Marwat *et al. EURASIP Journal on Wireless Communications and Networking* (2016) 2016:110
DOI 10.1186/s13638-016-0598-0

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Data aggregation of mobile M2M traffic in relay enhanced LTE-A networks

Safdar Nawaz Khan Marwat[1], Yasir Mehmood[2*], Carmelita Görg[2] and Andreas Timm-Giel[3]

Abstract

Machine-to-machine (M2M) communication is becoming an increasingly essential part of mobile traffic and thus also a major focus of the latest 4G and upcoming 5G mobile networks. M2M communication offers various ubiquitous services and is one of the main enablers of the Internet-of-things (IoTs) vision. Nevertheless, the concept of mobile M2M communication has emerged due to the wide range, coverage provisioning, high reliability as well as decreasing costs of future mobile networks. Resultantly, M2M traffic poses drastic challenges to mobile networks, particularly due to the expected large number of devices sending small-sized data. Moreover, mobile M2M traffic is anticipated to degrade the performance of traditional cellular traffic due to inefficient utilization of the scarce radio spectrum. This paper presents a novel data aggregation and multiplexing scheme for mobile M2M traffic and thus focuses on the latest 3GPP (3^{rd} Generation Partnership Project) tong-term-evolution-advanced (LTE-A) networks. 3GPP standardized layer 3 inband Relay Nodes (RNs) are used to aggregate uplink M2M traffic by sharing the Physical Resource Blocks (PRBs) among several devices. The proposed scheme is validated through extensive system level simulations in an LTE-A based implementation for the Riverbed Modeler simulator. Our simulation results show that besides coverage extensions, RNs serve approximately 40 % more M2M devices with the proposed data multiplexing scheme compared to the conventional without multiplexing approach. Moreover, in this paper an analytical model is developed to compute the multiplexing transition probabilities. In the end, the simulation and analytical results of multiplexing transition probabilities are compared in order to analyze the multiplexing scheme.

Keywords: Machine-to-machine, 4G, 5G, Internet-of-things, 3^{rd} Generation partnership project, Long-term-evolution-advanced

1 Introduction

The contest towards developing next-generation 5G networks has already started. Resultantly, an increasing interest is shown by several operators, service providers, automotive industry as well as academia to provide a general framework for 5G. On the one hand, the motivation behind the increasing interest in 5G is the thousand fold rise in mobile traffic expected in 2020 [1]. On the other hand, supporting future massive machine-to-machine (M2M) communication is also one of the major focuses of 5G networks. Thus, massive M2M communication is considered as one of the potential horizontal topics in 5G networks [2]. This increasing importance is due to the fact that the collaborative communication between intelligent systems through mobile networks has recently achieved a significant importance. One of the resultant emerging domains is mobile M2M communication. M2M communication is a pattern which identifies the evolving paradigm of interconnected devices communicating with each other without or with limited human interaction [3]. Additionally, advancements in mobile M2M communication have also played a prominent role in improving human life standards by enabling numerous ubiquitous solutions towards the Internet-of-things (IoT) vision. The idea behind the increasing use of M2M devices is that devices are more valuable when these are networked. Similarly, the network attains more importance when the number of connected machines is larger.

The characteristics of M2M traffic differs from existing human-to-human/human-to-machine (H2H/H2M) traffic. The first detailed characterizations of M2M were performed in [4], in which the authors collected traffic

*Correspondence: ym@comnets.uni-bremen.de
[2]Communication Networks, University of Bremen, Bremen, Germany
Full list of author information is available at the end of the article

Springer Open

© 2016 Marwat et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

data for 1 week from one of the major TIER-1 cellular service providers in the USA. The authors concluded that there is a significant difference between the characteristics of M2M and traditional mobile traffic. Their experimental results showed that a large part of the M2M traffic has a set of challenging characteristics such as small-sized data transmissions, group-based communication, time controlled, and delay-tolerant communication. Moreover, their findings showed that unlike conventional mobile traffic, M2M is an uplink dominant traffic. Additionally, existing mobile standards such as the Global System for Mobile Communications (GSM), the Universal Mobile Telecommunication System (UMTS), and LTE are providing services to numerous M2M applications. However, it is expected that the number of devices will approximately grow to 50 billion by the end of 2020, according to ERICSSON [5]. Thus, current mobile standards are expected to run out of capacity and will not be able to support such a large number of devices. Resultantly, mobile M2M communication may not be facilitated proficiently.

Cellular standards were particularly designed and optimized to support broadband traffic with a relatively small number of UEs (Users Equipments) per cell. To support a high peak data rate and enable high mobility, 3GPP introduced LTE-A in Release 10 with some enhancements such as carrier aggregation (CA), massive multiple-input-multiple-output (MIMO) [6, 7], Coordinated multi-point (CoMP), and relaying in Release 11 [8]. LTE-A introduces IP networks with improved system capacity, increased uplink and downlink peak data rates and scalable bandwidth. Therefore, LTE-A networks are expected to support future M2M data traffic [9]. However, the dilemma is that M2M traffic patterns vary in diverse application domains and in most of the applications, the M2M device data is only a few bytes [9–11]. Thus, it is quite crucial to achieve full spectrum efficiency in the presence of a large number of devices with small-sized data packets.

Researchers as well as standardization bodies are currently working on proposals to support mobile M2M traffic efficiently without degrading the performance of traditional users such as voice/data connections. The major issue related to mobile M2M traffic is to support a potentially large number of devices for which the mobile networks such as LTE and LTE-A are not yet optimized. 3GPP has started working on evolving LTE-A in order to support M2M traffic efficiently. Till now, several open issues as well as M2M network architectures have been proposed by 3GPP and ETSI (European Telecommunication Standard Institute) in [12, 13], respectively. Additionally, the Institute of Electrical and Electronics Engineers (IEEE) has standardized 802.16p (for vehicular applications) and 802.16.1b to modify existing standards for supporting future M2M capillary networks. Additionally, major efforts related to cost and access schemes were accomplished by the two Seventh Framework Programme (FP7) projects namely EXpAnding LTE for Devices (EXALTED) [3], and Achieving LOw-LAtency in Wireless Communications (LOLA) [14]. The Mobile and wireless communications Enablers for the Twenty-twenty (METIS 2020), one of the biggest EU (European Union) research projects also initiated to shape upcoming 5G networks by focusing on M2M communications. In METIS 2020, M2M is considered as a front line research topic by covering its radio link challenges [15]. However, new and advanced schemes are still required to support the myriad of devices simultaneously sending small-sized data.

The rest of the paper is structured as follows. Section 2 describes existing issues in the conventional radio resource utilization for M2M traffic. Then, Section 3 presents an overview of data aggregation. The system model for the proposed data traffic aggregation scheme is discussed in Section 4. The simulation environment along with parameter settings is given in Section 5. Section 6 presents the simulation results and discussions. In Section 7, an analytical model is developed to analyze the data aggregation scheme in terms of the packet arrival and departure probabilities (multiplexing transition probabilities). In the end, conclusions and outlook are drawn in Section 8.

2 Problem statements

The distinctive characteristics of M2M traffic pose serious challenges for cellular service providers and mobile network researchers. The most challenging problem is the expected large number of small-sized data generated by a large number of M2M devices in the near future. Moreover, it is also observed that in most M2M applications, the group behavior is similar to that of the botnet, see [11], as a large number of devices will access the network simultaneously. Besides, several M2M applications such as logistics, intelligent transportation, and mobile e-healthcare demand mobility support. Existing mobile standards are neither designed to handle small-sized data efficiently nor to support simultaneous access of thousands of devices. Consequently, this will lead to congestion over the radio link. Thus, the current mobile standards must be revised in order to support a large number of devices. According to the 3GPP standardization, the smallest resource element for LTE/LTE-A is the PRB with a dimension of 180 kHz in the frequency domain and 1 ms (millisecond) in the time domain. Moreover, a PRB can transmit several hundred bits under favorable channel conditions [16]. For instance, 712 bits are sent in a TTI (transmission time interval) with an MCS (modulation and coding scheme) of 26. However, allocating 1 PRB to

a single device can significantly degrade PRB utilization [11, 17–19].

3 Literature review

Data aggregation has remained a potential research topic particularly in wireless sensor networks. It is generally applied to reduce the number of sensor transmissions and thus to improve the network lifetime. In recent years, notable research efforts have been made for the topic of data aggregation. For instance, authors of [20, 21] proposed aggregation tree algorithms to find aggregation node in order to improve network lifetime. In [22], authors explored the benefits of data-centric routing algorithms and proposed a routing algorithm based on the tree and query model. Furthermore, authors in [23] enhanced the network bandwidth by selecting multiple suitable aggregation nodes through clustering algorithms which reduce the number of transmitted packets to the gateway.

