Consideration for High-performance WAN

Quan Xiong(ZTE)

IETF 120 HP-WAN Side Meeting, July 2024

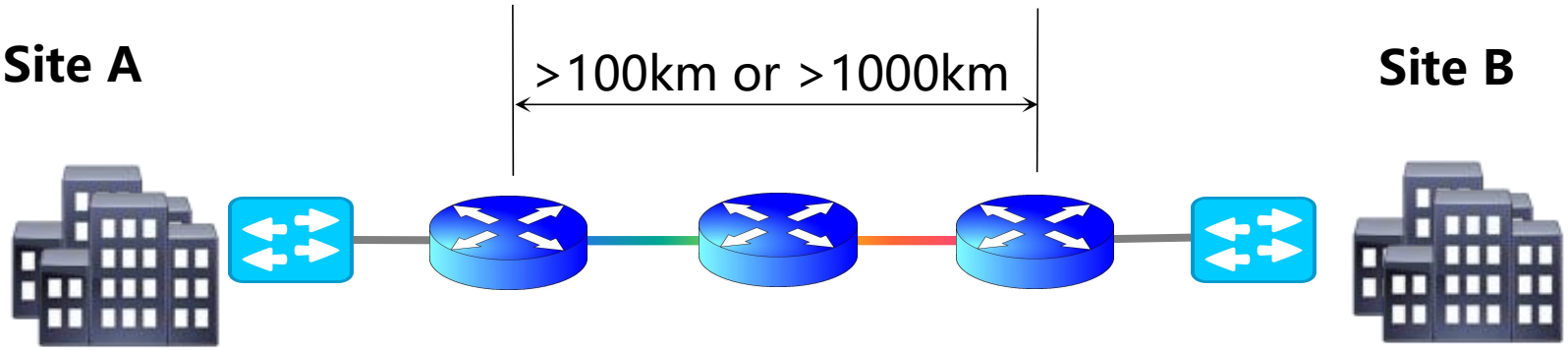
Motivation and Problem Statement

- With the rapid development of big data and intelligent computing, there are many applications requiring massive data transmission through long-distance connection, such as cloud storage and backup of industrial internet data, digital twin modeling, Artificial Intelligence Generated Content (AIGC), multimedia content production, distributed training, High Performance Computing (HPC) for scientific research and so on.

- The industries need to solve the problems such as long distance, slow feedback, course-grained load balance, low throughput and so on.

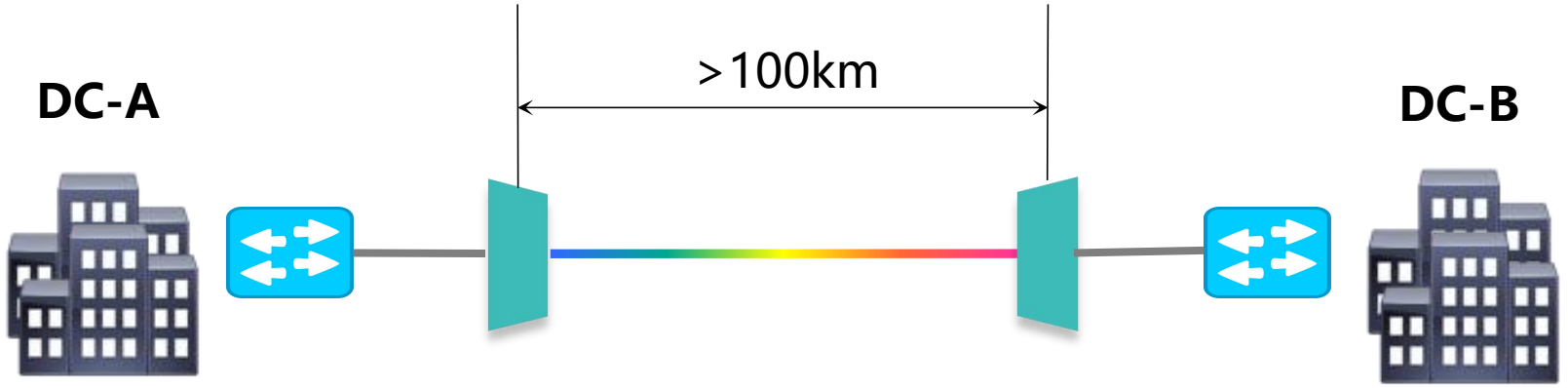
- It is required for High-performance WANs (HP-WAN) to achieve effective and reliable transmission which demands higher performance such as ultra-high bandwidth utilization, ultra-low packet loss ratio and low latency and jitter in WANs.

