Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: 13 January 2022

G. Li
China Mobile Research Institute
S. Randriamasy
Nokia Bell Labs
C. Xiong
Tencent

12 July 2021

ALTO Uses Cases for Cellular Networks
draft-li-alto-cellular-use-cases-00

Abstract

This draft presents a number of use cases of applications running on endpoints located in cellular networks and whose performances highly depend on network information. This document first, shows how the performances of these applications can be further improved with ALTO provided abstracted network information and transportation means thereof. Second, upon reviewing the existing ALTO capabilities, it lists the ALTO features that need to be extended or defined to support the presented use cases.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 13 January 2022.

Table of Contents

1. Introduction 3
1.1. Requirements Language 5
1.2. Terminology 5
2. Motivation And First Considerations 5
2.1. Overview of challenges for applications on cellular networks 5
2.2. Benefits expected from using ALTO to expose network topology to applications 6
2.3. First considerations on ALTO information features for wireless network 7
3. Example applications and use cases 8
3.1. Use case 1: rate adaptation for cloud VR/gaming 8
3.1.1. Application needs in information capabilities from network 9
3.1.2. Missing ALTO information and features 10
3.2. Use case 2: Video-conferencing applications 11
3.2.1. Application needs in information capabilities 12
3.2.2. How the application can get the 5G network information from ALTO 13
3.3. Use case 3: ALTO supporting applications on UEs 14
3.3.1. Use case: Access-aware AEP selection from UE with cascaded ALTO Servers 14
3.3.2. Scenario and assumptions 14
3.3.3. Missing ALTO information and features 15
4. Highlights on 3GPP Information Useful to ALTO 15
5. Gap analysis with Existing ALTO features 16
5.1. ALTO limits w.r.t. Cellular Network Information 16
5.2. ALTO Limits on network information transport: gap analysis with ALTO SSE 16
6. Summarizing ALTO added value and gaps for cellular networks 17
6.1. Summarizing ALTO added value to cellular use cases . . . 17
6.2. Summarizing new ALTO features needed to support cellular use cases 17
6.2.1. ALTO Cellular Network Information 18
6.2.2. Efficient transport for ALTO Cellular Network Information based on SSE 18
6.2.3. Time constraints on ALTO-provided Cellular Network Information 19

6.2.4. ALTO notifications to non-GBR as well as GBR traffic 20
7. Acknowledgements 20
8. IANA Considerations 20
9. Security Considerations 20
10. References 20
10.1. Normative References 20
10.2. Informative References 20
Appendix A. Additional Stuff 21
Authors' Addresses 21

1. Introduction

The ALTO protocol has been defined as modern network traffic started to convey a majority of user-initiated multimedia flows comprising essentially video payload. The purpose of ALTO is to alleviate network costs, congestion and load while maintaining application performances. To this end, ALTO provides to applications that have a choice among several endpoint location. This guidance consists for ALTO in exposing a Network Map that is a subjective abstraction of the Internet provider network. The Network Map is an arbitrary provider-defined partition of the provider topology in zones that have a human-readable identifier named PID and that gather network endpoints that may be treated similarly, see RFC7285 sec 5.1. Among these PIDs, the ALTO Cost Map defines abstracted network costs.

The design of ALTO is flexible and generic enough to support further evolutions of application traffic and enable lightweight protocol extensions. In particular as per section 5.2 of RFC7285, "There are many types of addresses, such as IP addresses, MAC addresses, or overlay IDs.", while the 7285 "document specifies (in Section 10.4) how to specify IPv4/IPv6 addresses or prefixes." Likewise, the time granularity of ALTO path costs was intended in the initial requirements of RFC6708 to be in the order of days or more, extensions such as RFC 8688 specifies the value encoding in floating numbers, allowing thus smaller time interval durations. Nevertheless, ALTO is by no means meant to provide information in real-time.

In the first two WG charters, the ALTO applications and use cases have focused on applications using network maps, cost maps defined on IP Networks. Nevertheless, while the central use case in the base protocol was peer-to-peer file sharing, the ALTO WG progressively moved to CDN use cases, following thus the usage trend and the needs of the service providers exposing their network abstraction. This has produced extensions supporting finer grained information in terms of network capabilities and time span. Applications are today evolving to require more resources, performance, service presence and reactivity.

