EVALUATION OF LATENCY-AWARE SCHEDULING TECHNIQUES FOR M2M TRAFFIC OVER LTE

I.M. Delgado-Luque, F. Blánquez-Casado, M. Garcia Fuertes, G. Gomez, M.C. Aguayo-Torres, J.T. Entrambasaguas, J. Baños

AT4 wireless, S.A., Málaga, Spain
University of Málaga, Málaga, Spain

ABSTRACT

Machine to Machine (M2M) communications are expected to grow dramatically in next years. Scheduling techniques are determinant to achieve high spectral efficiency in wireless systems and to provide QoS guarantees to system users. In this work, several scheduling algorithms are evaluated in order to accommodate delay limited M2M communications over an LTE system. Simulation results show a reduction of mean and 95th percentile packet delay.

Index Terms— Scheduling, LTE, Machine to Machine, delay

1. INTRODUCTION

Some estimations foresee the world could have a trillion communicating devices (including sensor and RFID networks) in a decade [1]. It is expected that most of them will be wireless and a high percentage will not be operated directly by humans. This Machine to Machine (M2M) communications have certain characteristics which makes them hard to accommodate in mobile networks. In particular, certain flows are very sensitive to delay.

There are M2M applications, such as vehicle collision detection and avoidance, sensor-based alarms and remote control, etc., that require extremely low latency values. In remote video surveillance type applications, for instance, Unmanned Ground Vehicles (UGV) and Unmanned Aerial Vehicles (UAV), devices carrying video cameras (usually robots) are remotely controlled by a human, or by an intelligent control system, based on the video information captured by the camera and transferred to the control point. In terms of latency the video flow is more critical than the control signal. Assuming the use of optimized video codecs for real-time video transmission, the latency introduced by the network, and in particular by the Medium Access Control (MAC) and physical (PHY) layers [2], becomes important in the latency budget.

3GPP is working to improve its Long Term Evolution (LTE) standard performance for M2M applications in next LTE-Advanced releases [3]. Efforts are being put on the provision of enhancements for RAN overload control for Machine-Type Communications (MTC) [4] and on the provision of low-cost MTC User Equipment based on LTE [5]. However, there is little or no work identified on the provision of enhancements for latency-constrained M2M applications. The use of latency-aware and optimized MAC schedulers is therefore crucial so that these latency-constrained applications are allocated with the necessary physical resources to ensure proper operation.

In a multiuser system, an optimized MAC layer should allocate radio resources to users according to several parameters, including traffic source characteristics, Quality of Service (QoS) needs [6] and the frequency, time and space diversity of the mobile radio channel.

When traffic sources are variable-rate and their transmission resources requirements fluctuate asynchronously for different users, exploiting the source multiuser diversity (statistical gain) allows more users to be accommodated on the system. A MAC design adaptable to the changing traffic and channel characteristics and to the specific QoS requirements, i.e. a cross-layer design, improves the system performance by exploiting the radio resources more efficiently [7].

Several scheduling algorithms over LTE have been proposed in the literature [8]. Literature is also extensive on heuristic approaches which take into account the source behaviour (see [9] and [10]). The chosen techniques should be sufficiently flexible to accommodate traditional traffic sources as well and the existence of other sources with different QoS requirements [6]. In LTE, the eNodeB is responsible for implementing the scheduling policies both in the downlink and uplink scenarios and the UEs are informed about the resource allocation decisions on a subframe basis, through control channels.

In this paper, we focus on the evaluation of resource allocation algorithms for M2M communications over LTE. Work is partially result of EU FP7 Project LOLA

(Achieving Low-Latency in Wireless Communications), whose traffic analysis for M2M has been presented in [11]. Advanced scheduling algorithms have been chosen considering different aspects like the instantaneous channel conditions, latency requirements or pending retransmissions in order to adequately allocate transmission turns to users.

The remainder of this paper is organized as follows. Section II briefly describes the proposed resource allocation algorithms. In Section III the link-based simulator used to evaluate these algorithms is introduced. Scheduler performance is then presented in Section IV. Finally, some concluding remarks are given in Section V.

2. DESCRIPTION OF THE PROPOSED SCHEDULERS

Three scheduling algorithms chosen from the literature have been analyzed for delay-dependent traffic coming from a M2M traffic source: Opportunistic hard priority [12][13], Channel Dependent Earliest Deadline Due (CD-EDD) [14] and CD-EDD with postponed EDD term [15].

