**3GPP TSG-SA WG1 Meeting #108 S1-244505**

**Orlando, USA, 18-22 November 2024** *(revision of S1-244085 and S1-244086)*

Title: New use case on Immersive Service Discovery and Coordination

Agenda Item: 8.1.4

Source: Nokia

Contact: Hideaki Takahashi; hideaki dot takahashi at nokia dot com

*Abstract: This contribution proposes a new use case on immersive service discovery and coordination to discover service instances of immersive communication present in the network and offload one or more service processing from UE to the service instances of immersive communication.*

Updates in S1-244505;

- S1-244085 (immersive service discovery) and S1-244086 (immersive service coordination) are merged.

- Proposed requirements are generalized not specific to split rendering, but generic to immersive services

- Added editor’s note to open the door for any other requirements.

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

## 8.X Use case on Immersive Service Coordination

### 8.X.1 Description

Immersive communication is an emerging paradigm to enable immersive user experience. For immersive communication services, multiple computing and data storage resources are deployed at various locations aiming at e.g., reduced latency, bandwidth efficiency, enhanced security, reliability, etc. XR is a prominent application where multiple compute resources are needed to perform edge/cloud rendering, i.e., rendering at the closer site to a user or a central site in the cloud. Each computing service instance, depending on its allocated compute resources, can handle a specific number of users. Whilst these computing resources can be scaled up to accommodate more users, the capacity of computing site ultimately limits the scalability.

The existing 5GS is capable of offering networking assistance to Edge Computing services, with the primary focus being fulfilling or optimizing a target network performance, e.g., latency. With this approach each user (UE, application client) is provided with compute resources on a site closest to the user. This may lead to situations where some compute sites get overloaded while others are underutilized. In order to optimally reserve or balance the use of computing resources across the compute sites while ensuring the application’s end-to-end performance and user experience, the compute site or application server selection should be performed based on both network and compute metrics, with any site or server meeting the application’s networking requirements considered a candidate. Figure x.1.1 illustrates a conceptual scenario that an optimal compute site is discovered by means of both network and compute metrics.

In a conventional edge computing scenario, a computing instance is always in-operation and running at least in one of the multiple computing sites. In the future, an advance scenario can be envisaged that a computing instance can be deployed and in-operation in an on-demand basis as per service request from applications or users. The motivations behind the on-demand service are as follows:

- The application servers required by each individual application need not be pre-deployed on every computing site, but this can be done on-demand especially for the applications that are more rarely used.

- To save energy for running computing servers, sites, etc.

Some technologies allow for rapid instantiation such as serverless, WASM, etc.

The on-demand instance deployment use case aims to select an optimal compute site, to deploy a computing service instance, that meets network and compute requirements of the application or the user. For instance, in AI co-inference scenarios, a UE could utilize a pruned model (i.e., a model with less parameters to reduce its size) locally that is suited for most frequent tasks, while the larger model is stored to be deployed on-demand as per the request of the user. When the application starts requesting non-frequent tasks, the large model could be deployed at compute site that meets both network and compute requirements of the user. This approach would reduce resource consumption for both the UE (reduced local processing), the compute site (activated only when needed), and the application provider (cost incurred only during active use). Figure x.1.1 also covers the on-demand use case that some computing service instances are in-operation in the on-demand basis in addition to the pre-deployed service instances running all the time.



Figure 8.x.1-1: Computing Service Discovery by means of network and compute metrics

In immersive communications, it is typical that the whole application service running in the user device (i.e., UE) is composed of one or more software modules. Depending on dynamic compute and network capabilities and availability, a particular software module could run in the UE or be “offloaded” to a compute site. The motivations of such a use case are as follows:

- By offloading a software module to the compute site, the UE could better reduce its resource consumption (e.g., energy and local processing).

- As the computing resource in the network has more resources than UE typically, the offloaded software module could run more efficiently, resulting in enhanced Quality of Service.

-

For instance, some XR applications use different modules such as ‘scene manager’ module, ‘object detection’ module, ‘rendering’ module, etc. With the consideration of the imaging technologies of live service migration, these modules could be designed to support relocation to a target server that is also designed to support offload. In this regard, the UE would be able to reduce its consumption by offloading a module to a compute site as long as the required QoS is retained.

Figure x.1.1 illustrates a conceptual workflow on how the software module is relocated and offloaded according to the dynamic statistics, e.g., UE power consumption, network latency, etc.



