A Multi-queue Aggregation Framework for M2M Traffic in LTE-A and Beyond Networks

Wen Feng, Hongjia Li(✉), Ding Tang(✉), Liming Wang, and Zhen Xu

State Key Laboratory of Information Security, Institute of Information Engineering,
Chinese Academy of Sciences, Beijing 100093, China
{lihongjia,tangding}@iie.ac.cn

Abstract. Traffic aggregation has been considered as an effective approach to improve the radio resource utilization for M2M communication in LTE-A and beyond networks. In the LTE-A specification, the Relay Node (RN) is recommended to aggregate uplink M2M small-sized packets. However, the delay brought by the packets aggregation is inevitably increased, which is a vital metric for M2M packets with low delay requirement, such as emergency alerting. In this paper, we propose a new framework for optimal aggregation implemented in the PDCP of RN, which features balancing a tradeoff between QoS requirements of packets and the utilization efficiency of Physical Radio Blocks (PRBs). Specifically, (1) the RN dispatches the new arrival M2M packets into corresponding virtual queues according to their priorities set by M2M devices. Then, an Optimal Aggregating Scheme (OAS) is designed to minimize the PRB usage in condition satisfying the specific restriction of waiting time of packets in virtual queues. (2) The optimal aggregating problem is proved to be a NP-hard problem, which is solved by the Priority Branch and Bound Algorithm (PBBA) and the Priority Aggregating Heuristic. Numerical results illustrate that OAS achieves a tradeoff of QoS and PRB utilization efficiency in comparison with four existing schemes.

Keywords: M2M · Quality of service · Packet aggregation · LTE-A

1 Introduction

Massive Machine-to-Machine (M2M) communication is considered as the potential important research direction in 5G networks. M2M communication is a pattern which identifies the evolving paradigm of interconnected devices communicating with each other without or with limited human interaction. The application domain of M2M traffic includes smart metering, e-health, surveillance and security, intelligent transportation, city automation, smart monitoring and many more. M2M traffic patterns vary in diverse application domains and in most of the applications, and especially some M2M devices mainly are small packets that consist of a few bytes. Because the payload of the data packets associated with

K. Long et al. (Eds.): 5GWN 2017, LNICST 211, pp. 609–621, 2018.
https://doi.org/10.1007/978-3-319-72823-0_56

M2M applications is usually smaller than a Physical Resource Block (PRB) [1]. Thus, M2M traffic is supposed to degrade the utilization of radio spectrum. In addition, M2M devices can have different delay tolerances based on their applications, ranging from a few milliseconds (ms) to several minutes or even hours, e.g., emergency alerting may need to provide a very stringent low delay, while temperature monitoring can owe a great delay tolerance [2]. Therefore, it is a significant problem for the evolution of M2M communication to improve the utilization of radio spectrum and satisfy delay requirement.

Packets or traffic aggregation, which collects and accumulates data packets from multiple nodes before transmitting to the next hop, is supposed to effectively improve the utilization of radio spectrum for M2M traffic in future 5G networks because of reducing the extra overhead and adding the size of data in a PRB. For example, packets aggregation significantly improves the PRB utilization compared to the conventional without multiplexing approach [3]. However, packets aggregation inevitably results in the delay increasing of M2M application. Because the incoming packets can be transmitted only when certain aggregation conditions are satisfied. Therefore, it is a fundamental problem to obtain a tradeoff between Quality-of-Service (QoS) requirements of packets and the utilization efficiency of Physical Radio Blocks (PRBs).

Some efforts have been separately made to improve radio spectrum utilization and reduce time delay of M2M communication. For instance, in [4], the authors propose a data aggregation scheme which aggregates uplink M2M traffic by sharing the PRBs to increase the number of M2M packet in a PRB. Reference [5] efficiently integrates M2M traffic into cellular networks to take advantage of uplink transmission time slot and reduce resource wastage at both the network and device. Authors in [6] propose an optimal aggregation for multirate WLANs to minimize overall transmission time. In [7], a packet chunking is introduced which need multiple buffers and then classifies the arrival packets to one buffer based on their acceptable waiting time to take the best use of the core router's forwarding capacity. However, few arts have delved into the combination of improving radio spectrum utilization and reducing time delay of M2M packets.

