TRANSACTIONS ON EMERGING TELECOMMUNICATIONS TECHNOLOGIES
Trans. Emerging Tel. Tech. 0000; **00**:1–12
DOI: 10.1002/ett

Loss and Delay Analysis of Non-Poisson M2M Traffic over LTE Networks

Morteza Zarrini and Abdorasoul Ghasemi
Faculty of Computer Engineering, K.N. Toosi University of Technology, Tehran, Iran
Email: mzarrini@ee.kntu.ac.ir, arghasemi@kntu.ac.ir

ABSTRACT

In this work, loss rate and delay of non-Poisson machine-to-machine (M2M) traffic in LTE networks have been studied. In contrast to prior works which used Poisson process model for M2M traffic in their analysis, more realistic traffic models have been proposed which follow the traffic patterns of M2M communications reported lately by 3GPP. Markov Modulated Poisson Process (MMPP) and an approximated Coupled MMPP (CMMPP) are adopted for modeling the uncoordinated and coordinated M2M network traffic, respectively. Also, Fixed-Access Grant Time Interval (AGTI) algorithm is used as a low-overhead cluster based scheduling algorithm to serve packets. Using some approximations, we first derive analytical results for the delay violation and packet loss probabilities of M2M traffic in LTE networks in different scenarios. We then investigate the effect of buffer size on the quality of service of M2M user equipments (UEs). Specifically, in contrast to uncoordinated traffic model, it is shown that for the coordinated traffic model increasing the buffer size above a threshold is not effective in decreasing the loss probability of UEs. Simulation results are provided to justify the analysis and the effect of buffer size on delay and loss probabilities.

1. INTRODUCTION

Machine-to-Machine (M2M) communications is an inseparable part of the modern communication networks to offer services such as Internet of things (IoT), E-Health, fleet management, and monitoring systems. The number of M2M devices connected to cellular networks worldwide is estimated to be more than hundreds of millions and in comparison to traditional communication devices like smartphones it is growing at a rapid pace [1, 2]. On the other hand, Long Term Evolution (LTE) networks provide an appropriate infrastructure for M2M communications due to their advantages such as ubiquitous coverage, low latency, high capacity, and all-IP technology [3]. Therefore, it is expected that M2M traffic comprises a considerable volume of the LTE networks traffic in the near future.

In order to accommodate M2M communications in LTE networks, it is important to consider the peculiarities of this type of communications as LTE networks are inherently developed for Human-to-Human (H2H) communications. Some of these peculiarities include high uplink to downlink traffic ratio, small packet size, less mobility, diverse quality of service (QoS) requirements for different M2M applications, the huge number of devices trying to access the network, and possible synchronization of UEs' activation which intensifies the burstiness of the aggregated traffic [1]. Dealing with synchronized traffic is one of the important issues in the analysis of M2M communications which is addressed in this paper. Due to these peculiarities, efficient scheduling algorithms in Radio Access Network (RAN) are required to support M2M user equipment (UE) and avoid QoS degradation of H2H communications.

The scheduling algorithm is a key component of efficient resource management in the RAN of LTE networks. Resources can be scheduled for each UE individually or for each cluster of UEs which have the same QoS requirements. The former performs better in terms of efficient utilization of RAN resources, however, suffers from the signaling overhead and high complexity since it needs the channel condition and delay requirements of each UE. The latter scheme, which is also adopted in this paper, does not need to exchange control data with each UE individually and allocates resources to each cluster of M2M UEs. In this work, Fixed-Access Grant Time Interval (AGTI) is used as a

Prepared using ettauth.cls [Version: 2012/06/19 v2.10]

promising cluster based scheduling scheme. AGTI requires very low signaling overhead and supports the required QoS of the clusters provided that UEs have been clustered appropriately [4–7].

On the other hand, the performance analysis of scheduling algorithms depends on the assumed traffic model of UEs. The traditional Poisson process traffic modeling which is typically assumed in the designing of the scheduling algorithms, is not accurate for modeling traffic sources that show spatial and temporal correlation as in the case of M2M UEs [8]. Since each UE may operate in regular and alarm modes, Markov Modulated Poisson Process (MMPP) is a good suggestion for modeling this traffic behavior [8–10]. Specifically, for the uncoordinated traffic in which the UEs of each cluster operate in an asynchronous manner, the generated traffic by each cluster can be modeled by a number of separate MMPPs. In contrast, for the coordinated traffic the UEs may operate synchronously when they respond to events in a coordinated manner. In this case, the behavior of the cluster's traffic can be characterized by Coupled Markov Modulated Poisson Process (CMMPP) [8, 10].

In this paper, assuming fixed AGTI scheduling algorithm, the QoS of UEs in terms of delay violation and packet loss probabilities for uncoordinated and coordinated M2M traffics are analyzed. The MMPP and an approximated CMMPP are considered to model the traffic of each cluster in the uncoordinated and coordinated scenarios respectively. Finally, the MMPP/D/1/K queueing model is used to evaluate delay and packet loss for M2M UEs for different buffer sizes.

The rest of this paper is organized as follows. After reviewing some related works in section 2, the system and traffic models are presented in section 3. In section 4, the performance evaluation of the system for uncoordinated traffic is discussed. In section 5, by adopting a simplified traffic model the system performance for coordinated traffic is analyzed. Numerical studies to justify the analysis and discussion are provided in section 6 before concluding in section 7.

2. RELATED WORK

Delay-aware and channel-aware M2M scheduling schemes over LTE networks are discussed in [3, 11] where packets are prioritized according to the remaining tolerable delay and reported channel quality by each UE. Due to small size packets in M2M communications, these schemes incur a high signaling overhead especially for a scenario with a large number of UEs. Considering this issue, some works propose the cluster based scheduling in which UEs are clustered according to their required QoS. In the cluster-based approach, exchanging of the control information for each UE is not required and packets could be served using a deterministic scheduling scheme [2, 4–7].

In [4, 5] assuming a constant time interval for packet arrival, UEs are clustered according to their required QoS where a priority and a fixed AGTI are assigned to each cluster. This simplifying assumption is in contrast to the random nature of M2M traffic [2, 6, 7]. That is, the performance of scheduling schemes strongly depends on the offered traffic behavior of UEs and hence more attention is required in M2M traffic modeling.

In [7], M2M traffic is modeled by a Poisson process and the probability of the delay violation for a given threshold in single-class and multi-class scenarios are analyzed and evaluated by simulation. Although, Poisson process is a typical model for the traffic modeling of H2H communications, recent studies [8–10] and 3GPP reports [12] show that individual M2M UEs and their aggregated traffic do not follow Poisson process model. Specifically, UEs may operate in the regular and alarm modes where the distribution of packets' inter-arrival times and their sizes are different in each mode [8–10]. Regarding a single UE traffic, at least two Poisson processes with different rates are required to efficiently model the generated traffic in the uncoordinated model. However, in the coordinated model, M2M UEs exhibit temporal and spatial synchronism [8] and more complicated models should be used.

Considering these facts, the 3GPP proposes two traffic models for the aggregated traffic of M2M UEs named as the model I and model II. The arrival times of packets in these models follow uniform(0,1) and Beta(3,4) distributions, respectively [12]. Uniform distribution is used to model the traffic in the uncoordinated model and beta distribution is used for the coordinated model in which UEs change their traffic mode synchronously, i.e., UEs go to the alarm mode synchronously. M2M traffic modeling is also discussed in [8, 10], where traffic is modeled by the MMPP and the CMMPP for uncoordinated and coordinated models, respectively. Also, it is shown that these models converge to the corresponding model I and model II of the 3GPP.

Most of the existing works on scheduling M2M traffic over LTE networks do not consider these peculiarities of M2M traffic. In this paper, more realistic traffic models are used in analyzing the AGTI scheduling scheme.

