**3GPP TSG-WG SA4 Meeting #123-e *S4-230504***

**E-meeting, April 17 – 21, 2023 (revision of S4-230xxx)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *CR-Form-v12.2* | | | | | | | | |
| **Pseudo CHANGE REQUEST** | | | | | | | | |
|  | | | | | | | | |
|  | **26.806** | **CR** | **##** | **rev** |  | **Current version:** | **1.1.0** |  |
|  | | | | | | | | |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* | | | | | | | | |
|  | | | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME | **X** | Radio Access Network |  | Core Network | **X** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | |
| ***Title:*** | Clarification on non-5G delay measurement | | | | | | | | | |
|  |  | | | | | | | | | |
| ***Source to WG:*** | Huawei, HiSilicon | | | | | | | | | |
| ***Source to TSG:*** | SA4 | | | | | | | | | |
|  |  | | | | | | | | | |
| ***Work item code:*** | FS\_SmarTAR | | | | |  | ***Date:*** | | | 2023-04-07 |
|  |  | | | |  | |  | | |  |
| ***Category:*** | **C** |  | | | | | ***Release:*** | | | Rel-18 |
|  | *Use one of the following categories:* ***F*** *(correction)* ***A*** *(mirror corresponding to a change in an earlier release)* ***B*** *(addition of feature),* ***C*** *(functional modification of feature)* ***D*** *(editorial modification)*  Detailed explanations of the above categories can be found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | | | | | | | | *Use one of the following releases: Rel-8 (Release 8) Rel-9 (Release 9) Rel-10 (Release 10) Rel-11 (Release 11) … Rel-16 (Release 16) Rel-17 (Release 17) Rel-18 (Release 18) Rel-19 (Release 19)* | |
|  |  | | | | | | | | | |
| ***Reason for change:*** | | In KI#2, several ways of non-5G delay measurements are proposed as candidate solitions. This paper intends to clarify the details for the delay measurement methods. | | | | | | | | |
|  | |  | | | | | | | | |
| ***Summary of change:*** | | Clarification on the details for the delay measurement methods. | | | | | | | | |
|  | |  | | | | | | | | |
| ***Consequences if not approved:*** | | KI cannot be addressed with detailed and correct solution. | | | | | | | | |
|  | |  | | | | | | | | |
| ***Clauses affected:*** | | 6.2 | | | | | | | | |
|  | |  | | | | | | | | |
|  | | **Y** | **N** |  | | | |  | | |
| ***Other specs*** | |  | **X** | Other core specifications | | | | TS/TR ... CR ... | | |
| ***affected:*** | |  | **X** | Test specifications | | | | TS/TR ... CR ... | | |
| ***(show related CRs)*** | |  | **X** | O&M Specifications | | | | TS/TR ... CR ... | | |
|  | |  | | | | | | | | |
| ***Other comments:*** | |  | | | | | | | | |
|  | |  | | | | | | | | |
| ***This CR's revision history:*** | |  | | | | | | | | |

\* \* \* \* First change \* \* \* \*

## 6.2 Key Issue #2: How to determine the non-5G delay for the 5G relay architecture

### 6.2.1 Description of the key issue

In an end-to-end connection that includes a tethering link (e.g., Wi-Fi link), a 5G network and the Internet, the Wi-Fi segment and the Internet segment typically cannot guarantee latency. To achieve low end-to-end latency, one approach is to make the latency in the 5G network very conservative such that the end-to-end latency is below a target value. This, however, comes at a cost, because provisioning an unnecessarily low latency in the 5G network means excessive resource allocation (e.g., to support a more robust modulation-and-coding scheme (MCS)) or pre-empting many other traffic flows.

An alternative approach is to dynamically adjust the delay in the 5G network in accordance with the total delay incurred elsewhere on the end-to-end path. The delay on a Wi-Fi link may change over time depending on the interference generated by other nearby Wi-Fi networks operating on the same frequency. Similarly, the delay between the UPF and the application server depends on the location of this selected UPF and the network congestion level. Therefore, measurements may be used to estimate these time-varying delays on the non-5G segments.

There are two ways to measure the latency and they fill in the details for step 10 in Figure 5.2-5 in clause 5.2.

### 6.2.2 Solution: Segment-by-segment delay measurement

The delay on Wi-Fi link and the delay between the UPF and the application server are measured separately. One simple solution is to use the ICMP ping protocol (ICMP Echo and Echo Reply, IETF RFC792). The 5G phone sends a ping request to the AR glasses, which replies with a ping response. The 5G phone then obtains the RTT over the Wi-Fi link. Similarly, the UPF sends a ping request to the application server, which replies with a ping response, and the UPF obtains the RTT between the UPF and the application server. The respective RTTs can then be halved to get estimates of the one-way delays for the two non-5G segments.

In step 4, the MAF reports the one-way delay estimate to the AF.

In step 7, the UPF reports the one-way delay estimate to the SMF, which forwards the estimate to the AF.

NOTE: How UPF retrieves the RTT between the UPF and the application server and further exposes the latency results to the AF are not supported in SA2 in current release.