Modern Applications now span endpoints in both fixed and cellular networks. They are usually catered by Servers located at the edge or within the core network. Servers view and manage IP flows while Clients are associated to one or more flows defined according to their hosting network technology. For example, a Client located in 3GPP defined cell is mapped to a so-called QoS flow, that may map to one or more IP flows, either because the application involves say voice or video, or because the user equipment (UE) is running several applications simultaneously. These applications are also the most Quality of Experience (QoE) demanding ones.

The COVID period has witnessed frequent service degradation and interruption in applications such as videoconferencing or gaming on a mobile phone. The Servers need to be reactive enough to correctly adapt their resources to the challenged users. To do so, they need to have a reliable view of the network capabilities and bottlenecks, while that view should cover application paths end to end and not only at the user access network. There is a strong need for application vendors to sustain user QoE, especially low latency and acceptable throughput. Given the number of user flows, the complexity of the network and sensitivity of user and network information, the application server should get all and only what information it needs. In other words, the network abstraction exposed to applications must be reliable, preserve confidentiality, suit the application needs and minimize the data exchange volume.

This draft emphasizes the importance of abstracted cellular information, as the cell resides in the last IP hop, which is usually the bottleneck. Therefore, last hop network information is critical. It is important to have end to end information for apps having cellular user footprint that covers both the edge and core Internet and the access plus the cloud. The current networking SDOs such as 3GPP, ETSI, IETF do not cover all these areas simultaneously but provide each a separate focused view. Note that it is the same for Open Source organizations such as Open RAN (ORAN) or Facebook sponsored - Open Computing Platform (OCP) providing in-band telemetry info for DC networks.

This draft focuses on popular applications that have a user footprint in cellular networks. It explores on one hand the requirements of such applications in terms of network information, the capabilities of 3GPP network information functionalities such as the Network Exposure Function (NEF), and their the missing. On the other hand, it explores what ALTO may potentially provide to compensate and extend the scope of 3GPP information. The draft is thus motivated by the following "gaps" between ALTO and 3GPP:

* Cellular information provided by 5G is limited in 2 ways: RAN scope only, and QoS negotiation for only given type of traffic (GBR),

* ALTO on the other hand is generic regarding the type of access. However, it lacks small grain information and its dynamicity is currently limited. Above all, the ALTO protocol and its extensions are specified for IP networks.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.2. Terminology

TO BE COMPLETED

QoS, ABR, RTT, GBR, XR/VR, AR, CGS

2. Motivation And First Considerations

2.1. Overview of challenges for applications on cellular networks

This section lists some challenges faced by applications running on cellular networks that motivate the definition of ALTO-based network topology exposure to improve application performance.

* New popular applications running on mobile networks: The current ALTO protocol exposes endpoint addresses or path information. ALTO was initially designed for P2P file or chunked data download applications. Nowadays, P2P-like applications are no more widely used. Instead, mobile internet got more popular, the price to buy legal music/video/ebook files for download or streaming from a dedicated service provider have dramatically reduced. Consequently, there is little need to allow a user to change the destination server address the path to destination server. The major challenge now is to better support applications running on

the mobile internet. In 5G networking, selection of an destination application server and related path is no major issue: the 3GPP has defined a mechanism in [TS23.548] to let the application change the Edge Compute server on the fly during the download, where this change is automatic and transparent to the user. A bigger issue, in 5G networks, is how to adapt the application behavior, based on the data rate fluctuation in the mobile network.

* 5G low-level network information too complex for application developers: Too detailed radio level (low layer) parameters are very hard to be understood by the application developers and are thus very hard to be tested for the application development and deployment.

* Notifications of RAN changes restricted to GBR traffic: In 5G networks, notifications to CGS of QoS level change are currently sent only for Guaranteed Bit Rate (GBR) traffic whereas they should be used for non-GBR traffic as well.

