Cross-layer approach enabling communication of high number of devices in 5G mobile networks

Jan Plachy, Zdenek Becvar
Department of Telecommunication Engineering, FEE,
Czech Technical University in Prague
Technicka 2, 166 27 Prague, Czech Republic
jan.plachy@fel.cvut.cz, zdenek.becvar@fel.cvut.cz

Emilio Calvanese Strinati
CEA-Leti
17 rue des Martyrs, 38000 Grenoble, France
emilio.calvanese-strinati@cea.fr

***Abstract*—Introduction of Internet of Things and Machine Type Communication to future mobile networks will cause significant increase in the number of connected devices. At the same time, the connected devices can change traffic patterns as frequent transmission of small volumes of data is expected from sensors and machines. Transmission of such data is very inefficient due to redundancy of signaling information. In this paper, we analyze limits for the number of devices and machines communicating in current 4G mobile network. Then, we propose a novel solution, which shifts the current limits of the number of communicating devices towards requirements on 5G mobile networks. The proposed solution exploits cross-layer approach considering buffering of data and clustering of nearby users in order to minimize overhead and improve transmission efficiency. This way, we can increase the number of devices served by a single cell up to 24 times comparing to the state of the art solution.**

Keywords—connected devices, mobile network, 5G, Machine Type Communication, Internet of Things, mobile edge computing

I. INTRODUCTION

Mobile networks of the fifth generation (5G) are foreseen to enable communication of a huge number of connected devices (e.g., smartphones, tablets, sensors, or machines), either communicating or sensing data. The 5G should enable communication of trillions of devices in 2020 [1]. The increase in the number of devices is based on introduction of Internet of Things (IoT) and Machine Type Communication (MTC) and on continuous increase in the number of conventional users' devices such as smartphones or tablets. It is expected that a single base station would serve between ten and hundred thousand machine type devices and thousands of conventional user equipment (UE) [2]. Also, at the same time, the number of network parameters, which need to be configured in order to optimize network performance is expected to increase from 1500 in 4G to 2000 in 5G [3]. This motivates self-optimization of the mobile networks [3]. For efficient self-optimization of mobile networks, huge amount of information has to be collected and further processed. Such information may range from radio parameters, such as signal quality to information related to position or speed of UE or sensors. However, the common indicator of all expected information is small volume of data collected from many devices with relatively high frequency.

Collected data has to be processed either centrally or using distributed computing resources deployed closer to the edge of mobile network. An example of such solution is mobile edge computing represented by, e.g., Small Cell Cloud [4][5]. For both centralized as well as distributed processing of data, load of the radio channels will increase significantly due to a need for gathering of small payloads with significant overhead from a large number of devices [6].

The architecture of LTE-A mobile network is designed to support high speed data transmission of a large payload (e.g. video, file sharing, etc.). Nevertheless, when it is required to serve a large amount of devices sending or receiving relatively small volumes of data, performance of LTE-A becomes limiting [7][8]. In current mobile networks, majority of traffic is being transmitted in downlink rather than in uplink [9]. However, collection of data from devices would significantly increase uplink utilization.

Transmission of small payloads by a high number of densely spread devices is currently an issue for LTE-A as its transmission protocol stack is not prepared to handle it [10]. Limitations are seen in maximal number of devices to be scheduled within a single subframe [7] as well as in collisions of devices trying to connect to the network [8]. Also, an important problem comes from transmission of significant overhead at all layers of the protocol stack. This problem can be partly solved by Robust Header Compression (ROHC) [11], which reduces Transmission Control Protocol/Internet Protocol (TCP/IP) overhead. Transmitted overhead can be reduced by buffering of several payloads to send them at once [12]. Other possible solution is to cluster nearby users and send their payload merged into single packet with less overhead as expected in, for example, wireless sensor networks [13].

The contribution of this paper consists in analysis of the overhead introduced at bottom layers – physical, control and TCP/IP – of radio interface in LTE-A mobile networks. Then, we derive the limits introduced by current LTE-A transmission protocol stack on the number of devices frequently transmitting small amount of data. Further, we introduce a new solution improving efficiency of frequent transmission of small payloads from large number of devices over radio interface in mobile networks, to increase the number of devices served per base station (eNB). Our solution exploits information related to the device (e.g., device's source and destination address, payload size, etc.), that is already known to the eNB serving the device. Furthermore, we consider information related to the delivered content (e.g., Time To Live (TTL)) to buffer payloads and send multiple payloads at once [14]. Finally, we exploit also possibility to cluster users in proximity of each other and let them exchange payload directly among themselves by means of device-to-device (D2D) communication in LTE-A (see [15][16] for more details). This way, we merge the content from several neighboring devices and send it to the eNB via single device, which is denoted as a cluster head. The proposed solution is designed with awareness of backward compatibility with 4G networks.

The rest of this paper is organized as follows. In the next section, we explain limitations of the LTE-A transmission protocol stack and introduce existing solutions coping with these limitations. In Section III, we describe novel cross-layer approach minimizing amount of communication overhead and enabling more devices to send their data of a small size. In Section IV, we compare the state of the art solutions with our proposed approach. In the last section, we summarize simulation results and provide future work plans.