3. SYSTEM MODEL

3.1. System model and problem statement

A single cell of an LTE network is considered in which M2M UEs directly connect to the eNB of the cell. The eNB schedules UEs in a centralized manner. The bandwidth of the cell is divided into some sub-channels where m sub-channels are dedicated to M2M communications. Each LTE frame consists of the corresponding spectrum resources of these m sub-channels for the duration of 10 ms. Each frame, in turn, is divided into 10 sub-frames called as the transmission time interval (TTI) with the duration of $1ms$. Hence the smallest allocable resource

Trans. Emerging Tel. Tech. 0000; **00**:1–12 © 0000 John Wiley & Sons, Ltd.
DOI: 10.1002/ett
Prepared using ettauth.cls

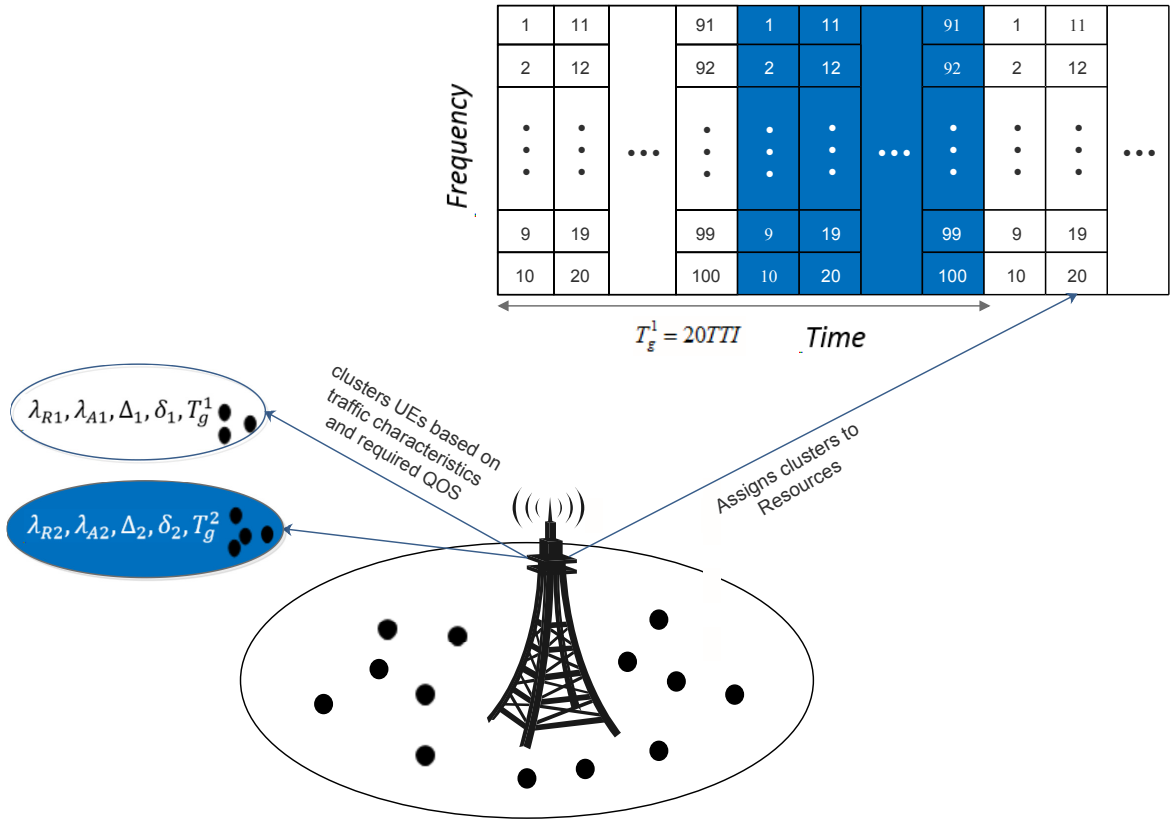


Figure 1. System Model: eNB assigns the registered UEs to different clusters according to their traffic characteristics and QoS requirements. Resources are assigned to UEs according to the assigned service time interval to the corresponding cluster.

element to each UE in a service time consists of one sub-channel in the frequency domain and one TTI in the time domain.

The scheduler assigns the registered UEs in appropriate clusters according to their traffic characteristics and QoS requirements. It is assumed that UEs are clustered into two different groups as it is shown in Fig. 1. Also, an appropriate service time interval, T_g, is assigned to each cluster by the eNB according to its requested QoS, e.g., $T_g^1 = 20$ means that UEs of cluster 1 are served every 20 TTI. The required QoS for UEs in cluster i is specified by the maximum tolerable delay for data transmission and is denoted by Δ_i. That is:

$$\Pr(Delay > \Delta_i) < \delta_i, \qquad i = 1, 2 \tag{1}$$

where δ_i is a service level parameter.

It is assumed that each UE operates in regular and alarm modes. When no event is detected in the vicinity of a UE, the UE operates in the regular mode and its generated traffic is a low rate periodic traffic with long inter-arrival time between two consecutive requests. Upon detecting an event, the UE goes to the alarm mode and requests are generated in a more intensive way. Hence, the traffic of each UE cannot be modeled by a single Poisson process as discussed in [8–10]. Furthermore, the traffic of each UE has spatial and temporal correlations with the nearby UEs.

The objective is to find out the effects of non-Poisson traffic model on the delay violation and packet loss probabilities in a specific service time duration. For this purpose, proper modeling of the M2M traffic and considering the possible temporal and spatial correlation between the UEs are the key factors which are discussed in the next subsection.

3.2. M2M Traffic Modeling

The traffic of each UE is characterized by the arrival process of its requests. Let λ_{R_i} and λ_{A_i} denote, respectively, the arrival rates of UEs' requests in regular and alarm modes for cluster i. Each UE switches between these modes and it is assumed that the interval times between the mode switching follows an exponentially distributed random variable with parameter λ_s.

The transition matrix of mode switching for UE_i is denoted by P_i. To consider the spatial correlation of UE_i's traffic, parameter $\alpha_i, 0 \leq \alpha_i \leq 1$ is dedicated to this UE according to its distance to the source of events. This parameter would be greater for UEs which are located closer to the source of events. On the other hand, the temporal correlation in traffic generation of UEs is

modeled by a piecewise constant function $\theta[t]$, i.e., its value is constant in each LTE frame. In other words, $\theta[t]$ shows the degree of temporal correlation of UEs' traffic at time instance t.

Summing up, the synchronicity parameter of UE_i with other UEs at time instance t, $\theta_i[t]$, is given by $\theta_i[t] = \alpha_i\theta[t]$. Having the synchronicity parameter of UE_i, the transition probability matrix of this device at time t is given by (2).

$$P_i[t] = \theta_i[t]P_C + (1 - \theta_i[t])P_U \tag{2}$$

where P_C and P_U are given by:

$$P_C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, P_U = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}. \tag{3}$$

In (3), P_C and P_U are the transition probability matrices for fully coordinated and fully uncoordinated UEs, respectively. The elements of row i and column j of these matrices show the transition probability from state i to state j where regular and alarm modes of UEs are denoted by indices 1 and 2, respectively. That is, $P_{C_{12}}$ shows the transition probability from regular to alarm mode. A fully coordinated UE oscillates between the alarm and regular modes at each transition point and an uncoordinated UE stays in the regular mode permanently.

4. UNCOORDINATED TRAFFIC MODEL WITH INFINITE BUFFER

In the uncoordinated model, the traffic of each UE is generated independently from other UEs, so it follows a constant traffic behavior over the time. In this case, the aggregated traffic of UEs follows Model I [12] of the 3GPP which corresponds to $Uniform(0, 1)$ probability density function (pdf) as discussed in [10]. Hence, the temporal correlation, θ has the uniform pdf in the duration that each UE is active, i.e., $\theta[t] = 1$ for all t. Also, the transition probability matrix of each UE is time independent and hence the traffic generation can be modeled by a two-state MMPP as shown in Fig. 2. In Fig. 2 R and A show regular and alarm modes, respectively. Notice that in the regular and alarm states the traffic of each UE is generated according to Poisson processes with parameter λ_R and λ_A, respectively.

