hpwan
Internet-Draft
Intended status: Informational
Expires: 24 August 2025

K. Yao
China Mobile
Q. Xiong
ZTE Corporation

20 February 2025

High Performance Wide Area Network (HPWAN) Use Cases and Requirements -- From Public Operator's View

draft-yx-hpwan-uc-requirements-public-operator-00

Abstract

Bulk data transfer is a long-lived service over the past twenty years. High Performance Wide Area Networks (HP-WANs) are the backbone of global network infrastructure, enabling the seamless transfer of vast amounts of data and supporting advanced scientific collaborations worldwide. Many of the state-of-the-art dedicated networks have been mentioned in [I-D.kcrh-hpwan-state-of-art]. For non-dedicated networks like public operator's network, the case is different in terms of QoS policies, security policies, etc. This document presents use cases and requirements of HPWAN from public operator's view.

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 24 August 2025.

Table of Contents

1. Introduction

Bulk data transfer is a long-lived service over the past twenty years. Some dedicated networks have been designed to carry these kind of data transfer services, like CERN, internet2, etc. Many of the state-of-the-art dedicated networks have been mentioned in [I-D.kcrh-hpwan-state-of-art].

In these dedicated networks, policies and network provisioning are all designed specifically for bulk data transfer services, which means these services have high SLA guarantee. In non-dedicated networks, for example, the operator's networks, large amount of data transfer service has grown quickly, with the increase of network bandwidth. The difference is that, in these non-dedicated networks, bulk data transfer flows need to share bandwidth with Internet traffic, which inevitably leads to resource contention and requires further planning and protocol optimization.

When there are multiple data transfer requests coming to the operator's network, the scheduling of different data transfer flows will influence the completion time for each job. The scheduling primarily comes from two aspects. One is the job scheduling or orchestration, that is, at the service orchestrator, different data transmission requests need to be scheduled on bandwidth reservation and its corresponding time occupancy, according to the priorities of each job. The other is the traffic scheduling during data transmission. The second aspect needs coordination of both routing and transport technologies. The routing issues related to bulk data transfer include traffic engineering, and load balancing, etc. While transport issues cover more, including flow control, congestion control, admission control, and proxy. Due to the characteristics of a shared network like public operator's network, long-distance transmission of bulk data with packet loss rate over 0.1% and traffic contention, transport optimizations are key for bulk data transmission services.

This document presents some typical bulk data transfer use cases the non-dedicated networks currently deploy, and propose some requirements for transport layer.

2. Definition of Terms

This document uses the following terms defined in [I-D.kcrh-hpwan-state-of-art]:

* High Performance Wide Area Network (HPWAN)

* Remote direct memory access (RDMA)

3. Use Cases

3.1. Large File Transfer Over Operator's Shared Network

Astronomy and biology data have been growing fastly, with the development of advanced instruments for data collection. For example, The volume of the data generation per year of an astronomical observatory in China is around 500 terabytes(TBs), and there are around 200 observation jobs, resulting in an average transmission volume of 2.5 TB per job.

The customer (astronomical observatory) currently rents leased line (10Gbps) from public operator like China Mobile. Considering there are other background traffic, the real transmission speed of an astronomical observatory job is 7Gbps, leading to a total job completion time around 47.6 minutes.

Customers want to reduce cost, and they want to be charged by the time of bandwidth occupancy, rather than renting leased line every month. In the current operator's shared network, like China Mobile backbone network (CMnet), the average packet loss rate is around 0.1%, in some cases, the packet loss rate will increase to 0.5%. Customers currently use upper layer protocol like FTP [RFC959] or data transfer tools like RSYNC [RSYNC], the underlying transport protocol is TCP. TCP is sensitive to packet loss, especially in long range transmission, e.g., over 2000 kilometers(km). Under such circumstance, the transmission rate will sharply decrease to around hundreds of Mbps, e.g, 700 Mbps, leading to an increase of job completion time to 476 minutes, which can not be accepted by customers. Therefore, operators want to optimize transport behaviors to reduce the overall data transmission duration.

To sum up, in the case of large file transfer over operator's shared network, the transport protocol is TCP, the typical volume of transmission job is several TBs, the completion time is required to be within several hours. The packet loss rate is around 0.1% to 0.5%. The objective is to increase the transmission rate over long range like 2000 km.