2.2. Benefits expected from using ALTO to expose network topology to applications

The first and fundamental benefit expected from ALTO is the ability to expose simple and abstracted network topology information to applications. An ALTO Client collocated with an application server could easily get concise and safe information allowing adaptive application behavior and QoE maintenance. Besides, with ALTO and its handy JSON contents, application developers can take advantage of the ALTO services to rapidly develop and update wireless applications with less changes in their code.

Another important benefit is the ability of ALTO to support to exposition to applications of the QoS supported by a path, given that the supported QoS may vary over time. A fundamental ALTO service to support the 5G would be the notification of different bit rates supported in the access network. Providing such path QoS notification to is of highest importance, as it allows the applications to appropriately adapt their behavior, that is, in the present case, their bitrate. Given that the path QoS can vary very quickly, the information exchange must be fast, with simple terms, while moderating the volume of exchanged information.

A first step in this direction is to introduce the definition of three levels of path QoS supported for applications, that can be defined in abstracted from 3GPP in ALTO terms. Each level may draw specific requirements on the interaction between ALTO Client and ALTO Server.

* Basic Level: for e.g. video streaming; normal data rate, and loose latency requirements (e.g. network transport latency lower that 150ms)). In such case, the ALTO interaction can use the basic QoS Notification Control (QNC) and alternative QoS profile (different bit rate) defined in 5G.

* Medium Level: for e.g. conference call; high data rate (e.g. higher than 5Mbps), but moderate vlatency requirement (e.g. network transport lower than 150 ms). In this case the ALTO interaction can use the QNC and small measurement time.

* Hard/complex Level: for e.g. cloud gaming, XR/VR services; high data rate (e.g. higher than 5Mbps) and low latency (lower than 10ms for network transport, lower than 100ms end to end in service level) services. More technologies are being studied in 3GPP R18. e.g. new QoS Notification Type.

It is desirable to have the same data model in ALTO to express the different QoS values or levels for different the use cases.

2.3. First considerations on ALTO information features for wireless network

This draft does not aim at defining ALTO extensions to support applications on cellular networks. This section lists some initial considerations that would characterize such extensions.

Firstly, the use cases and associated needs of this draft do not restrict to 5G cellular networks. They cover potential application queries for ALTO network information that happen within the first IP hop. Such needs may be true for 4G and 5G networks.

Currently in ALTO, the only way for an application to get a more adequate QoS level is to use another path by selecting another endpoint. In a wireless network, the application must use the same path in which the QoS can change over time. The end user cannot create or select another path during an application session. But because of the physics of the wireless, the QoS of the wireless connection may change quickly, so the application in the end user and in the application server needs to detect the change of the QoS value and adapt the application bit rate accordingly. Therefore, if the ALTO service is used in a wireless network, we need to extend ALTO to support only one path but with changes in the path cost or supported QoS that occur very quickly, e.g. within seconds or sub-seconds.

A first architectural assumption is needed here for the ALTO Server that would expose "below 1st hop" network information. We assume the availability of a Local ALTO Server, that manages the information

covering the "below" first IP hop area delineated by the UEs and the Packet Data Network (PDN). In the use cases of this draft a Local ALTO Server may cover a 4G or 5G topology. It is queried by ALTO Clients that may be associated to application functions in the network edge or in end systems. It is not scalable to provide ALTO information on paths from all UEs to all application endpoints in the Internet. Therefore, it is desirable to use "cascaded" ALTO Servers, where a local server covers the relevant access area and can, if needed compose its information with a "core" ALTO Server that manages information at an IP hop level or above. Such a solution is proposed in sections 5.1 of [alto-cost-context] and section 2.3.3 of RFC 7971 on deployment cases. Thus:

* A Local ALTO Server (LAOS): covers a local and restricted part of the network. It is typically located before the Internet gateway, in the access network. For example, it can be collocated with the NEF, as in the Cloud Gaming use case or with a gateway. It hosts the information on the local 4G/5G network, covering the paths between e.g. the UEs and the cells or the PGWs/UPFs. It may host an ALTO Client that sends an ALTO request to a "core" ALTO Server, covering the zone beyond the PGW/UPF.

* A "core" ALTO Server covers the whole ISP network view, at the IP and beyond level, as it would if the "local ALTO Service" is not available or deactivated. That is, it does not see the details below the (UE, PGW/UPF) hop.

