# HPWAN Use Cases and Requirements from Public Operators’ View

**Kehan Yao, China Mobile**

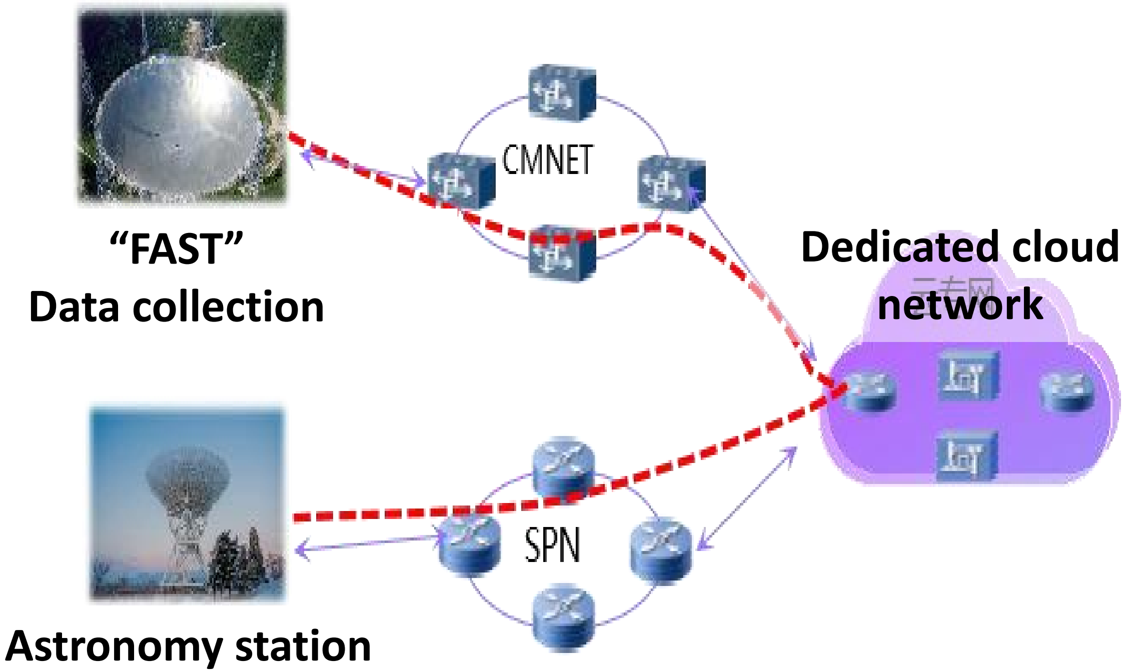
Quan xiong, ZTE

Yangyang Wang, Tsinghua

IETF 122 HPWAN Side meeting

# Use Case #1: Large Volume Data Transfer over Shared Network

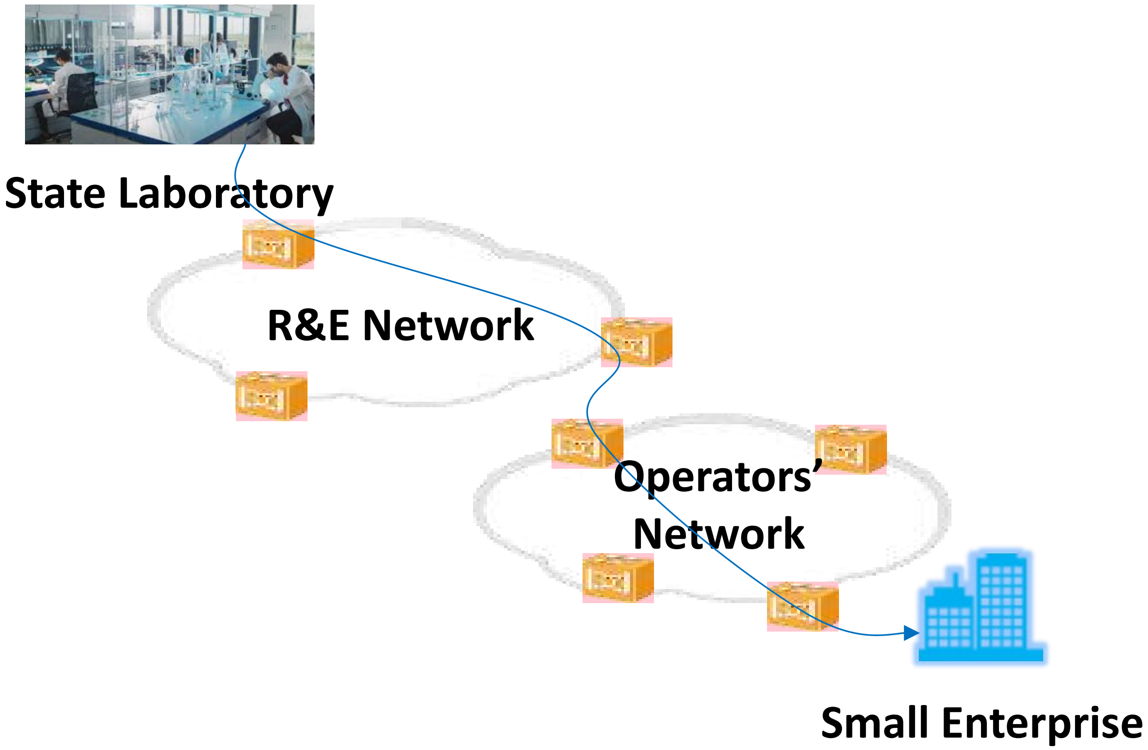
- Large volume data transfer (LVDT) includes scenarios like biology and astronomy observation, etc.
- For example, Five-hundred-meter Aperture Spherical Telescope (FAST) tansfers astronomy to astronomy stations
- In shared operators' networks, the total transmission duration is influenced by packet loss, resource contention, QoS policies, etc.




Large Volume Data Transfer over Shared Operator's Network

# Use Case #2: Cross Multi-entities' Network Transfer

- Small enterprises want to request for research data owned by colleges or state laboratories.
- The data is transferred through R&E network (dedicated) as well as operator's network (non-dedicated)
- R&E networks like ESNET can guarantee efficient transmission, but through the SDN-based scheduling of workflows [1]
- There maybe different QoS policies for multiple jobs, so it's hard to guarantee E2E traffic scheduling



