Centralized Bandwidth Management in Multi-Radio Access Networks

Balázs Héder
Nokia Siemens Networks
Budapest, Hungary
balazs.heder@nsn.com

Péter Szilágyi
Nokia Siemens Networks
Budapest, Hungary
peter.1.szilagyi@nsn.com

Csaba Vulkán
Nokia Siemens Networks
Budapest, Hungary
csaba.vulkan@nsn.com

***Abstract*—Radio access technology evolution resulted in two alternative architectural solutions: Evolved HSPA (High Speed Packet Access) systems with centralized architecture and LTE (Long Term Evolution) systems with distributed, full packet based architecture. Both systems are capable of providing high data rates and low latency to the users. Due to factors such as the need to preserve existing investments and reduced operational costs, for the time being these systems will coexist by sharing a common transport infrastructure and by providing services over the same areas. Good user experience over these systems requires harmonized QoS (Quality of Service) architectures and fair resource sharing mechanisms even in case of transport congestion. Technological and architectural differences of HSPA and LTE systems result in fairness problems that are not handled well by existing mechanisms designed for homogeneous environments. This paper proposes a comprehensive solution which, as simulation results indicate, has superior performance and handles the fairness and QoS issues efficiently.**

***Keywords*-HSPA, LTE, CC, multi-RAN, QoS**

I. INTRODUCTION

Smart phones are able to provide true multimedia experience and access to the multitude of Internet based applications and services such as streaming multimedia, mobile mail, web browsing, instant messaging, micro blogging, etc., which dominantly use TCP (Transmission Control Protocol) as transport protocol. This generates continuously growing demand for increased radio access system capacity, high user data rates and reduced latency. In parallel with the penetration of smart devices, the radio access technology is evolving as well. There are two main tracks of this evolution defined by the 3GPP (3rd Generation Partnership Project): evolved HSPA and LTE. On the one hand, evolved HSPA improves the radio and transport capability of the WCDMA (Wideband Code Division Multiple Access) systems via additional functionalities mainly implemented at the Node B without changing the system architecture. On the other hand, LTE proposes a full packet based technology with new, flat architecture where the radio and the transport network layers are packet switched and radio protocols are terminated at the eNBs (evolved Node Bs). In LTE, the latency of packet transmission is low because there are no Radio Layer 2 RTXs (retransmissions) over the transport network as opposed to the WCDMA/HSPA. Existing radio access networks based on WCDMA/HSPA technology will not necessarily be replaced by LTE but will coexist with it in a heterogeneous environment, where in certain locations multiple radio access possibilities (WCDMA, HSPA, LTE, etc.) will be provided to the users. This coexistence increases the system capacity and diversity, preserves the existing investments and provides a fall-back possibility and redundancy. As the LTE transport network layer is already packet based and HSPA is being migrated over packet technology, the deployment of a common transport network to be shared by the coexisting radio access systems is an obvious choice that allows efficient management and resource usage. These heterogeneous systems are referred to as multi-RANs (Multi-Radio Access Networks) in this paper. Harmonized QoS over multi-RANs is an important enabler of proper user experience. Users should have the same experience regardless of their point of attachment, that is, they should be able to use their applications with acceptable quality both over HSPA and LTE. Harmonized QoS has two important enablers: consistent HSPA and LTE QoS parameters, and QoS enforcement mechanisms able to provide fair resource usage over the shared transport. The former means that HSPA and LTE UP (user plane) bearers providing the same service should have a set of compatible QoS parameters. The latter requires coherent mapping to transport services. Assuming packet transport with DiffServ (Differentiated Services) based QoS architecture, this can be achieved by marking packets of the same application/service with the same DSCP (DiffServ Code Point) regardless of the access technology (HSPA or LTE). While the definition of harmonized HSPA and LTE QoS parameters and mapping rules is a simple management task for operators, QoS enforcement also raises problems that are not of administrative nature. Transport congestion that might occur in packet based networks (especially on the capacity limited backhaul links such as microwave radio) is handled differently in legacy (HSPA) and flat (LTE) systems. This is due to the difference in architecture and to technological constraints, such as the operation of the Radio Layer 2 protocols in HSPA systems. The HSPA CC (congestion control) mechanism, introduced by 3GPP [1], has the additional scope to prevent RLC AM (Radio Link Control Acknowledged Mode) RTXs over the Iub interface

 ISBN: 978-1-61208-203-5

[2] as these can cause significant efficiency degradation. LTE has no such standardized solution; currently, it relies on the TCP CC mechanism, that, together with RED (Random Early Detection), is able to resolve congestion and enforce fairness among the connections. In LTE, this might be enough but not for HSPA as it is not able to prevent RLC AM RTXs [3].

