Radio-Aware Service-Level Scheduling to Minimize Downlink Traffic Delay Through Mobile Edge Computing

Jose Oscar Fajardo(✉), Ianire Taboada, and Fidel Liberal

University of the Basque Country (UPV/EHU), ETSI Bilbao, 48013 Bizkaia, Spain
{joseoscar.fajardo,ianire.taboada,fidel.liberal}@ehu.eus
http://det.bi.ehu.es/NQAS

Abstract. One of the most challenging problems in mobile broadband networks is how to assign the available radio resources among the different mobile users. Traditionally, research proposals are either specific to some type of traffic or deal with computationally intensive algorithms aimed at optimizing the delivery of general purpose traffic. Consequently, commercial networks do not incorporate these mechanisms due to the limited hardware resources at the mobile edge. Emerging 5G architectures introduce cloud computing principles to add flexible computational resources to Radio Access Networks. This paper makes use of the Mobile Edge Computing concepts to introduce a new element, denoted as Mobile Edge Scheduler, aimed at minimizing the mean delay of general traffic flows in the LTE downlink. This element runs close to the eNodeB element and implements a novel flow-aware and channel-aware scheduling policy in order to accommodate the transmissions to the available channel quality of end users.

Keywords: Mobile networks · Flow scheduling · Mobile edge computing · 5G

1 Introduction

Complex future content centric Internet and the 5G paradigm, including Network Function Virtualization (NFV) and Software Defined Networking (SDN), suggest the introduction of intelligent network nodes that will enable more powerful adaptation and prioritization frameworks over the whole transmission chain, and especially at the edge of mobile network segments [1].

The classical issue in current Radio Access Networks (RAN), including 4th Generation (4G) Long Term Evolution (LTE), is related to the problem of how to handle the assignment of shared radio resources among multiple mobile users in order to maximize the overall service experience. Many research proposals have dealt with the problem, aimed at optimizing the provision of certain services and traffic patterns [2]. Most of these works are based on the introduction of

R. Agüero et al. (Eds.): MONAMI 2015, LNICST 158, pp. 121–134, 2015.
DOI: 10.1007/978-3-319-26925-2_10

service and channel awareness in the scheduling function of the eNodeB, considering channel awareness as fine-grain Channel Quality Indicator (CQI) feedbacks provided by mobile devices.

Unfortunately, most of these types of scheduling functions are based on ideal channel awareness and are driven by complex mathematical logic. The associated complexity prevents its implementation into real-world eNodeBs, which need to determine the multi-user traffic assignments at Transmission Time Interval (TTI) slots (i.e., 1 ms in LTE).

In the framework of future 5G mobile networks, different proposals are emerging aimed at overcoming the capacity limitations of current Radio Access Networks (RAN). Cloud computing principles are being proposed to design the future RAN, in order to create Cloud Radio Access Networks (C-RAN) with increased flexible capacity [3], and to deploy service instances within the C-RAN enabling increased adaptability to mobile users' context.

One of the emerging technologies to cope with more personalized and user-centric service provisioning is the novel Mobile Edge Computing (MEC) industry initiative [4], a promising approach to solve these types of problems from an operator-supported approach. This initiative proposes that mobile network operators would provide an API to third-party partners, offering them access to critical features such as location awareness and network context information. This information may be exploited to deploy proximity-enabled services with close-to-zero latency characteristics, in order to optimize the management of future mobile networks.

Figure 1 illustrates different alternatives for the centralization of RAN functions. On one hand, fully centralized RAN entails that all the processing is deployed in centralized data centers. In 3GPP terminology, the BaseBand Unit (BBU) would perform all the required functions including the Radio Resource Management (RRM), while the Remote Radio Header (RRH) would only transmit the generated radio signals. This architecture enables a full virtualization of RAN functions, but requires reliable high speed connections in the fronthaul. Other alternatives, as discussed in [5], allow the centralization of specific RAN functions. As shown in Fig. 1, the RRM functions may still run at the eNodeB specific hardware in order to guarantee accurate channel state information. Meanwhile, other functions are more feasible for centralization/virtualization, usually associated to the higher layers.

Regardless the adopted architecture, MEC-driven service instances must be deployed over the cloud resources available at the RAN. Thus, the degree of coupling between service-level instances and RRM functions may differ.

