International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 411-421

Addressing the Emerging challenges in Modeling and Simulating the Performance of LTE Networks

Belal Abuhaija
Member IEEE, Department of Computer Engineering
Faculty of Computers and Information Technology,
University of Tabuk, Saudia Arabia.
E-mail: babuhaija@ut.edu.sa

Abstract
There is several challenges in LTE system capacity and performance evaluation. Among other things, traffic modeling and distributions, the quality of services (QoS) requirements of such traffic, the instantaneous propagation conditions, and the number of users in the system can influence the system capacity a great deal. The convergence between the internet traffic and the LTE traffic can complicate the performance evaluation of LTE system. Therefore, in this contribution, we are aim to provide an accurate LTE system evaluation and capacity based on addressing the identified challenges. Another contribution is the utilization of the internet traffic profiles for the various applications. We are presenting a mathematical model based on Continuous Time Markov Chain (CTMC). We are utilizing a system level enhanced simulation tool by the author for numerical analysis.

Keywords: LTE Performance analysis, traffic modeling, and CTMC.

Introduction

Long Term Evolution (LTE) networks and systems as specified by 3GPP are the technology of choice to provide the solution for the last mile bottleneck of cellular networks. 3GPP community has standardized an all-IP technology based on orthogonal frequency division multiplexing (OFDMA) in the downlink. LTE architecture has two main components, The Evolved Terrestrial Radio Access Network (E-UTRAN) and the Evolved Packet core. The base station (E-NodeB) is handling all radio functionality and feedback messages. Such a flat architecture along with the realization of Multiple Input Multiple Output (MIMO) antenna technology is poised to provide hundreds of megabits in the downlink (in excess of 300Mbps), low latency, and improved mobility.

Performance analysis of complex systems such as LTE can be a very difficult task. Accounting for many challenges and issues such as the types of traffic and characteristics, the propagation conditions (slow fading), the instantaneous channel conditions (fast fading), antenna system employed and quality of services (QoS) can complicate the performance evaluation and capacity of LTE systems.

The SGi interface is connecting LTE networks and the Public data networks gateway (LTE Architecture is outside the scope of this article). Therefore, the convergence between the internet traffic and LTE traffic is essential and mandated. In recent years the internet traffic and characteristics has shown a noticeable change from the traditional FTP download, email and web browsing to a more demanding anytime anywhere multimedia applications that requires guaranteed QoS. The internet data customers are consuming huge resources for multimedia traffic as the applications are becoming elaborate and competitive for resources. The convergence of the internet and the cellular telecommunication systems (such as LTE systems) has imposed a huge demand on cellular system capacity and throughput. To keep up with the wired internet traffic; LTE is a promising technology to bridge the gap between the broadband internet and wireless communication broadband traffic. Therefore, we need to address the convergence between the internet traffic and the cellular systems traffic.

The newly developed heterogeneous applications of different traffic classes have varying parameters of quality of service (QoS) and can impose huge burden on network management. Packet switched networks (i.e. the internet) was built on best effort services and have gradually migrated to provide guaranteed bit rate (GBR) to the newly developed multimedia services such as VOIP, video on demand and video conferencing. The efficient support of various packet switched services such as VOIP and multimedia has been one of the goals of cellular systems as outlined in the standards [4]. 3GPP has standardized four major traffic types namely conversational, streaming, inter-active services, and background services [3]. The elastic traffic such as FTP and interactive traffic has been modeled as heavy tail log normal distribution while the VOIP and video conferencing are more suitable to be modeled with negative exponential distribution as explained in section 3 below. Therefore, modeling and evaluating the performance of LTE traffic capacity based on the experience gained from the internet traffic shall provide a realistic and accurate results when studying LTE system capacity. The proper traffic services' modeling is one of these challenges because of the diversity of the traffic [3]. However, in the presence of heterogeneous applications competing for the resources, the issue of ensuring the QoS for different applications with different requirements is becoming a challenge. Thus, LTE system needs to guarantee QoS while preventing other applications from bandwidth starvation in the presence of multi traffic.

Another challenge is the standardization of Multi Input Multi Output (MIMO) by 3GPP [10-13]. In any realistic analytical or simulation studies, MIMO system shall be present. Utilizing MIMO antenna systems can enhance the capacity and throughput of the base station and the network in general.

Many other issues can influence the performance of LTE system, such as radio resource scheduling algorithms, power allocation, electrical and mechanical tilting and many more [5-9].

The research community has addressed the above challenges separately; however, we have been unable to find a comprehensive study to address these issues collectively at present. Therefore, this paper as outlined above presents and identifies some of the technical challenges in evaluating and modeling LTE network systems analytically and through simulation. Another contribution is the utilization of the internet traffic profiles for the various applications. To the best of our knowledge, no study shows the internet traffic modeling impact on the performance analysis and capacity of LTE system. We also aim to show the collective impact of the above issues on the performance of LTE system.

Related Work

A number of research papers have offered an analytical model for analyzing the LTE system performance and capacity [1-2]. However, the authors in [1] and based on extension to the work done in [2] have considered three time domain schedulers; namely; maximum throughput, blind equal throughput and optimized service aware. They realized only one type of service, which is elastic service (FTP), and when the authors expand the services; the extension was for the same elastic service (FTP) but with different departure time. As explained in the previous section the traffic services are more diverse than the considered services by the authors. Such models do not hold in real implementations; in other words; with the internet heterogeneous applications; the model should reflect such diverse traffic. Another issue with the findings of the authors is that, how this service behaves in the network in the presence of other traffic and the impact on the time domain schedulers mentioned above. However, the authors in [2] did not include any propagation model or the instantaneous channel conditions and its impact on the system throughput as articulated in [19] as the dwell time depends on many factors, among others, the propagation conditions, and the velocity of the mobile as well as the communication range. The model also did not include MIMO antenna system and its impact on the capacity and user throughput. In our model, we realized the instantaneous channel conditions as suggested by [19], as well as several services as outlined in the standards [3, 16-17].

In [15], the authors have analyzed empirical data from cellular operators to study the inter arrival and the holding times. The authors have concluded that the inter arrival time is very well represented by an exponential distribution as cell calls follow a Poison process for more than 90% of the call . However, for holding time and call duration the findings were inconclusive. They found out that the call duration modeled with short tail distributions such as Erlang distribution is not suitable. The lognormal distribution is more suited for the holding time. In our model, we are utilizing lognormal distributions for the performance analysis of LTE in order to bring a better and accurate system representation and produce sound results.