Scenario 1: Long-distance Transmission



- Typical Use Cases：
  - *HPC for Scientific Research*
    - generates 60~100 GB of data every five minutes, requiring data transmission from one laboratory to another for analysis.
  - *Distributed Storage*
    - move data from one storage system to another and needs to maintain data consistency across the distributed storage systems.
  - *Data Express Service*
    - requires task-based data transmission, point-to-point model, high resilience and throughput, with single data ranging from TB to 100TB.
  - *Multimedia Content Production*
    - the raw material data of a large-scale variety show or film with a single transmission of data in the range of 10TB to 100TB and needs to be transmitted between data centers or different storage sites.
  - *Data backup and Disaster Recovery*
    - the master and slave data centers are built in different locations and requires long distance and massive data transmission for disaster recovery.

Scenario 2: Collaborative Distributed Computing



- Typical Use Cases：
  - *Collaborative AI Training across Multiple DCs*
    - provide on-demand task allocation to different clusters, sufficient bandwidth, low latency, high throughput, and extremely high network availability and reliability for data centers communications.
    - parameters exchange significantly increases the amount of data transmission across DCs, typically from tens to hundreds of TBPS.
  - *Collaborative Cloud Computing across Multiple DCs*
    - cloud computing where resources and services are distributed and managed across multiple data centers, often located in different geographical locations.

Objectives

- What are the characteristics for HP-WAN?
  - Massive elephant flows data with large burst, multiple concurrent services co-existed with dynamic flows *(e.g. 10G~400Gbit/s)*
  - Long distances, multiple hops, paths and domains between DCs *(e.g. >100km, >1000km)*
- What are the objectives and goals for HP-WAN?
  - The primary goal is higher performance such as sufficient bandwidth, effective high-throughput and low latency transmission, the performance indicators includes
    - *ultra-high bandwidth utilization*
      - efficient use of available network capacity to maximize data transfer rates
    - *ultra-low packet loss ratio*
      - the packet loss negatively correlates with throughput
    - *low latency and jitter*
      - the RTT negatively correlates with throughput
      - collaborative computing demands low latency to synchronize the state of the system across multiple DCs

Impact of the Performance Indicator

- The computing method of throughput using TCP and QUIC transmission protocol is as following shown.

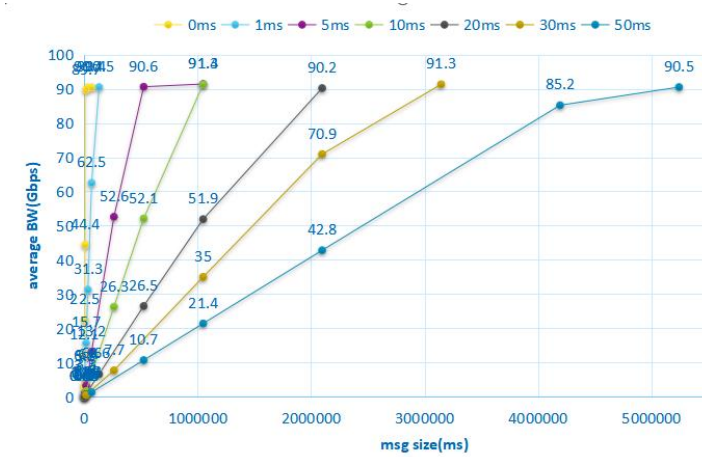
$$Throughput_{TCP} = \min\{BW, \frac{WindowSize}{RTT}, \frac{MSS}{RTT} \times \frac{C}{\sqrt{p}}\}$$

