Optimized Energy Management for Mixed Uplink Traffic in LTE UE

Vinod Mirchandani and Peter Bertok
School of Computer Science & IT, RMIT University, Melbourne, Australia
Email: {vinod.mirchandani, peter.bertok}@rmit.edu.au

Abstract—Battery life is a major issue for any mobile equipment, and reducing energy consumption via energy management in 3G LTE user equipment (UE) will be essential for the delivery of a variety of services. Discontinuous transmission (DTX) and reception (DRX) have been designed to facilitate power management, but they can provide energy savings only via proper tuning. Relevant work in the literature mainly pertains only to discontinuous reception mode (DRX) for downlink data. However, today’s increasingly powerful UEs can generate and upload significant amount of data. This paper proposes an energy management framework applicable to both discontinuous transmission (DTX) and DRX power saving modes. In particular, in DTX mode it can reduce UE energy consumption for uplink intensive applications like telemedicine or social networking. The proposed novel energy management framework is based on jointly using a-priori analytical evaluation of a M/G/1/K finite uplink queue system for mixed traffic with an optimized DTX/DRX algorithm. DTX mode is modeled by an expression, through which the impact of quality of service (QoS) parameters on the UE’s mean energy consumption for uplink transfer is determined. The model extracts and operates on the values computed for the M/G/1/K queue. Finally, a dynamic energy management algorithm for DRX/DTX modes is proposed for energy consumption optimization based on an integrated Analytical Hierarchy Process (AHP) and Grey Relational Analysis (GRA). Analytical evaluation has shown that using our algorithm to tune DTX can achieve 49-73% energy saving over not using DTX.

Index Terms— Energy management, Performance evaluation, LTE, M/G/1/K, AHP, GRA, DTX

I. INTRODUCTION

3G Long Term Evolution (LTE) is the first true next generation mobile network (NGMN) alliance compliant mobile network technology [1]. It is a high capacity all IP based wireless network that is expected to be used for many years to come [2]. Notable attributes of 3G LTE are: uplink and downlink peak data rates of 50 Mb/s and 100 Mb/s respectively, better spectrum efficiency [3] and flexibility, cost effectiveness of infrastructure, low power requirement by wireless terminals such as mobile phone or laptop computer, robustness to channel variations etc [4]. These attributes make 3G LTE network a strong candidate for the support of mission critical and commercial services such as m-health, e-commerce, m-gaming and smart grids.

In 3G LTE systems, there will be a high prevalence of diverse mobile terminals such as smart phones, laptops and iPads. Each one of these devices has a different battery capacity to support different applications over varying lengths of time. In the development of mobile communications technology battery performance has not kept pace with the advancements made in computing power. The wide variety of applications that may run simultaneously on 3G LTE handsets necessitates the UE to be used over prolonged periods of time. Reducing power consumption by switching off the UE transmitter has been a commonly accepted method, as transmitting circuits waste little power during switch-on/switch-off [5]. There is an increasing motivation for the UE to make use of the DRX/DTX (discontinuous reception/ transmission) framework feature provided by 3GPP to conserve battery power and cope with energy requirements of the applications. The main aim of our work was the creation of a dynamic algorithm based on DTX/DRX, for effective energy savings by the UE.

Our work is client-centered i.e. based on UE, and can be applied both to the DRX and DTX modes. We will explain our work in the context of DTX mode for uplink traffic transfer, because: (i) the emergence of applications, such as those related to telemedicine and social networking are expected to generate substantial uplink traffic and (ii) not much work to date seems to have addressed DTX mode. Our work will facilitate the conservation of battery power in the UE, and for this it focuses on using the DTX feature along with the quality of service (QoS) metrics computed for the finite M/G/1/K uplink queue system that serves traffic mixes of video, VoIP, and general TCP. (Note: In section III, we justify the assumption for M/G/1/K model in our work.) Specifically, the expected waiting time, blocking probability and throughput of the mixed traffic are determined through an analytical method. These computed parameters are then used to modulate the DTX cycle i.e. ON/OFF times, by taking into account the QoS bounds of the traffic – number of packets that can be
dropped, how long the packets can be buffered i.e. delay. By estimating these parameters, we can dynamically shape the traffic i.e. increase delay, blocking probability within the allowable bounds of QoS, even prior to the arrival of the traffic packets in the UE buffer. We then optimize the duty cycle of the DTX mode by applying an integrated analytical hierarchy process (AHP) and grey relational analysis (GRA), which to our knowledge has not been applied so far in this area. (Note: DTX cycles consist of power ON/OFF periods i.e. periods during which the UE goes into active (awake) and sleep modes.)

An indispensable part of the uplink communication in the UE is the mobile RAM of finite capacity that buffers the packets for transmission to the base station (eNodeB) in LTE. Mobile RAMs are expensive [6] and consume power, so a simple, cost-effective and power efficient strategy will be to have a single buffer of suitable size to serve the mix of heterogeneous services’ packets originating from the UE instead of having multiple RAMs to buffer each traffic type or have an arbitrarily large size RAM. The traffic for heterogeneous services generated by the UE has a characteristic packet inter-arrival distribution, packet size distribution and ON/OFF duration distribution. TCP traffic burstiness is generally characterized by long-range dependent (LRD) distribution such as Pareto or lognormal. Thus, if Poisson arrival of packets are assumed then we can consider the queue system formed in a finite buffer that stores a mix of TCP and UDP data traffic to have a generalized service time distribution and be of M/G/1/K type [7].

Motivated by observations on the M/G/1/K buffer and in particular its limitations [8], we have conducted a study of the M/G/1/K finite buffer for a mix of VoIP, video and TCP data. Although the buffer size study presented in this paper is directly applicable to LTE UE uplink buffer, it can also be suitably applied to other wireless technologies. The results of our work can lead to cost savings of mobile RAMs through proper dimensioning as well as in the savings of the limited power supply that is generally available in the UE. To the best of our knowledge, no one else has conducted such a study for a M/G/1/K queue system formed of composite heterogeneous traffic in an uplink UE buffer of a wireless network. Also, most of the work in the literature is not client (UE) centered.

Our work, in so far as M/G/1/K queue systems are concerned, is based on an embedded Markov chain that uses traffic models for VoIP, video and TCP data as specified by 3rd generation partnership project (3GPP), NGMN alliances or made use of in widely acknowledged literature. Our contributions in this paper are listed below; they complement the work of [9, 10].

i) An equation to obtain the transition probabilities and hence obtain transition probability matrix ‘P’ in a M/G/1/K queue has been derived based on [9] analytical framework. Our analysis is much deeper and exhaustive than that carried out in [9], as in our work we consider a mix of three flows i.e. VoIP, video and TCP data.

ii) With the help of steady state probabilities \( \pi_n \) (n=0, 1,...,K) obtained by solving the vector equation \( \pi_n = \pi_n P \) of a system in state ‘n’, we compute:
   - Mean delay of a packet in the UE’s M/G/1/K uplink queue.
   - Combined blocking probability of VoIP and video traffic as a function of buffer size.
   - TCP throughput as a function of buffer sizes for a round trip time (RTT) and packet error rate (PER).

iii) Analysis and study of the impact of different percentage compositions of data, video and VoIP traffic ratios on the performance. These results also help to obtain the impact of real-time traffic on TCP throughput.

iv) An expression has been created for power consumed by the UE, while the packets are waiting in the queue and being transmitted. From this expression, we obtain (a) the impact of increase in average waiting time of a packet in the queue on energy consumption (b) the impact of increase in blocking probability of the traffic on energy consumption and (c) the impact of increase in transmitter’s (RF Modem) ON time on energy consumption.

v) An algorithm for efficient energy management of the DTX mode for UE uplink data transfer. The algorithm is based on the combination of Analytical Hierarchy Process (AHP) and Grey Relational Analysis (GRA).