The authors in [9] derived a model that includes fast fading, shadowing, and attenuation while modeling guaranteed Bit Rate (GBR) and non-guaranteed Bit Rate (N-GBR) services. The model assumed a priority for the GBR users over the N-GBR users by assigning the required radio resources for the GBR first and the rest to N-GBR. The author concluded that the GBR users could not cause any problem if the data rate requirement is below 1Mbps. However, if the data rate requirements are in the range of 3Mbps, it may cause problems to N-GBR users because of the high demands on radio resources if the GBR users have low signal to interference plus noise ratio (SINR) especially at the cell edge. The authors compared three different scheduling algorithms, round robin, proportional fair and Max SINR. However, the traffic profiles considered do not take in considerations the internet traffic mix as outlined in [16-17]. At the same time, the authors concluded that the GBR users and applications can cause problem for the elastic services but did not offer any solution to such problem.

In [20 and the references therein], the authors proposed a stochastic model for user throughput predictions. The authors proposed a model to represent the impact of estimation errors on network bandwidth availability in order to study the network resource's optimization problem under some uncertainties such as user mobility and instantaneous channel conditions in the system. The authors grouped the available solutions into three categories; network group, cell group and user group predictors. The network group predictor's model has been concerned with the whole network at once and the average throughput achievable in a location of the user most likely location. The cell group predictor models aim at identifying the next cell a user is likely to visit and the load of the cell. The user predictor models group has been concerned with empirical data and aims at fast bandwidth variations experienced by the users. The authors concluded that in the three categories mentioned there has been no single predictor that satisfies the three categories together.

In [30], the authors argued that one of the Myths in wireless communication systems is deriving the system capacity with respect to SINR measurements. They argue that the achievable system data rate is entrenched behind , where B is the bandwidth. Therefore, increasing the allocated power or increasing the bandwidth can achieve better data rates, allocating more bandwidth depends on the spectrum reframing strategy of the operator while interference and the log in Shannon formula limit increasing the power. The authors argued that the load of the base station and the number of users serviced by the base station influence the base station throughput. Enhancing the propagation channel by other methods other than increasing E-Node power such as utilizing MIMO antenna system as articulated in [31].

In [32], the authors derived a mathematical model to estimate the capacity of heterogeneous wireless networks based on the SINR constraint. They realized LTE, WLAN and Wireless Mesh Networks dedicated to internet access. The data rate experienced by the user is dependent on the distance from the access point or the base station in general (cellular or wireless). However, there are two issues with this approach; first, the authors did not address instantaneous channel (fast fading). Second, the authors considered two types of traffic services, real time (12.2 kbps) and non-real time at (128kbps) which is not realistic for the heterogeneous wireless networks.

In our proposal, SINR is measured and dynamically influenced by the instantaneous channel conditions and shadowing. In other words, the location of the customer is not the only deciding factor of the signal quality and therefore the scheduling resources. The signal quality depends on attenuation, channel fading and terrain. As far as

the number of users is concerned, we need to estimate a balance between the number of users, the service type, and base station load.

As seen from above, deriving an accurate analytical model of a complex system such as LTE can be a daunting task and depends on the aim of the study. However, none of the above research has focused on the convergence between the internet traffic modeling and the LTE system. At the same time, many issues arise when studying the performance analysis of LTE system as outlined above. Therefore, a mathematical model based on multi-dimensional continuous time Markov chain is proposed and an enhanced simulation framework for performance analysis that includes most of the challenges outlined above. Simulation models are the preferred method for performance evaluation of cellular networks to verify performance analysis throughput and capacity. To the best of our knowledge, this complete model includes GBR services and N-GBR services.

This paper is organized as follows, section 3 illustrates the model assumptions and services, and section 4 presents a mathematical formulation of the system while section 5 describes the simulation system. Discussion of the simulation results in section 6 and we conclude with section 7.

Proposed System

This section introduces the system model for the performance evaluation of LTE system, with the following model assumptions:

A. The system bit rate

The wide spread of LTE networks deployment has positively influenced the acceptance of LTE as the 4G cellular system for the future. Orthogonal Frequency Division Multiple Access (OFDMA) technology has been the technology of choice for LTE networks. LTE supports data rates up to 300Mbps in a 4x4 MIMO deployment. However, the number of physical radio resources (PRB), cell power as well as deployment scenarios have a major influence on cells bit rate and therefore the data rate is limited by the amount of interference and noise in the network. In frequency domain, resources are grouped in 12 subcarriers or 180 kHz. Thus, the scheduled physical radio resource block PRB = {1, 2, 3… r} is 180 kHz for a sub-frame duration = 1ms. Each subcarrier can carry 6 or 7 resource elements depending on the cyclic prefix. We define a set of code rate as CR = {1, 2, 3… C} and modulation M = {1, 2, 4 and 6} bits per symbol as defined in table 1. The UE transmits the Channel coding indicator (CQI) by mapping between the MCS value, transport block size, and the CQI indices as illustrated in table 1. However, the calculation of the CQI value and accordingly the determination of the number of resources to use are vender specific design issues. However, the factors that play a role in the calculations are, among other factors, the signal to interference plus noise ratio (SINR) and the instantaneous channel conditions is considered in this paper. The maximum achievable bit rate for a user depends on both the modulation and coding scheme (MCS) influenced by the instantaneous channel condition and the number of users in the cell. From the above and table 1, the user data rate can be given as in equation (1) below

$$Ri = \frac{CRi}{Ts} \log_2 Mi \sum_{1}^{r} PRBr * 12 \quad (1)$$

Ri is the user data rate; the duration of the sub-frame is 1ms, the code rate as in table 1 and the modulation depends on the signal quality.

Table 1: Simulation Parameters

CQI index	Modulation	Bits/Symbol	Coding
0	Out of range		
1	QPSK	2	0.1523
2	QPSK	2	0.2344
3	QPSK	2	0.3770
4	QPSK	2	0.6016
5	QPSK	2	0.8770
6	QPSK	2	1.1758
7	16QAM	4	1.4766
8	16QAM	4	1.9141
9	16QAM	4	2.4063
10	64QAM	6	2.7305
11	64QAM	6	3.3223
12	64QAM	6	3.9023
13	64QAM	6	4.5234
14	64QAM	6	5.1152
15	64QAM	6	5.5547

B. Modulations

In general, terms, we assume that the system provides different levels of modulation to compensate for the different levels of signal quality as specified in 3GPP standards. In LTE standards, the system allows many different modulation modes such as, (64QAM, 16QAM, QPSK and BPSK) [21]. The model design to realize different modulations per customer per service and based on the CQI value reported as in table 1.

The MCS used for each user depends on the terrain and the propagation conditions in the cell at that instance of time, reflected by signal to noise ratio (SNR) as, depicted in the following equation:

$$SNR = \frac{PtGtGr}{N0NfLFm} \quad (2)$$

Pt is the transmitted power, Gt is the gain of the transmitter, Gr is the receiver gain, Nf is the noise figure,N0 the spectral noise density, L is the measured path loss between receiver and transmitter as well as other losses (i.e. penetration losses, etc.) and Fm is the channel fade margin. For fast fading, Rayleigh model is employed as a complex Gaussian random process with zero mean and double sided unit variances N (0; 1)[10, 22].

