Enabling RAN Moderation and Dynamic Traffic Steering in 5G

Athul Prasad[1], Fernando Sanchez Moya[2], Mårten Ericson[3], Roberto Fantini[4], Ömer Bulakci[5]
[1]Nokia Bell Labs, Finland; [2]Nokia Bell Labs, Poland; [3]Ericsson Research, Sweden; [4]Telecom Italia, Italy; [5]Huawei ERC, Germany
Email: {Athul.Prasad, Fernando.Sanchez_Moya}@Nokia.com; Marten.Ericson@Ericsson.com;
Roberto.Fantini@TelecomItalia.it; Oemer.Bulakci@Huawei.com

***Abstract*—The exponential increase in capacity and data rate demands, along with the diversification of use cases and verticals that are planning to use cellular radio access networks (RANs) to provide connectivity, has prompted the development of the fifth generation (5G) of radio access technology. Traffic steering, which aims at optimum mapping of the data flows to the appropriate RAN access points (APs), is considered to be one of the key enablers for supporting diverse set of requirements ranging from 1000 times higher capacity than 2010 figures to 99.999 % reliability. Furthermore, to constrain power consumption due to the ultra-densification of the network in 5G, enhanced mechanisms are required that can exploit multi-connectivity. In this paper, we build upon the envisioned connectivity provided through multiple radio links and provide schemes to enable dynamic traffic steering and energy-efficient RAN moderation in 5G. First, we present the various protocol options for tight integration of long-term evolution (LTE) and 5G networks. We then investigate how faster traffic steering over the multiple radio links can be enabled, which can both reduce the packet delivery time by increasing the capacity and reliability, and also reduce the power consumption of the network through efficient operation. Detailed performance evaluations done on the various presented mechanisms indicate the technology potential of the proposed enhancements.**

Keywords—5G; Energy Efficiency; METIS-II; RAN Design; RAN Moderation; Resource Management; Traffic Steering

I. INTRODUCTION

The fifth generation (5G) radio access technology, through the support for extreme mobile broadband and ultra-reliable communication, is expected to address the significant increase in capacity and data rate demands that network operators are expecting during the coming years. Due to the wide range of frequency bands used and the need to tailor the air interface parameters depending on the frequency band, the 5G landscape is expected to consist of multiple air interface variants (AIVs), which could include evolved legacy technologies, e.g., long-term evolution advanced (LTE-A) air interface as one component [1]. Since 5G needs to support a wide range of diverse use cases and requirements, such as extreme mobile broadband with 1000 times higher capacity, ultra-reliability of 99.999 % and low-latency of less than 1 ms over the air interface [2], it is expected that the network shall be optimized for the target use case and the associated requirements, as well.

One of the possible key enhancements required in 5G networks is a more dynamic mechanism for traffic steering, which could be complemented by a dynamic definition and enforcement of quality of service (QoS). In legacy networks, traffic steering was considered as a key enabler for load balancing and improving user throughput [3]. Various mobility-based traffic steering strategies for LTE-A heterogeneous networks were studied in [4], where each user is connected to the layer that can best serve it. The work done in [5] proposes traffic steering using self-tuning controllers in a densely deployed LTE femtocell network.

Due to the ultra-densification needs of the network in 5G [6], it is expected a significant impact on the power consumption of the network, which needs to be optimized. One approach for enabling short-term power savings is the reduction of always-on signals in the 5G system design [1]. This enables longer base station discontinuous transmission power-saving modes compared to what is possible in LTE. In addition, another complementary feature, which could increase the opportunities for RAN access points (APs) to attain power saving, is intelligent RAN moderation, especially in the context of multi-connectivity where the target service requirements are fulfilled by a subset of available radio links through joint coordination. Such enhancements utilizing fast traffic steering over multiple active radio links to achieve energy savings have received limited attention in the available literature.