3. Example applications and use cases

This section presents emblematic examples of application use cases on cellular networks. For each use case, it lists the network information needed to maintain or improve application performances, which one is available from 3GPP and from ALTO, which ones are missing. Current examples applications are: Cloud gaming, Conferencing, and use cases are divided in network-based decision making and UE-based decision making.

3.1. Use case 1: rate adaptation for cloud VR/gaming

FOR FURTHER VERSIONS: Need artwork based on slide 1 of "ALTO Recharter for Wireless Use cases-V6SR-2004"

5G is beginning to be commercialized globally since 2020, and there is a great improvement regarding bandwidth enhancement and latency reduction. Cloud-based interactive streaming applications such as cloud Virtual Reality (VR) and cloud gaming are booming.

This type of applications requires low latency and highly reliable transmission of motion tracking instructions from user to server in the cloud, while the cloud is required to perform the rendering of pictures per frame and deliver them to users with usually low latency and high data rate. The required end-to-end transmission delay may be as low as 20ms. The less the latency, the better the user QoE as, for instance, a slight extra delay may cause user dizziness. The downlink bandwidth normally depends on different parameter settings such as DoF (Degree of Freedom), image resolution, frame rate, adapted rendering and compression algorithm. For example, a high definition with 1080p with 60 frames per seconds may require at least 20Mbps and ultra-high definition with 4K may require more than 40Mbps.

Cloud VR/gaming is regarded as one of killer applications as well as major traffic contributor to cellular 5G networks. The major advantages of cloud VR/gaming are easy and quick start since there is no need to download and install a big volume of software in the user device beforehand, and also it is cost effective and demands too moderate processing load in the user device. Last, it is also regarded as a more trusted solution. Thus, cloud gaming becomes a competitive replacement for console gaming using cheaper PCs or laptops. On one hand, the above cloud-based interactive applications normally require high bandwidth and low latency, on the other hand, a larger radio bandwidth implies larger variations since radio resources are shared and competed by mobile users in a cell. Therefore, the last mile radio link is viewed as the bottleneck of the QoE.

To address this problem, the application usually estimates available bandwidth based on application-level measurements such as traffic throughput, RTT and latency. It then uses an adaptive bitrate (ABR) mechanism to change the video encoding bitrate so as to match the fluctuation of radio bandwidth. However, it is hard to accurately estimate or predict user cellular link status only from the application's perspective since only the cellular network has a full understanding of all users' radio channel status, traffic fluctuation, moving in and out of cells. Moreover, bandwidth estimation based on application level traffic throughput statistics, rather than radio channel capabilities may be inaccurate.

3.1.1. Application needs in information capabilities from network

Based on the above use case analysis, it would be beneficial if the cellular network could inform the cloud streaming application on the cellular network link status. The following approaches may be considered.

* The application uses request/response to query the link status from the cellular network for a specific user. This may cause extra latency since two way signaling is required.

* The cellular network periodically reports network link status to the application. This may cause extra signaling overhead if the reporting period threshold is too frequent.

* The application uses a pub/sub-like mechanism, specifically, the application may subscribe a certain condition (for example, whether the QoS requirement is satisfied or significant QoS change happens) for cellular network. As soon as the predefined condition is fulfilled, the notification will be triggered immediately and if the condition is not triggered, no signaling is involved. Obviously, this approach is more cost effective from signaling point of view and timely compared with request/response.

The application may inform the cellular network of the following information (not exhaustive list, new information may be added later):

* QoS requirements of application, e.g., min and/or max bandwidth, latency

* Significant QoS changes, e.g., 30% drop of bandwidth change

The cellular network link status may include the following information (not exhaustive list, new information may be added later):

* The available bandwidth and latency of the cellular network

* Whether QoS requirements of application are satisfied or not by the radio network

* Whether significant QoS changes happen

3.1.2. Missing ALTO information and features

However, the current ALTO protocol and extensions are missing the following information.

* The ALTO representation of the QoS of an application path requires to represent the path endpoints. That is the path source and destination. It the ALTO Endpoint Cost Service (ECS) is used, it requires to indicate the endpoints. This is possible for the destination, as it is the CGS, usually identified with an IP address. However, the UE address would to be identified with an identifier relating to cellular address. And the current ALTO protocol only supports IP addresses.

