SPECIAL ISSUE - M2M

Analytical modelling and performance evaluation of realistic time-controlled M2M scheduling over LTE cellular networks
Antonis G. Gotsis*, Athanasios S. Lioumpas and Angeliki Alexiou
Department of Digital Systems, University of Piraeus, Piraeus, Greece

ABSTRACT
Supporting emerging machine-to-machine (M2M) communications over Long-term Evolution (LTE)/LTE Advanced cellular networks in an efficient way will be beneficial for both telecommunication communities. The first step to migrate to an M2M-enabled cellular standard is to provide these new services through the existing architectures and protocols, while maintaining seamless backward compatibility. To this end, we thoroughly examined a key LTE Medium Access Control entity, which is the packet scheduler, and proposed solutions based on the time-controlled M2M feature, to deal with the diverse M2M traffic characteristics and quality-of-service requirements. Starting from the single M2M class case, we extended our study to more realistic scenarios, involving more M2M classes with diverse quality-of-service requirements. We defined analytical models for predicting the system performance on the basis of queueing theory concepts and considered the interaction between classes with different priorities. The proposed analytical models are validated through extensive system-level simulations. On the basis of the insight obtained from our analytical approach, we modified an existing scheduling algorithm to improve the performance of low-priority M2M device groups, and we demonstrated its superior performance both experimentally and analytically. Copyright © 2013 John Wiley & Sons, Ltd.

*Correspondence
Antonis G. Gotsis, Department of Digital Systems, University of Piraeus, 126 Grigoriou Lampraki Street, Office 501, GR18532, Piraeus, Greece.
E-mail: agotsis@unipi.gr

Received 28 May 2012; Revised 30 November 2012; Accepted 2 February 2013

1. INTRODUCTION
The machine-to-machine (M2M) communications paradigm complementing the classical human-to-human (H2H) interaction model constitutes an emerging market, for which recent market analyses predict more than 500 million embedded M2M connections by 2014 [1, 2]. Mobile broadband stakeholders, including standardisation organisations, have identified the need for supporting M2M applications through cellular networks. The 3rd Generation Partnership Project (3GPP) has formed dedicated study items under the umbrella term of ‘machine-type communications’ (MTC) for Release 10 and beyond [3–5], whereas IEEE has initiated the IEEE 802.16p and 802.16.1b projects oriented towards amending the latest IEEE 802.16m standard with M2M support (for example, [6]). The M2M ecosystem spans a wide range of applications, including smart metering (industrial and energy/smart grid), health monitoring and alerting, and intelligent transportation (fleet management, car-to-car collision avoidance, etc.) [7], characterised by the following: (i) the large number of simultaneously connected devices; (ii) small data volume transmissions; and (iii) vastly diverse quality-of-service (QoS) requirements compared with classical mobile broadband services.

In this work, we focus on how M2M communications could be efficiently supported by current and future 3GPP mobile broadband standards beyond Wideband Code Division Multiple Access/High-Speed Packet Access, providing a smooth evolution of Long-term Evolution (LTE) and LTE Advanced towards M2M-enabled cellular networks. Packet scheduling is the key radio resource management mechanism for guaranteeing QoS requirements while minimising the overall resource usage [8]. Assuming that the existing cellular resources will initially support both H2H and M2M services and taking into account that M2M applications significantly differ from H2H applications, in terms of traffic and QoS profiles, designing efficient M2M-over-LTE scheduling schemes constitutes a major challenge.
1.1. Motivation—prior work

The packet scheduling framework considered herein endorses three key features, namely compatibility with LTE scheduling principles, time-controlled M2M transmissions, and low signalling overhead. The first feature indicates that although specific scheduling algorithms are beyond the scope of the standardisation procedure, the associated scheduling decisions should be based on a set of standardised QoS indicators, or under the 3GPP terminology, ‘QoS class identifiers’ (QCIs), defined in [9] and including priority levels, packet delay budget, and dropped packet rate requirements due to channel conditions. The second feature dictates that M2M devices may not be granted resources at any frame but at predefined time intervals. The last one presumes that low-complexity periodic-like scheduling grants should be allocated to M2M devices, to reduce both forward (information grant messages) and feedback (scheduling request messages) signalling, especially when M2M devices are multiplexed together with H2H users within an Orthogonal Frequency Division Multiplexing (OFDMA) frame as in LTE.