Herein, we highlight enabling technologies for RAN moderation and dynamic traffic steering within the framework of agile resource management investigated by the 5G Public Private Partnership (5G PPP) METIS-II project [7]. To this end, we consider a next generation heterogeneous cellular network with connectivity provided through a diverse set of AIVs. Basic connectivity is assumed to be provided to the user equipment (UE) using the multi-connectivity paradigm [6], where the UE would have active links with multiple APs, including different AIVs. Multi-connectivity is considered essential in 5G networks, especially due to the low reliability of the higher frequency bands such as millimeter wave (mmW), which necessitates multiple active radio links to provide seamless connectivity. Such a network provides an opportunity to enable faster traffic steering between the various active air interface links for a user, thereby acting as a key enabler to satisfy the diverse set of 5G requirements. Such advanced traffic steering mechanisms would also enable the energy efficient operation of the networks, since it could optimize the actual active operation time of the APs in the network.

In this paper, first, various protocol options for tightly integrating LTE and 5G radio access networks are presented. The throughput gains from having the various options are also shown. Thereafter, the impact of fast traffic steering (based on channel quality feedback from the active air interface links and a dynamic QoS mechanism enforced in the RAN) on the delivery time of packets is also discussed, along with the performance evaluations that show the technology potential. Means for achieving energy efficiency through advanced traffic steering mechanisms, along with traditional interference coordination and mitigation techniques, are investigated as well. Performance evaluations conducted on network power consumption particularly under low-load conditions are presented.

The rest of the paper is structured as follows. Section II gives an overview of the system model used, and Section III discusses various protocol options through which LTE and 5G can be tightly integrated, along with some performance evaluation results. Section IV presents how the fast traffic steering mechanism can lead to significant reductions in packet transmission delays. Section V focuses on how the traffic steering mechanism can be used to achieve energy efficiency in 5G networks. Finally, Section VI concludes the paper and provides some future research directions.

II. SYSTEM MODEL

The overall system model considered is as shown in Fig. 1, with the possible functional decomposition linked to the architecture presented in [8]. Here, it is assumed that the QoS Class Identifiers (QCIs) [9] informed are sent from the core network (CN) to the access network – outer (AN-O) layer, which could be mapped to the business enablement layer in [8]. The AN-O layer does the traffic aggregation and steering towards the access network – inner (AN-I) layer based on real-time feedback from the radio APs. The AN-I layer is assumed to be mapped to the infrastructure resources layer in [8]. The AN-I layer is operated dynamically with optimal number of active nodes, based on the real-time traffic steering strategies, for energy efficiency. The traffic steering would enable the enforcement of key 5G requirements such as high data rates for extreme mobile broadband and high reliability for ultra-reliable communication.

In the considered system, the UEs move within the coverage area of the network, engaging in multi-connectivity with the strongest available links. The maximum number of links and signal quality for UEs to engage in multi-connectivity are configured by the network. The signal quality is assumed to be different for different AIVs, similar to the assumptions in [10]. Multi-connectivity is assumed to take place over the same or different frequency bands.

A holistic resource management framework should provide inherent support for dynamic QoS. The diverse requirements in the scope of 5G systems will require native support for dynamic QoS in order to ensure end-to-end (e2e) QoS delivery flexibility and to fulfill the QoS targets and attributes enforced by higher-layer protocol entities.

One potential paradigm change in 5G is the protocol stack layer on which traffic steering is performed, i.e., the assignment of services to AIVs. In existing systems, the assignment to Radio Access Technologies (RATs) (e.g., 3G and 4G) and cells takes place via hand-over between cells, i.e. radio resource control (RRC)-level mechanisms. Additionally, a device may be served in multi-connectivity within a radio technology (e.g. LTE Release 12 dual connectivity), where traffic is then further steered on packet data convergence protocol (PDCP) level to the different radio legs. In 5G, considering more stringent service requirements, more degrees of freedom and larger radio dynamics especially in higher frequency bands, it may be beneficial to perform traffic steering further down in the protocol stack, e.g. on radio link control (RLC) or medium access control (MAC) layer, and hence, on a faster time scale overcoming the semi-static and time-consuming bearer modification in legacy systems [1].

