3GPP TSG SA WG5 Meeting 134-e S5-206166

**electronic meeting, online, 16th - 25th November 2020**

**Source: Nokia**

**Title: Load Balancing Optimization Solutions based on Resource Consumption and Service Specifics**

**Document for: Approval**

**Agenda Item: 6.5.1**

# 1 Decision/action requested

***Propose to update use case and solution in subclause 6.5.3***

# 2 References

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TS 28.550: "Management and orchestration; Performance assurance".

[3] 3GPP TS 28.533: "Management and orchestration; Architecture framework".

[4] 3GPP TS 28.530: "Management and orchestration; Concepts, use cases and requirements".

[5] 3GPP TR 28.861: "Study on the Self-Organizing Networks (SON) for 5G networks".

[6] 3GPP TR 28.805: "Study on management aspects of communication services".

[7] 3GPP TS 28.554: "5G end to end Key Performance Indicators (KPI)".

[8] 3GPP TS 28.552: "Management and orchestration; 5G performance measurements".

[9] 3GPP TS 22.101: "service aspects; service principles".

[10] 3GPP TS 32.500: "Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements".

[11] 3GPP TS 37.816: "Study on RAN-centric data collection and utilization for LTE and NR".

[12] 3GPP TS 37.320: "Radio measurement collection for Minimization of Drive Tests (MDT); Overall description".

[13] 3GPP TS 23.501: "System Architecture for the 5G System (5GS); Stage 2".

[14] 3GPP TS 28.310: "Energy efficiency of 5G".

[15] 3GPP TR 21.866: "Study on Energy Efficiency Aspects of 3GPP Standards".

[16] 3GPP TS 26.247: "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)".

[17] 3GPP TS 26.114: "IP Multimedia Subsystem (IMS); Multimedia Telephony; Media handling and interaction".

[18] 3GPP TS 23.288: "Architecture enhancements for 5G System (5GS) to support network data analytics services".

[19] 3GPP TS 28.313: "Self-Organizing Networks (SON) for 5G networks".

[20] 3GPP TS 28.541: "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3".

[21] 3GPP TS 38.304 NR: "User Equipment (UE) procedures in idle mode and in RRC Inactive state".

[22] 3GPP TS 28.545: " Management and orchestration; Fault Supervision (FS) ".

# 3 Rationale

This contribution enhances the current load balancing optimization use case, requiremetns and proposes two solution.

# 4 Detailed proposal

**First Change**

### 6.5.3 Load Balancing optimization

#### 6.5.3.1 Use case

The rapid traffic growth and multiple frequency bands utilized in a commercial network make it challenging to steer the traffic in a balanced distribution. To address the problem, load balancing had been proposed. The objective of load sharing and load balancing is to distribute cell load evenly among cells or to transfer part of the traffic from congested cell, or to offload users from one cell or carrier or RAT to improve the network resource utilization efficiency and achieve network energy saving. This can be done by means of optimization of cell reselection/handover parameters and handover actions.

To ensure the service performance and user experience, the load balancing action based on handovers highly depends on the measurement report (MR) from the UE. For example, the inter-frequency scenarios with the deployment of multiple different frequency bands, the MR configuration and UE MR reports may cause amount of signalling overhead over Uu interface. The frequent inter-frequency measurement will cause huge UE power consumption and severely impact on running service by the data interruption for inter-frequency measurement gap, e.g. the gap time in LTE is number of frequency\*60ms per 480ms period and the gap time in NR also depends on SMTC period. The gap assistant inter-frequency measurements mechanism will bring delay of the measurement and decrease the data transmission rate. Solutions are desired to improve the effectiveness of the MR configuration and report, which may help to greatly reduce the MR signalling overhead, UE power consumption and data interruption of running service, and improve the convergence speed of the load balancing.

The MDA can help to predict the measurement results of cell on neighboring frequencies for each UE without the GAP assisted measurement. Via analyzing the historical intra-frequency and inter-frequency measurement from both the serving cell and the neighbour cell, the MDA can construct the network “radio finger print”, which characterize the network intra-frequency and inter-frequency coverage quality. The “radio finger print” information is composed of multiple virtual grid. The grid index is to identify a specific virtual grid and this index consists of cell ID and corresponding coverage quality, e.g., RSRP, of at least three intra-frequency cells. The attributes of the grid are used to describe the wireless characteristics of the grid, such as coverage of inter-frequency neighbor cells, including RSRP, reference signal receiving quality (RSRQ), received signal strength indication (RSSI), channel quality indicator (CQI), modulation and coding scheme (MCS), beam ID, etc.

The MDA producer provides the analytics report on “radio finger print” information to the gNB, gNB can directly predict the measurement values of cells on neighboring frequencies for each UE based on the well-constructed “radio finger print” and the real-time intra-frequency measurement. In this case, the GAP assisted inter-frequency measurement is avoided, and the gNB can make proper load balancing actions based on the predictions, which helps to reduce the data interruption of running services and improve the load balancing speed. Moreover, the MDA producer can provide authorised consumers with radio configuration options to perform MLB based handover according to the Service Experience Type, e.g. on voice, video, other, associated with the offloaded UE from one cell to another to ensure the desired service experience.

The MDA producer may also provide the traffic load prediction report to the authorized consumers, e.g., gNB, to enable the proactive load balancing actions. This would help to prevent the user experience degradation in advance compared to the reactive optimizations based on the delayed load information measurement and exchange.

