Network Working Group
Internet-Draft
Intended status: Informational
Expires: 7 November 2026

P. Psenak
J. Horn
Cisco Systems
B. Decraene
G. Gryszata
Orange

6 May 2026

Distributed Congestion Mitigation
draft-psenak-lsr-igp-dcm-00

Abstract

This document describes the Distributed Congestion Mitigation (DCM) mechanism using the Interior Gateway Protocols (IGPs) such as IS-IS [RFC1195], OSPFv2 [RFC2328], or OSPFv3 [RFC5340]. DCM is a tactical, distributed mechanism, designed to mitigate network congestion by offloading traffic to an alternate, congestion-free paths. DCM is fully integrated in IGPs.

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 7 November 2026.

Copyright Notice

Copyright (c) 2026 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 (https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction

Network capacity planning is a proactive strategy to deal with the network congestion. Even with the proper capacity planning, network congestion arises from oversubscription, link or node failures, and from the shifting traffic patterns. DCM provides a reactive, distributed mechanism to mitigate local congestion by leveraging the interface utilization monitoring, IGP link administrative groups [RFC8919], [RFC8920], and Flex-Algo [RFC9350] path computation and forwarding.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

ABR: Area Border Router.

ASBR: Autonomous System Border Router.

DCM: Distributed Congestion Mitigation.

FAD: Flex-algo Definition [RFC9350].

OFA: Offloading Flex-Algo, used for traffic offloading.

Congestion Affinity: Administrative group used to exclude congested links from the OFA topology.

High Utilization Affinity: Administrative group used to signal high utilization and prevent additional offload traffic.

Congestion Threshold: The utilization level at which a link is marked with Congestion affinity. Traffic offloading is initiated for the local link.

Non-Congestion Threshold: The utilization level at which the additional offloading on the link stops. If no traffic is offloaded from the local link, the Congestion affinity is removed.

High Utilization Threshold: The utilization level at which a link is marked with High Utilization affinity to stabilize the load.

Restore Threshold: The utilization level at which traffic restoration is initiated.

Low Utilization Threshold: The utilization level used to remove High Utilization affinity.

LSP: Link State Packet.

LSA: Link State Advertisement.

4. DCM Requirements

DCM aims to offload traffic from the locally congested links. Some of the requirements of DCM are listed below:

* DCM offloaded traffic MUST avoid any congested link.

* DCM offloaded traffic MUST NOT create new congestion in the network.

* DCM MUST support multiple congested links in the network.

* Offloaded traffic MAY be protected via LFA [RFC5286] or TI-LFA [RFC9855]. A microloop avoidance MAY be used for the offloaded traffic.

5. Control Plane

DCM provisions a dedicated OFA. OFA's FAD is set to exclude the Congestion affinity. Any link declared congested MUST be excluded from the OFA topology by setting the Congestion affinity. This allows the OFA to natively route traffic around all congested links.

OFA is only used to carry the offloaded traffic.

DCM can be used to offload Algorithm 0 traffic or traffic of any Flex-Algo, which is not used as OFA itself. If the DCM is used for Flex-algo traffic, the OFA, on top of using the Congestion affinity exclude rule, SHOULD inherit the FAD algorithm-type, constraints, and metric type from the original Flex-algo for which the DCM is done.

Multiple OFAs can coexist inside the IGP area.

6. Local Congestion Monitoring and Detection

DCM utilizes precise congestion monitoring and detection mechanisms for the local interfaces on the router. Some of the characteristics of such monitoring are:

* Interface output rates are collected periodically and are statistically adjusted to avoid oscillations and ensure stability. An Exponential Weighted Moving Average (EWMA) is an example of such adjustments, used for smoothing the output rate samples.

* Trend monitoring may optionally be employed to enhance detection performance. When an upward trend is detected, a more aggressive weighting may be applied to react faster to significant changes, thereby preventing or minimizing any traffic drop. Stabilization windows and trend thresholds can be used to detect and confirm sustained upward trends in the link utilization.

* Some noise filtering is required for short-duration utilization spikes.

* A more granular traffic rate data, like per destination prefix or per flow, MAY be collected to find out the destinations that are significantly contributing to the local congestion.

7. Traffic Offloading and Restoring

Traffic offloading is performed to divert the traffic onto the shortest path that avoids any congested links. The offloading process adheres to the following principles:

* Only traffic from locally congested interfaces are offloaded at each node.

* Offloading MUST NOT be done for Flex-algo that is used as OFA.

* Offloading MUST NOT be done if there is no valid path to the destination node in the OFA topology.

* Additional offloading MUST NOT be done, if the path to the destination node in the OFA topology crosses any link that is advertised with the High Utilization Affinity.

* The offload path is calculated for each prefix individually.

* Any destination with the primary path over the locally congested link is eligible for offloading. Offloading may be subject to user-defined filtering.

* A more granular traffic rate data, like per destination prefix or per flow, MAY be used to decide which traffic is offloaded.

* Traffic is diverted using the Unequal Cost Multi-Path (UCMP) across the primary path and the offload path.

