**3GPP TSG-SA WG1 Meeting #98e S1-221076r4**

**Electronic Meeting, 9 – 19 May 2022** *(revision of S1-22xxxx)*

**Source: China Mobile, China Telecom, vivo**

**Title: New SID on supporting computing aware network**

**Document for: Approval**

**Agenda Item: 4**

3GPP™ Work Item Description

Information on Work Items can be found at <http://www.3gpp.org/Work-Items>
See also the [3GPP Working Procedures](http://www.3gpp.org/specifications-groups/working-procedures), article 39 and the TSG Working Methods in [3GPP TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm)

Title: Study on computing-aware network

Acronym: FS\_Compute

Unique identifier:

Potential target Release: Rel-19

# 1 Impacts

{For Normative work, identify the anticipated impacts. For a Study, identify the scope of the study}

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Affects: | UICC apps | ME | AN | CN | Others (specify) |
| Yes |  |  |  | x |  |
| No |  | x | x |  |  |
| Don't know | x |  |  |  |  |

# 2 Classification of the Work Item and linked work items

## 2.1 Primary classification

### This work item is a

|  |  |
| --- | --- |
|  | Feature |
|  | Building Block |
|  | Work Task |
| X | Study Item |

## 2.2 Parent Work Item

For a brand-new topic, use “N/A” in the table below. Otherwise indicate the parent Work Item.

|  |
| --- |
| Parent Work / Study Items  |
| Acronym | Working Group | Unique ID | Title (as in 3GPP Work Plan) |
| N/A | N/A | N/A | N/A |

### 2.3 Other related Work Items and dependencies

|  |
| --- |
| Other related Work /Study Items (if any) |
| Unique ID | Title | Nature of relationship |
| 860009 | Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS | Have related use case in computation offloading between UE and network |

# 3 Justification

With the development of 5G network, new services such as AR/VR, cloud gaming, Metaverse, V2X, remote healthcare, Industrial Internet, etc. have much higher requirements on both computing aspect and network aspect. For the network aspect, these scenarios require more in high bandwidth, low latency, high reliability, and high flexibility. For the computing aspect, computing resources of different categories (including CPU, GPU, and FPGA) and storage resources are to be supported, and cloud computing and edge computing are introduced.

However, the computation aspect and network communication aspect are considered separately now, for services which have high demand of computation, 3rd party usually evaluate the end-to-end service, and provide communication requirements to 5GS, for example, if a VR service need 30ms end-to-end latency, application will tell 5GS to satisfy 10ms latency from UE to UPF, and they use the rest 20ms for transportation after N6, do the rendering and processing. From this example we can see that in this VR service, the application tells 5GS which is the target server so that 5GS can do the optimization within itself to satisfy 10ms, and the rest part is for the application itself to do the optimization in rendering, processing etc. In this case, the coordination between this two parts is in a low-efficient way, because it takes some time to transfer some key information from the application on the target server to the network. As neither of network nor the application has the global view of both network topology and service endpoint, it will cause delay when service endpoints need to be updated, and also will cause the end-to-end path is not the best for user experience. So it’s better that some information on service endpoint could be exposed to 5GS, so that there’s global optimization of both network and service processing, which could help achieving balance among different dimensions, lower the cost of computing resource, memory resource and total processing time consumption.

From another aspect, as we know about the energy consumption of data centres are really high. For an individual cloud service provider, to provide better service experience to customer with wider coverage and lower latency, it has to set up more data centres, which leads to higher economic cost and energy consumption. So it’s valuable if we can use the existing data centre more efficiently. However, now different cloud data centres and edge compute resources belong to different service providers, they are not intercommunicating with each other, and there is no platform which can be aware of all these computation capabilities, it’s will be beneficial if these computation capabilities could be exposed and complement each other. Considering this work is a complicate and difficult, it is proper to start with the computation resources controlled by operators.

Based on the above reasons, it’s valuable for operators to have this kind of platform to provide joint optimization of network and computation resources. As operators have better deployment in edge, the whole picture of end-to-end path will be clear and easy to optimize from operator perspective. With the information about service endpoints or servers exposing to operators, the connection could change from "no business perception, pure network" to "perceived user business needs, and establishes an on-demand connection between data and computation services".

So we introduce computing aware network, which provides the capability to be aware of computation capability within network (e.g. computation capability on the edge), so that 5GS can consider both network topology and service endpoints to optimise network path, offload computing task to available computing resource.

# 4 Objective

This study is aiming at identifying use cases, providing gap analysis and defining potential requirements in the following aspects regarding computing aware network.

The objectives include:

* Study use cases and potential requirements related to opportunities to add standardized support to mobile network controlled network and computation resources used to deploy the 5GS, including:
	+ Potential requirements for enforcing and coordinating service performance (e.g. latency, etc.) of network and operator controlled computing resources (e.g. edge resource) in order to achieve overall service efficiency.
	+ Potential enhancement in information exposure between network and applications on computing related information for optimal end-to-end service.
* Identify gaps between the use cases above and existing support specified in stage 1, orchestration and performance management specifications in SA5 and orchestration and management standards specified in ETSI ISG NFV.
* Identify KPIs relating to computing capability including latency (e.g. network latency and processing latency), reliability, etc.
* Other aspects include security, charging and privacy.

# 5 Expected Output and Time scale

|  |
| --- |
| New specifications {One line per specification. Create/delete lines as needed} |
| Type  | TS/TR number | Title | For info at TSG#  | For approval at TSG# | Rapporteur |
| TR |  | Study on supporting computing aware network | SA#98 (Dec 2022) | SA#99 (Mar 2023) | Xiaonan Shi |

|  |
| --- |
| Impacted existing TS/TR {One line per specification. Create/delete lines as needed} |
| TS/TR No. | Description of change  | Target completion plenary# | Remarks |
|  |  |  |  |

# 6 Work item Rapporteur(s)

Xiaonan Shi, CMCC, shixiaonan@chinamobile.com, Primary Rapporteur,

Jiayifan Liu, China Telecom, liujyf@chinatelecom.cn, Secondary Rapporteur.

# 7 Work item leadership

SA1

# 8 Aspects that involve other WGs

None.

# 9 Supporting Individual Members

{At least 4 supporting Individual Members are needed. There is an expectation that these companies will provide resources to progress the work. Note that having 4 supporting companies is a necessary but not sufficient condition: the usual TSG approval process by consensus is needed for the WID approval}

|  |
| --- |
| Supporting IM name |
| China Mobile |
| China Telecom |
| CATT |
| ZTE |
| Spreadtrum Communications |
| CEPRI |
| Alibaba Group |
| Xiaomi |
| vivo |
| OPPO |
| Asia Info |
| Lenovo |
| Motorola Mobility |
| InterDigital |
|  |