In this model, the propagation conditions is influenced by the fast fading due to the channel coherence time being small compare to delay constraint of the channel and slow fading which is influenced by shadowing, log of the distance and the terrain. The intensity of the users' arrival to the system is independent of the channel conditions and / or the MCS. In other words, we are not splitting the coverage area into zones. Our approach is more practical to accommodate various cell deployment scenarios.

C. Propagation model

Wireless channel is characterized by small and large-scale channel variations. The large-scale channel variation components are path loss, which is distance and frequency dependent for the most part while shadowing is terrain and environment dependent. In this paper for path loss we used the extended Hata model as in [10, 22] COST 231, considering the general propagation path loss in the environment as in equation (3)

$$PLe = 46.3 + 33.9 * \log(f) - 13.8 * \log(hb) - a(hm) + (44.9 - 6.55 * \log(hb)) * \log(d) + CFe \quad (3)$$

$_{PLe}$ is the path loss in dB, f is the frequency in MHz, h_b is the height of the base station in meters and h_m is the mobile height in meters while $_{CFe}$ is the correction factor which is 3dB for dense urban and zero everywhere else.

$_{a(hm)}$ is given by the following equation (4) according to [22]

$$a(hm) = (1.1 * \log(f) - 0.7) * h(m) - (1.56 * \log(f) - 0.8)dB \quad (4)$$

The second component of large scale is shadowing and the lognormal distribution model with 7dB variance. A Penetration loss of 12dB due to buildings and obstacles is considered.

For fast fading, since the imaginary and real parts of the signaling elements are independent and identically distributed (i.i.d), Rayleigh fading channel modeled as Gaussian random process is used with zero mean and variance σ =1/2 as illustrated in equation (5) below

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (5)$$

D. Services offered

i. Interactive Traffic using HTTP Web browsing (WB).

Packet switched services in general is of two flavors, one that does not establish a session before sending data which utilizes UDP protocol. Such that, no call establishment is required and by the size of the file or group of files associated with the webpage and the actual data rate in the cell at the time of download determines the length of "session". The second uses TCP protocol and requires a connection establishment before sending or receiving data. According to 3GPP standards, HTTP traffic is an interactive service of the first flavor. Before the wide spread of the internet and its applications, telecommunications traffic has been for a long time modeled as a Poisson process where the inter arrival and holding times of the session are exponentially distributed. This might be acceptable when only one type of calls, voice calls, is the service on the telecommunications systems. However, the internet traffic is more diverse in both the applications and the networks. Therefore, with the evolution of networks and applications in general; the inter arrival time of the Poisson process is still widely accepted while the holding time has evolved and more studies have proven that for data traffic the holding time follows long tail distributions for data traffic while exponential distribution is still acceptable for voice calls. From telecommunication empirical data analysis [14-19], the HTTP traffic (data traffic) follow a log normal distribution.

HTTP traffic has shown to be more complex than other protocols, as an interactive traffic with request response pattern characterized by low round trip, time delay as well as negligible content errors. The large majority of the page responses consist of small objects, which reference other traffic. The distribution of the main page size (SWB) is heavy-tailed lognormal distribution. Each web page consists of a number of embedded objects, embedded images, style sheets and executable java applets or plug-ins. There is also a parsing time and a reading time between pages.

Therefore, we assume that the length of a web session in bytes follow truncated lognormal distribution with mean of 8.35 and 1.37 variance as specified in the standards and literature [14, 16, and 23] for HTTP traffic and presented in the following probability distribution function

$$f(x) = \left(\frac{1}{\sigma x\sqrt{2\pi}}\right) \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) \quad (6)$$

The session size follow a lognormal distribution with minimum of 100 bytes and a maximum of 2Mbytes in the system. The Parsing time (TP) and the Reading time (TD) are both negative exponentially distributed with different intensities (Lambda). Thus, the session period for HTTP traffic based on the above will follow the equation below:

$$Twb = \frac{Swb * 8}{ABR} \quad (7)$$

Where WB is the average time unit required by the system to service the request, Swb is the session length in bytes; ABR is the available bit rate for the Web Session services at that instance of time. Since the number of embedded objects are following the Pareto distribution, but the size of the objects are following truncated lognormal distribution and to avoid modeling complexity; we are not considering any embedded objects associated within the Web sessions [16-17].

Based on the above assumptions, and the granularity and variability of web sessions, the average total service rate TWB as shown in Equation (4); is a mix between negative exponential distribution and a long tail lognormal distribution that depends on the session length in bytes. Both Reading time (T_B) and Parsing time (T_D) are negative exponentially distributed as given in equation (8), Therefore, the total holding time ($1/\mu_{WB}$) is $T_{WB} = T_{WB} + T_D + T_B$ (8).

Since the browsing is an interactive service, HTTP traffic is an ON-OFF traffic. However, in our model we are more concerned with the overall system performance rather than an individual modeling of the service. Therefore, we are assuming that the whole (ON-OFF) processes shall collapse to two states (arrival and departure).

ii. FTP traffic applications

An FTP session consists of a sequence of files separated by reading times. The FTP session parameters are the size of a file to be transferred following the lognormal distribution with σ and μ as specified in [16, 23]. The reading time is following the negative exponential distribution.

Table 2: FTP Parameters

Parameter	Distribution
File Size	Truncated lognormal $f(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp(-\frac{(\ln(x)-\mu)^2}{2\sigma^2})$ $\mu = 14.45$ and $\sigma = .35$ and $x > 0$. Mean = 2Mbytes, concatenated between minimum = 100bytes and Maximum = 5Mbytes
Reading time TD	Exponential distribution with $\lambda = 180$ s.

Therefore, the session period of FTP is given by $TS=\frac{SFTP*8}{ABR}$ while the total best effort holding time ($1/\mu_{FTP}$) session including reading time is given by

$$T_{FTP} = T_S + T_D \quad (9).$$

The holding time of the FTP session is a mix of negative exponential distribution and a long tail lognormal distribution. This is a more practical and realistic approach than what have been suggested in [1-2].

From mathematical point of view; when modeling the above two services; we can obtain the pdf function of the lognormal distribution function as stated above (equation 6 and table 2) when $\ln(X)$ has a normal distribution with $\sigma \in (0,\infty)$ and $\mu \in \Re$ for any random variable X . Therefore, the cumulative distribution function $F(x)=\Phi(\frac{\ln(x)-\mu}{\sigma})$ for $x\in(0,\infty)$ and the lognormal quantile function is given by $F^{-1}(p)=\exp\left[\mu+\sigma\Phi^{-1}(p)\right]$, for $p\in(0,1)$ [29].