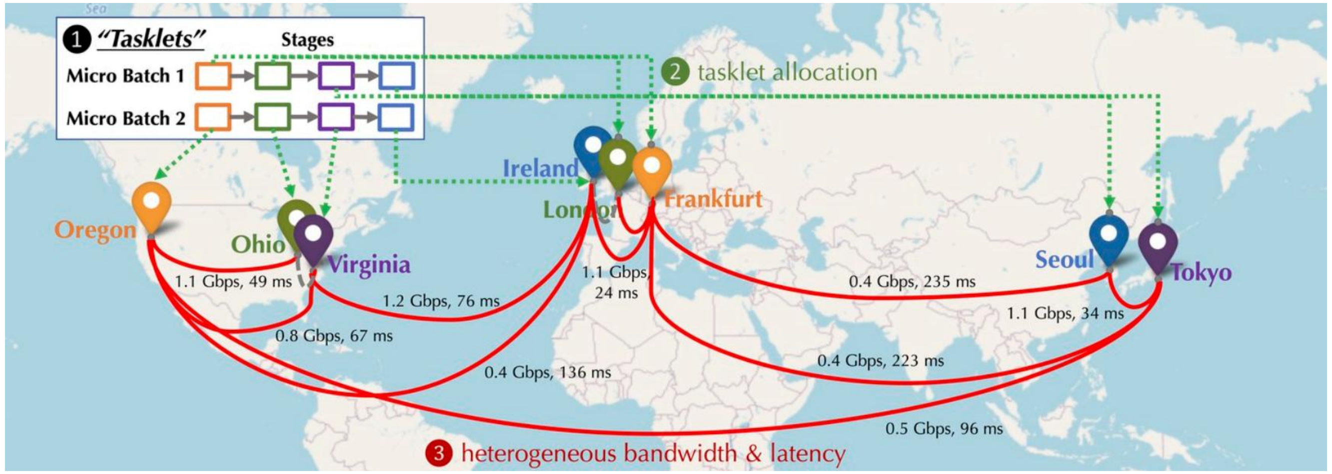

**Cross Multi-entitie' Network Large Volume Data Transfer**

[1] https://sense.es.net/architecture

# Use Case #3: WAN-grade Decentralized Training

- It's been technically proved that WAN-grade training is possible, considering some coordination design on parallel algorithms and the compute resources.

- To facilitate the speed of the overall training efficiency, data transmission across nodes must be **as timely as possible**.

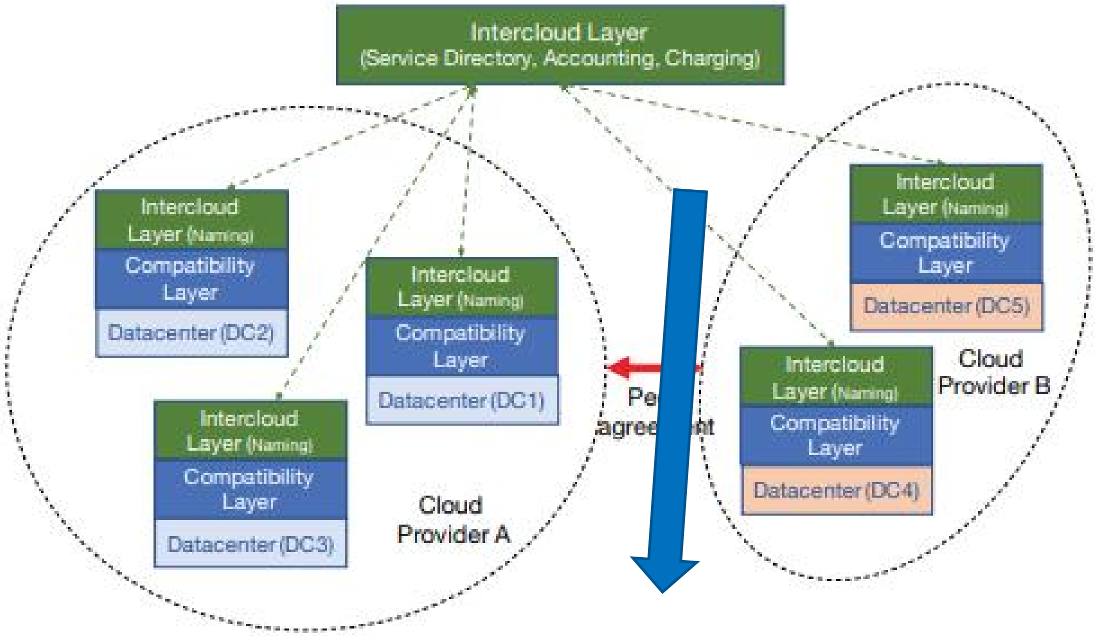



**Worldwide Decentralized Training [2]**

[1] https://www.primeintellect.ai/blog/intellect-1-release
[2] https://www.together.ai/blog/neurips-2022-overcoming-communication-bottlenecks-for-decentralized-training-12

# Use Case #4: Multi-clouds Sky Computing

- Muti-cloud Sky computing seems more a use case from hyperscalers.
- Public operators also build large data centers, like telecom cloud.
- Small enterprises need to rent heterogeneous cloud resources for computation and even cross-clouds computation [1].
- Jobs require efficient transmission across clouds and between clients and clouds.