* One possible workaround would be to identify a cell as a PID. However, a PID MUST specify the network addresses it contains and only IP addresses are supported. The path cost from a UE/Cell to a CGS should be specific to a cell, but the IP address that is provided to the UE is not. The ALTO Server should ensure that the metric used to indicate the quality of the path to a CGS reflects the specifics of a cellular network.

3.2. Use case 2: Video-conferencing applications

In the current context of massive teleworking, reliable video-conferencing tools are of utmost importance. Poor experience or service interruption occur more than often and may be caused by factors impacting functions at both the Server end and the UE end.

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 is able to provide feedback to the anchorwoman or the anchorman of the game. A gaming show has almost the same QoS requirements as videoconferencing. It broadcasts the game playing to all the audience, and 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 QoS information of the transport layer of the wireless.

3.2.1. Application needs in information capabilities

The bit rate of radio link changes quickly. If the bit rate is downgraded and the application still uses the previous bit rate to send in the downlink to the user, the RAN will soon be in congestion state, the user video will be frozen and user's QoE will be very bad. So, the application needs to detect the bit rates of the transport network. The DASH-based mechanism defined in MPEG can be used. However, while the DASH based mechanism is normally for the file download type video streaming, it is not suitable for the interactive video. In the interactive video case, the transport can provide the supported bit rate to the application, and application can adaptively change its video bit rate down to the supported network bit rate and there is no congestion anymore [TS 23.501]. The QoS Flow in the 5G cellular network is established to transport the video streaming for the conference, and application server may request the 5G network to provide network information by the steps as described below, and specified in [TS23.501]:

* The application requests the 5G network to notify the link status and indicate whether the required GFBR (Guaranteed Flow Bit Rate) can no longer be guaranteed or can be guaranteed again.

* The cellular network will notify the application that "GFBR can no longer be guaranteed" if the radio link cannot provide the required guaranteed bit rate. However, it does not specify the amount of guaranteed bit rate. In this case, the cloud video server (normally the medio server) will downgrade the video streaming bit rate in order to avoid the video service being frozen. However, since the video server does not know the currently supported bit rate, the downgraded bit rate may be still higher than the supported bit rate. After receiving the above notification, the video service may still be frozen frequently. If the video server downgrades the bit rate two much, the quality of the video may be too downgraded and the user QoE becomes unacceptable.

* The cellular network will notify the application that "GFBR can no longer be guaranteed and the supported bit rate is XXX" if the radio link cannot provide the required guaranteed bit rate, but the cellular network can provide additional information of supported guaranteed bit rate. Upon receiving this notification, cloud video server can downgrade the video streaming bit rate below the indicated bit rate. In such case, the user QoE does not downgraded too much.

* The cellular network will notify the application that "GFBR can be guaranteed again" if the radio link can provide the required guaranteed bit rate again (for example, the user handovers to another radio station). After receiving this notification, the cloud video server can upgrade the video streaming bit rate to the previous required guaranteed bit rate. In such case, the user QoE can be recover to the best.

The application may inform the cellular network of the following, given that the list below will be completed in the future:

* QoS requirements of application, e.g., the guaranteed bit rate, the alternative guaranteed bit rate.

* The measurement time for the guaranteed bit rate, e.g. 2 seconds (i.e. 2000ms) or 200/500 ms (to improve the response time, i.e. more quickly to provide QoS Notification from the 5G RAN).

The cellular network link status exposed to the application may include the following, given that the list below will be completed in the future:

* "GFBR can no longer be guaranteed",

* "GFBR can no longer be guaranteed and the supported bit rate is XXX",

* "GFBR can be guaranteed again".

All in all, the network should more quickly provide QoS Notification to the application

3.2.2. How the application can get the 5G network information from ALTO

The application (i.e. the ALTO client) can use the restful API to subscribe and be notified by the message from the ALTO Server, the application can adaptively change its encoding scheme to make the bit rate just below the provided network bitrate notified by the network. Also, the application can be pushed to get the message from the ALTO server using the SSE as defined in RFC 8895[RFC 8895].

