ALTO
Internet-Draft
Intended status: Standards Track
Expires: 26 April 2023

Y. Jia
Y. Zhang
Tencent
Y. R. Yang
Yale University
G. Li
CMRI
Y. Lei
Y. Han
Tencent
Sabine. Randriamasy
Nokia

23 October 2022

MoWIE for Network Aware Application
draft-huang-alto-mowie-for-network-aware-app-05

Abstract

With the deployment of 5G networks, cloud-based interactive services such as cloud gaming have gained substantial attention. 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 quality perceived by a user with mobile and wireless device may vary, as a function of many factors, and unhandled changes can substantially compromise the user's QoE. In this document, we investigate network-aware applications (NAA), which realize cloud-based interactive services with improved QoE, by efficient utilization of a solution named 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 an information, the applications can then adapt key control knobs such as media codec schemes, encapsulation, and application layer processing to minimize QoE distortion. Based on the evaluations, we discuss how the MoWIE features can define extensions of the ALTO protocol, to expose more lower-layer and finer grain network dynamics.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft will expire on 26 April 2023.

Table of Contents

1. Introduction of Network-aware Applications

With the deployment of 5G networks [draft-ietf-dmm-5g-uplane-analysis], more applications are available as remote cloud-based applications (e.g., cloud AR/VR/MR). Many conventional interactive, daily business applications are also becoming widely used as cloud applications, with the help of mobile networks and cloud, e.g., cloud video conference. Especially, during the coronavirus pandemic in 2020, many people had to stay at home and work/study remotely, and the usage of cloud applications, including cloud-based online courses, cloud-based conferencing, and cloud gaming, have surged significantly.

To optimize QoE for end users using mobile networks (a.k.a., cellular networks), many cloud applications utilize information about the mobile network status, e.g., delay, bandwidth, and jitter, to dynamically balance the generated media traffic and the rendering/ mixing in the cloud. Currently, such an application assumes the network as a link and continuously uses client or server measurements to detect network characteristics, and then adaptively changes its network-related configuration parameters which may impact QoS handling as well as logical functions to support the upper layer application. However, when only application information is utilized, the QoE that an application can achive can be limited in some cases. First, information from the application side (i.e., the 3rd party application server) may have relatively long delay. When a user enters a location with a bad network connectivity, such as an elevator or an underground garage, the application will not receive such an 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 cannot know how many resources it can get and when it will change. If other users enter the cell and compete on the usage of 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 the network layer like physical resource block (PRB) information and utilization rate can help applications to describe how many resources the user will get and how many users are competing with it. Such an information is helpful to predict the network and streaming videos. However, the application cannot get those kinds of information yet. Because 5G network deployed by the mobile network

operation which is normally in a different domain compared with internet-based applications with 3rd party, the application server may not be able to directly get the network internal status without network exposure. Therefore, there have been continuous efforts in 3GPP enable network exposure to serve the 3rd party applications in a better way in terms of network resource utilization and user's experiences.

The 3GPP mobile networks are always adhering to standard solutions to get network dynamic indicators that can be used by applications. In 3GPP, many IP-based QoS mechanisms are used. For example, ECN [RFC3168] has been supported by the 4G radio station (eNB) to provide CE(Congestion Encountered) information to the IMS application to perform Adaptive Bitrate (ABR) [TS26.114]. The application can downgrade the bit rate after receiving the CE indication, but does not know the exact bit rate to be selected. DSCP [RFC2474] is used to identify the QoS class and paging strategy [TS23.501], but typically the application cannot dynamically change the DSCP to improve the bit rate based on the network status by mapping DSCP with 4G QCI or 5G 5QI can be a good approach to realize end-to-end QoS. DASH [MPEGDASH] is an MPEG standard widely used for the applications 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 re-buffering. SAND-DASH [TS26.247] defines the mechanism that the network/server can provide available throughput to the applications; in such case, the better bitrate can be selected by DASH application.

There are some early works on network status exposure to TCP server, which may also indirectly impact application layer optimization, but did not consider 5G networks [Sprecher_mobile], [flinck_mobile] In 5G cellular networks, network capability exposure has been specified to allow the 5G system (5GS) to expose user device location and network status towards the 3rd party application servers modeled as AF (Application Function) [TS23.501]. In such a case, the AF can request the 5GS to establish a dedicated QoS Flow to transport an IP flow with the AF-provided QoS requirements. Via certain measurements, network internal status including congestion can be exposed and optimization can be carried out for the cellular network [PBECC].

