rtgwg
Internet-Draft
Intended status: Standards Track
Expires: 15 August 2025

Y. Wang
China Telecom
C. Lin
New H3C Technologies
A. Wang
China Telecom

15 February 2025

IGP Prefix Independent Convergence
draft-wang-rtgwg-igp-pic-02

Abstract

In many cases, a large number of routes can be reached by multiple next hops. When a link fails, route calculation needs to be performed and a new reachable path needs to be calculated. If all routes are re-calculated and refreshed, the calculation time increases linearly as the number of routes increases, resulting in a long time for route convergence. This document describes an architecture where the number of prefixes is independent. This architecture allows routes to be recalculated when paths change, regardless of the number of IGP routes.

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 15 August 2025.

Copyright Notice

Table of Contents

1. Introduction

In modern networks, it is not uncommon to have a prefix reachable via multiple paths. When the primary link fails, routes must be converged again as soon as possible.

For the OSPF route calculation process, see [RFC2328].

1) Calculate the shortest path (spf) tree from the root node to all routing nodes based on the link status.

2) The cost of each prefix is calculated according to the distance between the root node and the router node in the shortest path tree.

When the number of prefixes increases, route convergence slows down.

This document proposes a hierarchical shared forwarding chain organization that allows traffic to be restored in time periods independent of prefix number. This technology relies on internal router behavior that is completely transparent to operators and can be deployed and enabled progressively without operator intervention.

2. 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 [RFC2119] .

3. Terminology

The following terms are defined in this draft:

* IGP prefix: A prefix P/m (of any AFI/SAFI) that is learnt via an Interior Gateway Protocol, such as OSPF and ISIS, has a path for. The prefix may be learnt directly through the IGP or redistributed from other protocol(s)

* OSPF ABR Node: OSPF Area Boundary Router, A OSPF router between multiple areas

* OSPF ASBR Node: OSPF AS boundary router, A OSPF router that exchanges routing information with routers in other AS

* OSPF Node: A node is associated with a real OSPF router or the combination of multiple OSPF routers that advertise the same prefix. Real OSPF Routers include OSPF ABR Node, OSPF ASBR Node, and OSPF ordinary Node.

* ISIS Node: A node is associated with a real ISIS router or the combination of multiple ISIS routers that advertise the same prefix

* IGP Node: including OSPF Node and ISIS Node

4. Overview

The idea of IGP-PIC is based on two pillars,

1) A shared forwarding Chain: Instead of having q separate list of next-hops for each destination, all destinations sharing the same list of next-hops can point to a single copy of this thereby allowing fast convergence by making changes to a single shared list of next-hops rather than possibly a large number of destinations.

2) A forwarding plan that support multiple levels of indirection: A forwarding that starts with a destination and ends with an outgoing interface is not a simple flat structure. Instead a forwarding entry is constructed via multiple levels of dependency.

Designing a forwarding plane that constructs multi-level forwarding chains with maximal sharing of forwarding objects allows rerouting a large number of destinations by modifying a small number of objects thereby achieving convergence in a time frame that does not depend on the number of destinations.

Similar to the implementation of BGP-PIC, see[I-D.ietf-rtgwg-bgp-pic]chapter 2 for details.

4.1. Dependency

This section describes the required functionality in the forwarding and control planes to support IGP-PIC described in the document.

IGP PIC requires a hierarchical hardware FIB support: for each IGP forwarded packet, a destination is looked up, then an IGP Node, then an Adjacency.

4.2. FRR Consideration

As per [RFC5286] Rapid failure repair is achieved through use of precalculated backup next-hops that are loop-free and safe to use until the distributed network convergence process completes. So based on backing up the next hop of the current route in advance, FRR can achieve rapid switching of faulty links.

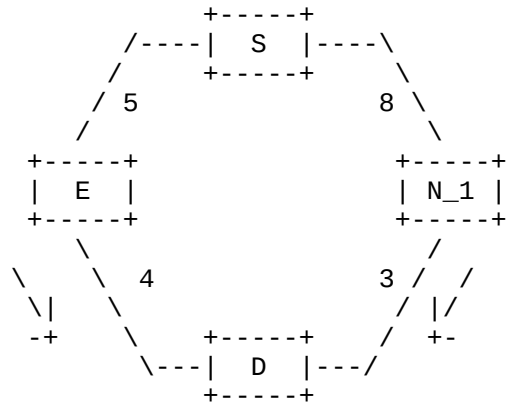

Figure 1: Node Protection Topology

As shown in the figure, the optimal next hop from original device S to D is E. If we take N_1 as the next hop for backup from S to E, when there is a fault between S and E, the data packet to D is handed over to N_1. It can be forwarded to D normally, so N_ 1 has the qualification for backup next hop from S to E. But if the COST value of the direct link from N_1 to D is greater than 17,before the route on N_1 converges again, the next jump from N_1 to D is S instead of D thus forming a temporary loop. So as per [RFC5286]

