Routing Area Working Group
Internet-Draft
Intended status: Informational
Expires: 3 September 2025

Jiayuan. Hu, Ed.
China Telecom

2 March 2025

Credit-based Flow Control for Cross-AIDC WAN transmission Based on RSVP
draft-hu-rtgwg-cbfc-rsvp-00

Abstract

This draft defines the Credit-based flow control mechanism for WAN based on the RSVP protocol. With the increasing demand for AI computing power, the computing power of a single AIDC can no longer meet the needs of large model training. This has given rise to cross-AIDC distributed model training, driving the demand for transmitting RoCEv2 packets over WAN networks. AI training is extremely sensitive to network packet loss, and even a small amount of packet loss may lead to a significant decline in training efficiency. In addition, the elephant flow and extreme concurrent traffic also place higher demands on network performance. Credit-based flow control is a Backpressure-based traffic management technology, which has high reliability and stability in practical applications. It can provide high-throughput and zero-packet-loss transmission guarantees for RoCEv2 traffic, effectively ensuring the efficiency of cross-data center AI training.

This draft focuses on the scenario where RoCEv2 packets are transmitted through SRv6 tunnels in the WAN and further expands the capabilities of the RSVP protocol in WAN. This draft introduces the Credit-based flow control mechanism into the RSVP protocol to achieve precise traffic control and provides processing analysis.

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 3 September 2025.

Table of Contents

1. Introduction

The exponential growth of AI computing power demands, especially for large-scale model training, has transformed the data center landscape. The current single AIDC can no longer meet the needs of large-scale model training. The training parameters of large models have skyrocketed in the past five years, reaching the trillion level, and are expected to increase a hundred-fold in the next five years, reaching the quadrillion level. This has led to the rise of cross-AIDC distributed model training, which, in turn, drives the need to transmit RoCEv2 packets across WANs.

AI training is highly sensitive to network packet loss. Even a small amount of packet loss can significantly reduce training efficiency. Additionally, the presence of elephant flow and extreme concurrent traffic poses greater challenges to network performance. To address these issues, this Draft focuses on a Credit-based flow control mechanism for WAN transmission.

Credit-based flow control is a Backpressure-based traffic management technology. It has demonstrated high reliability and stability in practical applications, capable of providing high-speed and zero-packet-loss transmission guarantees for RoCEv2 traffic. This effectively ensures the efficiency of cross-AIDC AI training.

This draft centers on the scenario where RoCEv2 packets are transmitted through SRv6 tunnels in the WAN. It aims to expand the capabilities of the RSVP in WAN environments. By introducing the Credit-based flow control mechanism into the RSVP protocol, the draft enables precise traffic control and provides in-depth processing analysis. The RSVP, while originally designed to reserve resources for data flows, has limitations such as scalability issues with increasing reservations, assuming a static network topology, and potential resource waste and service interruptions in case of node failures. The proposed solution redefines an RSVP option to implement the Credit-based flow control, which is detailed in subsequent sections through the initial process, data transmission process, and termination process.

2. Conventions Used in This Document

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. Abbreviations

AIDC: Artificial Intelligence Data Center

RoCEv2: RDMA over Converged Ethernet version 2

RSVP: Resource Reservation Protocol

MB: Megabytes

3. Scenarios for distributed AI training network

3.1. Distributed model training for difference AIDC

Industry insiders can all feel the accelerating pace of the development of large AI models. Mainstream technology companies are developing large models and iterating new versions as quickly as possible, in the hope of gaining a head start in this brand-new industry. The training parameters of large models have increased a hundredfold in the past five years and have reached the trillion level. It is expected that the parameters will increase a hundredfold again in the next five years, reaching the quadrillion level. The intelligent computing power has also been rapidly upgraded accordingly. Currently, a single data center has reached the scale of a ten-thousand GPU cluster to try its best to meet the almost endless AI computing demands.