When the transport is shared by the LTE and HSPA traffic, congestion may cause fairness problems as HSPA traffic is not TCP friendly, i.e, the TCP sources can achieve only a limited throughput when competing for transport resources with TCP unfriendly traffic [4].

The coexistence of GSM (Global System for Mobile Communications), WCDMA and LTE on a shared transport is mentioned in [5] but it does not deal with the fairness problems in multi-RAN. An idea to use TCP friendly rate control in HSDPA (High Speed Downlink Packet Access) is described in [6] but considering only a homogeneous environment. An alternative HSDPA CC algorithm based on PDCP (Packet Data Convergence Protocol) / RLC packet discard was presented in [7] that solves the fairness problem only in case of DL congestion. Also, the applicability of the solution is limited to TCP.

This paper discusses the problems of inter-system fairness over capacity limited transport networks shared by multi-RAN systems. A novel centralized CC and bandwidth management algorithm is proposed, capable of resolving congestion and enforcing the right level of QoS and fairness. TCP and UDP (User Datagram Protocol) based user traffic are handled in the same way, without compromising the QoS and fairness. The solution is flexible, i.e., it can be used both in homogeneous and heterogeneous systems. The actions of the CC are based on the actual status of the system, the available resources, the topology, the QoS and fairness policies.

The rest of the paper is organized as follows. Section II provides a detailed overview of the multi-RAN systems, defines the fairness criteria and QoS requirements and deals with the fairness problem in case of transport congestion. Section III describes the proposed centralized CC algorithm. Performance evaluation is given in Section IV and finally Section V concludes the paper.

II. System Overview

Multi-RAN systems are based on the cooperation of the HSPA and LTE network elements. HSPA and LTE specific architectural elements impose special fairness and QoS aspects whereas transport congestion requires a common CC.

A. The System Architecture of Multi-RAN Systems

The architecture of a multi-RAN system [8] (Fig. 1) consists of HSPA and LTE network elements connected by user and control plane interfaces. Access to the packet services is granted through the SAE-GW (System Architecture Evolution Gateway). The eNBs are connected directly to the SAE-GW via the S1-U interface. HSPA traffic can reach the CN (core network) through the Iub that connects the Node Bs to the RNC (Radio Network Controller). The RNC is connected to the SGSN (Serving GPRS Support Node) via the Iu-PS interface. The S1-C and S11 interfaces provide the LTE control plane connectivity. The MME (Mobility Management Entity) is responsible for the UE authentication, location tracking and subscription profile management within the LTE system. Inter-system control plane connectivity is available via the S3 interface, whereas the S4 interface provides mobility and control support between the SGSN and the SAE-GW. From the RNC's point of view, the SAE-GW takes the role of the GGSN (Gateway GPRS Support Node). The RNC is connected to the SAE-GW via the S12 interface when direct tunnel is established and indirectly via the Iu-PS and S4 interfaces when no direct tunnel is established. The S12 is based on the Gn-u interface between the SGSN and GGSN in the legacy architecture (not shown).

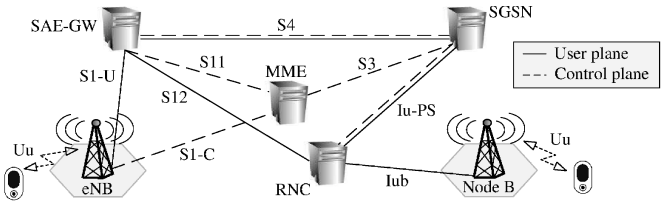

Figure 1. System architecture of a heterogeneous multi-RAN system

B. Harmonized QoS in Multi-RAN Systems

The HSPA and LTE QoS architectures are bearer centric, that is, the QoS parameters are defined and enforced on bearer level. HSPA bearers and LTE EPS (Evolved Packet System) bearers responsible for the UP connectivity between the CN and the UE are mapped to RABs (Radio Access Bearers) by the Radio Network Layer protocols terminated at the RNC (HSPA) and at the eNB (LTE), respectively. In both systems, the air interface packet scheduler has key role in the QoS enforcement, therefore at bearer/RAB setup, the related QoS parameters are signaled to the Node B/eNB. The Node B receives the RAB specific QoS parameters through the RNC: the SPI (scheduling priority indicator), the GBR (guaranteed bit rate) and the DT (discard timer) [9]. The SPI allows the definition of at most 16 distinct priorities. For each SPI, a GBR and DT value can be defined. HSPA flow and congestion control mechanisms support the packet scheduler in the QoS enforcement. The LTE systems allow the definition of 9 distinct QoS classes, referred to as QCI (Quality Class Identifier) classes, that is, upon setup, each EPS data bearer and the corresponding LTE RAB are mapped to a QCI [10]. For each QCI, and thus for each bearer, a GBR value can also be defined. At the transport network, HSPA bearers, RABs and EPS bearers are mapped to the transport QoS classes by DSCP