On the other hand, the requests of each UE are served according to the fixed-AGTI scheduling scheme which allocates one RB to each UE in constant time intervals. That is, the service rate for each UE is constant. Therefore, MMPP(2)/D/1 queueing model can be used for system performance analysis in this scenario.

A simple and efficient approximation for MMPP(2)/D/1 queuing model for the Continuous Time Markov Chain (CTMC) is proposed in [13]. The transition rate matrix for the corresponding CTMC model of our Discrete Time

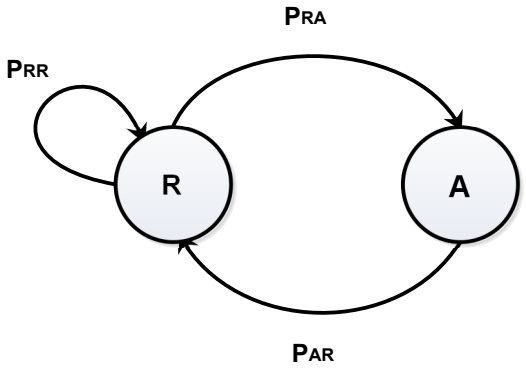


Figure 2. Traffic generation of each UE in the uncoordinated scenario is modeled by a two-state MMPP.

Markov Chain (DTMC) model is given by [14]:

$$R = \lambda_c(P - I), \tag{4}$$

where R, I, and λ_c are, respectively, the infinitesimal generator matrix which shows the transition rate between states, the unity matrix, and the rate of changes in the system. The system is observed at each service time and the transition between states happens at switching time, hence we have:

$$\lambda_c = \frac{T_g}{T_s}, \tag{5}$$

where T_g and T_s are, respectively, service (grant) time and the mean switching time, i.e., 1 LTE frame.

Let $W(x)$ denote the Cumulative Distribution Function (CDF) of each request waiting time in the MMPP(2)/D/1 queuing model. Then, the complement CDF, $V(x) = \Pr(W > x) = 1 - W(x)$ is approximated by:

$$V_{appr}(x) = a_1 e^{b_1 x} + a_2 e^{b_2 x}, \tag{6}$$

where x is the waiting time in terms of the number of service times that each request waits in the queue. The Laplace transform of $V(x)$ in (6) is given by:

$$\hat{V}_{appr}(s) = \frac{a_1}{s - b_1} + \frac{a_2}{s - b_2}, \tag{7}$$

where $a_1 > 0$, $a_2 > 0$, and $b_2 < b_1 < 0$ and b_1 and b_2 are asymptotic decay rates parameters and a_1 and a_2 are asymptotic constants. In fact, b_1 and b_2 are the first and second negative poles of $\hat{V}(s)$ and are computed using (8).

$$det[sI + R - \Lambda + \Lambda e^{-s}] = 0, \tag{8}$$

where matrix R is the infinitesimal generator matrix as before and Λ is the arrival rate matrix and is defined by:

$$\Lambda = \begin{bmatrix} \lambda_R & 0 \\ 0 & \lambda_A \end{bmatrix}. \tag{9}$$

Therefore, the waiting time for each request in the uncoordinated model can be found when buffer size is infinite.

Trans. Emerging Tel. Tech. 0000; **00**:1–12 © 0000 John Wiley & Sons, Ltd.
DOI: 10.1002/ett
Prepared using ettauth.cls

5. SYSTEM PERFORMANCE EVALUATION FOR COORDINATED TRAFFIC MODEL

In the coordinated model, UEs mutually affect each other's traffic behavior. So, each UE affects the traffic pattern of adjacent UEs and its traffic is affected by them. Coupled Markov Modulated Poisson Process (CMMPP) is proposed to model these bidirectional connections in each cluster. However, as the number of UEs in the cluster is increased, the number of interconnections among UEs increases rapidly and subsequently the complexity of CMMPP grows excessively.

In order to find out a more scalable and tractable traffic model for analysis, the effect of these interactions in the developed approximated MMPP traffic model of each UE is considered. Specifically, according to the Model II of 3GPP, the aggregated traffic of the coordinated UEs follows $Beta(3,4)$ probability distribution function (pdf) [12]. Furthermore, in [8,10] it is shown that the aggregated traffic of UEs which individually follows $Beta(3,4)$ pdf, is the same as the aggregated traffic in the coordinated model. Hence, to take into account this behavior in each UE traffic model, an approximation of the $Beta(3,4)$ pdf is considered in the θ function which reflects the temporal correlation of UEs' traffic. This temporal correlation directly affects synchronicity parameter which in turn constructs the transition probability matrix and subsequently affects the UEs' traffic. Therefore, the traffic of each UE in the coordinated scenario could be modeled by an MMPP where its transition matrix is varying over the time. That is, we approximate the initial CMMPP by an approximated MMPP model in which each UE is independently modeled by an MMPP. Accordingly, the MMPP/D/1/K queueing model can be used to investigate the QoS of each UE in subsequent analysis.

In the following subsections, a numerical approximation of MMPP/D/1/K queueing model is applied to calculate waiting time and packet loss probabilities. Then an illustrative example is provided in order to exemplify the process of traffic modeling and performance evaluation.

5.1. Approximation of Coordinated Traffic Model

In the coordinated model, the transition probability matrix of each UE follows Beta(3,4) pdf in time. This pdf can be approximated by piecewise constant function in consecutive short time intervals or slots. The transition probability matrix of the corresponding MMPP in each slot then would be constant.

Let the total activation interval of coordinated UEs, $[0,T]$, be divided into $z=\frac{T}{\Delta t}$ time slots of duration Δt, and assume that the transition matrix of UE_i during the j^{th} slot, $[(j-1)\Delta t, j\Delta t]$, is constant. Therefore, the traffic of each UE can be modeled by z two-state MMPPs in consecutive time slots as it is shown in Fig. 3. To ensure that this model follows Beta(3,4) distribution over the time, the sojourn time in each MMPP and then transition to the next MMPP should be adjusted properly. In the following the corresponding MMPP of time slot j is denoted by $MMPP_j$ and the sojourn time of the system in this MMPP is denoted by Δt.

Notice that the system would be at $MMPP_j$ for $(\frac{\Delta_t}{T_s}-1)$ LTE frames and then moves to the next MMPP at $\frac{\Delta_t}{T_s}^{th}$ transition time. Hence, the probabilities of staying in current MMPP and moving to the next MMPP are given by $1-\frac{T_s}{\Delta_t}$ and $\frac{T_s}{\Delta_t}$, respectively. For example, if $T_s=1$ LTE frame, $\Delta_t=50$ LTE frames, and the pdf of $Beta(3,4)$ is approximated by a piecewise constant function with $z=20$, then the probabilities of staying in current MMPP and moving to the next MMPP would be $49/50$ and $1/50$, respectively.

As an illustrative example consider a simple scenario in which two UEs are served in a cluster in an LTE cell and the input traffic lasts for 100 LTE frames or two slots. Also, assume that the temporal correlation is approximated by 0.6 and 0.2 in slot 1 and 2, respectively. The spatial correlations for UE_1 and UE_2 are 0.5 and 0.3. Hence, the transition matrices for UE_1 and UE_2 in slots 1 and 2 are given by:

$$P_1[1]=\begin{bmatrix}0.7 & 0.3\\ 1 & 0\end{bmatrix} \quad P_1[2]=\begin{bmatrix}0.9 & 0.1\\ 1 & 0\end{bmatrix}$$
$$P_2[1]=\begin{bmatrix}0.82 & 0.18\\ 1 & 0\end{bmatrix} \quad P_2[2]=\begin{bmatrix}0.94 & 0.06\\ 1 & 0\end{bmatrix}.$$

The corresponding approximated traffic model for UE_1 in the activation period is depicted in Fig. 4 as a 4-state MMPP. In this figure, R_i and A_i denote the regular and alarm states in the ith slot respectively. A similar traffic model can be approximated for UE_2.