II. Analysis of Overhead and Limits on Number of Communicating Devices in 4G

In this section, we analyze overhead and other limitations at each of bottom three layers of LTE-A radio interface of the device (physical, control, TCP/IP layers) imposed by protocol stack as shown in Figure 1. Each layer of the protocol stack adds its respective header. This header is added to user data coming from upper layers. The transmission protocol stack of LTE-A radio interface consists of: physical layer, Medium Access Control (MAC), Radio Link Control (RLC), Packed Data Convergence Protocol (PDCP) and TCP/IP. Each layer creates a Protocol Data Unit (PDU), which consists of a header of the layer and data coming from an upper layer denoted as

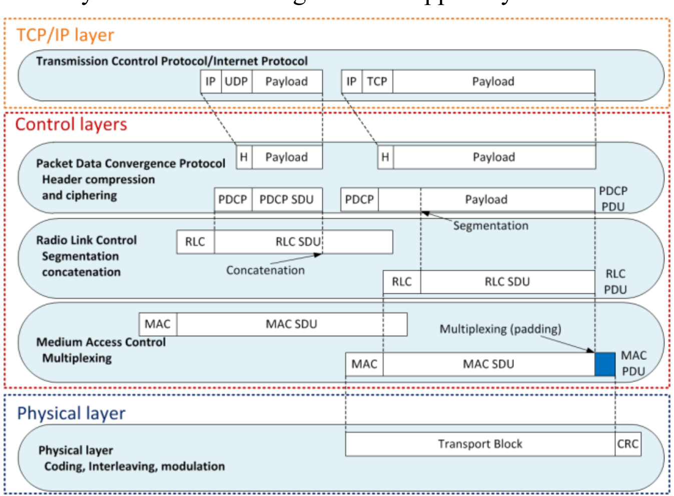

Figure 1. LTE-A transmission protocol stack at the device.

Service Data Unit (SDU). At the physical layer, MAC PDU is fitted in to the Transport Block (TB), and supplemented with Cyclic Redundancy Check (CRC). Finally, data is ready for the transmission.

Headers from all layers form overhead, which is added to the payload mainly for routing and QoS purposes. If we consider "small" payloads, the overhead is much larger than the payload itself as shown in Figure 2. This leads to inefficient communication with significant amount of redundant signaling overhead. For example, if we consider payload size of 20B, then overhead ratio is 200%.

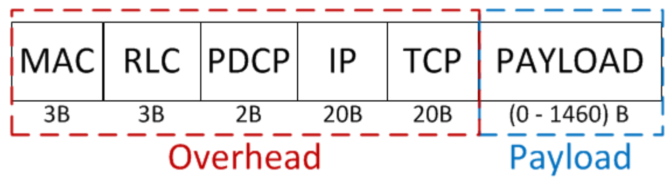

Figure 2. Overhead ratio example

In the following subsections, we analyze overhead at respective layers and we provide discussion of potential limitations of each layer on the number of devices. In addition, we also describe existing solutions for suppressing potential limitations at each layer.

A. Physical layer

The first limitation on the maximal number of served devices is implied by the physical layer as it defines the amount of bits the devices can transmit within a time period. On the radio, LTE-A defines a frame with a duration of 10ms. The frame is divided into 10 subframes (each with duration of 1ms). The minimum amount of bits allocated to one device is defined by the minimal amount of resources per subframe allocated to the device. Each subframe is composed of two resource blocks (RB) in time domain. In LTE-A, at least two time consecutive RBs (i.e., subframe) must be allocated to the device [17]. Depending on used Modulation and Coding Scheme (MCS), the device can send between 32 and 616 bits per subframe in one TB using QPSK and 64 QAM, respectively (see [18] for more details on relation between TB size, number of RBs, and MCS). The TB contains the device's payload and headers added by all layers as shown in Figure 1. If the device is willing to send less bits than the amount, which can be transmitted in RBs allocated to the device, the MCS for transmission can be lowered to reduce transmission error rate. However, from spectral efficiency point of view, this approach is very inefficient and leads to wasting of radio resources.

To enable transmission of less than two RBs, LTE-M has been proposed. The LTE-M aims on the MTC and reduces transmission bandwidth to enable use of single RB (i.e., a half of the subframe) [19][20]. However, this requires to use Generalized Frequency Division Multiple Access (GFDM) (see more details in [21]) for multiplexing instead of Single Carrier Frequency Division Multiple Access (SC-FDMA), which is defined for uplink in LTE-A.

Next limit for the uplink transmission originates from the Random Access Procedure (RAP), which is used to initiate communication with the eNB. The procedure consists of the randomly selected preamble sent by the device to identify itself over the Physical Random Access Channel (PRACH). As there is a limited number of preambles to distinguish each device, collision may occur at the PRACH [8]. Collision probability can be reduced by use of extended access barring (EAB), which is barring communication of low-priority devices. Results from analysis in [22] show that the EAB decreases collision probability but with the cost of increased delivery delay, which prohibits its use for our purposes. Different way to avoid the collisions is to reduce the number of the RAP by buffering of several payloads from the device [12] as shown in Figure 3. Instead of utilizing the RAP each time the device has a payload to be send (Figure 3a) the RAP is used once for transmission of multiple buffered payloads (Figure 3b). Nevertheless, buffering must respect delay constraints of each type of payload. Another way to overcome the PRACH limitation is to dynamically allocate more resources to the PRACH [23]. On one hand, it enables more devices to use RAP. On the other hand, this consumes resources commonly allocated to the device for communication. Consequently, a higher number of the devices can be able to access radio resources, but these resources might not be available to all of them in required quantity. Finally, the number of devices being able to transmit required payload by single eNB might not be increased sufficiently.

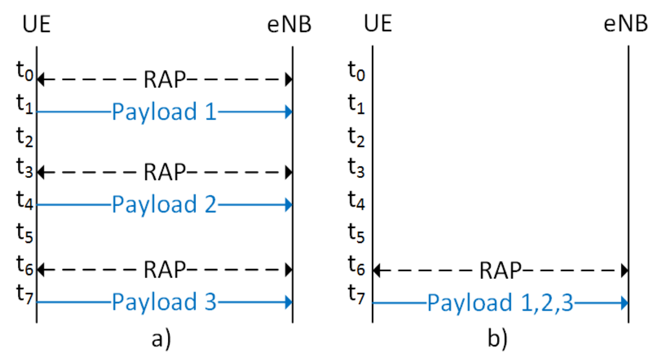

Figure 3. Transmission of payloads by one device in case of (a) conventional transmission without buffering, and (b) with buffering.