In [10, 11], a massive access management framework was proposed, based on the following: (i) organising the M2M devices into clusters or classes sharing the same QoS requirements and (ii) allocating fixed access grant time intervals (AGTIs) to each class, on the basis of the traffic rate and the priority of each class. Our work follows this particular framework because it adopts the time-controlled/periodic-like scheduling feature but substantially extends it to overcome its major limitations: (i) only constant-rate traffic patterns were considered, which stand in contrast to the randomness of real M2M traffic [12]; and (ii) only a loose deterministic bound was proposed for the average experienced delay, whereas it is more realistic to guarantee probabilistic packet delay and dropped rate performance indicators (e.g. percentile values compared with a threshold) [9].

Related works conducted in the context of LTE scheduling (see for example [13, 14] for LTE-only traffic and [15] for mixed LTE/M2M traffic) provide fully dynamic (per transmission time interval, TTI) QoS-aware, queue-aware, and channel-aware scheduling solutions. Although the performance of those schemes is quite high, it comes at the expense of huge signalling loads because of the following: (i) at each TTI, each MTC device should report its buffer status along with its channel quality to the base station; and (ii) the base station should inform each MTC device which resource block (RB) is granted. We have to point out that our work differs from the preceding LTE-related approaches because the key feature of our scheduling framework is the low signalling overhead. This is accomplished by the following: (i) basing the scheduling decisions only on QCIs (which are known at the beginning of an M2M connection); (ii) avoiding feedback signalling for buffer status and channel quality reporting; and (iii) utilising minimum forward signalling for just communicating to each M2M device the decided scheduling period, and not the exact granted TTIs.

In [16], several resource allocation schemes for M2M communications over OFDMA are proposed. Nevertheless, the authors focused on two-hop architectures (M2M gateway/LTE base station) instead of single hop, as in our work, and targeted energy savings instead of satisfying prescribed QCIs. In addition to the aforementioned works, which are more or less directly applicable to existing standards, there are several works that aim at theoretically exploring the limits of OFDMA-based scheduling under throughput [17], fairness [18], or delay-based QoS constraints [19, 20]. For example, the authors in [19] investigated the performance of an optimal off-line scheduler, which is aware of the future requests of all users. Such works could serve as benchmark solutions for practical algorithm performance evaluation.

1.2. Contributions

A realistic scheduling framework is extensively covered in this work, which does the following:

- adopts the time-controlled 3GPP-MTC feature through fixed pre-allocated pattern grant decisions,
- handles realistic heterogeneous event-driven M2M traffic (Poisson-modelled bursts), and
- serves traffic according to LTE-related QCIs, namely priority levels along with statistical packet delay/dropped packet rate requirements.

The contribution of this paper is threefold. First, considering M2M devices sharing homogeneous traffic characteristics and QoS requirements, we propose an exact analytical model that combines statistical QoS performance indicators (in terms of a delay threshold violation probability) with the allocated grant period. This model could be utilised either for conducting a performance assessment study of M2M over LTE without the need of running time-consuming simulations or for estimating the minimum grant period to meet prescribed statistical QoS requirements. The latter is essentially an accurate bandwidth estimation method for M2M-type traffic. Second, building on the previous model, we examine the more realistic multiple-M2M-class scenario, where lower-priority M2M devices may postpone their transmissions in the presence of a higher-priority M2M device in the same grant. Regarding this scenario, we propose an approximating analytical model for quantifying the QoS performance loss due to this ‘grant collision’ phenomenon. Third, we propose a simple modification to the considered periodic-like fixed grant-scheduling algorithm, to improve the performance of low-priority (LP) M2M devices, along with an approximating model that predicts the improvement levels. The developed analytical models are based on concepts from queueing theory and validated through extensive system-level simulation experiments.
The rest of this paper is structured as follows. In Section 2, the system model is introduced along with the assumptions and the followed notations. In Section 3, the single-M2M-class scenario is considered, whereas in Section 4, the multiple-M2M-class scenario is thoroughly covered. Numerical results along with discussion points are presented in Section 5, whereas Section 6 summarises our work and states open problems.