3.2. Time Constrained Traffic Across Data Centers

Another case is the traffic between large data centers which are owned by public operators. Traditionally, the primary traffic across data centers is cloud file backup. Cloud file backup may have similar requirements like the case mentioned in the previous section. Currently, data centers have advanced to accommodate more artificial intelligence (AI) jobs, like AI training and inference. Cross data center traffic pattern has emerged to carry more AI job traffic, for example, cross data center training.

With rapid development of large language models (LLMs), more compute instances need to be deployed in data centers, approaching to the energy and physical space limit of each single data center. Under such circumstance, operators try to train models across data centers over long distance. According to the parallel computing strategies, there are three major types, data parallel(DP), tensor parallel(TP), and pipeline parallel(PP). TP consumes more bandwidth resource, e.g., hundreds of MB for a single flow and , and has strict requirements on latency, e.g., microseconds level. therefore, TP is primarily deployed within data centers. While PP and DP consumes less bandwidth than TP, and less strict latency requirements than TP. Therefore, operators consider to deploy PP and DP traffic across data centers.

For example, considering training a LLM with tens of billions of parameters like Llama3 with a thousand of Graphics Processing Units (GPUs) across two data centers, that is, deploying around 500 GPUs in each data center. The bandwidth of a single network interface card (NIC) is 200Gbps, and the total bandwidth requirements is 102.4Tbps. Under hybrid parallel computing strategies, the total volume of the cross data center traffic is the product of the number of TPs, the number of PPs, and the GPU bandwidth, which equals to 12.8Tb. These traffic is not time sensitive to microseconds level, but still needs to be transmitted as quickly as possible, thereby the overall training efficiency can be improved.

In this case, the transport protocol is based on Remote Direct Memory Access (RDMA). Traditionally, in controlled environments like data centers, these traffic is implemented over RDMA over Converged Ethernet version 2 (RoCEv2). However, since these traffic will be transmitted over shared network or even Internet, the performance as well as security both require guarantee. The iWARP protocol suite works over TCP and can provide security guarantee, but its performance in throughput optimizations need further improvement.

3.3. Sharing Traffic Between Dedicated Network and Non-dedicated Network

This case is more on sharing bulk data transfer service between dedicated network and non-dedicated network. Institutes and Universities are usually connected to dedicated networks like education and research networks, but data needs to be spread across public operator’s networks to remote consumers. In this case, there exist some difference regarding to QoS and security policies.

In dedicated networks, traffic types can be relatively simple, and the QoS mechanisms like bandwidth and priority guarantee can be flexibly adjusted. While in non-dedicated networks, the QoS mechanisms might not be the same as dedicated networks when traffic types vary a lot. Priorities of different flows may change when traffic are transmitted across a dedicated network and another non-dedicated network. Meanwhile, the security policies differ in two separate domains. The firewall or traffic shaper deployed in each domain may have some access requirements for traffic from other networks.

In this case, policies related to transport layer need further coordination between two different networks, especially on QoS guarantee and security.

3.4. Summarization of the Characteristics of Use Cases

This section summarizes the differences of the three aforementioned use cases from the following aspects. Figure 1 shows the comparison.

* What transport protocols they use?
* How large are the volume of the data that need to be transmitted?
* What are the time constraints?
* What security aspect do they care?
* What data transfer applications are currently been in use?

The basic performance indicators of the shared operator network environment is as follows.

* Packet loss rate, around 0.1%.
* Number of hops, 5 to 20.
* Transmission distance, over 1000 km.
* bandwidth, 1 to 10 Gbps at the access network, 10Gbps to 100Gbps at the core network.

Use Cases	Protocol	Time constraints	Data Volume	Security	Transfer tools
Large File Transfer over Shared Network	TCP	tens of minutes to several hours	TBs ~ 10 TBs /job	Data integrity	FTP/RSYNC
Cross Data Center Traffic	RoCEv2/ iWARP	as timely as possible	~ 10 TBs/job	Data integrity	Depend on AI Framework
Non-dedicated Network to Dedicated Network	TCP/ QUIC	tens of minutes to several hours	TBs ~ 10 TBs /job	Data encryption	FTP/RSYNC

Figure 1: Comparison among Use Cases

4. Requirements

According to the three different use cases mentioned in the previous section, some requirements on transport layer are proposed in this section.

