RTGWG
Internet-Draft
Intended status: Informational
Expires: 1 September 2026

Z. Li
Z. Du
China Mobile
W. Cheng
J. Wang
G. Zhang
Centec Networks
28 February 2026

Congestion-Aware Adaptive Flow Table Switching for ECMP
draft-li-rtgwg-congestion-aware-flowset-switching-00

Abstract

This document defines a congestion-aware adaptive flow table switching mechanism for Equal-Cost Multi-Path (ECMP) routing. The mechanism periodically assesses the congestion state of egress ports and progressively adjusts flow table mappings based on quantified congestion levels. This addresses the port congestion issues that occur in traditional ECMP load balancing when traffic patterns change suddenly or multicast traffic is present, while maintaining packet ordering within flows.

Table of Contents

1. Introduction

Equal-Cost Multi-Path (ECMP) routing is a widely deployed load balancing technology in data center networks [RFC2991]. Traditional ECMP distributes traffic across multiple equal-cost paths by hashing packet header fields, typically the five-tuple. To ensure packet ordering within a flow, the mapping between a flow and its egress port typically remains unchanged throughout the flow's lifetime.

However, this static mapping approach exhibits significant limitations in the following scenarios:

Traffic Surge Scenario: Network traffic is highly dynamic and may cause sudden increases on certain ports. The flow table mapping cannot be adjusted in time to alleviate congestion.

Multicast Traffic Scenario: The replication characteristics of multicast traffic may cause it to concentrate on a small number of ports, exacerbating load imbalance.

Existing congestion response strategies typically adopt two extreme approaches: either no switching (maintaining the original mapping until flow aging) or full switching (simultaneously migrating all flows on a congested port). The former cannot respond to congestion in a timely manner, while the latter may cause congestion transfer and resource fluctuations.

This document defines a congestion-aware adaptive flow table switching mechanism that quantifies port congestion levels and progressively adjusts flow table mappings to achieve dynamic optimization of load balancing while preserving packet ordering.

2. Terminology and Conventions

2.1. 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.

2.2. Definitions

ECMP (Equal-Cost Multi-Path): A routing strategy that distributes traffic across multiple paths of equal cost.

Flow Table: A data structure that stores the mapping between flow identifiers and egress ports, ensuring that packets of the same flow are forwarded from the same port.

Congestion Quantification Index (CQI): A quantified value representing the degree of port congestion, ranging from 0 to a configured maximum. A CQI of 0 indicates no congestion.

Assessment Interval: The time interval for port congestion state assessment.

Flow Table Migration: The operation of remapping a flow table entry from one egress port to another.

3. Problem Statement

3.1. Limitations of Traditional ECMP

Traditional ECMP load balancing uses static hash mapping. Once a flow is assigned to a port, the mapping remains unchanged throughout the flow’s lifetime. This design has the following deficiencies:

Delayed Response: When a port becomes congested, flows already mapped to that port cannot be migrated in time, causing congestion to persist.

Load Imbalance: The randomness of traffic and the presence of elephant flows may cause severe load imbalance between ports.

3.2. Inadequacy of Existing Solutions

Flowlet Switching: This mechanism switches based on inter-packet gaps within a flow and relies on manually configured time thresholds. If the threshold is too large, it degrades to traditional ECMP; if too small, it may cause packet reordering.

Full-Switch Strategy: Migrating all relevant flows simultaneously when congestion is detected may cause the target port to be instantly overloaded, resulting in congestion transfer.

3.3. Requirements Summary

A mechanism is needed that can:

1. Perceive port congestion state in real-time

2. Progressively adjust flow table mappings based on congestion level

3. Avoid congestion transfer and resource fluctuation

4. Preserve packet ordering

4. Solution Overview

This mechanism defines two core functional components:

Port Congestion Assessment: Periodically assesses the congestion state of each egress port and generates a Congestion Quantification Index (CQI).

Adaptive Flow Table Migration: Progressively migrates flow table entries from congested ports to less loaded ports based on the CQI value.

The fundamental design principle is that the higher the CQI value, the more flow table entries are allowed to migrate in the current assessment interval. For each entry migrated, the CQI is decremented by 1 until the CQI reaches zero or no more entries need migration.

5. Protocol Specification

5.1. Port Congestion Assessment

5.1.1. Assessment Interval

Implementations MUST support a configurable assessment interval. The RECOMMENDED default value is between 10ms and 100ms.

Implementations MAY adaptively adjust the assessment interval based on overall traffic levels: shortening the interval during high traffic to improve responsiveness, and lengthening it during low traffic to reduce overhead.

5.1.2. Congestion Quantification Index Calculation

CQI calculation SHOULD be based on one or more of the following metrics: port egress queue depth, port buffer utilization, and port packet drop counter increment.

The CQI value range is 0 to CQI_MAX. The RECOMMENDED value for CQI_MAX is 16.

The recommended CQI calculation method is:

CQI = min(CQI_MAX, floor(queue_depth / congestion_threshold))

where congestion_threshold is the congestion determination threshold, RECOMMENDED to be 10% of queue capacity.

5.1.3. State Advertisement

At the end of each assessment interval, the Port Congestion Assessment component MUST synchronize each port’s CQI value to the Flow Table Migration component.