**Underlying Network between Multi-clouds Require HPWAN Capabilities**

[1] https://sky.cs.berkeley.edu/

# Use Cases Summarization

中国移动 China Mobile

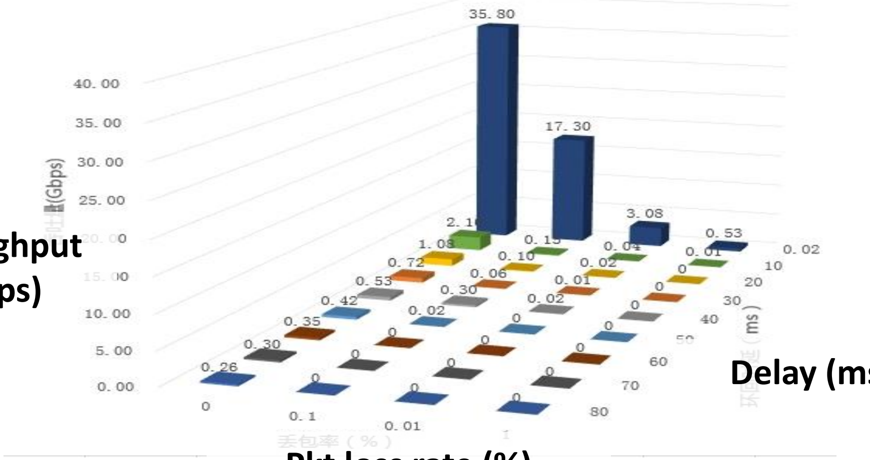
| Use Cases | Protocol | Time constraints /QoE expectation | Data Volume | Transfer tools | Workflow Scheduling | Extremely low packet loss rate |
|---|---|---|---|---|---|---|
| **LVDT over Shared Network** | TCP | tens of minutes to several hours | TBs ~ 10 TBs /job | FTP/RSYNC | Centralized | as low as possible |
| **Cross Multi-entities' Network** | TCP | tens of minutes to several hours | TBs ~ 10 TBs /job | FTP/RSYNC | Decentralized | as low as possible |
| **WAN-grade Decentralized Training** | RDMA /TCP | as timely as possible | TBs /job | Open-sourced AI framework | Decentralized /Centralized | as low as possible |
| **Multi-clouds Sky Computing** | RDMA /TCP | as timely as possible | TBs /job | Open-sourced AI framework / Big data platform | Decentralized | as low as possible |

Some typical settings in public operators' shared networks:

- 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.

# Existing Approaches - TCP

- FTP/RSYNC are based on TCP.
- TCP doesn’t scale well when BDP increases.
- TCP + BBR may work well in well-scheduled networks, e.g., Google’s effingo.
- But in networks where measurement is not stable, it may require more time to converge and for the throughput to be increased.
- BBR also requires large buffer queues and not friendly to fairness [1].
- Can’t efficiently utilize the network capacity.




| | TCP+BBRv1 0.1%Pkt loss | TCP+BBRv1 1%Pkt loss | TCP+CUBIC 0.1%Pkt loss | TCP+CUBIC 1%Pkt loss |
|---|---|---|---|---|
| Single Stream | 14Gbps | 10Gbps | 8.6Mbps | Null |
| 3 Streams | 41Gbps | 24.5Gbps | 28Mbps | Null |
| 10 Streams | 70Gbps | 61Gbps | 91Mbps | Null |
| 25 Streams | 84Gbps | 84.7Gbps | Null | Null |

Figure 2: TCP throughput performance,RTT=70ms,MTU=1500

**TCP throughput performance test**
**(With TOE enabled)**


[1] Yi Cao, Arpit Jain, Kriti Sharma, Aruna Balasubramanian, and Anshul Gandhi. 2019. **When to use and when not to use BBR: An empirical analysis and evaluation study**. In Proceedings of the Internet Measurement Conference (IMC '19). Association for Computing Machinery, New York, NY, USA, 130–136. https://doi.org/10.1145/3355369.3355579

# Can QUIC be a better substitute solution?

中国移动 China Mobile

- QUIC does have requirements for high volume data transmission, since QUIC-based traffic are increasing and QUIC guarantees E2E data encryption.
- QUIC primarily reduces the latency in initial session establishment, but not perform very well in throughput.
- BBR can help QUIC increase E2E throughput performance, but worse than TCP.
- It occupies large CPU resources and the throughput is not good even CPU is saturated.