As mentioned above, in different scenarios, there are different transport protocols carrying bulk data transfer services. For example, TCP, QUIC, RDMA, iWARP. Traffic with different transport protocols may run over the same public operator's network. Optimizing each transport protocol may incur much overhead, including congestion control algorithms design and parameter tuning, hardware adaptation, QoS policies, etc. To avoid such overhead, operators think about more flexible deployment solutions and propose some new requirements on transport protocol proxy.

R1: It MUST support transport protocol proxy which adapts one transport protocol to another, for example, TCP to iWARP or QUIC to iWARP. This is to guarantee that the transport layer mechanisms like congestion control and flow control remain consistent between two proxies, when applications at endpoints run different transport protocols.

R2: The proxy MUST support traffic classification, flows or sessions that need acceleration (increase priorities or guarantee bandwidth) can be selected at the proxy, while some other normal flow can be transmitted transparently.

R3: The proxy SHOULD support the aggregation of some mouse flows or the split of an elephant flow into multiple flows, based on the traffic classification.

When implementing iWARP in the proxy, the hardware resource needs to be considered to guarantee transmission performance. Some work may choose to implement simplified iWARP stack in the proxy, therefore,

R4: If iWARP is selected as the transport protocol, SHOULD support the transform of operation types, for example, from Unreliable Delivery (UD) mode to Reliable Connection (RC) mode.

Congestion control is always an important issue for data transmission over shared network where congestion might not be precisely located and predicted. Usually, HPWAN flows are large and may incur congestions into the network. There are already many congestion control algorithms that have been standardized or being standardized, like [I-D.ietf-ccwg-bbr], [RFC9438], and [RFC9330]. But they have poor convergence speed based on blind transmission with rate adjusting due to the unpredictable behaviour of non-dedicated network. The network should collaborate with the client to perform active congestion avoidance such as resource scheduling, rate-based acknowledgement to achieve the completion time.

The proxy can get information (topology, link bandwidth, queue and buffer) from a centralized controller of the WAN, for example, a SDN controller. The proxy can serve as an agent of the SDN controller of each network domain, and can exchange information with clients. The proxy can also serve as an intermediate node for congestion information sharing with clients.

R5: It MUST support signaling from client to the proxy, including the QoS guarantee request for scheduled traffic, for example, bandwidth, traffic pattern. Then the transport protocol proxy can negotiate with the SDN controller to design more flexible QoS policies and the corresponding resource planning (like bandwidth) and traffic scheduling.

R6: It MUST support signaling from the proxy to the client, including the response of negotiated rate for the client to send traffic and the fast and accurate quantitative feedback when proxy performs active admission control.

Some of the use cases mentioned in section three have strong security requirements, for example, biology data, financial data, and AI training data, etc. Therefore,

R7: The data integrity MUST be guaranteed for these services, especially transferring data over non-dedicated network where data may be exposed to attacks.

Also the security policies may differ when traffic is transmitted over dedicated network as well as non-dedicated network. Thus,

R8: The security policies of dedicated network as well as non-dedicated network SHOULD be exchanged, although this might not be implemented on transport layer. A practical approach is to coordinate at the orchestrator or exchange information through border gateways.

5. Security Considerations

TBD.

6. IANA Considerations

TBD.

7. References

7.1. Normative References

[RFC9330] Briscoe, B., Ed., De Schepper, K., Bagnulo, M., and G. White, "Low Latency, Low Loss, and Scalable Throughput (L4S) Internet Service: Architecture", RFC 9330, DOI 10.17487/RFC9330, January 2023, <https://www.rfc-editor.org/rfc/rfc9330>.

[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/rfc/rfc9438>.

[RFC959] Postel, J. and J. Reynolds, "File Transfer Protocol", STD 9, RFC 959, DOI 10.17487/RFC0959, October 1985, <https://www.rfc-editor.org/rfc/rfc959>.

7.2. Informative References

[I-D.ietf-ccwg-bbr] Cardwell, N., Swett, I., and J. Beshay, "BBR Congestion Control", Work in Progress, Internet-Draft, draft-ietf-ccwg-bbr-01, 21 October 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-ccwg-bbr-01>.

[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>.

[RSYNC] "RSYNC", n.d., <https://github.com/RsyncProject/rsync>.

Contributors

Hongwei Yang
China Mobile
Email: yanghongwei@chinamobile.com

Guangping Huang
ZTE Corporation
Email: huang.guangping@zte.com.cn

Authors' Addresses

Kehan Yao
China Mobile
Email: yaokehan@chinamobile.com

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