Now, assume that the packets' arrival rates in regular and alarm modes are $\lambda_R=0.0125$ pkt/ms and $\lambda_A=0.125$ pkt/ms and the system is able to serve one packet every 20 ms. Also, we have $\lambda_s=0.1$ switch/ms since the switching among states occurs at the end of each LTE frame. Adopting indices 1 to 4 for states R_1, A_1, R_2, and A_2, the corresponding system transition and rate matrices for UE_1 are given by:

$$P_1=\begin{bmatrix}0.686 & 0.294 & 0.02 & 0\\ 0.98 & 0 & 0.02 & 0\\ 0.02 & 0 & 0.882 & 0.098\\ 0.02 & 0 & 0.98 & 0\end{bmatrix},$$
$$\Lambda_1=\begin{bmatrix}0.25 & 0 & 0 & 0\\ 0 & 2.5 & 0 & 0\\ 0 & 0 & 0.25 & 0\\ 0 & 0 & 0 & 2.5\end{bmatrix}, \tag{10}$$

where Λ equates to Λ/μ since service time is assumed unity.

5.2. Deploying MMPP/D/1/K for Analysis

MMPP/D/1/K queueing model can be deployed to evaluate the probabilities of delay violation and packet loss for each

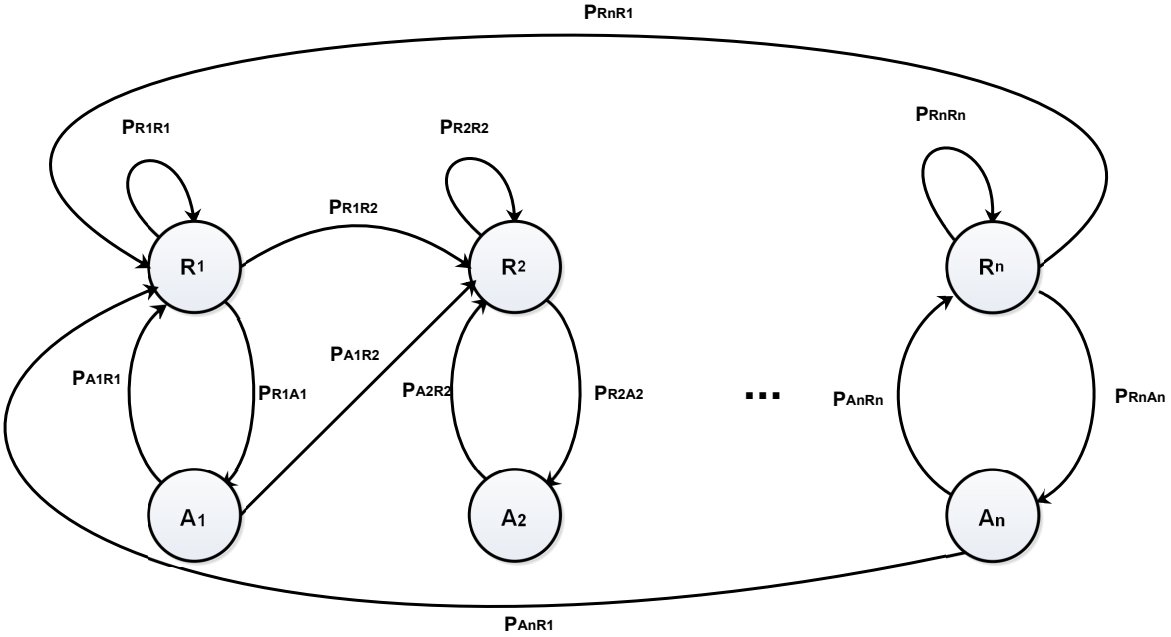


Figure 3. Traffic modeling for each UE in coordinated model using consecutive two-state MMPPs

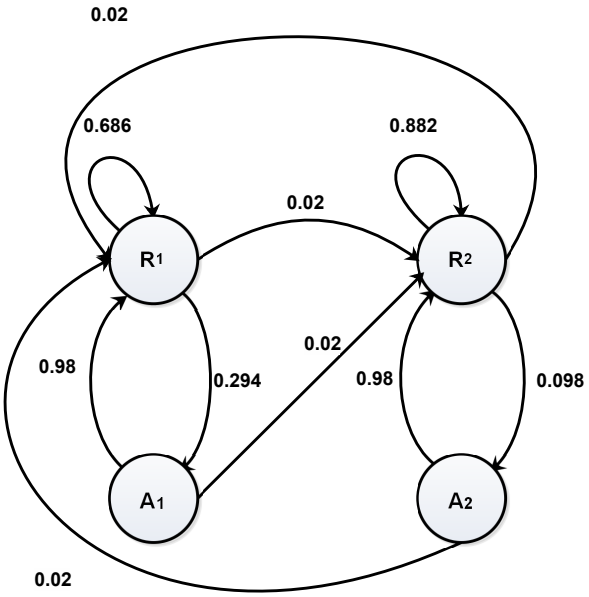


Figure 4. Approximate traffic modeling of UE_1 in the illustrative example

UE using the approximated traffic model in the previous section where K is the finite buffer size of each UE. Also, notice that the results of this section can be used for performance evaluation in the uncoordinated model with finite buffer size as well.

In [15] a numerically efficient method for MMPP/D/1/K queue via Padé approximation is proposed. A brief overview of this method is provided in Appendix A.

Let the waiting time vector be defined as:

$$\mathbf{w}(x) = [w_1(x), w_2(x), \ldots, w_n(x)], \qquad (11)$$

where $w_j(x)$ is the stationary probability that at an arbitrary time the arrival process be in state j and the waiting time of an unfinished work be at most x, i.e., the waiting time be less than or equal to x service times. The Laplace transform of $\mathbf{w}(x)$ is given by:

$$\hat{\mathbf{w}}(s) = \mathbf{y}_0[sI + R - \Lambda + \Lambda e^{-s}]^{-1}, \qquad (12)$$

Using the Padé approximation the irrational e^{-s} term in (12) can be approximated by $\frac{\hat{R}_a(s)}{\hat{Q}_a(s)}$.

Now, the waiting time for MMPP/D/1/K is given by (13):

$$\mathbf{w}(x) = \begin{cases} \mathbf{y}_{0,K} B_1 e^{A_1 x} C_1, & 0 \le x \le K, \\ \mathbf{y}_{0,K} B_1 e^{A_1 K} e^{A_2 (x-K)} C_1, & K \le x \le K+1, \end{cases} \qquad (13)$$

where $\mathbf{y}_{0,K}$ represents the stationary probability that with finite buffer of size K, at an arbitrary time the arrival process be in state j and the number of packets in queue be zero. Let $\boldsymbol{\pi}$ denote the stationary probability vector of being in each state. we have:

$$\mathbf{y}_{0,K} B_1 e^{A_1 K} e^{A_2} C_1 = \boldsymbol{\pi}. \qquad (14)$$

In (14) A_1, B_1, A_2, B_2 are square matrices computed according to the polynomials $\hat{R}_a(s)$ and $\hat{Q}_a(s)$, and C_1 is a constant column matrix.