In order to reduce energy consumption, bandwidth efficiency and increased system throughput, data aggregation collects and accumulates data packets from multiple nodes before transmitting to the next hop. For instance, an energy-aware aggregation algorithm is proposed in [24] in which the authors discussed through experimental analysis that data aggregation improves the network lifetime by 80 %. On the other hand, the authors of [24] indicated the drawbacks of data aggregation such as increased transmission time. Resultantly, authors in [25] studied two timing control methods which include periodic per-hop and periodic simple. In the former case, packets are aggregated until the number of buffered packets reaches a maximum defined threshold. Whereas in the latter case, the aggregation node retrieves packets until the expiry of the timer. Besides, authors in [26] analyzed the effect of bufferbloat on delay and system throughput in IEEE 802.11n wireless networks. Their experimental results showed that large buffers can degrade the fairness in multi-hop networks based on parking lot. Similarly, authors in [27] proposed a queue management scheme called WQM for wireless networks which adapts the size of the buffer based on measured link features and network load.

In literature, limited research efforts have been made to study the impacts of data aggregation in the context of mobile M2M communication. For instance, authors in [28] evaluated the performance of data aggregation in terms of energy efficiency, thus to increase the lifetime of capillary M2M networks. However, the authors neglected the emerging mobile M2M applications in their work. Besides, authors in [29] proposed a method of bundling M2M data packets at the macro station also called Donor eNBs (DeNBs) to reduce the risks of congestion in backbone networks. However, authors completely neglected the possible risks of congestion over the radio link which can be a bottleneck due to the finite frequency spectrum.

Due to the unique characteristics of mobile M2M traffic such as high density of devices simultaneously sending small-sized payloads, data aggregation can notably improve the radio resource utilization over the radio link, and thus can significantly reduce the risks of network congestion. However, to the best of the authors knowledge, the influence of data aggregation on the bandwidth utilization in the context of mobile M2M communication has not yet been addressed in literature. This paper presents a novel data aggregation and multiplexing scheme by considering both periodic simple and period per-hop control methods, as discussed in [25]. For this purpose, a recently 3GPP standardized wireless layer 3 inband RN is used for enhancing coverage and aggregating uplink M2M traffic.

4 System model

This section presents details of implementation and functioning of the proposed data aggregation and multiplexing scheme. Since the authors of [4] stated that M2M traffic is an uplink dominant traffic, therefore uplink M2M small-sized data packets are aggregated in this work. Moreover, in the simulation model, both *periodic simple* and *periodic per-hop* control methods discussed in [25] are implemented. Additionally, an analytical model is developed to analyze the proposed scheme in terms of multiplexing transition probabilities particularly in low-load scenarios. This is due to the fact that the multiplexing process does not affect packet delay as well as PRB utilization in high-load scenarios.

4.1 RNs implementation and scheduling

In this work, a fixed RN is implemented in the LTE-A model according to the 3GPP specifications given in [30]. Traffic from M2M devices located in the proximity of an RN is accumulated at the RN. According to 3GPP specifications [8], the access link (Uu) and the backhaul link (Un) antennas of the RN are assumed to be well separated in order to avoid self-interference. For instance, the Uu antenna towards M2M devices is assumed to be located inside a warehouse building and the Un antenna is placed outside the building to ensure RN inband operation without time division of the resources. At the DeNB MAC (medium access control), a channel and QoS aware (CQA) scheduler is implemented to schedule the RN along with the regular UEs/M2M devices. The time and frequency domain uplink scheduler are decoupled. Candidate lists for guaranteed bit rate (GBR) and non-GBR bearers are created in the time domain and are sorted according to the time domain packet scheduling (TDPS) metric values of the bearers. In the scheduling process at the DeNB MAC, the priorities are given to

UE/M2M device bearers which are placed at the top of the list.

$$W_{i,a}(t) = \frac{R_{min,a}}{R_{avgi,a}(t)} \frac{\tau_{i,a}(t)}{\tau_{max,a}} \rho_a(t) \quad (1)$$

$$W_{i,a}(t) = \frac{R_{min,a}}{R_{avgi,a}(t)} \quad (2)$$

The $W_{i,a}(t)$ parameter indicating the quality of service (QoS) weight of bearer **a** of user **i** at time **t** of the weighted proportional fair (W-PF) as given in (1). The $R_{min,a}$ is the bit rate budget (target data rate) and $\tau_{max,a}$ is the end-to-end delay budget (target delay) of QoS class of bearer **a**, $R_{avgi,a}(t)$ is the average throughput of bearer **a** and $\tau_{i,a}(t)$ is the packet delay of bearer **a** of UE **i**, $\rho_a(t)$ is a variable with value set to, e.g., 10 if $\tau_{i,a}(t)$ is above the threshold value of bearer **a** at time **t** (as the bearer packet delay has reached the threshold), otherwise equal to 1. The value 10 of $\rho_a(t)$ raises the metric value of bearer **a** by 10 times and ensures immediate scheduling.

In (2), the parameters related to bearer delay are eliminated from (1) due to the fact that in uplink data packet transmission, the information about packet delay at the packet data convergence protocol (PDCP) buffer of the UE (or RN) is not available at the DeNB. The buffer status report (BSR) provides information only about the length of the buffer at the eNodeB. Although, the length of the buffer can be used to estimate the delay [31], but since the scheduling scheme in the CQA scheduler is already designed to prioritize the delay-sensitive GBR bearers by creating a separate GBR bearer list, so the delay estimation is not critical. Hence, the delay estimation is not considered for the scheduling decisions. Furthermore, the RN MAC scheduler is designed to allocate radio resources to the UEs and M2M devices connected to the RN at the Uu link. The round robin scheduling approach is adopted for such type of M2M traffic. All the active devices ready to transmit packets are scheduled in a round robin fashion.

4.2 Proposed M2M data aggregation and multiplexing

In the proposed approach, small data packets are aggregated at the PDCP layer of the RN in order to maximize the multiplexing gain without aggregating the additional headers such as those from the PDCP, radio link control (RLC) and MAC, as shown in Fig. 1. The PDCP layer exists in the UE, DeNB as well as in the RN; it is a part of LTE air interface control and user plane protocols. The major functionalities and services of PDCP layer for the user plane include header compression and decompression, user data transfer, delivery of upper layer packet data units (PDUs) in sequence as well as retransmission of the lost PDCP service data units (SDUs), etc. On the other hand, the control plane services include ciphering and integrity protection and transfer of control plane data. The proposed aggregation and multiplexing scheme can also be applied either at the lower layers of the RN Uu interface such as RLC and MAC, or in the PDCP of RN Un interface. However, the multiplexing gain will be reduced due to the multiplexing of additional headers along with the original IP packets. An aggregation buffer B is created at the PDCP layer of the RN which aggregates the incoming packets according to the size of available transport block − RN Un protocol overhead, see Fig. 2. The RN Un protocol overhead includes GPRS tunneling protocol ($GTP = 12$ *bytes*), user datagram protocol ($UDP/IP = 28$ *bytes*), and layer 2 (4 *bytes*). Later, the aggregated packet is sent to the RN GTP over the Un link. The additional overheads such as the GTP, UDP/IP, PDCP, and RLC are added. Then, from the physical (PHY) layer of the RN Un interface, the aggregated packets are sent to the DeNB.

The proposed approach of multiplexing significantly improves the PRB utilization [32]. However, there are certain constraints regarding latency requirements of high priority M2M traffic such as emergency alerting. Each packet waits until the size of the buffer B reaches the

Algorithm 1: An overview of the proposed data aggregation and multiplexing algorithm in the RN PDCP.