5G also can provide QNC (QoS Notification Control) to an 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 5G enhances the QNC by providing a list of AQPs (alternative QoS profiles). With AQP, a 5G network provides a subset of supported AQPs with the QNC, and then the AF selects a bit rate

from 5G network supported AQPs. In such a case, the GBR can be fulfilled again if the radio state of the UE is changed. QoS prediction 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 to an entity such as application server.

However, both network capability exposure and QoS prediction solutions are designed for 5G access and core networks, which cannot cover the whole end-to-end network. How to enable the application which locates the internal DN (data network) to be aware of the lower layer networks within 5G network are an important area for both industrial and academic researchers, that is addressed by MoWIE.

MoWIE is a solution that aims to realize real-time provisioning of cellular radio network information by networks to applications, thus helping service providers to achieve a better policy control and to improve user experience. The benefits of the MoWIE concept/solution have been experimented on several use cases, detailed in Section 4.1.

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

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

2.1. Cloud Gaming

As mentioned above, cloud gaming is widely used, and this kind of games requires low latency and highly reliable transmission of motion tracking data from the user to the 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 a high-volume software in the user device), less cost and process load in the user device and it is also regarded as an anti-cheating measure. Thus, cloud gaming has become a competitive replacement for console gaming using cheaper PC or laptop. To support high quality cloud gaming services, the application needs 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 close interaction between the application and the network.

Compared with conventional broadcast services, this service is interactive, which means that 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 audiences, and it also requires playing game interaction between the anchor and audiences. A delay lower than 100 ms 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 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 20 ms, 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 (FPS, frame per second) 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.

With the typical parameters setting, cloud gaming generally needs a bandwidth of 20~60 Mbps; we also consider that significant lagging happens when the latency is larger than 200 ms, depending on the types of games (e.g. 40 ms for First Person Shoot games, 80 ms for Action games, and 200 ms for Puzzle games). In order to avoid bad

user experiences, lagging is better when it is lower, and can be as low as zero (in an optimal QoE). For low latency live shows, 20~50 Mbps bandwidth may be needed and the end-to-end latency requirement 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 on the parameter settings and they are provided to illustrate the order of magnitude of these parameters for the aforementioned 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, and packet loss rate.

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, with 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 25 Mbps, which brings huge burden to the network, even for 5G network. Those ROI-based video compression methods are mainly applied to high concurrency networks 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 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 information, including rate, download time, buffer size or network level information which can reflect the performance, are useful to reinforcement learning. Since those data can reflect the network dynamics, they have been used to help the applications to know how to change the rate and improve 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. The lack of standardized approach to acquire these data makes it difficult to make this usable for different applications for large scale deployment. Meanwhile, since these data reflect the real-time network status, they may change rapidly and randomly, and hence can be 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 are not rich, direct, real-time; they also lack predictability, especially when in the mobile and wireless network scenarios, which results in long reaction 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 network service providers to realize a better policy control and to improve users' experience.

A possible MoWIE architecture includes 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; further processing on these collected data and the exposure of Network information are provided to the application Server.

An application server can send network information request about 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 enhance the application functions e.g. bit rate, latency, and jitter.

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

Cell level Information:

* The number of Downlink PRBs (Physical Resource Blocks) occupied during sampling period;

* the cell load;

* the downklink (DL) MAC data rate per cell;

* the channel status (e.g. RSRP (Reference Signal Received Power) and CQI (Channel Quality Indicator));

* the DL data rate;

* the PDCP (Packet Data Convergence Protocol) buffer status;

UE level information (without privacy information):

* The Downklink Signal to Inference plus Noise Ratio (SINR);

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

* The number of packets occupied in PDCP buffer;

* The number of downlink PDCP Service Data Unit (SDU) packets;

* The number of lost PDCP SDU packets;

* The per UE downlink MAC data rate;

* the per-UE channel status (e.g. RSRP (Reference Signal Received Power) and CQI (Channel Quality Indicator));

* the per-UE DL data rate;

* the per-UE PDCP (Packet Data Convergence Protocol) buffer status;

The network information listed here can also be found in 3GPP (PRB [TS38.211], cell load [TS38.300] PDCP for 5G [TS38.323] RSRP, RSRQ, RSSI [TS38.331], MCS, CQI [TS38.214], The number of packets occupied in PDCP buffer, the number of Downlink PDCP SDU packets, the number of PDCP SDU packets lost, the per-UE PDCP buffer status [TS38.323]), to demonstrate the potential benefits of MoWIE for network-application integration over cellular network. Figure 3-1 and Figure 3-2 list the data types correponding to the cell-level information and UE-level information, respectively.