This paper focuses on the network-assisted optimization of downlink traffic flows in the latter scenarios, as a solution to introduce network intelligence in partially centralized RAN deployments. Rather than applying complex scheduling logic at the eNodeB hardware, this paper explores the possibility to deploy user-aware service instances within the C-RAN in order to optimize the delivery of traffic flows based on close-to-the-user channel awareness.

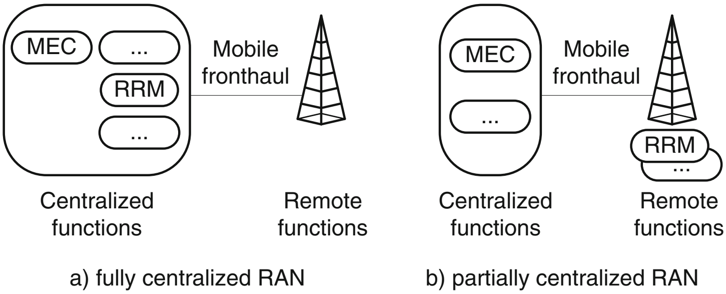

Fig. 1. Fully centralized vs. partially centralized RAN.

1.1 Contributions and Schedule of the Paper

In this paper, we analyze the possibility of extracting the intelligent scheduling logic out of the eNodeB specific hardware to the cloud-enabled RAN. This novel approach emerges as a non-optimal but implementable solution, bringing service optimization close to mobile end users while reducing the complexity required by radio hardware elements.

The main contributions of this paper are:

- A novel service delivery architecture, based on the principles of C-RAN and MEC, which allows deploying intelligent scheduling logic at the mobile edge without requiring high computational resources at the eNodeB hardware.
- The study of the resulting two-round channel-aware scheduling process, comprising a first traffic conformance deployed as a MEC instance and a latter RRM at the eNodeB hardware.
- The analysis of some preliminary results for the proposed solution obtained through simulations, which incorporates authors previous results in the field of intelligent scheduling and experimental LTE channel measurements.

Section 2 describes the proposed architecture, where a novel flow- and channel-aware scheduling policy is located at the mobile edge just before the eNodeB scheduling function. The new scheduler is aimed at implementing context-aware traffic shaping to minimize the average delay of traffic flows; meanwhile, eNodeB scheduling is driven by a classical implementable scheduling policy [6]. Section 2 also discusses some technological issues that determine the achievable performance in this kind of multi-user mobile service provisioning, such as the partially observable channel problem.

Section 3 focuses on the scheduling policy proposed to drive the cloud-enabled traffic shaping. The proposed scheduling policy is based on [7], where authors define a near-optimal index policy named Attained Service dependent Potential Improvement (ASPI). ASPI solves the opportunistic scheduling problem for general file size distributions and multiple channel states, restricted to i.i.d. channels and single user transmissions. In this paper, we apply ASPI to multi-user simultaneous transmissions leading to Multi-user Attained Service Potential Improvement (MASPI). Additionally, we analyze the problem of moving the channel awareness to the cloud-enabled RAN and the implications of using non i.i.d. radio channels. In this sense, the paper does not propose a novel scheduling policy specifically designed to the problem, but it analyzes the implications of deploying state of the art radio scheduling functions out of the eNodeB.

Section 4 focuses on the performance evaluation, comparing the delay results of three different architectures:

- eNodeB running classical scheduling policy aimed at maximizing cell throughput.
- eNodeB running novel scheduling policy aimed at optimizing mean flow delays.
- MEC function running novel scheduling policy aimed at optimizing mean flow delays together with eNodeB running classical scheduling policy aimed at maximizing cell throughput.

As derived from the performance evaluation and discussed in Sect. 5, the proposed novel architecture allows improving the overall performance of the mobile services without increasing the computational requirements of eNodeB elements.

2 MEC-assisted Traffic Scheduling

The proposed novel service element to be deployed within the C-RAN is intended to incorporate novel concepts of complex radio scheduling logic through software-based Mobile Edge Computing instances. This way, the new service element provides the required user awareness to deploy optimized service delivery over current eNodeB scheduling functions.

Figure 2 illustrates the overall network architecture of the resulting solution.