An AI cluster is formed by connecting multiple computer nodes to create a collaborative computing environment, thus providing powerful computing power and data processing capabilities for artificial intelligence applications. However, any specific thing has its limits, and computing power clusters are no exception. A single AI cluster cannot expand without limit, as it will be affected by factors such as power supply and the size of the geographical location. In addition, cloud computing has a peak-valley effect, and the computing power of a single cluster faces the problem of fragmented deployment, making it difficult to bear large-scale AI training business and leading to a decrease in resource utilization.

Facing the problem of fragmented AI resources on the cloud, Microsoft has proposed the "Singularity" framework, which enables planet-scale pre-emptible, migratable, and elastically scalable AI task scheduling. This framework can achieve high elasticity and migratability in resource scheduling and increase the utilization rate of AI resources on the cloud, but it lacks attention to the training performance across clusters. Facing the problem of heterogeneous AI training networks in public clouds, AWS has proposed the MiCS solution, which can make full use of heterogeneous network bandwidth. By reducing the network traffic on slower links, it can amortize the expensive global gradient synchronization overhead. To solve the problem of the high cost of building AI training clusters, Meta has proposed decentralized heterogeneous training. It uses distributed, heterogeneous, and low-bandwidth interconnected AI training resources to train basic large models, reducing the training cost. Therefore, model training across AIDCs is an important stage in the development of artificial intelligence, and corresponding network capabilities are required to ensure its implementation.

3.2. Separated storage and model training

In the traditional data processing architecture, computing and storage are often closely coupled. Although this integrated computing and storage architecture performed excellently in the early data requirements, with the surge in business complexity and data volume, some insurmountable defects have gradually emerged: When adding a new computing node, data (including MetaData and Data) needs to be synchronized between nodes, resulting in low expansion efficiency and increasing the complexity and management cost of the system. Since computing and storage resources are closely coupled, when the business load changes, it is impossible to dynamically allocate resources according to the business load, flexibly adjust computing and storage resources, resulting in low resource utilization and difficult cost control.

In the scenario of separating storage from computing, even a tiny data packet loss can seriously disrupt the progress of training. Therefore, efficient storage data transmission and zero packet loss are of great importance, which also places higher demands on the network.

4. Forwarding Plan Solution

4.1. RSVP protocol extend

RSVP is a signaling protocol that enables end systems and network devices to reserve resources along a data path. [RFC2205] It allows applications to request specific levels of QoS, such as bandwidth, delay, and packet loss rate, for their data flows. However, It has scalability issues as the number of reservations increases. Additionally, RSVP assumes a static network topology, and it may not adapt well to highly dynamic networks. The following RSVP common header has been provided in RFC2205. [RFC2205]

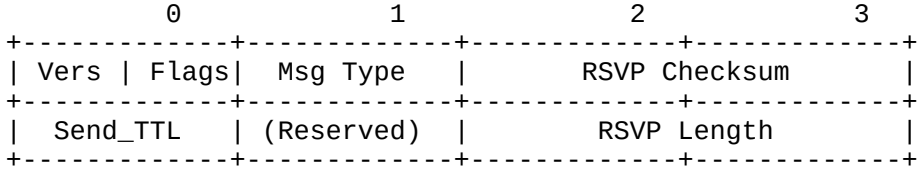

Figure 1: Common Header

At the same time, in the resource reservation process of the RSVP, even if the data flow is not successfully established in the end, the reserved resources will not be released, resulting in resource waste. The RSVP protocol relies on specific nodes for the management and maintenance of resource reservation. If these nodes fail, it may lead to the interruption of the entire resource reservation process. In order to solve the problems of the RSVP protocol, such as its inability to dynamically adjust resource reservation and the service guarantee issues that may occur due to possible failures, this draft defines a new RSVP option, which is used to implement credit-based flow control. The detail process is divided into three steps: the initial process, the data transmission process, and the termination process.

