Internet Draft
Intended status: Proposed Standard
Expires: January 2021

W. Huang
Y. Zhang
Tencent
R.Yang
Yale University
C. Xiong
Y. Lei
Y. Han
Tencent
G. Li
CMRI
July 13, 2020

MoWIE for Network Aware Application

draft-huang-alto-mowie-for-network-aware-app-01

Abstract

With the quick deployment of 5G networks in the world, cloud based interactive services such as clouding gaming have gained substantial attention and are regarded as potential killer applications. To ensure users' quality of experience (QoE), a cloud interactive service may require not only high bandwidth (e.g., high-resolution media transmission) but also low delay (e.g., low latency and low lagging). However, the bandwidth and delay experienced by a mobile and wireless user can be dynamic, as a function of many factors, and unhandled changes can substantially compromise users’ QoE. In this document, we investigate network-aware applications (NAA), which realize cloud based interactive services with improved QoE, by efficient utilization of Mobile and Wireless Information Exposure (MoWIE) . In particular, this document demonstrates, through realistic evaluations, that mobile network information such as MCS (Modulation and Coding Scheme) can effectively expose the dynamicity of the underlying network and can be made available to applications through MoWIE; using such information, the applications can then adapt key control knobs such as media codec scheme, encapsulation and application logical function to minimize QoE deduction. Based on the evaluations, we discuss how MoWIE can be a systematic extension of the ALTO protocol, to expose more lower-layer and finer grain network dynamics.

Status of this Memo

Copyright and License Notice

Table of Contents

1. Introduction of Network-aware Applications

With the quick and widely deployment of 5G network in the world, more and more applications are now moving to the remote cloud-based application, e.g., cloud office, cloud education and cloud gaming. Some new and amazing applications are created and hosted in the remote cloud, e.g., cloud AR/VR/MR. What's more a lot of traditional niche interactive applications are becoming widely used in daily business with the help of mobile network and cloud, e.g., cloud video conference. Especially, during the coronavirus pandemic in 2020, many peoples have to stay at home and work/study remotely, the usage of cloud applications, including cloud-based online courses, cloud-based conferencing, and cloud gaming, has surged significant.

To provide acceptable QoE to the end users via the mobile network, the cloud application needs to know the mobile network status, e.g., delay, bandwidth, jitter to dynamically balance the generated media traffic and the rendering/mixing in the cloud. Currently, the application assumes the network as a black box and continuously uses client or server measurement to detect the network characteristics, and then adaptively change the parameters as well as logical function of the application. However, when only application information is utilized, the application can’t guarantee a good QoE in some cases. First, information from application side may have delay. When a user enters some place with bad network such as elevator or underground garage, the application will not receive such information immediately. As a result, the buffer of video application may have a high chance to run out. Then the screen will freeze and users QoE will be harmed. Besides, the application does not have information about other users in the cell. Thus, it can't know how many resources it can get and when it will change. If other users enter the cell and compete the resource, the application layer may misjudge the resource and request a high bitrate. Then the delay will increase and QoE will drop. Some information from network layer like physical resource block (PRB) information and utilization rate can help to describe how many resources the user will get and how many users are competing with him. Such information is helpful to predict the network and streaming videos. However, the application can't get those kinds of information yet.

Mobile network is always pursuing standard solutions to get network dynamic indicators that can be used by applications. In 3GPP, a lot of IP-based QoE mechanism are reused. The ECN[RFC3168] has been supported by the 4G radio station (eNB) to provide CE(Congestion Encountered) information to the IMS application to perform the Adaptive Bitrate (ABR) [TS26.114].The application can downgrade the bit rate after receiving the CE indication, but does not know exact bit rate to be selected. The DSCP[RFC2474] is used to difference the QoS class and paging strategy[TS23.501],normally the application cannot dynamically change the DSCP to improve bit rate based on the network status. DASH [MPEG DASH] is a MPEG standard widely used for the application to detect the throughput of the network based on the current throughput and buffering states and adaptively select the next segment of video streaming with a suitable bitrate in order to avoid the re-buffering. SAND-DASH[TS26.247] defines the mechanism that the network/server can provide available throughput to the application, in such case, the better bitrate can be selected by DASH application.