A neighbor N_1 can provide a loop-free alternate (LFA) if and only if Distance_opt(N_1, D) < Distance_opt(N_1, S) + Distance_opt(S, D)

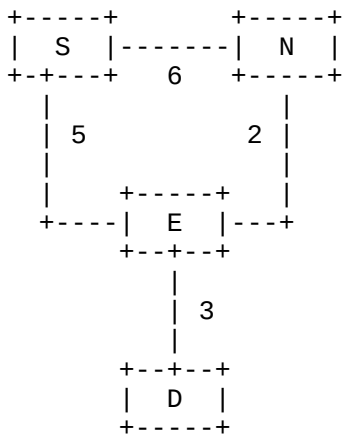

Figure 2: Link Protection Topology

Another typical scenario is shown in figure 2. When S and N Both have enabled IP FRR, so S and N will treat each other as their backup to the next hop of the D main path. At this time, when downstream node E fails, S and N will send messages to D to each other and resulting in a microloop. So the priority of node protection is higher than that of link protection.

4.3. IGP-PIC Illustration

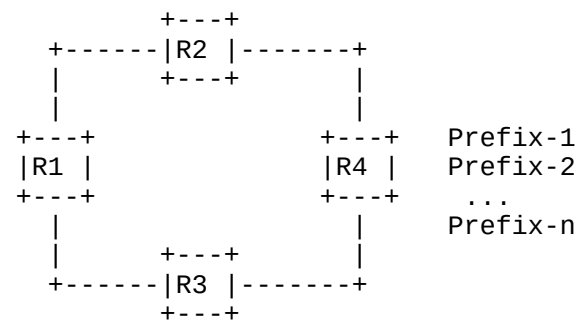

Figure 3: Single source PIC network diagram

As shown in the figure 1, R4 advertides n prefix routes. R1->R2->R4, R1->R3->R4. When the link between R1 and R2 is faulty, route calculation is performed again. Topology calculation is performed first to calculate the path to R4 from the original equal-cost path to the single path R1->R3->R4. Routes from prefix-1 to prefix-n are recalculated, and forwarding entries are updated for all routes.

When the number of prefix-1 to prefix-n increases, the time for route calculation and forwarding table update increases as the number of routes increases, which slows route convergence.

For prefix-1 to prefix-n routes, since they are all advertised by R4, their paths are the same after switching. In route calculation, the change of the route to R4 only needs to be calculated once, and the forwarding table to R4 needs to be updated to the new forwarding path. The route from Prefix-1 to Prefix-n can be updated. This is the convergence of prefix-independent routes.

Before PIC route calculation, the prefix needs to be associated with the IGP Node. In the current example, the IGP node is the real router R4.

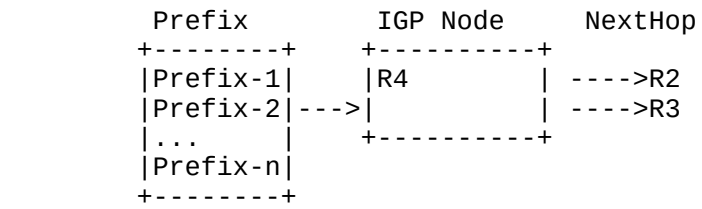

Figure 4: Single source PIC Forward

When path switching occurs, only the forwarding path of the IGP node needs to be updated from the equal-cost route ECMP path R2+R3 to R3, without recalculating and updating all prefixes. This saves the time

of route calculation and forwarding table update, and improves the speed of route convergence. In the process of PIC route calculation update, that is, the next hop information to the corresponding IGP node is updated regardless of the specific prefix.

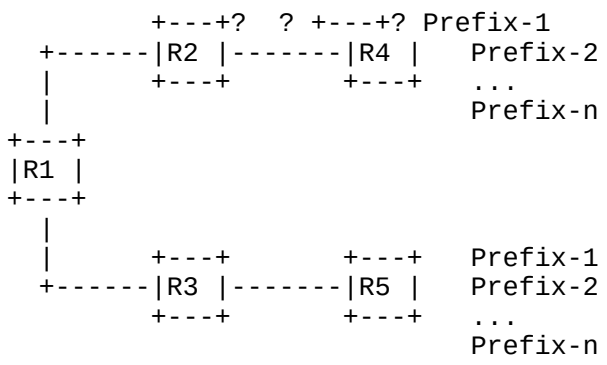

Figure 5: Multi-source PIC network diagram

In the case of multiple sources, the multiple destination nodes are combined into combined IGP node and the path is calculated for this combined node.