The rest of the paper is organized as follows: Section II reviews the related literature associated with mixed traffic performance in wireless networks and with energy management. Section III explains the assumption for M/G/1/K model, our analytical model for heterogeneous traffic mix in a M/G/1/K queue system as well as simulation models. It also presents and discusses the results of the M/G/1K study. Section IV presents our analytical expression for power usage in UE LTE and discusses key results obtained from it and simulations. It also explains our energy management algorithm and demonstrates the effectiveness of the integrated AHP and GRA approach in our algorithm to accomplish energy optimization. Section V concludes the paper.

II. LITERATURE REVIEW

Below, we first discuss significant literature on the performance of mixed traffic primarily in cellular wireless networks, after which we discuss the literature for energy management in wireless networks including 3G LTE.

A. Performance of Mixed Traffic in Wireless Networks

Chatterjee et al. [11] conducted a system wide performance study of the mixed services in a CDMA2000 cellular network to determine that both voice and data can be effectively supported. Since their work the traffic models have evolved, for example more recently the
NGMN recommends web traffic packet sizes to be Pareto distributed in sharp contrast to [11] who have considered the inter-arrival time to be Pareto distributed and not the packet sizes. Their study also neither considers video traffic in the mix nor carries out an analytical/simulation study to evaluate the implications of buffer size in the on the performance. Huang et al. [12] conducted a thorough simulation study of the feasibility of CDMA2000 to support the co-existence of voice, data and streaming video in the traffic. They progressively determined the performance of the system for each type of traffic present individually in the system, and then extended their study for mixed traffic. Their study lacks any evaluation of the impact of buffer size, either by analysis or by simulations. Also, it must be pointed out that the streaming video model is different from a live video model and hence their study is not valid for uplink traffic, as generally video streaming occurs over the downlink.

A notable and a detailed study of UDP and TCP traffic mix for Internet optical switches was carried out by Vishwanath et al. [9]. A key finding of their work is that they observe an anomalous behavior of UDP packet losses with an increase in buffer size, the anomaly being that in the range of 9KB-24KB the UDP packet loss surprisingly increases with an increase in the core routers buffer size. On the downside, the authors of [9] did not consider any specific traffic models for the UDP traffic and they considered the FTP file durations to be Pareto distributions, which is contrary to the recommendations of NGMN [1] and 3GPP [12]. These shortcomings in [9] provided us with an opportunity to: (i) carry out the performance study for a more likely M/G/1/K queue system serving a traffic mix in the UE uplink (ii) consider a mix of three heterogeneous flows - VoIP, video and TCP data, (iii) carefully adopt traffic models that have been recommended by 3GPP and NGMN and (iv) Evaluate the TCP performance over 3G LTE uplink.

The interaction between Web traffic and VoIP traffic in a 1X EVDO (3G) network was studied by Sun et al. [12] and the results obtained are used as a basis for a suitable scheduling and call admission control mechanism. Their traffic model encompasses a hierarchical construction of the traffic in terms of user session, TCP session (for web traffic only), bursts within TCP/UDP connection and the packets inside the burst. They approach the study by first obtaining the TCP/UDP connection and the packets inside the burst. They do not consider any specific traffic models for the UDP traffic and they consider the FTP file durations to be Pareto distributions, which is contrary to the recommendations of NGMN [1] and 3GPP [13]. These shortcomings in [9] provided us with an opportunity to: (i) carry out the performance study for a more likely M/G/1/K queue system serving a traffic mix in the UE uplink (ii) consider a mix of three heterogeneous flows - VoIP, video and TCP data, (iii) carefully adopt traffic models that have been recommended by 3GPP and NGMN and (iv) Evaluate the TCP performance over 3G LTE uplink.

B. Energy Management in Wireless Networks

The literature pertaining to DRX/DTX mechanism for energy management of the UE is fairly limited as it is still an emerging area for research. Almost all the work found so far in the literature pertains to the DRX mechanism to conserve power. DRX operates on the downlink data sent from eNodeB and received by the UE. The parameters to operate the DRX mechanism at the receiver are evaluated by eNodeB and passed on to the receiver through radio resource control (RRC) signaling [16]. This in turn requires interaction with the scheduling mechanism to meet the QoS requirements of the traffic for different applications. The topic of energy/power management in the client wireless devices has gained popularity due to the widespread prevalence of IEEE WLANs, WiMax, UMTS (3G) networks. The main hardware responsible for a significant use of power in client handheld devices is the wireless network interface card (WNIC) [6]. A fundamental approach to decrease energy consumption in a mobile handheld is to transmit the data in bursts, which increases the packet transfer delay as the periodic transmitter ON (TX/RX active) state is punctuated with an OFF (Tx/Rx sleep) state(s).

The issue of energy/power management in wireless networks is a challenging one, which will continue to be an important area of research. In [6] the focus is on creating an energy efficient approach for streaming applications running from a server on the Internet via a proxy to the handheld client. The architecture is server-proxy/client and the proxy can be located at the access point. The main issue here is meeting the real-time constraints of the streaming application. Here the proxy is used for maintaining an ON/OFF schedule and for informing the client of the time when there will not be any transmission, so that the client can enter a sleep period. The proxy also adjusts the schedule based on the dynamic conditions of the channel i.e. the available bandwidth and the buffer size available at the client.

A number of techniques for power management have been created around this approach, such as power saving mode (PSM) for 802.11b [17], dynamic voltage scaling [17] and disk spindown [17]. PSM is not effective when application data is received at a frequent and steady rate, and it also increases the RTT to multiples of 100ms [17]. Efforts have been made earlier to achieve power savings by controlling the TCP sender’s behavior. For example, Chan et al in [17] have proposed regulation of ACKs back to the sender in 3G networks. One solution is based on TCP congestion control, and it creates a burst of TCP data by the receiver’s advertising to the sender a TCP buffer size of zero to delay the transfer of packets.
The receiver indicating the appropriate buffer space to the sender later, the sender can release the packets in a burst. This way a smooth TCP stream is manipulated into a bursty stream that generates many sleep periods, which results in energy savings. Their scheme performs 64% better as compared to the baseline TCP in terms of Energy*Delay product. The Energy*Delay product is used in [17] as a metric to compare different energy management schemes. This seems to be a valid performance metric for energy management schemes in view of the fact that in general increasing the energy savings results in additional delay.

Complex circuitry in the 3G LTE UE is a major power drain affecting battery life [18]. Discontinuous Reception (DRX) is a method in 3G LTE to increase battery lifetime drain affecting battery life [18]. Discontinuous Reception (DRX) is a method in 3G LTE to increase battery lifetime by powering down most of its circuitry when packets are not transferred. Furthermore, the discontinuous nature of the DRX mechanism over the air interface helps to make resource utilization more efficient for applications that operate in an ON/OFF manner [18]. There is a tradeoff between the energy savings achieved due to the DRX mode and the delay that results, due to prolonging the transfer of data during the OFF periods. Therefore, it is vital to select the ON/OFF duration of the DRX cycle to decrease the impact on the QoS of the application. DRX has two cycles: Short and Long. Once the UE enters into DRX mode, the optional short DRX cycle is used for a predefined time and is succeeded with a constant long DRX cycle. Details of DRX in the context of timers and their values are well explained and given in [16, 18].

Predicting the behavior of the traffic was also proposed to reduce power in 3G LTE without affecting the user’s experience in terms of delay and throughput QoS [16]. This prediction information could then be used to switch the RF modem of the 3G LTE UE into ON (active) and OFF (sleep) periods. This may be feasible for constant bit rate (CBR) traffic such as Voice, as it is fairly predictable. However, other traffic, such as web browsing, is less predictable, and the varying traffic characteristics combined with changing traffic load in the network makes the prediction difficult.

