July 2023

TECHNIQUES TO ENHANCE USER PLANE FUNCTION (UPF) LOAD METRICS TO IMPROVE UPF SELECTION

Swapnil Barai

Suyog Subhash Belsare

Bhavin Desai

Ravi Shekhar

TECHNIQUES TO ENHANCE USER PLANE FUNCTION (UPF) LOAD METRICS TO IMPROVE UPF SELECTION

AUTHORS:
Swapnil Barai
Suyog Subhash Belsare
Bhavin Desai
Ravi Shekhar

ABSTRACT

Mobile network architectures, such as Fifth Generation Core (5GC) architectures are designed to serve a variety of use cases in which each of the use cases can demand very different network resource allocation and traffic treatment. The introduction of the Control and User Plane separation (CUPS) architecture in the Evolved Packet Core (EPC) mobile network architecture, requires the control plane (CP) to select an appropriate User Plane Function (UPF) to serve a service request from a user equipment (UE) with a desired Quality of Service (QoS). This proposal provides techniques through which optimum UPF usage can be determined by considering multiple load factors, which can aid in performing UPF selection for 5GC and CUPS/EPC mobile network architectures.

DETAILED DESCRIPTION

Currently, UPF load level is one of the factors along with other factors, such as location, etc. that are often considered for UPF selection by control plane (CP) to serve a UE session. Current Third Generation Partnership Project (3GPP) standards, such as 3GPP Technical Specification (TS) 29.244 define a load metric as a single value (called 'load-level') that a UPF is to send to the CP, which is to interpret the value for UPF selection operations.

There are many factors that a UPF may consider in calculating a load-level value and the CP is typically unaware of the factors considered by a UPF in calculating the value. Thus, the CP is not aware on the weightage of various factors that are used in calculating the load-level at the UPF.

Assuming the number of sessions will be equally distributed across all UPFs provided for a network, there is a possibility that some UPFs may end up having low data throughput sessions whereas others may have high data throughput sessions. Such discrepancies could lead to some UPFs being under-utilized in terms of data throughput and whereas others would be over utilized, which could lead to overall performance degradation for the network. Such discrepancies can further lead to incorrect UPF capacity projections that could trigger adding more UPFs in the network unnecessarily, which could increase operational, maintenance, and serviceability costs for a mobile network operator (MNO).

Additionally, per current 3GPP standards, the frequency of inclusion of load control information sent from a UPF to the CP is implementation specific and only the UPF can determine when to send load information to the CP; the CP does not have control regarding when it may receive such information. If, for example, in a mesh network where a single CP node is attached to multiple UPFs, each performing its own determination regarding the frequency of inclusion of load control information, it can become difficult for the CP to prevent UPF overload situations.

Thus, it would be desirable to determine proper load-metrics for UPFs that may consider additional factors, such as session-based factors, throughput factors, etc. that could be utilized by the CP to facilitate proper UPF selection for serving UE service (session) requests. Further, it would be desirable for the CP to have some input regarding the frequency at which load control information is received from UPFs for a deployment.

This proposal provides techniques through which optimum UPF usage can be determined by considering multiple load factors, which can aid in performing UPF selection for 5GC and CUPS/EPC mobile network architectures. In at least one instance, techniques herein may facilitate intelligent load-based UPF selection that allows for selecting a UPF that might be loaded based on bandwidth but not based on number of devices, thus, making the UPF available to be selected for devices that may be numerous but may have low throughput requirements, such as Internet of Things (IoT) devices.

Broadly, techniques of this proposal seek to enhance the current load-metric to include more key factors that can be utilized to determine a current UPF load. Such techniques may be helpful for the CP to select an appropriate UPF for serving a service request at a desired QoS.

In accordance with techniques of this proposal two new information elements (IEs) are introduced that can facilitate enhanced load communications between UPFs and CP functions. The two new IEs include a “Query UPF Load Metrics” IE and a “UPF Load Metrics” IE. Using these IEs, CP functions can be provided both control as well as visibility on different parameters that can impact load calculations at UPF. Through such control and visibility, CP functions can be enabled to perform intelligent load based UPF selection. Further, the CP functions can also control how often the load control information is received from UPFs (e.g., controlling the frequency at which load control information is provided by UPFs), which may help to prevent overload situations for any associated UPFs.