Both services mentioned above are elastic services or none guaranteed bit rate services (N-GBR). The holding time depends on the availability of the resources and scheduling polices, particularly when the cell is fully loaded with services that have stringent delay requirements such as video conferencing and voice over IP services; such services is described as guaranteed bit rate services (GBR).

iii. Video conferencing traffic (VC)

VC has different requirements for video and audio traffic. Several studies have shown that for video traffic, the required compressed rate is between 32kbps and 1Mbps while for audio is between16-64kbps [17, 23]. However, a typical connection requires 384kbps for average session duration of 3600s. Based on this, we have elected to model this traffic based on 384kbps of bandwidth for holding time negative exponentially distributed random value of μ_{VC} = 1/3600s [23]. The instantaneous channel condition at the time of establishing the session influences the holding time of the service. The burst nature of the streaming traffic can impose a significant burden on the network from increased traffic congestion to an increased packet delays and losses. Therefore, we are more concerned in this model with the system level view or service view rather than individual packet view.

iv. VOIP Traffic as semi real time traffic

The traffic is modeled by a two state ON-OFF or ideal state and active state since the human nature in any conversation is to alternate between talking and listening. Internet protocols namely RTP/UDP/IP encapsulates the VOIP data traffic. The real time protocol (RTP) is an essential part of the system. Many encoding techniques such as adaptive multi rate audio (AMR) in the forms of G.711, G.722 and G729 among others are an examples of such codecs which results in different bandwidth requirements. The call holding time for VOIP is negative exponentially distributed random value of μ_{VOIP} =1/210s [23, 26]. The above traffic sources are deduced from the research literature and the standards as in [16-17, 23-24].

The distribution of the above mentioned services is following the traffic mix as described in [17], such that

$$\Phi(\text{WB}) + \Phi(\text{FTP}) + \Phi(\text{VC}) + \Phi(\text{VOIP}) = 1 \quad (10)$$

In LTE as it is the case with all 3GPP standards, services are mainly of two categories, guaranteed and non-guaranteed bit rates based on the established bearers for the service. Although the quality of service indicator (QCI) has been standardized [21] with scalar values to indicate the level of quality of service, it is a guideline for node pre-configurations. In a multi-vender environment, the QCI aims at ensuring that the minimum requirements of QoS.

In general, we can have any number of services in the system; however, increasing the number of services would raise the concern of practicality problems in solving the model because of the state space explosion problem.

Performance model

A. The Model Formulation

The model describes a single LTE cell serving different services. The model can accommodate multi cell scenarios where interference shall affect the system throughput and cell capacity. Let us denote M as the number of modulation states that are present in the system, where M = {0, 1... k}; such that, the number of WB sessions with a given modulation state is denoted by nWBk (t). In other words, at time t we have a number of WB sessions with modulation state k. let us keep in mind that the modulation state is influenced by the instantaneous channel conditions based on the propagation model at the time of admission.

In the same manner, nFTPk (t) is the number of FTP sessions in any modulation state k at a specific instance of time. Keeping with this methodology, we denote video conference customers' nVCk (t) in any modulation state k at a specific instance of time and the same goes to nVOIPk (t).

From the above, we can deduce that the total number of customers in the system as follows:

$$N=\sum_{0}^{n} WBK+\sum_{0}^{n} FTPK+\sum_{0}^{n} VCK+\sum_{0}^{n} VOIPK \quad (11)$$

In our model, we are assuming that the system service arrival is following a Poisson process, where the service arrivals are identical and independent. In [15, 18], the authors have deduced from empirical data that this holds true for the arrival process but is not necessarily true for the call durations (holding time). Holding time for the services is following negative exponential random values as explained

above. The reasoning behind this assumption is of two folds; first, two of the services (VC and VOIP) are modeled in the literature as negative exponential distributed random variables. Second, if we look at the FTP and HTTP sessions, we notice that the mix of the distribution is lognormal for the file size (i.e. number of bytes) and exponential distribution for reading and parsing time.

In our model, we are assuming an exponentially distributed inter arrival rate such that:

$$\lambda = \alpha 0\lambda WB + \alpha 1\lambda FTP + \alpha 2\lambda VC + \alpha 3\lambda VOIP \qquad (12)$$

$$\alpha 0 + \alpha 1 + \alpha 2 + \alpha 3 = 1 \qquad (13)$$

Which represents the total inter arrival intensity of the sessions. Since a Poisson process is split into n processes with probabilities α. Such processes are independent Poisson process with parameter αiλ, where i represent the probability of service occurrence.

The model transition is as depicted in fig. 1 below, therefore, and keeping with [1-2], the state space consists of the four services assumed in the model as individual Continuous Time Markov Chain (CTMC). The instantaneous channel conditions and path loss as articulated in section 3.3, influences the service data rate. The capacity of LTE cell is varying between the peak data rates (PDR) and the average bit rates (AVR) when the radio resources are fixed. Fig.1 illustrates the individual CMTC state transition diagram.

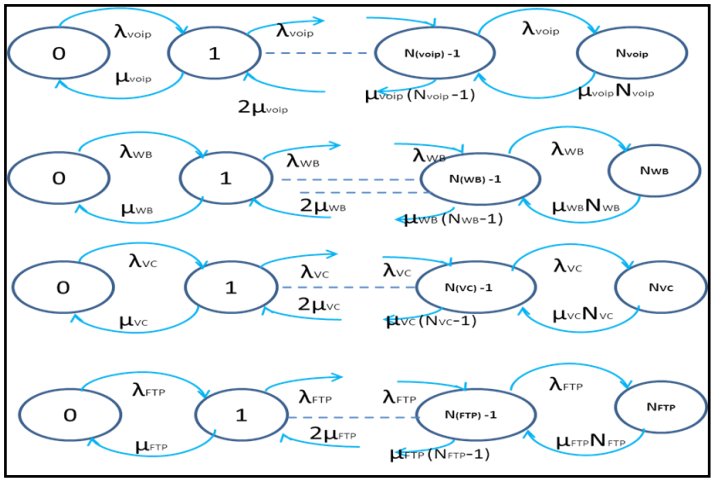


Figure 1: The individual state transition

Fig.2 illustrates the system state transition diagram, with initial state as all zero customers available in the system. With the intensity λ, the services start to arrive. However, inside the service itself, the MCS depends on the SINR value for the user based on equation (2), (3) and (4) above. This distinguishes our model in the sense that the system is not dividing the cell into zones of capacity in order to bring results that are more accurate to the system as each user is individually considered.

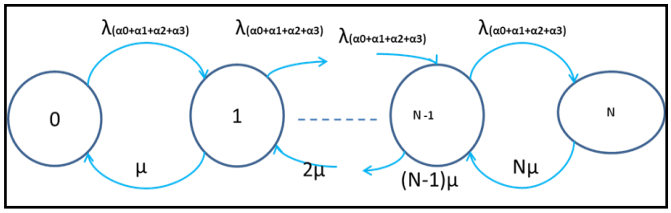


Figure 2: The system state transition

Fig.2 represents the service time μ as $\mu=\mu voip+\mu vc+\mu FTP+\mu WB$. With the traffic distributions as illustrated in the previous section, we should keep in mind two issues, the sojourn time for the two GBR traffic (i.e. VC and VOIP) having different μ while for the sojourn time for elastic services N-GBR traffic actually depends on the number of GBR customers in the system. However, the model as illustrated in figure 1 and 2 mimics the multi-dimensional continuous time Markov chain. Each state can be a mix of number of services where not only differs in the number of customers for each service but also differ in the modulation and coding rate inside the service itself. In other words, the state of the system can be any number of VOIP services with different MCS, plus any number of VC services with different MCS, plus any number of FTP services with different MCS and any number of WB services with different MCS.