4.1.1. Initial process

Before transmitting RoCEv2 flow between servers in different AIDC, a transmission channel with service guarantee capabilities needs to be established between the servers. The traditional RSVP can only guarantee the quality of service in a fixed manner according to specific service information. The RSVP based on credit is similar to the traditional RSVP in initialization. The sender sends a "Credit-based PATH" detection message to the receiver, this message will flood to every path that can reach the receiver, and this message contains the data flow identifier. When the devices on the path receive the RSVP message packet, they will all send a "Credit-based RESV" message to the upstream device in the path with the credit value. The credit value is stored in the object contents field of the RSVP message. It represents the buffer reserved by the device for the forwarding task, and the credit value should be less than the buffer.

The new msg type fields in the common header are as follows:

28 = Credit-based PATH

29 = Credit-based RESV

The RSVP protocol defines the Object field for expansion for different message types. Every object consists of one or more 32-bit words with a one-word header, and the format is as follows:

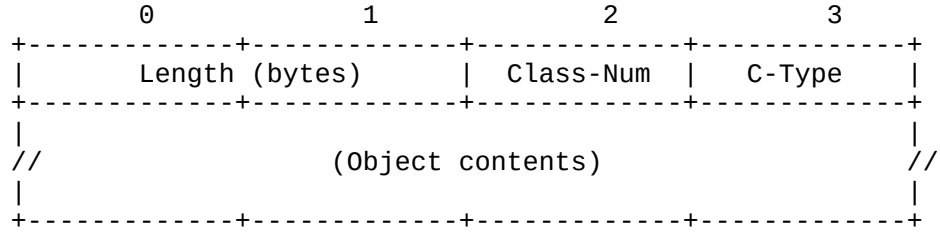

Figure 2: Object Formats

The object information in the Credit-based PATH message is the same as that in the traditional PATH message, but the Credit-based RESV message is different. There are newly defined types in Class-Num and C-Type, the following new classes are defined in Class-Num:

Credit : It represents the initial cache reserved by the device for the forwarding task, with the unit being MB.

4.1.2. Data transmission process

After the resource reservation initial process is completed, each network device will maintain a credit value. This credit value is equal to the credit value replied by the next-hop device on the path during initialization. At this time, the sender server can start sending data. The forwarding devices in the path randomly forward data packets smaller than the credit value they maintain. After sending, they subtract the size of the forwarded data packet from the credit value they maintain. When the credit value is not 0, they will continue to send; when the credit value is 0, they will stop sending.

In contrast, after the next-hop forwarding device receives the data packet, according to the first in first out principle, the device forwards the data block in the buffer, it will reply with a Credit-based RESV packet to the previous-hop forwarding device. This packet carries a new credit value, which is the size of the data packet it has just forwarded (indicating that a new buffer space has become available).

After each data transmission, the device should receive a message carrying the credit value in reply from the next-hop device. When a failure occurs in the link or the next-hop device, it may not be possible to receive the reply message. If the reply message is still not received after the preset heartbeat time elapses, it can be considered that a failure has occurred in the next-hop device or the link. The forwarding device will forward the traffic to the backup

path according to the acknowledgment message received during the initialization process, thereby ensuring the non-loss transmission of the traffic.

4.1.3. Termination process

After the data transmission is completed, the source device needs to send an RSVP message representing the end of the task to flood along all reachable paths. Upon receiving this message, the devices on the paths will terminate the corresponding guarantee tasks.

TBC

5. IANA Considerations

TBC

6. Security Considerations

TBC

7. References

7.1. 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>.

[RFC2205] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, DOI 10.17487/RFC2205, September 1997, <https://www.rfc-editor.org/info/rfc2205>.

Contributors

Thanks to all of the contributors.

Author’s Address

Jiayuan Hu (editor)
China Telecom
109, West Zhongshan Road, Tianhe District

Guangzhou
Guangzhou, 510000
China
Email: hujy5@chinatelecom.cn