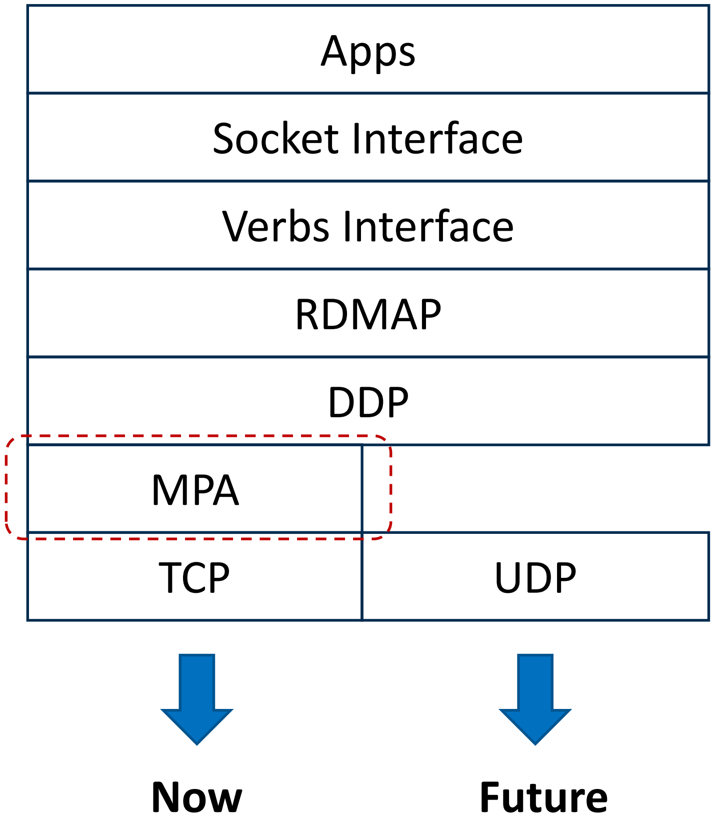
| | QUIC+BBRv1 0.1%Pkt loss 40 cores | QUIC+BBRv1 0.1%Pkt loss 80 cores |
|---|---|---|
| 40 Streams | 47.2Gbps | 52.8Gbps |
| 60 Streams | 42.4Gbps | 57.2Gbps |
| 80 Streams | 51.2Gbps | 62.4Gbps |
| 100 Streams | NULL | 63.2Gbps |

**QUIC throughput performance test (With TOE enabled)**

# RDMA(iWARP) in a glance

中国移动 China Mobile

- MPA layer is the bottleneck for iWARP throughput performance
- Need some modifications in each layer to bypass MPA, like UDP and Socket adaptation
- Improving congestion control and flow control on top of UDP
- Hard to implement since the stack is closely related to hardware
- Need more implementation results on RDMA over modified QUIC or RDMA over enhanced UDP



# Considerations on Architetural Aspects and Requirements

中国移动 China Mobile

- Congestion control -- coordination with CCWG
- End-to-end flow control, like HBH backpressure for ultra low pkt loss (PFC, IEEE std [1])
- Host-to-network signalling [2]
    - admission control -- coordination with TSVWG
    - traffic classification
    - traffic aggregation
    - traffic scheduling/dispatching
- Network-to-host signalling -- coordination with SCONE
    - rate control
- Proxying,
    - protocol transformation for
    - exposure of information for better traffic
- RDMA capabilities -- coordination with NFSv4
- Encryption and security

[1] https://1.ieee802.org/dcb/802-1qbb/
[2] https://datatracker.ietf.org/doc/draft-kwbdgrr-tsvwg-net-collab-rqmts/