**3GPP TSG-SA/WG2 Meeting #139-e *S2-2003594r01***

**Electronic meeting, 2020-06-01 – 2020-06-12**

**Source: Ericsson**

**Title: KI#3, New Sol: Network Information Provisioning using the IP path**

**Document for: Approval**

**Agenda Item: 8.3**

**Work Item / Release: FS\_enh\_EC / Rel-17**

***Abstract of the contribution:****This contribution proposes to use application measurement, ECN and L4S to improve latency for an application*

# 1 Introduction

This contribution proposes mechanisms to improve latency for an application

# 2 Discussion

KI#3 is about providing network information with low latency. The network is 5GS. The KI#3 description in clause 5.3.1 mainly talks about the latency between edge application server and application client in UE, which is the key element for low latency edge applications together with the bandwidth required by the application. The only way an application can adjust to higher latency in the network is to reduce its bandwidth so that queues in the network becomes shorter. If adjustment is not viable option, then the only other option is to terminate.

One of the key causes of latency in the network are queues. This fact is recognized by IETF draft draft-ietf-tsvwg-l4s-arch-06 [xx], and referenced document in the draft. RAN, UPF and routers all have queues that can build up during high load of the network. One way of rectifying this to have short queues, which may result in packet drop. A queue that has reached its maximum queue length for an interface means that the interface is congested. The fastest way to indicate congestion currently is to use Explicit Congestion Notification. By ECN, the application can immediately be notified of any congestion, and adjust its flow accordingly. For low latency traffic, we should assume this traffic to have a separate QoS flow. Admission control in RAN should make sure that in normal circumstances the low latency traffic should not be congested, but if congestion occurs, ECN is a fast way to indicate to the endpoints that the network is near its limits and that endpoint must reduce their traffic.

To improve ECN, the IETF is working on L4S (Low Latency, Low Loss and Scalable throughput). See IETF draft draft-ietf-tsvwg-l4s-arch-06.

From the draft:

"The L4S architecture primarily concerns incremental deployment. It defines mechanisms that allow both classes of congestion control to coexist in a shared network. These mechanisms aim to ensure that the latency and throughput performance using an L4S-compliant congestion controller is usually much better (and never worse) than the performance would have been using a 'Classic' congestion controller, and that competing flows continuing to use 'Classic' controllers are typically not impacted by the presence of L4S. These characteristics are important to encourage adoption of L4S congestion control algorithms and L4S compliant network elements. The L4S architecture consists of three components: network support to isolate L4S traffic from classic traffic and to provide appropriate congestion signalling to both types; protocol features that allow network elements to identify L4S traffic and allow for communication of congestion signalling; and host support for immediate congestion signalling with an appropriate congestion response that enables scalable performance."

# 3 Proposal

We propose to adopt 3 principles for the application to adjust to network latency situations:

1. The application does its own latency measurements and adjust accordingly
2. Use of ECN.
3. Use of L4S when available.

\*\*\*\*\*\*\*\*\*\*\*\*\* Start Changes \*\*\*\*\*\*\*\*\*\*\*\*\*

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

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

[2] 3GPP TS 23.501: "System Architecture for the 5G System".

[3] 3GPP TS 23.502: "Procedures for the 5G System".

[4] 3GPP TS 23.503: "Policy and Charging Control Framework for the 5G System".

[5] 3GPP TR 23.758: "Study on application architecture for enabling Edge Applications".

[6] 3GPP TS 22.261: "Service requirements for the 5G system".

[7] IETF RFC 7871: "Client Subnet in DNS Queries".

[x1] 3GPP TS 38.300: "NR; Overall description; Stage-2".

[x2] IETF RFC 8311: "Relaxing Restrictions on Explicit Congestion Notification (ECN) Experimentation".

[x3] IETF draft draft-ietf-tsvwg-l4s-arch-06: " Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service: Architecture ".

[x4] IETF draft draft-ietf-tsvwg-ecn-l4s-id-10: "Identifying Modified Explicit Congestion Notification (ECN) Semantics for Ultra-Low Queuing Delay (L4S)".

