**3GPP TSG-SA WG1 Meeting #92e S1-204159r1**

**Electronic Meeting, 16 – 25 November 2020** *(revision of S1-20xxxx)*

Title: Update to AMMT use case - Session-specific model transfer split computation decision operation

Agenda Item: 7.4.1 (FS\_AMMT)

Source: OPPO, Samsung?

Contact: Yang Xu <xuyang@oppo.com>

*Abstract:*

*This contribution is a merge of S1-204048 and S1-204159.*

Discussion

This is a merge of S1-204048 and s1-204159. Proposal in 4048 is that the split point is adjusted based on UE performance and proposal in 4159 is that the split point is adjusted based on communication performance. Both are reasonable scenario leading to split point change.

**In S1-204048**

At SA1 91e, the Session-specific model transfer split computation operations enabler use case was added to TR 22.874. That use case focussed on communication resources as input to the policy decision point determining the split computation strategy.

This P-CR further elaborates the input for strategy determination to include UE analytics that are essential for proper formation of a strategy to split computation operations. This is because the split computation strategy can only proceed successfully to the extent that the UE can perform as expected.

As noted in the use case below, the UE has ‘the glasses have limited computational capacity.’ Examining this in more detail, capacities of the UE relevant to the split computation strategy include

* The computation capacity (speed, measured in a metric of work throughput)
	+ The computational capacity will vary over time. The UE may have other work it needs to perform. A UE may only be able to perform work at a peak performance for limited periods of time, as this generates excessive heat.
* The storage capacity of the UE (measured in some units of byte storage, possibly including the latency of the storage – including hardware and I/O performance)
* Energy – how much operating time does the UE have left given its current battery charge, how quickly is it running out of energy, or is the UE plugged in (with external charge)

The ‘UE capabilities’ are not enough for the split computation strategy to be determined, since the capabilities listed above vary over time. The status, that is *current* capability of UEs in the 5G system are captured by the system enabler, to provide the ‘policy decision point’ with crucial information for successful split computation.

**In S1-204159**

As described in the subcaluse 5.5 and also subclause 5.1, each split point has a specific requirement for UL/DL data rate and the computation load (e.g. the number of layers computed in UE and NW side). Generally, the normal regulation for the splitting is “more layer computed in UE, less data rate required” for transmission as illustrated in the table for example (it can also be better described in Figure 5.1.1-1 and Table 5.1.1-1 of usecase in clause 5.1). Thus, when the Guaranteed Bit Rate (GBR) in QoS is changed, to keep the required E2E latency unchanged, the application shall adjust the split point to fulfil the new QoS (GBR).

**Table in subclause 5.5.2 and the corresponding behaviour when GBR is changed**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Approximate outputUL data rate (mbps) | Computation load in UE | Behavior when GBR is changed from 200 to 30 Mbps (example described in subclause 5.5.2) |
| Candidate split point 1  | 120 | Low | Applies when GBR is 200Mbps |
| Candidate split point 2 | 24 | High | Applies when GBR is 30Mbps |

However currently, the change of QoS is notified to UE only when this change has happened, which means some packets will not be handled correctly before the Application layer make an adjustment. In other words, the split point cannot be changed at the same time the new QoS is executed. An example is illustrated in Fig.1 below, where the yellow packets are processed using AI model split point-1 and the green packets are processed using split point-2 due to the change of QoS. There will be a time difference between the occurrence time of new QoS execution and adjustment of split point, because the App does not know the exact time point for the change of QoS. Then the grey packets are still be processed using the old split point-1 but transferred via the connection with new QoS, and the required latecy cannot be fulfilled, i.e. interruption happened for latency sensitive service like AR, VR.



Fig. 1

To avoid the interruption during the new QoS execution, the Appliaction needs to be awared of the condition information (e.g. specific time point, or location) for new QoS execution, so that the split point can be adjusted BEFORE the time that the new QoS is applied.