Given the above system assumptions, the stochastic process can define the state at any instant of time as

$$X(t)=(\mathrm{nWBk}(t), \mathrm{nFTPk}(t), \mathrm{nVCk}(t), \mathrm{nVOIPk}(t)) \qquad (14),$$

where n=0,1,…,N.

Each state above shows that the model is irreducible and the state transition may occur at any arbitrary instant of time. Furthermore, it is a continuous time Markov chain for each service (even though the states move according to the under laying discrete events of customer arrival and departure) as it is in line with the Markov property that the future behavior of the model does not depend on the historical behavior but depends only on the current state. If we consider the states of the CTMC of the heterogeneous traffic as $X=(\mathrm{nWB, nFTP, nVC, nVOIP})$; then the next state X' can be either an arrival of any type of service and can be in any of the following state

$$X' = \begin{Bmatrix} (n+1)WB_k, nVOIP_k, nVC_k, nFTP_k \\ nWB_k, (n+1)VOIP_k, nVC_k, nFTP_k \\ nWB_k, nVOIP_k, (n+1)VC_k, nFTP_k \\ nWB_k, nVOIP_k, nVC_k, (n+1)FTP_k \end{Bmatrix} \qquad (15)$$

or in case of service completion of any type the next state can be in any of the following state:

$$X' = \begin{Bmatrix} (n-1)WB_k, nVOIP_k, nVC_k, nFTP_k \\ nWB_k, (n-1)VOIP_k, nVC_k, nFTP_k \\ nWB_k, nVOIP_k, (n-1)VC_k, nFTP_k \\ nWB_k, nVOIP_k, nVC_k, (n-1)FTP_k \end{Bmatrix} \qquad (16)$$

As such, the process is controlled by a multi-dimensional infinitesimal generator matrix , where i and j are two states of the process and the rules for constructing the rate transition matrix are given in Table 3 below. For any two states i and j, by setting the system number of states to any countable order K, the infinitesimal transition matrix of the system is the transition rate from state i to state j such that i,j = 0,….., K; where represents the probability that the chain will be in state j in some units of time given that it is in state i now.

Let P = {p_r (t)} (r = 0. . . K) denote the steady state probability distribution vector of the process X(t) at any time instant t. The stationary probability distribution vector (r = 0. . . K) where $\mathbf{P}=\lim_{t\to\infty} P(t)$ and $\sum_j pj=1$, can be the solution of a

well-known matrix equation , where is the vector of size K with all zero elements as articulated in [28].

Table 3: Transition Rates

Modulation mi, i = 0,…., M	**Current state**	**Next State**	**Condition**	**Transition Rate**
New Arrival of WB to mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	(n+1)WBmi, nFTPmi, nVCmi, nVOIPmi	$ABR \geq BR_{av\_wb}$	α0λWB
New Arrival of BE to mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	nWBmi, (n+1)FTPmi, nVCmi, nVOIPmi	$ABR \geq BR_{av\_be}$	α1λBE
New Arrival of VC to mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	nWBmi, nFTPmi, (n+1)VCmi, nVOIPmi	$ABR \geq BR_{av\_vc}$	α2λVC
New Arrival of VOIP to mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	nWBmi, nFTPmi, nVCmi, (n+1)VOIPmi	$ABR \geq BR_{av\_voip}$	α3λVOIP
Departure of WB from mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	(n-1)WBmi, nFTPmi, nVCmi, nVOIPmi	$nWB\_i \geq 1$	μ_{wb} $ABR_i / (n_{wb} \cdot n_{rb})$
Departure of BE from mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	nWBmi, (n-1)FTPmi, nVCmi, nVOIPmi	$nFTP\_i \geq 1$	μ_{be} $ABR_i / (n_{rb} + n_{be})$
Departure of VC from mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	nWBmi, nFTPmi, (n-1)VCmi, nVOIPmi	$nVC\_i \geq 1$	$\mu_{VCi} * n_{VCi}$
Departure of VOIP from mi	nWBmi, nFTPmi, nVCmi, nVOIPmi	nWBmi, nFTPmi, nVCmi, (n-1)VOIPmi	$nVOIP\_i \geq 1$	$\mu_{VOIPi} * n_{VOIPi}$

The intensity of call arrivals (λ) to the system is according equation 12 and 13. The services as presented above have different QoS. Evaluating the performance of a complex system such as LTE depends on many factors. SINR measurement is one of such factors but it is not the ultimate deciding factor as articulated in [30]. The type of services, the number of users the cell serve, the availability of resources to serve users and the channel conditions including fast fading can influence the performance of the system. In the above formulation; we have two main types of services; one with guaranteed QoS as in VOIP and video conferencing (GBR) and second type of customers with elastic QoS as in FTP and WB services (N-GBR).

Based on the above discussion and formulation, the performance measurements for LTE system is as follows. We are presenting the performance evaluation for two cells, which should not affect the general model.

The Simulation System

The simulator is a modular system based on the discrete event simulator as described in [26-27] with LTE networks extensions:

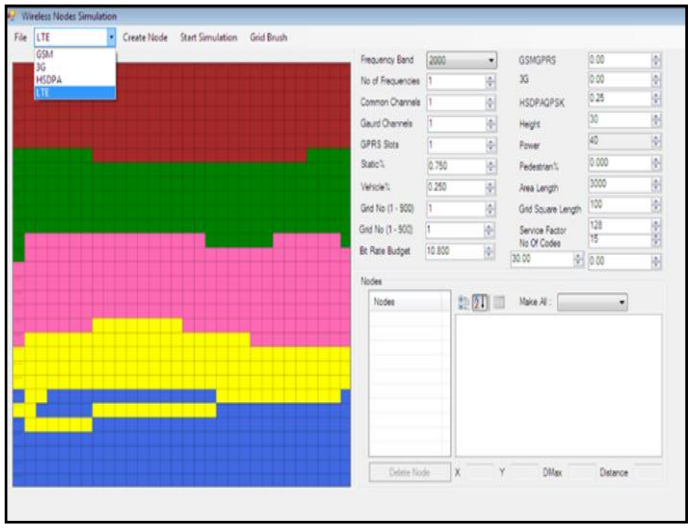


Figure 3: The Simulator

As illustrated in fig. 3, the system simulator comprises of 3GPP technologies including LTE networks. The system simulator is a modular system, which means that each component is a module by itself in order to ease the modification process. It consists of five different propagation environments, free space, rural, suburban, urban and dense urban. The deployment model for this study is portrait in the dense urban environment but can be evaluated at other environments as well because of the module structure of the system. The path loss module as described in the link module above is in accordance with [22]. The link level portion of the simulator is feeding the system level simulator with the required data such as the channel condition, user location and SINR measurements.