 ISBN: 978-1-61208-203-5

marking. For each SPI or QCI, a separate DSCP can be used. Note that the transport network QoS architecture should be configured so that it gives full support to the HSPA or LTE QoS. These parameters and mechanisms are sufficient for QoS enforcement in homogeneous radio access systems. Fairness is achieved if at a given Node B or eNB, bearers having the same SPI or QCI respectively receive the same level of service whereas bearers having different SPI or QCI receive service proportional to their QoS parameters. First, the packet scheduler should enforce the GBR of the bearers, whereas the remaining air interface resource should be distributed by considering the priority of the bearers. Throughout this paper, we assume that both the HSPA and LTE air interface packet schedulers implement the PF-RAD (Proportional Fair with Required Activity Detection) discipline [11], which is able to achieve optimal air interface usage and QoS differentiation. In order to facilitate the relative prioritization of the bearers, for each SPI/QCI an additional parameter, the scheduling weight (*wSPI* and *wQCI* respectively) is configured at each Node B/eNB. For the sake of simplicity and without loss of generality, we assume in this paper that the GBR of the bearers is zero, that is, QoS differentiation is enforced solely based on the *wSPI* and *wQCI* parameters. Fairness and QoS differentiation between the QoS classes i and j is achieved if the following expression is true: $\tau_i/\tau_j \approx w_i/w_j$, where τ_i and w_i denote the average measured throughput and the weight of QoS class i, i.e., the *wSPI* in case of the HSPA and the *wQCI* in case of the LTE. In multi-RAN systems, not only the intra- but the inter-system fairness must be achieved as well, i.e., user traffic belonging to the same application should receive the same relative service both through HSPA and LTE. One possibility is to give global meaning to the system specific QoS parameters, i.e., within the multi-RAN system, common QoS classes are defined with a set of well defined common data bearer and RAB level QoS parameters (GBR, weight, etc.). HSPA and LTE bearers are mapped to these classes and their own parameters are derived from these common QoS parameters. The inter-system fairness criteria is that $\tau_i/\tau_j \approx w_i/w_j$, $\forall i,j \in$ HSPA or LTE bearer, that is, the inter-system fairness is met if τ_i/w_i (the measured and weighted average throughput) is approximately the same for each QoS class in each radio access technology. In this setup, there is no need for dedicated bandwidth allocation to HSPA or LTE traffic over the transport network, thus the transport network is truly a shared resource, allowing the maximization of the multiplexing gain. That is, the resources can be dynamically shared by the HSPA and EPS bearers.

C. The Impact of Transport Congestion

In heterogeneous systems, LTE and HSPA share the same transport network as deploying separate transport for each RAN is not a realistic option due to cost, efficiency and manageability reasons. Despite the capabilities of the backhaul transport protocols (resilience, high data rate, low latency, QoS differentiation, etc.), transient congestion may occur due to the capacity limited links such as microwave radio or due to the overbooking of the high capacity aggregation links. During congestion, connections experience increased delay, packet drops and reduced throughput; additionally, it may deteriorate the intra- and inter-system fairness as well. Therefore, efficient CC mechanisms are needed. TCP, the dominant transport protocol used by the majority of data applications, has its own CC mechanism that reacts to congestion by reducing the rate of the connection and by retransmitting the data that is assumed to be lost. Together with RED, it is able to enforce fairness as well. In flat systems such as LTE, where packet drops due to transport congestion are transparent to the Radio Network Layer protocols, TCP's end-to-end CC mechanism is sufficient provided that its latency or the experienced RTT (Round Trip Time) is acceptable. In contrast, packet drops on the transport links connecting the Node Bs to the RNC trigger RLC AM RTX that has negative impact on the overall HSPA performance. The functionality of the HSPA systems has been extended by 3GPP [1] with means of detecting congestion without specifying the CC algorithm itself. The specified framework reuses the existing features of the HSPA systems and, despite the technical differences, provides similar solutions for UL (HSUPA, High Speed Uplink Packet Access) and DL (HSDPA). The HSPA CCE (CC Entity) is located at the Node B and it controls the rate of the connections either via capacity allocations sent to the RNC (HSDPA) or via grants issued to the UEs (HSUPA). Congestion detection is possibly based on the Delay Reference Time and Sequence Number IEs (Information Elements) included in the HS-DSCH (High Speed Downlink Shared Channel) and E-DCH (Enhanced Dedicated Channel) FP (Frame Protocol) data frame headers. The information provided by these IEs are used to detect delay build up (a common solution is to compare the estimated delay against thresholds) or packet drop (as frames are delivered in sequence, a missing sequence number indicates a drop).

