Abstract—Mobile data traffic has exploded over the last few years, especially with the introduction of smart phones in conjunction with increased capacity of mobile networks (HSDPA, LTE). It is commonly admitted that this continuous growth of traffic will induce congestion on the radio segment, usually considered as the bottleneck of the mobile chain. QoS mechanisms adapted to the radio segment are then potentially required when differentiated QoS is targeted. An associated QoS architecture has been defined by the 3GPP (e.g. use of multiple radio bearers per terminal, Guaranteed Bit Rate bearers). It is however inherited from a circuit-based model, which was suitable when voice services were predominant, but which is not so adapted to new packet oriented data services and to open multitask devices. In this paper, after a short summary of the weaknesses of the existing circuit-based QoS model, we introduce the concept of “IP aware radio scheduling”, an IP centric approach which aims at adapting some of the concepts of DiffServ to a wireless interface. The flexibility of this “all IP paradigm” and its relevance to mobile networks are highlighted. Initial simulations results in an LTE environment are then reported, before concluding on the aspects still to be investigated and the next steps.

Keywords—radio congestion; IP aware; mobile Internet, scheduling; DiffServ; QoS

I. INTRODUCTION

New devices and increased radio capacities have dramatically altered the mobile ecosystem, which should now accommodate to rapidly changing customer behaviors. The ubiquitous Internet should be understood not only as a universal access to mobile resources, but also as an open system allowing to dispose seamlessly of Internet resources, whatever the device or location. These changes in usage patterns will undoubtedly lead to a huge traffic increase. Even if 4G systems will relax the current constraints related to radio capacity, the radio segment will probably remain a potential bottleneck at least in some areas – installing new antennas is not always possible.

In order to face this mobile data traffic increase while limiting capacity upgrades, interest has been growing for applying QoS mechanisms on the radio segment. These mechanisms enable the operator to give a higher priority to either sensitive real-time services or to “premium users” who might pay higher fee for an assured service. Indeed, QoS management ensures optimization of the radio resources and also differentiated Quality of Service rather than “Best Effort for all” capacities.

The 3GPP has anticipated the need for differentiated QoS with the R8 release. However, the associated QoS model took into account the mobile ecosystem of that time (2008). Since then, a vast number of new devices, usages and applications have come to market. LTE dongles, MIFI, and tethering applications pave the way towards mobile Internet access on open devices such as PC. In this context, multitasking became common when the prevailing usage/devices of years 2000s were single tasking smart phones (limited screen size, battery life optimization). Offering differentiated QoS to these new multitask devices via the 3GPP circuit-based model may become a challenge as further analyzed in section IV of this paper.

Moreover, the increasing robustness of audiovisual applications (e.g. HTTP adaptive streaming) has opened the door to OTT (Over The Top) services, at the expense of Telco oriented ones based on IMS-like architecture. In a context of fierce global competition, the world of mobile data communications seems to move towards the open “best effort” Internet model, generally associated with a significant loss of value for mobile access providers.

In this context, it makes sense to investigate whether well-known IP QoS management functions could be adapted to fit the wireless segment. Indeed, these IP mechanisms had proven to be flexible, cheap, scalable, easy-to-configure and well adapted to open ecosystems. A new “IP centric” QoS model based on these mechanisms is then proposed in section V. Its behavior has been simulated and its performances have been evaluated in section VI.

II. 3GPP QoS MANAGEMENT

The QoS management scheme promoted by the 3GPP is clearly circuit oriented: its architecture relies on the use of several bearers per terminal (evolved User Equipment, eUE). Each bearer is dedicated to the transport of a particular QoS level. Bearers are operated in connected mode, that is, established, modified or disconnected via mobile control plane signaling protocols. Incoming traffic is oriented towards one of these bearers at bearer endpoints, in the P-GW for downstream flows and in the eUE for upstream traffic. Each bearer is
further supported by a GTP tunnel initiated by the P-GW. The GTP tunnel is in turn supported by the IP transport network, through encapsulation in UDP/IP packets.

The associated layering model (user plane) is depicted on Figure 1. Note that two IP layers are typically in presence, the IP end-to-end layer (IP E2E, in yellow) which is the layer 3 support of the customer applications and the IP transport network layer (IP TNL, in orange) which supports GTP tunnels (bearers) between P-GW and eNB.

In the 3GPP model, each EPS bearer is associated with a QoS level, defined through mobile QoS parameters (e.g. QCI, QoS Class Identifier). The QoS parameters of each bearer are conveyed by signalling protocols and are used as inputs to configure the scheduler which allocates radio resources to the different bearers. The scheduler is then the corner stone of QoS over the radio segment, and consequently, it dramatically impacts the end-to-end customer experience. A recall on radio schedulers is then briefly presented hereafter.