In 5G cellular networks, network capability exposure has been specified which allows the 5G system to expose the QoS Flow establishment with AF provided QoS requirements, user device location, network status towards the 3rd party application servers modeled as AF (Application Function) [TS23.501].In such case, the AF can request the 5G to establish a dedicated QoS Flow to transport an IP flow with the AF provided QoS requirements. The 5G also can provide QNC (QoS Notification Control) to the AF if the GBR(Guaranteed Bitrate) of the established GBR QoS Flow cannot be fulfilled, and the AF can change the bitrate after receiving the QNC notification. But the AF still does not know which bitrate to be selected. So the 5G enhances the QNC with providing a list of AQPs(alternative QoS profile). with this AQP, the 5G network provides a subset of supported AQPs with the QNC, then the AF selects a bit rate from 5G network supported AQPs, in such case, the GBR can fulfilled again if the radio state of user is changed. QoS predication is realized by network function inside 5GC to collect and analyze the status and parameters from the 5G network entities, and deliver the analytics results towards the entity such as application server. However, both network capability exposure and QoS predication solutions are designed for 5G access and core network, which cannot cover the whole end-to-end network. How to enable the application to be aware of the lower layer networks in Internet scenario is an important area for both industrial and academic researchers.

2. Use Cases of Network-Aware Application (NAA)

There are three typical NAAs, cloud gaming, low delay live show, and cloud VR, whose QoE can be largely enhanced with the help of MoWIE.

2.1. Cloud Gaming

As mentioned above, cloud gaming becomes more and more popular recently. This kind of games requires low latency and highly reliable transmission of motion tracking data from user to gaming server in the cloud, as well as low latency and high data rate transmission of processed visual content from gaming server cloud to the user devices. Cloud gaming is regarded as one major killer application as well as traffic contributor to wireless and cellular networks including 5G. The major advantages of cloud gaming are easy & quick starting (no/less need to download and install big volume of software in the user device), less cost and process load in user device and it is also regarded as anti-cheating measure. Thus, the kind of gaming becomes a competitive replacement for console gaming using cheaper PC or laptop. In order to support high quality cloud gaming services, the application need to get the information from the network layer, e.g., the data rate value or range which lower layer can provide in order to perform rendering and encoding, during which the application in the cloud can adopt different parameters to adjust the size of produced visual content within a time period.

2.2. Low Delay Live Show

In 2019, over 500 million active users were using online personal live show services in China and there are 4 million simultaneous online audience watching a celebrity's show. Low delay live show requires the close interaction between application and network. Compared with conventional broadcast services. This service is interactive which means the audience can be involved and they are able to provide feedback to the anchor. For example, a gaming show broadcasts the gaming playing to all audience, and it also requires playing game interaction between the anchor and the audience. A delay lower than 100ms is desired. If the delay is too large, there will be undesirable degradation on user experiences especially in a large-scale show. To lower the latency and provide size-adjustable show content, the application also requires the real-time lower layer information.

2.3. Cloud VR

Cloud VR data volume is large which is related to different parameter settings like DoF (Degree of Freedom), resolution and adopted rendering and compression algorithm. The rendering can be performed at the cloud/network side or a mix of the cloud and the user device side. Because the latency in cloud VR is even as low as 20ms, the application may need to interact with network to get the information about the segmentation or transport block information, and these lower layers information may be dependent on different layer 2 and layer 3 wireless protocol designs.

2.4. Performance Requirements of these Use Cases

There are different bandwidth, latency and lagging requirements for the above services which are characterized as parameter range. The reason of using a range is because such requirements are related to a group of parameter settings including resolution, frame rate and the compression mechanism. We consider 1080p~4K as the resolution range, 60-120 FPS (Frames per second) as the frame rate and H.265 as an example compression algorithm. The end-to-end latency requirement is not only related to FPS but also the property of the service, i.e., for weak interactive and strong interactive services [GSMA].