Trans. Emerging Tel. Tech. 0000; **00**:1–12 © 0000 John Wiley & Sons, Ltd.
DOI: 10.1002/ett
Prepared using ettauth.cls

Finally, the waiting time violation, $V(x)$, is given by:

$$V(x) = \Pr(W > x) = (1/\rho)(1 - W(x)), \qquad (15)$$

where $W(x) = \mathbf{w}(x)\mathbf{e}$, $1/\rho$ is the fraction of times that the server is busy, and $\mathbf{e}$ is the unitary column vector.

Also, the probability of loss, P_{loss}, is given by:

$$P_{loss} = \frac{(\boldsymbol{\pi} - \mathbf{w}(K))\Lambda\mathbf{e}}{\bar{\lambda}}, \qquad (16)$$

where $\bar{\lambda} = \boldsymbol{\pi}\Lambda$ is the mean arrival rate and $\mathbf{e}$ is the unitary column vector.

5.3. Illustrative Example

To follow how the MMPP/D/1/K analysis can be used for computing the waiting time and packet loss probability, in this subsection these performance metrics are computed for UE_1 and the corresponding cluster of the explained example in section 5.1.

Using (10) and (4), the infinitesimal generator is given by:

$$R = \begin{bmatrix} -0.628 & 0.588 & 0.04 & 0 \\ 1.96 & -2 & 0.04 & 0 \\ 0.04 & 0 & -0.236 & 0.196 \\ 0.04 & 0 & 1.96 & -2 \end{bmatrix}.$$

Assuming $o = 2$ and $l = 2$ as the degrees of polynomials $\hat{R}_a(s)$, and $\hat{Q}_a(s)$ in the Padé approximation we have:

$$\hat{R}_a(s) = s^2 - 6s + 12$$
$$\hat{Q}_a(s) = s^2 + 6s + 12.$$

Also, we will have $d = 3$, and matrices $\hat{H}_a(s)$, and $\hat{G}_a(s)$ using (21), and (27) are given by:

$$\hat{H}_a(s) = s^3 + \begin{bmatrix} 5.372 & 0.588 & 0.04 & 0 \\ 1.96 & 4 & 0.04 & 0 \\ 0.04 & 0 & 5.764 & 0.196 \\ 0.04 & 0 & 1.96 & 4 \end{bmatrix} s^2 + \begin{bmatrix} 5.232 & 3.528 & 0.24 & 0 \\ 11.76 & -30 & 0.24 & 0 \\ 0.24 & 0 & 7.584 & 1.176 \\ 0.24 & 0 & 11.76 & -30 \end{bmatrix} s + \begin{bmatrix} -7.536 & 7.056 & 0.48 & 0 \\ 23.52 & -24 & 0.48 & 0 \\ 0.48 & 0 & -2.832 & 2.352 \\ 0.48 & 0 & 23.52 & -24 \end{bmatrix},$$

$$\hat{G}_a(s) = s^3 + \begin{bmatrix} 5.372 & 0.588 & 0.04 & 0 \\ 1.96 & 1 & 0.04 & 0 \\ 0.04 & 0 & 5.764 & 0.196 \\ 0.04 & 0 & 1.96 & 1 \end{bmatrix} s^2 + \begin{bmatrix} 8.232 & 3.528 & 0.24 & 0 \\ 11.76 & 4 & 0.24 & 0 \\ 0.24 & 0 & 10.584 & 1.176 \\ 0.24 & 0 & 11.76 & 4 \end{bmatrix} s + \begin{bmatrix} -7.536 & 7.056 & 0.48 & 0 \\ 23.52 & -24 & 0.48 & 0 \\ 0.48 & 0 & -2.832 & 2.352 \\ 0.48 & 0 & 23.52 & -24 \end{bmatrix}.$$

Then matrices A_1, B_1, A_2 are computed using (23), (24), and (26) respectively.

The corresponding probability vector $\boldsymbol{\pi}$ for infinitesimal generator matrix R is given by:

$$\boldsymbol{\pi} = \begin{bmatrix} 0.3864 & 0.1136 & 0.4554 & 0.0446 \end{bmatrix}.$$

For $K = 6$, i.e., buffer size is 6, using (14) $\mathbf{y}_{0,K}$ is given by:

$$\mathbf{y}_{0,K} = \begin{bmatrix} 0.1331 & 0.0206 & 0.2487 & 0.0124 \end{bmatrix},$$

and the CDF of the stationary waiting time is:

$$\begin{bmatrix} \Pr(W^1 \leq \Delta) = \\ 0.6309 \quad 0.7693 \quad 0.8559 \quad 0.9157 \quad 0.958 \quad 0.9891 \\ \Delta = 1, 2, ..., 6 \end{bmatrix},$$

where, $W^1(x)$ is the waiting time for UE_1. According to (15), we have:

$$\begin{bmatrix} \Pr(W^1 > \Delta) = \\ 0.6306 \quad 0.3942 \quad 0.2463 \quad 0.1441 \quad 0.0717 \quad 0.0186 \\ \Delta = 1, 2, ..., 6. \end{bmatrix},$$

The same procedures can be applied for the second UE, and we have:

$$\begin{bmatrix} \Pr(W^2 > \Delta) = \\ 0.5252 \quad 0.2810 \quad 0.1571 \quad 0.0833 \quad 0.0379 \quad 0.0092 \\ \Delta = 1, 2, ..., 6. \end{bmatrix},$$

Also, the probability that the weighted average waiting time, W^T, of the cluster exceeding a certain threshold, can be derived. Since the fractions of packets which belong to the first and second UEs are 0.56 and 0.44, respectively, we have:

$$\begin{bmatrix} \Pr(W^T > \Delta) = \\ 0.5842 \quad 0.3444 \quad 0.2071 \quad 0.1173 \quad 0.0568 \quad 0.0145 \\ \Delta = 1, 2, ..., 6. \end{bmatrix},$$

Performing a simulation study, the corresponding probabilities are also derived in a scenario in which packets are generated for 10 periods or 100 seconds for a network consists of two UEs with traffic parameters as mentioned in section 5.1 and buffer size of 6. The network has been realized for 1000 times. There is one frequency sub-channel and UEs get service every 20 ms. The results are:

Table I. Common parameter settings used for simulations

Parameter	Value
m	10
λ_s	1 switch/LTE-frame (0.1switch/ms)
α_i	Normal(0.25,0.1) within [0.02,0.48]
Number of UEs in each cluster	100
Network Realizations	100

$$\begin{matrix} & & & \Pr(W^1>\Delta)= & & \\ [0.6158 & 0.3984 & 0.2545 & 0.1555 & 0.0865 & 0.0378] \\ & & & \Pr(W^2>\Delta)= & & \\ [0.5018 & 0.2830 & 0.1622 & 0.0897 & 0.0455 & 0.0180] \\ & & & \Pr(W^T>\Delta)= & & \\ [0.5664 & 0.3485 & 0.2148 & 0.1272 & 0.0688 & 0.0292] \\ & & & \Delta=1,2,...,6. & & \end{matrix}$$

Also, the probabilities of packet loss using (16) are 0.0279, 0.0135, and 0.0216 for UE_1, UE_2, and the cluster, respectively. The simulation results show that the corresponding values are 0.0253, 0.0112, and 0.0193, respectively, which are consistent with the analysis.

6. RESULTS AND DISCUSSION

We simulate an LTE cell in three different scenarios where 10 sub-channels are shared between two M2M clusters. Each cluster has 100 UEs which are served in specific time intervals as it is shown in Fig. 1. In the first scenario, the UEs of both clusters are uncoordinated and UEs have unlimited buffer size. For the second and third scenarios, assuming limited buffer size, the UEs operate according to the uncoordinated and coordinated models respectively. The general simulation parameters are summarized in table I and the specific parameters for each scenario are given in table II. The reported results are the average of 100 runs where the corresponding 95 percent confidence intervals are also reported.

