The needed existence of a PEP in an encrypted world

Marcus Ihlar, Salvatore Loreto, Robert Skog

Ericsson

Abstract

Characteristics of cellular networks can have detrimental effects on TCP performance. These effects can be mitigated substantially by the use of congestion control mechanisms better tuned for cellular networks. Such congestion control is typically implemented in Performance Enhancing Proxies.

New transport protocols are currently implemented on top of UDP and a layer of encryption. We argue that PEPs will be needed for new transport protocols as well as TCP. Important future work will be how to maintain the benefits of end-to-end encryption while allowing the usage of PEPs.

1 Introduction

Wikipedia defines PEP as a **Performance-Enhancing Proxy** designed to improve the end-to-end performance of some communications protocol [1]. A well know communications protocol is TCP.
RFC3135 talks about Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations [2].

Many Mobile Operators have implemented TCP PEP to handle the transport characteristics in a Radio Network.
One example of such a characteristic is that a dropped TCP packet is not always a sign of congestion in the radio network, rather a signal that the buffers are full. Therefore, many TCP PEP, located in Mobile Network, implements a Congestion Control Algorithm adapted to Radio Network.

Ideas of new Transport Protocol are based on the usage of UDP as base with the addition of new and needed transport protocols functionality on top.
In the strive for end-users privacy many parts in the Transport Protocol headers are encrypted.

In this position paper we argue that upcoming requirement on new transport protocols should take in consideration the characteristics of cellular networks in conjunction with encryption requirement.

The importance of focusing on cellular network performance can be seen in the Ericsson Mobility Report which shows that global mobile data use (both download and upload) has increased from a few hundred Petabytes in 2010 to over three thousand Petabytes by the end of 2014 [3] .

2 Cellular network characteristics

The main source of latency in cellular networks is the power conservation characteristics of devices. A device radio operating at maximum capacity will deplete batteries fast. Therefore, devices try to minimize the time where the radio is active. This is formalized in Radio Resource Management (RRM) schemes where different radio states are defined. The cost of RRM is increased latency when switching from an idle to an active state. Such state transitions also lead to more variable latency characteristics than in fixed networks. The power conserving schemes differ somewhat between 3G and 4G networks, but the basic principle is the same. The noticeable difference is that the time it takes to switch from an idle to an active state is lower in 4G networks, roughly 400ms compared to above second times in 3G.
Furthermore, radio networks base their transmission decisions on timely and accurate feedback about channel conditions, buffer status, interference and other users activity. This is possible through a very fast and dynamic scheduling mechanism, always trying to optimize resource utilization, throughput and user resource fairness. The transmission over the air interface is therefore rather bursty. An important prerequisite for optimal RAN utilization is that sufficient data is stored in the radio layer buffers. If the buffer runs empty, or not enough data is available, the radio channels may be underutilized.
Yet another source of latency in radio networks is the local retransmission mechanisms used in radio links. These mechanisms reduce the amount of Transport Protocol retransmissions at the cost of highly variable latency.
When combining the Transport Protocol and the radio network control mechanisms, there is a risk for sub-optimizations due to the slow increase and fast and abrupt decrease of, for example, the TCP send rate. The consequence could be an unnecessary low throughput, followed by slower end-user response and load times. It is also possible that TCP occasionally transmits data too aggressively.
In such a case, data will be discarded in the buffers followed by retransmissions and a decreased TCP send rate. Latency characteristics in wired networks are usually stable. However, in cellular networks this is not the case; transitions between idle and active states induce high latency variability. A typical TCP connection retains its RTT measurements when switching from an idle state, this can lead to misconfigured retransmission timeout (RTO) values and unnecessary performance degradation as a result. Seamless handover between radio access types also pose problems for TCP since it can lead to radically different latency and throughput profiles, mid-connection.

The problems described above are expected to be mitigated by advances in wireless access technology. 5G will put new and more rigid requirement on latency and user experienced data rate (bit/s in the application layer).

Any new Transport Protocol based on UDP, that implements a Congestion Control Algorithms (CCA) will be subject to above cellular network characteristics.

3 New Transport Protocols

Cellular networks exhibit rapidly-varying conditions, both in terms of latency and packet loss. Such variability causes problems for New Transport Protocol. These types of problems can be mitigated by improved access technology; LTE already performs better than 3G and 5G will likely mitigate the problems further.
However, large scale deployment of new infrastructure takes a lot of time and is a very costly affair. A complementary approach is to make the transport layer more robust to cellular network conditions. This can be done in several non-exclusive ways such as reducing the amount of round trips, mitigating the effect of spurious retransmissions and deploying delay-based Congestion Control Algorithms (CCA).
Delay-based CCAs can be tuned specifically for different access types, accounting for the specific characteristics of that network.

If transport protocols headers are encrypted, then an access network PEP cannot use any access specific CCA.
One can argue that the solution is to have a solution for end-to-end negotiation of which CCA to use; fixed network CCA or mobile 4G CCA or mobile 5G CCA, etc.

However, besides being closer to the device from RTT point of view, the access network has more information that can be used by the transport protocol; cell load, user subscription type, radio technology, etc. These types of information are often limited to usage inside the access network and will not be shared outside the access network domain, either because of privacy reasons or fast variation of data.
In conjunction, an Operators Radio Network is a shared resource that shall be efficient and fair used by many types of devices, from high bandwidth demanding 4K media devices to low bandwidth demanding, but many, IoT devices.

4 Position

From a Transport Protocol of view, cellular network characteristics are demanding and any new Transport Protocol should mitigate this by allowing for efficient usage of an access network PEP, this in balance with end-user privacy.

5 References

[1] http://en.wikipedia.org/wiki/Performance-enhancing_proxy

[2] http://www.rfc-editor.org/rfc/rfc3135.txt

[3] Ericsson AB. Ericsson mobility report November 2014. http://www.ericsson.com/res/docs/2014/ericsson-mobility-report-november-2014.pdf, November 2014.