Three network algorithms have been proposed to adapt the various DRX parameters and have been shown that with prudent setting of parameters and their adaptation efficient power savings can be achieved [16]. The three algorithms proposed are Static, Semi-Static and Dynamic DRX. In Static DRX, the DRX parameters and DRX cycle length are kept constant for the entire web-browsing session. In the Semi-Static case the DRX cycle is kept constant but the DRX ON duration parameters are optimized. Finally, in case of Dynamic DRX the inactivity timer is used and the ON duration is set to one transmission time interval (TTI). It has been shown that Dynamic DRX is the most effective amongst the three algorithms, as it involves the use of the inactivity timer [16].

In [19] the use of short DRX is proposed in addition to the regular DRX cycle. The short DRX cycle essentially is a cycle with a shorter period than the regular DRX cycle, while ON duration is the same for both. By adjusting the short DTX cycle to the burstiness of the data, optimal result for energy saving can be achieved.

A study has also been conducted to evaluate the performance of DRX to achieve energy savings for VoIP traffic under two different scheduling strategies – Dynamic scheduling and Semi-Persistent scheduling (SPS) [20]. In dynamic scheduling, the scheduler determines the users to multiplex and their assignments in the frequency domain during each transmission time interval (TTI). Whereas, in the case of SPS, that is introduced by 3GPP, the frequency resources are assigned for a period longer than one TTI, i.e. in a persistent fashion. The main advantage of SPS is a decrease in signaling overhead. The authors of [20] do not recommend the use of short DRX for VoIP call, as it will result in packets being dropped and incur higher downlink delays.

III. PERFORMANCE EVALUATION–COMPOSITE TRAFFIC M/G/1/K

A. Analytical Modelling

Although, M/G/1/K system has been well studied and analyzed for an individual traffic type in [8,23], we have analyzed it for a mix of heterogeneous traffic i.e. VoIP, video and TCP. To the best of our knowledge this has not been done so in the literature.

Markov chain representing states of a M/G/1/K queue system for composite heterogeneous traffic mix is shown below in Fig. 1. Dark thick curved lines show the bunch of transitions occurring from state 1 onwards to each of the successive states. In LTE UE the traffic will be a mix of TCP data traffic with UDP traffic such as VoIP or video or interactive video telephony.

![Figure 1. Markov chain transitions in M/G/1/K composite traffic queue](image-url)

As seen from Fig 1, the arrival of a VoIP, video or TCP packet that causes a change of state of the buffer system is well embedded in the Markov chain. This prevents any decoupling between traffic types and is also accounted for in the calculation of the transition probabilities later in this section.

The traffic models adopted by us are: (i) TCP- Poisson arrival and Pareto/Log normal service time distributions [21, 22] (ii) Video- ON/OFF with Poisson packet arrivals and deterministic service time distributions (fixed packet size) [30] and (iii) VoIP- ON/OFF talk spurts and silence periods [1]. Although VoIP traffic is CBR, it can be considered as Poisson arrival based on the impact of the approximation, and aggregation level of VoIP flows e.g. in Internet [9]. We consider VoIP traffic to have Poisson
arrival [32] and deterministic service time distributions (fixed packet size). It is shown later in section III (C) that this approximation has a negligible impact on the performance results (Fig. 3 and Fig 4). This is because the graphs for G/G/1/K queue system (VoIP arrival deterministic, video and TCP Poisson arrivals) and M/G/1/K (VoIP, video and TCP data as Poisson arrivals) overlap. This is fairly plausible in our model, because (i) VoIP traffic intensity that we have used is much lower than either TCP or video traffic and (ii) VoIP packet sizes are much smaller than either TCP or video packets and (iii) our work is not concerned with computing the jitter of VoIP packets that is influenced by packet inter-arrival distribution. In order to assess the performance of such traffic mixes in a finite size buffer, we consider the queue system to be modeled as M/G/1/K.

The TCP protocol transports the http data (web traffic) and as such TCP application traffic is representative of the web traffic that is considered herein. The truncated Pareto service time distribution for web traffic model as stipulated in [1,21,22] uses the maximum packet size ‘m’ in equation (5) to be 66,666 bytes. Equation (5) is used to obtain the mean TCP packet size that is transporting the web data. The probability $s_j$ of $j$ job arrivals during a service time is given by [23]:

$$s_j = \int_{\alpha}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t} b(t) dt$$  \hspace{1cm} (1)

where $b(t)$=service time probability density function. We represent the transition probabilities for a traffic mix by

$$p_{ij} = \sum_{y=1}^{n} \sum_{z=1}^{n} \int_{\alpha}^{\infty} \left[ \frac{\lambda y}{z!} e^{-\lambda t} b(t) dt \right] \int_{\alpha}^{\infty} \left[ \frac{\lambda z}{y!} e^{-\lambda t} b(t) dt \right]$$

where $D$ denotes TCP packets, $V$ is video packets and $A$ is audio/VoIP packets in the mix. The indices of the summations are suitably adjusted so as to generate all different combinations of packet mixes based on the final state $j$ that the transition ‘$p_{ij}$’ in the Markov chain (refer Fig. 1) passes into on arrival of $i$ jobs (packets). For TCP data, $b(t)$=Pareto service time distribution given by [21]:

$$b(t) = \frac{ck^\alpha}{(\alpha-1)}$$  \hspace{1cm} (3)

Based on literature [25], the packet sizes for VoIP and video traffic were considered to be 100 and 400 bytes (mean size), respectively and TCP mean packet size to be 1000 bytes [9]. Permissible packet sizes were carefully selected to be an integral multiple of the VoIP packet size so as to represent any packet size as a multiple of a VoIP packet. This facilitates combinations between heterogeneous service packets when the system transits from one state to another. Further, we have considered two cases for the proportion of VoIP, video and TCP data traffic intensities (0.04:0.16:0.79 and 0.10:0.40:0.49) for the graphs for G/G/1/K queue system (VoIP arrival deterministic, video and TCP Poisson arrivals) and M/G/1/K (VoIP, video and TCP data as Poisson arrivals) to illustrate the influence of real-time traffic on throughput. The mean packet size of a Pareto distributed packet is [21]:

$$\mu = \int_{\infty}^{x} f(x)dx = \int_{k}^{\infty} f(x)dx + \int_{0}^{k} f(x)dx$$

where,

$$f(x) = \frac{cx^\alpha}{m^{\alpha+1}}$$  \hspace{1cm} (4)

and $x < m$ and ‘m’ is the maximum allowed packet size 66,666 bytes [21] and the shape parameter $\alpha = 1.5$, ‘k’ is the location parameter.

From equation (5), we obtain the location parameter ‘k’ to be 2666 bits for a TCP mean packet size of 1000 bytes.

We denote the state of system as the number of packets (jobs) left in the system after a departing packet (job) has been served. As such, state zero means that zero packets are left after a departing job (packet). The probabilities of packet (job) arrivals are calculated distinctly for each traffic type from equation (1), which is a standard equation for a M/G/1/K queue system. The arrival rates for the TCP, VoIP and video traffic are considered to be mutually independent Poisson arrival processes. This helps to express the packet arrivals of each traffic type as an independent event. The change of state of a system (Fig. 1) can occur due to different combinations of traffic type arrivals, thus we have to take all possible combinations into account for calculating the transition probabilities. Note: As stated earlier the change of state of the queue system was represented through combinations of different packet types expressed as an integral multiple of VoIP packet size of 100 bytes.