2. SYSTEM & TRAFFIC MODEL

We consider a single-cell area served by a base station through an LTE-like air interface. Both M2M devices and H2H terminals (user equipment, UEs) are multiplexed in an OFDMA frame comprising hundreds of time-frequency resource elements, organised in RBs. Generated packets are stored in a queue, and transmissions occur in time bursts or TTIs. Part of the LTE bandwidth is reserved for traditional LTE-H2H UEs, and the remaining may be allocated to M2M devices. Although this fixed resource allocation scheme may be suboptimal, it prevents the deterioration of legacy LTE-UE performance due to massive M2M communications. M2M devices are organised in classes (or equivalently groups or clusters), where for each class, a traffic profile along with a QCI-based QoS profile is strictly defined. In particular, assuming an arbitrary (out of $M$) M2M class and for the sake of simplicity omitting the class index, each class is fully characterised by the following set of features:

- The number of M2M devices, $K$
- The average packet arrival intensity per device, $\lambda$ (in packets/TTI units)
- The priority index, $P \in \{1, 2, \ldots, M\}$
- The maximum delay threshold $\Delta$ (in TTI units) for each queued packet
- The probability that the preceding delay threshold is violated, $\delta$ (%)
- If $W$ is the waiting time for a queued packet, then a packet-delay-budget QoS criterion is expressed through $\text{Prob}\{W > \Delta\} \leq \delta$. We notice that two QoS approaches are supported: (i) no packet drops occur; hence, QoS is defined as the probability of the waiting time exceeding a predefined threshold; and (ii) packet drops occur when the delay threshold is violated; hence, QoS is given by the dropped packet rate.
- The scheduled grant period for each device $T_g$ assuming the AGTI-based algorithm [10, 11] is utilised by the base station.
- The offered load per device $\rho$ defined by $\lambda \cdot T_g = \lambda / \mu$, where $\mu$ is the average service rate in queueing terms.

Designing an effective scheduling algorithm requires the consideration of an accurate traffic model. M2M traffic modelling is a very active research area, and an exact model has not been defined yet, because of the vast M2M application range. However, recent works conducted in the context of standardisation efforts (3GPP MTC and IEEE 802.16p) [21, 22] and ongoing ICT-FP7 research projects (LOLA) [12] indicate the usage of Poisson-modelled traffic arrivals for M2M evaluation studies. This is justified by the fact that a significant amount of expected M2M traffic will be triggered by random events. Typical event-based M2M applications include autopilot vehicle collision detection and avoidance, virtual bicycle race, emergency team tracking, and sensor-based alarm or event detection [12]. Hence, we adopt the Poisson traffic model, and the results presented hereafter may be applied in any of the aforementioned application scenarios. Note that constant-rate [21] and uniform [22] traffic models have also been lately proposed in addition to the Poisson model.

We assume that each RB carries a single packet, is allocated a fixed amount of power, and is tuned according to a fixed single-bit modulation and coding mode. In summary, each class is associated with a tuple $\langle K, \lambda, P, \Delta, \delta \rangle$. To exemplify the system description, an indicative dual-class multiplexing scenario is illustrated in Figure 1. One easily observes the repetitive scheduling patterns for M2M device groups and that part of the second lower-priority M2M class grants are postponed for TTIs later than the prescribed ones because of collisions with high-priority (HP) grants. The latter will be elaborated in the multiclass scenario description in Section 4.

3. THE SINGLE-CLASS SCENARIO

We first deal with a scenario where a single class of M2M devices requests service from LTE. In [11], where the fixed grant-scheduling algorithm was first introduced, constant-rate M2M traffic was assumed; thus, the grant period $T_g$ was easily computed as the inverse generated packet rate. In addition, the authors proposed a deterministic bound on the average experienced delay of each class. In what follows, we advance that work by defining the minimum grant period $T_g$ to guarantee specific statistical QoS guarantees dictated by $\{\Delta, \delta\}$, when event-driven Poisson-like M2M traffic with intensity $\lambda$ is offered to the system.

Following the Kendall notation from queuing theory (for example, [23]), the particular problem may be modelled as an $M / D / 1$, where $M$ stands for Poisson traffic arrival (or equivalently exponential interarrival times) and $D$ for deterministic service time given by $T_g$. Notice that although the original queuing model presumes that the service time for each M2M packet will last $T_g$ TTIs, whereas in our case, it means that each M2M device will be served every $T_g$ TTIs, the mathematical analysis still holds.

