Research Statement

Zili Meng, HKUST

The metaverse is coming – we are closer than any time to embrace real-time interactive multimedia streaming applications (e.g., cloud gaming, remote surgery, and virtual reality). However, while some applications (e.g., video-conferencing) have entered our lives, the others are still far from *a part of our life*. For example, we cannot wear VR headsets all day as normal glasses. People suffer from frequent stalls[1], feel sick due to motion lags[2], or have fears[3].

As a network researcher, my goal is to build networked systems to bring those real-time multimedia applications to our lives. This is challenging because:

i **Extreme latency requirements of applications.** Internet provides high performance in the median[4], but the *tail latency* still sucks. Yet, real-time multimedia applications require an extremely *consistent* low latency (e.g., the 99.99th percentile): Even stuttering of 0.01% has drastic consequences on critical applications such as remote surgery or remote-driving cars.

ii **Complicated causes of tail latency.** When we focus on the 99.99th percentile, the contributing factors of latency have changed. Causes can be competition from other users, interference from other devices, or fluctuation in the client, some of which might be dependent. The scenario will be even more complicated when the video quality goes up.

To this end, I work across the whole network stack to analyze potential causes of latency spikes at the tail for real-time multimedia applications, and challenge several long-standing assumptions.

- Previous work optimizes the *data path latency* (the time that packets travel from the sender to the receiver). I show that the *control path latency* (the time from the network changes to the sender reacts) is also critical (§1).
- Current designs react to network changes (*e.g.,* packet losses or flow competitions) individually in different layers. I find that the lack of coordination can lead to a huge transient increase of the delay (§2).

Impact: Owing to my intrigue for *real-world* systems, my research has received attention widely. Several of my research has been used in production by Tencent, Alibaba, etc, serving millions of users. I'm also honored to receive the Microsoft Research Fellowship[5], the ByteDance PhD Fellowship[6], and several paper awards[7] from the community.

1 *Control Path Latency*

Existing real-time multimedia applications can degrade transiently due to the delayed signal on the control path, which mainly contains two parts:

- Latency of measurement feedback: the time from a change happening in the network to the sender knowing it happened.
- Latency of decision making: the time from the sender knowing the network change, to the sender finally reacting to the change.

Measurement Feedback. Ideally, senders will reduce the bitrate when bandwidth reduces and congestion occurs. Unfortunately, senders are fundamentally limited in how quickly they can react because congestion signals need to be carried back from the network. I showed that the inflation of feedback delay can drastically compromise the tail latency on the data path. In response, Zhuge[8] for the first time accelerates the feedback time by manipulating the intervals between feedback packets, and significantly reduces stalls for real-time multimedia applications. Zhuge attracted attention from the media[9].

[1] Future of cloud gaming | Deloitte Insights
[2] Virtual reality sickness - Wikipedia
[3] Public's perception of robotic surgery
[4] For example, the maximum throughput for state-of-the-art WiFi is 9.6 Gbps.

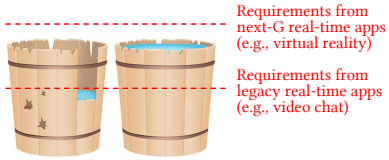

Figure 1: How much water the barrel can contain is decided by the shortest board. The network used to be the shortest board in latency optimization. However, many more components can be the shortest board when optimizing the tail latency.

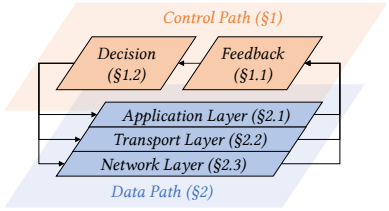

Figure 2: Network researchers usually split different functionality into different layers. My prior (§1, §2) research works on all layers from the network layer and above to optimize end-to-end latency. I have also recently started to collaborate with sister communities (§3) to optimize the latency.

[5] Microsoft Research Fellows (2020)
[6] ByteDance Scholar Program (2022)
[7] SIGCOMM 2018 SRC Gold Medal, IEEE/ACM IWQoS 2021 Best Paper Award, IEEE ICC 2020 Best Paper Award

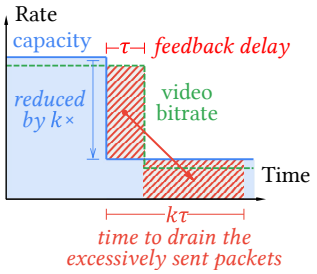

Figure 3: The feedback loop plays a dominant role when the network capacity drops. The solid blue line denotes the drop in link capacity in the bottleneck, and the dashed green line denotes how the sender reconverges to the new sending rate. Users suffer from interaction lags when draining excessively sent packets during the feedback.