6.1. Uncoordinated UEs with unlimited buffer

In the first scenario, the traffic of UEs are generated according to the uniform traffic model with deterministic service time intervals for each cluster. In Fig. 5 the probabilities of delay violation in terms of required number of service times using (6) and simulations are shown. As this figure shows the results of approximate analysis are consistent with the simulations.

The gap between the analytical and empirical results backs to this fact that in our analysis it is assumed that the traffic of each cluster is generated for infinite time and the system is in steady state. However, in our simulation, the 3GPP specification is followed for traffic generation

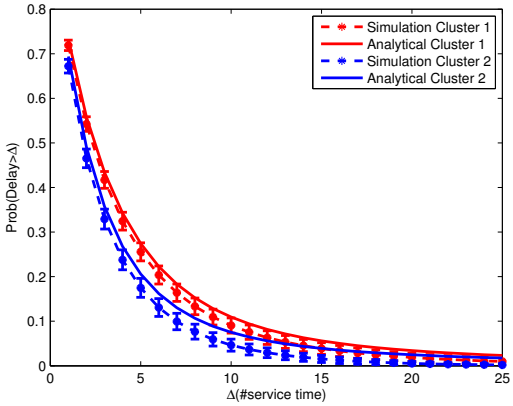


Figure 5. Delay violation probability of uniform traffic for the unlimited buffer size

Table II. Parameter settings for three different scenarios - S1, S2, and S3 denote scenario 1, scenario 2, and scenario 3, respectively.

Scenario	Parameter	Value
S1 and S2	θ	uniform(0,1)
	simulation time	60 s
S2 and S3	Buffer Size	2, 4, 6, 8
	heavy loaded $(T_g, \lambda_R, \lambda_A)$	20ms, 0.0125 pkt/ms, 0.125 pkt/ms
	lightly loaded $(T_g, \lambda_R, \lambda_A)$	20ms, 0.01 pkt/ms, 0.1 pkt/ms
S1	cluster 1 $(T_g, \lambda_R, \lambda_A)$	20ms, 0.0125 pkt/ms, 0.125 pkt/ms
	cluster 2 $(T_g, \lambda_R, \lambda_A)$	40ms, 0.00625 pkt/ms, 0.0625 pkt/ms
	θ	Beta(3,4)
S3	simulation time	10 s

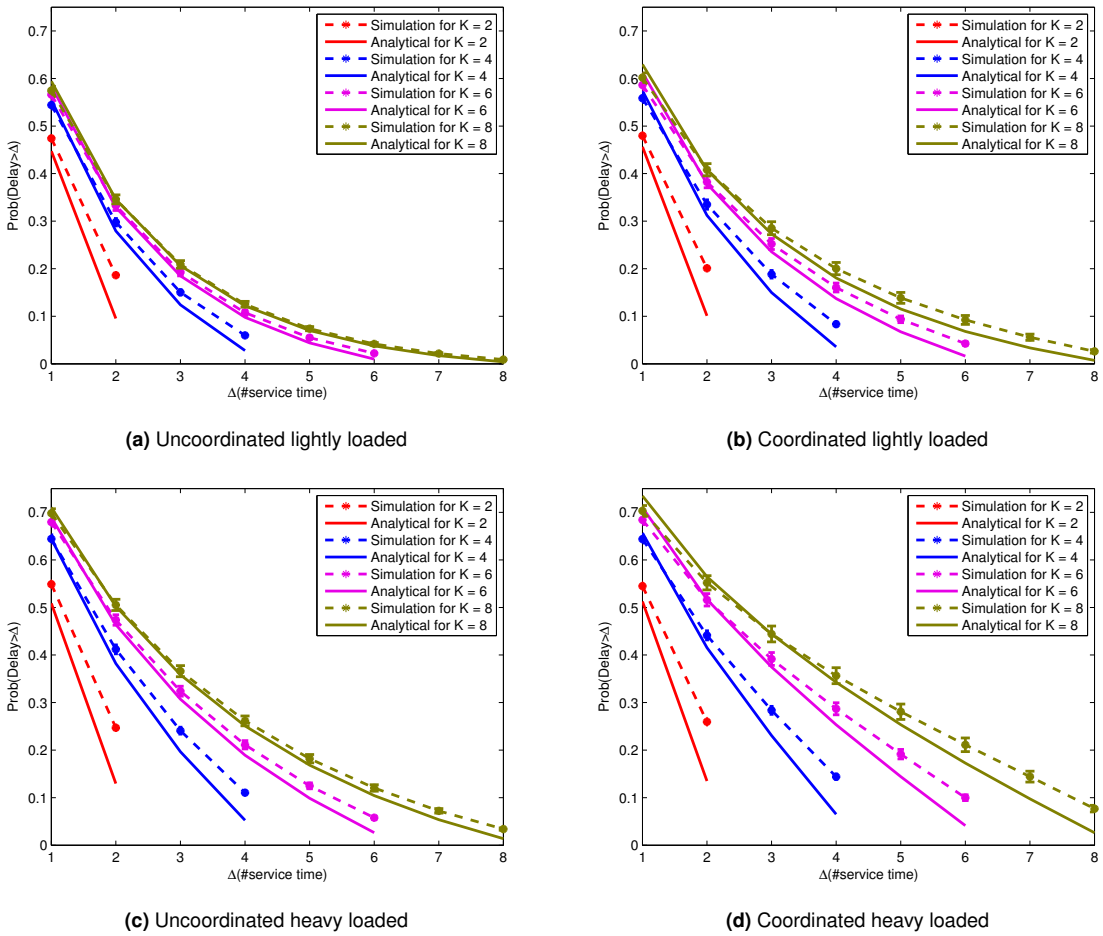
where the network is simulated for 60 seconds or 6000 LTE frames (see table II). Hence, by increasing the simulation time the results become more consistent.

Also, it has to be noticed that the system busy time $\rho = \bar{\lambda} T_g$ is the same for both clusters. Recall that λ_c indicates the rate of changes and using (5), we have $\lambda_c = 2$ and $\lambda_c = 4$ for cluster 1 and 2, respectively. As the rate of changes is increased the UE has enough chance to be in both traffic modes, so its traffic converges to a Poisson process with the mean rate of two underlying Poisson processes. Therefore, for a higher rate of changes, less QoS degradation is expected as it is shown in Fig. 5.

6.2. Uncoordinated and Coordinated UEs with limited buffer

In the second simulation scenario, it is assumed that UEs have uncoordinated traffic with finite buffer size. The results are reported for four different buffer sizes as given in table II. The traffic of each UE is generated according

Trans. Emerging Tel. Tech. 0000; **00**:1–12 © 0000 John Wiley & Sons, Ltd.
DOI: 10.1002/ett
Prepared using ettauth.cls



(a) Uncoordinated lightly loaded

(b) Coordinated lightly loaded

(c) Uncoordinated heavy loaded

(d) Coordinated heavy loaded

Figure 6. Delay violation probability of coordinated and uncoordinated traffic models for lightly and heavy loaded scenarios with limited buffer size

to the $uniform(0,1)$ pdf and the probability of delay violation in terms of the number of required service times are shown for light and heavy load clusters in Fig. 6 (a) and Fig. 6 (c) respectively. In Fig. 7 (a), the loss probability for different buffer sizes are shown for both light and heavy traffic.