With the typical parameters setting, cloud gaming generally needs a bandwidth of 20~60 Mbps , we also consider the lagging significantly happens when the latency is larger than 40~200ms, depending on the types of games (e.g. 40ms for First Person Shoot games, 80ms for Action games, and 200ms for Puzzle games).. In order to avoid bad user experiences, the lagging rate is better to be as low as zero (in an optimal QoE). For low latency live show, 20~50 Mbps bandwidth may be needed and the end-to-end latency requirements is less than 100 ms. Cloud VR service generally requires 100~500 Mbps bandwidth and 20~50 ms end-to-end latency. It is noted that these values are dependent with the parameter settings and they are provided to illustrate the order of magnitude of these parameters for the afore-mentioned use cases. These value range may be updated according to specific scenarios and requirements.

3. Current (Indirect) Technologies on NAA

The applications have tried to increase QoE with the help of network information captured from the application layer to guess the network dynamics, such as bitrate, buffer status, packet loss rate and so on. For example, adaptive bitrate (ABR) and buffer control methods to reduce delay, and application layer forward error scheme (AL-FEC) to avoid packet losing are proposed. This document focuses on two novel approaches, which have achieved good performance in practice. One is video encoding based on ROI, the other is reinforcement learning based adaptive bitrate.

3.1. Video Compression Based on ROI (Region of Interest)

A foveated mechanism [Saccadic] in the Human Visual System indicates that only small fovea region captures most visual attention at high resolution, while other peripheral regions receive little attention at low resolution. And we call those regions which attract users most, the regions of interest (ROI)[Fahad].

To predict human attention or ROI, saliency detection has been widely studied in recent years [Borji], with a lot of applications in object recognition, object segmentation, action recognition, image caption, image/video compression, etc.

Since there exists the region of interest in a video, the cloud server can give the ROI region higher rate while making other regions a lower rate. As a result, the whole rate of the video is reduced while the watching experience will not be harmed.

This method means to detect the ROI and re-allocate the coding scheme for interested and non-interested regions in order to save the bandwidth without sacrificing user's QoE. In recent years, the ever-increasing video size has become a big problem to applications. The data rate of a cloud gaming video in 1080P can reach 25Mbps, which brings huge burden to the network, even for 5G network. Those ROI-based video compression methods are mainly applied to the high concurrency network to relive the burden of networks and then keep QoE in an acceptable range.

However, current methods utilize application information like application rate and application buffer size as the indicators to roughly adjust the algorithm in interactive video services. That information is hard to reflect the real-time network status precisely. Therefore, it is hard to balance the QoE and bandwidth saving in real-time scenario. More direct information is helpful for those ROI methods to improve the performance.

3.2. AI-based Adaptive Bitrate

This method intends to reduce lagging and ensure the acceptable picture quality.

Applications such as video live streaming and cloud gaming employ adaptive bitrate (ABR) algorithms to optimize user QoE [MPC][CS2P]. Despite the abundance of recently proposed schemes, state-of-the-art AI based ABR algorithms suffer from a key limitation. They use fixed control rules based on simplified or inaccurate models of the deployment environment. As a result, existing schemes inevitably fail to achieve optimal performance across a broad set of network conditions and QoE objectives.

A reinforcement learning based ABR algorithm named Pensieve was proposed [Hongzi] recently. Unlike traditional ABR algorithms that use fixed heuristics or inaccurate system models, Pensieve's ABR algorithms are generated using observations of the resulting performance of past decisions across a large number of video streaming experiments. This allows Pensieve to optimize its policy for different network characteristics and QoE metrics directly from experience. Over a broad set of network conditions and QoE metrics, it has been proven that Pensieve outperformed existing ABR algorithms by 12%~25%.

For this method and those methods built upon this, it has been proven that all the information, such as rate, download time, buffer size or network level information which can reflect the performance are useful to the reinforcement learning [Hongzi2]. Since those data can reflect the network dynamics, they have been used to help the applications to know how to change the rate and promote the users' QoE.

However, all these data are obtained from the client side or the server side. In reality, it is not easy to obtain such data in an effective and efficient way. Lack of standardized approach to acquire these data, is difficult to make this usable for different applications for large scale deployment. Meanwhile, these data which reflect the real-time network status change rapidly and randomly which is hard to use a theoretical model to characterize.