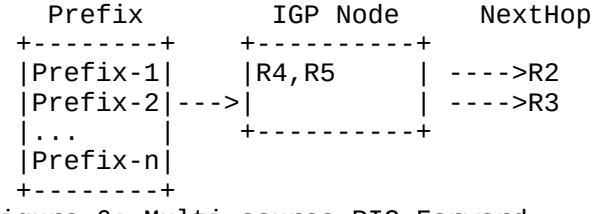

Figure 6: Multi-source PIC Forward

When the path changes, route calculation is performed again for the combined node (R4,R5), and the forwarding path is updated from the original R2+R3 to R3 without route calculation for all prefixes and forwarding table flushing.

5. ISIS PIC

5.1. Maintenance of ISIS IGP-nodes

For single-source prefixes, when an ISIS LSP is received carries the prefix TLV, an ISIS IGP Node is created and associated with the prefix. The key of ISIS IGP Node is system-id, level, and topo.

If the prefix is advertised by the LSP of the pseudo node, the key of ISIS IGP Node is system-id, pseudo node ID, level, and topo.

For multi-source prefixes, Multiple ISIS routers advertise the same prefix through LSPs, a combined ISIS IGP node is create and associated with the prefix. The key of the combined ISIS IGP node is multiple (system-id, level, and topo).

5.2. PIC Route Compute

The procedure for route calculation is as follows,

(1) Calculating the shortest-path tree for Level-1 and Level-2.

(2) Calculate each routes for Level-1 and Level-2.

When support PIC Route Compute, The procedure for route calculation is as follows,

(1) Calculating the shortest-path tree for Level-1 and Level-2.

(2) Instead of calculating routes based on each prefix, the next hop information is updated based on IGP-node.

6. OSPF PIC

6.1. Maintenance of OSPF IGP-nodes

The key of OSPF IGP-node is router-id, area, and topo.

When the prefix is advertised through a router-LSA, the OSPF IGP-node is create and the key is router-id, area, and topo.

When the prefix is advertised through a network-LSA, the key of OSPF IGP-node is router-id, DR IP-Address, area, and topo.

When the prefix is advertised through Type-3 summary-LSA, the key of OSPF IGP-node is ABR router-id, area, and topo.

When the prefix is advertised through Type-5 AS-external-LSA, the key of OSPF IGP-node is ASBR router-id, Forwarding Address, and topo.

For multi-source prefixes, Multiple OSPF routers advertise the same prefix through LSAs, a combined OSPF IGP-node is create and associated with the prefix. The key of the combined OSPF IGP-node is multiple (router-id, area, and topo).

6.2. PIC Route Compute

For OSPF route calculation, see [RFC2328], chapter 16, Calculation of the routing table. The procedure for route calculation is as follows,

(1) Calculating the shortest-path tree for an area, and then calculate the intra-area routes.

(2) Calculating the inter-area routes by examining summary-LSAs.

(3) Examining transit areas' summary-LSAs.

(4) Calculating AS external routes.

When support PIC Route Compute, The procedure for route calculation is as follows,

(1) Calculating the shortest-path tree for an area, and then calculate the intra-area routes. Instead of calculating intra-area routes based on each prefix, the next hop information is updated based on IGP-node.

(2) Calculating the inter-area routes by examining summary-LSAs. If the ABR IGP-node has been updated, the inter-area routes do not need to be recalculated.

(3) Examining transit areas' summary-LSAs. Instead of calculating routes based on each prefix, the next hop information is updated based on Intra IGP-node and ABR IGP-node.

(4) Calculating AS external routes. If the ASBR IGP-node has been updated, the AS external routes do not need to be recalculated.

7. Example

7.1. ISIS PIC Route

When the link to the IGP node changes, the topology is re-calculated and the corresponding next hop list is updated, without updating the forwarding table for each prefix.