B. *Control layers*

After successful RAP, the device has to be scheduled in order to transmit its payload. In general, three options of scheduling are known: persistent, semi-persistent, or non-persistent (or dynamic) [24].

The persistent scheduling allocates resources to the device for a given period defined by the number of Transmission Time Intervals (TTI). An advantage is that the resources are allocated once for the period of multiple TTI, which leads to transmission of less signaling overhead. However, if the device is not transmitting any data in some TTIs, its RBs cannot be reallocated to another device and these RBs are wasted.

In case of non-persistent scheduling, the number of devices scheduled in one TTI is limited to 10 due to the limitations imposed by the Uplink grant (UL grant) information carried in Physical Downlink Control Channel (PDCCH) [7]. The limit of 10 devices is due to the number of resources available for the UL grant. To overcome the limit of 10 devices scheduled by the PDCCH, 3GPP Release 11 introduces Enhanced Physical Downlink Control Channel (EPDCCH) [18][25]. The EPDCCH overcomes the limitation of PDCCH by utilizing more resources from Physical Downlink Shared Channel (PDSCH) (used for user data transmission in downlink) for the purposes of the UL grant transmission.

The last type of scheduling is semi-persistent. The semi-persistent scheduling periodically allocates RBs for the device. This is used for VoIP as it has deterministic payload size and regular periodicity of transmission [26]. Since the payloads in semi-persistent scheduling are periodical and of defined size, we can utilize this scheduling to overcome limitations of PRACH and PDCCH. However, this requires constant size of data transmitted by the device. If data is not of constant size, a part of resources has to be reserved for the dynamic scheduling to accommodate bits not fitting to the resources allocated by semi-persistent scheduling.

To schedule adequate number of resources to each device, the eNB can exploit knowledge of the device's buffer status (how many bytes are ready to be sent by the device) obtained via Buffer Status Report (BSR) message [27]. The BSR is send by the device in Logical Channel ID field within MAC header. The BSR is send if: i) new data is in buffer of the device, ii) the eNB requests BSR, or iii) there would be more padding bits in MAC header than the length of the BSR itself. However, the BSR can report only specific ranges of payload sizes in the buffer as specified by 3GPP [28]. If less than 10 bytes are in the buffer of the device, the BSR informs the eNB that the device has between 1 and 10 bytes of payload in the buffer. Consequently, 10 bytes are allocated to the device. However, 10 bytes allocated to the device can be more than what the device actually requires. This, then, leads to wasting of resources and less devices can be served.

C. *TCP/IP layer*

At the TCP/IP layer, the payload of device is encapsulated in the TCP and IP to enable communication through the IP based networks. The TCP header is typically of 20 bytes, whereas a size of the IP header depends on version of the IP used for communication. For the IPv4 and IPv6 headers, 20 and 40 bytes are required, respectively. The TCP/IP header can be compressed by the ROHC [11]. The ROHC avoids transmission of full TCP/IP headers if the device sends multiple packets to the same destination. Note that the ROHC can be used only for point-to-point connections. The ROHC sends only the dynamically changing parts of the TCP/IP headers to reduce the overhead. The ROHC can work in three modes: Unidirectional, Bidirectional Optimistic, and Reliable. In case of uplink connection without the need of correct delivery acknowledgement, the ROHC works in Unidirectional mode. Other modes utilize downlink for transmission of additional signaling (acknowledgement of ROHC signaling), thus, we focus on Unidirectional mode only. In this mode, the device (after a given number of packets in a given state) switches periodically between one of the three ROHC states [11]: the Initialization and Refresh (IR), the First Order (FO) and the Second Order (SO). In each state, the ROHC sends different signaling with different size in order to provide sufficient robustness. This periodic switching between the states causes problems to scheduling as not all resources can be scheduled using semi-persistent scheduling. Part of the resources has to be reserved for non-persistent scheduling to send the bits not fitting to the resources allocated by semi-persistent scheduling.

III. CROSS-LAYER OVERHEAD OPTIMIZATION

All above-mentioned existing solutions are not efficient for transmission of small payloads in future mobile networks because of the ratio of the overhead and data. Thus, we propose to combine all available information related to the device and known to the eNB across layers in order to further reduce the overhead beyond ROHC [11]. Description of the proposed scheme is divided into two subsections. The first subsection provides high level overview of the proposed concept, while in the second subsection, we describe details of the proposed scheme and related signaling. The proposed signaling is designed to ensure backward compatibility with LTE-A networks.

A. Concept of cross-layer optimization

In this subsection, we describe concept of the proposed scheme. Our objective is to reduce overhead and keep its amount constant for each transmission in order to simplify scheduling. Therefore, we target to use semi-persistent scheduling without the need for additional resource allocation using non-persistent scheduling. To enable semi-persistent scheduling we propose new signaling (described latter) that enables the device to inform the eNB about the payload size in more precise way than the BSR. Overhead is further reduced by using buffering and enabling collection of payload from more devices. Buffering allows us to send multiple payloads per single signaling message. However, as mentioned before; we have to respect the TTL of the payload.