To summarize, current practices can make some improvements by indirectly measuring network status and react in the application. However, the network status data is not rich, direct, real-time, also lacks predictability, especially when in the mobile and wireless network scenarios, which results in long react delay or high QoE fluctuations.

4. Preliminary QoE Improvement Based on MoWIE

4.1. MoWIE Architecture and Network Information exposure

The fundamental idea of MoWIE is to achieve on demand and periodic network information from network to applications, helping the service provider to do a better policy control to improve user experience.

A possible MoWIE architecture include three core components, the Client Application, the Mobile Network and the Application Server. The raw data are collected firstly from the radio network and core network and then further processing on these collected data and exposed Network information are provided to the application Server. These functions are defined as The network information service (NIS)and the NIS can be deployed at MEC (Mobile Edge Computing). The application server can send the NIS request on UE/Cell level information, and obtain the NIS response on network information from the mobile network. After user data pre-processing, the application server will make best use of the network information to perform analytics and directly influent the application functions e.g. bit rate, data amount etc.

Typically, the network information includes two types of information as below:

Cell level Information:

- The number of Downlink PRBs (Physical Resource Block) occupied during sampling period; and

- the Downlink MAC data rate per cell;

UE level information (without privacy information):

- The Uplink SINR (Signal to Inference plus Noise Ratio);

- MCS: The index of MCS (Modulation and Coding Scheme);

- The number of packets occupied in PDCP buffer; The number of Downlink PDCP SDU packets;

- The number of PDCP SDU packets lost;

- The Downlink MAC data rate per UE.

4.2. RAN assisted TCP optimization based on MoWIE

The RAN information are used to assist TCP sending window adjustment rather than traditional transport layer measurement and acknowledgement. The RAN proactively predicts available radio bandwidth and the buffer status per UE in a time granularity of RTT level (e.g. 100ms) and then piggybacks such information in TCP ACK. We have conducted trial in real mobile network. It is observed that for the UE with good SINR, the throughput is significantly improved by nearly 100%, and the UE with medium SINR can achieve approximately 50% gain.

4.3. NAA QoE Test based on MoWIE

Different from traditional video streaming, cloud gaming has no buffer to accommodate and re-arrange the received data. It must display the stream once the stream is received. Any late stream is of no use for the player. Cloud gaming performs not well in the existing public 4G network according to our actual measurements. The end to end delay is often greater than 100ms for a gaming client in Shenzhen to a gaming server in Shanghai, coupled with the codec delay. Here the delay is defined as the total delay from the user's operation instruction to show the response picture on user's screen.

Once the network fluctuates, users will experience a longer delay. The poor user experience is not only because of the relative low network throughput, but also because the server cannot adapt the application logical policies (e.g. codec scheme and data bitrate).

The popularity of 4K and even higher resolution and increasing FPS for cloud gaming and AR/VR services require both high bandwidth and low latency in wireless and cellular networks. The increasing resolution would incur a higher encoding and decoding delay. However, users' tolerance to delay will not increase with the resolution, which means the application needs to adapt to the network dynamics in a more efficient way. The higher resolution, the larger range of the rate adaptation can be used.

In this section, we make experiments based on the methods described in section 3 to improve the QoE of cloud gaming. The performance between network-aware and native non-network-aware mechanisms are compared.

4.4. ROI Detection with Network Information

The first experiment is based on the ROI detection. We will investigate the impact of network perception.

Saliency detection method has successfully reduced the size of videos and improve the QoE of users in video downloading [Saliency]. However, it is not effective when applied to real-time interactive streaming such as cloud gaming.

As we know, more accurate saliency region detection algorithm needs more time to obtain the result. However, when the users are suffering a bad performance network in cloud gaming, this precise detection may incur more delay to the system. As a result, it will harm the final QoE.

If the application can learn the network well in a real-time manner, it can choose the algorithm based on how much delay the system can tolerate. If the network condition is good enough, it can adopt an algorithm which has deeper learning network and the added delay will not be perceived by the end users. Thus, it can save huge bandwidth without harming the QoE. On the other side, in a network with bad condition, the server can use the fastest method to avoid extra delay.

We make the experiments to show how the network information will influence the total QoE and bandwidth saving in ROI detection.

The following 4 methods are compared:

1) The original video, without using ROI method. This acts as a baseline.