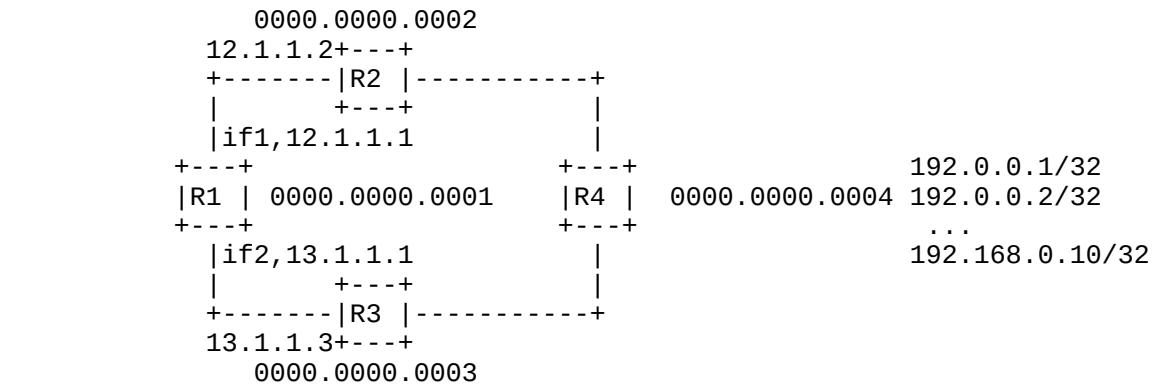

Figure 7: Single source ISIS PIC network diagram

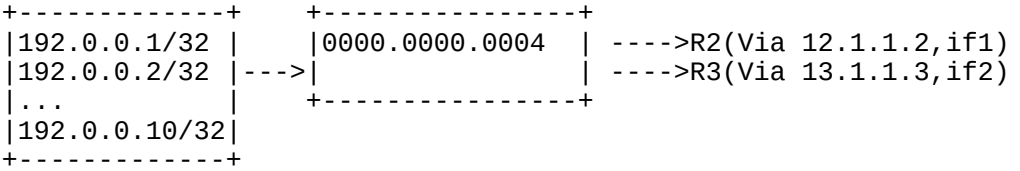

Figure 8: Single source ISIS PIC Forward

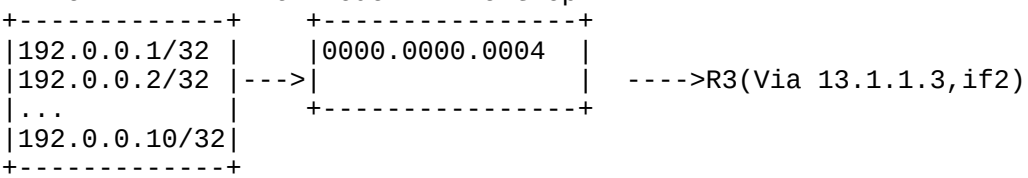

Figure 7: Single source ISIS PIC Forward

If the path to R2 is faulty, re-calculate the route and update the next hop information of the IGP node associated with R4.

7.2. OSPF PIC Route

When the link to the IGP node changes, the topology is re-calculated and the corresponding next hop list is updated, without updating the forwarding table for each prefix.

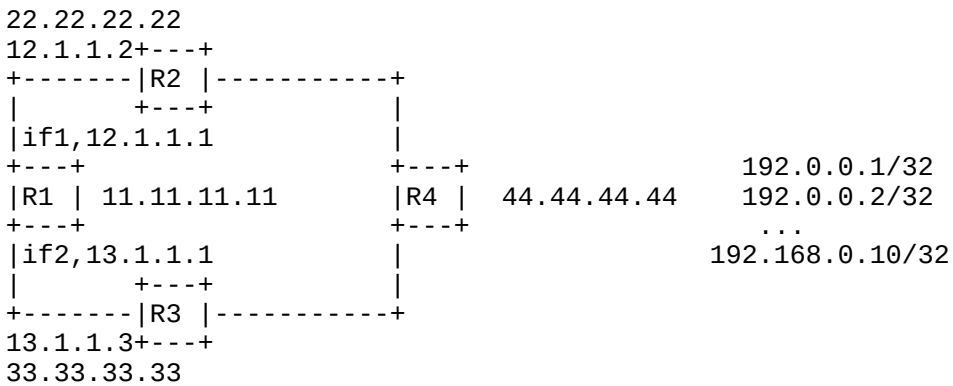

Figure 9: Single source OSPF PIC network diagram

Prefix	IGP Node	NextHop
192.0.0.1/32	44.44.44.44	---->R2(Via 12.1.1.2,if1)
192.0.0.2/32		---->R3(Via 13.1.1.3,if2)
...		
192.0.0.10/32		

Figure 10: Single source OSPF PIC Forward

Prefix	IGP Node	NextHop
192.0.0.1/32	44.44.44.44	
192.0.0.2/32		---->R3(Via 13.1.1.3,if2)
...		
192.0.0.10/32		

Figure 11: Single source OSPF PIC Forward

If the path to R2 is faulty, re-calculate the route and update the next hop information of the IGP node associated with R4.

8. Normative References

[I-D.ietf-rtgwg-bgp-pic]
Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", Work in Progress, Internet-Draft, draft-ietf-rtgwg-bgp-pic-19, 1 April 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-rtgwg-bgp-pic-19>.

[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>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[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>.

Authors' Addresses

Yue Wang
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: wangy73@chinatelecom.cn

Changwang Lin
New H3C Technologies
China
Email: linchangwang.04414@h3c.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: wangaj3@chinatelecom.cn