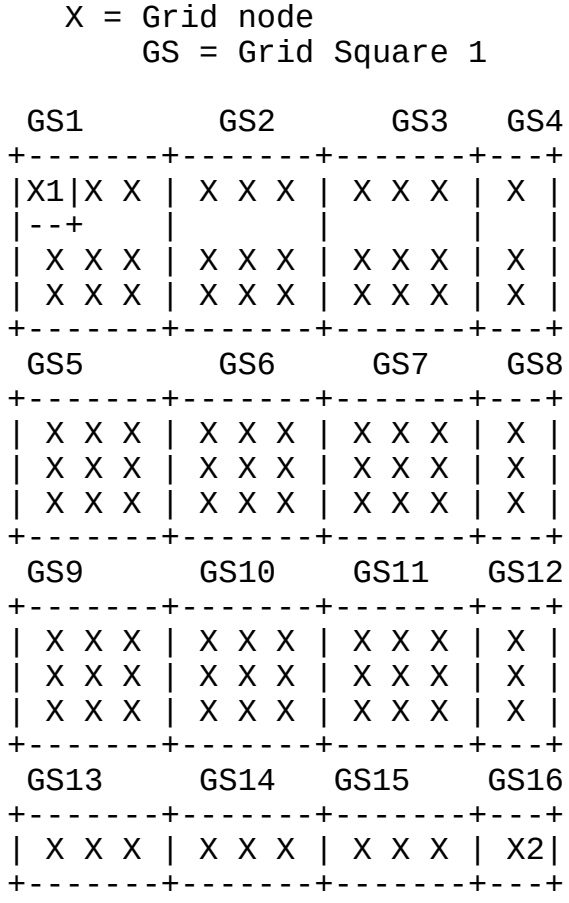

Figure 6

Figure 7: Worse Case for 10X10 Grid

5. Multiple simultaneous link and node failures

Part of these IPRAN network topologies exist in data centers with power and connective, but some do not. Ring routers are more likely

to be at remote sites where power loss can occur. However, some ring-grid routers or grid-only routers may be in remote sites.

In some geographic locations, power losses can be rolling blackouts that cause multiple link and node outages during the failure. These outages may be unpredictable due to weather or natural disasters, or semi-predictable due to brownouts. Upon attempts to restore power, the restorations may have mixed combinations of links and nodes up. Multiple simultaneous link and node failures may impact both the ring topologies and the grid topologies in the IPRAN network.

For simplicity of this discussion, I will present the node outages as the outages of all links. A node outage may take far longer if rebooting the routers or reconfiguring spare ring routers takes a long time. For this initial pass on this document, I will simply treat node outages as failure of all links for a time period that clear all valid paths.

Most fast re-route technology such LFA [RFC5286] or MRT [RFC7812] set-up IP backup paths to route around a single link or node failure. In fact, the MRT architecture explicitly states that

> "MRT-FRR creates two alternative forwarding trees that ... are maximally diverse from one another, providing link and node protect for 100% of paths and failures as long as the failures do not cut the network into multiple pieces"

5.1. Multiple link failures on Ring

Ring routers may be located at sites that may lose connection to the ring or to a grid-ring router. A single link failure may cut the ring, but leave all nodes attached if the failed link is between one of the ring routers (single on ring) or between the a ring-grid routers and a ring router.

Multiple link failures on a ring will cause the ring to partition, isolating some nodes. One way to handle this is to ignore the convergence on the partitioned rings. Since local phone service during these outages may be useful, it may be important for the IGPs on the isolated portions of the rings to continue to operate. During the restoration phase, additional links may appear to go up and down as the partitions heal. Several isolated portions of the ring may be restored to form a larger isolated portion of the ring. Eventually, the isolated parts should reconnect to a fully connected ring.

5.2. Multiple link failures on Grid

Multiple link failures can occur on the ring-grid routers or grid-only routers. These failures may dramatically impact the data forwarding pathways through the grid and the flooding pathways. Fast convergence of the grid depends on an algorithm tuned for the grid topologies.

The failures on the grid can impact different parts of the IGP convergence algorithm.