Thus motivated, we propose a new framework for optimal aggregation, which features balancing a tradeoff between QoS requirements of packets and the utilization efficiency of PRBs. For this purpose, a Relay Node (RN) of LTE-A are introduced to aggregate uplink M2M traffic. The main contributions are summarized below. (1) The RN dispatches the new arrival M2M packets into corresponding virtual queues according to their priorities set by the M2M devices. Then, an Optimal Aggregating Scheme (OAS) is designed to minimize the PRB usage in condition satisfying the specific restriction of waiting time of packets in virtual queues. (2) In order to solve the optimal aggregating problem which is proved to be a NP-hard problem, we propose a Priority Branch and Bound Algorithm (PBBA). Due to the low computational efficiency of PBBA, the Priority Aggregating Heuristic is introduced.

The rest of this paper is organized as follows. Section 2 overviews the framework of optimal traffic aggregation. Section 3 presents the optimal aggregating problem formulation and the solution algorithms. Section 4 evaluates the performance of our proposed algorithms. Finally, Sect. 5 concludes the paper.

2 The Framework of Optimal Traffic Aggregation

2.1 System Framework

In this part, we propose a new framework for optimal aggregation implemented in the Packet Data Convergence Protocol (PDCP) of RN as shown in Figs. 1 and 2. Traffic from M2M devices located in the proximity of an RN is accumulated at the RN. According to 3GPP specification [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. We have made two changes in RN. First, in order to satisfy delay requirement of diverse applications, multiple queues are introduced in the PDCP of RN, which are used to distinguish various applications. Second, the operation of de-multiplexes is made in the GPRS Tunnelling Protocol (GTP) of Public Data Network GateWay (P-GW) to reduce the additional overheads.

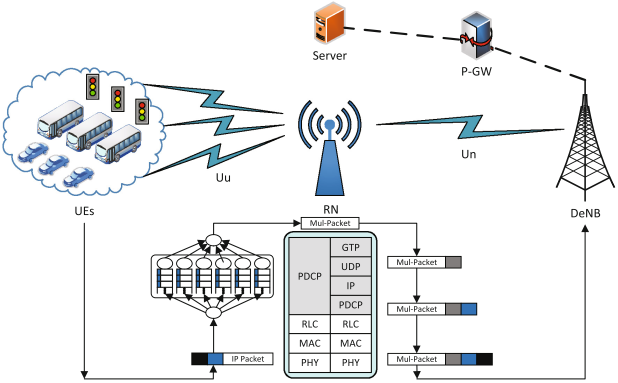

Fig. 1. An illustration of M2M packet flow with the packets aggregation in from UE to DeNB.

We consider an uplink priority aggregated scenario in RN. As shown in Fig. 1, packets of User Equipment (UE) are aggregated by the RN, and then are transited to DeNB. Packets are aggregated at the PDCP layer of RN in order to maximize the multiplexing gain. The main functionalities and services of PDCP layer for the user plane include header compression and decomposition, 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.

When arriving at the PDCP layer of RN, the incoming data packet is scheduled to one of six virtual queues which are mapping six priority levels as $p \in \{1, 2, 3, 4, 5, 6\}$. We assume that $p = 1$ (resp. $p = 6$) corresponds to the highest (resp. lowest) priority level. The six virtual queues are used to distinguish six services as shown in Table 1. According to 3GPP TS 36.107 [9], the typical Human-to-Human (H2H) services are divided into four different QoS classes, namely conversational, streaming, interactive and background, in which the differentiation have mainly considered the delay requirement. However, some of M2M applications cannot be properly mapped to the four QoS classes, especially emergency alerting and some applications of delay tolerant as shown in [10]. Thus, we extend the typical H2H services classification scheme. An indicative classification scheme, adding two service classes (i.e. emergency alerting and time tolerant), is shown in Table 1.