#### 6.5.3.2 Potential requirements

**REQ-MLB\_OPT\_CON-1** The MDAS producer should have a capability to provide the analytics report describing the radio measurement information to authorized consumers , e.g., gNB.

**REQ-MLB\_OPT\_CON-2** The analytics report describing the radio measurement information should contain the following information:

- the applied cell ID;

- the time period(s) of the original data used for deriving the analytics report;

- the serving cell and its inter-frequency/intra-frequency neighboring cell’s cell ID and corresponding radio measurement information, e.g., CSI-RSRP, SS-RSRP, etc;

- Indication on whether the gNB is suitable to be selected as the target gNB for the MLB based handover based on the radio signal qualities.

**REQ-MLB\_OPT\_CON-3** The analytics report describing the predicted resource utilization status of gNB should contain the following information:

- predicted virtual, radio, and transport resources utilizations for potential MLB source and target gNBs in the near future;

- Indication on whether the gNB is needed to activate the MLB operation;

- Indication on whether the gNB is suitable to be selected as the target gNB for the MLB based handover.

**REQ-MLB\_OPT\_CON-4** The MDAS producer should have the capability to provide authorized consumers, e.g. gNB, with the analytics report describing the service specific radio configuration options needed to perform MLB based handover.

The solution considers resource consumption both in terms of virtual and radio resource for the target gNB. The current resource consumption is analysed with the future/predicative resource consumption to decide if the target gNB is optimal for handover or not.

#### 6.5.3.3 Possible solutions

##### 

##### 6.5.3.3.1 Solution description

The solution considers resource consumption both in terms of virtual and radio resource for the target gNB. The current resource consumption is analysed with the future/predicative resource consumption to decide if the target gNB is optimal for handover or not.

##### 6.5.3.3.2 Data required

The following data is required to do the required analysis.

|  |  |
| --- | --- |
| **Data category** | **Required data** |
| Allocated Virtual Resource of the source and target gNB | Allocated Compute: This describes the number of vCPUs allocated to the virtual machine on which the gNB VNF is hosted.  Allocated Memory: This describes the number of vMemory allocated to the virtual machine on which the gNB VNF is hosted.  Allocated Storage: This describes the number of vStorage allocated to the virtual machine on which the gNB VNF is hosted. |
| Consumed Virtual Resource of the source and target gNB | Consumed Compute: This describes the number of total aggregated compute resource consumption at a particular point of time.  Consumed Memory: This describes the number of total aggregated memory consumption at a particular point of time.  Consumed Storage: This describes the number of total aggregated storage consumption at a particular point of time. |
| Consumed Radio Resource of the source and target gNB | Radio resource utilization: The physical radio resource utilization, see clause 5.1.1.2 of TS 28.552[8]; |
| MDT data | UE measurements related to RSRP, RSRQ, SINR time-stamped. |
| Service in use data analytics | Service in use statistics and service in use predictions provided by NWDAF per UE (source/target cell) |

##### 6.5.3.3.3 Analytics report on gNB resource consumption

The gNB resource analytics report contains the following information for the source and target gNBs.

|  |  |  |
| --- | --- | --- |
| **Analytics Report of gNB resource consumption** | **Attribute Name** | **Description** |
| Allocated Virtual Resource | Allocated Compute: This describes the number of vCPUs allocated to the virtual machine on which the gNB VNF is hosted.  Allocated Memory: This describes the number of virtual vMemory allocated to the virtual machine on which the gNB VNF is hosted.  Allocated Storage: This describes the number of vStorage allocated to the virtual machine on which the gNB VNF is hosted. |
| Consumed Virtual Resource | Consumed Compute: This describes the number of total aggregated compute resource consumption at a particular point of time.  Consumed Memory: This describes the number of total aggregated memory consumption at a particular point of time.  Consumed Storage: This describes the number of total aggregated storage consumption at a particular point of time. |
| Predicted Virtual Resource consumption | predicted Compute: This describes the number of total predicted compute resource consumption at a particular point of time.  predicted Memory: This describes the number of total predicted memory consumption at a particular point of time.  predicted Storage: This describes the number of total predicted storage consumption at a particular point of time.  Timestamp: Time for which the prediction is made. |
| Assigned radio resources | The physical radio resource assignment to the target gNB. |
| Consumed radio resource | The physical radio resource utilization of the target gNB. |
| Predicted radio resource | The physical radio resource predicted utilization of the target gNB. |
| isMLBNeeded | Indication on whether the target gNB needs to activate the MLB operation. |
| isOptimal | Indication on whether the target gNB is suitable to be selected as the target gNB for the MLB based handover. |

##### 6.5.3.3.4 Analytics report on service specific radio configuration for MLB based handover

The service specific radio configuration for MLB based handover contains the following information.

|  |  |  |
| --- | --- | --- |
| **Analytics Report for service specific radio configuration for MLB based handover** | **Attribute Name** | **Description** |
| Service list | List of supported service classes or corresponding slice IDs |
| Recommended actions | Recommendation for optimal gNB configuration and/or target gNB selection/prioritization based on the service class or slice ID. |

#### 6.5.3.4 Gap Analysis

* Introduce service-identifier for each radio configuration to support a service-specific MLB based handover
* Input PM data: Virtual Resource usage, RAN utilization are specified in TS 28.552 [8].
* MDT data is specified in TS 37.320 [12].
* Service experience analytics is specified in TS 23.288 [18].
* MDAS resport on gNB resource consumption.
* MDAS resport on service specific radio configuration for MLB based handover.

**Open Issues:** None.

**Feasibility**: Solution is feasible based on current performance measurements.

**End of Change**