Figure 1, below, illustrates example details for the Query UPF Load Metrics IE as provided in accordance with this proposal; the IE Type can be set as [Type = 399].

	Bits							
Octets	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
1 to 2	Type = 399 (decimal)							
3 to 4	Length = n							
5	Spare					RES	SESS	B/W
	Tolerance value							
k to (n+4)	These octet(s) is/are present only if explicitly specified							

Figure 1: Query UPF Load Metrics IE

During operation, Query UPF Load Metric IE can be sent from the CP to a UPF as part of a Packet Forwarding Control Protocol (PFCP) Association Setup Request or a PFCP Association Setup Response, based on whether the association is initiated by CP or UPF.

The Query UPF Load Metric IE will contain multiple flags with each flag corresponding to a load metric type (e.g., used session capacity of the UPF function, CPU/memory load in the UPF function, as used data throughput capacity of the UPF function) that is to be reported by the UPF to the CP via the UPF Load Metrics IE (discussed below) that can be included in the load control information (LCI) IE sent from the UPF to the CP. One or more of flags may be set in the Query UPF Load Metric IE. The Query UPF Load Metric IE can also include a tolerance value that can be used to control the frequency of inclusion of load control information sent from the UPF to the CP.

For example, consider various example details of the Query UPF Load Metric IE in which one or more flags can be encoded within Octet 5, as follows:

- Bit 1 – B/W: If this bit is set to "1" the UPF is to report to the CP a percentage indicating used bandwidth load information (via the UPF Load Metric IE to be included within the LCI IE);
- Bit 2 – SESS: If this bit is set to "1" the UPF is to report to the CP a percentage indicating used session capacity load information (via the UPF Load Metric IE to be included within the LCI IE);
- Bit 3 – RES: If this bit is set to "1" the UPF is to report to the CP (via the UPF Load Metric IE to be included within the LCI IE) a percentage indicating used resource capacity load information (e.g., central processing unit (CPU) utilization, port utilization, or memory utilization, whichever is higher); and
- Bit 4 to 8: Spare, for future use and set to 0.

The tolerance value will be coded within Octet 6. Whenever there is a change in any of the load metric values equal to the received tolerance value, the UPF is to send to the CP the LCI IE, in which the updated load metric values can be included within the new UPF Load Metrics IE.

As noted above, Query UPF Load Metrics IE can be sent to the UPF through a PFCP Association Setup Request or a PFCP Association Setup Response, based on whether the association is initiated by CP or UPF. Figures 2A and 2B, shown below, are example call flows illustrating example details through which the Query UPF Load Metrics IE can be communicated to a UPF.

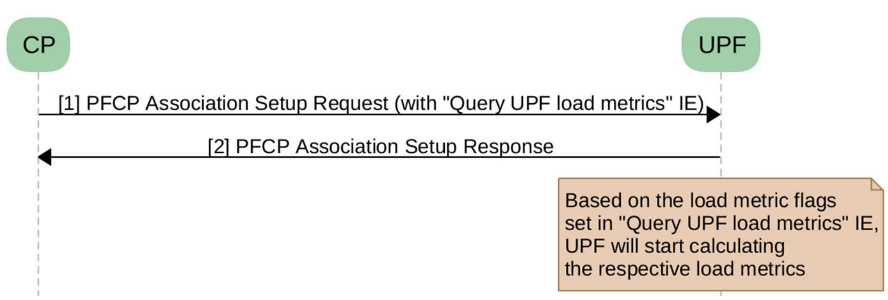

Figure 2A: PFCP Association Setup (Initiated by CP)

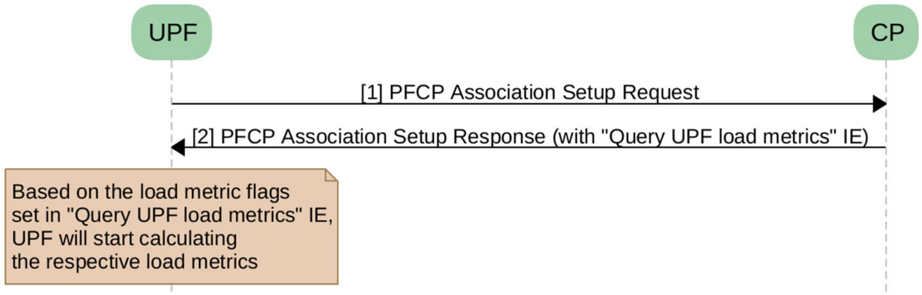

Figure 2B: PFCP Association Setup (Initiated by UPF)

Tables 1 and 2, below, illustrate various changes that may be provided for PFCP messaging in order to support the Query UPF Load Metrics IE as proposed herein.