Table 1. M2M applications classification

Priority level	Service classes	Waiting time
1	Emergency Alerting	0 ms
2	Conversational	10 ms
3	Streaming	100 ms
4	Interactive	1 s
5	Background	10 s
6	Time Tolerant	100 s

By scheduling, some packets in six virtual queues are aggregated and sent to GTP. Then, the RN adds the additional overheads such as the GTP, User Datagram Protocol/Internet Protocol (UDP/IP), PDCP and Radio Link Control (RLC), and sends to P-GW via a GTP tunnel. A GTP option is added before packet is sent to P-GW, which can provide the needed information for extracting the original small packets in the aggregated packet. The aggregated packet flow from RN to P-GW when adding the GTP option is given in Fig. 2.

When receiving the aggregated packet, the P-GW de-multiplexes and extracts the original transmitted standalone packets according to the GTP option information, and then sends them to Public Data Network (PDN). Finally, the original small packets are sent to the application servers. The P-GW also serves the regular LTE-A traffic. If the packets are received from regular users, then the P-GW directly forwards them to PDN. On the other hand, if the aggregated packet are received from the RN, the original small packets are extracted by P-GW, and then sent to PDN.

2.2 Optimal Aggregating Scheme (OAS)

In this part, an Optimal Aggregating Scheme (OAS) is designed, which minimize the PRB usage in condition satisfying the specific restriction of waiting time of

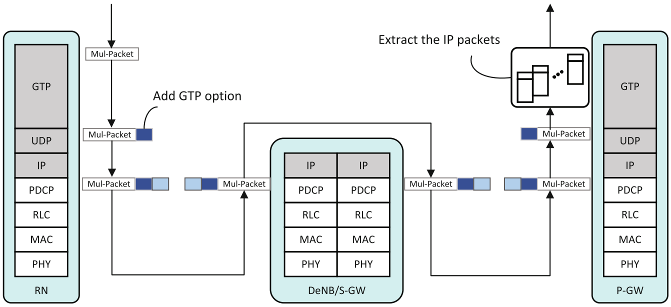

Fig. 2. The aggregated packet flow from RN to P-GW when adding the GTP option.

packets in virtual queues. Here, we denote L as the total size of all the packets in the virtual queues, and L_{max} as the maximum size that RN can aggregates the incoming M2M packets according to the size of available transport block - RN Un protocol overhead, see Fig. 1. The size of available transport block is determined by the number of PRB allocated by DeNB in a transmission time interval (TTI). According to [11], The minimum scheduling resource unit in LTE that are allocated to a single UE is a PRB in a TTI. A PRB can transmit several hundred bits under favorable channel conditions [12]. For example, 712 bits are sent in a TTI with an Modulation and Coding Scheme (MCS) of 26 when a PRB allocated by DeNB. The RN Un protocol overhead is 46 bytes: 12 bytes for GTP, 8 bytes for UDP, 20 bytes for IP, 1 byte for PDCP, 2 bytes for RLC, and 3 bytes for the MAC overhead. The payload of the data packets associated with M2M applications is IP packet. it will increases when the GTP option is added.

Priority Classification when arriving at the PDCP layer of RN, the incoming data packet is scheduled as shown in Step A of Fig. 3, which dispatchs it to one of the virtual queues according to its priority level. Before sent to RN, the packet's priority level p would be set at the Type of Service (ToS) segment of IP header by M2M devices. In the Transmission Control Protocol/Internet Protocol (TCP/IP), ToS segment of IP header has three bits that indicate six priority levels as shown by Table 1.

Packets Scheduling after priority classification, the M2M packet is placed into the tail of the corresponding virtual queue as shown in Step B of Fig. 3. Each arrival packet starts a waiting timer when in the virtual queue, which records the waiting time t of the incoming packet. The waiting time t is the time that the packet stay in PDCP of RN. Due to different delay requirements for the virtual queues, a waiting time T is introduced for the virtual queues, which denote the current waiting time of the virtual queues and is equal to the waiting time of head-of-line packet in that virtual queue. Without loss of generality, let T_{max} denotes the maximum waiting time of the virtual queues.