To send more payloads at once, clustering concept is exploited. The clustering enables to form clusters of nearby devices. For each cluster, a cluster head is selected out of all devices in the cluster. The cluster head collects payloads from the devices within the cluster and transmits them to the eNB. The clustering is further enhanced by buffering as shown in Figure 4. The cluster head (in Figure 4 denoted as *cl_head*) buffers payloads from the devices within the cluster. The devices inform the cluster head about TTL of their payloads in order to schedule transmission of individual payloads properly. This information is delivered from the device to the cluster head by means of D2D communication. To transmit buffered payloads to the eNB, signaling message is added and sent by the cluster head. The scheme merging new signaling, buffering, and clustering is labeled as Cross-layer Optimization (CLO).

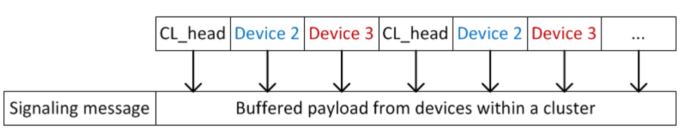

Figure 4. Principle of buffering within cluster.

Both clustering and buffering reduce the number of transmissions from devices. Clustering reduces the number of transmitting devices in the space domain (devices within a specific area transmit as one device to the eNB) while buffering in the time domain (device transmits once per multiple TTI). Therefore they can be considered complimentary.

In Figure 5, we show high-level overview of the proposed approach used for collection of payloads from devices in the network. We assume a single eNB to which all devices are connected, either directly or via the cluster head. In this paper, we assume basic clustering to show lower-bound of the gain introduced by our proposed scheme. Clustering is, therefore, based on distance (cl_{dist}). It means that the cluster is formed as a set of devices with mutual distance up to cl_{dist}. Advanced clustering approach can further improve performance, but it is left for future research. In Figure 5, DEV denotes device and represents common user's device, such as smartphone or tablet, as well as a sensor or a machine. The DEV_4 and DEV_8 are the cluster heads of Clusters 1 and 2, respectively, as they are closest to the eNB. The DEV_5 and DEV_6 are not members of any cluster as they are not in vicinity of other devices. These devices can be seen as the cluster heads of their own clusters with only themselves in the cluster. If a new *DEV* would require transmission of its payload within the proximity of DEV_5 or DEV_6, it could join either DEV_5 or DEV_6 and form a new cluster together.

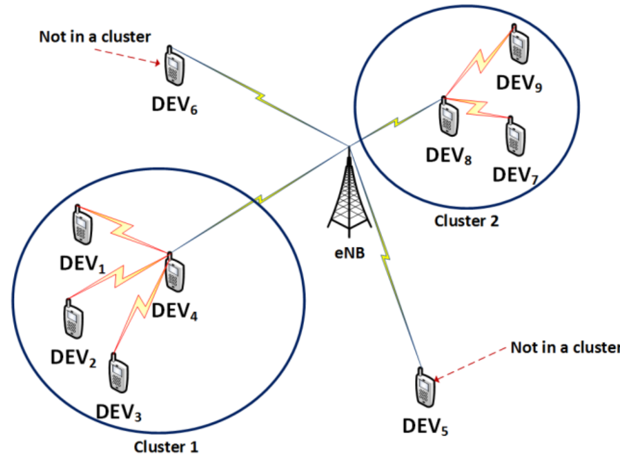

Figure 5. Scenario of the proposed approach for collection of information from devices.

B. Management of the proposed scheme

In this subsection, we describe management procedure of the CLO as shown in Figure 6. This figure shows a procedure used by each device, which wants to start sending its payloads. The procedure begins with checking whether there is a cluster head in the proximity of the device willing to transmit data. Based on this checking, there are three options for the device: a) the device becomes the cluster head if there is no cluster head in the vicinity, b) the device joins existing cluster, or c) the device is selected as the new cluster head for existing cluster. Communication between devices within the cluster exploits D2D communication. Using D2D communication, the devices form and manage clusters and transmit their payloads to the cluster head [15]. In case (a), where no cluster exists in the device's vicinity; the device buffers its payloads (respecting TTL of the content) and then starts RAP to obtain radio resources. Afterwards, the control message and payload are sent. Finally semi-persistent scheduling is initiated and the device sends buffered payloads. In case (b), the device joins existing cluster head and starts transmission of the payloads to the cluster head using D2D communication. In the case (c), when the device is selected as the new cluster head (the device is closer to the eNB than the existing cluster head), all devices within the cluster are informed about new cluster head. This information is issued by the former cluster head. After this, the device, which becomes the cluster head starts receiving payloads from the devices within the cluster and initiates RAP. Then, the cluster head sends control message and initiates semi–persistent scheduling. Finally, the cluster head transmits collected payloads to the eNB.

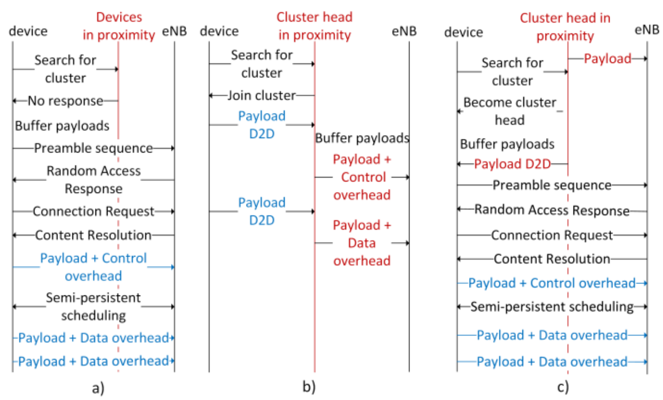

Figure 6. Procedure for the device beginning its transmission of payload in case of (a) no cluster in proximity, (b) joining cluster in proximity, (c) becoming cluster head