In DL, congestion is detected at the Node B [3], [12], whereas UL congestion is detected at the RNC that informs the Node B about it through the E-DCH FP CI (Congestion Indication) control frame messages [13]. The CCE at the Node B reacts to the detected DL congestion by reducing the resource grants of the flows via Capacity Allocation messages sent to the RNC. In a similar way, upon the reception of the CI, the Node B reduces the UL air interface resource grants to be provided to the UEs.

Efficient HSPA CC algorithms are not only being able to resolve transport congestion but can also support the HSPA QoS architecture by considering the QoS parameters of the active bearers at CC decisions. The delay measurement is an important element of the HSPA CC: delay must be kept low so that random discards by RED are avoided and

 ISBN: 978-1-61208-203-5

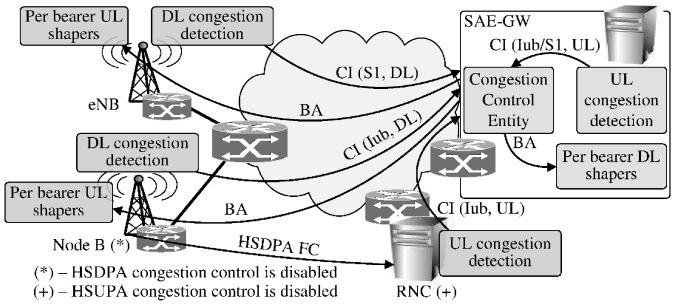

Figure 2. Concept of the centralized congestion control

(or) RLC timer expiration is prevented, i.e., the CC should keep the transport buffers under moderate load in order to prevent RLC AM RTXs. For further details on HSPA CC implementation, the readers are referred to [3].

As for the reasons discussed above and because the Node B and the RNC are topologically closer to each other than the UE and the content servers, the HSPA CC feedback loop is shorter than the end-to-end TCP CC loop. Therefore, in case the narrow link is shared by HSPA and LTE, the rate of the HSPA bearers is reduced first upon congestion. The unused bandwidth is taken by TCP connections over LTE, which continues until the total starvation of the HSPA bearers [7]. Disabling the HSPA CC in multi-RAN environments in order to prevent the self-starvation of HSPA bearers is not a good option either as at congestion, RLC AM RTXs over the Iub cause not only HSPA performance degradation but as the rate of the HSPA bearers is not reduced any more, now the LTE connections are going to starve [7]. Without CC, the Node B defines the resource grants allocated to the bearers so that the air interface resources are not wasted, which might even increase the transport congestion.

As explained above, HSPA CC is needed but the existing solutions are causing serious fairness problems in multi-RAN systems. This paper proposes an alternative solution that achieves fair operation by adapting the rate of both HSPA and EPS bearers sharing the congested link.

III. The Centralized Congestion Control

The proposed centralized CC and resource management solution is an efficient, flexible and versatile mechanism that is capable of resolving DL and UL congestion in multi-RAN (HSPA and LTE) and homogeneous (HSPA- or LTE-only) systems, being a feasible alternative of the existing HSPA CC mechanisms. It provides the enforcement of the HSPA/LTE QoS architectures (or any other bandwidth sharing or QoS differentiation policy) and it is able to guarantee the intra- and inter-system fairness.

The architecture of the solution is shown in Fig. 2. For the sake of simplicity, the description assumes that congestion can occur only on the last mile and aggregation links, i.e., it can affect only the traffic on the S1 and Iub interfaces. This is a reasonable assumption as the backbone network is not capacity limited due to the built in redundancy. In multi-RAN or HSPA-only systems, the HSPA CC mechanisms are replaced by the centralized CC, i.e., it takes over the bandwidth control functionalities, whereas the HSPA flow control mechanisms are only responsible to grant resources to the HSPA RABs according to the needs of the packet scheduler. The solution consists of the following elements: DL congestion detection entities located in the Node Bs and eNBs; UL congestion detection entities located in the RNC and in the SAE-GW; the centralized CCE, the topology database and DL per HSPA and EPS bearer shapers located at the SAE-GW; UL per HSPA and EPS bearer shapers located at the Node Bs and eNBs, respectively. One possible mechanism to detect congestion is to use the features of the ECN (Explicit Congestion Notification) [14] but the centralized CC is expected to work with any other congestion detection method as well. Congestion is detected when the ratio of the received CE (Congestion Experienced) marked IP packets exceeds a predefined detection threshold. The benefit of the ECN is that it is an already existing standardized functionality that provides explicit congestion indication by setting the relevant fields within the IP packet header [14]. The DL congestion detection entities residing in the Node Bs/eNBs communicate directly with the CCE via CI messages. The CCE identifies the source of the CI messages indicating DL congestion based on the ID of the sender coded into the message. For detecting UL congestion at the Iub interfaces, the CCE uses the services of the detection entity residing at the RNC, which sends a separate CI message per each Iub interface whenever it detects congestion. The ID of the Node B with congested Iub interface is encoded to this message. Finally, UL congestion on the S1 interfaces is detected by the detection entity located at the SAE-GW that sends CI to the CCE.