2) Quick saliency detection and encoding method, which is not accuracy in some cases. It only brings 10ms delay [Minbarrier].

3) A relative accuracy saliency detection method. In general, if an algorithm is more precise, it will take more time to get the results. And the complexity of the picture will also influence the detection time and accuracy. Based on our test video, we adopt the method which brings delay about 40~70ms [LSTM].

4) The application server in the cloud has the current bandwidth information which derived from the wireless LAN NIC. Here it is a simulation that all the collected bandwidth traces are already known by the server. Thus, it can use the bandwidth traces to compute transmission delay. Then the server can change the saliency detection

algorithm based on this information and then encode the video. Although the result of future bandwidth prediction is not always accurate in real environment, the assumption here will not influence the final results much. Since in cloud gaming the server encodes the stream based on ROI information frame by frame instead of in a grain of chunks, the future bandwidth prediction window size doesn’t have to be long. Therefore, even the server can only get the bandwidth or delay prediction for a short time window, the server can still use this method with network information.

Test environment:

A 720P game video segment with a rate of 6.8Mbps. This is not a very high bandwidth requirement example in cloud gaming. We just show how it will benefit from MoWIE. High bandwidth requirement case will benefit more if the bandwidth fluctuates much.

The three different networks are all wireless networks and the available bandwidth is varied frequently, where

Network 1: The overall network condition is not very good, the average network bandwidth is 7.1Mbps, but it continues to fluctuate, and the minimum is only 3.9Mbps.

Network 2: The overall network condition is good, with an average network bandwidth of 12Mbps and a minimum of 6.4Mbps.

Network 3: The network fluctuates dramatically, with an average network bandwidth of 8.4Mbps and a minimum network bandwidth of 3.7Mbps

Test content:

The four methods are conducted on the original video under each three networks. After re-encoding based on the saliency detection, we calculate the new QoE and the saved bandwidth. The results are shown in the Figure 4-1:

The QoE value is the MOS as standardized in the ITU.

	Network 1		Network 2		Network 3	
	QoE	BW Saving	QoE	BW Saving	QoE	BW Saving
1	3.8	0	4.8	0	4.3	0
2	3.8	5%	4.8	9%	4.3	7%
3	2.2	2.1%	4.6	38%	3.1	34%
4	3.6	9%	4.7	33%	4.3	25%

Figure 4-1: QoE and Bandwidth Saving

Conclusion:

It can be seen that the methods such as method 2 and method 3 that do not rely on the network information directly, have certain limitations.

Though the method 2 is simple and time-consuming, it can only detect a small part of region of interest accurately. Thus, even if the network condition is very good, it can only save a small amount of bandwidth, and sometimes there are some incorrect ROI detection. The QoE will be reduced without hitting the ROI region.

For Method 3, the algorithm is complicated, and it can correctly detect the user's area of interest, so that it can re-allocate encoding scheme and save a lot of bandwidth. However, its algorithm will introduce higher delay. When the user network condition is poor, the extra delay will cause even worst user's QoE. Although the bandwidth is saved, it affects the user experience seriously.

Method 4 is based on the application's awareness of the network. If the application can know certain network information, it can balance the complexity of the algorithm (introducing delay) and the accuracy of the algorithm (saving bandwidth) according to the actual network conditions. As can be seen from the experiment, method 4 can ensure the user's QoE and save the bandwidth greatly at the same time.

4.5. Adaptive Bitrate with Network Capability Exposure

This experiment is AI-based rate adaption by utilizing the network information provided by the cellular base station (eNB) in cellular network.

Tencent has launched real network testing of NAA-enabled cloud gaming in China Mobile LTE network, with the enhancement in eNB supporting base station information exposure.

To enable the NAA mechanism, some cellular network information from eNBs are collected in an adaptive interval based on the change rate of network status. There information is categorized in two levels, i.e., cell level and UE level. Cell level information are common for all the UEs under a serving LTE cell and UE level information is specific for different UEs. 3GPP LTE specifications have specified how the PDCP (Packet Data Convergence Protocol), RLC (Radio Link Control), MAC (Medium Access Control) and PHY (Physical) protocols operate and this information are very essential statistics from these protocol layers.