Cell-level Information	Data type/Range
PRB	Uint16
CQI	Uint8
RSRP	Uint8
RSRQ	Uint8
Cell load	[0,1]

Figure 4-1: Cell level data type

UE-level Information	Data type/Range
Downlink SINR	Uint16
MCS	Uint8
Downlink PDCP SDU packets	Uint8
PDCP SDU packets lost	Uint8
Packets occupied in PDCP buffer	[0,1]
CQI	Uint8
RSRP	Uint8
RSRQ	Uint8

Figure 4-2: UE level data type

4.2. RAN-assisted TCP optimization based on MoWIE

The RAN information is 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. 100 ms) and then piggybacks such information in TCP ACK.

We have conducted trial in real mobile networks. 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 100 ms 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 that 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.

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.

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 4-3: 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 (gNB) in cellular network.

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

To enable the NAA mechanism, some cellular network information from gNBs 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, 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 gNB, and the gNB 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 [MengZ], 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 gNB [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 gNB 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 4-4: 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 gNB 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 gNB 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 gNB 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. Standardization Considerations of MoWIE as an Extension to ALTO

In 3GPP, network information exposure based on control plane mechanism is introduced in 4G and 5G systems. In 3GPP Release 17, there is a work item named 5G_AIS (Advanced Interactive Services) which focuses on QoS enhancements for interactive services including cloud gaming, XR, remote driving and real-time digital twin. There is also continuous work in Release 18 XRM (XR and media services), which focuses on strengthening the interaction and collaboration between applications and networks by improving the ability of the

network information exposure especially for XR and multimedia services. MoWIE is closedly related to this study as it can be a tool to support exposure of 5G network status information. MoWIE can also support QoS prediction technologies which is being applied to 5G-enabled automated driving which is being developed in 5GAA [PQoS_white_paper], this solution can use network information exposure to predict the future network changes to the application layer in order to be prepared for such changes and make adaptations in application layer to improve the QoE of users. There have been close collaboration between 5GAA and 3GPP to provide E2E solutions on this area.

Among these above mentioned work items, one important way to support QoS (MoWIE and PQoS) enhancements is to expose the network status information to the application layer and the application layer can take measures to adapt according to the network status information. The network information can include the radio network statistics as has been elaborated in Section 4.1 and also the parameters specified in [ALTO_METRICS]. In Section 4.1, the parameters which MoWIE proposes to expose can provide real-time status and rich information about the wireless link which can be utilized by AI-ML (Machine Learning) algorithms to predict the available network resources in the subsequent transmission opportunities, which can help the application layer to adjust its traffic pattern or codec profile to optimize the user's experience. By mapping these parameters with the ALTO metrics which has been proposed or defining potential new ALTO metrics, it is possible to extent current ALTO protocols to provide better support for real-time immersive services. In ETSI MEC, RNIS [ETSI_MEC] has proposed to expose physical layer, Layer 2 and higher layer parameters including 4G and 5G. There are some common parameters like RSRP, RSRQ and RSSI which MoWIE also proposes; however, RNIS is based on the MEC architecture, and MoWIE is not restricted to MEC case.

It should be noticed that the previous mechanisms may also work on IEEE 802.11 standards (e.g., EHT), helping SP to have 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 cycles and related CQI information (e.g., MCS, and bandwidth) 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 data path. One issue with data path method is that MoWIE intends to convey more complex and rich information than current methods. To piggyback such complex and rich information in the data path will consume significant data path resource. But the data path-based method can provide frequently changed network information and it is technically simpler to synchronize the network information and user data at the same time scale. Normally, there is less user data in the uplink direction and the free "space" within the MTU can be used to piggyback the network information to the application, without the additional overhead of creating a second communication channel between the application and network. However, the data path 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 data path. Like the ICMP for IP and RTCP for RTP, this second path can be used to provide additional information associated with the data path. 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 this draft, 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. 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 considerations: 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 considerations: 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.

