Network Working Group
Internet-Draft
Intended status: Informational
Expires: 29 August 2025

Q. Xiong
ZTE Corporation
K. Yao
China Mobile
C. Huang
China Telecom
Z. Han
China Unicom
J. Zhao
CAICT

25 February 2025

Problem Statement for High Performance Wide Area Networks
draft-xiong-hpwan-problem-statement-02

Abstract

High Performance Wide Area Network (HP-WAN) is designed for many applications such as scientific research, academia, education and other data-intensive applications which demand high-speed data transmission over WANs, and it needs to provide efficient transmission services within a completion time. This document outlines the problems for HP-WANs.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

As described in [I-D.kcrh-hpwan-state-of-art], data is fundamental for research, academia, education, industrial and other data-intensive applications, such as High Performance Computing (HPC) for scientific research, cloud storage and backup of industrial internet data, distributed training of Artificial Intelligence (AI), and so on. The use cases in non-dedicated networks from public operators such as large file transfer, traffic across data centers and sharing traffic between dedicated network and non-dedicated network are also described in [I-D.yx-hpwan-uc-requirements-public-operator].

Within these applications, they may generate huge volumes of data by using advanced instruments and high-end computing devices. They need to be connected between research institutions, universities, and data centers across large geographical areas over long-distance links. For example, sharing data between research institutes must transfer over hundreds or thousands of kilometers. It needs to ensure large-scale data transfer and provide stable and efficient transmission services over Wide Area Networks (WANs). These applications may require a periodic or on-demand high-speed transfer with variable start time, data volume and transmission patterns, which demanding data transmission within a completion time.

More recently, the massive data transmission and long-distance connection over WANs have become a key factor affecting the performance of existing transport layer protocols such as Transfer Control Protocol (TCP), Quick UDP Internet Connections (QUIC), Remote Direct Memory Access (RDMA) and so on. Different transport protocols carrying massive data transfer requests will co-exist in the same network and the multiple transport protocols optimizations may incur much overhead, including congestion control algorithms redesign and parameter tuning, hardware adaptation and QoS policies, etc. The transport protocol proxy may be deployed to adapt the functionality for different transport protocols.

Moreover, the traditional congestion control algorithms are typically implemented at the host (sender and receiver) perform blind transmission by controlling the size of the congestion window with rate adjusting by detection of overloaded links. It will be difficult to predict the performance due to the unpredictable behaviour of the WANs. For example, for the host, without awareness of network capability, it will lead to a poor convergence speed impacting the completion time due to the slow start and passive rates adjusting. It will also lead to RTT fluctuation due to large buffer and long queues upon long feedback loop. For the network, it will transfer the unscheduled traffic with low bandwidth utilization due to the bottleneck links and instantaneous congestion. All of above will impact the performance and result in the untimely transmission of high-volume data.

High Performance Wide Area Network (HP-WAN) is designed for many applications such as scientific research, academia, education and other data-intensive applications which demand high-speed data transmission over WANs, and it needs to provide efficient transmission services within a completion time. A variety of problems about what are specifically in the way for HP-WAN requirements are outlined in this document.

1.1. 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. Terminology

This document adopts the terminology defined in [I-D.kcrh-hpwan-state-of-art].

It also makes use of the following abbreviations and definitions in this document:

BDP: Bandwidth Delay Product

DC: Data Center

DCI: Data Centers Interconnection

HPC: High Performance Computing

WAN: Wide Area Networks

PFC: Priority Flow Control

ECN: Explicit Congestion Notification

ECMP: Equal-Cost Multipath

RTT: Round-Trip Time

TCP: Transfer Control Protocol

RDMA: Remote Direct Memory Access

QUIC: Quick UDP Internet Connections

3. Technical Goals for HP-WANs

The services need to be provided in HP-WANs mainly focus on massive data with timely transmission while multiple services may co-exist over long-distance WANs as described below.

* Massive data transmission, high-volume data with high-speed transfer, e.g. the data speed of a flow could be at 2Gbps~1Tbps.

* Requested completion time, the data transmission should be completed within a requested completion time, e.g. the completion time could be minutes~milliseconds.

* Scheduled transmission, traffic patterns could be scheduled by the sender, e.g. data volume, start time, finish time, service type.

* Long-distance transmission over non-dedicated WANs, with multiple hops and domains, long RTT latency, routing changes, network congestion, packet loss, and link quality fluctuations, e.g. the distance between two sites or DCs could be more than 100km or 1000km.

* Multiple services are co-existed with concurrent flows, with different transport protocols for data transmission, such as QUIC, TCP and RDMA etc.