* Offloading and restoring is executed progressively in periodic iterations, utilizing a jittered delay between each step, to ensure network stability.

* At each iteration, a small increment of traffic (e.g., 5%) is offloaded or restored.

* The specific amount of traffic to be offloaded is calculated based on the lowest capacity link identified on the path from the offloading node to the destination node inside the OFA.

* Offloading starts when the Congestion Threshold is exceeded on the local interface. Additional offloading ceases when the utilization on the local interface drops below Non-Congestion Threshold.

* If the calculated OFA path from the offloading node to the destination node changes (i.e., any topological modification affecting the end-to-end path) while traffic for a prefix destined to that node is being actively offloaded, the offloading for that prefix MUST be terminated. The offloading process for that prefix MUST then be re-initiated from the initial state, following the standard iterative process.

* Restoration of the offloaded traffic to the primary path starts when the utilization on the local interface drops below the Restore Threshold. Restoration ceases when the utilization on the local interface exceeds the Non-congestion threshold, or if all traffic is restored.

8. Offloaded Traffic Forwarding

Offloaded traffic is routed via OFA paths, requiring the switching of the traffic's algorithm to the OFA. The forwarding behavior is defined as follows:

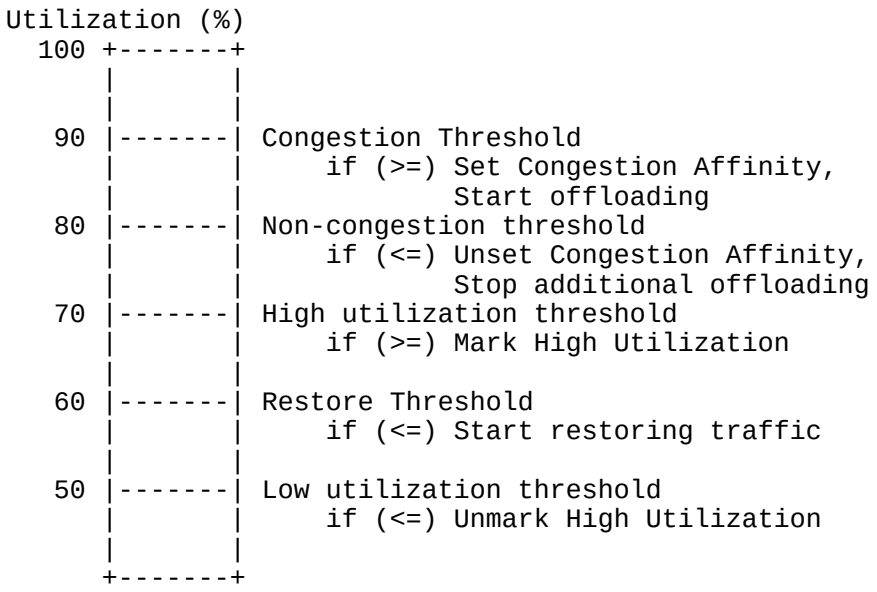
* In the MPLS network (intra-area): The top label is replaced by the OFA label of the destination node.

* In the MPLS networks (inter-area/inter-domain): The OFA label of the destination ABR/ASBR is pushed on top of the label stack.

* In the SRv6 networks: The mechanism requires the push of the OFA node SID (e.g., uN, END) of the destination node. Alternatively, an adjacency SID (e.g., uA, END-X) of the penultimate hop of the destination node towards the destination node itself in the OFA topology may be used, to avoid the presence of multiple local SIDs at the destination node.

9. Oscillation Avoidance

To limit the likelihood of oscillations, DCM uses two affinity-based signals based on link utilization thresholds as illustrated below:



The logic of the two affinities is as follows:

* Congestion Affinity: Removes the link from the OFA topology. Any offloaded traffic, local or remote, is removed from the link.

* High Utilization Affinity: Signals routers in the area to stop sending new offload traffic to the link without impacting existing offloaded traffic. Traffic that has already been offloaded over the link can stay there.

10. Oscillation Mitigation

While the affinity-based signaling described in Section 9 effectively mitigates large-scale oscillations, localized instabilities may still occur due to the following:

* Elephant Flows: If UCMP hashing results in a high-bandwidth flow being steered onto an offload path, it may trigger secondary congestion on that path.

* Simultaneous Offloading: In a distributed environment, multiple routers may simultaneously detect congestion and initiate offloading for their locally congested links, leading to rapid, coordinated shifts in traffic patterns that exceed the network’s convergence stability.

To address these scenarios, implementations MUST employ a damping mechanism for prefix-specific offloading:

* Path Change Monitoring: Routers SHOULD monitor the frequency of OFA path changes for each prefix. If the path changes more than N times within a defined interval T, the router SHOULD suppress further offloading for that prefix.

* Exponential Backoff: To further enhance stability, an exponential backoff mechanism MAY be applied, increasing the duration for which offloading is suppressed each time the threshold is exceeded. This ensures that prefixes causing persistent path churn are excluded from the DCM process until the network state stabilizes.

11. Implementation Scope and Discretion

While this document defines the mandatory protocol behaviors required to ensure interoperability and network stability, certain aspects of the DCM mechanism are left to the implementation.