We first obtained the transition probability square matrix ‘P’ by calculating the transition probabilities from equation (2) for a mix of traffic flows. Note: In equation (1) and (2) a job arrival corresponds to a packet size of 100 bytes i.e. 1 VoIP packet. Thus, we can express 1 TCP packet = 10 VoIP packets= 10 job arrivals and 1 Video packet=4 VoIP packets= 4 job arrivals. We denote $A$=Audio/VoIP packet, $D$=TCP packet and $V$=Video packet: Transition probabilities were computed as follows: $p_{0,0} = P(A \text{ in state } 0) \times P(D=0) \times P(V=0) =>$ probability that a Audio packet arrives in state 0 and while being served, probability that NO TCP and NO Video packets arrive.
\[ p_{0,k} = P(A \text{ in state } 0) \times P(A=1) \times P(D=0) \times P(V=0) \Rightarrow \text{while } A \text{ is being served, exactly one } A \text{ arrives but no } V \text{ and } D \text{ packets arrive.} \]

\[ p_{0,2} \text{ and } p_{0,3} \text{ can be computed similarly. However, for transition to higher states such as } p_{0,4}, \text{ means that either while an } A \text{ is being served, } A^4 \text{ and } D^V = 0 \text{ OR } V=0 \text{ and } A^D = 0 \text{ (Because one Video packet is } 4X \text{ audio packet size hence } 4 \text{ jobs). As the number of job arrivals i.e. } j \text{ increases, with increasing buffer size, then the number of tri-modal packet combinations to consider increases, and is non-trivial. For example transitions occurring from state } i=1 \text{ say to } j=11 \text{ where } j^K \text{ means the } 11 \text{ arrivals could be } A=11 \text{ and } V=D=0 \text{ OR } V=1 \text{ and } A=7 \text{ and } D=0 \text{ OR } V=2, A=3 \text{ and } D=0 \text{ OR } D=1 \text{ and } A=1 \text{ and } V=0. \]

From Fig. 1, \[ p_{i,j} = 0 \text{ if } i-j > 1. \]

Also, transition probability from state \( i \geq 1 \) to state \( K \) is

\[ 1 - \sum_{j=0}^{i} p_{i,j} \quad \text{and} \quad p_{0,K} = 1 - \sum_{j=0}^{K-1} p_{0,j}. \]

In the calculations of the transition probabilities, we have accounted for the arrival of VoIP/video packet’s ON/OFF characteristic by multiplying the probability of voice/video packets arrivals obtained from equation (1) with the estimated activity factors i.e. the ratio of ON time to the sum of ON and OFF times. The values of these parameters for voice were obtained from [1] and for video from [30]. The packet size distribution of the packet that is undergoing service is accounted for in equation (1) by the term \( b(t) \).

After computing the transition probability matrix, we used it in \( \pi^T \pi \Pi [26] \) to compute the steady-state probability vector \( \pi \), for each state in buffer size of ‘K’ states. For this step, we used MATLAB to obtain the Eigen vector and Eigen value of the transpose of matrix \( \pi \Pi \) from which we got the steady-state probabilities of the probability vector [27]. We then normalized the steady-state vector to obtain the normalized steady-state probabilities. This process helped us to solve the equations for steady-state vector probability quickly and efficiently even for a large number of states. The normalized steady-state probabilities are then used in equation (6) to compute the average delay of a packet in the M/G/1/K queue and equation (7) is used to compute the combined blocking probability of voice and video traffic [23]. The average delay of a packet in the M/G/1/K queue system is given by [23]:

\[ T = \frac{1}{\rho + \pi_0} \]

where, \( \pi_0 \) is the steady-state probability of state 0. In our case, \( \lambda = \lambda_{\text{VoIP}} + \lambda_{\text{video}} + \lambda_{\text{TCP}} \) and \( \rho = \frac{\lambda_{\text{VoIP}} + \lambda_{\text{video}} + \lambda_{\text{TCP}}}{\pi_0} \).

\[ P_b = 1 - \frac{1}{\rho + \pi_0} \]

One of the main challenges is to create the transition probability matrix for large number of states, which occurs progressively with an increase in buffer size. (Note: 1KB of buffer size corresponds to \( K=10 \), as each state represents 100 bytes of VoIP packet. The corresponding transition probability matrix for 1 KB buffer size thus has 10X10=100 transition probabilities).

The **dense and sparse** functions in MATLAB helped us to first generate the matrix as per required buffer size (i.e. value of \( K \) states). However, values of transition probabilities of the order of \( 10^{-8} \) or less were approximated to zero. We limited the buffer size to 15 KB, which on one hand kept the complexity of evaluating the transition probability matrix within tractable limits. On the other hand, as shown later in Fig.3, a buffer size of 15 KB, the mean packet delay is limited to around 12 ms, and this is conducive in providing quality of service (QoS) support for real-time services such as VoIP which are sensitive to end-to-end packet latency. Larger buffer sizes would not result in QoS improvement, so the cost to benefit ratio does not justify increasing the buffer size further. There is also literature [28-29] on approximate methods for the analysis of large state Markov chains, but none seemed available for the composite traffic analysis of embedded Markov chain with a large number of states.

### B. Simulation Modelling

The simulation plays a vital role in determining any spikes in waiting times/delays that cannot be obtained otherwise, such as through steady-state based analysis.

We used OPNET 16.1 Modeler to validate all the key results via exhaustive simulations. In our simulations, we used the traditional dumbbell shaped simulation topology as shown in Fig. 2.

**Figure 2.** Simulation Topology

Most of our simulations for M/G/1/K buffer size performance for a mix of TCP and UDP traffic are confined to the shaded block and its elements, as our work is focusing on the User Equipment. For TCP performance, we also included the eNodeB protocol stack.

1) **M/G/1/K Model**

We have modeled and validated the performance of composite traffic belonging to the (i) Interactive class in two cases: continuous uplink data and FTP, and (ii) Conversational class: live video and VoIP. We considered the arrival rates of the data (TCP) packets to be generated by a Poisson process with exponentially distributed inter-arrival times \([9]\), the arrival process is Markovian in the M/G/1/K system. To generate a general service time schedule for the traffic mix of different traffic types, we considered the data packet sizes to be Pareto distributed for http traffic \([1,21]\). The expression for truncated Pareto distribution is given in equation (4). In the simulation, we have generated TCP truncated Pareto distributed packets with a mean size of 1000 bytes and VoIP packets of fixed size 100 bytes and video packets of fixed size 400 bytes \([30]\). The TCP packet sizes are generated by using the following parameters of \( P(k, \alpha) \) from equation (5) - Pareto (2666, 1.5). The traffic
models and their key parameters used in the performance evaluation are summarized in Table I. For data traffic, we consider two scenarios to carry out the performance evaluation: (i) full load in which the traffic is produced continuously, such as for bulk data transfer at the end of business hours to the head quarters and (ii) batch mode, such as FTP traffic uploading individual files. Note: We ensured that the combined utilization of VoIP, video and TCP traffic in the system was < 1, so that steady state condition of the queue was maintained.