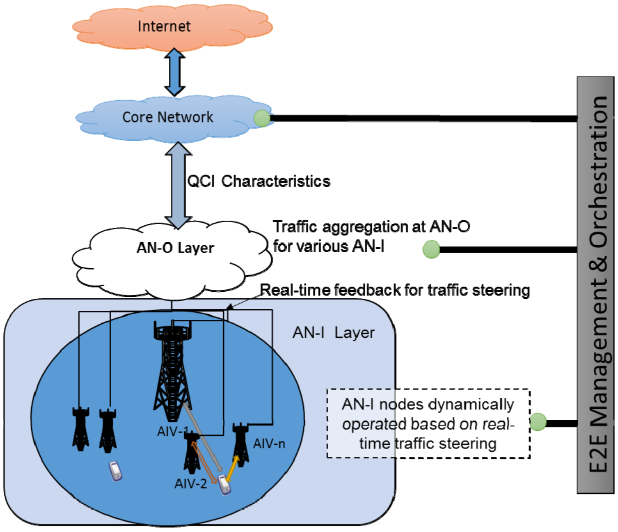

Fig. 1: The overall system model considered.

The traffic steering framework resides in the RAN and translates the AIV-agnostic QoS metrics to AIV-specific ones, based on the real-time feedback received from the AIVs. This feedback could, for instance, contain AIV-specific radio conditions and radio resource utilization values. Aggregated feedback of radio parameters (e.g., channel state information (CSI), load and resource usage) can be part of the signaling flow to higher layers. Additionally, the logical channel and user prioritization, and the multiplexing functions, can be made more generic and less radio dependent.

The requirements imposed by dynamic QoS support on MAC layer can be outlined as:

- The QoS targets can be configured/reconfigured with a higher granularity than in LTE, i.e.,, from bearer-centric to more granular application-centric QoS architecture (tuned to the end-user/application needs).
- The adaptation and configuration/reconfiguration of QoS parameters can be more frequent based on application/use case needs.
- The control signaling shall be lightweight (in LTE 30% of resources are taken by control signaling) and preferably in-band while still being sufficiently dynamic and granular.

- Aggregated radio feedback from different radio interfaces to higher layers will be required to adapt the parameters accordingly, striving for an overall optimization rather than separately enforcing individual targets.

III. TIGHT INTEGRATION OF LTE-5G WITH VARIOUS PROTOCOL OPTIONS

In order to realize the tight integration of LTE and the new 5G air interface, an architecture relying on common protocols (also called integration layers or AN-O) framework is considered. In [12], a common protocol framework for tight integration between LTE and 5G is proposed. The conclusion in [12] was that a common PDCP for the user plane and a common RRC for the control plane design were the most suitable alternatives. In contrast to PHY, MAC and RLC functions, the PDCP functions do not have strict constraints in terms of synchronicity with the lower layers. In other words, a separate design for PHY/RLC/MAC functionalities for LTE and the new 5G air interface would likely not impose any problems for a common PDCP layer. In addition to this, such integration would work in both co-located and non-collocated network deployment scenarios, making it more general and future proof. Therefore, using a common PDCP layer can be a suitable choice for LTE-5G tight integration. To investigate this, system-level simulations are performed for different tight integration cases studied: hard handover (HH, i.e., inter-RAT handover), fast switch (FS, fast user plane, UP, switch) and user plane aggregation (similar to dual connectivity, DC, in LTE).

A. Hard Handover

HH enables users with poor coverage to switch to another RAT with better coverage and users in LTE can switch to the novel 5G air interface to get higher throughput if there is coverage. The HH requires rather extensive RRC signaling and CN signaling as well as cell search and synchronization, which results in relatively long interruption delays. With the increased amount of 5G frequency bands and the usage of massive number of beams, it is expected that the cell search time will increase compared to the current situation. Another drawback with inter-RAT handover is the rather low reliability as the user can only be connected to a single RAT at each time. To model this in the simulations, a 300 ms service interruption delay is added when an HH is performed.

B. Fast Switch

The FS of the UP assumes that the control plane (CP) is using "DC" between LTE and 5G, while the UP is switched at PDCP level to either LTE or 5G. If the CP is connected to both the LTE node and the 5G node, no signaling is required when a UP switch is performed. Further on, the UP switch may be almost instantaneous as soon as the channel quality has changed (improved or deteriorated). Another benefit with CP in dual/multi connectivity is the increased reliability. A drawback is the user multiple flows of the CP and the increased overhead. The FS is modeled as a normal HH but with a 0 ms service interruption delay.