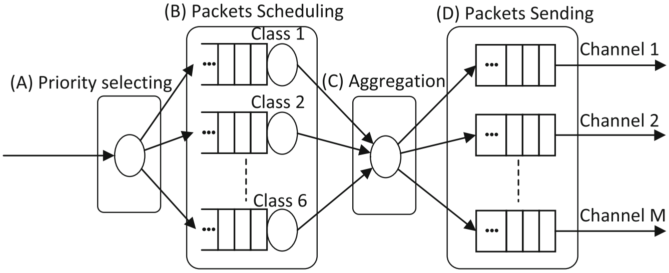

Fig. 3. The M2M schedule procession of packets at the PDCP layer of RN.

Aggregation when some aggregating conditions are satisfied, a optimal aggregating sequence of M2M packets is selected and aggregated into a large aggregated packet as shown in Step C of Fig. 3. We set two aggregating conditions: (a) The total size L exceed the maximum aggregated size of RN L_{max}; (b) The waiting time T_p ($p \in \{1, 2, 3, 4, 5, 6\}$) is greater than the maximum waiting time T^p_{max}. T_{max} is set a suitable value for six virtual queues according to characteristic of M2M packets in virtual queue. For example, the packet of the highest priority will be served immediately when incoming, and the packet of the lowest priority can be set a high value due to delay tolerance.

Packets Sending the aggregated packet is sent through one of M channels that are mapping the size of transport block allocated by DeNB, as shown in Step D of Fig. 3. Which channels are selected is determined by the number of PRB allocated by DeNB.

3 Optimal Aggregating Problem Formulation and Solution Algorithm

3.1 Optimal Aggregating Problem Formulation

We denote the set of all the packets in the PDCP of RN as I with integer index $i \in \{1, 2, 3, ...\}$, and each packet has two attributes: the priority p_i and the length l_i. We assume that 0-1 decision variable ρ_i indicate whether the data packet i is selected by RN, which is defined as

$$\rho_i = \begin{cases} 1, & \text{if the packet } i \text{ is selected} \\ 0, & \text{else} \end{cases}. \tag{1}$$

Meanwhile, the weight value of each UE's packet i in virtual queues, denoted by v_i, is defined as follows

$$v_i = \frac{t_i}{p_i}, \tag{2}$$

where p_i and t_i separately are the priority level and the waiting time of the packet i, and v_i represents the weight value of the packet i that is a significant parameter that distinct priority packets can be treated differently. v_i has two main purposes: let the high priority packet can be served faster than the low, and increase the probability that the low priority packets are served.

The goal of the following formulated problem is to obtain the optimal aggregating sequence of M2M packets such that the total weight value $\mathcal{P}(\boldsymbol{\rho})$ is maximized at any TTI.

$$\begin{aligned} \underset{\boldsymbol{\rho}}{max} \quad & \mathcal{P}(\boldsymbol{\rho}) = \sum_{i=1}^{n} v_i \rho_i, \\ s.t. \quad & \sum_{i=1}^{n} l_i \rho_i \le L_{max}, (A) \\ & \rho_i \in \{0, 1\}, \forall i \in I, (B) \end{aligned} , \tag{3}$$

where $\boldsymbol{\rho} = \{\rho_i, i \in I\}$ is the set of decision variables for the UE's packet in the PDCP of RN, and n is the total number of all UE's packets in the virtual queues. Due to constraint (B), this problem has been well proved NP-hard in [13].