6. Problem with Flat ISIS areas

Abstraction in an IGP can provide a logical means to scale IGPs. Creating 2 levels of topology in the IPRAN network based on ISIS areas could reduce the network depth and the the size of the topology database in level devices.

However, as Li states in [I-D.li-area-abstraction] the ISIS concepts work well if:

- "the Level 1 area is tangential to the Level 2 area", or
- if "there are a number of routers in both level 1 and level 2 and they are adjacent".

However it does not work well if Level 1 area needs to provide transit for level 2 traffic.

Suppose all ring routers networks were placed in level 1 areas, and grid-only routers were in level 2. The ring-grid routers are in both level 1 and 2. This reduces the current topology to a topology similar to the spine-leaf topology. While this reduces the amount of LSP stored, it may not significantly improve IGP convergence. The flooding topology must be examined to determine the maximum network depth, and the router operations must be examined to determine the per IGP flooding time.

It also restricts repair of an L2 Grid path via a L1 Ring. This repair might be necessary in the multi-failure scenario.

The area abstraction described in [I-D.li-area-abstraction] could be used to remove these restrictions.

Additional levels of hierarchy described by Li in [I-D.li-hierarchical-isis] could be utilized in the grid to allow additional levels of abstractions. These levels could reduce the network depth that IGP flooding passes through.

One difficulty with using abstraction provided by areas and levels is the configuration of the appropriate network topology with multiple levels, and reconfigurations of these levels. To be effective for 100X100 grids, it would be beneficial to automate the configuration of areas.

7. Problems with Dense Flooding Algorithm

- spine-leaves - rings may be leaves, but grid is not spine-leave topology.
- sparse link flooding - Grid may have too little or too much. Top priority is fast convergence not reduced load of LSPF, but fast convergence.
- preferred path graph - goal is preferred path reduction of the number of preferred paths through network. Fast re-route also sets up paths. The preferred path graph needs to be carefully integrated with any fast reroute scheme.
- flooding of 802.1aq - is designed for dense mesh.
  - The algorithm’s two tree structure of 802.1aq provide complete coverage in the presence of a single link failure while constraining the number of LSAs.
  - Both trees in the two structure have the same convergence properties in the IPRAN ring and grid.

8. References

8.1. Normative References:

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

8.2. Informative References

[I-D.allan-lsr-flooding-algorithm] Allan, D., "A Distributed Algorithm for Constrained Flooding of IGP Advertisements", draft-allan-lsr-flooding-algorithm-00 (work in progress), October 2018.

[I-D.cc-ospf-flooding-reduction]
Chen, H., Cheng, D., Toy, M., and Y. Yang, "LS Flooding Reduction", draft-cc-ospf-flooding-reduction-04 (work in progress), September 2018.

[I-D.ce-lsr-ppr-graph]
Chunduri, U. and T. Eckert, "Preferred Path Route Graph Structure", draft-ce-lsr-ppr-graph-01 (work in progress), October 2018.

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-01 (work in progress), November 2018.

[I-D.li-area-abstraction]
Li, T., "Level 1 Area Abstraction for IS-IS", draft-li-area-abstraction-00 (work in progress), June 2018.

[I-D.li-hierarchical-isis]
Li, T., "Hierarchical IS-IS", draft-li-hierarchical-isis-00 (work in progress), June 2018.

[I-D.li-lsr-dynamic-flooding]
Li, T., Psenak, P., Ginsberg, L., Przygienda, T., Cooper, D., Jalil, L., and S. Dontula, "Dynamic Flooding on Dense Graphs", draft-li-lsr-dynamic-flooding-02 (work in progress), December 2018.

[I-D.shen-isis-spine-leaf-ext]
Shen, N., Ginsberg, L., and S. Thyamagundalu, "IS-IS Routing for Spine-Leaf Topology", draft-shen-isis-spine-leaf-ext-07 (work in progress), October 2018.

[RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <https://www.rfc-editor.org/info/rfc5286>.

[RFC7812] Atlas, A., Bowers, C., and G. Enyedi, "An Architecture for IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR)", RFC 7812, DOI 10.17487/RFC7812, June 2016, <https://www.rfc-editor.org/info/rfc7812>.

Author’s Address

Susan Hares
Huawei
Saline
US

Email: shares@ndzh.com