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

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

3 March 2025

Congestion Detection Optimization
draft-li-congestion-detection-optimization-00

Abstract

This draft proposes an adaptive congestion detection mechanism for high-throughput data transmission in wide area networks (WANs). With increasing network bandwidth (up to 800Gbps) and challenges in traditional TCP-based protocols (e.g., throughput degradation over long distances and high packet loss rates), the solution focuses on optimizing congestion identification while minimizing bandwidth overhead.

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.

Copyright Notice

Table of Contents

1. Introduction

With the rapid development of key computing infrastructures, to meet the growing demand for the transmission of massive amounts of data across wide area networks, the bandwidth of data network transmission has also been continuously upgraded from 10 Gigabits to 25Gbps, 100Gbps, 200Gbps, 400Gbps and even 800Gbps. However, traditional transmission control protocols(eg TCP) experience a sharp decline in good throughput as transmission distances increase and packet loss rates rise. Therefore,design a high-throughput data transmission solution over wide area networks to improve data transmission efficiency is of significant importance.

Traffic control and congestion control determine the efficiency of data transmission and are key technologies in high-throughput network transmission. Identifying congestion points in the network quickly, accurately and not costly is of great significance for traffic and congestion control mechanisms.

2. Problem Statement

Traditional schemes for identifying congestion points are as follows: 1.Based on packet loss, delay, etc., require a detection time of at least one RTT (Round-Trip Time). 2.Based on active feedback from intermediate congestion nodes, the detection time can be compressed to less than one RTT (depending on the location of the congestion node), but this introduces new problems: 2.1 Feedback messages bring additional link bandwidth overhead, resulting in low bandwidth utilization, and can even cause new congestion on the back path, leading to detection time greater than one RTT. 2.2 To ensure the

timely detection, congestion status report messages are sent frequently, wasting bandwidth. 2.3 Maintaining the state of traffic flows at intermediate nodes requires high-performance equipment at these nodes, affecting scalability.

3. Solution

This draft introduces a mechanism that, when adjacent nodes communicate frequently (with an adjustable threshold, by default, a sending interval between two consecutive service message packets of no less than 0.5 RTT), utilizes normal service traffic packets to carry congestion information with the flow. When communication is infrequent (with an adjustable threshold, by default, a sending interval between two consecutive service packets of less than 0.5 RTT), it actively generates congestion indication packets, ensuring zero bandwidth overhead during heavy load and timely perception of downstream node’s congestion during light load or idle times.

When utilizing normal service traffic messages to carry information, this can be accomplished by reusing certain fields in the packet header, such as the flow label of an IPv6 message; When actively generating congestion indication packets for notification, it will directly generate a packet that is recognized by the signal source, such as a RoCEv2 CNP message

Define some value(eg A55A) as the congestion indication magic number when utilizing normal service traffic packets to carry congestion information. The congestion indication magic number can be transmitted using the ToS field of two consecutive IPv4 service packets or the TC bits of two consecutive IPv6 service packets. If service messages happen to transmit A5 and end without subsequent packets (within 0.5 RTT), the congested node replicates the packet header that sent the A5 magic number, constructs a payload of all 0s in a 64-byte packet, and modifies the ToS or TC field to 5A, completing the transmission of the congestion indication magic number. If the congested node does not have any service messages to send for 0.5 RTT or more, it proactively generates congestion indication packets such as CNP and sends back.

The sending frequency of the two types of congestion indication methods is not within the scope of this draft , and can be based on the mechanisms of existing congestion control algorithms, such as determining the sending frequency of packets based on the degree of congestion in the queue.

4. Security Considerations

TBD.

5. IANA Considerations

TBD.

Authors' Addresses

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

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

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