3.2 Priority Branch and Bound Algorithm

In order to solve the proposed optimal aggregating problem in (3), we propose a priority branch and bound algorithm (PBBA) as given in Algorithm 1. Branch and bound algorithm (BBA) has the attractive feature to reach the optimal solution, it is sufficient to enumerate only some of the possible by branching and bounding strategies. Specifically, the proposed algorithm consists of two stages: the waiting stage and the aggregating stage. In the waiting stage, the proposed algorithm distributes the incoming packet to one of the virtual queues, and checks the aggregating conditions. If satisfying the aggregating condition, it comes into the aggregating stage. If not, it go to the next cycle. In the aggregating stage, it sorts packets descending order by the weight value, and then get the aggregated packet with $\boldsymbol{\rho}$ by BBA. Finally it send the aggregated packet to DeNB.

The principle behind BBA is perform a systematic enumeration of candidate solutions. The search process is done in a tree structure manner, which starts at the root node and goes down in the tree. Taking problem (3) as the root problem, we denote the optimal variables as $\boldsymbol{\rho}_i^{*(0)}$ and its optimal solution as $\mathcal{P}^{*(0)}$. Then if all $\rho_i^{*(0)}$ in $\boldsymbol{\rho}_i^{*(0)}$ are binary, the root problem is terminated and the optimal to the original problem (3) is found. If not, the root problem on the first non-integer ρ_i, which is named as the branching variable, will be split into two more subproblems $S(\mathcal{P}_1^{(0)})$ and $S(\mathcal{P}_2^{(0)})$ by adding two upper bound and lower bound constraints. The new formed sub-problems can be generally expressed as

Algorithm 1. PBBA

```
Set $L_{max}$ = (available TBS - RN Un overhead) and $L = 0$;
Initialize $T_p$ and $T^p_{max}$ for each queue $p$;
while packet arrival == $TRUE$ do
    The incoming packet $i$ is put into one of virtual queue $p$ according to it's priority;
    Starts the waiting timer $t_i$ for the incoming packet $i$ and update $L$ and $T_p$;
    if $L \geq L_{max}$ && $T_p \geq T^p_{max}$ then
        Update the value $v$ for all the incoming packets;
        Get $\boldsymbol{\rho}$ via algorithm 2;
        Aggregate all the packets that satisfy $\rho_i = 1$;
        Send large aggregated packet to RN PHY via RN Un protocols;
        Add RN Un protocols overhead;
        Route multiplexed packet to DeNB in next TTI;
        Update $L_{max}$ from the DeNB;
        Break;
    end if
end while
```

$$\begin{aligned} \max_{\boldsymbol{\rho}} \quad & \mathcal{P}(\boldsymbol{\rho}), \\ s.t. \quad & \sum_{i=1}^{n} l_i \rho_i \leq L_{max}, \forall i, (A) \\ & \rho_i \geq 0, \forall i \backslash (i'), \forall i \in I, (B') \\ & \rho_{i'} = 0 \end{aligned} \tag{4}$$

and

$$\begin{aligned} \max_{\boldsymbol{\rho}} \quad & \mathcal{P}(\boldsymbol{\rho}), \\ s.t. \quad & \sum_{i=1}^{n} l_i \rho_i \leq L_{max}, \forall i, (A) \\ & \rho_i \geq 0, \forall i \backslash (i'), \forall i \in I, (B') \\ & \rho_{i'} = 1 \end{aligned} , \tag{5}$$

where i' is the index of the branching variable. The depth first strategy [14] is adopted, where the search goes down the tree until it reaches the first binary solution or reaches infeasibility, and then it backtracks the non-visited nodes recorded by a last-in-first-out stack. Therefore, we first go down to sub-problem (4). The optimal variables and optimal objective function is solved as $\boldsymbol{\rho}^{*(1)}$ and $\mathcal{P}^{*(1)}$ respectively. Again if any value of $\boldsymbol{\rho}^{*(1)}$ is not binary, problem (4) is split into two more sub-problems. This branch and bound process will be repeated until the optimal solution to the relaxed sub-problem satisfies all constraints with maximum objective function.