The CLO defines four types of signaling messages in order to replace ROHC (TCP/IP). Two types of control messages are used for initial transmission of the stand-alone device (option *a* in Figure 7) and cluster head (option *b*). The other two options (*c* and *d* in Figure 7) are used for transmission of the data from the stand-alone device (option *c*) and cluster head (option *d*). Using the control messages, the eNB initiates a record for the further transmissions of the device. This record is stored in a database in the eNB and is used to reconstruct the full TCP/IP header if the device's payload designation is in the Internet. The control message is always send as the first message. Then, data message is send in the subsequent transmissions. The first field in the proposed signaling messages is *D/C*. It specifies the type of the message in order to distinguish between data and control messages. The second field, *U/CL*, denotes whether the device transmits data to the eNB by itself or via the cluster head. These first two fields are the same for all four types of messages. Following fields in the signaling messages are defined depending on the type of message. Fields *DEST* and *SRC* denote destination and source address for the payload, *SEQ* is the sequence number of the transmitted payload. Information about the payload size is carried in *Payload size* field. The last field of each message is Cyclic Redundancy Check (*CRC*), which ensures correct delivery of the signaling message. For communication within the cluster, *i* identifies each device and ranges from 1 to the number of devices within the cluster (*N_DEV*). The flag *A/S* is used to inform the eNB, that payload of each device within the cluster is included, or if the payload from selected devices is included. If payloads from not all devices are included, filed *Bitmap* is included. This field identifies devices, from which the payloads are being transmitted.

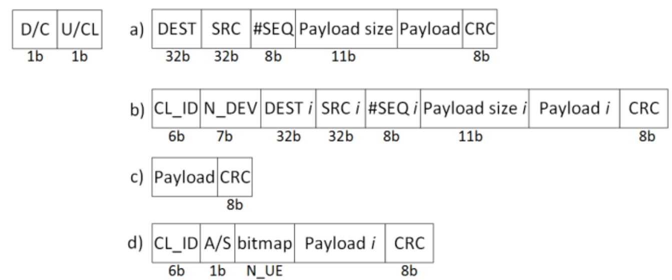

Figure 7. Proposed signaling messages enabling cross-layer optimization of frequent transmission of small payloads

To assure correct order of transmitted packets and reception of every transmitted packet if this is required by the service or application, we utilize Hybrid Automatic Repeat request (HARQ) and Automatic Repeat reQuest (ARQ) to check and repair received data. This enables to send constant message size after the control transmission and simplifies the semi-persistent scheduling. Thus, we determine resource allocation once and we do not use non-persistent scheduling even if reliable delivery of data is required as the ROHC does.

IV. Performance Evaluation

In this section, models and scenario for performance evaluation are defined. The evaluation is carried out by means of simulations in MATLAB.

A. Simulation scenario and models

Major parameters of the simulation, presented in TABLE I, are in line with recommendations defined by 3GPP [29]. We also follow parameters of the physical layer and frame structure for LTE-A mobile networks defined in the same document.

Signal propagation is modeled according to [29]. The devices are static in the simulation, thus, we consider MCS of each RB in the frame to be the same. Within the simulation area, devices (representing any type of device generating frequently small payloads) are randomly deployed with uniform distribution at the beginning of the simulation. In the center of the simulation area, one eNB is placed. We use 10 MHz bandwidth for uplink (i.e., 50 RBs per subframe). We assume that PUCCH, PRACH, reference and sounding signals occupy 20% of uplink resources [18][30]. In the simulation, the devices generate payload every 100ms. This interval is further referred to as a duty cycle. We assume no buffering, buffering of two payloads and buffering of three payloads. For clustering, the cluster radius (cl_{dist}) is set to 20, 40, and 50 meters. In simulation, we utilize semi-persistent scheduling with BSR. Thus, each device is allocated with enough resources for its transmission if the resources are available. We find maximal number of devices served by single eNB by continuous increasing of the number of devices until the eNB is able to schedule resources for all devices.

TABLE I. SIMULATION PARAMETERS.

Parameter	Value
Simulation area	400m x 400m
Carrier frequency	2000 MHz
UL bandwidth	10 MHz
LTE-A PHY layer overhead 10 MHz	20%
IP overhead (version IPv4)	20 B
Overhead of RCL/MAC/PDCP/TCP layers	2/3/2/20 B
ROHC signaling send in FO state	6 B
Scheduler	BSR, semi-persistent
D2D communication	outband
Cluster radius (cl_{dist})	20/40/50 m
Tx power of device	10 dBm
Number of eNB	1
Shadowing factor	6 dB
Number of interferers	10
Simulation step	1 ms
Device payload generation duty cycle	100 ms
Simulation time/Number of simulation drops	100 s / 50 drops

B. Simulation results

In this section, we analyze the number of devices that can be served in mobile networks and ratio of the signaling overhead. We compare the proposed CLO with two schemes: i) scheme without any overhead compression, i.e., sending full TCP/IP overhead with a size of 40 bytes (labeled as NC in following figures); and ii) the ROHC in the FO state [11]. The ROHC FO is selected instead of the ROHC SO as ROHC FO is send 5 times every 100 packets [11]. The ROHC in the IR state is not shown as its signaling is larger than for the NC [11]. We further include also results for our proposed signaling replacing of the ROHC but without buffering and clustering. This scheme is denoted as overhead reduction (OR) in all following figures. The LTE-M is not considered for performance comparison as it adopts different multiplexing and it is not backward compatible with the 4G.Note that the overhead in our simulations contains overhead introduced by all layers (TCP/IP, PDCP, MAC, and RLC).