It is required to achieve high-speed data transmission within a completion time. Moreover, it is also crucial to maximize bandwidth utilization while ensuring fairness among multiple services. This document outlines the technical goals for HP-WANs as described below.

* High throughput: ensuring the high-speed data transmission within a requested completion time for a flow, which could be impacted by the bandwidth, convergence speed, start time and RTT.

* Efficient use of capacity: efficiently using available network capacity with fairness to maximize data transfer rates and minimize the completion time for multiple flows.

4. Problem Statement

The specific requirements of HP-WANs may encompass a wide range of aspects. These include transport-related technologies such as proxy, flow control, QoS negotiation, congestion control, admission control and traffic scheduling. Additionally, they also involve routing-related technologies like traffic engineering, resource scheduling, and load balancing.

Existing network technologies face numerous challenges and fall short of meeting performance requirements. This document highlights the key issues associated with HP-WANs in the following sub-sections.

4.1. Poor Convergence Speed

The traditional congestion control mechanisms perform blind transmission by controlling the size of the congestion window with rate adjusting by detection of overloaded links. WAN is a black box to provide unpredictable behaviors for high-speed transmission due to the issues such as multiple hops and domains, long Round-Trip Time (RTT), routing changes, network congestion, packet loss, and link quality fluctuations. The BDP (Bandwidth Delay Product) which represents the maximum amount of data that can be in transit on the network at any given time is variable over WANs, so the inflight data is difficult to predict for host-based congestion control algorithms. It will lead to the poor convergence speed that the host always takes significantly long time to identify the optimal sending rate comparing to the requested completion time.

For example, it will use the slow start and blind detection with unawareness of network capability leading to long convergence time such as Cubic (e.g.over 50s), BBR (e.g.over 30s) and BBRv2 (e.g.30~50s). BBR divides the entire process into four stages, Startup, Drain, ProbeBW and ProbeRTT. The probe cycle of ProbeRTT state is long, e.g. 10s. The convergence time will be multiple probe cycle which will impact the completion time at seconds level. There is a significant transmission capacity gaps between the appropriate sending rate and the available network capacity. The transport protocols should signal and collaborate with the network to negotiate the rate for the host to send traffic.

4.2. Unscheduled Traffic

The host sending large unscheduled traffic without collaboration will lead to the instantaneous congestion in WANs. For multiple high-speed flows, the random arrival and departure of cross-traffic without scheduling creates significant fluctuations for available capacity in WANs. The network infrastructure may struggle to handle high-volume data transfers efficiently if applications do not proactively schedule the traffic. Without awareness of the traffic patterns, the network risks unscheduled resource allocation, leading to low bottleneck bandwidth utilization, reduced overall throughput, and uncontrolled completion time.

For example, for HPC applications, a large amount of data will be transmitted, e.g. the data volumes of a single flow may be from 10G to 1TB, the host sends the unscheduled large traffic leading to the instantaneous congestion, packet loss, and queuing delay within network devices in WANs, resulting in low throughput. Considering the multiple services with various types of flows, the optimal bandwidth and transmission time may be different and the traffic is random to join and leave without to be scheduled to multiple paths and fine-grained network resources, which can not achieve the timely transmission. The resource of WANs should be scheduled at the elements along the path to provide predictable capability for high-speed transmission.

4.3. Long Feedback Loop

The congestion algorithms are implemented by controlling the size of the congestion window and adjusting the sending rates upon the network status feedback. It will delay the network feedback due to the long-distance transmission delays and large RTT, resulting in the inability to adjust the transmission rate in a timely manner. It will be challenging for congestion control over WANs for controlling the total amount of data entering the network to maintain the traffic at an acceptable level, leading to RTT fluctuation due to long queues and large buffer at network devices with high-speed transmission upon the long network state feedback loop. Especially when multiple flows targeting an aggregating node, the maximum value is exceeding devices buffer capacity.

For example, the loss-based congestion control algorithms, such as Reno and CUBIC, depends on the congestion notification with packet loss. Explicit Congestion Notification (ECN) can be used to achieve an end-to-end congestion notification based on IP and transport layers. When a congestion occurred, the network may signal congestion by ECN markings or by dropping packets, and the receiver passes this information back to the sender in transport-layer acknowledgements, notifying the source to adjust the transmission rate. It will use the slow start, requiring large buffer which is impacted by multiple hops and long RTT latency over WANs.

And the congestion-based congestion control algorithms such as BBR, depends on the measurement of congestion, it actively measures bottleneck bandwidth (BtlBw) and round-trip propagation time (RTprop) based on the model to calculate the BDP and then to adjust the transmission rate to maximize throughput and minimize latency. But BBR relies on real-time measurement of the parameters, and will optimize the buffer overflow, but it is not significant under large RTT, e.g. retransmission will increase when the buffer size is less than two BDPs, thereby affecting the control precision of BBR in long-distance networks.