The CCE uses a time window based congestion control algorithm. During the window, the CIs are collected and the throughput of the active bearers are measured in both directions. At the end of each time window, provided that no CI was received, the CCE starts a new window. If a new CI was received, the CCE performs a CC action, consisting of the following four procedures: (a) congested link identification; (b) bandwidth recalculation for those interfaces that share the congested link; (c) sending the Bandwidth Allocation (BA) commands to the corresponding per bearer shapers; (d) execution of the BA commands. One CC action handles one congested link; if more congested links are identified by the CCE, a separate CC action is performed for each identified congested link. In this paper, the time interval in which the CC actions are performed is referred to as a CC period. During the CC period, no new CIs are accepted from the same source, i.e., the received CIs are ignored by the CCE.

Congested link identification. The CCE uses a topology database, which contains two entries for each link in the

 ISBN: 978-1-61208-203-5

network topology, one entry for each direction. Each entry contains the link ID, denoted by k, the link capacity C_k and a list of Node Bs/eNBs whose Iub or S1 traffic is routed via link k in the corresponding direction. For the sake of simplicity, it is assumed that each Node B/eNB has one S1 or Iub interface and that the links are symmetric, i.e., the link capacity is the same in both directions. The topology database is continuously updated by the CCE, i.e., entries are added or removed as the routes of the S1 and Iub change at the end of each window. To identify the congested link(s), the CCE ranks the links based on their likelihood of being congested. For that, the CCE uses a heuristic scoring method by which the following principles are considered: (a) link k is considered to be congested if CI has arrived from a Node B/eNB served by link k and $l_k > l^{(TH)}$, where $l_k = \tau_k / C_k$ is the load of link k, τ_k is the aggregated throughput of the active bearers routed through link k and $l^{(TH)}$ is a predefined threshold for the link load; (b) if for a given CI multiple links meet these conditions, the link at higher aggregation level is considered to be the congestion point, which provides a faster convergence to a congestion free state and better inter-node fairness. The aggregation level is represented by the number of served Node Bs/eNBs, denoted by $n^{(N)}$. If each CI received during the window resulted in the selection of a separate link, it does not matter which link is selected first because the others will also be selected later in the same CC period. The CCE calculates the score s_k of each link according to (1) and selects link k with the highest s_k, i.e., considers that link as being congested.

$$s_k = \begin{cases} n^{(N)} & \text{if } l_k > l^{(TH)} \text{ and CI is received} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Resource recalculation. After link k is selected, the CCE recalculates the shaping rates of the corresponding active bearers by considering the available resources, their QoS parameters and the fairness policies:

$$R_i = r \cdot \frac{w_i}{\sum_{j=1}^{n_k^{(b)}} w_j} \cdot C_k \quad \text{where} \quad r < l^{(TH)} < 1 \quad (2)$$

where R_i is the calculated shaping rate of bearer i, r is a multiplicative decrease factor, w_j is the weight of the bearer defined in Section II-B and $n_k^{(b)}$ denotes the number of bearers in the Node B/eNB set served by link k.

Sending the BA command. Based on the R_i shaping rates of the bearers calculated in the previous step, the bandwidth allocated to each affected Node B/eNB can be determined by summing up the rate of the bearers being served by the corresponding Node B/eNB. The bandwidth allocated to a Node B/eNB must not exceed the minimum of the link capacities along the route from the GW to the corresponding Node B/eNB. If this condition is not met, the minimum of link capacities must be allocated as the bandwidth to the Node B/eNB and the deficit must be reshared among the other Node Bs/eNBs. Here this method is referred to as deficit resharing. The allocated bandwidth is sent via the BA commands to the per bearer shapers.

