STUDYING PERFORMANCE OF LTE NETWORKS WITH QUEUING SYSTEMS

A Project Report Submitted for the Course

MA498 Project I

by

Prem Sujan Kotta (170123027)

and

Garvit Mehta (170123018)



to the

DEPARTMENT OF MATHEMATICS INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI GUWAHATI - 781039, INDIA

November 2020

CERTIFICATE

This is to certify that the work contained in this project report entitled

"Studying performance of LTE networks with Queuing systems" submitted

by Prem Sujan Kotta (170123027) and Garvit Mehta (170123018) to the De-

partment of Mathematics, Indian Institute of Technology Guwahati towards

partial requirement of Bachelor of Technology in Mathematics and Comput-

ing has been carried out by them under my supervision.

It is also certified that this report is a survey work based on the references

in the bibliography.

Turnitin Similarity: 21 %

Guwahati - 781 039

November 2020

(Prof. Selvaraju N.)

Project Supervisor

ii

ABSTRACT

The main aim of the project is to study various Queuing Models in the context of LTE Networks and establish various performance metrics for LTE Networks. We will study NBurst/M/1 queuing model which has a close resemblance to practical LTE networks. First, we establish the results of M/M/1 model for the LTE network and further we simulate the actual network traffic using NBurst/M/1 model to confirm them.

Contents

List of Figures						
1	Introduction					
	1.1	LTE s	ystems	1		
	1.2	Queui	ng systems	2		
	1.3	Impor	tance of queuing theory in LTE systems	2		
	1.4	Queui	ng models	4		
		1.4.1	Simple Markovian queues	4		
2	N-E	Burst N	Model for LTE Networks	9		
	2.1	Some	Key Words	9		
		2.1.1	Telecommunication Network	9		
		2.1.2	Self Similarity	10		
		2.1.3	Reliability Function	10		
		2.1.4	Bursts	10		
		2.1.5	Packet Delay	10		
		2.1.6	Buffer Overflow	10		
		2.1.7	Semi Markov process	11		
		2.1.8	Power tail distribution	11		
	2.2	LTE N	Network model	11		

Bi	Bibliography							
3	Fut	ure Work	20					
	2.6	LTE Performance Metrics	18					
	2.5	Limiting cases for $b \to 0$ and $b \to 1 \dots \dots \dots$	16					
	2.4	Distribution of Sub Processes	15					
	2.3	Description of Model	13					
		2.2.1 Requirements for an appropriate network model	13					

List of Figures

1.1	High-level architecture of LTE systems	2
1.2	Figure showing the transitions in birth-death process	3
1.3	M/M/1 queue	5
2.1	NBurst/M/1 queue	12
2.2	LTE network topology	12
2.3	On-OFF queue model	13

Chapter 1

Introduction

1.1 LTE systems

LTE stands for Long-Term Evolution and it is a standard wireless data communication technology for mobile devices and data terminals. The main aim of LTE is to facilitate low response time and high throughput supporting flexible bandwidth deployments.

The main components of LTE architecture are as follows:

UE: User Equipment represent the terminal devices which deal with modules like data communication and terminating data channels.

E-UTRAN: Evolved UMTS Terrestrial Radio Access Network are evolved base stations(eNB) which send and receive radio transmissions from the user equipment using the digital and analogue signal processing.

EPC: Evolved Packet Core deals with network management and gateway servicing modules.

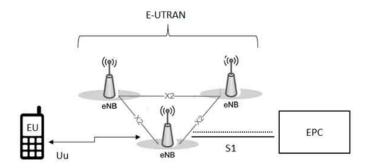


Figure 1.1: High-level architecture of LTE systems

1.2 Queuing systems

Queuing systems deal with the mathematical study of congestion. Generally, the study of servicing requests from customers/entities arriving in a queue fashion at a facility is called queuing theory. Examples: Ticket booking, customer service centers.

1.3 Importance of queuing theory in LTE systems

With growing modern internet applications and mobile devices, there is a spike in the usage of cellular networks. So it has become very important to investigate and analyze the performance of LTE networks.