Result: **Efficient utilization of PRBs among M2M devices**

initialization;
set expiry timer $T_{max} == 9$ *ms*;
set $B_{max} ==$ (*available TBS − RN Un overhead*);
set timer $T == 0$;
set multiplexing buffer $B == 0$;
schedule RN and allocate PRBs (e.g., 5 PRBS are set for RN to analyze multiplexing process);
schedule M2M devices within the coverage of RN for uplink transmission;
while *packet arrival == TRUE* **do**
  start multiplexing process based on the value of timer and the size of the multiplexing buffer;
  if $T < T_{max}$ && $B < B_{max}$ **then**
    accummulate incoming packet into buffer B;
    increment timer T;
  else
    re-assemble aggregated packet of size *available TBS − overhead* from buffer B;
    send large multiplexed packet to RN PHY via RN Un protocols;
    add RN Un protocols overhead;
    route multiplexed packet to DeNB in next TTI;
    reset timer T;
  end if
end while

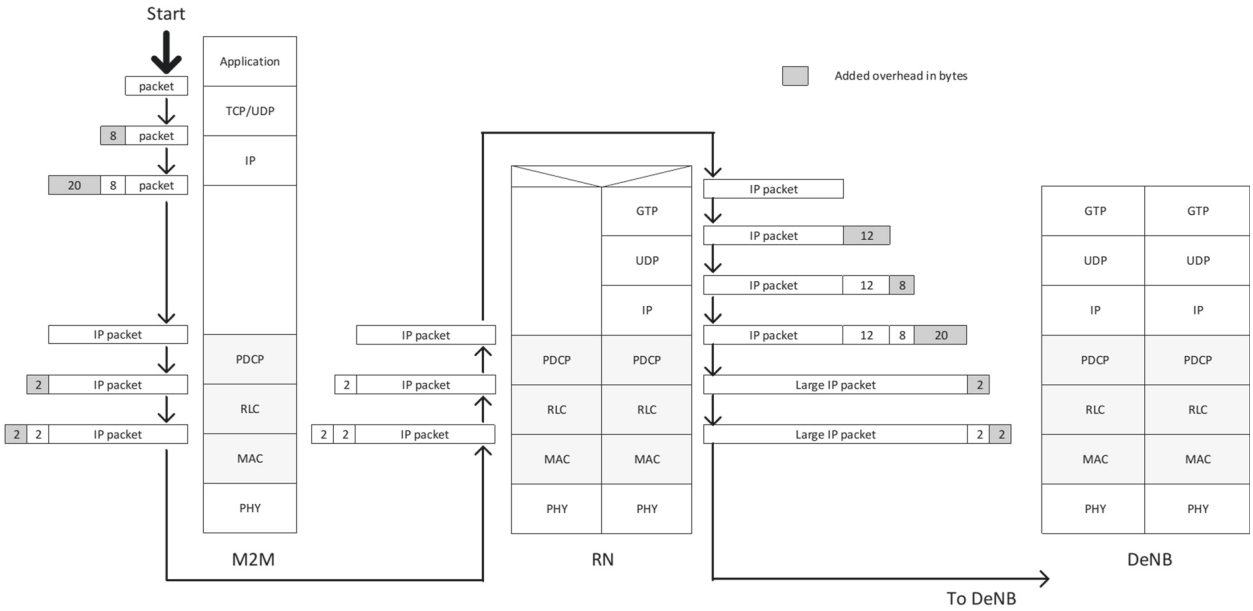

Fig. 1 An illustration of the packet flow and protocol stack of the M2M device, RN, DeNB and aGW

size of (TBS−Un protocol overhead). In a highly loaded scenario, the waiting time is not long due to a high arrival rate. However, in a low-load scenario, there is a comparatively longer delay due to the low arrival rate. Consequently, the performance of delay sensitive M2M applications such as e-healthcare and emergency alerting can be degraded. To tackle this issue, an expiry timer T_{max} is introduced. The timer is, e.g., set with a fixed value of 9 ms in the current implementation. This means that the buffer serves the aggregated packets after 9 ms at the latest. The value of the timer could also be adaptive, i.e., it can change its value adaptively according to priorities of the incoming packets. For this purpose, the algorithm must be fully aware of various priorities of M2M

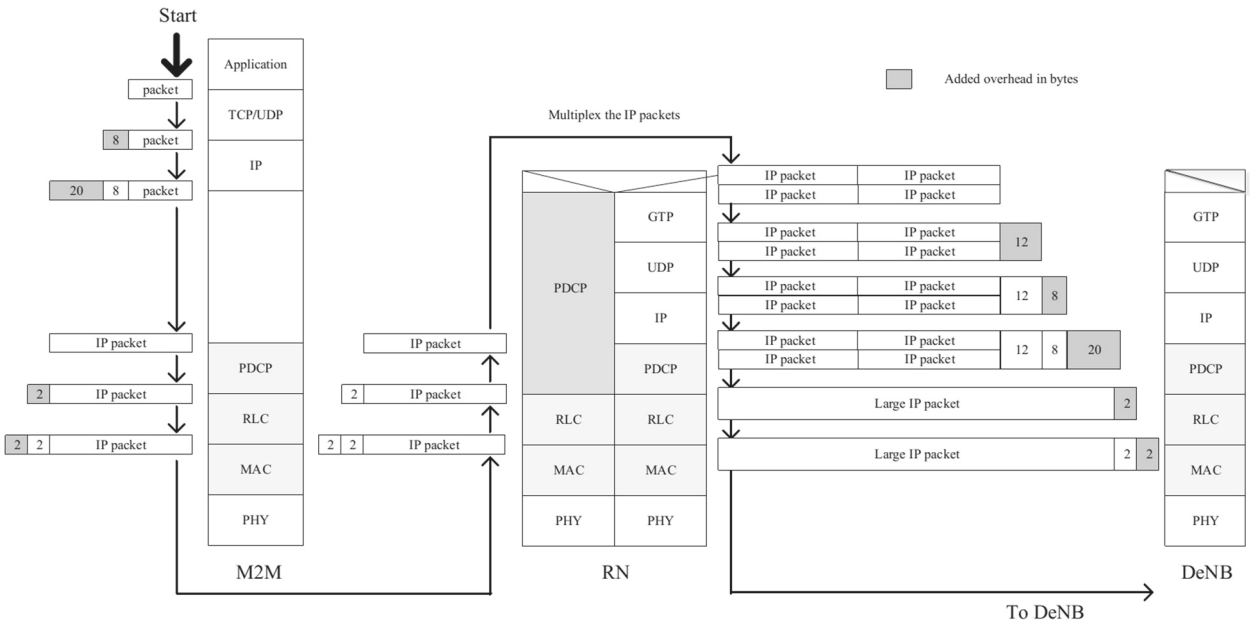

Fig. 2 An illustration of the proposed data aggregation and multiplexing scheme along with packet flow and protocol stack of the M2M device, RN, DeNB, and aGW

applications. The description of the proposed scheme is given in Algorithm 1.

When the aggregated packet is received at the DeNB PHY layer, it is sent to the DeNB GTP for de-multiplexing of the aggregated large IP packets. The original small IP packets are then sent to the application servers. The block diagram of the uplink data differentiation algorithm implemented at the DeNB is provided in Fig. 3 (left).

The DeNB also serves the regular LTE-A traffic such as voice and video UEs. The DeNB takes into account the source of the incoming packets. If the packets are received from a standalone LTE-A user, then the PDCP of the DeNB forwards packets to the GTP for tunneling. After tunneling, the packets are routed towards the access gateway (aGW). On the other hand, if the packets are received from the RN, the data packets are sent to the GTP via IP and UDP for de-multiplexing. Later, the original standalone packets are sent to the aGW. Similarly, in downlink, the DeNB GTP recognizes the source and destination identities (IDs) of the incoming packets. If the destination ID of incoming packet belongs to an RN, then the packets belong to a user which is served by the RN. On the other hand, packets belong to a user which is directly served by the DeNB. The block diagram of the downlink data differentiation algorithm implemented at the DeNB is given in Fig. 3 (right).