4.4. Multiple Transport Protocols Adaption

Multiple services are coexisted for massive data transmission over WANs with different transport protocols, such as QUIC, TCP and RDMA etc. Multiple transport protocols, each handling substantial data transfer requests, will coexist within the same network. Optimizing these diverse transport protocols can entail significant overhead. This encompasses issues such as redesigning congestion control algorithms, mapping parameters, adapting hardware components, and formulating QoS policies. To improve such significant overhead, a more flexible deployment strategy, such as the implementation of a transport protocol proxy, can be enabled for the adaptation of functionality to suit the requirements of different transport protocols. The proxy should support high-speed transmission such as traffic classification, packet processing, buffering, and implement the collaboration and interaction between proxy and hosts. Seamless communication between hosts and network infrastructure requires adaptive coordination across heterogeneous transport protocols (e.g., TCP, UDP, QUIC, RDMA).

Moreover, in some scenarios, it is difficult to simultaneously ensure both encrypted data and high-speed transmission. Encryption algorithms (e.g., AES, RSA) require intensive CPU operations, which reducing available capacity for data transmission. Edge computing nodes with limited CPU capabilities struggle to balance encryption and data processing. The proxy could perform optimizations (e.g., hardware acceleration, distributed encryption modules) to mitigate the bottlenecks.

5. Security Considerations

This document covers several of representative applications and network scenarios that are expected to make use of HP-WAN technologies. Each of the potential use cases does not raise any security concerns or issues, but may have security considerations from both the use-specific perspective and the technology-specific perspective.

6. IANA Considerations

This document makes no requests for IANA action.

7. Acknowledgements

The authors would like to acknowledge Guangping Huang, Yao Liu and Zheng Zhang for their thorough review and very helpful comments.

8. References

8.1. Normative References

[I-D.kcrh-hpwan-state-of-art]
King, D., Chown, T., Rapier, C., and D. Huang, "Current State of the Art for High Performance Wide Area Networks", Work in Progress, Internet-Draft, draft-kcrh-hpwan-state-of-art-01, 8 January 2025, <https://datatracker.ietf.org/doc/html/draft-kcrh-hpwan-state-of-art-01>.

[I-D.yx-hpwan-uc-requirements-public-operator]
Yao, K. and Q. Xiong, "High Performance Wide Area Network (HPWAN) Use Cases and Requirements -- From Public Operator’s View", Work in Progress, Internet-Draft, draft-yx-hpwan-uc-requirements-public-operator-00, 20 February 2025, <https://datatracker.ietf.org/doc/html/draft-yx-hpwan-uc-requirements-public-operator-00>.

[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>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC7424] Krishnan, R., Yong, L., Ghanwani, A., So, N., and B. Khasnabish, "Mechanisms for Optimizing Link Aggregation Group (LAG) and Equal-Cost Multipath (ECMP) Component Link Utilization in Networks", RFC 7424, DOI 10.17487/RFC7424, January 2015, <https://www.rfc-editor.org/info/rfc7424>.

[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>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC9232] Song, H., Qin, F., Martinez-Julia, P., Ciavaglia, L., and A. Wang, "Network Telemetry Framework", RFC 9232, DOI 10.17487/RFC9232, May 2022, <https://www.rfc-editor.org/info/rfc9232>.

[RFC9331] De Schepper, K. and B. Briscoe, Ed., "The Explicit Congestion Notification (ECN) Protocol for Low Latency, Low Loss, and Scalable Throughput (L4S)", RFC 9331, DOI 10.17487/RFC9331, January 2023, <https://www.rfc-editor.org/info/rfc9331>.

[RFC9438] Xu, L., Ha, S., Rhee, I., Goel, V., and L. Eggert, Ed., "CUBIC for Fast and Long-Distance Networks", RFC 9438, DOI 10.17487/RFC9438, August 2023, <https://www.rfc-editor.org/info/rfc9438>.

Authors' Addresses

Quan Xiong
ZTE Corporation
China
Email: xiong.quan@zte.com.cn

Kehan Yao
China Mobile
China
Email: yaokehan@chinamobile.com

Cancan Huang
China Telecom
China
Email: huangcanc@chinatelecom.cn

Zhengxin Han
China Unicom
China
Email: hanzx21@chinaunicom.cn

Junfeng Zhao
CAICT
Beijing
China
Email: zhaojunfeng@caict.ac.cn