3.3. Use case 3: ALTO supporting applications on UEs

This section presents use cases where a UE runs an application with the support of an ALTO Client. The application in the UE needs to decide to which application endpoint (AEP) in the Internet to connect. An AEP is assumed to be an application server for e.g. content download, videoconferencing, or other applications for which many servers are deployed. For now, the ALTO-based selection of an AEP is usually done in the network, sometimes with the help of an authoritative entity based on the cost or performance of the path from the UE to the AEP. However, the capabilities of the access network, in the first path IP hop, may highly impact the application performance and therefore needs a deeper insight. The use cases in this section illustrate how network abstraction within the first hop can be beneficial to optimize application path performance.

3.3.1. Use case: Access-aware AEP selection from UE with cascaded ALTO Servers

In this use case, a UE is located in a 4G or 5G network and may connect via several access technologies, e.g. 4G/5G Cellular or WiFi. It is assumed that the UE has subscribed to the same ISP for both fixed and mobile access with a given Service Level Agreement SLA. Users and ISPs tend usually prefer fixed or WiFi connection to cellular, because it is cheaper, more performant and cellular resources are limited. However, it is observed that in many places, including some urban areas in countries with a good average network infrastructure, the fixed network coverage is very poor and worse than the cellular coverage, so that users need to connect via a cell phone or a 4G/5G dongle. Sometimes also, MNOs and ISPs have spare data resources or offer them for free or at low price to users, depending on their SLA. For both parties, access-aware Endpoint selection for users is thus beneficial.

The major QoE challenges in wireless network arise in the access network, that is, in the first hop, between the UE and its one or more associated packet data network gateways (PGW for 4G) or user plane function (UPF for 5G). The path of a UE to its associated PGW/UPF impacts the path to the AEP and thus the application QoE. Therefore, once the PGW/UPF has been selected and will stay unchanged, it is beneficial to help the UE selecting between Cellular and WiFi access.

3.3.2. Scenario and assumptions

The end to end path from the UE to the AEP is considered in 2 parts

* the path from the UE to the PDN,

* the path from the PDN to the AEP.

FOR FUTURE VERSIONS: ARTWORK ON SCENARIO

We assume the availability of multiple cascaded ALTO Servers, as mentioned in section XXXX, to provide a (UE, AEP) path cost. In our "cascaded" use case, we define 2 types of servers involved in conveying the end to end ALTO (UE, AEP) path cost, as follows:

* A Local ALTO Server (LAOS): hosts the information restricted to the local 4G/5G network, covering the paths between e.g. the UEs and the cells or the PGWs/UPFs. It receives the ALTO request issued by the local ALTO Client LAOC associated to the UE. It May host an ALTO Client that can send an ALTO request to a "core" ALTO Server, covering the zone beyond the PGW/UPF, if needed by the application. It composes, when applicable, the response of the "core" ALTO Server with its own response to the LAOC query to obtain a better-informed end to end view of the application path.

* A "core" ALTO Server covers the whole ISP network view, as it would if the "local ALTO Service" is not available or deactivated. That is, it does not see the details below the (UE, PGW/UPF) hop.

3.3.3. Missing ALTO information and features

However, the ALTO information in the path between a UE and an AEP currently provides a cost for one single path only. It does not consider that multiple paths to the PGW/UPF are possible. Currently, the access technology is accounted in RFC 7971 (on deployment cases) in the last hop, to prefer selecting an AEP located in a fixed network over an AEP located in a mobile network. One way to achieve this is that ALTO provides a path cost that, for a given metric, takes multiple values each depending on parameters such as access technology and the access technology or SLA.

4. Highlights on 3GPP Information Useful to ALTO

This section lists the 3GPP information that can be used by ALTO. Either as it comes in, with a minimal encapsulation, or as input to an aggregated form.

In 3GPP specifications [TS23.501, 23.502, 23.503], the following mechanisms have been specified to enable the interaction between AF (application function) and NEF (network exposure function) from network's perspective.

* Alternative QoS profiles,

* notification control.