### Table I. Key Simulation Parameters for TCP, Video and VoIP Traffic

| Parameters   | Traffic Duration Distribution | Video Arrival Distribution | VoIP Arrival Distribution | TCP Distribution parameters | TCP Continuous Source rate | FTP File size | Mean packet size bytes | Source rate
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ON/OFF Exp. (H.265 Codec) [30]</td>
<td>ON/OFF Pareto</td>
<td>Poisson Arrival exp.</td>
<td>6.53 x 10^-2 sec (ON State)</td>
<td>4.766 x 10^-2 sec (OFF State)</td>
<td>P(0.144, 1.7) -ON State</td>
<td>P(0.127, 1.02)OFF State</td>
<td>File size 2MB avg file size, 0.722MB std dev, Avg read time 180s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P(2666, 1.5)</td>
<td></td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400 includes RTP/UDP/IP hdrs</td>
<td>100 with RTP/UDP/IP hdrs</td>
<td>1000</td>
<td>1000 Pareto dist.</td>
<td>200 kb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64 kb/s</td>
<td>1.26 Mbps</td>
<td></td>
</tr>
</tbody>
</table>

### C. Results and Discussion

In this sub-section, we present the analytical and simulation results for performance evaluation of a mix of VoIP, video and TCP data traffic obtained by using OPNET for its validation. The analysis becomes more and more complex with larger buffer sizes, because an increase in the number of states in the embedded Markov chain results in a squared increase in the size of the transition probability matrix. To overcome this, two steps were taken. First, low transition probabilities (below $10^{-4}$) were approximated to zero, which reduced the accuracy of the analysis as opposed to the simulation results. This can be observed as a gap between the simulation and analytical results, particularly in Fig. 4. Second, the buffer size range was limited to 15 KB, which did not affect QoS parameters adversely, as stated earlier. The simulation not having such limits, larger buffer sizes were also examined.

Fig. 3 shows the impact of buffer size on the mean packet delay in a mix of TCP, video and VoIP traffic when the traffic intensity proportions are 0.79/0.16/0.04. It can be seen that as the buffer size increases the mean delay time monotonically increases as well, due to the increase in queue length. We can also observe that the mean delay is higher in case of 0.79 TCP data in the mix (Fig. 3) than for 0.49 TCP data in the mix (Fig 5a.). This is because the TCP packet size is ten times the VoIP packet size and 2.5 times the video packet size and the TCP traffic intensity is higher, which increases the overall occupancy of the buffer relative to the case for 0.49 TCP proportion for the same buffer size. As Fig 3 shows, the analytical result from equation (6) closely follows the simulation result.

The blocking probabilities for VoIP and video packets in the traffic mix when the traffic intensity proportions are 0.79/0.16/0.04 is shown in Fig. 4.

![Figure 3: Mean packet delay vs Buffer size.](image)

![Figure 4: VoIP and Video packet blocking probability (shown in Log scale)](image)

It can be seen that as the buffer size increases the blocking probability decreases. It is worth noting that Fig. 4 does not indicate any anomalous behavior in the blocking probability, i.e. we can not observe blocking probability increase when buffer size increase in the range around 9 KB, which was reported in [9].

However, [9] considered hundreds to a thousand of TCP flows in an optical network based core router. Whereas, in our case, we have considered only 3 flows of which only one is TCP, as the UE is unlikely to generate many more flows. From this observation, we can infer that it needs a very large number (hundreds) of bursty TCP flows to result in anomalous blocking probability behavior.

The analytical result trend is similar to the simulation results in Fig. 4 and is obtained by using equation (7). The disparity between the simulation and the analytical results in Fig. 4 increases with an increase in the buffer size due to transition probability matrix becoming more approximate with a higher number of states in the Markov chain. The traffic intensity used in equation (7) is the combined traffic intensities for VoIP and video traffic mix, and the steady state probability for state zero is for the traffic mix of VoIP, video and TCP data traffic.
Fig 5 (a) shows the comparison between the full load traffic (long-lived) TCP connection transferring bulk traffic continuously and the ON/OFF FTP traffic.

It is interesting to see that the FTP traffic relative to the persistent TCP traffic with the same traffic composition of 0.49:0.40:0.10 initially has a lower average packet delay up to around 33 KB buffer size and thereafter it becomes higher. This can be explained with the help of Fig 5(b), which shows the blocking probability of real-time traffic in the mix for FTP traffic, thus we explain Fig 5(b) first.

FTP traffic has an ON/OFF characteristic where the OFF time has a mean duration of 180 secs and the ON time based on parameters considered in Table I is of around 10-12 sec in duration.

As TCP data in FTP occurs in short bursts, it does not obstruct the real-time traffic, which results in an extremely low blocking probability for real-time traffic observed in Fig 5(b). In Fig 5(a), for the case of FTP traffic, the queue gradually builds up and after around 33 KB buffer size the delay exceeds the long-lived TCP traffic of the same composition because of the ON/OFF nature of FTP traffic coupled with less real-time packets having been dropped.

The comparison between the TCP throughputs for the two different traffic compositions i.e. 0.79 TCP, 0.16 Video and 0.13 VoIP and 0.49 TCP, 0.4 Video and 0.10 VoIP is made clear in Fig. 6. It is obtained by using equation 8 [31]:

\[ S = \frac{5(2 - 30p + 63p^2)}{T + RTT} \]  

where, \( S \) is the approximate TCP throughput in packets/sec, \( p \) is the packet error probability, \( RTT \) is the average round trip time. The TCP throughput is higher in case of traffic composition of 0.49 TCP traffic intensity because, as explained earlier with regards to Fig 5(a), it has a lower mean packet delay relative to the traffic with a 0.79 TCP traffic component. Thus, for the same TCP window size, more packets per unit time are transferred in case of 0.49 TCP traffic share than with 0.79 TCP traffic.

### IV. ENERGY MANAGEMENT

#### A. Analytical Model

3GPP has defined the DRX/DTX power saving modes [18] in 3G LTE to reduce the power drain on battery and extend battery lifetime. The main concept behind the DRX/DTX is to control the time for which the receiver/transmitter is switched on (awake), as it is a major battery power consumer. The drawback of DRX/DTX modes is extending the mean transfer time of data, which results in reduced data throughput. Thus, the parameters for DRX/DTX mode such as the ON time, DRX cycle duration, inactivity timer and short DRX cycle are carefully chosen in the network and passed on to the UE by eNodeB over the downlink via RRC signaling. A typical DRX/DTX cycle is illustrated in Fig 7.

The allowed TON times are integral multiples of a TTI (transmission time interval) sub-frame time. A TTI sub-frame has a duration of 1 ms, and the typical TON times in the literature [18] are: 1, 2, 3, 4, 5, 6, 8, 10, 20, 30, 40, 50, 60, 80, 100, 200 ms. The percentage energy reduction ratio (efficiency) due to the DRX mode is given by [18]:

\[ \frac{ME_{save} + NE_{mode}}{(M + N)E_{save}} \times 100 \]  

where M and N are the number of frames during which the UE is in DRX mode and in normal mode, respectively.
We propose analytical equation (10) to model the DTX cycle operation on the M/G/1/K uplink queue system for composite traffic mix and to obtain the power/energy consumed by the uplink queue system. Equation (10), uses the fundamental concept behind equation (9) i.e. power consumption is based on the awake/sleep states of the UE and the number of frames (bits) in each state.