Table 1: Change in Information Elements in PFCP Association Setup Request

Information elements	P	Condition / Comment	Appl.				IE Type
			Sxa	Sxb	Sxc	N4	
Query UPF load metrics	M	This IE will be present to indicate to UPF, the load metric types that the UPF shall include while communicating its load information to CP. Also, it will indicate the tolerance value to decide the frequency of inclusion of LCI IE to CP.	X	X	X	X	**Query UPF load metrics**

Table 2: Change in Information Elements in PFCP Association Setup Response

Information elements	P	Condition / Comment	Appl.				IE Type
			Sxa	Sxb	Sxc	N4	
Query UPF load metrics	M	This IE will be present to indicate to UPF, the load metric types that the UPF shall include while communicating its load information to CP. Also, it will indicate the tolerance value to decide the frequency of inclusion of LCI IE to CP.	X	X	X	X	**Query UPF load metrics**

When the Query UPF Load Metric IE is received by the UPF from the CP, the UPF is to send the CP an indication of the requested load metrics using the UPF Load Metrics IE [Type = 400], which can be included in the LCI IE. Thus, using the Query

UPF Load Metrics IE, the CP can have control over what load metric values will be received from the UPF.

Figure 3, below, illustrates example details for the Query UPF Load Metrics IE as provided in accordance with this proposal.

	Bits							
Octets	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
1 to 2	Type = 21 (decimal)							
3 to 4	Length = n							
5	Spare					RES	SESS	B/W
6 to 9	B/W Load							
m to (m+3)	Session Load							
p to (p+15)	Resource Load							
k to (n+4)	These octet(s) is/are present only if explicitly specified							

Figure 3: UPF Load Metrics IE

The UPF Load Metrics IE will contain a flag for each of the load metric types to be reported and a corresponding load metric value for each metric type. For example, consider various example details of the UPF Load Metric IE in which one or more flags can be encoded within Octet 5, as follows:

- Bit 1 – B/W: If this bit is set to "1", then the B/W Load field shall be present, otherwise the B/W Load field shall not be present;
- Bit 2 – SESS: If this bit is set to "1", then the Session Load field shall be present, otherwise the Session Load field shall not be present;
- Bit 3 – RES: If this bit is set to "1", then the CPU/Memory Load field shall be present , otherwise the CPU/Memory Load field shall not be present; and
- Bit 4 to 8: Spare, for future use and set to 0.

One or more of the above flags can be set to "1 and encoded with corresponding loading information to be reported to the CP, as follows:

- B/W Load – percent used bandwidth load information;
- Session Load – percent used session capacity load information; and

- Resource Load – percent used resource capacity load information (e.g., CPU, port utilization, or memory, whichever is higher).

The UPF Load Metric IE can be sent from the UPF to the CP through a variety of PFCP communications, including through PFCP Session Establishment communications, PFCP Session Modification Communications, PFCP Session Deletion communications, and/or PFCP Session Report communications. Figures 4A – 4D, show below, are example call flows illustrating example details through which the UPF Load Metrics IE can be sent from a UPF to the CP.