In Figure 8, we show how many devices can be served by one eNB. With increasing payload size, the number of served devices decreases because more resources are required for the transmission of all devices. The NC enables eNB to serve the lowest amount of devices (1537 devices for 10 bits payloads) comparing to other schemes. The ROHC FO roughly doubles the number of served devices against the NC (gain up to 110.5%). The proposed OR, which is based only on optimization of overhead of ROHC without considering clustering and buffering, improves the number of served devices by additional up to 40% comparing to the ROHC FO. Far the best performance is achieved by the proposed CLO. The gain is more significant for small payloads as payloads from more devices can be buffered together and sent within one transmission. The CLO enables to serve more than 65 000 devices transmitting 10 bits payloads. It corresponds to improvement in the number of served devices up to 22.8 times and 16.3 times compared to the ROHC FO and the OR, respectively. This gain is a result of sending more payloads from nearby devices in one message, which is achieved by the combination of clustering and buffering. Moreover, by using clustering, only the devices closest to the eNB (using higher MCS) transmit and, thus, less resources are required for the transmission. From the results, we see that existing solutions NC and ROHC FO are not suitable for the 5G as the number of devices served if these approaches are adopted is lower than the expected number of devices connected to one eNB in 5G (10 000 to 100 000 devices, see [2]). However, our solution with only basic, not optimized, clustering enables to serve the required number of devices even for 10 MHz bandwidth, which is much lower that the bandwidth expected for 5G. Further increase in the number of served devices by our proposed approach can be reached by simple extension of bandwidth. This also shows that we can serve the required amount of devices with lower density of eNBs. Hence, the overall cost of the network deployment required for IoT or MTC can be lowered.

In Figure 9, we show the impact of the number of buffered payloads and cluster radius on the number of the served devices. Increase in the payload size leads to decrease in the number of served devices as more resources are required for the transmission. Impact of increasing number of buffered payloads and increasing cluster radius on the number of served devices is negligible as the difference is less than 7.2%. This 7.2% improvement represents further increase in gain with respect to ROHC FO in the number of served devices so that the proposed CLO increases the number of served devices by up to 24.4 times for the most efficient buffering and clustering combination (cl_{dist}=50m, 3 payloads). Impact of clustering is limited by cluster size (number of devices within the cluster) as a large

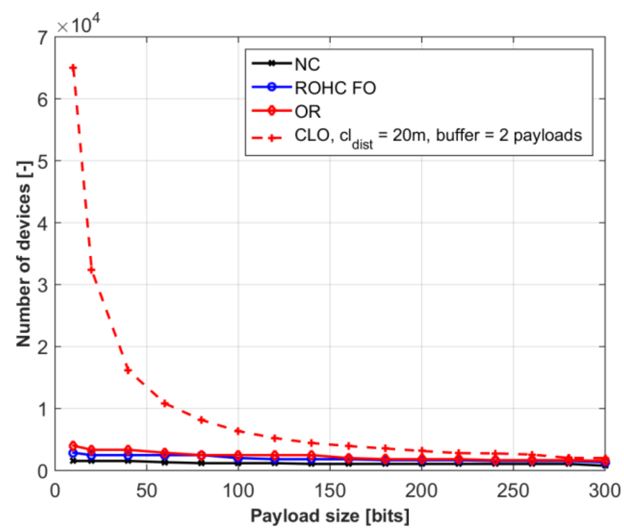

Figure 8. The number of served devices transmitting frequently small payloads.

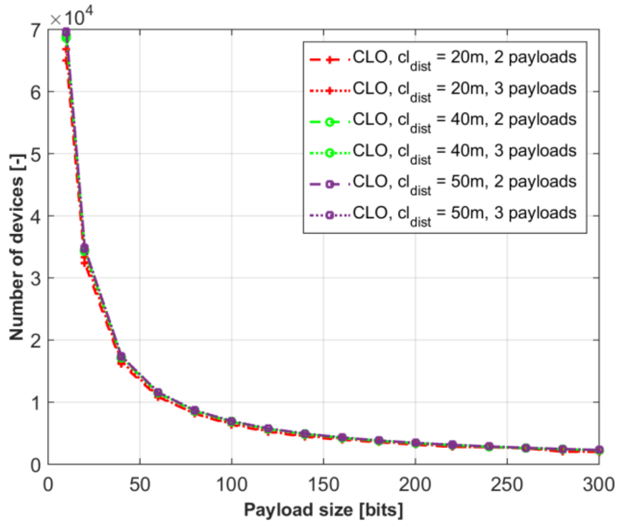

Figure 9. Impact of clustering and buffering on the number of devices.

cluster leads to the same problem as large number of devices. Impact of buffering is, on the other hand, limited by a need to respect TTL of the transmitted data.

In Figure 10, we compare overhead ratio, i.e., the ratio between the overhead and the payload. As expected, the overhead ratio decreases with increasing payload size. The ROHC FO decreases overhead ratio by 66% comparing to the NC. Further decrease in the overhead ratio is introduced by the OR. The OR reduces the overhead ratio by 48.5% comparing to the ROHC FO. However, still, the OR leads to significant ratio of the overhead to the payload (660% for 10 bits payload). Significant improvement is reached by the CLO, which reduces the overhead ratio to less than 10.5%. It corresponds to up to 68 times reduction comparing to the OR, up to 132 times comparing to the ROHC FO and up to 390 times comparing to the NC.