[8] **Zili Meng**, Y. Guo, C. Sun, B. Wang, J. Sherry, et al. Achieving Consistent Low Latency for Wireless Real-time Communications with the Shortest Control Loop. In *Proc. ACM SIGCOMM*, 2022
[9] APNIC Blog

Decision Making. The decision-making of rate adaptation algorithms is increasingly time-consuming and unreliable since the algorithms shift from heuristics to neural networks. This is critical because decisions in networks are made at a high frequency, and delayed or incorrect decisions will degrade the tail performance. I unveil that neural networks are used in networks not because they are *more expressive*, but *easier to optimize*. Thus, I decouple offline optimization and online execution, and convert complex models into lightweight decision trees for multimedia streaming, which reduces decision-making latency[10] and the unreliable decisions[11]. This work has been received well in the community[12], and also adopted by Tencent in production.

2 Data Path Latency

Network researchers split the communications into different layers (Figure 2) for modularity. My research revisits the causes of latency due to lack of corporation between different layers for real-time multimedia applications.

Application Layer. The high frame-rate and resolution of real-time multimedia applications increase the processing time at the application layer and create new bottlenecks between the network stack and the application. However, the buffer between them is not designed for low latency and lacks active management. Therefore, I propose AFR[13] to achieve ultra-low latency by actively managing the queue between the network stack and application. AFR coordinates the video codec and network stacks to reduce the queueing delay, and has been deployed in Tencent for one year.

Transport Layer. The recent development of new infrastructure (5G and WiFi 6) and new architecture (edge computing) have reduced the network propagation delay from 100-200ms to 10-20ms. Such a latency has now gone across a boundary – the median network latency is lower than the human's perception ability. This shakes a common belief in real-time streaming applications – now we should control the latency exactly what the application needs but not as low as possible. In the transport layer, I design a new packet loss recovery mechanism, Hairpin[14], which coordinates the requirement from applications with recovery choices to improve the tail latency.

Network Layer. Currently, it is always end hosts that take care of the behaviors of the capricious network. When the network changes, rate adaptations on the end hosts cover for the network to provide a consistent experience for the application. However, when we are looking at the 99.99th percentile, even the transient the covering-up process will lead to the latency increase. In this case, the network itself should be more responsible and consider end hosts' reactions during resource allocation. I then design Confucius[15], a new queue management mechanism, to retroactively make the network change to follow the sender's rate adaptation. Confucius also significantly improves the transient performance and is able to patch on most commercial routers.

3 Future Work

I intend to leverage my knowledge toward three main directions: facilitating cross-community knowledge (§3.1), generalizing to more real-time applications (§3.2), and bringing insights back to the Internet architecture (§3.3).

3.1 Cross-community Optimization

Real-time multimedia applications involve many sibling communities in both systematic and algorithmic aspects, which I would like to collaborate with.

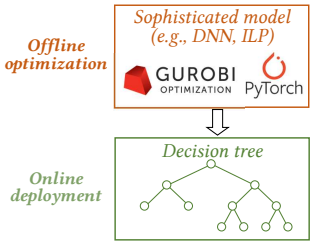

Figure 4: By decoupling the offline optimization from online deployment, my solutions can reduce the uncertainty in the decision-making in rate adaptations.

[10] **Zili Meng**, J. Chen, Y. Guo, C. Sun, H. Hu, et al. PiTree: Practical Implementation of ABR Algorithms Using Decision Trees. In *Proc. ACM Multimedia*, 2019

[11] **Zili Meng**, M. Wang, J. Bai, M. Xu, H. Mao, et al. Interpreting Deep Learning-based Networking Systems. In *Proc. ACM SIGCOMM*, 2020

[12] SIGCOMM '20 (Top 10 by citations).

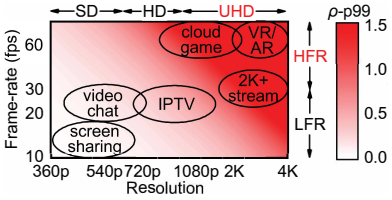

Figure 5: For next-generation real-time multimedia applications, with the increase of resolution and frame-rate, the queue between network and application will be overloaded. ρ is the load of that queue.

[13] **Zili Meng**, T. Wang, Y. Shen, B. Wang, M. Xu, et al. Enabling High Quality Real-Time Communications with Adaptive Frame-Rate. In *Proc. USENIX NSDI*, 2023

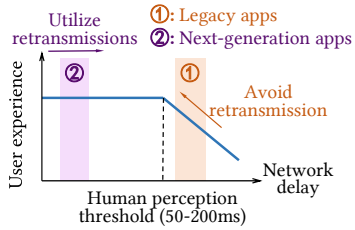

Figure 6: Existing solutions strive to avoid retransmission to reduce the end-to-end latency. With a reduced propagation delay, users can now tolerate retransmissions due to the perception ability of humans.