A number of Internet traffic sources deliver traffic flows towards the mobile end users following arbitrary flow distributions. These flows traverse the Internet segments and the mobile backhaul before arriving at the RAN.

In current 4G mobile network architectures the scheduling logic is run at the eNodeB elements, which also implement the radio air interface. Therefore, these eNodeB elements are the unique network nodes endowed with user-reported channel awareness. Channel awareness is indeed implemented by means of CQI reporting from mobile devices to the eNodeB, based on the configured CQI Reporting Rate (CRR). Although perfect channel awareness is not possible, CRR is usually configured to a low value (e.g., 5 ms) in order to avoid imprecise scheduling decisions.

In emerging C-RAN architectures, Internet applications would be able to gather users' channel feedbacks to some extent in order to optimize their traffic delivery to mobile users. As illustrated in Fig. 2, the key element in the MEC industry initiative is the MEC Server, which is integrated into the mobile operators' RAN in order to provide value-added capabilities to third party developers [4]. In the case of LTE, the MEC Server is directly integrated into the

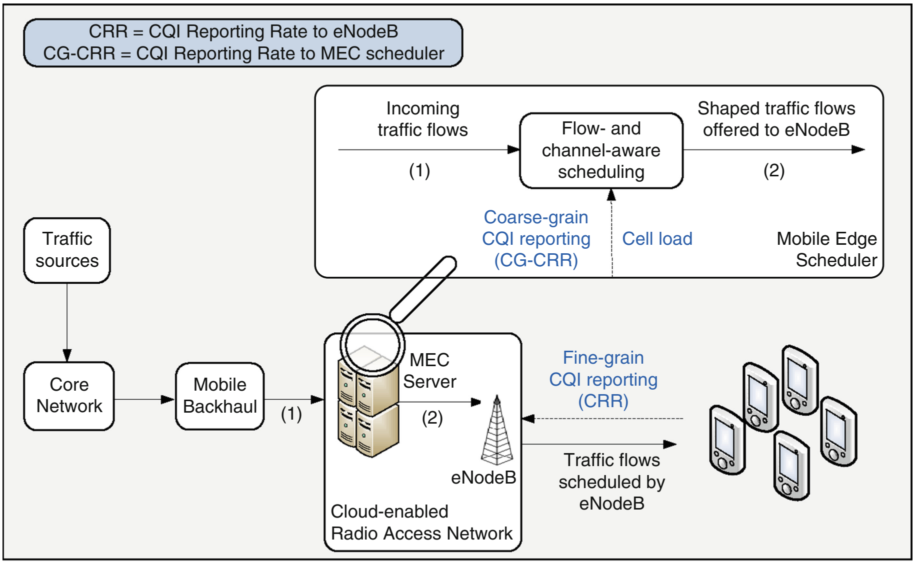

Fig. 2. Proposed MEC-assisted flow scheduling architecture.

eNodeB. The MEC Server is based on cloud computing principles to run third-party applications over a common hardware infrastructure. These applications may range from lightweight monitoring instances, which would provide relevant RAN-related information to optimize external OTT servers, to more complex applications, intended to handle and modify the traffic to/from the mobile users at RAN level.

Upon this MEC Server element, we propose to deploy the Mobile Edge Scheduler (MESch) as an intelligent node aimed at optimizing the traffic delivery to multiple mobile users. MESch is included in the service provisioning chain, enabling the required flow awareness concepts. At the same time, MESch interacts with the MEC Server in order to retrieve cell and user context information, such as the cell load or per-user radio performance statistics.

The MEC API is the designated interface for retrieving cell statistics and individual channel quality information. In multi-user scenarios, the CQI feedback granularity would entail a high load and traffic volume in the interface between MEC Server and MESch. Thus, a coarse-grain feedback statistic is defined in Fig. 2 as CG-CRR. Besides, MESch implements the flow- and channel-aware scheduling policy in order to maximize the service experience. This module combines flow awareness, user context awareness and cell status awareness, in order to run the required intelligent logic that will be detailed in the following section.

The output of MESch can be seen as a series of decomposed traffic flows, properly segmented in chunks according to the optimization logic and being forwarded to the eNodeB for its delivery to the end users. Consequently, the eNodeB will run a more basic scheduling logic at TTI level and making transmission decisions based on the updated channel feedbacks reported by the mobile devices. The whole operating mode, with two concatenated scheduling functions with non-ideal shared channel awareness, entails a series of limitations that will be analyzed in this paper.