5 RIVERBED simulation environment and parameter settings

The optimized network engineering tool (OPNET) modeler which is newly named as RIVERBED modeler is

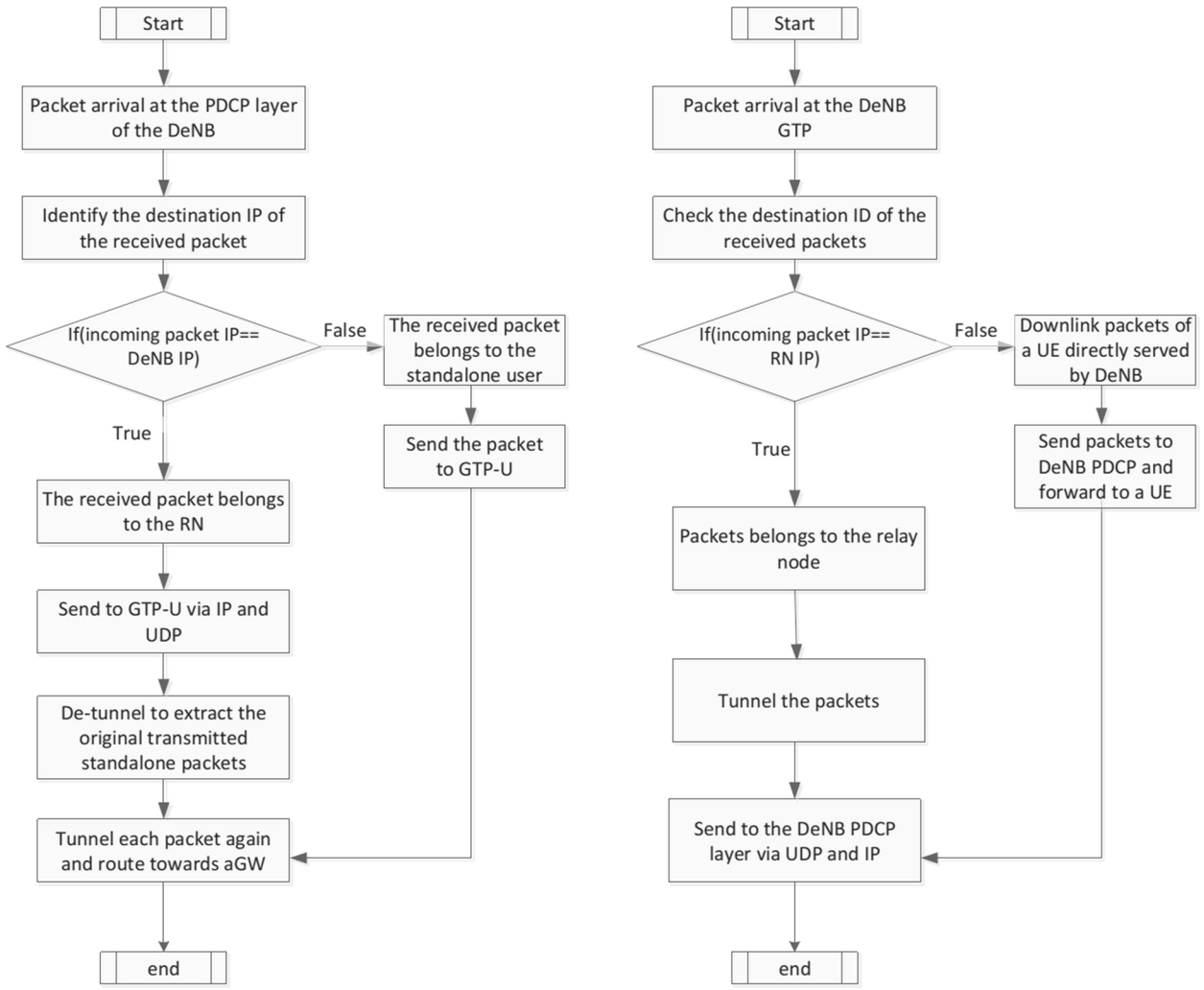

Fig. 3 Block diagram of the uplink data differentiating algorithm at the DeNB (*left*), Data differentiation algorithm for the downlink packets at the DeNB (*right*)

used as a primary modeling, simulation and analysis tool for this research. The RIVERBED modeler provides a simulation environment for the performance measurements of communication networks. A simulation model with seven cell LTE-A environment having single-sector DeNBs [33] is developed in the RIVERBED modeler with frequency reuse factor of 1 for each DeNB. The project editor of the LTE-A designed model with several nodes along with LTE-A functionalities and protocols is depicted in Fig. 4. Moreover, an RN is implemented and placed in the coverage of a DeNB to perform data aggregation and multiplexing. The RN acts as an eNB towards the M2M devices, whereas it acts as a UE towards the DeNB. The remote server supports the file transfer protocol (FTP), voice over Internet protocol (VoIP), video and M2M applications. The remote server and the aGW are interconnected with an Ethernet link with an average delay of 20 ms. The aGW node protocols include the Internet Protocol (IP) and Ethernet. The aGW and eNB nodes (eNB1..) communicate through IP routers (R1..). QoS parameters at the transport network (TN) ensures QoS parameterization and traffic differentiation. The user movement in a cell is emulated by the mobility model by periodically updating the location of the user. The user mobility information is stored in the global user database (Global-UE-List). The channel model parameters for the air interface include path loss, slow fading and fast fading models.

Moreover, in this paper, the simulation modeling mainly focuses on the user plane to perform end-to-end performance evaluations. The simulations are performed with the parameter settings given in Table 1. Moreover, the parameters of the traffic models of the LTE-A and M2M traffic are also given in Table 1. The node model implementation of an RN is depicted in Fig. 5. The layer 3 inband RN is developed in the LTE-A model with the capabilities of Uu and Un links. The Uu and Un protocol stacks are developed in accordance with the LTE-A protocols and end-user protocols. The Uu protocol stack consists of PDCP, RLC, MAC, and PHY layers which are named as as *pdcp*_0, *rlc*_0, *mac*_0, and *phy*_0, respectively as shown in Fig. 5. The RN scheduler is located at the MAC layer of the Uu interface where resources are allocated to the users served by the RN. The proposed scheme for packet data aggregation and multiplexing into large packets is implemented at the PDCP layer of the Uu interface. At the Un side, the protocols are the GTP, UDP/IP, PDCP, RLC, MAC, and PHY layers. The GTP protocol is implemented in such a way that the data packets are tunneled at the RN and detunneled again at the DeNB. Similarly, downlink packets are tunneled at the DeNB and detunneled at the RN.

6 Results and analysis

In this section, we discuss the simulation results in terms of coverage enhancement as well as efficient spectrum utilization with the proposed scheme.

6.1 Coverage enhancements

One of the major functionalities of the 3GPP LTE-A RNs includes coverage enhancements, especially for the UEs at the cell edge which usually experience poor channel conditions. The DeNB MAC scheduler allocates PRBs to each UE according to the received signal strength. The cell edge users usually experience poor signal strength which ultimately reduces the maximum utilization of the

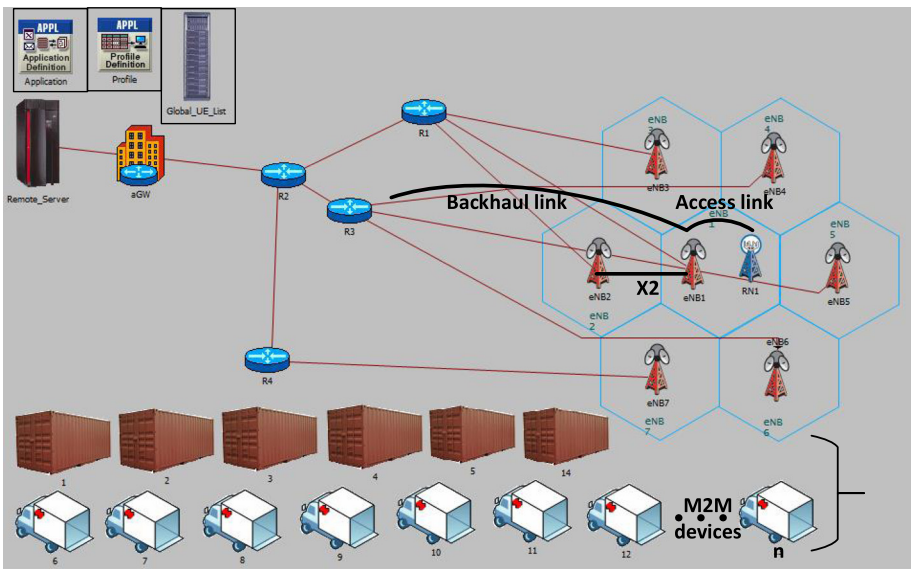

Fig. 4 Project editor of the LTE-A based model developed in the RIVERBED modeler simulator

Table 1 Simulation parameters and settings

Parameters	Values
Simulation length	1000 s
Donor eNBs (DeNBs)	7 eNBs with hexagonal coverage.
	500 m inter-eNB distance.
	Transmission power per PRB=26 dBm
Min. DeNB - UE distance	35 m
Terminal power	Maximum terminal power = 23 dBm
	(regular UE and M2M device).
Terminal speed	5 km/h (UEs)
TTI	1 ms
Mobility model	Random way point (RWP)
Frequency reuse factor	1
Transmission bandwidth	5 MHz
No. of PRBs	25
MCS	QPSK, 16QAM, 64QAM,
Channel models	Path loss: $128.1 + 37.6log_{10}(R)$, R in km,
	Slow fading: Log-normal shadowing,
	correlation 1, deviation 8 dB,
	Fast fading: 3GPP Pedestrian A.
RN parameters	
PRBs for RN	5 PRBs are allocated to RN by DeNB
	to evaluate PRB utilization.
Corresponding MCS	16
TBS capacity	1608 bits against MCS 16 and PRBs 5.
	available service rate $TBS - overhead$ (bits/TTI),
	1608 (TBS) − 352 ($overhead$) = 1256 bits/TTI
M2M packet size	RN Uu PHY receives 262.4 bits per TTI.
	RN Un PHY receives 656 bits per TTI.
Simulated scenarios	No multiplexing,
	multiplexing without timer, and
	multiplexing with timer
Timer expiry values	9 ms
Timer expiry values for subsection 6.3	4, 9, and 14 ms
Type of RN	Fixed
Video traffic model	
Size (frame)	1200 Bytes
Time (between frames)	75 ms
M2M traffic model [35].	
Message size	38 bytes at the device PDCP
Inter-send time	1 s

allocated PRBs. This inefficient utilization of PRBs significantly degrades the cell throughput. The RNs play a vital role in improving the QoS performance of the cell-edge UEs.