5.2. Adaptive Flow Table Migration

5.2.1. Migration Decision

When a packet arrives, implementations MUST process it according to the following rules:

Rule 1 (Flow Table Does Not Exist): Perform normal flow table learning and select the port with the lightest current load.

Rule 2 (Port Failure): If the flow table exists but the corresponding port is unavailable, a new port MUST be selected.

Rule 3 (No Congestion): If the flow table exists and the corresponding port’s CQI is 0, the implementation MUST continue using the current port and MUST NOT perform migration.

Rule 4 (Congestion Exists): If the flow table exists and the corresponding port’s CQI is greater than 0, the implementation SHOULD perform flow table migration.

5.2.2. Migration Operation

When migration is triggered, implementations MUST perform the following steps:

Step 1: Select the port with the smallest CQI from all available ports as the target. If multiple candidate ports have the same CQI, implementations MAY use random selection or round-robin.

Step 2: Update the flow table entry’s egress port to the target port.

Step 3: Decrement the original port’s CQI by 1.

5.2.3. Migration Quantity Control

A key property of this mechanism is that the migration quantity is proportional to the congestion level. When the CQI value is high, more flow table entries may be migrated within a single assessment interval. When the CQI value is low, the migration quantity decreases accordingly.

Implementations MUST ensure that within a single assessment interval, the number of flow table entries migrated from a port does not exceed that port’s initial CQI value.

5.2.4. Continuous Migration

If the CQI does not drop to 0 within an assessment interval, subsequent assessment intervals will recalculate the CQI. If congestion persists, migration will continue; if congestion is alleviated, migration will decrease or stop.

6. Data Structures

6.1. Flow Table Entry

A flow table entry MUST contain the following fields: flow identifier (obtained through hash calculation), egress port identifier, valid bit, and timestamp (for aging).

6.2. Port Status Table

The port status table MUST contain the following fields: port identifier, port status (UP/DOWN), current CQI value, and queue depth.

7. Operational Procedures

7.1. Initialization

Implementations MUST perform the following at startup:

1. Clear the flow table
2. Initialize all ports’ CQI to 0
3. Start the periodic assessment task

7.2. Packet Processing

The packet processing flow is as follows:

1. Packet arrives
2. Calculate flow identifier
3. Query flow table
4. If flow table does not exist: learn new entry, select lightest loaded port
5. If flow table exists: check port status and CQI, perform migration if needed
6. Forward packet

8. Relationship with Existing Mechanisms

8.1. Relationship with ECMP

This mechanism is an enhancement extension to traditional ECMP, adding congestion awareness and adaptive migration capabilities on top of ECMP. Implementations MAY overlay this mechanism on existing ECMP implementations.

8.2. Relationship with Flowlet

This mechanism MAY be used in conjunction with flowlet switching. Flowlet uses inter-packet gaps within a flow for switching, while this mechanism uses port congestion state to trigger switching. The two can be complementary.

8.3. Relationship with Congestion Control

This mechanism operates at the forwarding layer and is orthogonal to end-to-end congestion control mechanisms such as ECN and DCQCN. Implementations SHOULD consider coordination with congestion control mechanisms.

9. Security Considerations

9.1. Denial of Service Risk

Attackers may induce frequent migration by forging traffic, consuming device resources.

Mitigation Measures: Implementations SHOULD set a maximum number of migrations per unit time. Implementations SHOULD use smoothing algorithms for CQI calculation to avoid overreaction to instantaneous fluctuations.

9.2. Information Disclosure Risk

CQI values and migration decisions may reveal network topology or traffic pattern information.

Mitigation Measures: Implementations MUST implement access control for related data. Inter-module communication SHOULD use security mechanisms.

9.3. Configuration Integrity

Mitigation Measures: Implementations MUST ensure configuration parameter integrity. Implementations SHOULD log configuration changes.

10. IANA Considerations

This document does not require IANA to allocate any resources.

11. References

11.1. 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>.

[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>.

11.2. Informative References

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, DOI 10.17487/RFC2991, November 2000, <https://www.rfc-editor.org/info/rfc2991>.

[RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, DOI 10.17487/RFC6438, November 2011, <https://www.rfc-editor.org/info/rfc6438>.

[RFC7098] Carpenter, B., Jiang, S., and W. Tarreau, "Using the IPv6 Flow Label for Load Balancing in Server Farms", RFC 7098, DOI 10.17487/RFC7098, January 2014, <https://www.rfc-editor.org/info/rfc7098>.

Authors' Addresses

Zhiqiang Li
China Mobile
32 Xuanwumen West Street
Beijing
100053
China
Email: lizhiqiangyjy@chinamobile.com

Zongpeng Du
China Mobile
32 Xuanwumen West Street
Beijing
100053
China
Email: duzongpeng@chinamobile.com

Wei Cheng
Centec Networks
Suzhou
215000
China
Email: chengw@centec.com

Junjie Wang
Centec Networks
Suzhou
215000
China
Email: wangjj@centec.com

Guoying Zhang
Centec Networks
Suzhou
215000
China
Email: zhanggy@centec.com