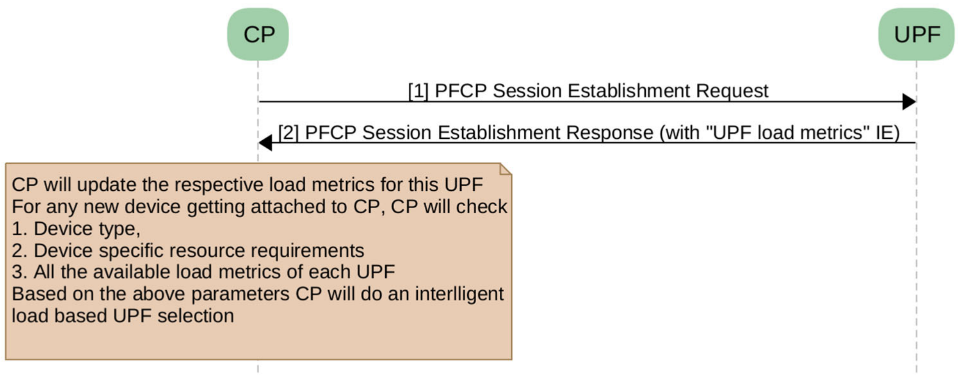

Figure 4A: PFCP Session Establishment for Reporting UPF Load Metrics

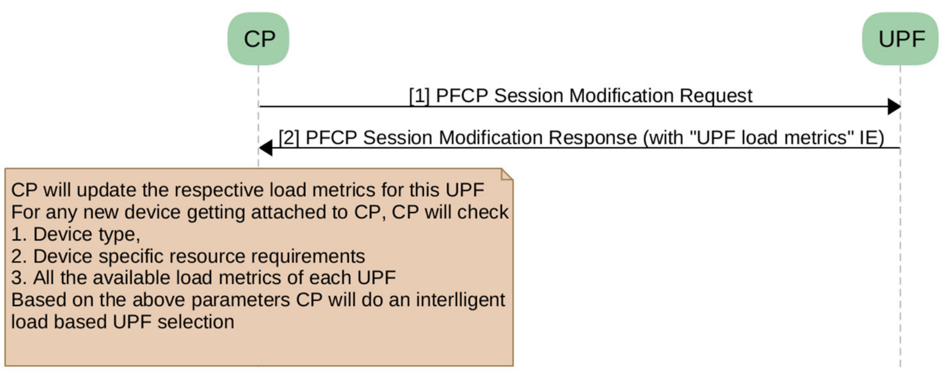

Figure 4B: PFCP Session Modification for Reporting UPF Load Metrics

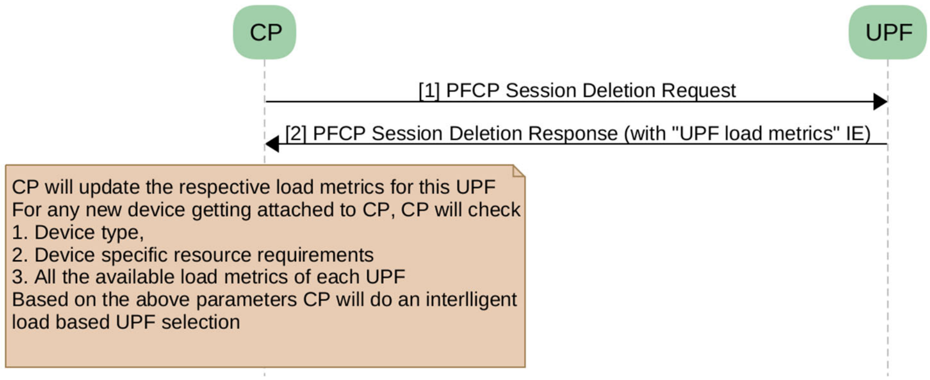

Figure 4C: PFCP Session Deletion for Reporting UPF Load Metrics

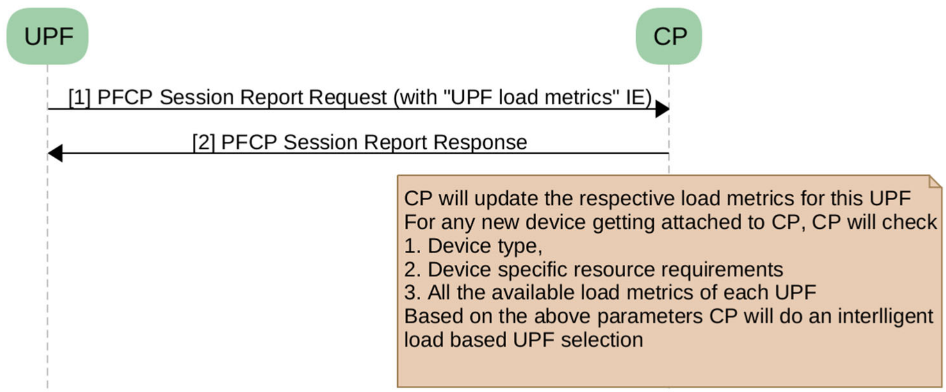

Figure 4D: PFCP Session Deletion for Reporting UPF Load Metrics

Table 3, below, illustrates various changes that may be provided for PFCP messaging in order to support the UPF Load Metrics IE as proposed herein.