In order to evaluate the impact of relaying on the QoS performance of the cell-edge UEs, the scenarios are categorized into two groups namely direct-DeNB access and relay-based-DeNB access. In the first category, all the UEs communicate with DeNB directly. Whereas in the second category, an RN acts as an access point for the cell-edge UEs and later routes the UEs data to the DeNB. In both the categories, the UEs and RN are assumed to be stationary which results into non-varying channel conditions. The position of RN corresponds to MCS 16 with a TBS 1608 bits per TTI of 1 ms duration.

Each category is further subdivided into various subscenarios in which LTE-A video UEs are simulated. In the first subscenario, 10 video UEs are simulated. The number of the video UEs is incremented by 10 in the subsequent subscenarios. Figure 6 illustrates the simulation results of the average PRBs used with 95 % CI (confidence interval) in all the categories describes above. In the case of *without relaying*, the number of average PRBs used in the low-load case is less than 25 (total number of PRBs). However, increasing the number of UEs within cell also increases the average number of PRBs used. In the case of relaying, less number of average PRBs are used due to improved coverage which ultimately results in improved signal strength. Consequently, PRBs are used with the maximum capacity. However, in the high-load scenarios, e.g., 60 UEs, all the 25 PRBs are used which is the total available bandwidth in a 5-MHz system.

The E2E delay statistics of video traffic with 95 % CI is depicted in Fig. 7. In the low-load scenarios, the simulation results illustrate slightly higher delay in the case of relay-based-DeNB access due to relaying. However, in the high-load scenarios, the delay time in direct-DeNB access scenarios rises exponentially. This is due to the almost complete utilization of the available bandwidth which can be seen in Fig. 6. However, in the case of relaying, the delay time is still acceptable which is due to the fact that the maximum PRBs utilization due to improved coverage.

The simulation results of average traffic received with 95 % CI in both categories are depicted in Fig. 8. The better utilization of PRBs in the case of *relaying* gives increased average uplink traffic received at the PHY layer of the eNB. In all the subscenarios of both categories, the average traffic received is slighty better in the case of relaying. From the given simulation results, it is concluded that relaying can significantly improve the network QoS performance in terms of PRB utilization, E2E delay as well as average traffic received.

6.2 Efficient radio resource utilization

In this subsection, the performance of the proposed data traffic aggregation and multiplexing is evaluated. The scenarios are simulated according to three major groups. In the first group, M2M data packets are relayed in uplink

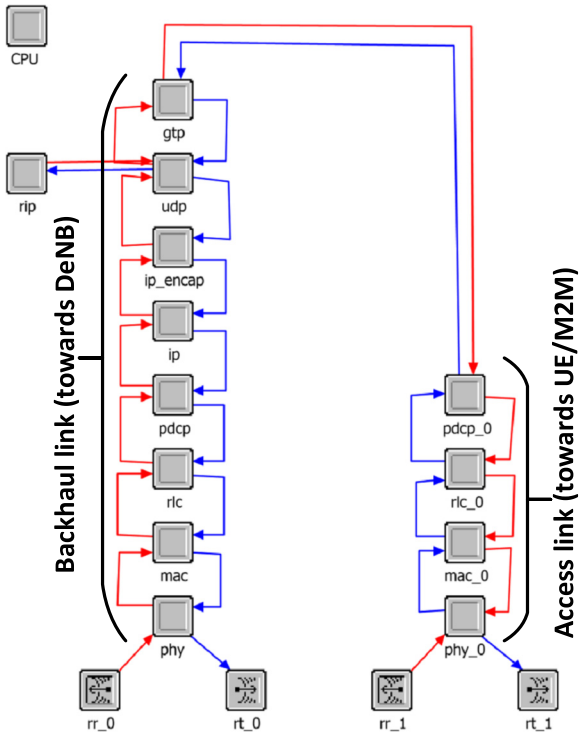

Fig. 5 The node model implementation of an RN in LTE-A-based model developed in the RIVERBED modeler simulator

without multiplexing. In the second group, the data packets from all the active M2M devices which are located in the proximity of the RN are aggregated at the Uu PDCP layer before being sent to the DeNB. However, only the periodic per-hop control scheme is used in which the large aggregated data packets are served when their size is equal to the available TBS − Un protocol overhead. In the third group, an expiry timer is introduced in order to limit the multiplexing delay especially in the low-load scenarios. In this case, the aggregated packet is served after T_{max} at the latest (periodic simple approach).

All the above mentioned groups are further subcategorized into various subscenarios. In the first subscenario, 200 M2M devices are placed in the proximity of the RN. The number of M2M devices is incremented by 200 in the subsequent subscenariors. The position of the RN corresponds to an MCS of 16 and a TBS of 1608 bits per TTI with the allocation of 5 PRBs by DeNB. The performance of the proposed scheme is computed in terms of mean number of PRBs used and packet E2E delay in uplink for all the aforementioned scenarios.

The simulation results for the mean number of PRBs used and packet E2E delay with 95 % CI are illustrated in Figs. 9 and 10, respectively. The values of the upper and lower bound of the confidence intervals are very small in most of the scenarios as depicted in Figs. 9 and 10. The

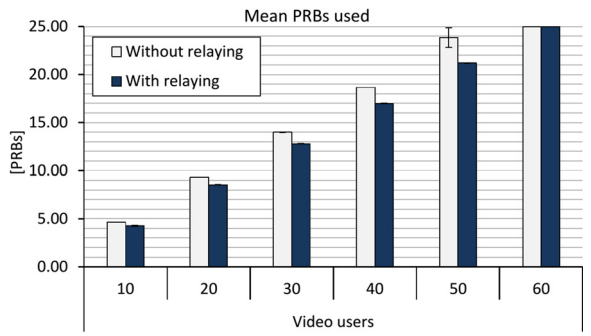

Fig. 6 Mean number of PRBs utilized with and without relaying

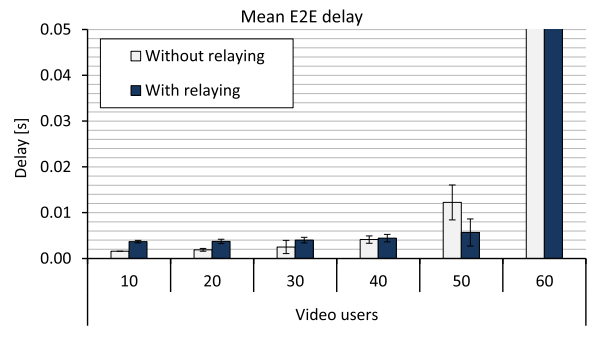

Fig. 7 Mean packet E2E delay with and without relaying for video traffic

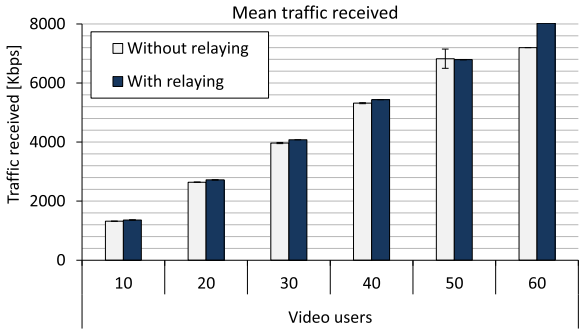

Fig. 8 Mean traffic received with and without relaying

simulation results in Fig. 9 clearly show the efficient utilization of PRBs in uplink with the proposed M2M data aggregation. For instance, in the "no multiplexing" scenario with 400 devices, the arrival rate at the RN Uu PHY layer is 262.4 bits per TTI which almost utilize 1 PRB. However, in the case of multiplexing, only half of the PRBs are used to serve the 400 devices. Similarly, without multiplexing the RN serves nearly 2400 devices with 5 PRBs in uplink, which actually depicts the limiting case of no multiplexing approach. Moreover, the system utilization, ρ can be determined in terms of arrival and service rate. The arrival rate is given as, $\lambda = (N \times 656)\ b/s$ where N is the number of active devices and 656 bits is the amount of data per device received at the RN Un PHY layer. Moreover, the service rate is $\mu = \mu_{PRB} \times 5\ (PRBs)\ b/s$. Thus, the system utilization, ρ and the maximum number of served devices, N_{max} can be determined according to $\rho = \frac{N \times 656}{321600}$ and $N_{max} = \frac{\rho \times 321600}{656}$, respectively. However, the present system parameter settings serve approximately $N_{max} = 2400$ devices, which also resembles with Fig. 9.

