

Routing Area Working Group
Internet-Draft
Intended status: Informational
Expires: 27 March 2026

Jiayuan. Hu, Ed.
Z. Hu, Ed.
T. Pi, Ed.
Y. Zhu, Ed.
China Telecom
23 September 2025

Credit-based Flow Control Based on RSVP for RDMA transmission in WAN
draft-hu-rtgwg-cbfc-rsvp-01

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 RDMA 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 RDMA traffic, effectively ensuring the efficiency of cross-data center AI training.

This draft focuses on the scenario where RDMA 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 27 March 2026.

Copyright Notice

Copyright (c) 2025 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	3
2. Conventions Used in This Document	3
2.1. Requirements Language	3
2.2. Abbreviations	4
3. Scenarios for distributed AI training network and problem statement	4
3.1. Transmitting parameter between multiple AIDCs	4
3.2. Directly transmitting sample data to AI servers	4
3.3. Coordinated model inference	5
4. Control Plan Solution	5
4.1. RSVP protocol extension	5
4.2. Initial process	7
4.3. Data transmission process	8
4.4. Termination process	9
5. IANA Considerations	10
6. Security Considerations	10
7. References	10
7.1. Normative References	10
Contributors	11
Authors' Addresses	11

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 RDMA 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 RDMA traffic. This effectively ensures the efficiency of cross-AIDC AI training.

This draft centers on the scenario where RDMA 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 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

RDMA: RDMA over Converged Ethernet version 2

RSVP: Resource Reservation Protocol

MB: Megabytes

3. Scenarios for distributed AI training network and problem statement

3.1. Transmitting parameter between multiple AIDCs

The computational scaling of individual AI Data Centers (AIDCs) faces physical constraints (e.g., power/cooling limits). To meet ZFLOPS-scale demands for foundational model training, distributed coordination across geographically dispersed AIDCs becomes essential. This scenario employs parallelization strategies (e.g., pipeline/data parallelism), requiring frequent synchronization of model parameter via RDMA protocols. [draft-hhz-fantel-sar-wan]

In this scenario, the synchronization traffic of the parameter which usually features some elephant flows. The characteristics of elephant flow are long duration and large data amount, which can easily cause network congestion. Therefore, WAN needs to have a flow control method in order to provide efficient and lossless transmission of parameter data.

3.2. Directly transmitting sample data to AI servers

For customers with stringent data security requirements, sample data must be streamed directly to training servers via RDMA protocols. Current RDMA implementations suffer severe performance degradation (up to 50% throughput loss at 0.1% packet loss) due to their Go-back-N retransmission mechanism. [draft-hhz-fantel-sar-wan]

In the scenario of transmitting sample data to AI servers, even a tiny data packet loss can seriously disrupt the progress of training. Therefore, efficient sample data transmission and zero packet loss are of great importance, which also places higher demands on the network.

3.3. Coordinated model inference

To address the growing demand for LLM inference while reducing infrastructure costs, many customers implement split-learning architectures between their data centers and third-party AIDCs. This solution can distribute computational loads (e.g., keeping prefill phase local while offloading decode phase to AIDCs) and maintains data privacy by retaining input/output layers on-premises.

In the scenario of coordinated model inference, similar to scenarios mention above, WAN need to have dynamic elephant flow control and fast network failure protection capabilities.

4. Control Plan Solution

4.1. RSVP protocol extension

RSVP is a signaling protocol that enables end systems and network devices to reserve resources along a data path. [RFC2205] 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].

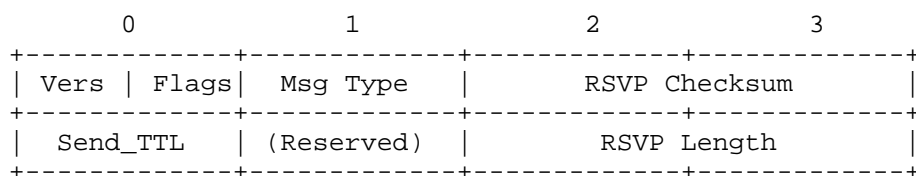


Figure 1: Common Header

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.