2.1. Opportunistic Hard Priority

This algorithm applies opportunistic priority to the transmissions of delay-sensitive flows if a maximum packet delay threshold is exceeded.

The algorithm sets the same priority to all packets as long as packet delay is below the threshold and sets high priority to packets exceeding the threshold until they are served. A delay threshold D_t and a delay budget D_b is assigned to each delay-sensitive flow. LTE defines a delay budget for the radio interface of 80 ms for VoIP traffic [16]. The delay threshold must be chosen to be lower than the delay budget with a sufficient margin so that the scheduler can serve packets exceeding this threshold before violating the delay budget. Packets exceeding the delay budget are discarded.

Initially, users are cyclically ordered based on arrival time to their transmission queues. The waiting time of the head of line (HOL) packet in each queue is continuously monitored. If the HOL packet delay of a user exceeds the delay threshold the Opportunistic Hard Priority scheduler will give priority to the transmission of such packet until it gets served (strict prioritization).In contrast, Proportional Fair (PF) policy can be seen as a kind of soft prioritization scheduling (assuming similar source traffic characteristics): users with worse average channel conditions are expected to suffer higher delay; to compensate it, PF policy indirectly prioritizes them through the inverse of the average throughput applied to the instantaneous potential rate.

The algorithm implemented by the Opportunistic Hard Priority scheduler is as follows:

1. Set the delay parameters for each data flow: delay budget and delay threshold
2. Set the priorities for each data flow:
   a) If (D^b – HOL packet delay) < 0 → the packet is discarded.
   b) else if (D^b – HOL packet delay) < D^t → priority = 1
   c) else priority = 0
3. Sort the list of data flows from highest to lowest priority.
   a) Allocate a set of free Physical Resource Blocks (PRB) to the flow with the highest priority in the list if, and only if, the following conditions are met:
      - There are non-assigned PRBs.
      - There is data waiting to be served.
      - There is at least one free HARQ channel.
      - The data flow has not received the maximum allowable number of assignments in the current Transmission Time Interval (TTI).

      If there is more than one data flow with the same priority, a Round Robin (RR) scheme is applied.
   b) Remove the served data flow from the list.
   c) If it is possible to continue, go to step 3.a).

2.2. Channel Dependent Earliest Deadline Due (CD-EDD)

The priority assigned by this scheme depends on two components: the delay-aware component (EDD) and the channel-aware component (PF). The EDD term works in such a way that users are prioritized as their HOL packet delay gets close to the delay budget. The PF term favours terminals with temporarily good channel conditions. The scheduling tag assigned to user k is therefore calculated according to the following formula:

$$\frac{T_k[n]}{R_k[n]} \cdot \frac{W_k[n]}{D_k^b - W_k[n]} \tag{1}$$

where:

- $T_k[n]$ is the throughput of user k at TTI n
- $R_k[n]$ is the average throughput of user k up to TTI n
- $W_k[n]$ is the waiting time of the HOL-packet in the queue of user k (expressed in TTIs)
- D_k^b is the maximum allowable delay (or delay budget) of user k (in TTIs)

The PF and EDD terms are well differentiated in the formula. The term associated to the PF algorithm is the first quotient ($\frac{T_k[n]}{R_k[n]}$) and the EDD term is the second quotient ($\frac{W_k[n]}{D_k^b - W_k[n]}$). As the delay of the HOL packet gets close to D_k^b, the EDD term quickly dominates the scheduling tag.

For low HOL packets delays, the PF term dominates the scheduling tag. As for the Opportunistic Hard Priority algorithm, packets exceeding the delay budget are discarded.

2.3. CD-EDD with postponed EDD term

This scheme is a modification of the previous algorithm, affecting the delay-aware term. The PF term will now dominate the scheduling tag as long as HOL packet delays are far from exceeding the delay budget. Such approach is based on a simple utility function:

$$\frac{T_k[n]}{R_k[n]} \cdot \left(\frac{\max(0, W_k[n] - D_k^t)}{D_k^b - W_k[n]} + 1 \right) \quad (2)$$

where D_k^t is the minimum delay (or delay threshold) associated to user k. The delay-aware term will provide priority to users whose HOL packets are greater than this threshold (D_k^t). This approach takes advantage of the flow delay tolerance in order to increase the system capacity. Again, packets violating the delay budget are discarded.