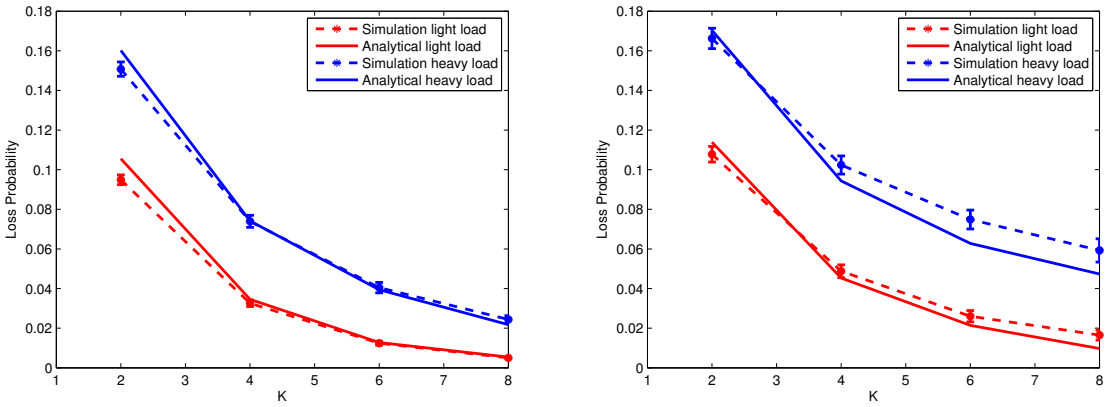
The gap between analytical and simulation results are rooted in the approximation error with MMPP/D/1/K model when we use Padé approximation. To improve the precision we can increase the degree of polynomials, o and l, in the Padé approximation at the cost of higher computational complexity. In this simulation, both parameters o, and l are equal to 2.

In the third simulation scenario, coordinated traffic model is considered where the traffic of each UE follows $Beta(3,4)$ distribution. UEs have a limited buffer size, and packets are served in deterministic time intervals as given in table II. The probability of delay violation for light load and heavy load traffic are shown in 6 (b) and 6 (d), respectively. As expected, compared to the uncoordinated scenario the probabilities of delay violation are higher due to the burstiness of the traffic in coordinated scenario. Also, Fig. 7 (b) shows that the probability of packet loss for the coordinated scenario is greater compared to the uncoordinated scenario for all buffer sizes and network load.

The performance gap between analysis and simulation in this scenario is also partly rooted in the Padé approximation. It also rooted in approximating the Beta(3,4) pdf with a piecewise constant function which has a constant value at each slot. Hence, the performance gap can be decreased by selecting smaller values for Δt at the cost of increasing the number of model states and hence the computational complexity. As a compromise between precision and the computational complexity, the slot size is set to 50 LTE frames.

6.3. Effect of buffer size

In Fig. 7 the effect of increasing the buffer size on decreasing the packet loss probability is shown. From



(a) Uncoordinated **(b)** coordinated

Figure 7. Loss probability of heavy and lightly loaded traffic for different buffer sizes

this figure, it is found that the improvement in the loss probability saturates as the buffer size is increased especially in the coordinated model. That is, in the coordinated model increasing the buffer size does not much help in decreasing the loss probability.

This observation suggests that the underlying traffic in the coordinated traffic model has a heavy tail distribution [16]. In this regard, a single M2M UE with parameters $\alpha = 0.25$, $\lambda_R = 0.0125$ pkt/ms, $\lambda_A = 0.125$ pkt/ms, and a UE with Poisson traffic and parameter $\lambda = 0.035$ pkt/ms is considered. Note that the average number of generated packets from both UEs is the same. In Fig. 8 the CDFs of the inter-arrival times of the generated traffics are shown. Note that the x-axis is logarithmically scaled. It is found that in the coordinated traffic model, the probabilities of observing short and long inter-arrival times are noticeable compared to the Poisson traffic. This behavior justifies why increasing the buffer size does not help much in the coordinated model.

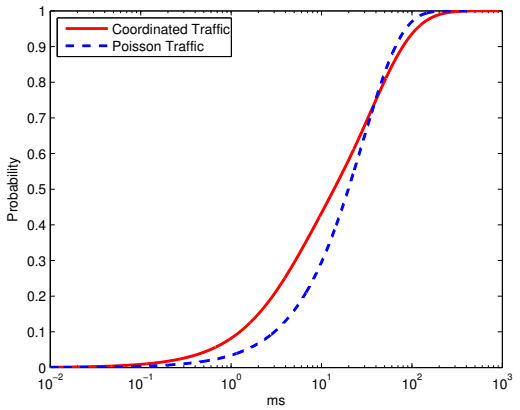


Figure 8. Empirical distributions of packet inter-arrival times for coordinated traffic model and a Poisson process with the same average number of arrivals.

7. CONCLUSION AND FUTURE WORKS

MMPP and CMMPP are used to model the non-Poisson M2M traffic of M2M communications in the uncoordinated and coordinated models. The proposed model is flexible and can be used to explain the recently reported behaviors in the traffic of M2M communications. Then Fixed AGTI is used as a simple scheduling scheme in M2M communications to evaluate the delay and loss of the packets in different scenarios using the machinery of queuing theory. The analysis and numerical results show that UEs with coordinated traffic model experience greater delay violation and packet loss probabilities compared to the UEs with uncoordinated traffic. In future works, more sophisticated scheduling algorithms are considered which take into account the channel or queue status of UEs and investigate its effect on decreasing the incurred packet delay and loss. Furthermore, by considering inter-cluster scheduling and exploiting the shared resources in a dynamic manner the loss probability could be decreased. We did not consider specific priority for each cluster which needs to use more sophisticated queueing theory models in analysis. Another interesting research direction is analyzing the coexistence scenario in which H2H and M2M communications exploit from shared resources and a group of uncoordinated or coordinated UEs could take advantage of underutilized resources by H2H communication in a dynamic manner.

Trans. Emerging Tel. Tech. 0000; **00**:1–12 © 0000 John Wiley & Sons, Ltd.
DOI: 10.1002/ett
Prepared using ettauth.cls

REFERENCES

[1] Muhammad Zubair Shafiq, Lusheng Ji, Alex X. Liu, Jeffrey Pang, and Jia Wang. A first look at cellular machine-to-machine traffic: Large scale measurement and characterization. *SIGMETRICS Perform. Eval. Rev.*, 40(1):65–76, June 2012.

[2] Antonis G Gotsis, Athanasios S Lioumpas, and Angeliki Alexiou. M2m scheduling over lte: Challenges and new perspectives. *Vehicular Technology Magazine, IEEE*, 7(3):34–39, 2012.

[3] Athanasios S Lioumpas and Angeliki Alexiou. Uplink scheduling for machine-to-machine communications in lte-based cellular systems. In *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, pages 353–357. IEEE, 2011.

[4] Shao-Yu Lien, Kwang-Cheng Chen, and Yonghua Lin. Toward ubiquitous massive accesses in 3gpp machine-to-machine communications. *Communications Magazine, IEEE*, 49(4):66–74, 2011.

[5] Shao-Yu Lien and Kwang-Cheng Chen. Massive access management for qos guarantees in 3gpp machine-to-machine communications. *Communications Letters, IEEE*, 15(3):311–313, 2011.

[6] Antonis G Gotsis, Athanasios S Lioumpas, and Angeliki Alexiou. Evolution of packet scheduling for machine-type communications over lte: Algorithmic design and performance analysis. In *Globecom Workshops (GC Wkshps), 2012 IEEE*, pages 1620–1625. IEEE, 2012.

[7] Antonis G Gotsis, Athanasios S Lioumpas, and Angeliki Alexiou. Analytical modelling and performance evaluation of realistic time-controlled m2m scheduling over lte cellular networks. *Transactions on Emerging Telecommunications Technologies*, 24(4):378–388, 2013.

[8] Katerina Smiljkovic, Vladimir Atanasovski, and Liljana Gavrilovska. Machine-to-machine traffic characterization: Models and case study on integration in lte. In *Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 2014 4th International Conference on*, pages 1–5. IEEE, 2014.

[9] Navid Nikaein, Markus Laner, Kaijie Zhou, Philipp Svoboda, Dejan Drajic, Milica Popovic, and Srdjan Krco. Simple traffic modeling framework for machine type communication. In *Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on*, pages 1–5. VDE, 2013.