Figure 8.x.1-2: Immersive Service Coordination

**Potential sustainability impacts of the use case**

|  |  |  |  |
| --- | --- | --- | --- |
|  | (the UN SDGs/GDC matching goals of each aspect within 3GPP context) | **Potential benefits of the use case (added value)** | **Potential areas of attention of the use case (risks to be mitigated)** |
| **Environmental sustainability aspects**(UN SDGs 12, 13, 14, 15 and indirectly 6, 7 & 11UN GDC “Develop principles for environmental sustainability of digital technologies”) | **Energy resources**(UN SDG 7, 11, 12) | * Reducing device energy consumption, by offloading one or more modules to compute sites
* Reducing the network energy consumption too by taking metrics of energy consumption into account.
 | - success of such use case may trigger more low cost end-user devices and hence as rebound effect generate more energy consumption overall if an additional personal IoT device and not a replacement device of the smartphone for end users |
| **Material resources**(UN SDG 11, 12) | * - Reducing complexity of end-user devices (and thus e-waste) by leveraging network/edge compute capabilities
 | * - success of such use case may trigger more low cost end-user devices and hence as rebound effect generate more material extraction to build these devices, as well as globally additional e-waste
 |
| **Emissions**(UN SDG 6, 7, 11, 12, 13, 14, 15) | * Enabling CO2 emission reduction in the overall system by means of reduced energy consumption, which sources of energy are reduced (network and compute sites, instead of UEs) and can be better managed towards low carbon.
 | * success of such use case may trigger more low cost end-user devices and hence as rebound effect generate more ghg emissions
 |
| **Socio-economic sustainability aspects**(UN SDGs 2, 3, 4, 5, 8, 9, 10, 11, 16 & 17 and indirectly 12)UN GDC “Closing Digital Divides and Accelerating SDG Progress”&“Expanding Digital Economy Inclusion”&“Fostering an Inclusive, Safe Digital Space”) |  |  |  |
| **Health**(UN SDG 3) | * - Offload enable lightweight devices that can thus be worn for a longer period
 | * effects of massive usage of wearables may have negative impact on health (EMF etc) or safety risks (be isolated from surrounding environment)
 |
| **Inclusion & Equality** **(**UN SDGs 11, 10, 4, 5 and indirectly 3, 16 & 17) | * offload enables low cost end devices more largely affordable by the population
* Wearable devices – eg handsfree, lightweight – can also be used by a larger population base beyond some disabilities
 |  |
| **Trustworthiness (**UN SDGs 11 and indirectly 3 & 17) | * Offload to the compute sites may provide various technical alternatives to local UE computing thus increasing the availability and resilience of the system towards service continuity
 | * privacy of data being offloaded (eg image/video of people’s face) has to be taken into consideration when offloading
 |
| **Work & income**(UN SDG 8 and indirectly 12) | * offload can enable longer battery time in operations, in particular for remote workers, thus increasing productivity
 |  |
| **Infrastructure**(UN SDG 9) | * ***-*** offload can benefit remote workers for many infrastructure sectors with longer battery times and lightweight, more affordable devices
 |  |
| **TCO reduction**(UN SDGs 8, 9 and 12) | * offload can benefit several sectors with more affordable devices, thus making some business cases of remote workers commercially viable
 | * extra cost for investment in end-user devices and subscriptions to access offload capabilities, which should be traded-off with productivity gains
 |

### 8.X.2 Pre-conditions

A Computing Service Provider (CoSP) is a provider to offer their computing resources, e.g. for immersive communications. MNO or the 3rd party can be a CoSP. The CoSP deploys one or more immersive application service instances (e.g., split rendering, etc.) at various physical sites.

Subject to CoSP’s policy, split rendering service instances are always in-operation, or instead, not all of them are always in-operation and some of them are deployed only when a service is requested by users, i.e., on an on-demand basis.

In CoSP’s platform, while a split rendering service instance is active, its computing resource status (e.g., CPU/memory load of the computing service) is measured, which is to be used as a metric of computing performance. Additionally, computing resource status of the compute site (e.g., available CPU/memory of the compute site) is also measured, which is to be used for on-demand deployment. In case MNO plays a role of CoSP, the MNO network is also capable of measuring network performance (e.g., latency between a UE and an application service instance at a given compute site) when requested for a given UE.

A user subscribes immersive application services (e.g., XR applications, etc.) provided by the MNO and/or the ASP by leveraging CoSP’s split rendering service instances at their physical sites.