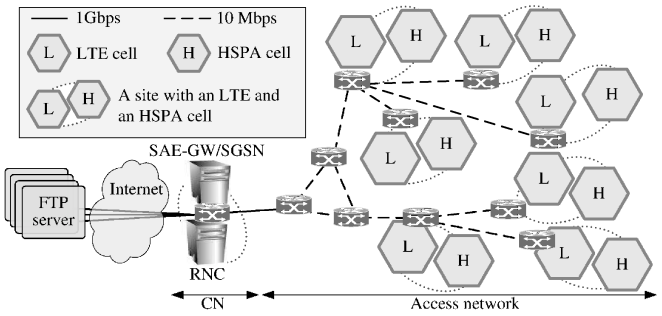

Figure 3. Simulation topology

Execution of the BA command. The shapers distribute the allocations among the active bearers (using a formula analogous to (2)) and initiate a prohibit timer. If the timer expires and no BA is received, the shapers start to increase the rate of the active bearers with an additive increase mechanism, clocked by the prohibit timer.

After BA commands are sent, the CIs of the Node Bs/eNBs served by the congested link are ignored. If there are remaining links with unhandled congestion, the CCE continues with new CC actions until all the congested links are handled, which is indicated by all link scores being zero. At that time, the CCE starts a new time window, accepting CIs again.

It is ensured by (1) that links with low load, which are not congested, are never selected by the CCE. It is also ensured that if the GW receives a CI, the CCE will perform a CC action, which will resolve the congestion by reshaping the corresponding bearers within a few CC periods. In addition, the deficit resharing mechanism ensures that the CC action does not induce further congestion on other links.

IV. Performance Evaluation

The performance of the centralized CC algorithm was analyzed with simulations. The simulation model implements in detail the UP protocols and interfaces (shown in Fig. 1), the Radio Layer 2 (PDCP/RLC/MAC) protocols, the transport network layer protocols (Iub: UDP/IP/Ethernet, S1, X2 and Iu-PS: GTP/UDP/IP/Ethernet, etc.) and the mobility procedures including the relevant control messages. Intra-system HOs (handovers) are modeled: hard HOs (HDSPA and LTE) and soft HOs (HSUPA). The details of the simulation models and the radio interface model can be found in [3].

The simulated logical topology (Fig. 3) consists of seven multi-RAN sites, each deployed both with an LTE eNB and a Node B. Each eNB and Node B is simulated with a one cell one sector configuration. The HSPA users are connected via HS-DSCH in DL and via E-DCH in UL to the RNC, whereas the LTE users are connected via DL-SCH (Downlink Shared Channel) in DL and via UL-SCH (Uplink Shared Channel)

 ISBN: 978-1-61208-203-5

in UL to the LTE eNBs. The SGSN, the SAE-GW, the MME and the RNC are considered to be co-sited. The FTP servers are connected to the SAE-GW/SGSN via the Internet. The CN consists of the RNC and the SAE-GW/SGSN/MME, interconnected through the core router. The access part of the network has a tree topology with 10 Mbit/s links. The access network is connected to the CN with a 1 Gbit/s link. The link capacities were selected in such a way that only the access links can be congested. The performance of the solution was analyzed by considering both DL (i.e., file downloads) and UL (i.e., file uploads) dominated traffic mix. Accordingly, at each simulation case, the users had either continuous file downloads or uploads to/from the FTP servers (located at the Internet) depending on the traffic mix. The TCP stack implemented the New Reno variant with 64 kB maximum advertised window size. At the transport layer, each bearer was mapped to the same PHB (Per-Hop Behavior). The minimum/maximum thresholds and the maximum drop probability parameters of the RED algorithm were set to 0.5, 1.0 and 0.1, respectively. At simulation start, the users were distributed evenly among the cells. In order to evaluate the performance of the solution under low, moderate and high load, the amount of users per cell was increased from 2 up to 6 in step of 1 that resulted in five distinct cases. The total amount of active HSPA and LTE users was equal in each case. The mobility model was random waypoint with velocity of 3 km/h. Users were executing intra-system HOs triggered according to the mobility procedures; therefore, the amount of users connected to a given Node B/eNB was changing depending on their actual location.

Three system alternatives were evaluated: (a) with no CC at all except the end-to-end TCP CC; (b) with HSPA CC only and (c) with centralized CC. When there is no CC in the system, HSPA users (both in DL and UL) receive much better service; their average throughput is at least 2.5 times of the throughput of the LTE users (Fig. 4). The reason is that the rate of FTP connections over LTE is reduced by the TCP CC whenever packet drops due to congestion are detected. In contrast, the RLC AM entity retransmits the dropped packets of the FTP connections over HSPA, which prevents TCP CC actions. The transport links are dominated by the HSPA connections that can achieve reasonable throughput whereas the RLC AM RTX rate is above 30% (Fig. 5). When there is only HSPA CC in the system, due to the shorter feedback loop, it detects congestion before the TCP CC and the rate of the HSDPA connections is reduced until their starvation (Fig. 6). This helps the LTE connections dominate the transport links. Note that in most of the cases, the HSUPA connections have lower throughput than the UL LTE connections but they are not starving. This is because the air interface capacity is narrower in UL than in DL, therefore the HSUPA and LTE air interface schedulers keep the rates of the UL flows at a lower level. Accordingly, the transport is less congested in UL than in DL.