The link level parameters as in table 4 below:

Table 4: Simulation Pareameters

Parameter	Value
Propagation model	Extended Hata COS231 [22]
Antenna Technology	1x1 SISO
Frequency	2GHz
Thermal noise density	-174 dB/Hz
Receiver noise figure	7dB
Bandwidth	10 MHz, 50 PRB
Broadband Service	MIX see Table 5
Subcarrier spacing	15Khz
Shadowing Log-Normal distribution	7dB
Fast Fading	Gaussian distribution
Modulation and coding	SINR Measurement dependant
Propagation Environment	Dense Urban Macro
Cell Size	500 m Radius
Cell Power	43dB
Building Penetration Loss	12 dB

Several services have been realized fin our model; we elected to follow 3GPP standards as well as what has been documented in the literature as illustrated in table 5. The simulator has been designed to guarantee minimum bandwidth for each GBR service, such that if the minimum requested bandwidth is not available then the system will reject the service. In other words, there is no negotiation on the minimum requested bandwidth. However, there is also no guarantee for the subscriber data rate if the service is N-GBR Web browsing or FTP downloads.

Table 5: Services Assumptions

Service type	Main Services	Percentage of users[14]	Requested Bandwidth(minimum)	Distribution
Conversational VOIP	Guaranteed Bit Rate(GBR)	30%	25kbps	Negative exponential distribution 1/210 s [14, 16-17]
Video Conference		40%(video streaming and gaming)	384kbps	Negative exponential distribution 1/3600s [14, 16- 17]
FTP traffic (Best Effort)	Non- Guaranteed Bit Rate(N-GBR)	10%	variable kbps	Mixed [23-25]
HTTP (Interactive)		20%	variable kbps	Mixed [23-25]

The elastic traffic such as FTP and HTTP is of lower priority than the guaranteed bit rate services. As the number of VOIP and VC users increase in the system, the less allocated bandwidth for the elastic services. However, we are adhering to the inter arrival time Poisson process with intensities as described in table 5 and [14].

Simulation results

The simulation deployment scenario is as in figure 4 below.

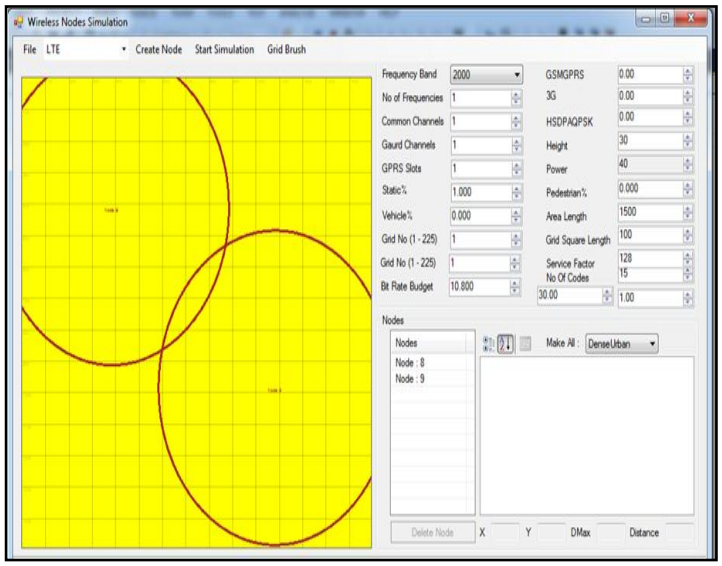


Figure 5: The deployment scenario

Fig.4 shows two LTE cells are deployment with overlapping coverage area and user static mobility. The inter site distance is 500m and power of 40W each. The deployment scenario is for dense urban area with service consideration is for users in the two cells coverage area. In the following, we analyze the performance of the system.

First, to validate the system we compare the analytical model with the simulation results by realizing the FTP services as in fig. 5,

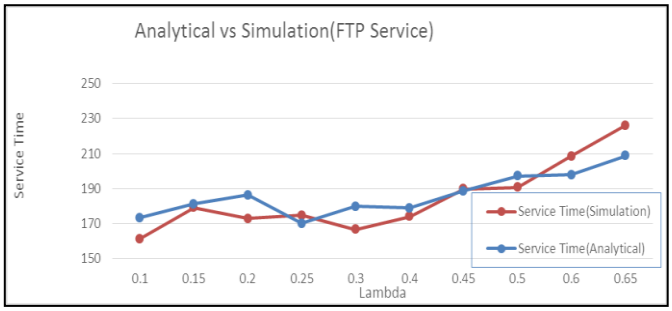


Figure 6: Inter arrival rate vs. Service completion time

Fig. 5 shows the average file download time comparison between the analytical model and simulations for various inter arrival rates. The results show that the model is not sensitive to different inter arrival rate. However, starting from traffic intensity of between .3 and .4, the download time starts to increase exponentially with the intensity of traffic. As the number of customers in the system, increase then the download time increases which degrades the quality of experience. The customer satisfaction will be at a minimum if the inter arrival rate is greater than 0.4.

Another way of validating the system performance by utilizing the intensity of the arrivals (λ) is in accordance with equations (12), (13) and table 5 without any restrictions on the elastic services data rate. The system is analyzed for the blocking probability considering the N-GBR (FTP and HTTP) services and GBR (VC and VOIP) services.

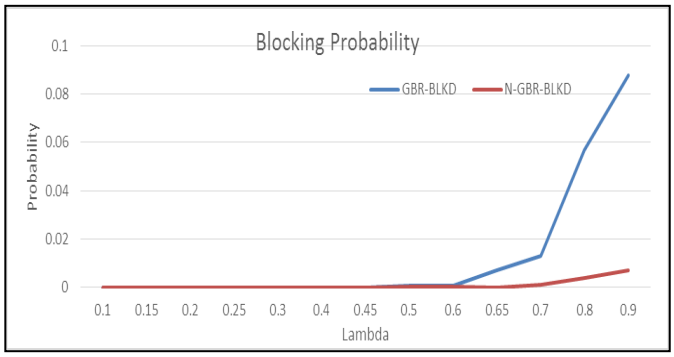


Figure 7: System Blocking Probability

The results as illustrated in figure 6 shows that it follows exponential distribution with blocking probability of 2% at the customer intensity of $\lambda = 0.7$ for the GBR services, the main reason for the blocking probability of the system at this rate is the demand for resources from video conference services. However, the VOIP service and the elastic services have a negligible blocking probability. To understand the system performance in a more detailed way, we need keep in mind that figure 6 reflects the intensity of arrivals as described in table 5 above. This measure might hold true for voice services but not necessarily true for elastic services. Therefore, we need a different measure to understand the behavior of the system, the measure we use is the service time (download time) that is required for the system to service the customer with a specific file size.