III. RADIO SCHEDULING ALGORITHMS

A. Radio Modulation and Coding Schemes

The quality of the radio channel of a given eUE is a key parameter determining his achievable throughput. Indeed, sophisticated radio Modulation and Coding Schemes (MCS) can be used when the eUE is in very good radio conditions, leading to a higher throughput per radio resource. On the contrary, in poor radio reception conditions, an eUE requires more robust MCS and experiences lower bit rates for the same amount of radio resources allocated. In other words, an eUE in poor radio conditions will require much more radio resources to reach the same throughput compared to an eUE in good radio conditions.

B. Basic radio scheduling algorithms

Every time slot, radio resources are dynamically allocated to the active eUEs of the cell according to a scheduling algorithm. Various examples of radio scheduling algorithms can be found in the literature.

Mobile network vendors generally implement adaptations of the Proportional Fair (PF) algorithm. This algorithm proposes a trade-off between cell throughput optimization and fairness (see [1], [2]). These schedulers prioritize users which offer the highest ratio of achievable instantaneous throughput normalized by its mean throughput. For example, a user in good radio conditions will be frequently scheduled at the beginning. But after a while, its mean throughput will increase, and the associated ratio will get lower than the ratio of another user with poorer channel quality which still has not transmitted yet. In the long term, fairness in terms of radio resources is therefore ensured.

C. QoS aware radio scheduling algorithms

As the initial PF algorithm does not consider any specific QoS requirements, various modifications have been proposed to give preferential treatment to specific flows (see [3]). For example, some schedulers propose to assign to certain eUEs a minimum guaranteed bit rate on the radio segment, almost independently of their radio conditions. This feature is called "Guaranteed Bit Rate" (GBR). They are generally vendor specific. In other schedulers, a weight is assigned to each eUE and is further used in the PF algorithm. In this case, the priority of eUEs is governed by the weights of favoured eUE against those of non-favoured eUEs.

D. Impact of QoS-aware radio scheduling

As a matter of fact, QoS-aware scheduling can have significant effects on the overall cell throughput capacity. Indeed, cell throughput reduction is a natural consequence of QoS-aware scheduling, since the introduction of QoS commitments is in contradiction with the trade-off between fairness and capacity of the original algorithm. In bad radio conditions, since less efficient coding schemes must be used, maintaining a minimum bit rate to a given eUE requires therefore allocating more radio resources to this eUE. Potentially, if no threshold is defined, all the radio resources can be allocated to a prioritized eUE facing very bad radio conditions.

While for fixed networks QoS-aware scheduling may effectively allow for managing the global bandwidth without reducing it, in mobile networks, it badly affects the overall cell capacity, especially when hard QoS commitments are sought (e.g. significant minimum guaranteed bit rate). The cell capacity reduction should then be compared to the extra revenues (direct or indirect) expected from the associated prioritized services. Voice-oriented services probably require such a hard commitment, but their impact on cell throughput should remain moderate, as the associated Guaranteed Bit Rates are very low. However, using such QoS aware scheduling mechanism for moderate to high Guaranteed Bit Rates is highly questionable.
IV. QoS POLICIES IN MOBILE NETWORKS OF TODAY

The circuit oriented QoS management scheme promoted by the 3GPP since R8 version was taking into account the usages and technical environment at that time (2008). It reveals now rather unsuitable to new multitask open devices which support simultaneously several running applications. In such configuration, the typical 3GPP QoS scheme would lead to open simultaneously several bearers per device, which may raise issues in terms of scalability (number of bearers), efficiency (signalling load to create new bearers) and performance (bearer setup delay).

Moreover, such telco oriented service models, based on connected architectures such as IMS, are always more attacked by OTT (Over The Top) business models. Indeed, in connected telco models, any new communication is preceded by signaling messages (e.g. SIP reservation protocol) thus requiring full interworking between service and network provider. On the contrary, sending data packets in Internet connectionless networks does not require any prior signaling, and Internet OTT players are particularly attached to this “ship in the night” approach.

Applicative detection through DPI (Deep Packet Inspection) has been also a popular policy among telcos in reaction to this lack of interworking. The idea was to detect the application through its applicative signature thanks to a DPI and to trigger a signaling procedure in order to create a new bearer (or to modify an existing one). However, such policy is not efficient because of the versatility of application signatures, added to the current growing trend to encryption.

It seems then reasonable to forecast that 3GPP architectures should evolve towards a low number of bearers per device, probably one (or two maximum when VoIP is supported) in an applicative connectionless model. Degraded QoE experience may happen when carrying multiple applications within the same bearer, even if the increasing robustness of applications to network impairments and resource fluctuations may limit the damage.