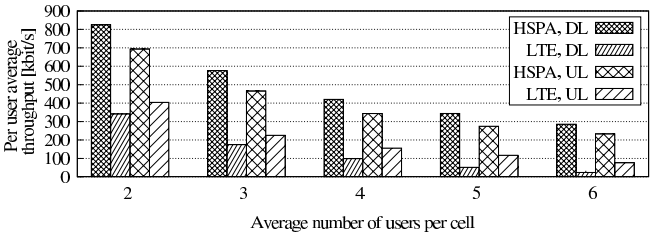

Figure 4. Per user average throughput if there is no CC in the system

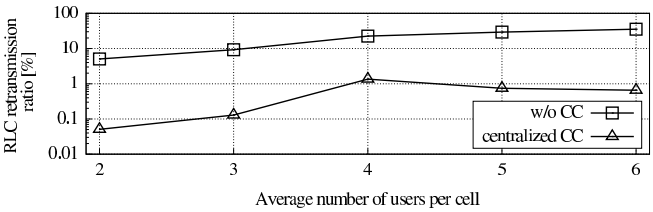

Figure 5. RLC RTX ratio over the Iub interface in DL. Results with only HSPA CC are omitted as HSPA connections are starving in that case.

The proposed centralized CC mechanism provides good level of service for both HSPA and LTE connections; their DL and UL average throughput is approximately the same (Fig. 7). The RLC AM RTX ratio is kept at reasonably low level (Fig. 5). If there is no CC or only HSPA CC used in the system, the intra-system fairness (evaluated with Jain's fairness index [15]) is poor in DL and a bit better in UL whereas the centralized CC is able to guarantee fair system operation both in DL and UL (Fig. 8).

The capability of harmonized QoS enforcement of the centralized CC was investigated in a scenario with two common QoS classes: high priority (HP) and low priority (LP). The SPI/QCI weights of the HSPA/LTE connections (bearers) were set to $wSPI_{HP} = wQCI_{HP} = w_{HP} = 2$ and to $wSPI_{LP} = wQCI_{LP} = w_{LP} = 1$. The meaning of the weights is defined in Section II-B. Three simulation cases were considered with 2 (1 HP, 1 LP), 4 (2 HP, 2 LP) and 6 (3 HP, 3 LP) users per cell according to low, moderate and high load (as before, the amount of HSPA and LTE users was equal). The results show that the centralized CC algorithm is able to provide harmonized QoS enforcement in case of DL traffic (Fig. 9): $\tau_{HP}/w_{HP} \approx \tau_{LP}/w_{LP}$ both in case of HSPA and LTE under each load (low, moderate and high), which is according to the expectations defined in Section II-B. Due to space limitations, the UL results, which are similar to DL ones, are not included.

V. Conclusion

This paper provides an overview of the aspects of QoS and fairness enforcement in multi-RAN systems sharing a common packet based transport network. At congestion, the users experience a fairness problem caused by technological and architectural differences of WCDMA/HSPA and LTE systems. WCDMA/HSPA networks with Radio

 ISBN: 978-1-61208-203-5

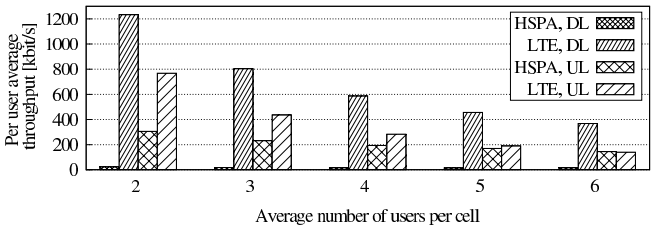

Figure 6. Per user average throughput with HSPA CC

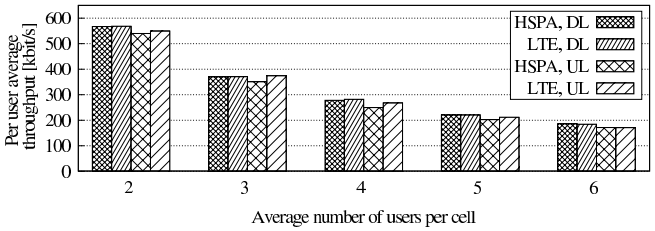

Figure 7. Per user average throughput if the centralized CC is used

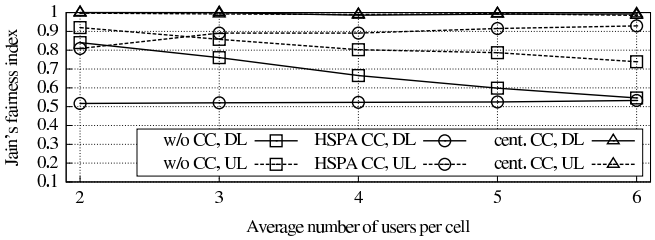

Figure 8. Jain's fairness index in DL and in UL. Index value close to 1 indicates high level of fairness.