3 Mobile Edge Scheduler

This section presents the scheduling policy proposed to perform the service-level traffic shaping at the mobile edge.

The proposed MESch includes a novel scheduling algorithm aimed at orchestrating the delivery of traffic flows to the different users present in the mobile cell. As a result, different flows (i.e., HTTP objects, media segments) arrive at the MESch, which needs to prioritize the delivery of these flows to minimize the overall mean flow delay. For this aim, the MESch takes into consideration the averaged radio conditions of each user in a CG-CRR period, as well as the current flow size-related information. The flow sizes are a priori unknown to the scheduler, but the flow size distribution of the traffic is known.

In the next subsections, we first analyze the state of the art of relevant scheduling policies and discuss the adopted alternatives. Later, the proposed scheduling policy is detailed.

3.1 Background on Flow Scheduling for Wireless Systems

Over the years the literature on performance evaluation and optimal scheduling of traffic flows in time-varying wireless channels has grown tremendously. Undoubtedly, the most studied resource allocation problems aim at minimizing the mean transfer delay of user flows. Although several channel-aware strategies exist with the objective of minimizing the mean flow delay, Best CQI and Proportional Fair [6] among the most popular ones, the achievement of the optimal solution for time-varying scheduling optimization problems is computationally and analytically unfeasible. To cope with this problem, flow-level channel-aware opportunistic scheduling in time-varying systems has been analyzed by approximate techniques [8,9] in order to design simple, tractable and implementable well-performing heuristic priority scheduling rules [7,9,10].

On the other hand, classical results show that prioritizing short flows minimizes the overall mean delay [11,12]. However, this is only applicable if there is a priori knowledge of flow sizes, which is not generally the case in current IP networks. For this reason, non-anticipating size-based scheduling disciplines make use of flow attained service, i.e. the bits that have been transferred of a flow, for making scheduling decisions. In this context, Gittins proposes an index rule that, based on the attained service of jobs, minimizes the mean delay when channel capacity is constant [13].

Therefore, an approach that combines both flow size awareness and channel awareness seems promising [8]. Our previous work in [7] proposes a simple, analytically founded, tractable and well-performing scheduling index rule that combines non-anticipating size awareness and channel awareness. The obtained priority-based discipline is called ASPI rule, which was derived by using Whittle approach [14]. However, this work assumes that a single user transmits in each transmission time interval, while modern cellular networks entail simultaneous transmission of multiple flows.Therefore, in the following subsection we propose a scheduler based on ASPI that overcomes the lack of the simultaneous transmission of flows.

3.2 MASPI Scheduling Algorithm

The ASPI index rule is our previous solution for solving an attained service-dependent and channel-aware Markov Decision Process by Whittle methodology [7]. To achieve this non-anticipating size-based and channel-aware solution, we first formulated the mean delay minimizing scheduling problem in the framework of Markovian decision processes. Then, since obtaining the optimal solution of this analytical model unachievable, we derived ASPI by using methodologies based on Whittle approach.

The Whittle method [14] consists in obtaining a metric per flow state that measures the dynamic transmission priority of a single flow. In this way, a simple scheduling index rule appears: at every slot, select for transmission the K flows with the highest current Whittle index value. With that aim, the non-anticipating size-based and channel-aware scheduling problem aimed at minimizing the mean delay can be relaxed by requiring to serve K jobs per slot on average, and further approached by Lagrangian methods [14]. As a result, a single-job price-based parametrized optimization problem is achieved, where the Lagrangian parameter can be interpreted as the per-slot cost of transmitting. Thus, the Whittle index is the break-even value of the Lagrangian parameter. For our case study, it measures the expected efficiency of transmitting a flow in each attained service and channel condition state. Note that the opportunistic and non-anticipating size-aware ASPI index was derived for an uncorrelated channel model and for the significant class of size distributions with Decreasing Hazard Rate (DHR).

The so-called ASPI index rule gives priority to users that are in their best channel conditions. The index value for users that are not in their maximum achievable channel condition equals the ratio between the instantaneous transmission completion probability and the expected potential improvement of the instantaneous transmission completion probability.