In this mono-bearer context, it appears then clearly that modifying bearer resource allocation through GBR or weighted Proportional Fair is not always the most relevant solution to improve customer experience. For example, when several applications are running on the same eUE (this situation is expected to be fairly common with tethering, MIFI, etc.), it may be much more efficient to properly schedule the user flows without modifying his global resources allocation. This implies intra-bearer QoS differentiation where sensitive flows of an eUE are favoured against less sensitive flows of this same eUE. Note that in this case, as the Proportional Fair allocation is not modified, the global cell capacity is left unchanged.

V. IP CENTRIC ARCHITECTURE - PACKET BASED QoS MANAGEMENT

As an alternative to the current 3GPP circuit oriented model, a packet oriented scheme is also possible, and starts being considered by eNB vendors, giving thus an IP flavor to mobile architectures. Indeed, IP networks natively operate packets, which are commonly conveyed in connectionless mode, without any setup procedure. IP QoS is then naturally managed on a packet by packet basis, as QoS parameters of each packet are contained in its header (DSCP/ToS field).

In the related IP centric model described in the present section, it is then proposed to distinguish eUE connectivity management, which is still operated in circuit mode through a bearer, from QoS management, which is performed on a packet basis. Several QoS levels are supported inside a unique bearer for each eUE; the IP packets are classified prior to the transmission over the radio segment according to their priority indicated in their DSCP/ToS field. For this purpose, an IP multiplexing stage is added before the radio scheduler.

These DSCP should be potentially taken into account by each node supporting the yellow IP E2E layer (i.e. P-GW, eNB, eUE) in order to schedule properly the packets of a given bearer/eUE. However, this requirement may be relaxed as QoS management is only effective on bottlenecks. As the wireless segment is generally considered as a bottleneck, an IP multiplexing stage should then be added before the radio scheduling in the eNB. The IP E2E layer required in the eNB reduces to a part of the User plane (queue management is included in standard forwarding functions) and does not imply control plane functions (e.g. no routing required).

It has to be noted that this scheme is more easily applicable to the downstream direction than to the upstream direction. However, upstream QoS management is generally not so critical, as congestion appears much more frequently on downstream traffic. Indeed, most services are today highly asymmetric (e.g. video) while they are supported by symmetrical capacities.

Different IP aware scheduler models are obviously possible. Some may consists in very simple and straight forward enhancements of the existing radio schedulers as to address the multi-application use case (user running multiple applications simultaneously), without changing the overall radio resources allocated to the eUE (intra-bearer arrangements).
Others may lead to deeper modifications of the radio scheduler in order to provide more radio resources to users operating high priority traffic (inter-bearer arrangements). In this case, the radio resources can possibly be allocated by the radio scheduler in the eNB taking into account the DSCP of the traffic mix of each customer, in addition to the parameters of the EPS bearer exchanged during signaling procedures. In this case, the radio scheduling algorithm is modified, with potential impacts on the cell throughput, as explained in the section III.

VI. SIMULATION RESULTS

A simulation tool has been elaborated in C++ in order to investigate the performance of an “IP aware radio scheduler” with an LTE wireless system.

The radio resource sharing algorithm chosen for the eNB is the Proportional Fair, widely implemented by vendors. An IP non-preemptive priority queuing system is added before this radio scheduler, without influencing it; it is composed of two finite priority queues: a high priority queue which has a strict priority on the Best Effort one. Both queues are operated according a non-preemptive service policy: whatever the queue it belongs too, once service on a packet has begun, it is continued without interruption until all the work associated with it has been completed (i.e. the packet has been transmitted).

The air interface model has been simplified in order to facilitate the interpretation of the results: we assume that each eUE transmits its Channel Quality Indicator (CQI) to the eNB each time slot (TTI, Transmission Time Interval) for each resource block and that the eNB allocates the Modulation and Coding Scheme according to this eUE’s feedback (that is, the only input data considered by the Proportional Fair algorithm is the CQI feedback). The radio channel is assumed not to suffer from any loss due to transmission.

The scenario simulated consists of two eUEs attached simultaneously to the eNB. For each eUE, two classes of packets arrive at the system in two independent Poisson streams. The first stream is marked in high priority while the second stream is marked in Best Effort, in order to emulate multi-tasking users (e.g. watching a video and downloading data simultaneously). Furthermore, we assume that the two terminals experience fairly constant radio conditions during the entire simulation: the first one evolves in good radio conditions (its CQIs vary uniformly between 10 and 15) and the second one evolves in bad radio conditions (its CQIs vary uniformly between 1 and 5). The overall CQI range is [1-15] in our simulation. The system parameters used in this simulation are summarized in Table I below. We focus on the congestion scenario since it is the most relevant (other scenarios without congestion have also been simulated).