Table 3: Change in Load Control Information IE within PFCP Messaging

Information elements	P	Condition / Comment	Appl.				IE Type
			Sxa	Sxb	Sxc	N4	
UPF Load Metrics	M	This IE will be present in the LCI IE to indicate the different UPF load metric values to CP	X	X	X	X	**UPF load metrics**

During operation, whenever a new session is received on the CP, the CP can perform intelligent UPF selection based on the device type, device specific resource requirements, and any load metrics received from UPFs managed by the CP.

Although certain load metric flags are discussed for the Query UPF Load Metric and UPF Load Metric IEs, it is to be understood that the IEs could be extended to include other load metric flags.

Further, as noted above, the tolerance value can be used by the CP to control the frequency for which the UPFs will include load metrics within load control information. When the tolerance value is received at a UPF, the UPF will start monitoring the change in the load metric values for each load metric type. When the change in value for any load metric type reaches the tolerance level, UPF is to communicate its load control information to the CP.

For example, consider an example scenario in which a tolerance value received at a given UPF is set to 10, discussed with reference to Table 4, shown below.

TABLE 4: Example UPF Load Information

Load Type	UPF Load (Time T)	UPF Load (Time T+20)	UPF Load (Time T+40)
Bandwidth Load	15%	18%	26%
Session Load	12%	16%	18%
Resource (CPU/Memory) Load	18%	22%	24%

With reference to Table 4, shown above, various sample load information is shown for the UPF with respect to different time intervals. For example, Time T illustrates initial load values. At time T+20, none of the values of the various load metric types have from the initial load to be greater than the received tolerance value of 10. Thus, no load control information is sent to the CP at time T+20. However, at the next time interval of T+40, the bandwidth load has changed to 26% from the initial value of 15%. In this example, since the change is more than tolerance value, the UPF would communicate its load control information to the CP in the next outgoing message to the CP.

Thus, techniques as proposed herein provide for enhancing the current single value load-level that is reported by a UPF to include multiple factors that can be used to determine UPF load. Such techniques can improve CP visibility and control of parameters that may be utilized by UPFs to perform load calculations, as well as report load information to the CP.

Various advantages over current UPF load reporting can be realized through the techniques of this proposal. For example, in many cases, the same UPF can serve many different network slices (serving different traffic needs) is a reality; thus, in such scenarios enhanced load metric values may be critical for optimal UPF operation/loading. Through the techniques of this proposal, the CP can perform intelligent load-based selection of UPFs based on device type of devices seeking session establish, along with any resource requirements that may be desired by the CP (e.g., for scenarios in which the CP may desire to load distribute traffic between UPF based on number of sessions, types of sessions, etc.).

Further, techniques as proposed herein can improve the CP's knowledge of key performance indicators (KPIs), which can improve UPF resource utilization and, thereby enhance end user experience. In edge deployments where the cost of adding a UPF is high, such techniques can also aid network planning processes such that UPF count can be accurately estimated for a network.

Still further, techniques as proposed herein can enable the CP to send tolerance values to the UPF for controlling the frequency of inclusion of load control information from the UPF. In some instances, the CP can decide to have a common tolerance value for all UPFs with which it is associated, which can be useful in detecting overload situations across all its associated UPFs before such situations occur.

Accordingly, techniques as proposed herein may help to reduce the overall operational, maintenance, and/or serviceability costs for MNOs, mobile virtual network operators (MVNOs), and/or enterprises deploying mobile network architectures that may cater to different kinds of devices (e.g., mobile phones, IoT devices, etc.). Additionally, techniques as provided herein may also enable better network planning in terms of the number of UPFs needed for a network, which is typically of utmost importance for edge UPF deployments.