C. User Plane Aggregation/Dual Connectivity

This alternative assumes that both UP and CP are connected to LTE and 5G and the UP data is aggregated, similar to LTE DC but using different RATs. The benefits of this feature is increased throughput, pooling of resources and support for reliable seamless mobility. The UP aggregation may have limited benefits for the user throughput when the air interfaces provide different latencies and throughput but will still provide improved reliability compared to HH.

D. Simulation Results

The simulation model for the new 5G air interface in this paper is called NX and is based on the Ericsson 5G test-bed [14]. The deployment model is the 3GPP Case 1 with typical urban channel model. The LTE and 5G nodes are co-sited and the frequency bands investigated are 2 GHz for LTE and 15 GHz for NX. The bandwidth is 20 MHz per radio access. Note that all signaling is ideal, i.e. all RRC signaling is always received correctly. The difference between LTE and NX is a shorter transmission time interval (TTI) for NX of 0.2 ms TTI as well as fewer sub-bands. Fig. 2 shows a summary of the user throughput for the three different LTE-NX tight integration cases. The stand-alone NX is used for comparison. It uses 15 GHz frequency and a bandwidth of 40 MHz (in contrast to 20+20 MHz for the tight integration cases). The simulation parameters are similar to the one used in [14]. The bars in Fig. 2 show the relative 10%-ile (worst users) and 90%-ile (best users) user throughput at medium load.

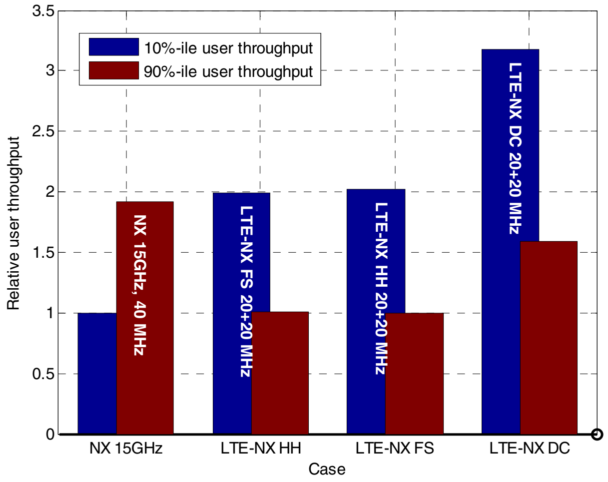

Fig 2: Summary of the LTE-NX tight integration user throughput at medium load.

The NX stand-alone case has the best 90%-ile user throughput, almost 90% better than HH and FS cases, and around 15-20% higher than the LTE-NX DC case. The main reason is the faster transmission control protocol (TCP) slow start due to lower TTI. However, for the worst users (10%-ile user throughput), the NX stand-alone case shows the worst performance. LTE-NX DC case shows more than 200% higher user throughput than the stand-alone NX case, and the FS and HH cases around 100% higher user throughput for the worst users.

IV. RAN Moderation and Dynamic Traffic Steering Mechanism with Multi-Connectivity

The main goal of RAN moderation and traffic steering mechanism is to serve the right application and related data flows for a UE using the most appropriate AIVs. One of the key limitations with the current LTE-A systems is the lack of dynamic QoS provisioning possible at the RAN. The current LTE bearer architecture is as shown in Fig. 3 [9], where the radio bearer QoS values are defined by the Packet data network-Gateway (P-GW) located at the edge of the CN. Here it is proposed to limit the QoS metrics definition at the AN-O layer based on the information from the CN (from the 5G P*-GW and related interfaces) and leave the enforcement of the QoS to the AN-I layers. This concept is as shown in Fig. 4, where the AN-I layer enforces the QoS metrics received from the CN over multiple APs.

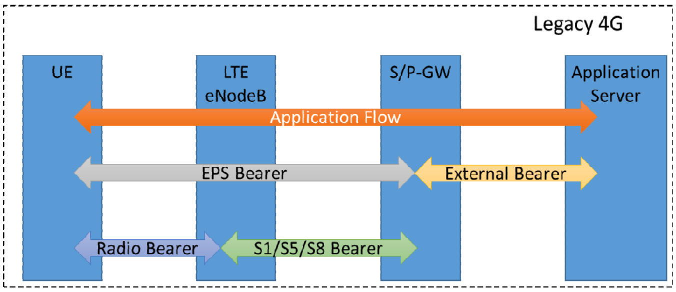

Fig. 3: LTE Bearer Architecture [9].