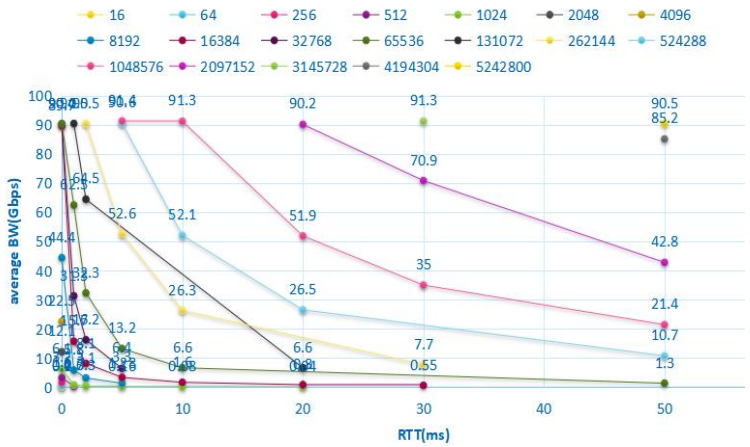
- The computing method of throughput using RDMA transmission protocol is as following shown.

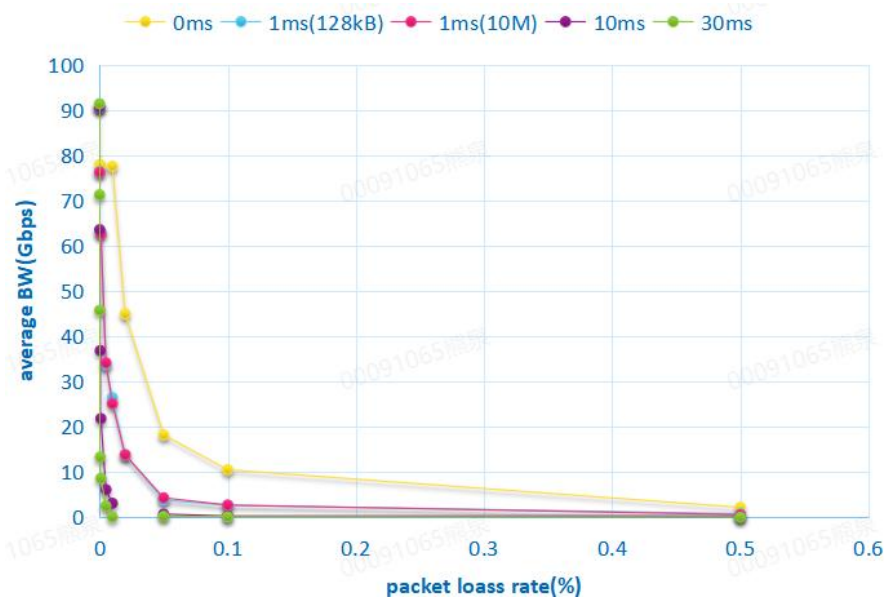
$$Throughput_{RoCE} = \frac{MSS}{RTT} \times \sum_{i=1}^{\frac{Messagesize}{MSS}} (1-p)^i$$

Performance Impact for RDMA Transmission

- Throughput increases linearly with message size.
- As RTT is low (<5ms), throughput decreases exponentially with latency and as RTT increased and it showed a linear decrease.
- The throughput decreases exponentially with packet loss rate.