In the initial configuration, some modules are processed at the UE (e.g., module #A and #B as in Fig. x.1.1), whereas the other is at a (central) physical site (module #C as in Fig.8. x.1-2), which is hosted by the CoSP.

In CoSP’s platform, while a split rendering service infrastructure is active, its computing resource status (e.g., CPU/memory load of the immersive service) and its energy consumption is measured, which is to be used as a metric of computing performance and (energy) cost. The MNO network is also capable of measuring network performance (e.g., latency between a UE and a computing instance) when requested for a given UE.

### 8.X.3 Service Flows

For immersive service discovery as illustrated in Fig. 8.x.1-1;

1. The user invokes the subscribed immersive application and request the service to MNO’s or ASP’s application host.

2. The ASP or MNO requests the CoSP to provide the compute metrics per split rendering service instance or site where the target immersive application can be provided for user’s UE requesting the application service.

3. In case a slit rendering service instance or site is in-service, User’s immersive application or ASP collects from CoSP, the computing resource status of the split rendering service instance or the compute site as well as the network performance in relation to user’s UE. In case MNO plays a role of CoSP, the network performance is obtained from 6G network, whereas compute performance is obtained via a different layer than the one where 6G network encompasses.

5. User’s immersive application or ASP selects an optimal split rendering service instance, taking into account the computing and network metrics for each application service instance or compute site obtained from CoSP.

6. If the selected split rendering instance is provided on an on-demand basis, user’s application or ASP requests on-demand computing service provisioning.

For immersive service coordination as illustrated in Fig. 8.x.1-2;

1. The UE processing software modules #A and #B recognises that power consumption is significantly soared.

2. The immersive application client at the UE requests MNO’s or ASP’s application server to offload these software modules from the UE, possibly indicating information related to the associated power consumption of these module at UE side.

3. The ASP or MNO requests the CoSP to provide the compute & energy metrics per compute site where the software modules #A and #B may be relocated.

4. User’s application or ASP collects from CoSP, the computing resource status and energy consumption of the compute site as well as the network performance in relation to user’s UE. In case MNO plays a role of CoSP, the network performance and energy consumption are obtained from 6G network, whereas compute performance is obtained via a different layer than the one where 6G network encompasses.

5. User’s application or ASP selects an optimal compute site to host the software module #B, taking into account the computing, energy and network metrics for each compute site obtained from CoSP . Based on the same information, and on the information provided by the UE, User’s application or ASP decides that software module #A should remain on the UE.

6. The software module #B is relocated from the UE to the selected compute site (the central site in Fig. x.1.1).

7. Later, the application client or server recognises that RTT becomes too long between UE and the compute site hosting software modules #B and #C, which hinders the processing of software module #B from satisfying QoS.

8. The application client or server again collects the compute performance and energy consumption of available compute sites and select an optimal site to host the software module #B fulfilling both computing, energy and network metrics.

9. The software module #B is again relocated to the selected compute site (the edge site in Fig. 8.x.1-2).

### 8.X.4 Post-conditions

The immersive application user can enjoy optimal performance and user experience by means of the metrics of computing and network. The immersive application client or server can relocate and offload one or more software modules in order to retain optimal QoS and user experience over dynamic compute, energy and network characteristics. In addition, the MNO and ASP can contribute to save energy consumption by means of the on-demand basis service provisioning and enhance their annual ESG report.

### 8.X.5 Existing features partly or fully covering the use case functionality

None.

### 8.X.6 Potential New Requirements needed to support the use case

[PR 8.X.6-1] Subject to operator’s policy and an authorized 3rd party and/or user consent, the 6G system shall be able to assist immersive applications in selecting an immersive service instance or site which is optimal, including based on network, energy and compute metrics, and immersive application’s networking and compute requirements.

[PR 8.X.6-2] The 6G network shall be able to receive information about immersive service instances or sites for which its selection assistance applies (e.g., identification, addressing and capability information of the immersive service instances or sites).

NOTE: The capability information of the immersive service instances or sites includes information on which metrics exposure or determination methods (e.g., measurement result of metrics) the immersive service instances or sites supports towards the 6G Network.

[PR 8.X.6-3] The 6G network shall be able to determine, e.g., by measurements using the measurement results of metrics supported by the immersive service instances and sites, the network metrics (e.g., latency) between a UE and each potential immersive service instance or site.

[PR 8.X.6-4] The 6G network shall be able to expose network metrics (e.g., latency) between a UE and each potential immersive service instance or site to an authorized 3rd party.

Editor’s note: FFS on further potential new requirements.