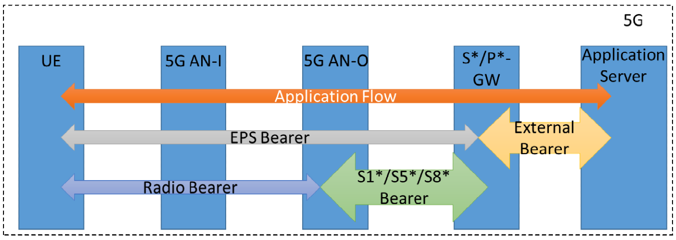

Fig. 4: Proposed 5G Service Flow Architecture.

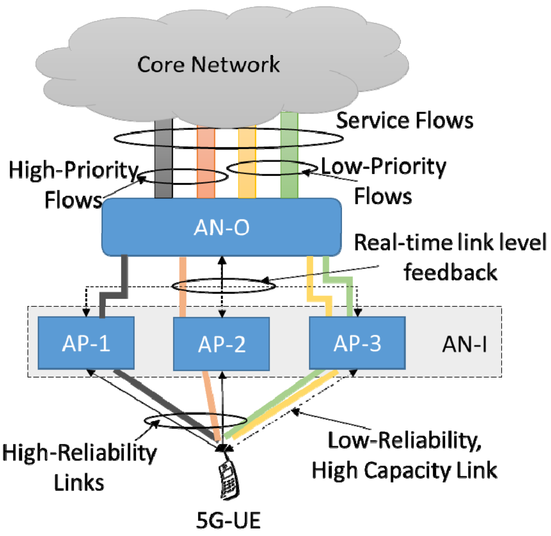

Fig. 5: Possible service flow delivery using the dynamic traffic steering framework.

An example of the service flow delivery mechanism in 5G is as shown in Fig. 5, where the AN-I layer steers high and low priority traffic (derived from the QCI values of the flow) into various 5G access points, taking the link quality into account. Thus, the high-priority traffic is delivered over more reliable links and the low-priority ones using low-reliability, but high capacity links. Here, the AN-O entity is assumed to receive the channel quality feedback on each TTI over all the active links for a user and mitigation mechanisms are activated as soon as a loss of connectivity over a link with high-priority traffic is detected. Since the AN-O has control over defining the radio bearer level QoS parameters, if a link is lost even temporarily, it can remove the associated radio bearer. This would be important especially for high-priority guaranteed bit rate (GBR) bearers, which requires resource reservation at the base station. Thus, having the dynamic QoS management at the AN-O can lead to efficient use of radio resources and reduction in delays, since the traffic can be routed dynamically, depending on real-time radio conditions experienced by each user.

A. Simulation Results

The concept was simulated using system-level simulations in an outdoor small cell deployment scenario, with the scenario dimensions the same as the one used in [10], excluding the buildings. 1000 UEs were uniformly dropped, with 100 multi-AIV deployed small cells that support 20 MHz LTE carrier at 2 GHz and 200 MHz mmW carriers at 30 and 60 GHz, with the path loss models following [15]. The probability was calculated using the NYU model defined in [16], and slow fading standard deviation of 3 dB for all the carriers. The value was an approximation based on the values presented in [16] for UEs in close vicinity of the transmitter.

Two packet sizes were assumed: a normal burst of 0.125 MB, based on the traffic model for open-areas or parks in [17], and a short burst were the packet is sent in two parts. The short burst assumption was made to simplify the dynamic traffic steering process by giving the AN-O layer to do a fast re-routing. It also enables the baseline mechanism to do faster re-routing, since packet sizes are smaller, the feedback of packet reception is also received faster. Since the data rates are expected to be significantly higher in 5G, the packet sizes are expected to be significantly higher than in LTE-A. Throughput calculations were done using the same method used in [10]. We assumed that non line-of-sight (NLOS) would lead to radio link failure and an additional delay of 200 ms was assumed in the baseline mechanism to route the traffic over the active links. It is also assumed that the UE can receive data over all the three active carriers using multi-connectivity. It is also assumed that the packets are delivered over a GBR bearer in 4G architecture, which leads to resource reservation at the eNB. In the optimized 5G mechanism, the AN-O layer which monitors all the links is assumed to tear down the GBR bearer while routing the traffic over to the remaining active links, leading to better resource utilization.