Thus, the ASPI scheduling discipline consists in, at every decision slot:

- choosing to transmit a flow from a user that is in the best channel condition with the highest value of $c\mu_{(a,N)}$;
- and if there is no user in the best channel condition, choosing the flow with the highest value of $\frac{c\mu_{(a,n)}}{\sum_{m>n} q_m(\mu_{(a,m)}-\mu_{(a,n)})}$.

Where q_m is the probability of being in channel state m, $\mu_{a,n}$ the flow transmission completion probability ($P(a < X \leq a + r_n | X > a)$), being X the flow size, a the instantaneous attained service and r_n the bits transmitted in channel condition n (being N the best channel condition that a user can achieve), and c the holding cost per slot while the flow transmission is not completed (when the objective is minimizing the mean delay the value of c is 1). Note that if the flow size distribution belongs to the DHR class, $\mu_{a,n}$ also decreases with attained service [10].

Nevertheless, the performance of this Whittle-based policy has only been validated considering the transmission of a single flow in each TTI, this is, assigning all the available network resources to the same flow, which is not realistic in current and future wireless networks. However, a Whittle-based policy allows the simultaneous transmission of multiple flows. In such situation, we extend ASPI proposal to a multi-user approach. We propose an enhanced ASPI scheduler, denoted as Multi-user ASPI or MASPI, which considers the transmission of multiple user traffic flows per TTI.

In this way, the MASPI discipline consists in: at every decision slot, having K network resources to assign:

- choosing to transmit K flows from users that are in their best channel condition with the highest value of $c\mu_{(a,N)}$;
- and if there are less than K user flows in their best channel condition, being K' flows in the best channel conditions, choosing to transmit the $K - K'$ flows with the highest value of $\frac{c\mu_{(a,n)}}{\sum\limits_{m>n} q_m(\mu_{(a,m)} - \mu_{(a,n)})}$.

4 Performance Evaluation

In this section, we analyze the performance of the proposed MEC-assisted scheduling. For that purpose, we compare the achieved performance in terms of mean delay of the following different approaches:

- Classical eNodeB scheduler (eNodeB(BC)): eNodeB running Best CQI (BC) discipline. This discipline gives priority to those users with higher reported CQI values.
- Near-optimal eNodeB scheduler (eNodeB(MASPI)): eNodeB running MASPI scheduler. This discipline aims at prioritizing those flows with lower expected time to finish.
- The proposed two-stage proposal (MESch-eNodeB(BC)): MASPI running at MESch and classical BC policy at eNodeB.

As can be observed, all these schemes are channel-aware. Based on the CQI values, eNodeB applies Adaptive Modulation and Coding (AMC) to select the most suitable Modulation and Coding Scheme (MCS) for each UE. Similarly, MASPI incorporates the coarse-grain channel awareness for decision making. Additionally, MASPI takes into account the monitored Block Error Rate (BLER) and would apply conservative MCS assignations when required in order to mitigate the effect of radio retransmissions under highly variable radio channel conditions.

For performance evaluation, we simulate the behavior of an LTE network configured with 20 MHz of bandwidth. Consequently, the eNodeB scheduling function arranges the transmissions of multi-user traffic at every TTI of 1ms, optimally allocating the set of the available 100 Resource Blocks (RBs). For simplicity, we assume that all the UEs are LTE Category 4 using LTE MIMO, with up to 150 Mbps of downlink peak rate.

In order to capture the effects of real-world radio channels, we employ actual CQI traces from live LTE networks as reported in [15]. Therefore, we use our own system-level simulation environment implemented in MATLAB. This implementation allows us to simulate the radio channel at CQI level, and to apply the required RRM functions. BLER is generated based on the difference between the last reported CQI and the actual CQI.

Table 1 illustrates the channel state probabilities of the experimental CQI traces captured for mobile users and used for performance evaluation. In our set up, CQI traces at eNodeB provide a granularity of 5 ms, while CQI traces available at MESch are generated from averaging these experimental traces with 1 s granularity. The average CQI value is 9.47 and, consequently, the average data rate individually achievable is 67.326 Mbps.

Table 1. Channel state probabilities for different CQIs.