Figure 3 below represents the evolution of the two queues’ state (High priority and Best Effort) against time for the eUE in bad radio conditions during the first 30 seconds of the simulation.

The Best Effort queue is almost always full leading to high packets loss rate, whereas the length of the high priority queue is significantly lower (maximum of 8 packets, except at the very beginning of the simulation, due the PF initialization period before entering in steady state), without any packet loss. In this simple “IP aware radio scheduling” model, the IP packets marked in high priority are effectively prioritized and sent over the radio segment with a much smaller latency and loss compared to the Best Effort packets.

<table>
<thead>
<tr>
<th>Table I. Parameters Considered During the Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Number of Physical Resource Blocks (PRB)</td>
</tr>
<tr>
<td>Number of terminals</td>
</tr>
<tr>
<td>Transmission Time Interval (TTI) duration</td>
</tr>
<tr>
<td>Data rate of prioritized traffic</td>
</tr>
<tr>
<td>Data rate of non-prioritized traffic</td>
</tr>
<tr>
<td>Packet size of prioritized traffic</td>
</tr>
<tr>
<td>Packet size of non-prioritized traffic</td>
</tr>
<tr>
<td>Queue size (prioritized and non-prioritized)</td>
</tr>
<tr>
<td>Simulation time</td>
</tr>
</tbody>
</table>

![Fig. 3. Filling levels of the two queues as a function of time for the eUE in bad radio conditions, in case of congestion](image)

Without such a mechanism, both traffics would be transmitted in Best Effort and would suffer from the same latency and loss, including sensitive flows. The introduction of an IP priority queuing stage before the radio scheduler decreases drastically jitter and latency of sensitive data in case of congestion, without any impact on the cell throughput.

VII. NEW TRENDS IN STANDARDIZATION BODIES

In November 2011, 3GPP initiated a study item on mobile user plane congestion (UPCON). The associated Technical Report [6] intends to study scenarios and use cases leading to user plane traffic congestion in the Radio Access Network (RAN), and to propose system enhancements for managing this congestion.
The initial proposal was to notify the congestion status of the RAN to the Core network, which may apply policies in order to reduce the traffic sent to this saturated RAN. Contrary to common queuing theory, UPCON excludes from its scope burst level congestions, thus restricting its studies to relatively long-lasting saturations.

Although IP aware QoS management gives an efficient answer to RAN congestion issues, it is not exactly in the UPCON initial scope as it relies only on the eNB local behaviour - no need for RAN congestion notification. Moreover, as in any IP node, IP aware QoS management is effective as soon as the queuing system is not empty, which covers the case of short duration congestions. However, it may reasonably be expected that this scheme could be discussed within the UPCON study group and possibly integrated in the associated Technical Report as a possible solution. Some initial proposals close to this IP aware scheme have been made in UPCON recently [7] as a solution to RAN congestion.

VIII. CONCLUSIONS AND FURTHER STUDIES

As a matter of fact, mobile operators have made very little use of the 3GPP circuit based QoS model, and the current trend is a “Best Effort for all” model for data traffics. This low revenue model - based on a unique bearer per customer – would probably lead to poor customer experience when carrying multiple applications within the same bearer.

The IP aware QoS management proposed in this paper allows a smooth QoS differentiation within such a mono-bearer scheme. This solution is less efficient than the initial 3GPP proposal in terms of customer experience, but much easier to deploy and operate at a marginal cost. Above all, it is perfectly in line with usual Internet paradigms, based on connectionless packet oriented networks.

In fact, the 3GPP circuit based QoS model is very similar to access architectures proposed in the late 90s for residential fixed services on ADSL. As a matter of fact, in less than a decade, Internet QoS paradigms (packet oriented QoS / DiffServ) have gradually replaced circuit based schemes inherited from circuit based networks (synchronous or ATM networks) in the vast majority of fixed networks, including those supporting voice. As fixed/mobile convergence is more and more perceivable in terms of usages and environment, a similar evolution can be reasonably expected in mobile networks, leading then to a graceful network functional convergence.

Next steps envisaged by the authors for introducing IP centric QoS schemes in mobile networks are to develop additional simulations with more complex scenarios (higher number of users, impact of TCP behavior, etc.). A first contribution is already submitted within the Study Group 12 of ITU-T [8]. It is also planned to submit IP centric models to other standardization bodies in order to address future releases of 3GPP and WIFI standards. Complementary work is also foreseen within IP standard bodies such as IETF and ITU-T in order to operate a smooth convergence between mobile and fixed ecosystems.

IX. REFERENCES