An example is shown in the Fig.2 below:



Fig 2

With such information (time point-a for QoS change), the Application can makea synchronous adjustment, i.e. change the split point just at the time point-a (as illustrated in Fig.3below). Then the interruption can be avoided since no grey packets (in Fig.1) exist.

Note: the candidate split poin-2 can be applies before time point-a (not necessarity at time point-a), that greatly helps to reduce requirement of accuracy of time sync.



Fig.3

**Updating proposal:**

update the usecase in subclause 5.5 based on the two contributions in the discussion above.

---------- Use Case template ----------

## 5.5 Session-specific model transfer split computation operations

### 5.5.1 Description

A UE, to achieve results for the user, employs split computation. Computation intensive tasks (machine learning, complex computation using input data and the model, etc.) can be fully or partially offloaded. This use case considers a particular use – rendering augmented reality in a headset with modest computational resources. The decision how to split the computation task between the UE and other computation resources depends in part on the conditions of the communication network.

NOTE: The decision of how to split computation is itself out of scope of 3GPP and not discussed here.

### 5.5.2 Pre-conditions

Abigail has Augmented Reality glasses, a UE with limited computational power. She leaves a bus and stands at the bus stop, where, behind a large advertisement display, a gNB is installed. Abigail’s glasses get access through the access point. She seeks to augment her view of the city with directions and annotations (opening hours, local history, description of businesses, etc.) Augmenting the visual scene of the city in real time is a computationally intensive task, accomplished by a model developed through ML. the model has two candidate split points, each candidate split point has a different workload and communication requirement shown as below. A strategy for splitting computation has been indtalled in UE and network server so that the split point can be adjusted based on change of communiactionperformance and/or UE’s capabilities.

|  |  |  |
| --- | --- | --- |
|  | Approximate outputUL data rate (mbps) | Computation load in UE |
| Candidate split point 1  | 120 | Low |
| Candidate split point 2 | 24 | High |

The glasses have limited computational capacity, and this capacity varies over time. A means for identifying the current status (capabilities) of the UE is available to the network. Initially it is determined that the UE has the capability to support either candidate split point 1 or 2.

The network communication resources are enormous, it is determined by the augmented reality service to apply candicate split point 1 so that computation is executed mainly in the network, receiving large quantities of data provided by her glasses and this helps reduce computation in UE. The large quantity of data is transmitted via the QoS flow with guaranteed data rate (GBR) 200 Mbps.

### 5.5.3 Service Flows

**Case-a (split pont adjusted based on communication performance):**

Abigail walks away from the bus stop and the vicinity of the hot spot.

As Abigail stood a few meters from the gNB hotspot, The insufficient communication resources leads to the serving gNB becoming unable to keep the QoS flow with GBR 200 Mbps any more Thus the policy decision point (which could be anywhere – we leave out what takes the decision and how) determines to downgrade the GBR from 200 Mbps to 30 Mbps and immediately notifies UE and Application server of this downgrade will be executed at the time point-a. The strategy for splitting computation for the AR application now must be adjusted, i.e. change to candidate split point 2, for which more computation needs to be done locally but the required bit rate for UL transmission is reduced to 24 mbps. . The strategy and constraints for the partition of work is out of scope of this use case. (These could include e.g. partial results could be sent to the UE, which could perform sub-optimally with reduced resources, can model information be sent in a lossy / compressed form that is still useful, etc.) In any case, one of the crucial inputs to the decision of how to split the work *is the current set of communication resources available.*

The network provides current network resource information concerning the UE to network communication performance such as new QoS paramteres (GBR=20mbps), the condition information (time point-a) for update of a new QoS, as well as end to end performance between the UE and the computation resources (e.g. in the Service Hosting Environment). This information is made available (exposed) to the split computation ‘policy decision point’ (which could be anywhere – in the UE, the edge, the cloud, etc., this is not relevant to the use case.)

**Case-b (split point adjusted based on UE’s capability):**

Originally candidate split point-2 is selected when UE’s capability can support a high work load.