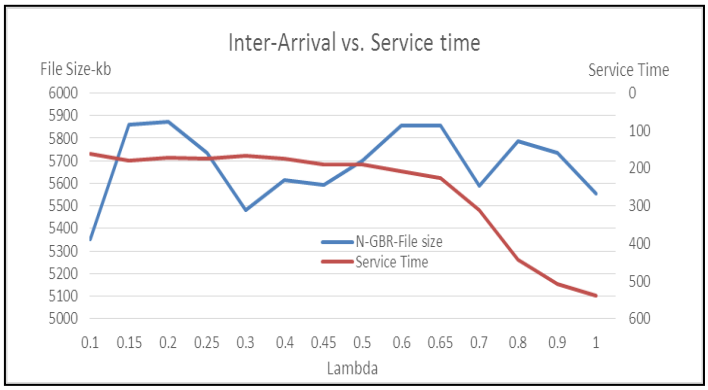


Figure 8: Elastic service time vs traffic intensity and Download time.

Fig. 7 shows that the elastic services file size is constant during the simulation period, which validate the simulation results. It is evident that the file size is averaging between 5.4 Mb to 5.6 Mb through the duration of the simulation; however, the service time is increasing from 100 unit time to 550 unit time or almost 4 folds. Figure 7 also confirms the fact that when the download file size is constant, the intensity of the traffic and the traffic mix in the system can affect the quality of experience, customer satisfaction and system performance and utilization. Such that, as the number of GBR service increase in the system, the N-GBR services starve for bandwidth. To prevent elastic services and applications from starvation the traffic intensity needs to be in the range of .3 and .4 and the system bandwidth needs to be split between elastic services and guaranteed services.

From the above we can conclude that even when we maximize the data rate or the capacity of LTE cells, the need still exists to protect elastic customers from bandwidth starvation. Therefore, the need to conduct another analysis for the system in order to strike a balance between the numbers of customers (services) the system can handle while preventing elastic services starvation. Therefore, we need to analyze the system performance when the data rate of the elastic services are constraint to a minimum of 200 kbps (including 30% of the data rate is for signaling) [10].

We start this analysis by measuring the service time (download time) of the elastic services against the inter arrival rate. Figure 8 shows the download time is almost constant with the inter arrival rate, this is due to bandwidth reservation for the elastic services. It is clear that when we reserve data rates for elastic services, the intensity of the traffic has little effect as we preserve the QoS for the elastic services and shall be steady throughout the deployment and optimization phase of the system. It also prevents the elastic services starvation for bandwidth, which in turn influence the customer satisfaction positively. In this way, we can guarantee the QoS and QoE for elastic services.

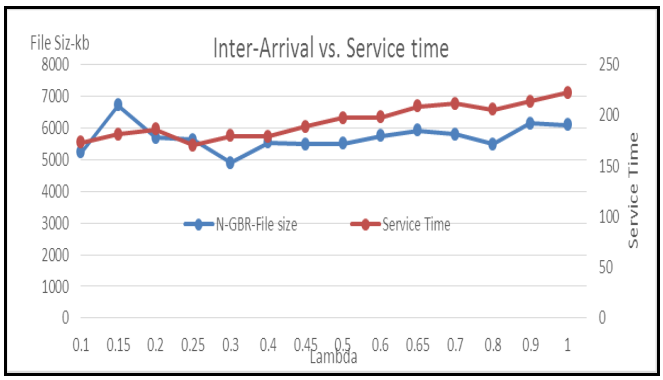


Figure 9: Inter-Arrival vs Service time and Elastic services file size

In a mixed services environment the quality of experience and customer satisfaction plays a major role when analyzing the system performance. It is also evident by comparing figure 7 and figure 8 that when there is no protection for the elastic services, the individual data rate can diminish completely which cause bandwidth starvation for the elastic services customers. If we to analyze the system performance we can see that the elastic services data rate is declining with the intensity of the users arriving to the system.

Figure 9, illustrates the reserved data rates in kbps for the GBR services and the elastic services with different inter arrival rates. The GBR represents the inclusion of the adaptive modulation and coding techniques employed in the simulation tool. Since the elastic services accounts for 30% of the services distribution, then it is expected to account for 30% of the total bandwidth, as such, the traffic intensity shall not exceed .3.

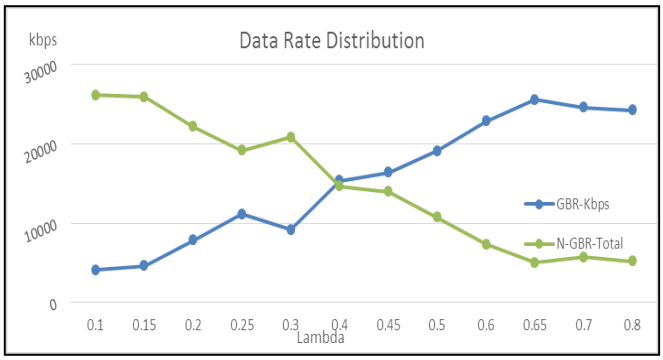
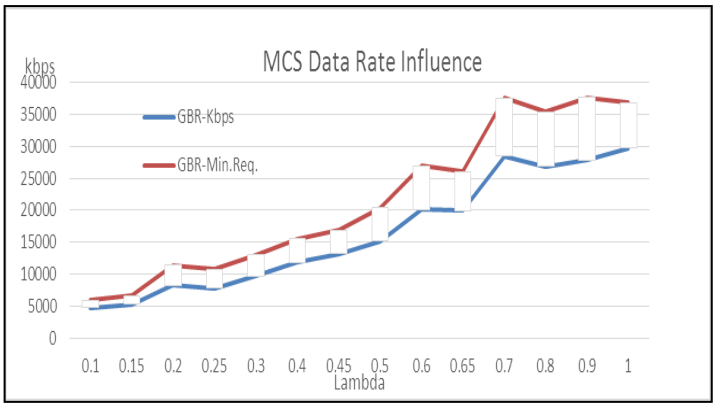


Figure 10: Data Rates in kbps for the mix of services

From the results, the individual data rate for the elastic services has been constraint to 200 kbps for the individual flow even with the influx of GBR services to the system. The total system data rate for the elastic services has been dropped to 5 Mbps. By examining table 5, in fact the reservation for the elastic services should be at 30% of the

total bandwidth. According to our performance study, the system should reserve 10Mbps for the elastic services that happens at traffic intensity of .3 with individual data rate of 200kbps for each customer.
Finally, the influence of modulation and coding scheme (MCS) as per table 1 has been assessed as per figure 10 below.