However, in case of multiplexing, the number of devices served by the RN nearly doubles. This is due to the fact that in the case of "no multiplexing", each data packet contains an additional Un air interface overhead of GTP, UDP, IP and layer 2. The additional overhead causes an extra PRB usage and overall reduces the PRB utilization efficiency. Moreover, in low-load scenarios, the average number of PRBs used is slightly higher in the case of "multiplexing with timer". This is due to the fact that the RN serves the traffic at the latest after 9 ms and thus the PRB is not necessarily used with its maximum capacity due to the low arrival rate. However, in high-load scenarios, the timer has almost no impact and nearly equal numbers of PRBs are used with and without timer.

Figure 10 depicts the simulation results of M2M mean packet E2E delay in all three aforementioned scenarios. The results show that the value of packet E2E delay is higher in the case of multiplexing for low loads. This increase in delay time is due to less arriving packets especially in the low-load scenarios. The buffer aggregates packets until its size + the additional RN Un overheads is equal to the offered load. However, the use of an expiry timer limits the delay by serving the aggregated packets at the latest after 9 ms. However, in high-load scenarios (e.g., 2000 devices) the E2E delay is slightly higher compared to the case of no multiplexing. This is due to high arrival rate and the buffer aggregates the incoming packets to make a large aggregated packet within less time. Moreover, the value of E2E delay is very large in fully loaded scenarios as depicted in Fig. 10, when the RN utilizes all five PRBs with the maximum capacity.

6.3 Impact of timer expiry, T_{max}

In this subsection, we investigate the impact of timer expiry, T_{max} on mean PRB utilization and E2E delay by varying input traffic. However, we only consider low-load scenarios due to the fact that the timer expiry has almost no affect on mean PRB utilization and E2E delay in high-load scenarios in the multiplexing process, as discussed

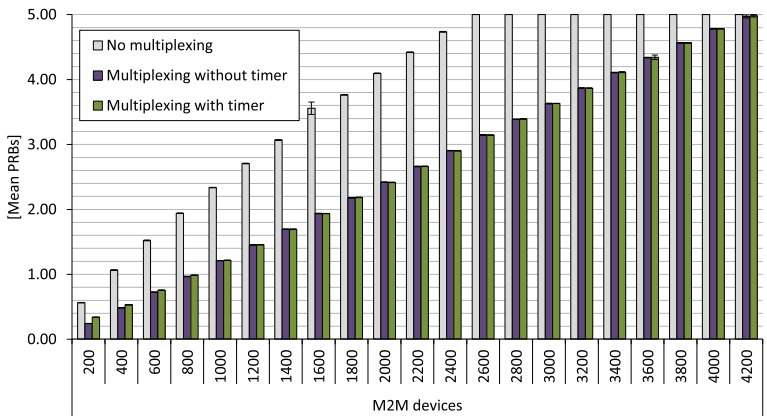

Fig. 9 Average number of PRBs used at the Un air interface away from the RN in the uplink

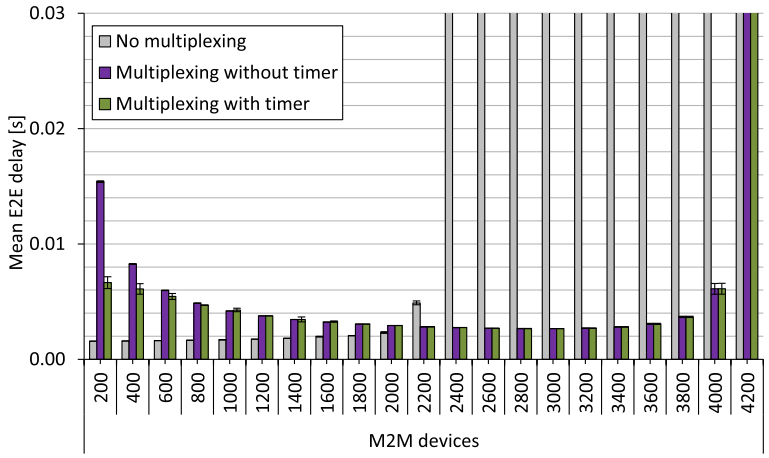

Fig. 10 Average data packets E2E delay with and without the proposed M2M data aggregation

earlier in Figs. 9 and 10. In general, the mean value of used PRBs decreases and E2E delay increases for larger values of the timer expiry. Additionally, this trade-off is heavily dependent on the arrival rate, as discussed earlier.

In this work, 10 scenarios are simulated in order to evaluate the impact of timer expiry on the multiplexing process. In the first scenario, 100 M2M devices are considered to generate input traffic. The number of devices in incremented by 100 in the subsequent scenarios. The above scenarios are simulated for the timer expiry values of 4, 9, and 14 ms. The simulation results of mean number of used PRBs and E2E delay are compared for the given values of maximum waiting time. Figures 11 and 12 compare the simulation results of mean PRB utilization and E2E delay, respectively. Figure 11 shows that less PRBs are used in the case of larger values of T_{max}, when the arrival rate is kept constant. For instance, in the first scenario with $N = 100$, average numbers of PRBs used are 0.48, 0.27, and 0.20 for the timer expiry values of 4, 9, and 14 ms, respectively. It shows that for the larger values of maximum waiting time, less number of PRBs are used on average, as it allows more packets to be multiplexed before timer expiry. Resultantly, it increases the multiplexing gain. Additionally, it is further noted that larger values of T_{max} has almost no affect on PRBs used in the case of high loads. For instance, when $N = 1000$, the average values of PRBs used are 1.219 and 1.213 for the maximum waiting time of 9 and 14 ms, respectively.

Similarly, Fig. 12 compares the simulation results of the mean E2E delay for the given timer expiry values. As discussed in the beginning, the larger values of waiting time, T_{max} increase the mean packet E2E delay, particulaly in the low-load scenarios. However, the effect reduces with the larger values of N. For instance, in the low-load scenario when $N = 100$, the given values of maximum waiting time introduce a mean E2E delay of 0.0045, 0.0070, and 0.0095 ms, respectively. Since the effect of timer on the multiplexing process decreases due to the increasing arrival rate, the mean E2E delay values are reduced to 0.00368, 0.00416, and 0.00419, respectively for $N = 1000$. Furthermore, it is noted that the values of mean E2E delay for T_{max} of 10 and 15 ms are almost similar. Moreover, the

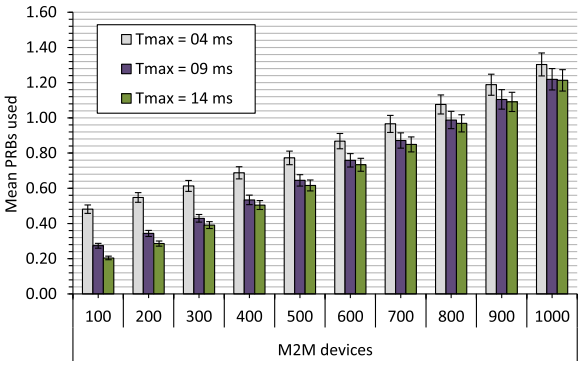

Fig. 11 A comparison of mean PRBs used

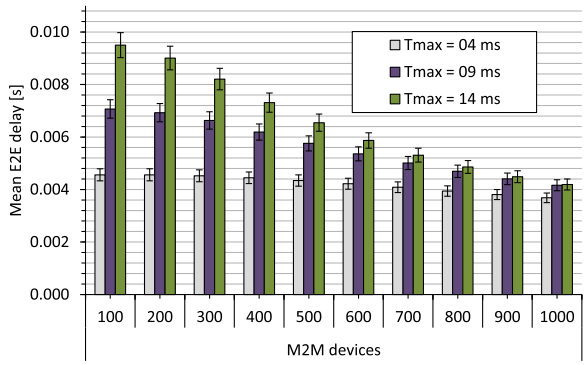

Fig. 12 A comparison of packet mean E2E delay

impact of timer expiry completely vanishes when input traffic load is increased beyond 1000, see Figs. 9 and 10.