Algorithm 2. BBA

```
Set the best lower bound $\mathcal{P}^{(low)} = 0$ and $\boldsymbol{\rho} = \emptyset$;
Initialize the problem stack with the root problem as $S = \{\widehat{S}(\mathcal{P}^{(0)})\}$;
while $S \neq \emptyset$ do
    Pop a node problem $\widehat{S}(\mathcal{P}^{(j)})$ from stack $S$;
    Solve $\widehat{S}(\mathcal{P}^{(j)})$ to obtain the optimal variables; $\boldsymbol{\rho}^{*(j)}$ and the optimal objective value $\mathcal{P}^{*(j)}$;
    if $\mathcal{P}^{*(j)} \geq \mathcal{P}^{(low)}$ then
        if $\boldsymbol{\rho}^{*(j)}$ are all integers then
            Set $\mathcal{P}^{(low)} = \mathcal{P}^{*(j)}$ and $\boldsymbol{\rho}^{*} = \boldsymbol{\rho}^{*(j)}$;
            Delete $\widehat{S}(\mathcal{P}^{(j)})$ from stack $S$, i.e., $S := S \backslash \widehat{S}(\mathcal{P}^{(j)})$;
            Continue;
        else
            Branch $\widehat{S}(\mathcal{P}^{(j)})$ into two sub-problems $\widehat{S}(\mathcal{P}_1^{(j)})$ and $\widehat{S}(\mathcal{P}_2^{(j)})$ as (4) and (5);
            Push $\widehat{S}(\mathcal{P}_1^{(j)})$ and $\widehat{S}(\mathcal{P}_2^{(j)})$ to $S$;
        end if
    end if
    Delete $\widehat{S}(\mathcal{P}^{(j)})$ from stack $S$, i.e., $S := S \backslash \widehat{S}(\mathcal{P}^{(j)})$;
end while
```

3.3 Priority Aggregating Heuristic

The OAS improves performance by maximizing the utilization efficiency of PRB, while the potential loss in performance comes from increase of the waiting time, especially in high arrival rate, a sharp increase of the waiting time. Therefore, we design a Priority Aggregating Heuristic (PAH) that can be implemented in real system. The PAH is identical to adopting the greedy algorithm (GA) instead of BBA in PBBA. The GA has given in Algorithm 3. It work as follows. First, define six pointers, which respectively point to the HOL packet of the six virtual queues. Then get the packet which is the largest weight value among six packets that the six pointers point to. If the size of the aggregating buffer is smaller than L_{max} when the packet is put into the aggregating buffer, the packet is marked as $\rho = 1$ and put into the aggregating buffer. If not, marked as $\rho = 0$ and the pointer of the corresponding packet is moved to the next packet, until all the packets of the virtual queues have been marked. This is a low-complexity implementation of OAS that does not consider complex BBA. The PAH is an online algorithm that generates one aggregated packet at a time. If there are n packets in the virtual queues, the algorithm is O(n).

4 Performance Evaluation

4.1 Simulation Setups

The network model Setup: The $1000 \times 1000\,\mathrm{m}^2$ square simulation scenario is set up, where RN is placed near the centre, UEs are uniformly and independently

Algorithm 3. GA

```
Set six pointers which point to the head-of-line (HOL) packet in virtual queues;
The current size $S$ == 0;
**for** all the packets in virtual queues **do**
    $CP$ == the packet of the maximum weight value $v$ among six pointers;
    **if** $S$ + the size of packet $CP > L_{max}$ **then**
        Set $\rho_{CP} = 1$;
        $S = S$ + the size of packet $CP$;
    **else**
        Set $\rho_{CP} = 0$;
    **end if**
    The corresponding pointer of $CP$ point to the next packet;
**end for**
```

distributed in RN's coverage, and DeNB is randomly dropped beyond RN's coverage but can smoothly communicate with RN to perform packets aggregation. RN's coverage radius is approximately 350 m. Each UE send packet to RN randomly with the access link (Uu). We assume that the number of packets arrived in RN is satisfied with poisson distribution. The small-sized M2M data packets are considered in the simulation, and have the randomly size between 21 and 120 bytes and the randomly priority between 1 and 6. The DeNB allocates only five PRBs to the RN, and the position of RN corresponds to MCS 20 with a TBS 2344 bits per TTI of 1 ms duration.