The effective traffic entering the queue system i.e. the throughput after taking into consideration the blocking probability is given by the numerator of the first term (outside the brackets) in equation (10). Under steady state conditions of the queue system, the effective throughput divided by the channel capacity is the utilization of the system. The multiplication of the first term within the brackets with the utilization represents the energy that is consumed by the packets (frames) that are awaiting service i.e. during the sleep period of the DTX cycle. The second term within the brackets after multiplication by utilization represents the energy that is consumed while serving the frames during the ON (awake) period of the DTX cycle i.e. duty cycle, by the transmitter. Equation 10 will follow the non-linear characteristic of the waiting time of packets (frames) for an increase with buffer size (refer Fig 3) and is valid under steady state condition of the queue system. The significance of our equation (10) is that it expresses the power/energy consumed by the UE in terms of QoS factors, namely blocking probability and mean waiting time of a packet in the M/G/1/K queue. It also takes into account the service time and the utilization of the traffic. In the literature, we could not find any similar equation expressing DRX/DTX power consumption in terms of QoS and other related key factors.

\[
(1 - P_b) \left( \frac{\lambda_A + \lambda_V + \lambda_T}{C} \right) \sum_{n=0}^{\infty} \frac{\left( \frac{T_{OFF}}{T_{ON}} \right)^n}{n!} e^{-\lambda \Delta t} = \frac{T_{OFF} - \Delta t}{T_{ON} + \Delta t} \cdot E_{sleep} \cdot X
\]

Further, in equation (10), we assume that the arrival rates for all the traffic generated in UE are independent and follow Poisson distribution. This assumption enables us to add the arrival rates for all three independently generated traffic flows in the UE i.e. audio (VoIP), video and TCP data. Also, it can be noticed in equation (10) that whatever amount of time we decrease T_ON by, we increase the mean waiting time of a packet in the M/G/1/K queue by the same amount. That is, essentially we decrease the T_ON time to decrease the power consumed, but at the cost of an increase in the mean waiting time of the packet in the M/G/1/K queue system. As stated earlier, this is the underlying principle behind the DTX/DRX power saving mode in 3G LTE, and our equation (10) satisfies it.

We investigate equation (10) by using typical values of the parameters as shown in Table II. We determine how energy consumption in the M/G/1/K queue system of LTE UE is governed by the factors such as blocking probability, increment in waiting time and DTX/DRX T_ON time in the DTX mode equation (10).

| **TABLE II. PARAMETERS AND THEIR TYPICAL VALUES FOR EQUATION (10)** |
|-----------------|-----------------|
| **Parameter**   | **Typical Value** |
| \( \lambda_A \)  | Arrival rate of Video traffic 384 kb/s |
| \( \lambda_V \)  | Arrival Rate of VoIP traffic 12.2 kb/s |
| \( \lambda_T \)  | Arrival Rate of TCP data 256 kb/s |
| \( W_{e} \)     | Mean waiting time in Queue 5 ms |
| \( E_{sleep} \)| Energy spent per TTI during Normal mode 3000mW [20] |
| \( E_{OFF} \)  | Energy spent per TTI during DRX/DTX 11mW [20] |
| \( P_b \)      | Blocking Probability 0.03 |
| \( T_{OFF} \)  | On duration of DRX/DTX cycle As in [18] |
| \( T_{ON} \)   | Total Length of DRX/DTX Cycle 80 ms [20] |
| \( C \)        | Channel data rate 1.6 Mbps |
| \( \Delta t \)  | Increment in time. 0 to 40 ms |
| \( X \)        | Mean service time per packet based on application or their combination. ms |

While [10, 20, 32] focused on energy consumption solely by VoIP transmission over the air interface, they did not evaluate the impact of QoS factors on energy consumption in the UE uplink/downlink system.

Energy consumed by a UE in terms of mWh (milli Watt hour) is primarily determined by the transmitter power-on time and secondarily by QoS factors. T_ON duration affects the time for which the transmitter will be in the active state, and the longer the transmitter is in the active state then more number of TTIs (sub-frames) can be transmitted resulting in an increase in energy consumption by the UE. This is observed in Fig. 8, which shows two graphs of the energy consumption as a function of (i) T_ON time indicated in red color and (ii) effect of mean packet waiting time (top X-axis) indicated in black color. From equation (10) it can easily be seen that a progressive increase in mean waiting time of a packet causes the net T_ON time to decrease by the same amount. As the net T_ON time decreases, the energy consumption by the UE for the uplink traffic decreases.

![Figure 8. Influence of variation in Transmit ON time (Red color graph) and mean packet delay (black color graph) on energy consumption by UE for uplink traffic.](image)

We found that the blocking probability did not impact much in the steady-state probability range (refer Fig. 4).

**B. Simulation Modelling**

We made use of OPNET Modeler 16.1 to conduct a simulation study of the energy consumption by UE uplink transmission (RF modem) in the 3G LTE system over a 600 second simulation time. We focused on the energy...
consumed by the transmitter, and ignored other components, such as display, CPU, memory. The 3G LTE system simulated consists of detailed models of eNodeB, UE, evolved packet core (EPC) and application server, all of which closely replicate the functionality of protocol stacks used in the LTE system. In the analytical model’s equation (10) we considered the main traffic’s energy requirement only, without the power consumed in a UE for signaling, scheduling, hybrid ARQ and other inherent routine activities in the 3G LTE system.

The simulation model plays a key role in complementing the analytical study by way of taking into consideration a number of attributes that are difficult to model analytically, such as path loss and fading due to distance separating the UE from eNodeB, modulation and coding scheme (MCS) used, signaling information exchanged and the topology of the scenario: regular hexagonal, grid or random layout.

1) Simulation Scenario

We considered a realistic scenario of the 3G LTE system consisting of eNodeB, UEs, EPC and a server in a hexagonal layout as shown in Fig. 9. The main objectives of the simulation study were to evaluate the:

- energy consumed by the UE for different commonly used applications-VoIP, Video, Email, web-browsing and FTP
- energy consumed by the UE in different locations within the hexagonal cell – closest to the eNodeB and at the cell edge.

Some of the general parameters and their values used for LTE physical layer and that for the UE are shown in Table III.

<table>
<thead>
<tr>
<th>TABLE III. LTE AND UE KEY SIMULATION PARAMETERS AND VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>eNodeB</td>
</tr>
<tr>
<td>Uplink – LTE 1.4 MHz FDD. 1920 MHz base frequency</td>
</tr>
<tr>
<td>1.4 MHz Bandwidth</td>
</tr>
<tr>
<td>Downlink – 2110 MHz, 1.4 MHz bandwidth</td>
</tr>
<tr>
<td>Unlimited Power</td>
</tr>
<tr>
<td>Cyclic Prefix type: 7 symbols per slot</td>
</tr>
</tbody>
</table>

In order to assess the impact of commonly used applications available in mobile handhelds on the energy consumed by the UE, we used models for applications such as VoIP, HTTP and FTP that have been recommended by 3GPP and NGMN [1,21,22]. We can see from Fig. 10 that a persistent and high bandwidth application like FTP has the highest energy consumed whereas bursty lower rate application like interactive VoIP has the lowest power consumed. It must be mentioned here that a VoIP packet is generated every 20 ms and so the power is spread over 20 TTIs. Therefore, the energy consumed by VoIP shown in Fig. 10 is after dividing it by 20. We also observed that an interactive application such as two-way conversation voice consumes more power than one-way voice communication. This is because the transmitter does not go much into the sleep mode.

Fig. 11 shows the energy consumed by the UE for uplink transfer of TCP data at different commonly used rates, when the UE is located closest to the eNodeB and when it is at the cell edge boundary (refer Fig. 9). It can be observed that for lower rates the UE at the cell edge consumes more power than the UE closer to eNodeB. The reason for this lies in the physical layer, a 3G LTE - UE at the cell edge uses a lower modulation and coding scheme (MCS) index than the UE closer to eNodeB. This is because of lower signal power received from the eNodeB due to attenuation. The lower MCS index corresponds to lower transport block size [33] and thus more number of TTIs need to be used, which results in more power being consumed to transmit the data.

Another, interesting observation is that the power consumed by the UE at the cell edge decreases at higher
The criteria/sub-criteria at are then multiple conflicting and subjective evaluation criteria. The alternatives could be options, policies or candidates. AHP [34] is a multi-criteria decision making (MCDM) method that facilitates the selection of the most favorable alternative in complex problems, which often have multiple conflicting and subjective evaluation criteria. The alternatives could be options, policies or candidates.