The objective of these three algorithms is to ensure that the instantaneous packet delay is kept below a certain value. If a packet waiting in the queue has exceeded the delay budget, the system discards such packet.

3. LINK-BASED SIMULATION ENVIROMENT

A proprietary simulation environment oriented to model and simulate complex MIMO-OFDM wireless systems has been used. The simulation environment is composed of a number of UEs connected to a Base Station (BS) through a frequency-selective Rayleigh fading channel. The high-level architecture for the direct link is depicted in Figure 1. The simulation environment has been adapted to implement and evaluate the proposed algorithms over an LTE-like system.

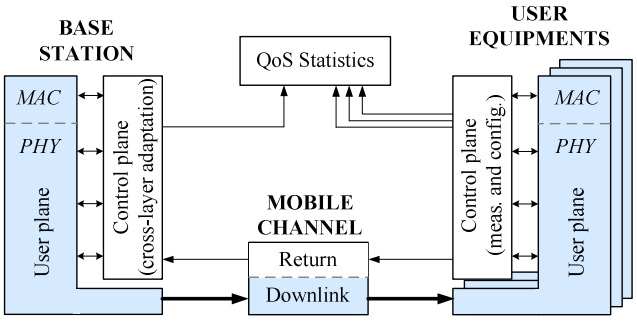

Figure 1. Simplified DL link-based simulation environment architecture

Baseline simulations have been done in order to evaluate the latency reduction achieved with different scheduling techniques [17]. A PF algorithm has been chosen as the baseline scheduling algorithm to compare latency results. A summary of the simulation parameters is shown in Table 1.

Table 1. Parameter settings

LTE Parameter	Value/Mode
Channel model	Extended pedestrian A (TS 36.803)
Mobile Speed	4 km/h (pedestrian)
Channel Bandwidth	20 MHz
OFDM symbols per TTI	14
PRB size	12 subcarriers
Carrier Frequency	2.5 GHz
Modulation schemes	QPSK, 16QAM and 64QAM
Target BLER	10%
ACK feedback delay	8 ms
CQI delay	2 ms
Nº of CQI bits	4
Max. Number of HARQ retransmissions	8
Number of parallel HARQ processes	8
MIMO mode	2x2 1 layer spatial multiplexing (Beamforming)
Codebook	TS 36.211
Channel Estimation	Non-ideal Zhao
MIMO detection	ZF
Noise power estimation	Error based
Signalling overhead	2/21
Number of users	10
Simulation length	20s
Type of traffic	M2M

The M2M traffic source used in this work is an IP video surveillance camera [11] transmitting a video flow for a delay-sensitive remote control M2M application. This traffic application can be mapped to the standardized QoS Class Identifier (QCI) number 7 defined in 3GPP specifications [16]. For this QCI, the maximum packet delay allowed is 100 ms. The following delay parameters have been associated for the algorithms to be analyzed:

Table 2. Parameters of the proposed schedulers

Opportunistic Hard Priority		CD-EDD		CD-EDD with postponed EDD term	
D^b	100 ms	D^b	100ms	D^b	100 ms
D^t	50 ms			D^t	50 ms

4. RESULTS

Results of mean and 95th percentile packet delay, packet loss rate and throughput per user are shown in Figure 2, Figure 3, Figure 4 and Figure 5, respectively, for the proposed scheduling algorithms. These results are discussed in following subsections.

When the HOL packet delay exceeds the delay budget, the three algorithms studied in this paper discard such packet

instead of incrementing the delay of the remaining packets to be served (which could be more damaging for the QoS).

Discarding packets is more likely to happen when the mean SNR is low (0-5 dB). In these conditions, the throughput per user is very low (see Figure 5) due to the high outage probability and the need of using robust coding schemes that ensure the target BLER. As a consequence, the probability of a delay budget violation increases and thus, a high number of packets have to be discarded so that delay results are kept below the delay budget value (100 ms), which implies a reduction near to 90% compared to the baseline results. However, for these simulation conditions, the discarding packet process does not imply a reduction of the average throughput compared with the baseline results (see Figure 5). This is because the achievable throughput in such conditions is lower than the load of the traffic source, so there will always be packets queued.