\*\*\*\*\*\*\*\*\*\*\* Next Change \*\*\*\*\*\*\*\*\*\*\*

## 6.0 Mapping of Solutions to Key Issues

Table 6.0-1: Mapping of Solutions to Key Issues

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Solutions | Key Issues | | | |
| 1 | 2 | 3 | 5 |
| #1: Provisioning URSP configuration to the UE to establish PDU Sessions for edge applications | X |  |  |  |
| #2: Local DNS based edge server address discovery | X |  |  |  |
| #3: DNS AF | X |  |  |  |
| #4: Providing the DNS authoritative server with IP addressing information about where the UE is located | X |  |  |  |
| #5: Server Discovery using DNS, IP Routing and URSP | X |  |  |  |
| #6: Discovery of EAS based on DNS | X |  |  |  |
| #7: SMF/I-SMF selection based on DNAI | X |  |  |  |
| #X: Network Information Provisioning using the IP path |  |  | X |  |

\*\*\*\*\*\*\*\*\*\*\* Next Change (all new) \*\*\*\*\*\*\*\*\*\*\*

## 6.X Solution #X: Network Information Provisioning using the IP path

### 6.x.1 Solution description

The solutions address Key Issue #3: Network Information Provisioning to Local Applications with low latency.

This solution uses ECN and later L4S to provide information about network status to the UE and AS.

The advantage of using ECN or L4S are that they are not 3GPP access specific but can be used by any access.

One of the key causes of latency in the network are queues. This fact is recognized by IETF draft draft-ietf-tsvwg-l4s-arch-06 [x3], and referenced document in the draft. RAN, UPF and routers all have queues that can build up during high load of the network. To get low latency for EC traffic, the queue length needs to be held short.

#### 6.X.1.1 Use of IP based congestion notification

Traffic that requires low latency will have their own QoS flow, e.g. QoS flow with 5QI=80. RAN will not admit more traffic than what it expects it can handle without getting the low latency traffic to be congested. Congestion in this sense will mean that RAN cannot sustain the required latency because of rare unforeseen events, like e.g. interference or fading. To provide fast information (feedback) with low latency, RAN will use ECN, which is a very fast indication to the endpoints that the network has trouble fulfilling the required characteristics (latency v.s. bandwidth). See TS 38.300 [x1] and RFC8311[x2].

By ECN, the application is immediately notified of any potential congestion (of this low latency traffic), and the application can adjust its flow accordingly.

The endpoints can make use of other back off algorithms than standard TCP, e.g. DCTCP. DCTCP and other scalable TCP-like methods have known compatibility issues with classic TCP when sharing the same queue, but separate queues can be employed in RAN for TCP and DCTCP if decided to be supported. A streaming application may use RTP. Here RTP receiver can take ECN into consideration when sending RTCP feedbacks to the source See RFC 8311 [x2].

NOTE: Endpoints that do not support ECN will have to be informed by packet drops about a congested situation. For TCP this will lead to that the dropped packet needs to be re-sent.

IETF is working on improving ECN the L4S as with will improve the characteristics of ECN.

L4S is best described in IETF draft draft-ietf-tsvwg-l4s-arch-06 [x3], and IETF draft draft-ietf-tsvwg-ecn-l4s-id-10 [x4].

L4S utilizes a separate queue for L4S traffic to support co-existence with classic ECN.

L4S Uses ECN bits in the IP header, but the setting of the EC (Explicit Congestion) code points does not mean congestion as per classical ECN. It rather means queues are getting longer, the longer a queue the more frequent notifications in the EC codepoint allowing endpoints to adjust their BW based on the frequency. This will result in queues for L4S traffic not to build up. See IETF draft draft-ietf-tsvwg-ecn-l4s-id-10 [x4].

### 6.X.2 Procedures

#### 6.X.2.1 Example, use of ECN with TCP



**Figure 6.X.2.1-1: Use of ECN with TCP**

0. PDU session is established to low latency Edge Computing DNN. A QoS flow for ECN traffic is established.

1. An ECN capable TCP connection is established by UE and EAS indicating ECN support to each other.

2. Application traffic over TCP.

3. RAN in 5GS detect incipient congestion on the e low latency QoS flows and starts to indicate congestion on these QoS flows (how and when RAN detect incipient congestion RAN implementation specific).

4. The EAS sends an IP packet towards the UE. The packet will travel trough5GS.

5. When IP packet is received in RAN, RAN congestion (CE) marks the packet.

6. The TCP stack in UE feedbacks the indicated congestion to the EAS.

7. EAS reduces its traffic towards the UE.

### 6.X.3 Impacts on Existing Nodes and Functionality

UE, EAS and NG-RAN need to support ECN or L4S. Applications may need to be able to use the feedback provided by IP protocol layer.

NOTE: This solution doesn't study how to support ECN and L4S in 5GS.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* End Changes \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*