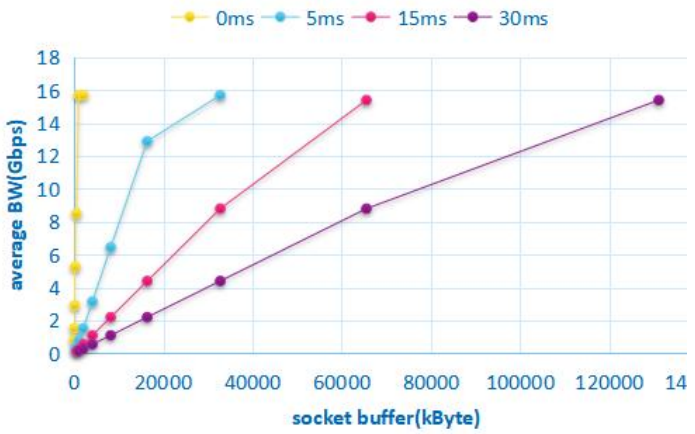


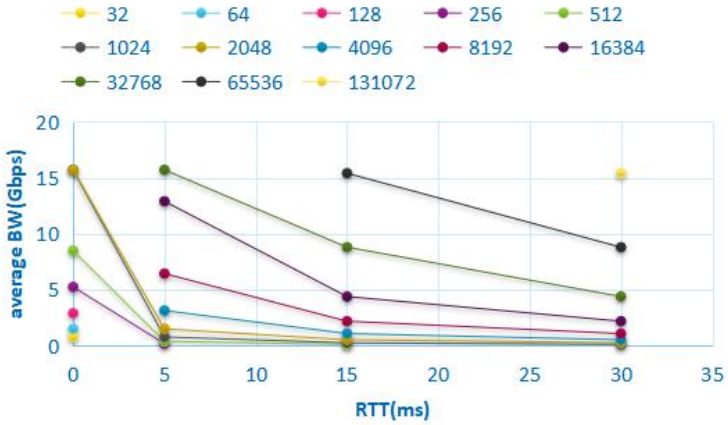


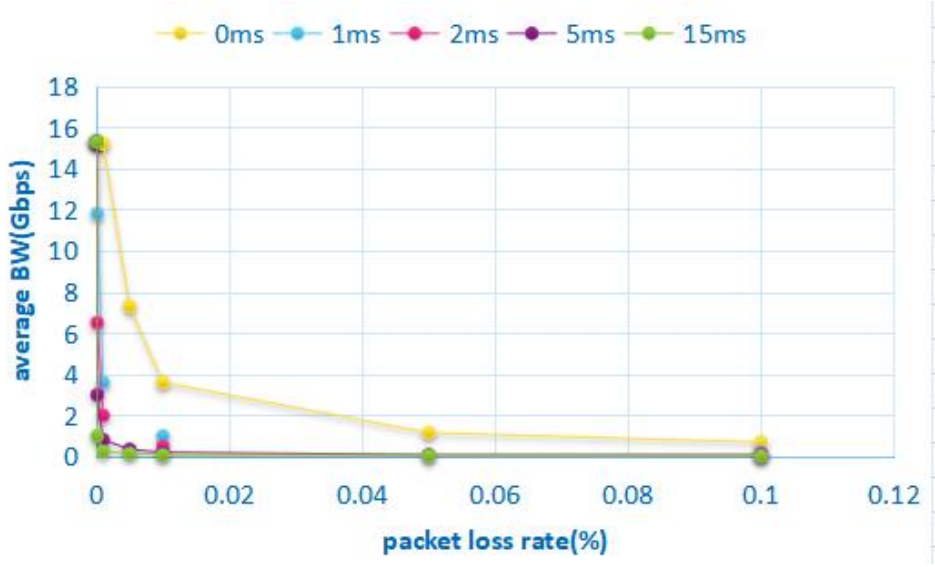


Performance Impact for TCP Transmission

- The throughput increases linearly with the TCP socket buffer and window size.
- As RTT increases, the throughput ramp up curve slows down.
- In long-distance transmission, extremely low packet loss rates can reduce TCP throughput by over 90%.
- TCP is more sensitive to packet loss rate compared to RDMA .







Gaps for Existing Technologies

- What are the gaps for existing technologies?
  - Optic Fiber direct connection *(e.g. OTN),*
    - *limited scale and deployment and high cost, requires using IP network resources*
  - DC Technologies *(e.g. PFC)*
    - *slow feedback and high Round-Trip Time (RTT) latency and jitter, requires improving flow control precision*
  - L3 Routing Technologies *(e.g. ECN/ECMP)*
    - *network is passive and unware of the status, requires coordination with the end systems*
    - *network resources utilization rate is low , requires fine-grained traffic scheduling*
    - *long-distance transimission requires ultra-low packet loss, long-distance latency and jitter guarantees*

Technical Requirements

- Support High-precision Flow Control
- Support Congestion Control based on End-network Coordination
- Support Muti-path Load Balance
- Support the Differentiated Traffic Scheduling
- Support Flow-based Network Monitoring

Related Drafts links for your reference

- https://datatracker.ietf.org/doc/draft-xiong-rtgwg-use-cases-hp-wan/
- https://datatracker.ietf.org/doc/draft-xiong-rtgwg-requirements-hp-wan/

Thank you!