First, we will express the cumulative waiting time distribution $\text{Prob}(W > i)$ ($i$ being an arbitrary TTI) for integer values of the grant period. According to [23, Section 10.4], this can be expressed as a function of the probabilities of equilibrium states $p(i)$, where an arbitrary
state $i$ corresponds to a volume of $i$ queued packets, as in Equation (1).

$$\text{Prob}(W > t) = 1 - \text{Prob}(W \leq t) = (1/\rho) \cdot (p(0) + p(1) + \ldots + p(i) + \ldots + p(t))$$

where $p(t)$ is the probability of being in state $t$, and $\rho$ is the arrival rate. The terms $p(0), p(1), \ldots, p(t)$ are given in Equation (2).

Equation (1) provides the probabilistic packet-loss rate (PLR) for a given TTI. The PLR is defined as the probability that a packet is lost due to the violation of the delay-budget QoS criterion when no drops occur. When packets are lost because of the violation of the $\Delta$ threshold, QoS is associated with the dropped packet rate. However, we can relate the two cases by a simple formula, found in [25, Equation (9.4.7)], that is,

$$\text{PLR} = (1 - e^{\lambda t}) = \frac{1}{1 - \lambda} \cdot \text{Prob}(W > t)$$

Combining all the preceding equations and given that $p(t)$ is the probability of being in state $t$, we obtain an analytical expression that relates the traffic intensity to the grant period and the QoS requirements. For known $\lambda$ and a target set of QoS criteria $[\Delta, \delta]$, the equations can be solved for $T_g$, and thus, the minimum grant period can be exactly estimated. Note that an explicit closed-form solution is not possible, and a numerical solver should be invoked as the solution is related to the Lambert W function [26].

4. THE MULTICLASS SCENARIO

In this section, extending the single-class case, we assume multiple coexisting M2M classes with a decreasing priority order served by a set of orthogonally shared resources. Without loss of generality, we consider a two-class scenario as depicted in Figure 1, where two independent device clusters are deployed over the same cell area. As will be explained later, extending our analysis to more than two classes is straightforward.
We first assume that the HP M2M class has been allocated the requested grants, to support the target QoS performance levels (refer to the single-class analysis). Then, the allocation to the LP class is performed, where one can observe that the constant service time pattern may be violated. As illustrated in Figure 1, for example, the M2M device 101, is granted RBs according to a repeat-interservice time pattern of \{18, 18, 24, 18, 18, 24, ...\} TTI instead of the desired \{18, 18, 18, ...\} TTI pattern. This leads to an inevitable deterioration of QoS performance compared with the target. In Figure 2, the actual QoS performance gap, extracted from simulations, is depicted for this particular use case with respect to the packet delay distribution. Note that for the simulation, we have assumed a Poisson traffic with an average rate of \( \lambda = 0.04 \) packets/TTI and the aforementioned distorted service pattern of \{18, 18, 24, 18, 18, 24, ...\} TTI due to the presence of a higher-priority class at every third granted RB. As for the analytical curve, we use Equations (1)–(3), with ideal values \( \lambda = 0.04, T_{g} = 18, \) and \( \rho = \lambda \cdot T_{g} \), as if no higher-priority class was present. Extensive results for such scenarios will be presented in Section 5.

In what follows, we will first show how we can predict this loss on the basis of an analytical model and then propose a slight modification to the original fixed grant algorithm to improve the lower-priority performance and better approximate the target.

### 4.1. An approximate analytical model for estimating the quality-of-service performance loss of low-priority classes

Because of possible conflicts between grants allocated to different classes, the \( M/D/1 \) model for the LP class is not valid anymore. The service model could now be described not by a unique service time \( T_{g} \), but by a discrete set of service times. Assuming \( N \) possible service times, \( T_{g}^{(i)} \), arbitrary service time, and \( p_{i} \) probability of occurrence of the particular service time, the LP problem may be now represented by an \( M/DN/1 \) (or \( M/iD/1 \) under a different notation) queueing model [27].

Although an exact expression for the steady-state waiting time distribution \( W_{q}(t) = 1 - \text{Prob}(W > t) \) along with a set of algorithms that produce it can be found in [27], an alternative, yet accurate, approach based on the Laplace–Stieltjes transform can be followed. Following [28] and given that the offered load is now expressed by \( \rho = \lambda \cdot \sum_{i=1}^{N} p_{i} T_{g}^{(i)} \), the Laplace–Stieltjes transform for the packet delay distribution is given by

\[
W_{q}^{*}(s) = \frac{1 - \lambda \cdot \sum_{i=1}^{N} p_{i} T_{g}^{(i)}}{s - \lambda \cdot \left(1 - \sum_{i=1}^{N} p_{i} e^{-s T_{g}^{(i)}}\right)}
\]

which leads to the desired result after an inversion is applied. However, because of the discontinuity on the service time cumulative distribution function, the inversion is not straightforward; a first-order numerical approximation method, which was developed in [28], can be also employed in the examined scenario and will be explained in detail later.