The application inform the cellular network of the specific QoS requirements, which may include a list of QRP(QoS requirement parameters, such as min/max bandwidth) in a prioritized order. Such requirements will be conveyed to the RAN and the RAN will always try to fulfil the QoS requirement in the list with highest priority. And in the meantime, whether QoS requirements is fulfilled or not in each item of the list will be notified to the application timely. In order to avoid too frequent signalling, it is assumed that RAN can apply hysteresis (e.g., via a configurable time interval) to reduce too much signalling overhead. On the other hand, the QoS values within the list of QoS requirements are required not to be too close to each other.

5. Gap analysis with Existing ALTO features

This section assesses the currently existing ALTO features against the needs in network information and related transport capabilities listed along with the use cases.

5.1. ALTO limits w.r.t. Cellular Network Information

ALTO is by design not expected to provide real time information. The initial use cases were defined for cost maps conveying BGP-based path costs. Later use cases for CDN and video download on individual end-systems support finer grained network information. Nothing though prevents ALTO information to be in minutes or sub-minutes frequency. ALTO may convey values that are valid for a couple of seconds. However, an interval of 2 seconds, in regard of a cellular, network may be too large.

5.2. ALTO Limits on network information transport: gap analysis with ALTO SSE

In RFC8895, ALTO Incremental Updates Using Server-Sent Events (SSE) introduces a mechanism to allow an ALTO server to push updates to ALTO clients to achieve two benefits:

* (1) updates can be incremental, in that if only a small section of an information resource changes, the ALTO server can send just the changes

* (2) updates can be immediate, in that the ALTO server can send updates as soon as they are available.

SSE is a well-shaped pub/sub-like mechanism based on subscription, which quite matches the requirements of proposed cellular use cased in section XXXX, for example, cost effective and immediate. Therefore, SSE can be used as a baseline for protocol extension of cellular use cases. The following message flows can be reused. In the following figure, init request of step 1 is used to subscribe QoS requirements. Data update message of step 2a is used to notify whether subscribed QoS requirements are satisfied or not.

FUTURE VERSIONS: ARTWORK ON SSE COMES HERE

Based on the existing SSE message flows, attributes should be extended for each message flow for the proposed cellular use cases. For example, subscribed QoS parameters and conditions in step 1, the corresponding notification/data update in step 2a.

6. Summarizing ALTO added value and gaps for cellular networks

6.1. Summarizing ALTO added value to cellular use cases

* ALTO abstraction covers several network technologies

* ALTO abstraction covers several network scopes (RAN, edge, core, transport, WAN)

* ALTO can aggregate network information over several technologies and scopes

ALTO can provide end to end path information with insight on selected parts. To our knowledge, standards covering the RAN, the edge, the transport and the WAN (cite 3gpp, ETSI, others) provide a separate network view to applications, if ever. There is therefore no way currently for applications to get an integrated view, allowing more informed decisions. Additionally, ALTO provides abstracted network information, that protects operator confidentiality by exposing only relevant information to applications. This last aspect adds simplicity to the efficiency gained with network-aware decisions. Simplicity all the more allows quick decisions, which is crucial in cellular networks, that have high dynamics.

6.2. Summarizing new ALTO features needed to support cellular use cases

The uses cases presented in this draft are stressing the need for new or extended ALTO features to convey cellular network information. This section also lists a number of concerns raised, in some cases, by the exposure of particular cellular information to third party applications. Other needs may be identified as the cellular use cases will be further investigated.

6.2.1. ALTO Cellular Network Information

ALTO features to represent identify a cell or a WiFi access point.

Abstracted and simplified metrics and costs for wireless networks: ALTO Clients can request a list of supported values for a given set of ALTO metrics. All these metrics are easy to be understood and to be tested by the application developer and application platform (service provider). Examples of useful metrics for our use cases include: throughput/bit rate, latency, priority, error rate, Jitter. Most of these metrics are being standardized in [alto-performance-metrics]. However, their values need to be abstracted from the data provided by the 5G network, typically via the NEF.

Other associated parameters associated to path cost metrics may be useful. For instance, a new type of useful metric would be an abstracted form of the time span to measure the bit rate. The network also needs to expose the attributes of supported alternative QoS. These include alternative throughput/bit rate, the time span to measure the bit rate, latency, priority, error rate, Jitter and associated priority.