In Figure 11, we show the impact of parameters of the CLO (number of buffered payloads and cluster size) on the overhead

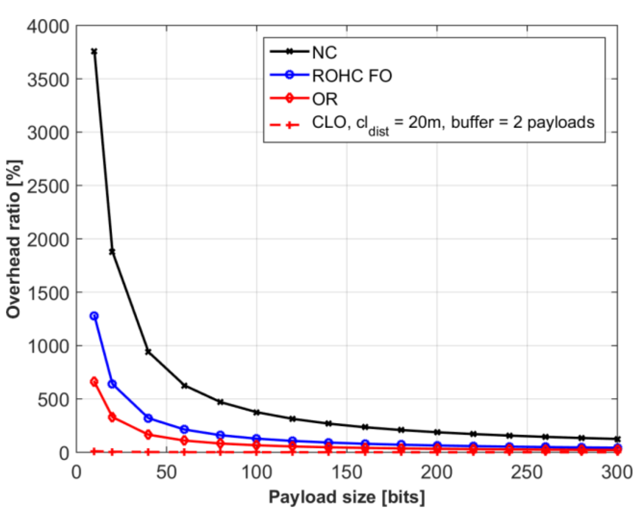

Figure 10. Overhead ratio by the proposal and competitive schemes.

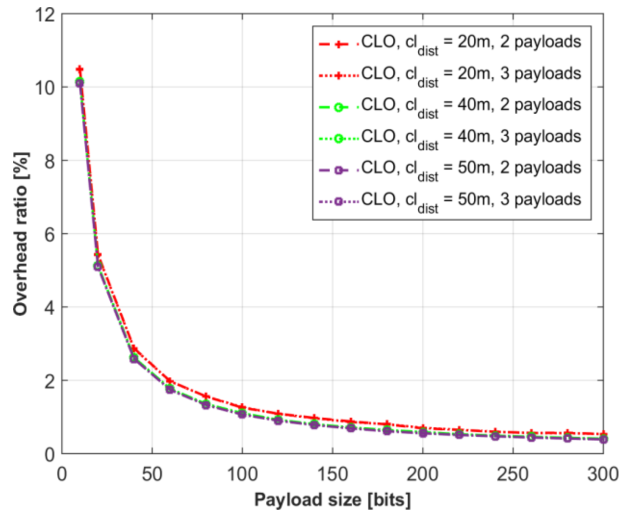

Figure 11. Impact of clustering and buffering on overhead ratio

ratio. Difference in the overhead ratios between configurations of parameters is minimal like for the number of devices. Improvement by using various cluster sizes or numbers of buffered payloads is less than 0.3%.

In Figure 12, we show the impact of different duty cycle time (interval between two consequent payloads generated by one device) on the number of the served devices for payload of 100 bits. The number of served devices increases linearly with duty cycle as the devices generate payload less often.

V. Conclusion

In this paper we have analyzed limitations for the uplink transmission of small payloads in 4G mobile networks at the physical, control, and TCP/IP layers. We have proposed a cross-layer solution to increase the number of devices that can be served by one eNB. The solution combines reduction of the TCP/IP overhead with buffering and clustering concepts in

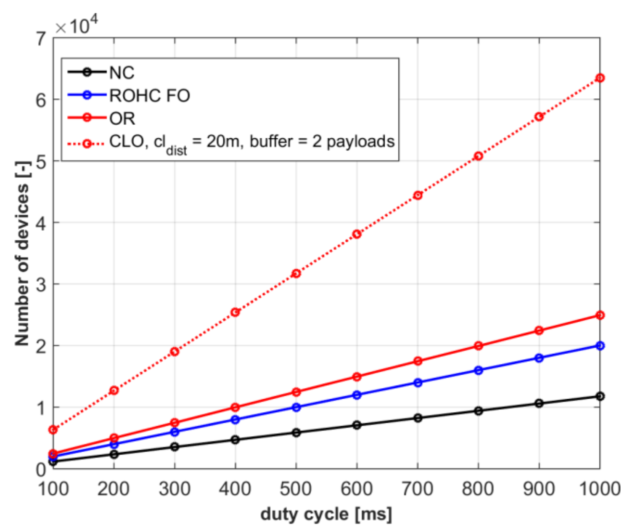

Figure 12. Number of served devices for different duty cycles for payload of 100 bits.

order to maximize efficiency of the transmission of small payloads by a high number of devices such as sensors, machines, or conventional user devices. The proposal enables to serve more than up to 65 000 devices by one eNB in case of a 10 MHz bandwidth. This represents 24.4 times increased number of devices with respect to the state of the art solutions. Even if the proposed solution is compatible with existing 4G networks, it enables to serve the number of devices expected to be connected in 5G networks only with 10 MHz bandwidth.

In the future, we plan to consider mobility of devices, optimize clustering, and consider information density to further improve limits for the number of devices served by one eNB.

ACKNOWLEDGMENT

This work has been supported by the grant of Czech Technical University in Prague No. SGS SGS13/199/OHK3/3T/13.

REFERENCES

[1] J. G. Andrews, S. Buzzi, Ch. Wan, S. V. Hanly, A. Lozano, A. C. K. Soong, J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications,* vol.32, no.6, pp.1065-1082, June 2014.

[2] NetWorld2020, "5G: Challenges, Research Priorities, and Recommendations", 2015.

[3] A. Imran, A. Zoha, "Challenges in 5G: how to empower SON with big data for enabling 5G", *IEEE Network,* vol.28, no.6, pp.27-33, Nov.-Dec. 2014.

[4] F. Lobillo, Z. Becvar, M. Puente, P. Mach, F. Lo Presti, F. Gambetti, M. Goldhamer, J. Vidal, A. Widiawan, E. Calvanesse, "An architecture for mobile computation offloading on cloud-enabled lte small cells", *IEEE Wireless Communications and Networking Conference Workshops (WCNCW),* 2014.

[5] O. Munoz-Medina, A. Pascual-Iserte, J. Vidal, "Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading", *IEEE Transactions on Vehicular Technology,PP (99)*, 2014.