7 Analytical modeling for multiplexing transition probabilities

In this section, an analytical model is developed based on the work done in [34] for multiplexing of small common part sub-layer (CPS) packets into large ATM cells with the AAL2 layer multiplexer. An ATM cell is made up of a 5-byte header and a 48-byte payload. Among the 48 bytes of payload, 1 byte is reserved for the start field. The rest of the 47 bytes are available for multiplexing the short CPS packets. The multiplexing approach presented in [34] is analogous to the aggregation and multiplexing scheme proposed for the small-sized M2M data packets. The list of the various symbols with a brief description is provided in Table 2.

The M2M data packets arrive at the Uu PDCP layer of the RN from various M2M devices. The arrival of a packet to the multiplexer at the RN PDCP layer sets a timer in such a way that the packet has to wait for a certain duration of timer T_{max} for the arrival of more packets to achieve maximum multiplexing gain. If the waiting time exceeds the timer expiry limit T_{max}, the multiplexing process is initialized, and the multiplexed data is sent to the Un PDCP buffer of the RN. If all the N_{max} PRBs are not required for transmission, the unused PRBs are allocated to other users requesting resources from the DeNB. In case the maximum waiting time T_{max} is not exceeded, but the accumulated size of the packets in the buffer exceeds the TBS capacity $n_{max} - overhead$ (bytes), then the multiplexing starts even if the timer has not expired. The overhead from GTP, UDP, IP, PDCP, RLC, MAC, and PHY layers of the Un interface of the RN have to be considered while determining the TBS capacity. The maximum number of M2M packets r that can be multiplexed into a large IP packet can be expressed as,

$$r = \frac{n_{max} - additional\ Un\ protocol\ overhead}{l}, \tag{3}$$

Table 2 List of symbols

Symbols	Descriptions
T_{max}	Maximum waiting time for aggregation and multiplexing.
N_{max}	Complete number of PRBs.
n_{max}	Number of packets in the buffer at any time.
l	Fixed size of the packet (656 bits).
l_k	Total length of the buffer.
p_k	Probability of next arrival before multiplexing with k existing packets.
q_k	Probability that multiplexing starts before next arrival.

where l denotes the fixed size of the M2M data packet received at the RN. For simplicity, we assume the fixed sized M2M data packets. If there are k M2M packets in the PDCP buffer, then $1 \leq k \leq r$ until the waiting time reaches T_{max} or the buffer size reaches $n_{max}overhead$. The total size of the buffer is l_k. The inter-arrival time of the M2M data packets is considered to be negative exponentially distributed with a mean value of λ. Therefore, the maximum waiting time for starting the multiplexing process is Erlangian. The probability of the next arrival before multiplexing with k packets already in the buffer is p_k where as the probability that the multiplexing process starts after the arrival of the k^{th} packet is q_k such that $p_k = 1 - q_k$. This process can be modelled as an r-stage Coxian process and is depicted in Fig. 13. Thus, the probability of the next arrival, p_k with k packets in the buffer can be determined as,

$$p_k = \begin{cases} 1 - \left(\sum_{j=0}^{k-1} \frac{e^{-\lambda T_{max}}}{j!} (\lambda T_{max})^j \right), & k < r \\ 0, & k \geq r \end{cases}. \tag{4}$$

The Laplace-Stieltjes Transform (LST) of the probability distribution function (PDF) of the r-stage Coxian process can be expressed as,

$$A_s = \sum_{k=1}^{r} \varphi_k \left(\frac{\lambda}{s+\lambda}\right)^k, \tag{5}$$

where φ_k is the probability that the multiplexing process would start after k arrivals. φ_k can be expressed as,

$$\phi_k = q_k \prod_{j=0}^{k-1} p_j \tag{6}$$

where $p_0 = 1$.

In this work, low-load scenarios are considered where each M2M device transmits 33 bytes towards the RN. The values of p_k are determined for some of the low-load scenarios. The comparison of the probabilities from p_1 to p_5 is given in Fig. 14. p_1 is the probability of the arrival of the 2^{nd} packet before the multiplexing process starts. In case of 100 M2M devices, the probability p_1 is around 0.6 for both the simulation and analytical results. This implies that the probability of the timer expiry before the 2^{nd} arrival is approximately 0.4. Increasing the traffic load increases the probability p_1 in subsequent scenarios. In case of 700 or more M2M devices in the RN coverage area, p_1 is almost equal to 1.0 which indicates that at low load, the probability of multiplexing before 2^{nd} arrival is very low. Similarly, p_2 is the probability of the arrival of 3^{rd} packet before the multiplexing process starts. The results for p_2 show that the arrvial of the 3^{rd} packet is highly probable in high-load scenarios as compared to low-load scenarios. As expected, p_2 is slightly less than p_1. The simulation and analytical results for probability p_3 in Fig. 14

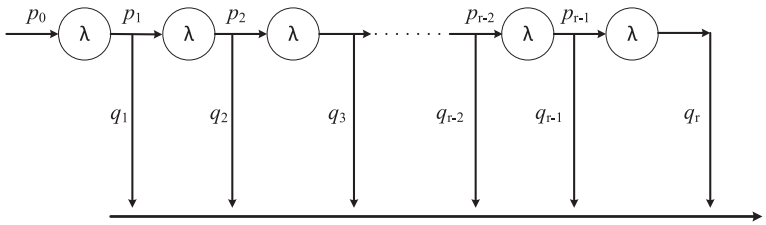

Fig. 13 r-stage Coxian distribution

show a similar pattern as the results of p_1 and p_2. At low load of 100 devices, p_3 is below 0.1. But, increasing the load results in higher values for p_3. In case of probability p_4, the results for low load of 100 M2M devices show that it is highly unlikely that a 5^{th} packet would arrive before the start of multiplexing process. However, the likeliness of a 5^{th} arrival increases with increasing the M2M traffic load.

8 Conclusions

Since radio spectrum will remain a scarce resource also in the future, new ideas for new traffic types such as M2M demand careful planning and evaluation. In addition, the IoT will drive new developments including new devices, applications as well as services. Thus, the future poses numerous challenges for all parties involved such as standardization bodies, network operators, and application providers.

This paper presents a data aggregation and multiplexing scheme for handling small-sized M2M traffic in order to improve LTE-A radio resource utilization. In the analysis, it is observed that the proposed scheme significantly improves the PRB utilization, particularly when no expiry timer, T_{max} is considered in the low-load scenarios. However, in the high-load scenarios, the timer T_{max} does not play any role and equal number of PRBs are used with and without an expiry timer. On average, approximately 40 % more M2M devices are served with the proposed multiplexing scheme. Given the simulation results, in the future recognizing narrowband M2M traffic could be a very important step for further bandwidth saving. Furthermore, the proposed scheme in this paper could be an efficient solution for handling such kind of bursty traffic.

In this work, M2M QoS is not considered, and all the devices are given equal priorities. Due to the wide application range, QoS provisioning in M2M communications is one of the most challenging tasks for cellular service providers as well as for mobile networks operators. As QoS provisioning is one of the major demands of the future mobile M2M applications such as e-healthcare and emergency alerting. Therefore, the data aggregation scheme can be upgraded by taking into account the QoS of several M2M applications. Furthermore, the data aggregation concept will be extended in the downlink.