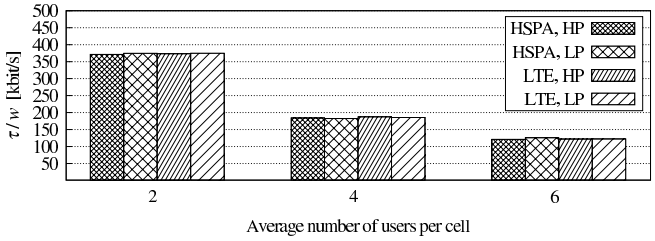

Figure 9. QoS differentiation capability of the centralized CC

Network Layer protocols such as RLC terminated at the RNC require special CC mechanisms in order to avoid performance degradation due to RLC AM RTXs over the Iub triggered by packet discards at transport congestion. The existing solutions work well in homogeneous HSPA systems but due to their intrinsic properties, they fail in multi-RAN environments. The centralized CC proposed by this paper provides a viable solution to the fairness problem combined with an efficient congestion handling and harmonized QoS differentiation capability, regardless of the traffic type. The solution is feasible both for DL and UL congestion control and can be applied in homogeneous HSPA or LTE networks as well. Simulation results confirmed that with centralized CC, the available bandwidth is shared in a fair way among the HSPA/LTE bearers regardless of the level of congestion. High fairness index, low RLC AM RTX rate and almost ideal QoS differentiation prove the superiority of the solution.

REFERENCES

[1] 3GPP, "Iub/Iur congestion control," TR 25.902 V7.1.0, 2007.

[2] ——, "Radio Link Control (RLC) protocol specification," TS 25.322 V10.1.0, 2011.

[3] L. Kőrössy and Cs. Vulkán, "QoS Aware HSDPA Congestion Control Algorithm," in *Proc. of IEEE International Conference on Wireless and Mobile Computing*, Avignon, France, Oct. 2008, pp. 404–409.

[4] B. Suter *et al.*, "Design considerations for supporting TCP with per-flow queueing," in *Proc. of INFOCOMM'98*, vol. 1, San Francisco, USA, Apr. 1998, pp. 299–306.

[5] "Optimizing global mobility through seamless coexistence and evolution of GSM, WCDMA and LTE," White Paper, Ericsson, Tech. Rep., Feb. 2009.

[6] K. D. Singh and D. Ros, "TCP-Friendly Rate Control over High-Speed Downlink Packet Access," in *Proc. of 12th IEEE Symposium on Computers and Communications*, Aveiro, Portugal, Jul. 2007, pp. 515–521.

[7] Cs. Vulkán and B. Héder, "Congestion Control in Evolved HSPA Systems," in *CD Proc. of IEEE 73rd Vehicular Technology Conference (VTC'11 Spring)*, Budapest, Hungary, May 2011, Paper No.: 1081710.

[8] 3GPP, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," TS 24.301 V11.0.0, 2011.

[9] K. I. Pedersen *et al.*, "Overview of QoS Options for HSDPA," *IEEE Commun. Mag.*, vol. 44, no. 7, pp. 100–105, 2006.

[10] H. Holma and A. Toskala, Eds., *LTE for UMTS: OFDMA and SC-FDMA Based Radio Access*. John Wiley & Sons, 2009.

[11] T. E. Kolding, "QoS-Aware Proportional Fair Packet Scheduling with Required Activity Detection," in *Proc. of IEEE 64th Vehicular Technology Conference (VTC'06 Fall)*, Montréal, Canada, Sep. 2006, pp. 1–5.

[12] S. Nádas *et al.*, "Providing Congestion Control in the Iub Transport Network for HSDPA," in *Proc. of GLOBECOMM'07*, Washington D.C., USA, Nov. 2007, pp. 5293–5297.

[13] 3GPP, "UTRAN Iub/Iur interface user plane protocol for DCH data streams," TS 25.247 V11.0.0, 2011.

[14] K. Ramakrishnan *et al.*, "The Addition of Explicit Congestion Notification (ECN) to IP," *IETF RFC 3168*, Sep. 2001.

[15] R. K. Jain *et al.*, "A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems," DEC-TR-301, Digital Equipment Corporation, Tech. Rep., September 1984.

 ISBN: 978-1-61208-203-5