Let \( h \) be a small (time) step size (e.g., \( h = 0.0001 \)) and \( m = [m_{1}, m_{2}, \ldots, m_{1}, \ldots, m_{N}] \) be a vector for which its \( i \)th arbitrary element is an expression of the \( i \)th possible service time that is \( m_{i} = [T_{g}^{(i)}/h] \). Time is sampled as \( nh \), where \( n = 1, 2, \ldots \), and the corresponding approximating sampled distribution values \( W_{q}(nh) \) are simply computed by Equation (5). The \( F_{q} \) terms in Equation (5) are in turn given by the recursive expression in Equation (6). To sum up, for an arbitrary delay threshold \( t \), we can compute the probability of exceeding it, that is, \( \text{Prob}(W > t) \), by performing the recursive computations of Equations (5) and (6) for \( n = 1 \) up to \( n = \lceil t/h \rceil \).

\[
W_{q}(nh) = \frac{1}{\rho} (1 - F_{n}), n = 1, 2, \ldots
\]

\[
F_{n} = \frac{1}{1 - \lambda h} \left(F_{n-1} - \lambda h \sum_{i=1}^{N} p_{i} F_{n-m_{i}}\right), n \geq 1
\]

\[
F_{0} = \frac{1 - \rho}{1 - \lambda h}, F_{n} = 0 \text{ for } n < 0
\]

Note that for more than two classes, the extension is straightforward because lower-priority classes need to take into account the allocation patterns of all the higher-priority classes.

**Example.** In the case study shown in Figure 1, there are two possible service times, \( T_{g}^{(2)} = 18 \) TTIs and...
\[ T_s^2 = 24 \text{ TTIs}, \text{ occurring with probabilities } p_1 = 2/3 \text{ and } p_1 = 1/3, \text{ respectively. Thus, an } M/D_2/1 \text{ queueing model should be used for the analysis.} \]

### 4.2. A modified fixed grant algorithm for improving quality of service of low-priority classes

Because the original fixed grant-scheduling algorithm suffers from a gap between the target and actual QoS performance for LP M2M devices, we propose a slight modification that deals with this problem and an analytical model for obtaining a good approximation of the performance of the modified algorithm.

#### 4.2.1. Description of the proposed modification.

We assume that the original algorithm has determined the grants for the LP classes. Then, the scheduler monitors the running average of the grant period and, when it detects that is lower than the target one, allocates the next available (if any) RB to the M2M class. Such an action distorts the original pattern, but the extra allocated RBs improve the QoS performance. We emphasise that this modification is not the optimal one, but it is very simple and effective as it will be demonstrated through both simulation experiments and queueing analysis. If there is enough bandwidth available, the average service time approaches the target.

**Example.** Revisiting the case study depicted in Figure 1, we see that the \{18, 18, 24, \ldots\} service time pattern is initially formed by the original algorithm. The scheduler begins to monitor the experienced grant period and detects that after the end of the third round of grants (101st TTI), the average target period of 18 TTIs has been violated (has become (18 + 18 + 24)/3 = 20 TTIs). Thus, extra grants should be allocated to the LP devices, starting from the first available TTI, which is the 102nd in our case, taking care that the new grants do not conflict with higher-priority ones.

#### 4.2.2. Analysis.

The resulting modified service time pattern for the LP class will depend on the available M2M bandwidth, the volume of higher-priority M2M devices, and the QoS requirements for all classes; hence, it is difficult to model it with a specific distribution. In other words, a generalised distribution for the service time is assumed and according to the Kendall notation, the problem fits an \( M/G/1 \) queueing model. This is the most difficult model to analyse; thus, we resort to a two-moment approximation approach [24, Section 4.2], where only the knowledge on the first two moments of the service distribution is utilised, and no other information or exact cumulative distribution function is required. We hereafter provide the analytical expressions for approximating the packet delay distribution, underlining that these are accurate for high percentile values (>90%), which are realistic in practical QoS scenarios.