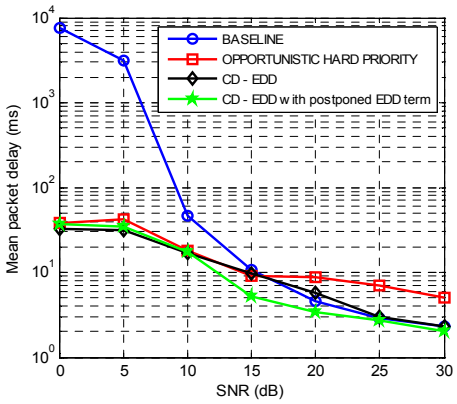

Figure 2. Mean packet delay results

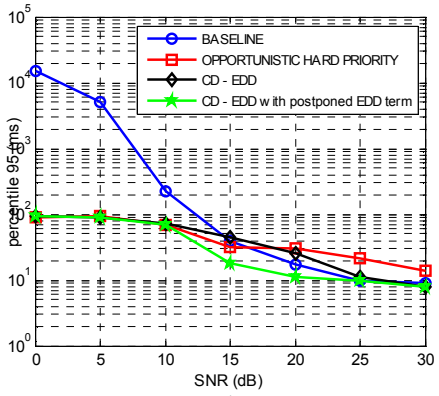

Figure 3. Percentile 95th of mean packet delay

When the mean SNR is increased, the packet loss rate decreases reaching 0 at a level of 15 dB (see Figure 4). Then, packet delay results are mostly influenced by the utility function of each algorithm. Also, for the three proposed algorithms, the average throughput is a bit lower than for the baseline (see Figure 5), as these algorithms favour users experiencing high packet delays. These users have in general worse channel conditions, so their throughput is consequently lower.

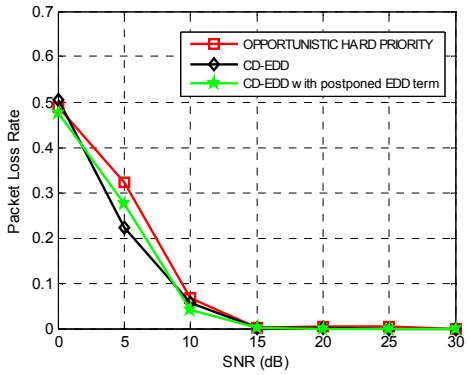

Figure 4. Packet loss rate

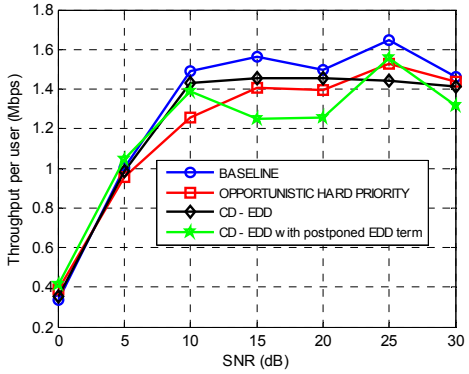

Figure 5. Throughput per user

4.1. Opportunistic Hard Priority

When mean SNR is increased, the performance of the opportunistic hard priority is similar to a RR algorithm: there are a lower number of packets whose waiting time exceeds the delay threshold; therefore, the algorithm will set the same priority to almost all data flows, which will be allocated in a cyclic order. This is the reason why the baseline configuration (based on a PF scheme) achieves a better performance for high SNR values.

4.2. Channel Dependent Earliest Deadline Due (CD-EDD)

For high SNR levels, mean packet delay associated to the CD-EDD technique gets slightly higher values than those associated to the baseline configuration. This is because HOL packet delays are low in such scenario and the EDD-term gives a rather low priority, thus degrading the performance compared to the case in which the PF-term dominates the scheduling decision. However, the CD-EDD improves the opportunistic hard priority algorithm.

4.3. CD-EDD with postponed EDD term

Packet delay results of the CD-EDD with postponed EDD term are also shown in Figure 2 and Figure 3. This algorithm

improves packet delay results for the whole range of SNR values, in contrast to the results obtained with previous algorithms. As the value of the SNR is increased, HOL packet delays decrease, so the probability of exceeding the D^t value is also decreased. Thus, the CD-EDD algorithm works as a PF algorithm. When the HOL packet delay of a user is above the D^t the EDD term gives higher priority to that user, which decreases the average packet delay results (especially for medium SNR values). Therefore, SNR values do not lead to lower priorities, as it occurs for the CD-EDD.

Regarding the configuration values of D^b and D^t parameters, a reduction of these values would improve delay results as higher priority would be given to users experiencing high packet delays (generally due to worse channel conditions) at the expense of a system throughput reduction and an increase of the packet loss rate.