In step 9, the AF determines the desired value for the delay in the 5G network needed to compensate for the variation in the delay in the non-5G segments in order to meet the end-to-end latency requirement for the application, and sends a delay request to the PCF.



Figure 6.2.2-1: segment-by-segment delay measurement

### 6.2.3 Solution: End-to-end delay measurement

The delay measurement is carried out in an end-to-end fashion. This avoids the potential rejection of a measurement message that originates from the UPF and reaches the application server. The AR glasses sends a ping request message to the application server, which replies with a ping response. The AR glasses then estimate the UL one-way end-to-end delay by halving the RTT. The 5G network estimates the UL one-way delay within the 5G network , e.g., by recording the time when the ping request arrives at the phone and the time when the ping request reaches the UPF and takes the difference or using the QoS monitoring mechanism to obtain the UL one-way delay within the 5G network.

Editor’s Note: How UPF detects and reports the arrival time of the ping test is not supported yet in SA2.

The estimated UL one-way delay on the non-5G segments is then .

   
Figure 6.2.3-1: End-to-end delay measurement

TS 23.501 [TS23.501, V16.4.0] offers two measurement methods for measuring the delay in the 5G system , originally intended for QoS monitoring to assist URLLC service. The first method, termed “Per QoS Flow per UE QoS Monitoring”, leverages the GTP-U headers to carry the timestamps, and the second method, termed “GTP-U Path Monitoring”, leverages the GTP-U Echo protocol. The first method is shown in Figure 6.2.2-2.



Figure 6.2.3-2: Measuring the delay in the 5G system: Per QoS Flow per UE QoS Monitoring in TS 23.501 [9]

The PCF generates the QoS monitoring policy based on the request from the AF (directly or via NEF) (step 2).

The SMF initiates a QoS monitoring request to the NG-RAN (step 3) and the PSA UPF (step 4).

Step 6: Time stamp T1 is taken in the PSA UPF, indicating the time when the PSA UPF sends a monitoring packet to the NG-RAN (i.e., gNB).

Step 7: the PSA UPF sends a monitoring packet to the NG-RAN, containing T1, QFI and QoS Monitoring Packet (QMP) indicator in the GTP-U header.

Step 8: Time stamp T2 is taken when the monitoring packet is received by the NG-RAN.

Step 10: Time stamp T3 is taken when the NG-RAN forwards an UL packet, or generate a dummy UL packet, where for either case the NG-RAN puts UL/DL packet delay results of RAN part, T1, T2, T3 and the QMP indicator in the GTP-U header.

Step 12: Time stamp T4 is taken when the UL packet is received.

Step 13: Between the NG-RAN and the PSA UPF, if they are synchronized, then the UL delay will be T4-T3, and the DL delay will be T2-T1. If they are not synchronized, then the procedure computes the average one-way delay (T2-T1 + T4-T3)/2.

Step 14: The delay on the access network (between the UE and the NG-RAN) can be added to the results in step 13 to get the total delays in the 5G system, i.e. UL, DL or RT latency.

Finally, the UPF reports the QoS monitoring results to SMF and SMF further reports to PCF. The PCF then exposes the QoS monitoring results to the AF directly or via NEF as requested and the AF eventually obtains the UL/DL or average delay within the 5G System.

Piggybacking timestamps rather than the timestamp message or timestamp reply message to an RTP data packet may reduce the communication overhead. For example, the Timestamp message or Timestamp Reply message are of the following format (RFC792):

0 1 2 3

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| **Type** | Code | Checksum |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| Identifier | Sequence Number |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| **Originate Timestamp** |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| **Receive Timestamp** |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| **Transmit Timestamp** |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Figure 6.2.3-3: The packet format of Timestamp Reply message in RFC792

For the Timestamp message, the size is 12 bytes. If only the timestamp portion – the Originate Timestamp – is piggybacked, the size is reduced to 4 bytes. To let the receiver know the type of the information (i.e., how many timestamps are contained), the Type field can be added. As a result, the total size is 5 bytes. The savings is 7 bytes. This may not look much but can be significant if frequent measurements are needed.

Similarly, for the Timestamp Reply message, the size is 20 bytes. If again only the timestamps and the Type information is included in the piggybacked RTP packet, the size is reduced from 20 bytes to 13 bytes.

Assuming the number of Timestamp messages are the same as the number of Timestamp Reply messages, the average saving is 44%.

### 6.2.4 Time measurement protocol

The ICMP ping protocol uses two timestamps generated at the transmitter to get an estimate of the RTT. The measured delay includes the time gap between the reception of the ping request message and the transmission of the ping response message at the receiver. The time gap contributes to the estimation error, and it depends on the operating system used at the receiver and may become significant for low latency applications.

Alternatively, ICMP timestamp approach (IETF RFC792) can be used, which, compared to ICMP ping, provides the source two timestamps, one for the reception of the timestamp message and the other for the transmission of the timestamp reply message. The two timestamps are carried back to the source, which can use the difference to calculate the time gap and get a more accurate estimate of the RTT.

\* \* \* \* End of changes \* \* \* \*