The normalized packet delivery delay distribution for the short and normal bursts is as shown in Fig. 6. The delay values are normalized to the maximum delay, in order to show relative gains. Here, the optimized 5G mechanism with dynamic traffic steering and RAN moderation performs better due to the fast rerouting of traffic over the active links where a radio link failure occurs on a mmW link. Due to the use of the dynamic QoS concept at the RAN, in the optimized 5G scheme, the resource reservation for GBR bearers are also removed as soon as an active link is lost, thereby enabling the reuse of resources for the other active users connected to a mmW 5G-NB. Significant gains of up to 20 % can be observed for the normal burst traffic can also be observed in the mean packet delay values shown in Fig. 7, relative to the baseline 4G mechanism. The gains are due to the fast rerouting by the AN-O layer with link quality monitoring and the use of dynamic QoS to avoid resource reservations for bearers that needs to be removed due to inactive radio links.

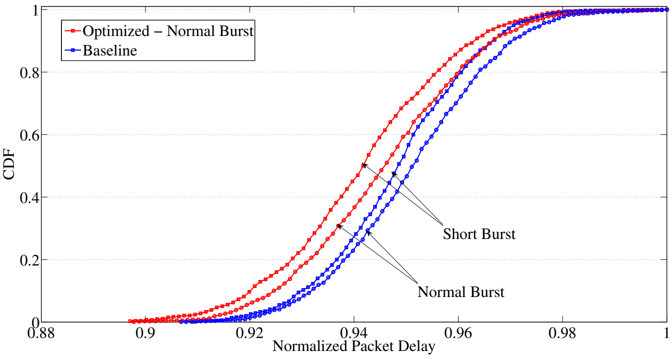

Fig. 6: Normalized CDF distribution of packet delivery delay.

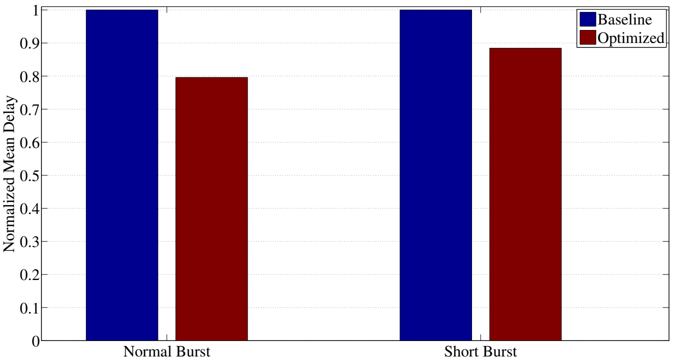

Fig. 7: Normalized mean packet delivery delay.

V. TRAFFIC STEERING FOR ENERGY EFFICIENCY

Real time traffic steering mechanism can be also designed to adapt the amount of active network resources to the actual amount of traffic that should be dispatched, thus improving the overall energy efficiency in the network. In [18] it was shown that typical multipoint coordination techniques as non-coherent Joint Transmission (JT) and Dynamic Point Selection/Dynamic Point Blanking (DPS/DPB) can be exploited in order to further reduce energy consumption when traffic is below its peak. In the proposed approach a centralized entity would dynamically select, based on traffic requests, which transmission nodes should be kept active, exploiting cooperation techniques between the selected nodes in order to switch off more nodes than it would be possible without multipoint coordination. The solution was evaluated assuming that base station overall power consumption, as a function of used resources, followed power models that were specified in the EARTH project [18] for 2010 equipment. METIS-II recently proposed similar power model for 2020 equipment [20] as shown in Fig. 8. According to these power models, the overall energy consumption of a macro or micro base station increases in first approximation linearly with the total amount of radio resources that are used for transmission. When no transmission is performed, the base station can enter in a sleep-mode that further reduces its consumption, so that a discontinuous point is shown in the figure for 0% resource usage. According to METIS-II power models, moving from years 2010 to 2020, not only the equipment will have better scalability of the energy consumption as a function of resource usage, and more effective sleep mode states, but the overall energy consumption at full resource usage will also be drastically lower. These new power models have been used in the same scenario as presented in [18] to assess the effectiveness of the proposed scheme when these more advanced equipment will be available.