Let \( S \) denote the stochastic service time and \( E\{S\} \) and \( E\{S^2\} \) the first two moments. The squared coefficient of variation of the specific distribution is defined as \( c_s^2 = E\{S^2\}/E^2\{S\} \) and should take small values (typically lower than 2) for the approximation to be valid. The offered load may be expressed, through the first moment of \( S \), as \( \rho = \lambda \cdot E\{S\} \). The core idea of the two-moment approach is that the percentile of the packet delay distribution is approximated by a weighted sum of the corresponding percentiles of two well-studied service models, namely an exponential one \( (M/M/1) \) and a deterministic one \( (M/D/1) \) with the same first moment.

Mathematically speaking, the \( p \)-th (close to unity) percentile \( \xi(p) \) may be expressed as follows [24, Equation (4.42)]:

\[
\xi(p) = (1 - c_s^2) \cdot \xi_{M/M/1}(p) + c_s^2 \cdot \xi_{M/D/1}(p) \quad (7)
\]

Regarding the \( M/M/1 \) percentile, it is easily calculated by the following:

\[
\xi_{M/M/1}(p) = E\{S\} \cdot \frac{1}{1 - \rho} \cdot \log \left( \frac{1}{1 - p} \right) \quad (8)
\]

On the other hand, the \( M/D/1 \) percentile is more complex to obtain. In particular, it is given by

\[
\xi_{M/D/1}(p) = \frac{1}{\delta_{M/D/1}} \cdot \log \left( \frac{YM/D/1}{1 - p} \right) \quad (9)
\]

where

\[
YM/D/1 = \frac{1 - \rho}{\delta_{M/D/1} \cdot E\{S\} - (1 - \rho)} \quad (10)
\]

and \( \delta_{M/D/1} \) is the unique positive solution of the following non-linear equation:

\[
e^{\delta_{M/D/1} \cdot E\{S\}} - \delta_{M/D/1} / \lambda - 1 = 0 \quad (11)
\]

In summary, for a known input traffic load expressed by \( \lambda \) or \( \rho \), a service distribution \( S \) with the first two moments known, and an arbitrary delay threshold \( t \), we have to perform the following steps to calculate the probability of exceeding this threshold: (i) solve Equation (11) and replace it into Equation (10); (ii) obtain the percentiles of Equations (8) and (9) for all the \( p \)-values of interest and replace them into Equation (7); (iii) associate the required delay threshold \( t \) with a percentile value \( p \) such that \( \xi(p) \rightarrow t \); and finally, (iv) obtain the required probability as \( 1 - p \).

### 5. RESULTS AND DISCUSSION

Following the analysis of the previous sections, regarding both the single-class and multiclass scenarios, we will now...
Table I. Simulation parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value/comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>General parameters</td>
<td></td>
</tr>
<tr>
<td>Overall LTE + M2M bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>M2M bandwidth</td>
<td>2 MHz</td>
</tr>
<tr>
<td>LTE frame duration</td>
<td>1 ms</td>
</tr>
<tr>
<td>Single-class scenarios</td>
<td></td>
</tr>
<tr>
<td>Number of M2M devices</td>
<td>100</td>
</tr>
<tr>
<td>Average packet arrival rate</td>
<td>0.02/0.04 packets/TTI</td>
</tr>
<tr>
<td>Scheduling period</td>
<td>25–45/13–22 TTIs</td>
</tr>
<tr>
<td>Network realisations</td>
<td>100</td>
</tr>
<tr>
<td>Realisation length</td>
<td>1000 s</td>
</tr>
<tr>
<td>Multiclass scenarios</td>
<td></td>
</tr>
<tr>
<td>Number of M2M classes</td>
<td>2 (HP/LP)</td>
</tr>
<tr>
<td>Number of M2M devices</td>
<td>100 (HP), 20 (LP)</td>
</tr>
<tr>
<td>Average HP-class packet arrival rate</td>
<td>0.03/0.01 packets/TTI</td>
</tr>
<tr>
<td>HP-class scheduling period</td>
<td>30/50 TTIs</td>
</tr>
<tr>
<td>Average LP-class packet arrival rate</td>
<td>0.03–0.045/0.025–0.035</td>
</tr>
<tr>
<td>LP-class scheduling period</td>
<td>18/20 TTIs</td>
</tr>
<tr>
<td>Network realisations</td>
<td>10</td>
</tr>
<tr>
<td>Realisation length</td>
<td>500 s</td>
</tr>
</tbody>
</table>