A significant strength of AHP is that the criteria can be qualitative as well as quantitative. AHP has been used in many fields since it was first proposed by Saaty [34]. For instance, it has been used to select between the UMTS (3G) network and WLAN i.e. most suitable network so as to provide the user with the best available QoS at any time in different scenarios [34], or to select intermediate nodes in application-specific routing in a wireless sensor network [35]. In AHP the problem is first structured into three main hierarchies.

Step 1: The topmost level is the ‘Goal’, the second level are all the main ‘Criteria’ on which the goal is based – Each of the criteria can be divided into the sub-criteria. The third level is the ‘Alternatives’ i.e. the different available choices from which one needs to be selected in order to achieve the Goal.

Step 2: In this step, weights are allocated to each of the criteria and sub-criteria with respect to the element in the level above. In AHP the weights are allocated on a 9 point relative scale [34]. As the weight assignment is based on subjective judgment, a consistency ratio (CR) checks the consistency of weight assignment later.

Step 3: The criteria/sub-criteria at each level are then compared pair-wise in a matrix.

Step 4: The relative priorities of each criteria/sub-criteria are calculated at each level by means of normalized Eigen vector of the matrix [34].

Step 5: The global priority/weight of each alternative choice is obtained from a synthesis of priority computed for each alternative choice across all the criteria and sub-criteria.

If the consistency ratio (CR) is less than or equal to 10%, then the degree of consistency is considered to be acceptable. If the CR > 10%, then the subjective judgment will need to be revised [34]. The principal Eigen value of the comparison matrix at each level helps to determine the consistency ratio.

2) Building Block 2: Grey Relational Analysis (GRA)

Ever since the Grey system theory was proposed by Deng [36], it has been used across many fields such as hiring of personnel, prediction of serial crime and stock selection. Grey Relational Analysis (GRA) is based on Grey System theory and its main advantages are that it is computationally fast, simple and can handle unclear and incomplete information precisely. The term “Grey” in GRA means that the information is between black and white, where black represents no information and white represents all information. Fundamental to the operation of GRA is that it reduces the multi-attribute decision making (MADM) problem into a single-attribute decision making problem by synthesizing all the attributes for every alternative into a single value. This makes the comparison and selection of an alternative computationally much simpler than AHP [36]. The key steps involved in GRA are:

Step1: the compared and a target reference sequence are generated by data pre-processing that involves normalization in order to make the compared and target sequence independent of the units and scale/compress the range value of the attributes.

The obtained parameters are then used in the simulation results with the result obtained by using the analytical model of equation (10). It can be seen that the analytically obtained result closely follows the simulation results for both the UE closer to eNodeB and the UE at the cell edge. However, at higher data rates i.e. above 2000 kbps (A to B in Fig. 11), our analytical model does not take into account the MCS feature and the path loss based on the distance between UE and eNodeB. In spite of this, our analytical equation is rather accurate in the commonly used data rate range, i.e. up to 2 Mbps.

C. Proposed Energy Management Method

The proposed energy management method obtains the QoS parameters, namely mean packet delay and blocking probability in the UE uplink queue system for a traffic mix of video, voice and TCP data as was explained in section III. The obtained parameters are then used in Table V, which are then processed by the GRA to select the best alternative i.e. the best set of QoS, channel and application parameters that will result in the goal of optimal power management i.e. table VI. To increase the accuracy of the selection, the GRA makes use of the weights obtained by using the AHP in table IV. The proposed framework also provides flexibility to the user to obtain the QoS, channel and application parameters as per their own method(s) and then incorporate within it.

1) Building Block 1: AHP (Analytical Hierarchy Process)

AHP [34] is a multi-criteria decision making (MCDM) method that facilitates the selection of the most favorable alternative in complex problems, which often have multiple conflicting and subjective evaluation criteria. The alternatives could be options, policies or candidates.
Step 2: the Grey relational coefficient (GRC) between the reference sequence and the sequences that it is compared with is calculated. It essentially determines the closeness of each attribute to that of the corresponding attribute in the reference sequence.

Step 3: the degree of Grey coefficient (Grey relational grade) ranks the alternatives and the alternative that has the highest Grey relational grade is selected. The highest Grey relational grade signifies the closeness of a comparability sequence for an alternative to the ideal sequence. The Grey Relational Grade is calculated by taking into account the weight of each attribute. If all the attributes have equal significance then the weight of each attribute is \(1/n\), where \(n\) is the number of attributes for each alternative. In general, however, all the attributes may not have same weight and we obtain the weight of each attribute through AHP. This is called an integrated AHP and GRA approach, and has been shown to be reliable and practically feasible [37].

D. Proposed Energy Management Algorithm – Key Steps

Step 1 – A software agent operating within the UE will inform the UE energy management algorithm regarding the applications that are currently active in it, such as VoIP, interactive video or FTP.

Step 2: The mean waiting time of a packet in the M/G/1/K queue system of the UE will be determined by our analytical method, as explained in section III.

Step 3: If the difference between the standard QoS parameter values (such as mean packet delay, blocking probability) and the corresponding estimated values for the actual traffic is lower than a particular threshold then no power management will be conducted. If the difference is higher than the threshold, then our energy management technique will be initiated.

Step 4: The DTX cycle ON time is modulated by the value of mean packet delay in UE’s M/G/1/K queue system and the blocking probability. That is, to reduce energy consumption, the value of mean packet delay and blocking probability are increased but within limits such that the resulting mean packet delay and the blocking probability will still be below the maximum values allowed by the QoS of the application. As explained earlier with regard to equation (10), if the mean packet delay value is increased by 1 unit of time, the \(T_{ON}\) time of the DTX cycle is decreased by the same amount and energy consumption is reduced.

Step 5: Generate a AHP matrix and compute the weights of all the pre-defined criteria and sub-criteria. We assume that the DTX duty cycle can dynamically switch between the following three alternative types defined by us: long (\(T_{ON}=40ms\)), medium (\(T_{ON}=30ms\)) and short (\(T_{ON}=20ms\)). We considered, short \(T_{ON}=20ms\) because this corresponds to the size of VoIP frame/TTI size [32]. It is stressed here, that the alternative types and the criteria/sub-criteria are entirely left at the implementer’s discretion. The weights in the AHP matrices may need to be slightly adjusted so that the consistency ratio is kept below 10%.

Step 6 – GRA is used to determine the degree of Grey coefficients (Grey relational grade) for each of the alternatives. The GRC gives the deviation of the attributes (parameters) of each alternative from the ideal reference. The weights obtained from the AHP for each of the criteria/sub-criteria are used in the computation of the degree of GRCs.

Step 7- If the degree of GRC\(_{new}\) of an alternative duty cycle is greater than the degree of GRC\(_{current}\) of the currently used duty cycle, then the currently used DTX duty cycle will be replaced with this alternative.

Step 8 – The algorithm then goes back to Step 1 and the repeats the process periodically.

The algorithm explained above is summarized through a flow chart in Fig. 12.

---

**Figure 12. Proposed power/energy management algorithm**

---

**Figure 13. Comparison of energy consumption by traffic mixes with and without DTX mode**
It compares the energy consumed by UE for (i) VoIP only (ii) combination of VoIP and video traffic and (iii) combination of VoIP, video and data traffic with the corresponding energy consumed by UE when DTX mode is not used for these traffic for 600 sec. The results have been obtained by using DTX analytical model equation (10). Fig. 13 also helps to determine the impact of the DTX mode on the energy saving at the UE. We calculated the energy savings as: VoIP only traffic approx. 73%, VoIP and video traffic 62%; and VoIP, video and data traffic 49%.

a) Power Optimization

We explain the power optimization feature in our algorithm for UE DTX mode by computing the parameters that pertain to AHP and GRC. To do this, the problem is first translated into an AHP hierarchy as shown Fig 14. Level 2 of the AHP represents the three alternatives from which one has to be selected for optimized energy management, and which results from a trade-off between the different criteria. We follow steps 2-5 stated under AHP, discussed earlier, to obtain the weights for the three criteria and the global priorities (weights) for the three sub-criteria. For this we use equations (11) and (12).

\[ [W, \lambda] = eig(A) \]  

In equation (11), the eigen value of the square matrix ‘A’ for AHP is obtained. The priority vector i.e. weights for the matrix criteria and sub-criteria are also obtained [34].

\[ CI = \frac{\lambda_{max} - n}{n-1} \]  

and

\[ CR = \frac{CI}{RI} \]

CI is the consistency index, ‘n’ is the number of comparisons made at each level in the square matrix, RI is the random consistency index that is 0.58 for n=3 and \( \lambda_{max} \) is principal Eigen value. We confirmed that our AHP matrices are consistent by checking that CR < 10%.