Comparison Scenarios: (a) In the first group, M2M data packets are forwarded in PDCP of RN without aggregating. (b) In the second group, the data packets are aggregated at the Uu PDCP layer, in which the aggregated packet are served when their size is equal to the maximum total sizes L_{max} for all the packets or the waiting time reach the maximum waiting time T_{max}. (c) In the third group, the data packets are aggregated at the Uu PDCP layer by PBBA, but set different T_{max} for each queue. (d) In the fourth group, we the data packets are aggregated by PBBA but use the GA to replace BBA. The four groups respectively are No Aggregating Scheme (NAS), Simple Aggregating Scheme (SAS), the proposed Optimal Aggregating Scheme (OAS) and the proposed Priority Aggregating Heuristic (PAH). The maximum waiting time of the second group are set as 10 ms. However, the maximum waiting time T^{p}_{max} are set as {0, 0.01, 0.1, 1, 10, 100} seconds for the virtual queues p={1, 2, 3, 4, 5, 6}.

4.2 Performance Analysis

The simulation results in Fig. 4 clearly show the efficient utilization of PRBs in all four aforementioned schemes. In the NAS scenario with 200 data packet, the average number of PRBs usage is almost utilize 1 PRB, and in the case of SAS, only half of the PRBs are used with 200 data packets. However, in the case of OAS, the average number of PRBs usage is the lowest in comparison of NAS, SAS and OAS. The average number of PRBs usage is slightly higher than the

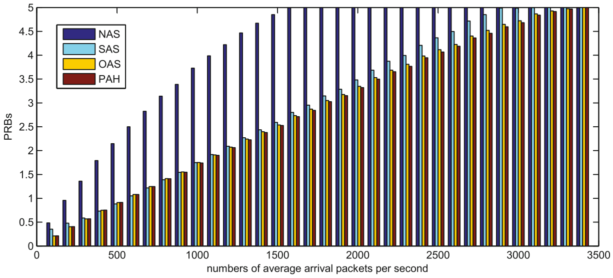

Fig. 4. Average number of PRBs usage for the four groups.

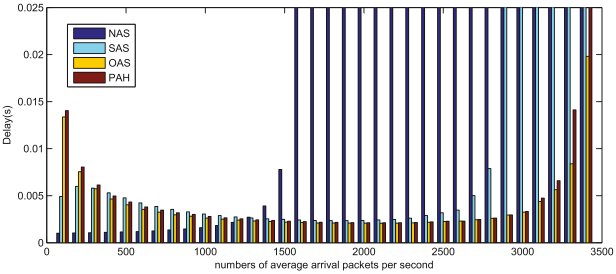

Fig. 5. Average waiting delay of all the packets for the four groups.

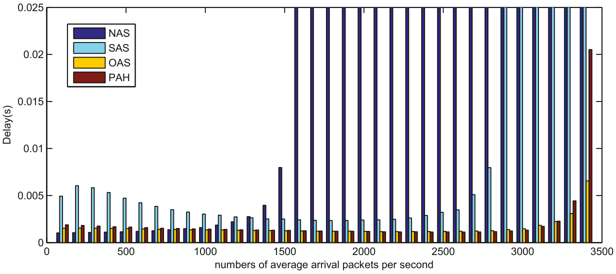

Fig. 6. Average waiting delay of the highest priority for the four groups.

case of SAS between 400 and 1000 data packets. Because the high priority data packets are sent quickly when reach the RN, and the size of all the packets aren't fill with the its maximum capacity. In particular, in low loaded scenarios, the average number of PRBs usage in the case of OAS is lower than the case of SAS, because the packets of lower priority are aggregated that the waiting time T_{max} of the lower priority is set more longer than the second group. Compared with the case of OAS, the average number of PRBs usage is slightly of lower in the PAH scenario, but the total delay is slightly higher. Because the purpose of OAS is to use the least amount of time, while PAH is to get to send more packets.