A. Simulation Results

Simulation results are shown in Fig. 9, obtained in the simplified Madrid Grid scenario that was proposed in METIS [18]. The scenario reproduces an urban environment with 3 macro base stations and 9 micro base stations, serving 10 users each, and transmitting on a 10 MHz signal bandwidth. To test the capability to switch off unnecessary nodes, different traffic load conditions have been considered, obtained assuming Constant Bit Rate (CBR) traffic sources that generate a given traffic data rate for each user. As a reference also full-buffer traffic condition has been considered, even if in this case the proposed algorithm cannot switch off any cell, so no gain in energy consumption reduction can be achieved. The figure reports the overall power consumption when no multipoint coordination is exploited between nodes (NoCoord) and when the proposed centralized entity can exploit JT and DPS/DPB to improve Energy Efficiency (EE JT).

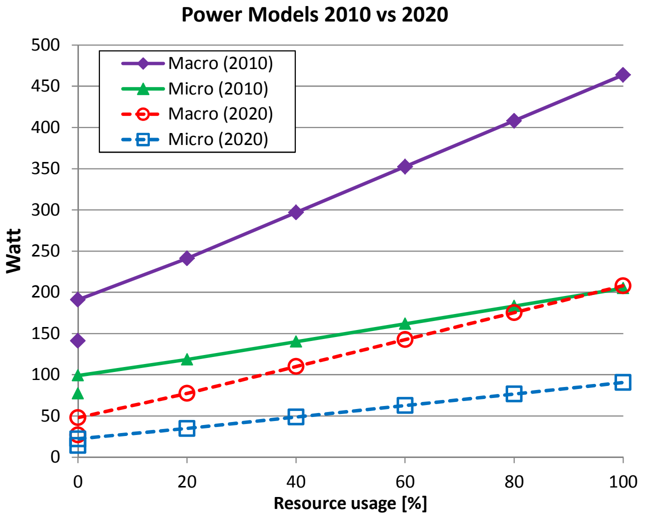

Fig. 8: Power models for Micro and Macro Base Station in 2010 and 2020 [20].

As it was expected, the higher energy efficiency of 2020 equipment reflects in a drastic reduction of the power consumption, both with and without the centralized entity for coordination. However, the higher dynamicity in power consumption offered by 2020 equipment can be better exploited with the proposed solution, so that in 2020 the power consumption reduction that can be obtained using the Energy Efficient (EE) JT scheme can be as high as 51 %, whereas in 2010 only savings up to 31 % could be achieved.

More in general, the results show that there is the possibility to exploit traffic steering together with traditional interference management techniques, such as multipoint coordination, to reduce energy consumption when traffic conditions are favorable.

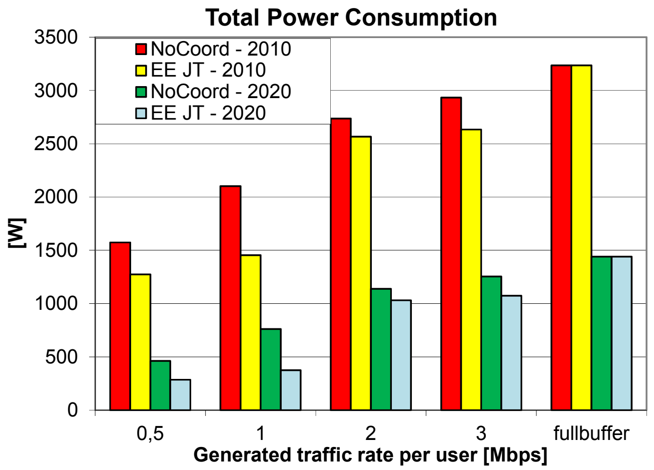

Fig. 9: Evolution of power consumption from 2010 and 2020, with and without the proposed coordination approach

VI. CONCLUSIONS

In this paper, various approaches to enable RAN moderation and dynamic traffic steering in 5G are discussed, which exploit the availability of multiple radio links. First, the various practical constraints and options for the tight integration of 4G LTE-A and 5G networks are presented. It was shown that the LTE-5G DC (multi-connectivity in the context of multiple AIVs) has the best performance in terms of 10th and 90th percentile user throughput values. Further, it was shown that fast traffic steering at RAN level with multi-connectivity can provide better performance than current state-of-the-art solutions in terms of packet delivery time. The gains from the use of fast traffic steering to achieve energy efficiency were also presented, with significant network power consumption reduction observed by using the coordinated transmission techniques, compared to the baseline scheme without such coordination.