[6] SK Telecom, "SK Telecom's View on 5G Vision, Architecture, Technology, and Spectrum", 2014.

[7] H. Harri, T. Antti, "LTE for UMTS : OFDMA and SC-FDMA Based Radio Access" *Wiley*, 2009.

[8] R. G. Cheng, W. Chia-Hung, T. Shiao-Li, F. Ren, "RACH Collision Probability for Machine-Type Communications" *IEEE 75th Vehicular Technology Conference (VTC Spring)*, vol., no., pp.1-5, 6-9 May 2012.

[9] Nokia, "What is going on in Mobile Broadband Networks? Smartphone Traffic Analysis and Solutions", 2015.

[10] 3GPP TS 36.300 V12.5.0, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2, (Release 12)", Technical specification, *3rd Generation Partnership Project (3GPP),* 2015.

[11] R. Fracchia, C. Gomez, A. Tripodi, "R-RoHC: A Single Adaptive Solution for Header Compression" *IEEE 73rd Vehicular Technology Conference (VTC Spring),* pp.1-5, 15-18 May 2011.

[12] K. Zhou, N. Nikaein, "Packet aggregation for machine type communications in LTE with random access channel" *IEEE Wireless Communications and Networking Conference (WCNC),* pp.262-267, 7-10 April 2013.

[13] J. Yunjie, L. Ming; Z. Song; D. Pengtao, "A Clustering Routing Algorithm Based on Energy and Distance in WSN" *International Conference on Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM),* pp.9-12, 5-6 March 2012.

[14] G. Corbellini, E. Calvanese Strinati, A. Duda, "DA-MAC: Density Aware MAC for Dynamic Wireless Sensor Networks", 22nd IEEE Personal Indoor Mobile Radio Communications (PIMRC'11 - LPAN), Toronto, Canada, September 2011.

[15] A. Asadi, V. Mancuso, "Energy efficient opportunistic uplink packet forwarding in hybrid wireless networks". *In Proceedings of the fourth international conference on Future energy systems (e-Energy '13). ACM*, New York, NY, USA, 261-262, 2013.

[16] P. Mach, Z. Becvar, Vanek T., "In-Band Device-to-Device Communication in OFDMA Cellular Networks: A Survey and Challenges", *IEEE Communications Surveys & Tutorials*, vol. PP, no. 99, pp.1-38, 2015.

[17] D. C. Dimitrova, H. van den Berg, R. Litjens, G. Heijenk, "Scheduling strategies for LTE uplink with flow behaviour analysis.", *In Fourth ERCIM Workshop on eMobility*, May 31, 2010, Lulea, Sweden (pp. 15-26).

[18] 3GPP TS 36.213 V12.5.0, "Evolved Universal Terrestrial Radio Access (E-UTRA) Physical layer procedures (Release 12)", Technical specification, *3rd Generation Partnership Project (3GPP)*, 2015.

[19] Deliverable D33, "Final report on LTE-M algorithms and procedures", *EXALTED project deliverable*, www.ict-exalted.eu, 2012

[20] R. Ratasuk, N. Mangalvedhe, A. Ghosh, B. Vejlgaard, "Narrowband LTE-M System for M2M Communication", *IEEE 80th Vehicular Technology Conference (VTC Fall),* pp.1-5, 14-17 Sept. 2014.

[21] N. Michailow, M. Matthe, I. S. Gaspar, A. N. Caldevilla, L. L. Mendes, A. Festag, G. Fettweis, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks", *IEEE Transactions on Communications,* vol.62, no.9, pp.3045-3061, Sept. 2014.

[22] R. G. Cheng, J. Chen, D. Chen, C. Wei, "Modeling and Analysis of an Extended Access Barring Algorithm for Machine-Type Communications in LTE-A Networks" *IEEE Transactions on Wireless Communications*, vol.14, no.6, pp.2956-2968, June 2015

[23] C. Úbeda, S. Pedraza, M. Regueira, J. Romero, "LTE FDD Physical Random Access Channel Dimensioning and Planning" *IEEE Vehicular Technology Conference (VTC Fall),* pp.1,5, 3-6 Sept. 2012.

[24] N. Abu-Ali, A. E. M. Taha, M. Salah, H. Hassanein, "Uplink Scheduling in LTE and LTE-Advanced: Tutorial, Survey and Evaluation Framework" *IEEE Communications Surveys & Tutorials*, vol.16, no.3, pp.1239-1265, Third Quarter 2014.

[25] S. Ye, S. H. Wong, C. Worrall, "Enhanced physical downlink control channel in LTE advanced Release 11", *IEEE Communications Magazine*, vol.51, no.2, pp.82-89, February 2013.

[26] J. Puttonen, T. Henttonen, N. Kolehmainen, K. Aschan, M. Moisio, P. Kela, "Voice-Over-IP Performance in UTRA Long Term Evolution Downlink" *IEEE Vehicular Technology Conference, VTC Spring* , pp.2502-2506, 11-14 May 2008.

[27] K. V. Pradap, V. Ramachandran, S. Kalyanasundaram, "Uplink Buffer Status Reporting for Delay Constrained Flows in 3GPP Long Term Evolution" *IEEE Wireless Communications and Networking Conference*, pp.1,6, 5-8 April 2009.

[28] 3GPP TS 36.321, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 12)", Technical Specification, *3rd Generation Partnership Project (3GPP)*, 2015.

[29] 3GPP TR 36.814, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)" Technical report, *3rd Generation Partnership Project (3GPP)*, 2010.

[30] T. Abe, "3GPP Self-evaluation Methodology and Results", *NTT DoCoMo*, 2009.