[14] **Zili Meng**, X. Kong, J. Chen, B. Wang, M. Xu, et al. Hairpin: Rethinking Packet Loss Recovery in Edge-based Interactive Video Streaming. In *Proc. USENIX NSDI*, 2024

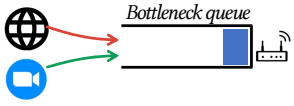

Figure 7: When a real-time multimedia application competes with a Web application, in the time when the Web application joins the competition, the real-time application suffers a transient latency spike.

[15] **Zili Meng**, N. Atre, M. Xu, J. Sherry, and M. Apostolaki. Confucius Queue Management: Be Fair But Not Too Fast. In *Submission*, 2023

Systematic – Multi-component coordination. The network is not the only component contributing to tail latency. For example, video encoding and decoding, CPU scheduling, and graphics rendering can also have latency jitters during runtime. I have some preliminary attempts to co-design with the video codec to improve latency in the application layer[16]. I am also eager to collaborate with the wireless and operating system community to see if coordination with the underlying wireless protocols and above system schedulers can bring additional benefits to real-time multimedia applications.

Algorithmic – Optimization with advanced algorithms. With more and more components that need to be considered when making decisions, the state and the action space of decision-making also increase. In this case, we do need some tools to automate the process of designing new algorithms. I made some preliminary attempts to apply deep learning-based algorithms in scheduling[17] and congestion control[18]. With efficient and optimal algorithms and solvers popping up, I am also working on how to better utilize these tools to solve problems in real-time streaming applications.

3.2 Generalization to Other Applications

The demand for real-time transport is beyond the scope of multimedia transport. Especially in recent years, many more applications require real-time transport, which I would also like to work on.

Vehicle network. In emerging applications such as autonomous driving and vehicle-to-everything (V2X), reliable transport inside or between vehicles is very critical. A delayed signal from sensors to the vehicle can cause the car to result in a crash. During my internship at a switching ASIC startup, I proposed a queue scheduling algorithm by utilizing the time-sensitive network to provide deterministic latency for data communications inside vehicles[19]. I also plan to extend my previous research efforts to vehicle networks.

Internet of Things. With the development of the Internet of Things (IoT), the sensors in daily life will significantly increase. Among them, real-time communication between the sensor and controllers is also necessary. However, a key feature of IoT is that the volume of traffic is massive – thousands of sensors are talking to each other. I intend to address the challenges and build real-time IoT systems by leveraging networking advancements.

3.3 Getting Insights on Internet Architecture

I am also rethinking whether the current Internet Architecture is capable of providing consistent low latency for real-time multimedia applications.

Layering architecture. Looking back to the history of several decades, one key to the success of the Internet is its layering architecture – researchers split the end-to-end transport into different layers, each responsible for certain functionality. However, to achieve extreme tail performance for real-time multimedia applications, many cross-layer designs in fact break through the layering architecture. I am thus motivated to rethink how we should split between different layers for more consistent performance.

Best effort design principle. The best-effort service on the network layer has been the fundamental assumption of the Internet for decades. However, such a service is not capable of delivering consistent latency with guarantees. For example, during my intern at a company for vehicle networks, their solution replaces the best-effort network with the time-sensitive network for latency guarantees. Thus, I also plan to investigate how the network layer can be optimized for real-time applications within the current infrastructure.

[16] T. Wang, **Zili Meng**, M. Xu, R. Han, and H. Liu. Enabling High Frame-rate UHD Real-Time Communication with Frame-skipping. In *Proceedings of the 3rd ACM Workshop on Hot Topics in Video Analytics and Intelligent Edges (HotEdgeVdeo '21), co-located with MobiCom '21*, 2021

[17] H. Mao, M. Schwarzkopf, S. Venkatakrishnan, **Zili Meng**, and M. Alizadeh. Learning Scheduling Algorithms for Data Processing Clusters. In *Proc. ACM SIGCOMM*, 2019

[18] J. Zhang, E. Dong, **Zili Meng**, Y. Yang, M. Xu, et al. WiseTrans: Adaptive Transport Protocol Selection for Mobile Web Service. In *Proceedings of the Web Conference (WWW '21)*, 2021

[19] Z. Ruan, **Zili Meng**, and Y. Huang. A Scheduling Method, Device, and Electronic Equipment, 2022. Chinese Patent CN113872887A. Granted: August 16, 2022

Figure 8: We are closer to a lot of real-time applications in our daily life than any other time in history. In the future, I will strive to extend my research to all real-time applications and build the Internet to be capable of delivering real-time contents (photo credit: great.gov.uk).