Figue 11: MCS Data Rate Influence

The utilization of different MCS rates can influence the user data rate based on path loss measurements. The results show that the system capacity and throughput can be enhanced by 2Mbps to 7Mbps especially when the system is at full load and that this due to user diversity gain.

Conclusion

In this paper, we have presented some challenges in the path of modeling LTE systems. We sought to address few challenges in the face of estimating the capacity of LTE networks in a mix traffic. We have identified the main problems that need to be addressed in such complex system. We developed a practical mathematical model for LTE network for performance analysis by employing Continuous Time Markov Chain (CTMC) that includes GBR services and N-GBR services. We compare the analytical model with the simulation tool, which prove that the model holds accurate. We choose the traffic modeling to be in accordance of the standards and what is reported in the literature for the traffic mix. The LTE cells bandwidth needs to be split by the different services as in table 5. However, the GBR services are greedy services and can consume the total bandwidth which might lead to unsatisfied customers. Therefore, to prevent elastic services from bandwidth starvation, the traffic intensity or inter arrival rate shall be in the range of .3 with certain amount of resources reserved for elastic services to preserve QoS, QoE and customer satisfaction. Furthermore, the need arises to design a resource reservation algorithm to split the bandwidth between the services. Therefore, in the future we need to design a resource reservation algorithm for bandwidth usage and investigate the various MIMO techniques impact on the performance of LTE networks from a customer satisfaction point of view of the offered services.

Acknowledgement

The author gratefully acknowledges the support for this work by the deanship of research under grant number s-0063-1436, University of Tabuk, Saudi Arabia.

References

[1] Y.Zaki et all, LTE Radio Schedulers Analytical Modeling using Continuous Time Markov Chains, WMNC 2013.
[2] S. Doirieux and B. Baynat and M. Maqbool and M. Coupechoux, An efficient analytical model for the dimensioning of WiMAX networks supporting multi-profile best effort traffic, Computer Communications, vol. 33, no. 10, pp. 1162-1179, 2010.
[3] 3GPP TS 22.105, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Services and service capabilities, V11.1.0 (2013-12).
[4] 3GPP TR 25.814, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Physical layer aspects for evolved Universal Terrestrial Radio Access (E-UTRA) (Release 7) V7.1.0 (2006-09)
[5] Fredrik Athley and Martin N. Johansson, Impact of Electrical and Mechanical Antenna Tilt on LTE Downlink System Performance. Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st.
[6] F. Capozzi, G. Piro, L.A. Grieco, G. Boggia, and P. Camarda, Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey. IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 15, NO. 2, SECOND QUARTER 2013.
[7] QiMei Cui, XueQing Huang, Bing Luo, XiaoFeng Tao, Jun Jiang, Capacity analysis and optimal power allocation for coordinated transmission in MIMO-OFDM systems. Science China Information Sciences June 2012, Volume 55, Issue 6, pp 1372-1387.
[8] Iana Siomina, Di Yuan, Analysis of Cell Load Coupling for LTE Network Planning and Optimization. IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 11, NO. 6, JUNE 2012.
[9] Osterbo, O "Scheduling and Capacity Estimation in LTE," Proc. IEEE ITC, 2011, pp. 63-70.
[10] S. Sesia, I. Toufik, and M. Baker, "Lte–the umts long term evolution," From Theory to Practice, published in, vol. 66, 2009.
[11] J. Lee, J.-K. Han, and J. Zhang "MIMO Technologies in 3GPP LTE and LTE-Advanced", EURASIP Journal on Wireless Communications and Networking, vol. 2009, pp.21
[12] http://lteworld.org/whitepaper/seven-modes-mimo-lte "The Seven Modes of MIMO in LTE"
[13] Abuhaija, B., "Performance analysis of LTE multiuser flat downlink power spectrum and radio resources scheduling", Journal of High Speed Networks, vol.18, no.3,pp: 173–184, 2012.
[14] 3GPP TR 37.852, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; RAN enhancements for UMTS/HSPA and LTE interworking (Release 12).

[15] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding traffic dynamics in cellular data networks. In INFOCOM'11, April 2011.
[16] 3GPP2-C50-EVAL 2001 022-0, "HTTP and FTP Traffic Models for 1xEV-DV Simulations", 2001.
[17] www.ngmn.org "Next Generation Mobile Networks Radio Access Performance Evaluation Methodology".
[18] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz. Primary user behavior in cellular networks and implications for dynamic spectrum access. IEEE Communications Magazine, 47(3):88-95, Mar. 2009.
[19] S. S. Rappaport, "Communications Traffic Performance for Cellular Systems with Mixed Platform Types" in Wireless Communications: Future Directions, Kluwer Academic Publishers, 1993.
[20] Bui, Nicola and Michelinakis, Foivos and Widmer, Joerg," A Model for Throughput Prediction for Mobile Users" In: European Wireless 2014.
[21] 3GPP TS 36.213, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 12).
[22] Motorola LTE planning guige at www.telecomsource.net - MOTOROLA-LTE-Rf-Pln-Guide.pdf.
[23] http://ieee802.org/16/tgm/docs/80216m-08_004r2.pdf.
[24] https://ortus.rtu.lv/science/en/publications/8781-Research.
[25] http://www.ieee802.org/16/tgm/contrib/C80216m-07_067.pdf.
[26] B. Abuhaija and K. Al-Begain. "Designing of internetworking simulator to enhance GSM/WCDMA internetworking". 24 June 2008, IEEE,9TH Annual conference, ISBN: 978-1-902560-19-9, Liverpool John Moores University, UK.
[27] B. Abuhaija and K. Al-Begain. "Enhanced Common Radio Resources Managements Algorithm in Heterogeneous Cellular Networks". IEEE. Next Generation Mobile Applications, Services and Technologies, International Conference on, Cardiff, UK. Pages 335-342, 2009.
[28] G. Bolch, S. Greiner, H. de Meer and K. Trivedi, "Queueing Networks and Markov Chains", John Wiley and Sons Inc, 1998.
[29] Bain, Lee J and Englehardt, Max. "Introduction to Probability and Mathematical Statistics (2nd edition)". PWS-Kent Publishing Company, 1992.
[30] J. Andrews, et all "An Overview Of Load Balancing in HetNet: Old Myths And Open Problems". IEEE Wireless Communications, April 2014.
[31] B. Abuhaija, "Uplink-Downlink LTE Multi Cell Capacity: A Performance Analysis in the Presence of ICI, Imperfect Channel Information and Reuse-1 Plan". International Journal of Computer Applications (0975 – 8887) Volume 89 – No 12, March 2014.
[32] W. MANSOURI,et.all "Capacity Analysis in the Next Generation Wireless Networks Based on SINR Constraints". Journal of Networks, Vol 10, No 01 (2015), 29-38, Feb 2015, doi:10.4304/jnw.10.01.29-38.