It is noted that in NAA mechanism, as the network information is from eNB, and the eNB has the real-time information of radio link quality statistics and layer 1 and layer 2 operation information, NAA mechanism can expose rich information to upper layer, e.g., it is capable to differentiate packet loss and congestion, which is very helpful to the applications in practice.

In order to compare the cases with and without NAA, the cloud gaming test environment is setup with 1080p resolution and around 20Mbps bitrate.

Test scenarios 1~5 are as follows.

Test scenarios 1: Weak network. This scenario is the case where radio link quality is low, e.g., in cell edge area and the bandwidth is not able to serve cloud gaming.

Test scenario 2: User competition scenario. This scenario is defined as the case when user amount is large thus the cellular network bandwidth cannot serve all the cloud gaming users.

Test scenario 3-5: Other scenarios with random user movement trace and user distribution.

Test method: To simplify to comparison, we just use the MCS (MCS index) information derived from the eNB [TS38.214]. The information is provided directly to the application, and the application then adjusts the bit rate according to this information. Here, MCS index shows the modulation (e.g. QPSK, 16QAM,...) and the coding rate used during physical layer transmission, which is relevant to the real data rate per UE. The benchmark method is adopting a constant bit rate without any information to help it predicting the network condition. We compare these scenarios and observe the reduction of delay when those eNB data are utilized.

For different scenarios, the lagging rate is defined as the performance indicator. In our experiments, we assume lagging happens when transmission delay is greater than 200ms and lagging rate is defined as the ratio between the number of frames greater than 200ms and the total number of frames.

Test Scenario	Reduction of Lagging Rate
1	46%
2	21%
3	37%
4	56%
5	32%

Figure 4-2: Reduction of Lagging Rate

It can be clearly seen that with the MCS information, the application can adjust the bit rate to decrease the lagging rate and then significantly improve the user QoE. In weak network scenario, 46% lagging can be avoided by NAA.

4.6. Analysis of the Experiments

The above-mentioned technologies demonstrate the performance gain of NAA with MoWIE.

Although application information can also help to predict the network and have already been used in adaptive bit rate methods, the application information is not as sensitive as eNB information at the very beginning in a lot of cases. For example, when more users enter the cell, the PRB information will first reflect that each user may get less bandwidth. However, the application information needs to react after there is a trend that the bitrate is decreasing. That is to say, the lower layer network information is more directly.

Without MoWIE, the application cannot get the lower layer network information directly and then try to detect "blindly" to adapt to the dynamics of the lower layer network, which cannot meet the requirements of cloud interactive applications like cloud gaming, low delay live show and Cloud VR.

It is noted that the more real-time network resource status the application can learn, the better it can predict how much network resource it can use within a prediction time window. However, there is tradeoff between network information collection frequency and its load and feasibility to the network devices. In principle, the total network resource consumed for such network status reporting is also designed in light-weight manner, e.g., by properly controlling the interval of report and also the number of bits needed to convey the reported information elements. In our experiments, the network status information can be obtained in an adaptive interval based on the change rate of network status, in order to provide good prediction with less load introduced in the network. In fact, not all scenarios need a very frequent information collection. If some information only changes in a very small range and won't influence the final decision, it is unnecessary to report such information all the time. However, if its value varies over the preset threshold, it will inform the application immediately.

The distribution and impact of the exposed data to the performance gain for different algorithm needs to be further studied. This draft is to give a guidance to figure out what kind of data needs to be exposed during initial deployment of these mechanisms.

In our current cloud gaming, the application information can help to reduce about 50% the lagging rate. The left 50% improvement room can be achieved by network information exposure with MoWIE. Actually, the effect of the two-layer information can be accumulated. However, due to current deployment limitation, we cannot collect the application information with the eNB information at the same time. Thus, in this version of the draft we compare the performance with and without MoWIE. We don't compare between application information assisted mode and network information assisted mode in this draft. This is our on-going work. Since both application and eNB information can reflect the network variation, we will compare the performance among application information assisted mode, network information assisted mode and the mode of utilizing both layer information.