CQI	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
q_n	0	0	0	0.00046	0.024	0.078	0.0047	0.015	0.1918	0.13	0.202	0.19	0.0842	0.056	0.0173	0.00132

Regarding traffic demands, we consider Pareto distributed flow sizes that belong to DHR class. It is known that Internet traffic flows are properly modeled by means of Pareto distributions [16]. In this case, we assume a mean flow size of 5 Mbit and a shape parameter of 1.3. These flows arrive according to a Poisson process, whose rate determines the network load. We consider six network states: low load ($\rho = 0.25, 0.375$), medium load ($\rho = 0.5, 0.75$) and high load ($\rho = 0.9, 0.95$).

4.1 Results

Now we analyze the performance in terms of delay for the aforedescribed wireless scenario, considering delay as the total time since the reception of a flow until its complete transmission to the mobile user. For that aim, we will show the mean delay results of the three schemes under study for the considered network loads. In addition to the mean value, the 95 % confidence intervals are included.

Figure 3 collects the mean delay results of the analyzed LTE settings. As can be observed, on the one hand, our two-phase scheduling proposal behaves notably better than a classical scheduler at the eNodeB such as BC. On the other hand, MESch-eNodeB(BC) provides worse results than a near-optimal eNodeB scheduling solution such as MASPI with BLER awareness. Moreover, it is worth to mention that both the difference in mean delay among schedulers and the confidence intervals increase with the network load value.

Therefore, the average delays achieved with our proposal enhance those obtained with eNodeB(BC), while not requiring high computing capabilities in the eNodeB hardware. These results are in the same scale as those achieved by the one-phase MASPI solution at the eNodeB, even if MESch-eNodeB(BC) makes use of more uncertain channel information due to the CG-CRR parameter.

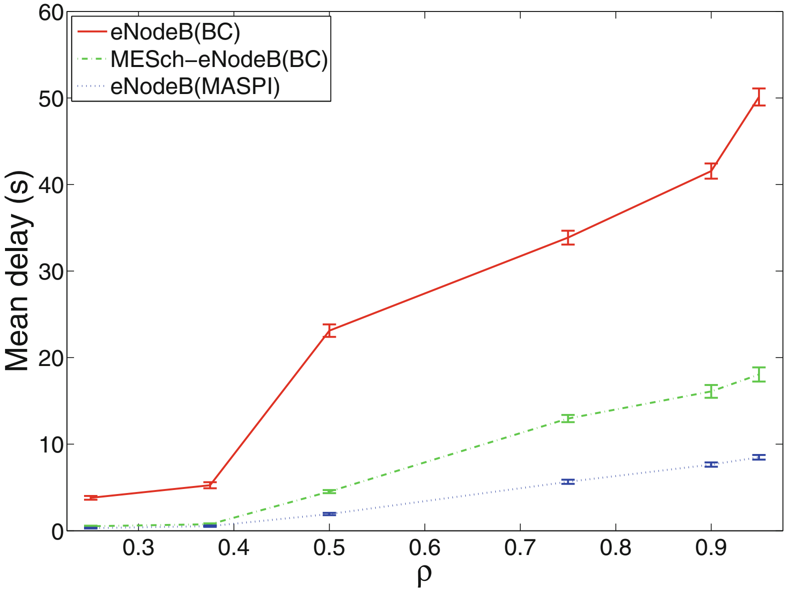

Fig. 3. Mean delay results.

Tables 2, 3, 4 show the detailed results for three representative test points, corresponding to low, medium and high radio network loads. Three parameters are shown for each test scenario: the ratio of completed flows in the simulation time, the average BLER through the whole simulation and the average flow delay experienced due to the additional MESch node. Logically, the latter value is only significant for the MESch-eNodeB(BC) case.

As can be observed, the ratio of flows completely transmitted to the UE follows the same behaviour than the average delays. For all the considered network loads, eNodeB(MASPI) and eNodeB(BC) provide the lowest and highest values respectively. The proposed MESch-eNodeB(BC) scheme achieves intermediate results.

Similarly, we can observe that experienced BLER is higher for eNodeB(BC), while eNodeB(MASPI) reduces the BLER significantly. The enhancement achieved in the proposed MESch-eNodeB(BC) scheme is considerable at low network loads, and converges to eNodeB(BC) at high loads.