Abigail’s UE’s capabilities are monitored. When the communication resources are sufficient to support Candidate split point 1, if the UE’s capabilities degraded sufficiently (e.g. due to depleted battery, lack of storage, reduced computation capacity) then this would be a reason to select Candidate split point 1.

Then, the split computation decision point then adjusts the split computation strategy:

**For case-a**:to avoid service interruption, the split computation decision point selects the new split point-2 before the time point-a arrives. How this is communicated or ‘enforced’ is out of scope of this use case and it is not suggested that this would be standardized.

**For case-b**: to guarantee the user experience, the split computation decision point selects the split point-1 as the UE’s capability is insufficient to support a high work load anymore.

### 5.5.4 Post-conditions

Abigail has no awareness of the change of model split point and continues to enjoy acceptable performance as she ventures into the city, even if perhaps it is not as good as when she stood at the bus stop. Note that this use case doesn’t conclude as long as Abigail continues to use the service – as the UE to network communication performance can change at any time.

### 5.5.5 Existing features partly or fully covering the use case functionality

22.261 v17.1.0 6.6.2

Based on operator policy, the 5G system shall support an efficient mechanism for selection of a content caching application (e.g. minimize utilization of radio, backhaul resources and/or application resource) for delivery of the cached content to the UE.

NOTE 1: The selection of content caching relies upon knowledge of communication resources (e.g. radio, backhaul, application) for delivery of content to the UE. Thus this requirement satisfies the requirements partly.

22.261 v17.1.0 6.7.2

The 5G system shall be able to provide the required QoS (e.g. reliability, end-to-end latency, and bandwidth) for a service and support prioritization of resources when necessary for that service.

The 5G system shall be able to support E2E (e.g. UE to UE) QoS for a service.

NOTE 2: E2E QoS needs to consider QoS in the access networks, backhaul, core network, and network to network interconnect.

The 5G system shall be able to support QoS for applications in a Service Hosting Environment.

22.261 v17.1.0 6.8

Based on operator policy, the 5G system shall support a real-time, dynamic, secure and efficient means for authorized entities (e.g. users, context aware network functionality) to modify the QoS and policy framework. Such modifications may have a variable duration.

Based on operator policy, the 5G system shall maintain a session when prioritization of that session changes in real time, provided that the new priority is above the threshold for maintaining the session.

22.261 v17.1.0 6.10.2

Based on operator policy, the 5G network shall provide suitable APIs to allow a trusted third-party application to request appropriate QoE from the network.

Based on operator policy, the 5G network shall expose a suitable API to an authorized third-party to provide the information regarding the availability status of a geographic location that is associated with that third-party.

Based on operator policy, the 5G network shall expose a suitable API to allow an authorized third-party to monitor the resource utilisation of the network service (radio access point and the transport network (front, backhaul)) that are associated with the third-party.

### 5.5.6 Potential New Requirements needed to support the use case

[P.R.5.5-001] Based on operator policy, the 5G network shall provide the means to allow an authorized third-party to monitor the resource utilisation of the network service (radio access point and the transport network (front, backhaul)) that are associated with the third-party. The resource utilization includes the realtime decision for change of QoS made by 5G network.

[P.R.5.5-002] Based on operator policy and subject to user consent, the 5G network and UE shall provide the means to allow an authorized third-party to monitor the current status of the UE in terms of its computation capacity (e.g. in units of work per unit of time), energy status (e.g. time of remaining charge, ‘external power supply,’) and available storage capacity (e.g. in kilobytes.)

[P.R.5.5-003] Based on operator policy, the 5G system shall be able to update a new QoS with a condition (e.g. at a time point).

[P.R.5.5-004] Based on operator policy, the 5G system shall be able to expose the condition information such as time point for updating a new QoS to an authorized 3rd party so that the 3rd party can make an adjustment of candidate split point before the update occurs.

NOTE: how the 3rd party adjust the candidate split point based on the condition is out of 3GPP scope.

.

Editor’s Note: It is FFS whether and how (units and measurement frequency) granularity of spatial information andresource utiliziation listed in the above requirements can be expressed in KPIs.