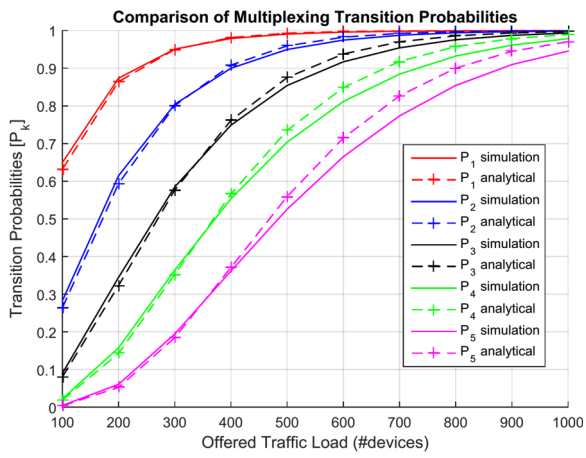

Fig. 14 Comparison of the probabilities, P_k

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
We thank the International Graduate School for Dynamics in Logistics (IGS), doctoral training group of LogDynamics, University of Bremen, Germany, and EU programme Erasmus Mundus FUSION (Featured Europe and South Asia Mobility Network) Strand 1, Lot 11, Grant Agreement Reference number 2013-2541/001-011 EM Action 2 partnerships, for the financial support of this work.

Author details
[1]Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan. [2]Communication Networks, University of Bremen, Bremen, Germany. [3]Institute of Communication Networks, University of Technology, Hamburg, Germany.

Received: 24 November 2015 Accepted: 5 April 2016
Published online: 18 April 2016

References

1. A Osseiran, V Braun, T Hidekazu, P Marsch, H Schotten, H Tullberg, MA Uusitalo, M Schellmann, in *IEEE VTC Spring Workshop*. The foundation of the mobile and wireless communications system for 2020 and beyond (IEEE, Dresden, Germany, 2013), pp. 1–5
2. METIS, Scenarios, requirements and KPIs for 5G mobile and wireless system (2013). Technical report, Deliverable Number ICT-317669-METIS/D1.1
3. Exalted, Expanding LTE for Devices. http://www.ict-exalted.eu/fileadmin/documents/EXALTED_WP2_D2.1.pdf. Accessed: 30 September 2013

4. MZ Shafiq, L Ji, AX Liu, J Pang, J Wang, in *ACM SIGMETRICS Performance Evaluation Review*. A first look at cellular machine-to-machine traffic: large scale measurement and characterization, vol. 40 (ACM New York, NY, USA, 2012), pp. 65–76
5. ERICSSON, Towards 50 billion connected devices. Technical report (2010). http://www.ericsson.com/au/res/region_RASO/docs/2010/ericsson_50_billion_paper.pdf. Accessed: 30 November 2014
6. Y Mehmood, W Afzal, F Ahmad, U Younas, I Rashid, I Mehmood, in *19th International Conference on Automation and Computing (ICAC)*. Large scaled multi-user mimo system so called massive mimo systems for future wireless communication networks (IEEE, UK, London, 2013), pp. 1–4
7. Y Mehmood, N Haider, W Afzal, U Younas, I Rashid, M Imran, in *IEEE Malaysia International Conference on Communications (MICC)*. Impact of massive MIMO systems on future M2M communication (IEEE, Kuala Lumpur, Malaysia, 2013), pp. 534–537
8. http://www.3gpp.org/technologies/keywords-acronyms/97-lte-advanced_links.html. Accessed: 28 May 2015
9. 3GPP, Standardization of machine-type communications. Technical report, V0.2.4 (2014). http://www.ericsson.com/au/res/region_RASO/docs/2010/ericsson_50_billion_paper.pdf. Accessed: 30 September 2015
10. 3GPP, System improvements for machine type communications (2011). Technical report, 3GPP TR 23.888
11. METIS, Novel radio link concepts and state of the art analysis (2013). Technical report, Deliverable Number ICT-317669-METIS/D2.2
12. 3GPP, Architecture enhancements to facilitate communications with packet data networks and applications (Release 13). Technical report, TS 23.682 V13.2.0 (2015)
13. ETSI, Machine-to-machine communications (M2M), Functional architecture (2013). Technical report, ETSI TS 102 690 V2.1.1
14. Achieving LOw-LAtency in Wireless Communications (LOLA). http://www.ict-lola.eu/_links.html. Page Accessed: 30 June 2015
15. Mobile and wireless communications enablers for twenty-twenty 2020. https://www.metis2020.com/?doing_wp_cron=1440943491.5532760620117187500000_links.html. Accessed: 28 August 2015
16. 3GPP, Technical specification group radio access network: evolved universal terrestrial radio access (E-UTRA); Physical layer procedures (Release 9) (2010). Technical report, TS 36.213 V9.2.0
17. T Pötsch, SNK Marwat, Y Zaki, Görg C, in *IEEE 6th Joint IFIP Wireless and Mobile Networking Conference (WMNC)*. Influence of future M2M communication on the LTE system (IEEE, Dubai, UAE, 2013), pp. 1–4
18. Y Mehmood, SNK Marwat, T Pötsch, F Ahmad, C Görg, I Rashid, in *LogDynamics International Conference (LDIC)*. Impact of machine-to-machine traffic on LTE data traffic performance (Springer, Bremen, Germany, 2014), pp. 259–269
19. Y Mehmood, C Görg, M Muehleisen, Timm-Giel A, Mobile M2M communication architectures, upcoming challenges, applications, and future directions. EURASIP J. Wireless Commun. Netw. **2015**(1), 1–37 (2015)
20. K Kalpakis, K Dasgupta, P Namjoshi, Maximum lifetime data gathering and aggregation in wireless sensor networks. Networks. **42**(6), 697–716 (2003)
21. D Luo, X Zhu, X Wu, G Chen, in *IEEE INFOCOM Proceedings*. Maximizing lifetime for the shortest path aggregation tree in wireless sensor networks (IEEE, Shanghai, China, 2011), pp. 1566–1574
22. S Madden, MJ Franklin, JM Hellerstein, W Hong, TAG: A tiny aggregation service for ad-hoc sensor networks. ACM SIGOPS Operating Syst. Rev. **36**(SI), 131–146 (2002)
23. V Pandey, A Kaur, N Chand, A review on data aggregation techniques in wireless sensor network. J. Electron Electr. Eng. **1**(2), 1–8 (2010)
24. L Mottola, GP Picco, MUSTER: Adaptive energy-aware multisink routing in wireless sensor networks. IEEE Trans. Mob. Comput. **10**(12), 1694–1709 (2011)
25. I Solis, K Obraczka, in *IEEE International Conference on Communications (ICC)*. The impact of timing in data aggregation for sensor networks (IEEE, Paris, France, 2004), pp. 3640–3645
26. A Showail, K Jamshaid, B Shihada, in *IEEE Wireless Communications and Networking Conference (WCNC)*. An empirical evaluation of bufferbloat in IEEE 802.11 n wireless networks (IEEE, Istanbul, Turkey, 2014), pp. 3088–3093
27. A Showail, K Jamshaid, B Shihada, in *Proceedings of the 2014 ACM SIGCOMM Workshop on Capacity Sharing Workshop*. WQM: An aggregation-aware queue management scheme for IEEE 802.11 n based networks (ACM New York, NY, USA, 2014), pp. 15–20
28. SY Tsai, SI Sou, MH Tsai, Reducing energy consumption by data aggregation in M2M networks. Wirel. Pers. Commun. **74**(4), 1231–1244 (2014)
29. SNS Jamadagni, RS Vaidya, SA Ganapathi, Method and system of transmitting packet data units of machine type communication devices over a network interface in a long term evolution network. Google Patents. US Patent App. 13/878,898 (2013). http://www.google.com/patents/US20130195017
30. 3GPP, Technical specification group radio access network: evolved universal terrestrial radio access (E-UTRA); Relay architectures for E-UTRA (LTE-Advanced) (2010). Technical report, TR 36.806 V9.0.0
31. A Baid, R Mad, A Sampath, in *10th IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*. Delay estimation and fast iterative scheduling policies for LTE uplink (IEEE, Tempe, Arizona, USA, 2012), pp. 89–96
32. Y Mehmood, SNK Marwat, Y Zaki, C Görg, A Timm-Giel, in *20 ITG Mobile Communication Conference*. Evaluation of M2M data traffic aggregation in LTE-A uplink (VDE, Osnabrück, Germany, 2015), pp. 24–29
33. Y Zaki, T Weerawardane, C Görg, A Timm-Giel, in *OPNET Workshop*. Long term evolution (LTE) model development within OPNET simulation environment (OPNET Technologies, Washington DC, USA, 2011)
34. G Ko, S Moon, A Ahmad, K Kim, Performance analysis of the ATM adaptation layer 2 (AAL2). Int. J. Electron. Commun. **58**(3), 193–199 (2004)
35. IEEE 802.16p-11/0014, IEEE 802.16p Machine to machine (M2M) evaluation methodology document (EMD) (2010). Technical report, IEEE 802.16 Broadband Wireless Access Working Group

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com