5. It should be noticed that the previous mechanisms may also work on IEEE 802.11 standards (e.g. EHT), helping SP having a better understanding for the network environment between AP and STAs. Based on the fact that 802.11 devices are working on unlicensed spectrums, and easily influenced by adjacent unlicensed devices, duty cycle and related CQI information (e.g. MCS, bandwidth, and etc.) are

considered very important network information here.Standardization Considerations of MoWIE as an Extension to ALTO

MoWIE can be a realistic, important extension to ALTO to serve the aforementioned use cases, in the setting of the newer generation (5G) of cellular network, which is a completely open IP based network where routers/UPF with IP connectivity will be deployed much closer to the users. One may consider not only the aforementioned cloud-based multimedia applications, but also other latency sensitive applications such as connected vehicles and automotive driving.

Extending ALTO with MoWIE, therefore, may allow ALTO to expose lower layer network information to ensure higher application QoE for a wide spectrum of applications.

One possible approach to standardizing the distribution of the network information used in the evaluations is to send such information as piggyback information in the datapath. One issue with datapath method is that MoWIE intends to convey more complex and rich information than current methods. To piggyback such complex and rich information in the datapath will take away a lot of datapath resource. But the datapath-based method can provide frequent changed network information and it is much easy to synchronize the network information and user data in the same time scale; Normally, there is less user data in the the uplink direction and the free "space" within the MTU can be used to piggyback the network informaiton to the application, in such case no additional create a second communicaiton channel between the application and network. However, the datapath design may bring out more limited privacy management, which is very important in MoWIE. The application cannot trust the network information if there is no message authentication mechanism for the piggyback network information. How the network inserts the network information in the data packet is also challengeable since a lot of transport layer protocol are encrypted and integration protected. Another method is to create an associated path aligned with datapath. Like the ICMP for IP and RTCP for RTP, this second path can be used to provide additional information associated with the datapath. But creating such second path is a big change to current widely used transport protocols and a lot of applications also need to change, this second path is also challengeable.

In 3GPP, network information exposure based on control plane mechanism is introduced in 4G and 5G systems. We mainly discuss ALTO extension-based design in tackling with this problem. Specifically, the MoWIE extension will reuse existing ALTO mechanisms including information resource directory, extensible performance metrics and calendaring, and unified properties. It also requires modular, reusable extensions, which we plan to specify in detail in a separate document. Below is an overview of key considerations; security considerations are in the following section.

- Network information selection and binding consideration: Instead of hardcoding only specific network information, a modular design of MoWIE is an ability for an ALTO client to select only the relevant information (e.g., cell DLOccupyPRBNum metric and UE MCS) and then request correspondingly. Existing ALTO information resource directory is a starting point, but the design needs to be generic, to provide abstraction for ease of use and extensibility. The security mechanisms of the existing ALTO protocol should also be extended to enforce proper authorization.

- Compact network information encoding consideration: One benefit of ALTO is its high-level JSON based encoding. When the update frequency increases, the existing base protocol and existing extensions (in particular the SSE extension), however, may have high bandwidth and processing overhead. Hence, encoding and processing overhead of MoWIE should be considered.

- Stability and reliability consideration: A key benefit of the MoWIE extension is the ability to allow more flexible, better coordinated control. Any control mechanism, however, should integrate fundamental overhead, stability and reliability mechanisms. .

6. Security Considerations

The collection, distribution of MoWIE information should consider the security requirements on information privacy and information integration protection and authentication in both sides. Since the network status is not directly related to any special user, there is currently no any privacy issue. But the information transmitted to the application can pass through a lot of middle box and can be changed by the man in the middle. To protect the network information, an end to end encryption and integration is needed. Also, the network needs to authenticate the information exposure provided to right applications. These security requirements can be implemented by the TLS and other security mechanisms.

7. References

7.1. Normative References

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

[RFC2474] K. Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers.", RFC 2474, <https://www.rfc-editor.org/info/rfc2474>.

