Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 4 September 2025

Z. Li
Z. Du
China Mobile
W. Cheng
J. Wang
Centec

3 March 2025

Adaptive Load Balance Enhancement
draft-li-adaptive-load-balance-enhancement-00

Abstract

This draft proposes an adaptive load-balancing mechanism to address high-throughput transmission challenges in East-West computing, data exchange, and DC interconnection services. Traditional methods-ECMP, RPS, and Flowlet-face limitations: ECMP suffers from hash collisions and load imbalance; RPS risks TCP packet reordering; Flowlet depends on impractical manual thresholds for burst interval configuration, leading to suboptimal performance.

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 .

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 4 September 2025.

Table of Contents

1. Introduction

East Data West Computing, Data Express, and DC Interconnection services require high-throughput network transmission. Current mainstream high-throughput interconnection and load balancing mechanisms in the industry include ECMP, RPS, and Flowlet.

ECMP transmits data flows as units. Its principle is when the first packet of a data flow arrives at the first switch from the host, the switch determines whether it is the first packet of the flow based on the five-tuple (source IP, destination IP, source port, destination port, protocol). If so, a hash algorithm selects a forwarding port (non-first packets follow the same port as previous packets). The advantage of ECMP is preventing TCP packet reordering and retransmission. However, poor hash design may cause collisions, leading to load imbalance and increased queuing delays. Even with optimal hashing, short flows may be assigned to ports handling long flows, worsening delays.

RPS transmits individual packets as units. Each packet arriving at the switch is randomly assigned a forwarding port, regardless of flow affiliation. RPS achieves near-perfect load balancing under high traffic, minimizing queuing delays. However, random port assignment risks TCP packet reordering, causing retransmissions. While rare, severe imbalance could amplify this issue.

Flowlet transmits flow segments (bursts of packets sent together in TCP). It leverages the time gap between TCP bursts: if the interval between two bursts exceeds the transmission delay threshold, the new flowlet is randomly assigned a port (otherwise, it follows the previous port). Flowlet balances ECMP's stability with RPS's load balancing. Its drawback lies in relying on manually set thresholds for transmission delay differences. Overly large thresholds mimic ECMP, while overly small ones mimic RPS.

2. Problem Statement

Throughput-sensitive services often involve elephant flows (long-duration flows). Traditional load balancing mechanisms like ECMP, which use five-tuple-based hashing, fail to fully utilize link bandwidth. Packet-level load balancing (RPS) introduces out-of-order delivery, and as bandwidth increases, the CPU and memory overhead for reordering packets at the receiver becomes prohibitive (or even unmanageable). Flowlet-based load balancing strikes a compromise between ECMP and RPS. However, determining the transmission time difference threshold is impractical, requiring manual configuration. Excessively large thresholds degrade Flowlet into ECMP, while overly small thresholds mimic RPS1. Meanwhile, the industry lacks standardized solutions for increasingly critical high-throughput transmission demands.

3. Solution

Edge Ingress Node: The entry point for high-throughput service flows. Based on service type (e.g., By DPI or directly use the service ID) and link conditions to the destination node (obtained via controller requests or distributed in-band detection), it implements multi-path parallel transmission. Specific steps include: 1.Query all available paths to the destination and select the path with lowest bandwidth utilization and lowest latency to send the first packet (or flowlet) . 2.For the second packet (or flowlet), choose a path with the lowest bandwidth utilization that ensures no out-of-order delivery after accounting for transmission delays . 3.Subsequent packets follow the same logic, dynamically selecting paths to balance bandwidth efficiency and sequence integrity . 4.When selecting packet sizes (e.g., single packet, flowlet, or subflow), ensure sufficient time margin to avoid out-of-order issues. For example, shorter subflows on high-latency paths and longer subflows on low-latency paths .

Intermediate Nodes: For high-throughput flows, these nodes perform hierarchical load balancing if instructed by packet headers; otherwise, they act as normal nodes for transparent forwarding.

Edge Egress Node: For strictly ordered services, it performs final in-order verification. If out-of-order packets occur (rarely), this node buffers and reorders them to ensure sequential delivery to the receive.

Interactive packets can be carried across multiple data plane protocols. Taking the IPv6 extension header as an example,the Next Header field for the high-throughput in-order transmission extension header is temporarily assigned the value 100 (to be updated after formal IANA allocation). The extension header length is 12 bytes, and its specific format follows the definition provided earlier.

4. Security Considerations

TBD.

5. IANA Considerations

TBD.

Authors' Addresses

Zhiqiang Li
China Mobile
Beijing
100053
China
Email: lizhiqiangyjy@chinamobile.com

Zongpeng Du
China Mobile
Beijing
100053
China
Email: duzongpeng@chinamobile.com

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

Junjie Wang
Centec
Suzhou
21500
China
Email: wangjj@centec.com