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.

The Centralized SON can have a bigger picture of the Network Congestion, by looking at a Cluster of cells. This can complement the load balancing at Distributed SON and make it possible to completely mitigate the risks of service degradations due to network congestions. Moreover, predictive analytics can make it possible to take anticipatory steps towards mitigating network congestions. The MDAS producer can predict the cells that will be congested in a near future, in a cluster of NR Cells, by performing data analytics on Cell Level Performance Measurements: DL PRB

Utilization, Latency related Performance Measurements, Average RRC Connected Users, DL Packet Drop Rate, DL PDCP SDU Drop Rate, Average Delay in DL in CU-UP, Average Delay on F1-U, IP Latency. Network Congestions can be predicted at Cell level as well as group of neighbouring cells. Based on this, a list of congested cells can be provided to the MDAS Consumer. Also, a list of cells for which it is predicted that the congestion will ease out, can be provided to the MDAS Consumer.

#### 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.

For this Solution, MDAS Consumer shall be C-SON. The MDAS producer can collect Performance Measurement data as described in Clause 6.5.3.3.2. Performance Measurement data shall be captured for a cluster of NR cells in Operator’s network. The MDAS Consumer can provide a mechanism for selecting a cluster of cells, depending upon geographic region, or, based on the network density. The periodicity of the Performance Measurement collections and the storage duration can be configurable. The prediction is required for a time period in advance, defined as Prediction Horizon, which can be a configurable parameter. The MDAS Producer can collect historic Performance Measurements in order to arrive at the congestion predictions.

Prediction shall be made well in advance, in order to compensate for possible latencies in collecting Performance Measurements as well as latencies in applying the mitigation actions to the network. The MDAS consumer can prevent Network congestion by applying cell reselection/handover parameter changes in advance, based on the Prediction Report from the MDAS producer.

##### 6.5.3.3.2 Data required

The following data is required to do the required analysis.

|  |  |
| --- | --- |
| **Data category** | **Required data** |
| Virtual Resource Measurements | For every Cell in the selected Cluster, Common performance measuremetns for NFs including: Virtualised resource usage, Virtual memory usage, Virtual disk usage, clause 5.7, TS 28.552 [8] |
| Resource Utilization Measurements | Radio resource utilization: The physical radio resource utilization of the source and target gNB, see clause 5.1.1.2 of TS 28.552 [8];  For every Cell in the selected Cluster, DL Total PRB Usage, clause 5.1.1.2.1, TS 28.552 [8] |
| Average RRC Connection Measurements | For every Cell in the selected Cluster, Mean number of RRC Connections, clause 5.1.1.4.1, TS 28.552 [8] |
| Packet Drop Measurements | For every Cell in the selected Cluster, DL Packet Drop Rate in gNB-DU, clause 5.1.3.2.2, TS 28.552 [8]  For every Cell in the selected Cluster, DL PDCP SDU Drop rate in gNB-CU-UP, clause 5.1.3.2.1, TS 28.552 [8] |
| Delay Measurements | For every Cell in the selected Cluster, Average delay DL in CU-UP, clause 5.1.3.3.1, TS 28.552 [8]  For every Cell in the selected Cluster, Average delay on F1-U, clause 5.1.3.3.2, TS 28.552 [8] |
| MDT data | UE measurements related to RSRP, RSRQ, SINR as specified in TS 37.320 [12] |
| Service experience data analytics | Service experience statistics and service in use predictions provided by NWDAF per UE (source/target cell) |
| NR Cell ID | Unique NR Cell ID For every Cell in the selected Cluster |
| Timestamp | Timestamp for each Performance Measurement reported |

##### 6.5.3.3.3 Analytics report

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

|  |  |  |
| --- | --- | --- |
| **Analytics Report of gNB resource consumption** | **Attribute Name** | **Description** |
| Predicted Virtual Resource consumption | Predicted Compute: This describes the average predicted compute resource consumption.  Predicted Memory: This describes the average predicted memory consumption.  Predicted Storage: This describes the average predicted storage consumption.  Timestamp: Time for which the prediction is made. |
| Predicted radio resource | The physical radio resource predicted utilization of the target gNB. |
| Prediction Horizon | Prediction is performed for a time window in advance. |
| isMLBNeeded | Indication on whether the target gNB needs to activate the MLB operation. |
| List of Predicted Congested Cells | List of cells or list of cells in the selected cluster, for which it is predicted that those will be congested at the Prediction Horizon. |
| List of possible target cells | Indication on whether the target gNB is suitable to be selected as the target gNB for the MLB based handover.  List of cells in the selected cluster, that are suitable for MLB based handover. |
|  | List of Predicted Congestion Easing out Cells | List of cells or list of cells in the selected cluster, for which it is predicted that the congestion shall ease out at the Prediction Horizon. |
|  | Prediction Confidence | The Prediction Confidence shall define the measured prediction accuracy. |
|  | Service list | List of supported service classes or corresponding slice IDs at the List of possible target cells. |

#### 6.5.3.4 Evaluation

The solution described in clause 6.5.3.3.1 requires the analytics inputs as described in clause 6.5.3.3.2, wherein:

* Input PM data: Virtual Resource measurements, RAN utilization, Average RRC Connection, Packet Drop Measurements and Delay Measurements 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].
* NR Cell ID and timestap are available.

With these analytics inputs which are already defined or accessible, the analytics output as described in 6.5.3.3.3 can be derived.

Therefore, this solution is a feasible candidate for load balancing optimization.

**End of Change**