Finally, it is remarkable that the average flow delay due to the added MESch scheduling phase also increases with the network load, reaching up to 1.5 s at high network load conditions.

As a result, we can state that the proposed MESch-eNodeB(BC) provides accurate performance outcomes, exhibiting average flow delays that are closer to the near-optimal eNodeB(MASPI) scheduling policy than to the classical eNodeB(BC) policy. At the same time, the complexity of the radio head is reduced to BC scheme and the complex logic is deployed at the cloud facilities of the RAN.

Table 2. Detailed results for $\rho = 0.25$.

	Finished flows	BLER	MESch delay
eNodeB(BC)	0.9805	0.2186	-
MESch-eNodeB(BC)	0.9991	0.1467	0.2061
eNodeB(MASPI)	0.9994	0.0932	-

Table 3. Detailed results for $\rho = 0.75$.

	Finished flows	BLER	MESch delay
eNodeB(BC)	0.9531	0.3019	-
MESch-eNodeB(BC)	0.9755	0.1709	0.6731
eNodeB(MASPI)	0.9896	0.1029	-

Table 4. Detailed results for $\rho = 0.95$.

	Finished flows	BLER	MESch delay
eNodeB(BC)	0.9299	0.3047	-
MESch-eNodeB(BC)	0.9549	0.3072	1.4491
eNodeB(MASPI)	0.9892	0.1176	-

5 Conclusions

This paper deals with the inclusion of complex scheduling algorithms in the context of emerging mobile broadband networks. The scenario analyzed includes a multi-user LTE network, where mobile users wander through the LTE cell experiencing realistic variable channel conditions. Since the paper focuses on the optimization of mean flow delays in general-purpose conditions, the traffic profile is assumed to be a Pareto distribution.

Taking into account several recent research results, flow- and channel-aware scheduling policies provide near-optimal results in these types of dynamic wireless scenarios. However, even the most efficient of those algorithms entail high computational requirements preventing their application in real-world eNodeB scheduling functions.

This paper proposes an alternative way of running this complex scheduling logic within the mobile edge but out of the eNodeB hardware element. This solution is based on the concepts of Mobile Edge Computing, which specifies the architecture to run personalized services close to the end mobile users and defines an interface to provide radio channel feedbacks to these service instances.

As a result, we propose a two-stage scheduling process. First, the Mobile Edge Scheduling -MESch- algorithm is aimed at shaping the traffic according to the actual flow and coarse-grain channel conditions. Second, the eNodeB runs its lightweight scheduling function with fine-grain channel information.

In this context, we compare the performance results obtained for (i) an eNodeB running the near-optimal MASPI scheduling policy, (ii) an eNodeB running the classical Best CQI scheduling policy, and (iii) the proposed solution with a double MESch-eNodeB scheduling.

As expected, eNodeB(MASPI) provides the lowest mean delays for the complete set of traffic flows. While the proposed architecture provides higher mean delays compared to eNodeB(MASPI), it significantly improves the results of the simple eNodeB(BC) approach. The total transmission delays in the proposed architecture are the combination of the scheduling function at MESch and the scheduling function at eNodeB. The first contribution provides slightly higher delay values than eNodeB(MASPI). Although the scheduling logic is the same, the results of MESch cannot capture the effects of radio retransmissions due to the variability of radio channels. The second contribution provides significantly lower average delay values compared to eNodeB(BC). MESch decomposes the Pareto traffic into a new train of flow segments according to the actual flow and channel characteristics. Thus, for example, heavy flows associated to users in good channel conditions may be delayed by MESch preventing higher delays for other flows with higher probability to finish.

MASPI is only based on the flow size distribution and the attained service of each flow. This characteristic allows implementing this kind of logic without the constraint of requiring the complete reception of flows to capture flow sizes. As a result, we show how this kind of intelligent scheduling logic may be implemented in emerging cloud-enabled multi-user mobile networks, relaxing the computing requirements of the eNodeB scheduling functions.

5.1 Future Work

The presented preliminary performance results anticipate the benefits of the proposed two-stage channel-aware scheduling architecture. Anyway, a more comprehensive performance analysis is required in order to identify the specific sources of delay and possible ways of improvement. This study would likely lead to a further enhancement of the MESch algorithm, in order to get closer to the optimal solution. Additionally, the computational and energy requirements in the different elements of the architecture should be thoroughly analyzed.