HP, high priority; LP, low priority; LTE, Long-term Evolution; M2M, machine to machine; TTI, transmission time interval.

demonstrate through system-level simulations the following: (i) the validity of the analytical models and (ii) the improvement on the QoS performance when employing the modified fixed grant periodic-like scheduling scheme. For the former, we will compare the QoS performance results extracted from extensive simulation experiments with that produced by the analytical models, whereas for the latter, we will compare the modified algorithm performance results (analytical and simulation based) with the QoS target levels. The key simulation parameters are summarised in Table I, whereas they are elaborated in what follows.

5.1. Single-class scenarios

Two experimental set-ups are considered to demonstrate the validity of the analytical models with respect to the packet delay distribution assuming infinite delay threshold [Equations (1)–(3)] and the dropped packet rate when a finite delay threshold is defined [Equation (4)]:

- Set-up 1: The average packet arrival period per device is 50 TTIs, that is, \( \lambda = 0.02 \) packets/TTIs, whereas the service rate (or the grant period) varies from \( T_g = 25 \) to 45 TTIs, covering a wide offered load area extending from \( \rho = 0.5 \) to 0.9.
- Set-up 2: A lower average packet arrival period is considered (25 TTIs or \( \lambda = 0.04 \) packets/TTIs), which is more demanding in terms of the QoS performance levels, whereas the grant period varies from 13 to 22 TTIs.

For both set-ups, 100 independent system instances were generated, each one simulating the network performance for 1000 s (1 000 000 LTE TTIs). At each system instance, a single M2M class is considered, comprising 100 M2M devices. Enough M2M bandwidth is considered; thus, the target and the actual QoS performance are identical. Figure 3(a, b) illustrates the performance results for the first set-up, whereas Figure 4(a, b) those for the second. In both set-up results, we observe the following: (i) the analytical modelling for the delay threshold violation probability is very accurate for various sets of operational parameters, that is, the offered traffic intensity, the scheduling rate capabilities, and the QoS requirements; (ii) the dropped packet rate analytical modelling through Equation (4) is also accurate; thus, both elastic and inelastic traffic scenarios may be handled by our common framework; and (iii) the QoS performance depends not only on the traffic utilisation \( \rho \) but also on the absolute traffic intensity and scheduling rate values.
In case of the multiclass scenario, we simulate two indicative set-ups:

- **Set-up 1:** We assume two classes of users (as in the case study of Figure 1), an HP one comprising 100 M2M devices and an LP one comprising 20 M2M devices. HP devices generate Poisson traffic with \( \lambda_{HP} = 0.03 \), whereas we vary the generated traffic of the LP class (from 0.03 to 0.045 packets/TTI) to control the overall offered load. The HP class is scheduled every 30 TTIs, whereas the LP every 18 TTIs.

- **Set-up 2:** The same general set-up is employed, but with \( \lambda_{HP} = 0.01 \), \( T_{LP}^{HP} = 50 \) TTIs, and \( T_{LP}^{LP} = 20 \) TTIs (leading to an LP service pattern of \( \{20, 30, 20, 30, \ldots \} \) TTIs), and \( \lambda_{LP} \) is tuned so that the LP load varies from 0.6 to 0.9.

For both set-ups, we simulate both the original fixed grant periodic scheduling algorithm (‘AGTI’) and the modified algorithm presented in Section 4.2 (‘AGTImod’). We generated 10 independent system realisations, each one capturing the network performance for 500 s (500,000 LTE TTIs). We also assume that 10% of the typical 20-MHz LTE bandwidth is reserved for M2M traffic, that is, 10 RBs per TTI. The scope of the particular experiments is threefold as reported in the following.