* Cost metrics considerations. In [ALTO_METRICS], some cost metrics are being standardized including throughput/bit rate, latency, priority, error rate, jitter. These parameters can be linked with cost metrics in 5G network entities like network exposure function (NEF) [TS23.501] or application function (AF) [TS23.501]. NEF or AF, which act as ALTO Clients, utilizing the network information exposure capability provided by 3GPP standards, can request to expose some of the proposed parameters with consideration of the ALTO performance metrics.

With the better utilization of the 5G network, IETF ALTO can well benefit from MoWIE architecture and QoE for the use cases of NAA applications can be achieved with the help of the convergence of 5G network architecture and IETF ALTO. For example, it was shown in Figure 5-1 that the ALTO client can act as AF or act as part of the functional module of AF to converge with the 5G system through the implementation in the AF. It can receive the cellular-based network information through interaction with NEF, identify the ALTO server

using ALTO service discovery, and request available ALTO information prepared by the ALTO server using ALTO Protocol [RFC7285]. The ALTO server can be implemented in other places through ALTO architecture and can interact with the third parties (e.g. Content providers) via external interfaces. Furthermore, the ALTO server can aggregate information from multiple functional entities to provide an abstract and unified view that can be more useful to applications. Examples of other systems include (but are not limited to) static network configuration databases, dynamic network information, routing protocols, provisioning policies, and interfaces to outside parties [RFC7285]. Please note that these components shown in Figure 5-1 are out of IETF ALTO scope and are just for completeness to put here. As a result, the ALTO Client can perform ALTO Service Discovery and interact with the ALTO server via ALTO Protocol to get the ALTO information and ALTO metrics regarding the proposed parameters with the help of MoWIE that can be updated dynamically based on network conditions from cellular aspects in Section 4.1. It is noted that in addition to AF-based convergence architecture which relies on control-plane interfaces e.g. N5/N33, MoWIE can also utilized user plane exposure via N6 or UPF which makes network exposure mechanisms more comprehensive and flexisble.