The future work in the area can include defining the protocol architecture required to support the proposed enhancements, and the performance evaluation using a diverse set of use cases and scenarios. The impact on backhaul load due to the feedback required for fast traffic steering is also an interesting area of further study.

ACKNOWLEDGMENT

This work has been performed in the framework of the H2020 project METIS-II co-funded by the EU. The views expressed are those of the authors and do not necessarily represent the project. The consortium is not liable for any use that may be made of any of the information contained therein.

REFERENCES

[1] METIS-II White Paper, “Preliminary Views and Initial Considerations on 5G RAN Architecture and Functional Design,” March 2016.

[2] A. Osseiran, et al. “Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS project,” IEEE Communications Magazine, 52(5), 2014, pp. 26-35.

[3] N.T.K. Jørgensen, , D. Laselva, and J. Wigard. "On the potentials of traffic steering techniques between HSDPA and LTE." IEEE VTC-Spring, 2011.

[4] P. Munoz, et al. "Mobility-based strategies for traffic steering in heterogeneous networks." IEEE Communications Magazine, 51(5), 2013, pp. 54-62.

[5] J.M. Ruiz-Avilés, et al. "Traffic steering by self-tuning controllers in enterprise LTE femtocells." EURASIP Journal on Wireless Communications and Networking, 2012.1, 2012, pp. 1-16.

[6] N. Bhushan, J. Li, D. Malladi, et al., "Network densification: the dominant theme for wireless evolution into 5G," IEEE Communications Magazine, 52, no. 2 (2014), pp. 82-89.

[7] Ömer Bulakci, Athul Prasad, Jakob Belschner; Marten Ericson, et al. “Agile Resource Management for 5G - A METIS-II Perspective,” IEEE Conference on Standards for Communications and Networking (CSCN) 2015, Tokyo, Japan.

[8] NGMN Alliance, “5G White Paper,” Feb. 2015. [Online]. Available: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[9] 3GPP TS 23.203, “Policy and Charging Control Architecture,” Dec. 2015, ver. 13.6.0.

[10] A. Prasad, et al., "Network Assisted Small Cell Discovery in Multi-Layer and mmWave Networks," IEEE Conference on Standards for Communication and Networking (CSCN), Oct. 2015.

[11] 3GPP TS 23.401, “GPRS Enhancements for E-UTRAN Access,” Dec. 2015, ver. 13.5.0.

[12] Icaro Da Silva, Gunnar Mildh et al.,”Tight integration of new 5G air interface and LTE to fulfill 5G requirements”, VTC 2015

[13] Peter Ökvist, Henrik Asplund, Arne Simonsson et al., “15 GHz Propagation Properties Assessed with 5G Radio Access Prototype”, PIMRC-2015

[14] M. Ericson et al, “Tight integration of LTE and novel air interface to fulfill 5G requirements: performance evaluation”, submitted to VTC 2016-Fall.

[15] M. Cudak et al., “Moving Towards mmWave-based Beyond-4G (B-4G) Technology,” in IEEE VTC-Spring, 2013, pp. 1–5.

[16] S. Sun, et al. "Path Loss, Shadow Fading, and Line-Of-Sight Probability Models for 5G Urban Macro-Cellular Scenarios." IEEE Globecom 2015.

[17] METIS Deliverable D6.1, “Simulation Guidelines,” Oct. 2013.

[18] ICT-317669 METIS project, “Final performance results and consolidated view on the most promising multi-node/multi-antenna transmission technologies,” Deliverable D3.3, Feb. 2015

[19] EARTH Project, see http://www.ict-earth.eu

[20] METIS-II, “Performance evaluation framework,” Deliverable D2.1, Jan. 2016