Other abstracted and simplified parameters, thresholds or attributes If radio level or low layer information are provided, a gateway is needed to translate or map these information to the high level terminology the ALTO Server. ALTO can then provide the abstracted information to the application. This gateway can be implemented in the southbound of the ALTO server and the gateway can be NEF or NWDA, see reference to 3GPP TS XXx, TS23.288.

6.2.2. Efficient transport for ALTO Cellular Network Information based on SSE

Improvements are necessary of the following ALTO SSE features:

FUTURE VERSIONS: ARTWORK ON EXTENDED SSE COMES HERE

Based on the existing SSE message flows, attributes should be extended for each message flow for the proposed cellular use cases. For step 1 in the procedure of "init request", the following parameters may be included:

* User IP flow ID, or other information relating to UE IP address

* DL/UL capabilities

* A list of QoS requirements and conditions including: Priority, Min bandwidth, Max bandwidth, delay threshold

For step 2a, the corresponding notification/data update can be included regarding subscribed QoS parameters and conditions in step 1.

* User IP flow ID, or other information relating to UE IP address

* DL/UL capabilities

* A list of QoS requirements and conditions including: Priority, current bandwidth or current delay

6.2.3. Time constraints on ALTO-provided Cellular Network Information

The main constraint when conveying ALTO information is speed. When a significant change occurs in the RAN it should be ideally be notified to an application in real-time. As this is not feasible with ALTO, it is necessary to specify constraints such as: acceptable notification delay, maximum delay beyond which the application performance would get degraded, parameters defining an acceptable degradation.

A typical example is as follows. The default measurement time for the guaranteed bit rate is 2000ms, and the maximum rate of notification is 1 time every 2 seconds in the case of data rate fluctuation. This causes too much latency for low latency (lower than 10ms) and/or high date rate (higher than 20mbps) services such as cloud gaming. But it is acceptable for normal latency (e.g. higher than 150ms) and normal data rate (lower than 5mbps) services. The measurement time can be changed to 500ms or 200ms to improve the response time, but the total number of notifications should not increase too much. That is, the total number of notification times should be the same level with the measurement time = 2000ms in a long time span such as 1 hour.

As opposed to fixed networks for which ALTO was initially specified, mobile networks, especially RAN have high traffic and network state dynamics. ALTO is by no means expected to provide real-time information. A careful design of 5G metrics abstraction and hysteresis thresholds triggering ALTO notifications is necessary. The resulting non-real time but highly dynamic ALTO information can reduce the volume of data exchanged with the applications and in some extent facilitate anticipation and reduce oscillations.

6.2.4. ALTO notifications to non-GBR as well as GBR traffic

In 5G networks, notifications to CGS are currently sent only for Guaranteed Bit Rate (GBR) traffic whereas they should be for non-GBR traffic as well. In practice nothing prevents an ALTO Server to do so. To this end, the ALTO Server needs to have the necessary identifiers of the IP flow that is impacted by the network conditions (or QoS level) change.

7. Acknowledgements

8. IANA Considerations

This draft includes no request to IANA.

9. Security Considerations

FUTURE VERSIONS: TBC

10. References

10.1. Normative References

[min_ref] authSurName, authInitials., "Minimal Reference", 2006.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

10.2. Informative References

[DOMINATION] Mad Dominators, Inc., "Ultimate Plan for Taking Over the World", 1984, <http://www.example.com/dominator.html>.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, DOI 10.17487/RFC2629, June 1999, <https://www.rfc-editor.org/info/rfc2629>.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, DOI 10.17487/RFC3552, July 2003, <https://www.rfc-editor.org/info/rfc3552>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <https://www.rfc-editor.org/info/rfc5226>.

Appendix A. Additional Stuff

This becomes an Appendix.

Authors' Addresses

Li Gang
China Mobile Research Institute
Beijing
China

Email: ligangyf@chinamobile.com

Sabine Randriamasy
Nokia Bell Labs
Nozay
France

Email: sabine.randriamasy@nokia-bell-labs.com

Chunshan Xiong
Tencent
Beijing
China

Email: chunshxiong@tencent.com