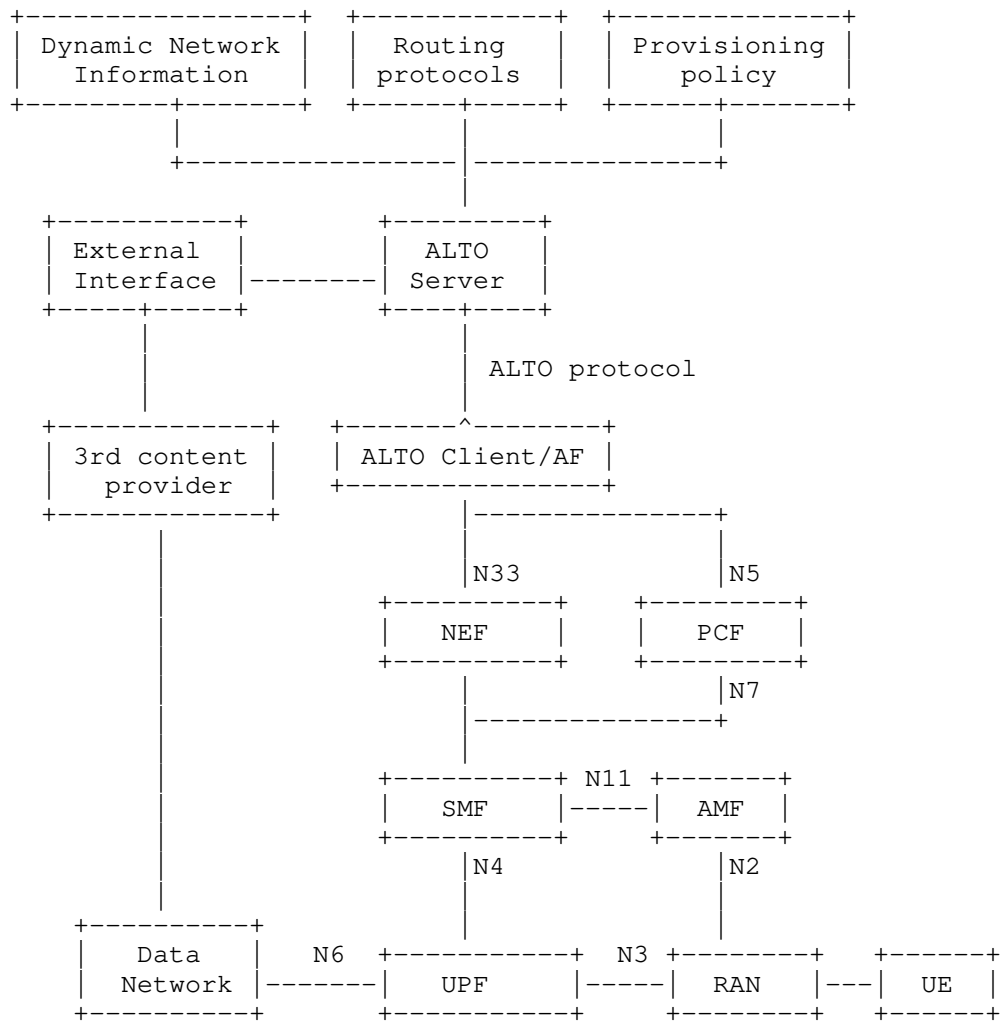

Figure 5-1: Convergence of 5G network architecture and IETF ALTO

By extending the exposure scope of network information beyond the cellular access, ALTO can help improve the QoE of several applications running on endpoints located in cellular networks. [ALTO_USE_CASES] is work in progress that investigates use cases where the performances of these applications can be further improved with abstracted network information and suitable transportation means provided by ALTO. Additionally, upon reviewing the existing ALTO capabilities, it lists the ALTO features that need to be extended or defined to support the presented use cases.

6. IANA Considerations

This document has no actions for IANA.

7. 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.

8. Acknowledgments

The authors would like to thank Huang Wei and Mohamed Boucadair for their contribution and technical review comments to the previous versions of this draft.

9. References

9.1. Normative References

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <https://www.rfc-editor.org/info/rfc2474>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC7285] Alimi, R., Ed., Penno, R., Ed., Yang, Y., Ed., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, DOI 10.17487/RFC7285, September 2014, <https://www.rfc-editor.org/info/rfc7285>.

9.2. Informative References

[ALTO_METRICS]
"Internet-Draft, draft-ietf-alto-performance-metrics-09", " ALTO Performance Cost Metrics", 2020, <https://tools.ietf.org/html/draft-ietf-alto-performance-metrics-09>.

[ALTO_USE_CASES]
"Internet-Draft, draft-li-alto-cellular-use-cases-00", " Requirements and reference architecture for Mobile Throughput Guidance Exposure", 2021, <https://datatracker.ietf.org/doc/html/draft-li-alto-cellular-use-cases>.

[CS2P] Sun, Yi., Yin, Xiaoqi., Jiang, Junchen., Sekar, Vyas., Lin, Fuyuan., Wang, Nanshu., Liu, Tao., and Bruno. Sinopoli, "CS2P: Improving Video Bitrate Selection and Adaptation with Data-Driven Throughput Prediction", DOI 10.1145/2934872.2934898, 2016, <https://doi.org/10.1145/2934872.2934898>.

[draft-ietf-dmm-5g-uplane-analysis]
"Internet-Draft, draft-ietf-dmm-5g-uplane-analysis-04", " ALTO Uses Cases for Cellular Networks", 2020, <https://datatracker.ietf.org/doc/html/draft-ietf-dmm-5g-uplane-analysis-04#section-4.1>.

[ETSI_MEC] "ETSI GS MEC 012", "Multi-access Edge Computing (MEC); Radio Network Information API", 2019, <https://www.etsi.org/deliver/etsi_gs/MEC/>.

[Fahad] Fazal Elahi Guraya, Fahad., Alaya Cheikh, Faouzi., and Victor. Medina, "A Novel Visual Saliency Model for Surveillance Video Compression", DOI 10.1109/SITIS.2011.84, 2011, <https://doi.org/10.1109/SITIS.2011.84>.

[flinck_mobile]
"Internet-Draft, draft-flinck-mobile-throughput-guidance-04", " User Plane Protocol and Architectural Analysis on 3GPP 5G System", 2017, <https://datatracker.ietf.org/doc/html/draft-flinck-mobile-throughput-guidance-04>.

[Hongzi] Mao, Hongzi., Netravali, Ravi., and Mohammad. Alizadeh, "Neural Adaptive Video Streaming with Pensieve", DOI 10.1145/3098822.3098843, 2017, <https://doi.org/10.1145/3098822.3098843>.