Table IV shows the weights obtained for the criteria and sub-criteria as a result of computations made for AHP. Table V gives the attributes and their values used for each alternative. We consider three scenarios corresponding to the applications used i.e. (i) VoIP only, (ii) VoIP and Video and (iii) VoIP, Video and Data. For each of these three scenarios there are three choices for DTX/DRX cycle defined by us i.e. short, medium or long (Refer step 5 of our algorithm). Table VI shows the degree of GRC obtained after following Steps 1-3 in the earlier discussion on GRA. The ranks in Table VI are based on the degree of closeness to the ideal operating condition to achieve the goal of optimized energy management, for the attributes specified in Table V. The table confirms that the most optimal energy saving mode is for Voice application with a short duty cycle i.e. short ON time. The Grey relation coefficient \( \chi \) is obtained from equation (14) [36,38]

\[ \chi = \frac{\Delta_{min} + \zeta \Delta_{max}}{\Delta_{max} + \zeta \Delta_{min}} \]  

where \( \Delta_{max} = \max \Delta_a(j) \), \( \Delta_{min} = \min \Delta_a(j) \)  

\( \zeta = 0.5 \) [36] The degree of Grey coefficient (Grey relational grade) is shown in the right most column and it is obtained by using equation (15) [36,38]

\[ \Gamma_a = \sum_{j=1}^{n} [W_j(j)] \chi_{ai} \]  

We have obtained the weights \( W_j(j) \) for each of the attribute’s sub-criteria and criteria by using AHP (Table IV). Therefore, our approach of integrating the AHP with GRA is much more accurate than would be the case if we had considered all the attributes to be of equal weight.
TABLE V. ATTRIBUTE VALUES OF CRITERIA/SUB-CRITERIA FOR DIFFERENT APPLICATIONS

<table>
<thead>
<tr>
<th>QoS</th>
<th>Channel</th>
<th>Power Usage - Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>PER</td>
<td>Average λ (kb/s)</td>
</tr>
<tr>
<td>λ</td>
<td>Frame size</td>
<td>T_ON (ms)</td>
</tr>
<tr>
<td>T_ON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short Cycle (Voice)</td>
<td>881</td>
<td>1</td>
</tr>
<tr>
<td>Med Cycle</td>
<td>881</td>
<td>1</td>
</tr>
<tr>
<td>Long Cycle</td>
<td>881</td>
<td>1</td>
</tr>
<tr>
<td>Short Cycle (Voice+Video)</td>
<td>808</td>
<td>2</td>
</tr>
<tr>
<td>Med Cycle</td>
<td>808</td>
<td>2</td>
</tr>
<tr>
<td>Long Cycle</td>
<td>808</td>
<td>2</td>
</tr>
<tr>
<td>Short Cycle (Voice+Video+Data)</td>
<td>692</td>
<td>4</td>
</tr>
<tr>
<td>Med Cycle</td>
<td>692</td>
<td>4</td>
</tr>
<tr>
<td>Long Cycle</td>
<td>692</td>
<td>4</td>
</tr>
</tbody>
</table>

TABLE VI. GREY RELATIONAL COEFFICIENTS FOR APPLICATIONS WITH DIFFERENT DTX DUTY CYCLE TYPES

<table>
<thead>
<tr>
<th>QoS</th>
<th>Channel</th>
<th>Power Usage - Applications</th>
<th>Degree of Grey Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘S’</td>
<td>pk/s</td>
<td>Delay (ms)</td>
<td>Blk Prob</td>
</tr>
<tr>
<td>Short Cycle (Voice)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Med Cycle</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Long Cycle</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Short Cycle (Voice+Video)</td>
<td>0.56</td>
<td>0.6</td>
<td>5</td>
</tr>
<tr>
<td>Med Cycle</td>
<td>0.56</td>
<td>0.6</td>
<td>5</td>
</tr>
<tr>
<td>Long Cycle</td>
<td>0.56</td>
<td>0.6</td>
<td>5</td>
</tr>
<tr>
<td>Short Cycle (Voice+Video+Data)</td>
<td>1</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Med Cycle</td>
<td>1</td>
<td>0.33</td>
<td>0.33</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

Work on energy management in 3G LTE has been limited, and deals with down load traffic via discontinuous reception (DRX). Increasingly powerful user devices can upload significant amount of data, which requires the use of discontinuous transmission (DTX). This paper proposed a energy management framework that can be used for both DRX and DTX modes. The framework has two main parts – (i) Evaluation of QoS metrics in M/G/1/K uplink queue system and (ii) algorithm for optimal energy management. First, using our analytical model for the evaluation of M/G/1/K queue system for heterogeneous traffic, we compute the QoS metrics, namely delay, blocking probability and throughput. These a-priori estimated values are periodically passed on to our energy management algorithm, which considers the tradeoffs between these parameters for the application and makes an optimal selection of the duty cycle for the DTX. The optimization of energy management is carried out via multi-criteria decision making provided by AHP and GRA. As part of our algorithm, we have also proposed an approximate analytical expression for energy consumption in terms of QoS metrics. We have validated this expression and the analytical model of M/G/1/K queue system by means of simulations. Our results will help to conserve the energy in UE and select buffer size accordingly.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their constructive comments. Furthermore, the authors acknowledge the role of OPNET Technologies in providing them with an Educational license for OPNET Modeler 16.1 to conduct simulations.

REFERENCES


[21] RACH Capacity Analysis-Packet 1,TSG1R16(99818), http://www.3gpp.org/ftp/tsg_ran/WG1_R1/L1/TSGR1_06/docs/Pdfs/r1-99818.pdf last retrieved on 20/02/2012.


Vinod Mirchandani was born in India. He has earned PhD degree in Computer Communications from the University of Melbourne, Australia in 1998, MS in EE from the University of Saskatchewan, Canada, BE in ECE from IISc, India and BSc degree from the University of Poona, India.

He has worked (i) as a Senior Research Engineer at Motorola Labs, Australia in the area of WLANs and Performance Evaluation (ii) at the University of Sydney in the area of B3G networks (iii) at the University of Technology, Sydney (UTS) on collaborative projects with Alcatel-Lucent, Bell Labs (Paris) in the areas Self organization of Wireless mesh networks and Grid Computing (iv) at the RMIT University, Melbourne in 3GPP LTE wireless network and data privacy. He has authored more than 35 IEEE/ACM and other international publications, 3 book chapters and one US Patent.

Dr. Mirchandani has received numerous research awards and University Fellowships.

Peter Bertok received his PhD from the University of Tokyo, Japan in the area of Computer Control, and his Master of Engineering from the Technical University Budapest, Hungary in the field of data and computer communication. Currently, he is an Associate Professor at the School of Computer Science & IT, RMIT University, Melbourne, Australia. He has worked in industry as well as at research institutions.

He is a member of IEEE and ACM and has over 100 scientific publications.