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

28 = Credit-based Path initialization

29 = Credit-based Reserve

30 = Credit-based End

31 = Credit-based End ACK

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:

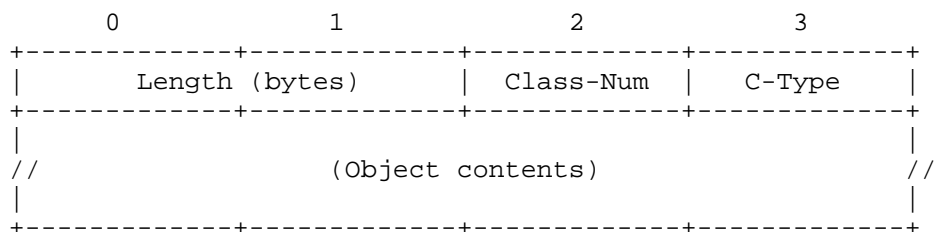


Figure 2: Object Formats

The object information in the "Credit-based Path initialization" message is the same as that in the traditional PATH message, but the "Credit-based Reserve 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 buffer reserved by the device for the forwarding task, with the unit being MB.

4.2. Initial process

Before transmitting RDMA 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 Reserve" 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. Initial process as shown in the following figure:

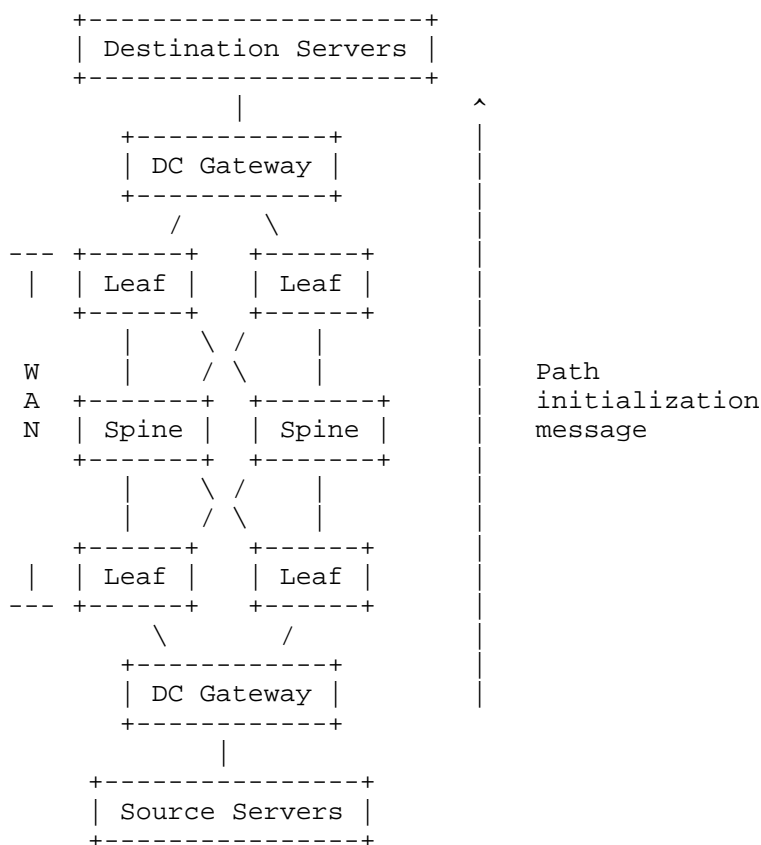


Figure 3: Initial process

After confirming the selected traffic control device, the receiver initiates resource reservation through an RSVP message, when the selected device receives the resource reservation message, they will all send a "Credit-based Reserve" message to the previous selected 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.

4.3. Data transmission process

After the resource reservation initial process is completed, each selected network device will maintain a credit value. This credit value is equal to the credit value replied by the next-selected 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-selected 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 Reserve packet to the previous-selected 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-selected device. When a failure occurs in the link or the next-hop device, it will not 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. Data transmission process as shown in the following figure:

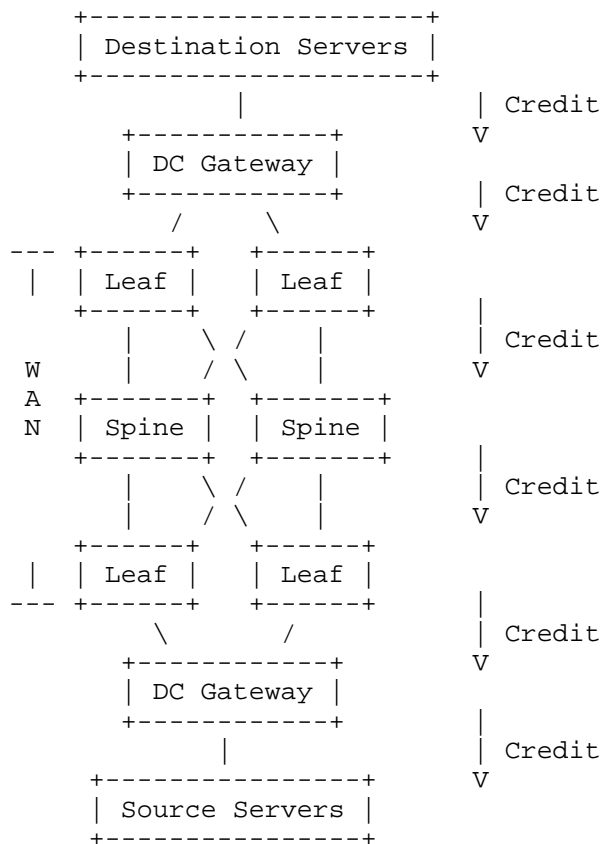


Figure 4: Data transmission process

4.4. Termination process

After the data transmission is completed, the source device needs to send a "Credit-based End" message representing the end of the task to flood along all reachable paths. Upon receiving this message, the selected devices on the paths will terminate the corresponding guarantee tasks and answer a "Credit-based End ACK" message to the previous selected device. If device not receive the "Credit-based End ACK" message from the next-selected device in within the preset time, it will repeatedly send 'Credit-based END' messages until it receives a "Credit-based End ACK" message from the next-selected device. Termination process as shown in the following figure:

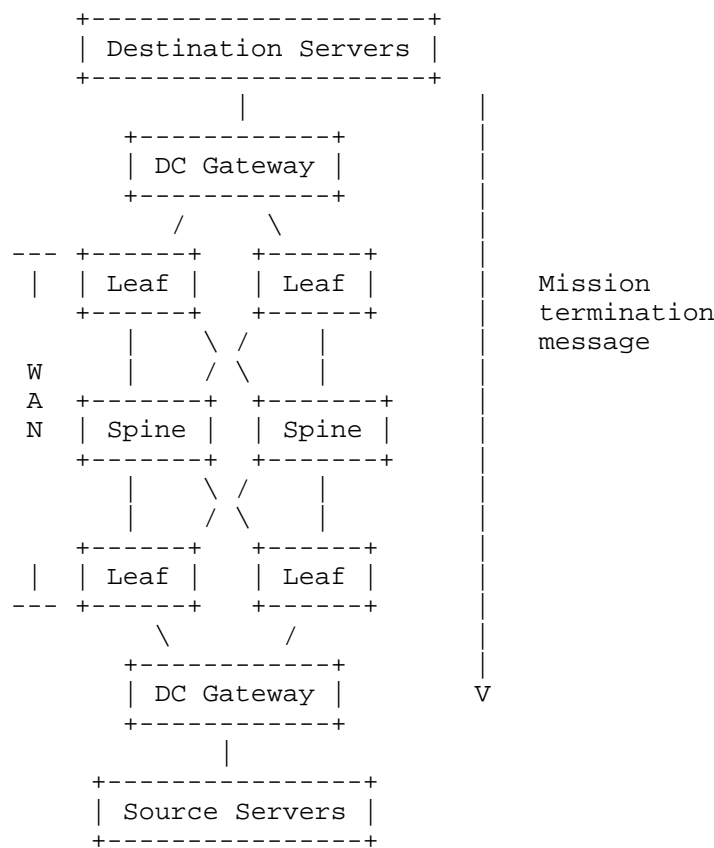


Figure 5: Termination process

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>>.
- [draft-hhz-fantel-sar-wan]
"FANTEL scenarios and requirements in Wide Area Network".

Contributors

Thanks to all the contributors.

Authors' Addresses

Jiayuan Hu (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangzhou, 510000
China
Email: hujy5@chinatelecom.cn

Zehua Hu (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangzhou, 510000
China
Email: huzh2@chinatelecom.cn

Tanxin Pi (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangzhou, 510000
China
Email: pitxl@chinatelecom.cn

Yongqing Zhu (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangzhou, 510000
China
Email: zhuyq8@chinatelecom.cn