Figure 5 plots the average waiting time in the PDCP of RN in all four aforementioned schemes. In the case of NAS, its average waiting time more than the waiting time of SAS and OAS when the average number of packet arrival exceed 1400 data packets, because its average number of PRBs usage reach the maximum capacity. However, in the case of SAS with 2900 data packets, the average number of packet arrival is much larger than the case of OAS, because its average number of PRBs usage also reach the maximum capacity. In contrast, in low loaded scenarios, the average waiting time of all the packets of OAS are higher than the cases of NAS and SAS, this is due to the maximum waiting time of lower priority queue is higher than the high priority queue and thus the whole waiting time increase. However, the average waiting time of the highest priority is maintained at a low level as depicted in Fig. 6, because the packets of the highest priority are immediately forwarded to DeNB when arriving at the RN.

5 Conclusion and Future Work

In this paper, we explored the problem of optimal packets aggregation among M2M applications. A new framework for optimal aggregation implemented in the PDCP of RN, which features balancing a tradeoff between QoS requirements of packets and the utilization efficiency of PRBs. Numerical results illustrate that OAS achieves a tradeoff of QoS and PRB utilization efficiency in comparison with four existing schemes. Last but not least, OAS provides an optimal scheduling scheme for uplink M2M traffics, and more key techniques will be systematically studied for the downlink M2M traffics in our future work.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61302108).

References

1. Wang, K., Alonso-Zarate, J., Dohler, M.: Energy-efficiency of LTE for small data machine-to-machine communications. In: Proceedings of the IEEE ICC, pp. 4120–4124 (2013)
2. Wu, G., Talwar, S., Johnsson, K., Himayat, N., Johnson, K.D.: M2M: From mobile to embedded internet. IEEE Commun. Mag. **49**(4), 36–43 (2011)
3. Mehmood, Y., Khan Marwat, S.N., Görg, C., et al.: Evaluation of M2M data traffic aggregation in LTE-A uplink. In: Proceedings of the ITG-Fachbericht-Mobilkommunikation, pp. 24–29, August 2015
4. Marwat, S.N.K., Mehmood, Y., Görg, C., Timm-Giel, A.: Data aggregation of mobile M2M traffic in relay enhanced LTE-A networks. EURASIP J. Wirel. Commun. Networking **2016**(1), 1–14 (2016)
5. Devi, U.M., Goyal, M., Madhavan, M., et al.: SERA: A hybrid scheduling framework for M2M transmission in cellular networks. In: Proceedings of the IEEE COMSNETS, pp. 1–8 (2015)
6. Majeed, A., Abu-Ghazaleh, N.B.: Packet aggregation in multi-rate wireless LANs. In: Proceedings of the IEEE SECON, pp. 452–460 (2012)

7. Sawabe, A., Tsukamoto, K., Oie, Y.: QoS-aware packet chunking schemes for M2M cloud services. In: Proceedings of the IEEE WAINA, pp. 166–173 (2014)
8. http://www.3gpp.org/technologies/keywords-acronyms/97-lte-advanced. Accessed 12 Sep 2016
9. 3GPP TS 23.107, 3rd Generation Partnership Project. Technical Specification Group Services and System Aspects. Quality of Service (QoS) concept and architecture (Release 13), V13.0.0, December 2015
10. Liu, R., Wu, W., Zhu, H., et al.: M2M-oriented QoS categorization in cellular network. In: Proceedings of the IEEE WiCOM, pp. 1–5 (2011)
11. 3GPP TS 36.300, 3rd Generation Partnership Project. Technical Specification Group Radio Access Network. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) (Release 12), V13.4.0, June 2016
12. 3GPP TS 36.213, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 13), V13.2.0, June 2016
13. Sahni, S., Gonzalez, T.: P-complete approximation problems. J. Assoc. Comput. Mach. **23**(3), 555–565 (1976)
14. Lawler, E.L., Wood, D.E.: Branch-and-bound methods: A survey. J. Oper. Res. **14**(4), 699–719 (1966)