11.1. Mandatory Requirements

Implementations MUST adhere to the following rules to ensure the integrity of the DCM mechanism:

* OFA Topology Exclusion: The OFA FAD MUST exclude Congestion Affinity.

* High Utilization Affinity Signaling: Routers MUST interpret the High Utilization Affinity as a signal to cease sending new offload traffic to the link, while allowing existing offloaded traffic to persist to avoid unnecessary churning.

* Routers participating in the OFA MUST monitor the local links utilization and advertise the Congestion Affinity when the Congestion Threshold is exceeded. They MUST stop advertising the Congestion Affinity if the utilization drops below the Non-Congestion Threshold and there is no traffic that is locally offloaded from the link.

* Routers participating in the OFA MUST monitor the local links utilization and advertise the High Utilization Affinity when the High Utilization Threshold is exceeded. They MUST stop advertising the High Utilization Affinity if the utilization drops below the Low Utilization Threshold.

* Advertising the Congestion and High Utilization Affinities is subject to the standard throttling used by the implementation when generating the LSP, or LSA update.

* The setting of the Congestion and High Utilization affinities SHOULD be performed more aggressively than the unsetting. Setting of these affinities is a protective measure against imminent performance degradation; therefore, it SHOULD be prioritized to prevent congestion. Implementations MAY use immediate (un-smoothed) utilization samples, or more aggressive statistical adjustments for setting these affinities. Conversely, unsetting these affinity is a restoration measure to return to optimal routing. A more cautious approach to unsetting is recommended to ensure the link has stabilized and usage of the smoothed, statistically adjusted utilization value is recommended.

* Algorithm Constraints: Offloading MUST NOT be performed for any Flex-Algo that is currently designated as an OFA.

11.2. Implementation-Specific Decisions

Implementations have the discretion to define the operational heuristics that trigger the protocol mechanisms, including:

* Congestion Detection Logic: The specific statistical methods used for link utilization adjustment (e.g., EWMA, trend analysis, or noise filtering) are implementation-dependent.

* Threshold Tuning: The specific values for Congestion, Non-Congestion, High Utilization, Restore, and Low Utilization thresholds are operational parameters that should be tuned based on network-specific capacity, stability and performance requirements.

* Offload Granularity: The iterative process, including the size of traffic increments, is left to the implementer to optimize for network stability.

* Offload Interval: The offload interval SHOULD be set to a value larger than the sum of the time required for the nodes in the network to detect high utilization or congestion, the time required to advertise the link affinities (including the LSP/LSA throttling), and the time required for those affinities to propagate across the network, plus some additional time to ensure network stability.

* Filtering Policies: While the protocol provides the mechanism for offloading, the policy governing which prefixes or traffic classes are eligible for offloading is a local configuration choice.

11.3. Deployment Considerations

DCM can be deployed incrementally in the network. Legacy routers, that do not support DCM, MUST not participate in the OFA.

DCM does not need to be enabled on all routers in the network. However, it must be enabled on all routers along the specific path towards the egress node, starting from the point where DCM is being used. To successfully offload traffic, the offload path must be contiguous. If any router along the path towards the egress node lacks DCM enablement, the OFA path may not be available.

12. IANA Considerations

This document makes no requests of IANA.

13. Security Considerations

DCM relies on standard IS-IS, OSPF, and OSPFv3 flooding mechanisms. Implementations MUST ensure that affinity configuration is consistent between routers participating in the DCM inside the area and protected against unauthorized modification, as malicious manipulation could lead to traffic drops or suboptimal routing.

14. 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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

15. Informative References

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <https://www.rfc-editor.org/info/rfc1195>.

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

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[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>.

[RFC9855] Bashandy, A., Litkowski, S., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute Using Segment Routing", RFC 9855, DOI 10.17487/RFC9855, October 2025, <https://www.rfc-editor.org/info/rfc9855>.

[RFC8919] Ginsberg, L., Psenak, P., Previdi, S., Henderickx, W., and J. Drake, "IS-IS Application-Specific Link Attributes", RFC 8919, DOI 10.17487/RFC8919, October 2020, <https://www.rfc-editor.org/info/rfc8919>.

[RFC8920] Psenak, P., Ed., Ginsberg, L., Henderickx, W., Tantsura, J., and J. Drake, "OSPF Application-Specific Link Attributes", RFC 8920, DOI 10.17487/RFC8920, October 2020, <https://www.rfc-editor.org/info/rfc8920>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

Contributors

The following people contributed to the content of this document and should be considered coauthors:

Francois Clad
Email: fclad@cisco.com

Authors' Addresses

Peter Psenak
Cisco Systems
Apollo Business Center
Mlynske nivy 43
82109 Bratislava
Slovakia
Email: ppsenak@cisco.com

Jakub Horn
Cisco Systems
Milpitas, CA 95035
United States of America
Email: jakuhorn@cisco.com

Bruno Decraene
Orange
France
Email: bruno.decraene@orange.com

Guillaume Gryszata
Orange
France
Email: guillaume.gryszata@orange.com