7.2. Informative References

[Fahad] Fahad Fazal Elahi Guraya ; Faouzi Alaya Cheikh ; Victor Medina; A Novel Visual Saliency Model for Surveillance Video Compression, 2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems

[Hongzi] Hongzi Mao; Ravi Netravali; Mohammad Alizadeh; Neural Adaptive Video Streaming with Pensieve; SIGCOMM '17: Proceedings of the Conference of the ACM Special Interest Group on Data Communication; August 2017 Pages 197-210

[Saccadic] E. Matin, Saccadic suppression: a review and an analysis, Psychological bulletin 81 (12) (1974) 899-917.

[Borji] A. Borji, L. Itti, State-of-the-art Analysis and Machine Intelligence, IEEE Transactions on 35 (1) (2013) 185-207.

[MPC] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli. 2015. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. In SIGCOMM. ACM.

[CS2P] Y. Sun et al. 2016. CS2P: Improving Video Bitrate Selection and Adaptation with Data-Driven Throughput Prediction. In SIGCOMM. ACM.

[Hongzi2] Hongzi Mao, Shannon Chen, Drew Dimmery, Shaun Singh, Drew Blaisdell, Yuandong Tian, Mohammad Alizadeh, Eytan Bakshy; Real-world Video Adaptation with Reinforcement Learning ; ICML 2 2019 Workshop RL4RealLife

[Saliency] Chenlei Guo, Liming Zhang; A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 19, NO. 1, JANUARY 2010

[Minbarrier] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, Radomir Mech; Minimum barrier salient object detection at 80 fps. The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1404-1412.

[LSTM]   Lai Jiang; Mai Xu; Zulin Wang; Predicting video Saliency with Object-to-Motion CNN and Two-layer Convolutional LSTM, arXiv:1709.06316v3 [cs.CV] 14 Jan 2019.

[TS23.501]   3GPP TS 23.501 System architecture for the 5G System (5GS), http://www.3gpp.org/ftp//Specs/archive/23_series/23.501/23501-g40.zip

[TS38.214] 3GPP TS 38.214, NR Physical layer procedures for data, http://www.3gpp.org/ftp//Specs/archive/38_series/38.214/38214-g00.zip

[TS26.114] 3GPP TS 26.114,   IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction, http://www.3gpp.org/ftp//Specs/archive/26_series/26.114/26114-g40.zip

[MPEG DASH] ISO/IEC 23009, Dynamic Adaptive Streaming over HTTP; https://mpeg.chiariglione.org/standards/mpeg-dash

[iiMedia] 2019-2020 China Online Live Streaming Market Research Report, https://www.iimedia.cn/c400/69017.html

[GSMA]   Cloud AR/VR Whitepaper, Last updated on April 26, 2019, https://www.gsma.com/futurenetworks/wiki/cloud-ar-vr-whitepaper/#

[5GAA]   https://5gaa.org/news/5gaa-releases-white-paper-on-making-5g-proactive-and-predictive-for-the-automotive-industry/

[TS23.287] 3GPP TS 23.287, Architecture enhancements for 5G System (5GS) to support Vehicle-to-Everything (V2X) services, http://www.3gpp.org/ftp//Specs/archive/23_series/23.287/23287-g20.zip

[TS26.247] 3GPP TS 26.247, Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)

Authors' Addresses

Wei Huang
Tencent Building,
No. 10000 Shennan Avenue, Nanshan District
Shenzhen, Guangdong, 518000

China

Email: wienhuang@tencent.com

Yunfei Zhang
Flat 9, No. 10 West Building.
Xi Bei Wang East Road
Beijing, 100090
China

Email: yanniszhang@tencent.com

Y. Richard Yang
Watson 208A,
51 Prospect Street
New Haven, CT 06511
USA

Email: yang.r.yang@yale.edu

Chunshan Xiong
Flat 9, No. 10 West Building.
Xi Bei Wang East Road
Beijing, 100090
China

Email: chunshxiong@tencent.com

Yixue Lei
Flat 9, No. 10 West Building.
Xi Bei Wang East Road
Beijing, 100090
China

Email: yixuelei@tencent.com

Yunbo Han
Tencent Building,
No. 10000 Shennan Avenue, Nanshan District
Shenzhen, Guangdong, 518000
China

Email: yunbohan@tencent.com

Gang Li
China Mobile Research Institute

No.32, Xuanwumenxi Ave, Xicheng District
Beijing 100053,
China

Email:ligangyf@chinamobile.com