First, we demonstrate that regarding the original AGTI algorithm, our proposed \( M/D_N/1 \) modelling of the LP-class QoS loss, compared with the target requirement (Section 4.2.2), is rather accurate. The target performance is the QoS behaviour if no higher-priority class was present in the system. Thus, it can be easily predicted from the single-class analytical model if we set the grant period equal to \( T_{LP}^{LP} \) in Equations (1)–(3). We also note here that we provide results in terms of the packet delay distribution, whereas dropped packet rate results can be obtained through Equation (4). Figure 5(a, b) illustrates the results for both set-ups, including the actual (simulated), predicted (analytical) and target performance of the LP class. One may clearly observe that for medium to moderate offered loads (\( \rho < 0.85 \)), the adopted \( M/D_N/1 \) model is a very accurate approximation of the QoS performance deterioration, which occurs because of the first-class prioritisation and the inevitable violation of the LP constant service rate scheduling. Thus, the model could serve as a valuable prediction tool for the scheduler entity when multiple conflicting M2M classes are present in the system.

Furthermore, we examine the performance of the modified AGTI algorithm we first proposed in Section 4.2.1. In Figure 6(a, b), we present the actual, predicted and target performance of AGTImod. Comparing these simulation results with the AGTI results in Figure 6(a, b), we deduce that the slight modification (blue curves) we applied to the original algorithm has led to a significant performance improvement. The modified algorithm achieves a delay performance very close to the required one set by the QoS targets (pink curves). The latter curves correspond to the optimal delay performance if a periodic grant scheme is applied. We point out that if a fully dynamic and queue-aware scheduling is applied, then the performance would be significantly enhanced, but at the expense of dramatically increased signalling overhead.

Finally, as far as the two-moment approximation of the \( M/G/1 \) model proposed in Section 4.2.2 is concerned, two are the main observations (if we compare the blue solid curves with the square marker curves in Figure 6(a, b)). First, for high but realistic percentiles (equivalently low probabilities of exceeding a target threshold), the particular model is a good approximation of the actual performance. The model fails at lower percentiles, but this is expected as stated in [24], because it is based on an exponential decay approximation that is valid for high delays. Empirically, we have observed that the approximation is accurate for percentile values \( 1 - p < 0.2 \cdot p \). As a second remark, the approximation is shown to be not so accurate in the second set-up;
6. CONCLUSION AND FUTURE WORK

The timely problem of efficiently scheduling M2M traffic over existing LTE/LTE Advanced frames was studied in this work. Towards achieving simple but efficient, time-controlled, and low-overhead M2M communication over LTE, we first proposed analytical models, which can be utilised either for predicting the QoS experience levels regarding realistic coexisting multiple-M2M-class scenarios or for tuning the scheduling algorithms. We also proposed a simple modification to an existing scheduling algorithm to improve lower-priority M2M device performance. The provided numerical results demonstrated, on the one hand, good accuracy of our queueing-based analytical models and, on the other hand, the efficiency of the modified scheduling algorithm we developed.

Several research problems remain open and are good candidates for future work. First, it would be interesting to incorporate to our framework packet drops due to link conditions (the last LTE-QCI characteristic), because we have considered so far drops only due to congestion. It seems reasonable that the cross-layer effective capacity model (for example, [29, 30]) would be a viable solution to analytically deal with this problem. Moreover, considering, from both simulation and analytical perspectives, other M2M traffic models, such as uniform or bursty, would allow optimisation of the proposed scheduling algorithms for different M2M applications. In addition, by incorporating simple 1-bit queue awareness to the original periodic algorithm, we could boost delay performance at the expense of overhead signalling; an analytical study of this trade-off would be important as it would reveal new possibilities to the scheduling designer. Finally, the dynamic split of LTE resources among legacy LTE-UEs and M2M devices would boost overall cellular performance and optimise system resource usage.
ACKNOWLEDGEMENTS

This work has been performed in the framework of the ICT project ICT-5-258512 EXALTED, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues, although the views expressed are those of the authors and do not necessarily represent the project. This research has been also cofinanced by the European Union (European Social Fund, ESF) and Greek national funds through the Operational Programme ‘Education and Lifelong Learning’ of the National Strategic Reference Framework (NSRF), Research Funding Programme THALIS-INTENTION.

REFERENCES

3. 3GPP TS 22.368. Service requirements for machine-type communications (MTC), September 2011. v.11.3.0 (release 11).
4. 3GPP TR 23.888. System improvements for machine-type communications, November 2011. v.1.6.0 (release 11).
5. 3GPP TR 37.868. Study on RAN improvements for machine-type communications, September 2011. v.11.0.0 (release 11).
9. 3GPP TS 23.203. Policy and charging control architecture, March 2012. v.10.6.0 (release 10).
23. Iversen VB. Teletraffic engineering and network planning, Technical University of Denmark, 2011.