Possible malfunction of the MESch shall also be analyzed. Since the MESch element is introduced as an add-on to the service chain, the effect of disabling this new function would result in falling back to the default situation, where only the eNodeB RRM is performed. However, the effect of highly variable CQI could lead to unstable situations where MESch applies inaccurate scheduling decisions leading to performance degradations.

In the current implementation of the MESch algorithm, the coarse-grain CQI values reported by the MEC Server are used during the whole CG-CRR period. In highly variable radio channels, this assumption may introduce further performance degradations due to inaccurate channel estimations. Therefore, the inclusion of partially observable channel models could enhance the performance of the proposed MESch in those scenarios.

Acknowledgments. The research leading to these results has received funding from the European Union's H2020 Research and Innovation Project SESAME, under the Grant Agreement H2020-ICT-671596, and from the Spanish Ministerio de Economia y Competitividad (MINECO) under grant TEC2013-46766-R: QoEverage - QoE - aware optimization mechanisms for next generation networks and services.

References

1. Soldani, D., Manzalini, A.: Horizon 2020 and beyond: on the 5G operating system for a true digital society. IEEE Veh. Technol. Mag. **10**(1), 32–42 (2015)
2. Fouziya Sulthana, S., Nakkeeran, R.: Study of downlink scheduling algorithms in LTE networks. J. Netw. **9**(12), 3381–3391 (2014)
3. Wu, J., Zhang, Z., Hong, Y., Wen, Y.: Cloud radio access network (C-RAN): a primer. IEEE Netw. **29**(1), 35–41 (2015)
4. Patel, M., et al.: Mobile-Edge Computing Introductory Technical White Paper. White Paper, Mobile-edge Computing (MEC) industry initiative (2014)
5. Small Cell Forum, Small cell virtualization functional splits and use cases, Document 159.05.1.01 (2015)
6. Aalto S., Lassila P.: Flow-level stability and performance of channel-aware priority-based schedulers. In: Proceeding of NGI 2010 (6th EURO-NF Conference on Next Generation Internet) (2010)
7. Taboada, I., Liberal, F., Jacko, P.: An opportunistic and non-anticipating size-aware scheduling proposal for mean holding cost minimization in time-varying channels. Perform. Eval. **79**, 90–103 (2014)
8. Aalto S., Penttinen A., Lassila P., Osti P.: On the optimal trade-off between SRPT and opportunistic scheduling. In: Proceedings of ACM Sigmetrics, pp. 185–196 (2011)
9. Ayesta, U., Erausquin, M., Jacko, P.: A modeling framework for optimizing the flow-level scheduling with time-varying channels. Perform. Eval. **67**, 1014–1029 (2010)

10. Taboada I., Jacko P., Ayesta U., Liberal F.: Opportunistic scheduling of flows with general size distribution in wireless time-varying channels. In: 26th IEEE International Teletraffic Congress (ITC), pp. 1–9 (2014)
11. Kleinrock, L.: Queueing Systems, Volume 2: Computer Applications. Wiley, New York (1976)
12. Avrachenkovt, K., Ayesta, U., Brown, P., Nyberg, E.: Differentiation between short and long TCP flows: predictability of the response time. In: IEEE Computer and Communications Societies, INFOCOM 2004, vol. 2, pp. 762–773 (2004)
13. Gittins, J., Glazebrook, K., Weber, R.: Multi-armed Bandit Allocation Indice. Wiley, New York (2011)
14. Nino-Mora, J.: Dynamic priority allocation via restless bandit marginal productivity indices. TOP **15**(2), 161–198 (2007)
15. Fajardo, J.O., Taboada, I., Liberal, F.: Analysis of CQI traces from LTE MIMO deployments and impact on classical schedulers. In: Aguayo-Torres, M.C., Gómez, G., Poncela, J. (eds.) WWIC 2015. LNCS, vol. 9071, pp. 60–73. Springer, Heidelberg (2015)
16. Thompson, K., Miller, G.J., Wilder, R.: Wide-area Internet traffic patterns and characteristics. IEEE Netw. **11**(6), 10–23 (1997)