5. CONCLUSSIONS

This paper presents a performance comparison between three delay-aware scheduling algorithms for M2M traffic over an LTE system.

Simulation results show that, for low SNR values (between 0 and 15 dB), the three delay-aware algorithms are able to reduce considerably the mean and 95th percentile packet delay at the expense of discarding those packets exceeding the allowable delay. For high SNR values (from 15 to 30 dB) packet delay results are mostly influenced by the utility function of each algorithm. In that sense, the algorithm called CD-EDD with postponed EDD term achieves the best performance in terms of delay, as the influence of the HOL packet delay is not always affecting the scheduling decision.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7) under the LOLA project (Achieving Low-Latency in Wireless Communications) grant agreement Nº 248993. This work has also been performed in the framework of the Junta de Andalucía (Proyecto de Excelencia P07-TIC-03226) and the Spanish Government under the project TEC2010-18451.

7. REFERENCES

[1] IDC Estimates, “The Expanding Digital Universe”, available at http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

[2] EU FP7 Project LOLA (Achieving Low-Latency in Wireless Communications), Project nº 248993, D2.1 Target application scenarios, v1.0, May 2010.

[3] 3GPP TR 36.912 V10.0.0 (2011-03), “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Feasibility study for Further Advancements for E-UTRA (LTE-Advanced) (Release 10)”.

[4] 3GPP TSG-RAN Meeting #53, RP-111373. “3GPP™ Work Item Description: RAN overload control for Machine-Type Communications”. Fukuoka (Japan), 13 – 16 September, 2011.

[5] 3GPP TSG-RAN Meeting #53, RP-111112. “3GPP™ Work Item Description: Provision of low-cost MTC UEs based on LTE”. Fukuoka (Japan), 13 – 16 September, 2011.

[6] 3GPP TS 23.107 V8.2.0 (2011-12), “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Quality of Service (QoS) concept and architecture (Release 8)”.

[7] J.T. Entrambasaguas, M.C. Aguayo-Torres, G. Gómez and J.F. Paris, “Multiuser capacity and fairness evaluation of channel/QoS-aware multiplexing algorithms,” IEEE Network, pp. 24-30, May-June 2007.

[8] H. Luo, S. Ci, D. Wu, J. Wu, H. Tang, "Quality-driven cross-layer optimized video delivery over LTE". IEEE Communications Magazine, vol. 48, no. 2, pp. 102-109, February, 2010.

[9] B. Sadiq, R. Madan, A. Sampath, “Downlink Scheduling for Multiclass Traffic in LTE”. EURASIP Journal on Wireless Communications and Networking, vol. 2009.

[10] Y. Qian, C. Ren, S. Tang, M. Chen, "Multi-service QoS guaranteed based downlink cross-layer resource block allocation algorithm in LTE systems". International Conference on Wireless Communications & Signal Processing, pp. 1-4, November, 2009.

[11] EU FP7 Project LOLA (Achieving Low-Latency in Wireless Communications), Project nº 248993, D3.2 Network related analysis of M2M and online-gaming traffic in HSPA, v1.0, June 2010.

[12] S. Choi, K. Jun, Y. Shin, S. Kang and B. Choi, “MAC Scheduling Scheme for VoIP Traffic Service in 3G LTE”. Proceedings of IEEE Vehicular Technology Conference (VTC-Fall). Baltimore (USA), October 2007.

[13] M. Andreozzi, G. Stea, A. Cacioccola and R. Rossi, “Flexible scheduling for real-time services in High-Speed Packet Access cellular networks”. European Wireless 2009, Aalborg (Denmark), May 2009.

[14] A. K. F. Khattab and K. M. F. Elsayed, ”Channel-quality Dependent Earliest Deadline Due Fair Scheduling Schemes for Wireless Multimedia Networks”. Proceedings of MSWiM 2004, Venice (Italy), October 2004.

[15] G. Barriac and J. Holtzman, ”Introducing Delay Sensitivity into Proportional Fair Algorithm for CDMA Downlink Scheduling”. Proceedings of ISSTA 2002, Vol. 3, pp. 652-656, Parsippany (USA), September 2002.

[16] 3GPP TS 23.203 V11.3.0 (2011-09), “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 11)”.

[17] EU FP7 Project LOLA (Achieving Low-Latency in Wireless Communications), Project nº 248993, D4.5 Scheduling Policies for M2M and Gaming Traffic, v2.0, January 2012.