[MengZ] Meng, Z., Guo, Y., Sun, C., Wang, B., Sherry, J., Liu, H. H., Xu, M. (2022), "Achieving Consistent Low Latency for Wireless Real-Time Communications with the Shortest Control Loop", DOI 10.1145/3544216.3544225, 2022, <https://zilimeng.com/papers/zhuge-sigcomm22.pdf>.

[MPC] Yin, Xiaoqi., Jindal, Abhishek., Sekar, Vyas., and Bruno. Sigopoli, "A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP", DOI 10.1145/2785956.2787486, 2015, <https://doi.org/10.1145/2785956.2787486>.

[MPEGDASH] ISO/IEC, "ISO/IEC 23009, Dynamic Adaptive Streaming over HTTP", 2020, <https://mpeg.chiariglione.org/standards/mpeg-dash>.

[PBECC] Xie, Yaxiong, Fan Yi, and Kyle Jamieson, "PBE-CC: Congestion control via endpoint-centric, physical-layer bandwidth measurements", DOI 10.1145/3387514.3405880, 2020, <https://dl.acm.org/doi/pdf/10.1145/3387514.3405880>.

[PQoS_white_paper] "Making 5G Proactive and Predictive for the Automotive Industry", " 5GAA Automotive Association, Tech. Rep.", 2020, <https://5gaa.org/wp-content/uploads/2020/01/5GAA_White-Paper_Proactive-and-Predictive_v04_8-Jan.-2020-003.pdf>.

[Saccadic] Matin, E., "Saccadic suppression: A review and an analysis", DOI 10.1037/h0037368, 1974, <https://doi.org/10.1037/h0037368>.

[Saliency] Guo, C. and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression", DOI 10.1109/TIP.2009.2030969", 2017, <https://doi.org/10.1109/TIP.2009.2030969>.

[Sprecher_mobile] "Internet-Draft, draft-sprecher-mobile-tg-exposure-req-arch-03", " Mobile Throughput Guidance Inband Signaling Protocol", 2017, <https://datatracker.ietf.org/doc/html/draft-sprecher-mobile-tg-exposure-req-arch-03>.

[TS23.501] "3rd Generation Partnership Project (3GPP)", "System architecture for the 5G System (5GS)", 2021, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>.

[TS26.114] "3rd Generation Partnership Project (3GPP)", "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction", 2021, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1404>.

[TS26.247] "3rd Generation Partnership Project (3GPP)", "Progressive Download and Dynamic Adaptive Streaming over HTTP(3GP-DASH)", 2020, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1444>.

[TS38.211] "3rd Generation Partnership Project (3GPP)", "NR; Physical channels and modulation", 2017, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213>.

[TS38.214] "3rd Generation Partnership Project (3GPP)", "NR; Physical layer procedures for data", 2021, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216>.

[TS38.300] "3rd Generation Partnership Project (3GPP)", "NR; NR and NG-RAN Overall description; Stage-2", 2017, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3191>.

[TS38.323] "3rd Generation Partnership Project (3GPP)", "NR; Packet Data Convergence Protocol (PDCP) specification", 2017, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3196>.

[TS38.331] "3rd Generation Partnership Project (3GPP)", "NR; Protocol specification", 2017, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197>.

Authors’ Addresses

Yuhang Jia
Tencent
Flat 9, No. 10 West Building, Xi Bei Wang East Road
Beijing
100090
China
Email: tonyjia@tencent.com

Yunfei Zhang
Tencent
Flat 9, No. 10 West Building,Xi Bei Wang East Road
Beijing
100090
China
Email: yanniszhang@tencent.com

Y. Richard Yang
Yale University
Watson 208A, 51 Prospect Street
New Haven, CT 06511
United States of America
Email: yang.r.yang@yale.edu

Gang Li
China Mobile Research Institute
No.32, Xuanwumenxi Ave, Xicheng District
Beijing
100053
China
Email: ligangyf@chinamobile.com

Yixue Lei
Tencent
Flat 9, No. 10 West Building,Xi Bei Wang East Road
Beijing
100090
China
Email: yixuelei@tencent.com

Yunbo Han
Tencent
Tencent Building, No. 10000 Shennan Avenue, Nanshan District
Shenzhen
518000
China
Email: yunbohan@tencent.com

S. Randriamasy
Nokia
Nokia Bell Labs
Nozay
United States of America
Email: sabine.randriamasy@nokia-bell-labs.com