[10] Markus Laner, Philipp Svoboda, Navid Nikaein, and Markus Rupp. Traffic models for machine type communications. In *Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on*, pages 1–5. VDE, 2013.

[11] IM Delgado-Luque, F Blanquez-Casado, M Garcia Fuertes, G Gomez, MC Aguayo-Torres, J Tomas Entrambasaguas, and J Banos. Evaluation of latency-aware scheduling techniques for m2m traffic over lte. In *Signal Processing Conference (EUSIPCO)*, pages 989–993, 2012.

[12] 3GPP. Study on ran improvements for machinetype communications. Technical Report, TR 37.868, 2011.

[13] Sang H Kang, Dan K Sung, and Bong D Choi. An empirical real-time approximation of waiting time distribution in mmpp (2)/d/1. *Communications Letters, IEEE*, 2(1):17–19, 1998.

[14] Ward Whitt. Continuous-time markov chains. 2012.

[15] Nail Akar and Erdal Arikan. A numerically efficient method for the map/d/1/k queue via rational approximations. *Queueing systems*, 22(1-2):97–120, 1996.

[16] Vern Paxson and Sally Floyd. Wide area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on Networking (ToN)*, 3(3):226–244, 1995.

A. ANALYSIS OF MMPP/D/1/K QUEUE

The performance analysis of MMPP/D/1/K queue via Padé approximation is proposed in [15]. In the following a brief review of the performance analysis of MMPP/D/1/K queue with First In First Out (FIFO) service discipline is given. Consider an MMPP/D/1 queue with n states in which the arrival rate in state j is denoted by λ_j. Let Λ be the diagonal matrix of the arrival rates defined in (17) and R is the infinitesimal generator matrix which determines the transition rate among the states.

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}. \tag{17}$$

Let the waiting time vector of MMPP/D/1 queue be defined as (18):

$$\mathbf{w}(x) = [w_1(x), w_2(x), \dots, w_n(x)], \tag{18}$$

where, $w_j(x)$ is the stationary probability that at an arbitrary time the arrival process be in state j and the waiting time of an unfinished work be at most x, i.e., the waiting time be less than or equal to x service times. Parameter n is the number of states. The Laplace-Stieltjes Transform (LST) of the waiting time vector, $\mathbf{w}(x)$, is given by:

$$\hat{\mathbf{w}}(s) = \mathbf{y}_0[sI + D_0 + D_1 e^{-s}]^{-1}, \tag{19}$$

DOI: 10.1002/ett
Prepared using ettauth.cls

where $D_1 = \Lambda$, and $D_0 = R - \Lambda$. The j^{th} element of vector $\mathbf{y}_0$, y_{0j}, is the stationary probability that at an arbitrary time the arrival process be in state j and the number of requests in queue be zero. Note that the e^{-s} term in (19) is the LST of the deterministic service time with unity service rate.

The irrational e^{-s} term can be accurately resubstituted using Padé approximation by $\frac{\hat{R}_a(s)}{\hat{Q}_a(s)}$ where $\hat{R}_a(s)$ and $\hat{Q}_a(s)$ are polynomials of degrees o and l, respectively and $o \leq l$. Hence (19) can be rewritten as (20).

$$\hat{\mathbf{w}}(s) = \mathbf{y}_0[sI + D_0 + D_1\frac{\hat{R}_a(s)}{\hat{Q}_a(s)}]^{-1}. \tag{20}$$

It is assumed that the coefficient of the highest degree term in $\hat{Q}_a(s)$ is one, i.e., $\hat{Q}_a(s) = s^l + q_{a,l-1}s^{l-1} + ... + q_{a,1}s + q_{a,0}$. Now, the $n \times n$ polynomial matrix $\hat{H}_a(s)$ is defined as:

$$\begin{aligned}\hat{H}_a(s) =&(sI + D_0)\hat{Q}_a(s) + D_1\hat{R}_a(s) = \\ &s^dI + H_{a,d-1}s^{d-1} + ... + H_{a,1}s + H_{a,0},\end{aligned} \tag{21}$$

where $d = l + 1$.

In the limited buffer case, MMPP/D/1/K, the waiting time vector can be calculated by (22):

$$\mathbf{w}(x) = \begin{cases}\mathbf{y}_{0,K}B_1e^{A_1x}C_1, & 0 \leq x \leq K,\\ \mathbf{y}_{0,K}B_1e^{A_1K}e^{A_2(x-K)}C_1, & K \leq x \leq K+1,\end{cases} \tag{22}$$

where the j^{th} element of vector $\mathbf{y}_{0,K}$, $y_{0,K}(j)$ is the stationary probability that at an arbitrary time the arrival process be in state j and the number of packets in the queue be zero given that the queue (or system) size is K.

In (22) A_1 is an $nd \times nd$ matrix and is computed by (23):

$$A_1 = \begin{bmatrix}0 & 0 & \dots & 0 & H_{a,0}\\ I & 0 & \dots & 0 & H_{a,1}\\ 0 & I & \dots & 0 & H_{a,2}\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \dots & I & H_{a,d-1}\end{bmatrix}. \tag{23}$$

Also, B_1 is an $n \times nd$ matrix calculated in a recursive way as given in (24):

$$\begin{aligned}&B_1 = \begin{bmatrix}B_{1,1} & B_{1,2} & \dots & B_{1,d}\end{bmatrix},\\ &B_{1,1} = I,\\ &B_{1,i} = q_{a,d-i}I - \sum_{j=1}^{i-1}B_{1,j}Ha, d-i+j \qquad i = 2,3,...,d.\end{aligned} \tag{24}$$

Finally, C_1 is an $nd \times n$ constant matrix.

$$C_1 = \begin{bmatrix}I\\ 0\\ 0\\ \vdots\\ 0\end{bmatrix}. \tag{25}$$

For the case that $K \leq x \leq K + 1$ in (22), the $nd \times nd$ matrix A_2 is given by:

$$A_2 = \begin{bmatrix}0 & 0 & \dots & 0 & G_{a,0}\\ I & 0 & \dots & 0 & G_{a,1}\\ 0 & I & \dots & 0 & G_{a,2}\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \dots & I & G_{a,d-1}\end{bmatrix}, \tag{26}$$

where $\hat{G}_a$ is computed by (27):

$$\begin{aligned}\hat{G}_a(s) =&(sI + (D_0 + D_1))\hat{Q}_a(s) = \\ &s^dI + G_{a,d-1}s^{d-1} + ... + G_{a,1}s + G_{a,0}.\end{aligned} \tag{27}$$

Vector $\mathbf{y}_{0,K}$ is given by solving linear equation (28):

$$\mathbf{y}_{0,K}B_1e^{A_1K}e^{A_2}C_1 = \boldsymbol{\pi}, \tag{28}$$

where $\boldsymbol{\pi}$ is the stationary probability vector of being in each state. For given matrices A_1, B_1, C_1, A_2, and vector $\mathbf{y}_{0,K}$, waiting time vector for MMPP/D/1/K queue can be calculated using (22).

Having the waiting time for each state, the CDF of waiting time for the system, i.e., for all states, $W(x)$ is computed by $W(x) = \mathbf{w}(x)\mathbf{e}$, where $\mathbf{e}$ is the unitary column vector. Also, the loss probability P_{loss} is obtained from

$$P_{loss} = \frac{(\boldsymbol{\pi} - \mathbf{w}(K))D_1\mathbf{e}}{\bar{\lambda}}, \tag{29}$$

where $\bar{\lambda}$ is the mean arrival rate.

Trans. Emerging Tel. Tech. 0000; **00**:1–12 © 0000 John Wiley & Sons, Ltd.
DOI: 10.1002/ett
Prepared using ettauth.cls