Here eNB acts as a server and requests from wireless devices arrive at eNB. So we can treat the LTE network as a queuing system. The terminal wireless devices use protocols like up-link and down-link for data transfer in LTE networks.

Some important metrics of LTE networks such as waiting delay and block probability are analyzed using queuing theory.

Thus it is very important to study Queuing systems to understand LTE systems and how applications with different burst patterns effect LTE network performance.

Definition 1.3.1. A process is memory-less if the system forgets the state constantly i.e. probabilities are not influenced by the history of the process. A random variable G is memory-less if the distribution satisfies the property:

$$P(G > a + b|G > a) = P(G > b)$$

This property is also referred to as Markov property.

Definition 1.3.2. A stochastic counting process S(t) with rate $\lambda > 0$ is a Poisson process if it satisfies with the following properties:

- a. S(0) = 0.
- b. S(t) has independent increments
- c. The count of arrivals in any time span of length t > 0 follows the distribution $Poisson(\lambda t)$.

Definition 1.3.3. Birth death Process: When a service request arrives at a queue system, it is allocated a resource and eventually, the customer leaves the system after the service. This is called a birth-death process. Each request arrival is considered as birth and each served customer is considered as death.

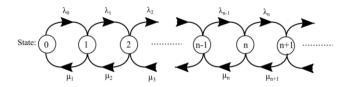


Figure 1.2: Figure showing the transitions in birth-death process

The arrival rate of customers (λ) and service rate (μ) are constant or state-dependent. Mathematical models of queuing systems are studied using the birth-death framework.

Definition 1.3.4. Little's Law: The fundamental relationship between the three parameters λ (arrival rate), L(length of the system), W(waiting time in the system)

$$L = \lambda W$$

1.4 Queuing models

Queuing systems are mainly characterized by the arrival of requests, the service mechanism, and queuing discipline

Definition 1.4.1. Kendall notation is a standard system used in queuing systems to describe a queuing model with three factors A/S/c. Sometimes also referred to as A/S/c/K/N/D.

A: Arrival process.

S: Servicing process.

c: Number of service channels.

K: Length of the queue.

N: Customer population

D: Service discipline

1.4.1 Simple Markovian queues

Single server Queues:

M/M/1 model: M/M/1 is a simple markovian birth-death process consisting a single server. Arrival and service rates are state independent. Inter-

arrival time and service times for $\mathrm{M}/\mathrm{M}/1$ queues follow exponential distributions.

$$d_{arr}(t) = \lambda e^{-\lambda t}$$

$$d_{ser}(t) = \mu e^{-\mu t}$$

 λ : The rate of arrival.

 μ : The rate of service.

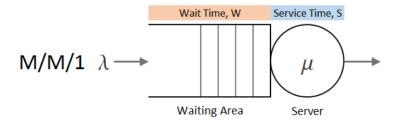


Figure 1.3: M/M/1 queue

Theorem 1.4.2. An exponential random variable S has the memory-less property.

Proof. From Bayes theorem,

$$P(S > a + b | S > b) = \frac{P(S > a + b, S > b)}{P(S > b)} = \frac{P(S > a + b)}{P(S > b)} = \frac{e^{-\lambda(a + b)}}{e^{-\lambda b}} = e^{-\lambda a}.$$

$$P(S > a + b|S > b) = P(S > a)$$

Implies exponential random variable is memoryless

Let P_n be the probability for the system in long-term to stay in state n as shown in Figure 1.2.

For a small period of time h, The steady state probability of the system to stay in state n at time t+h (infinitesimally small h) from the birth-death state figure 1.2 gives 3 cases:

- From state n-1 with one request and no service
- From state n+1 with no request and one service
- From state n with no request and no service

Mathematically written,

$$P_n(t+h) = P_{n-1}(t)\lambda h(1-\mu h) + P_{n+1}(t)(1-\lambda h)\mu h + P_n(t)(1-\lambda h)(1-\mu h)$$

For steady state,

$$\frac{P_n(t+h) - P_n(t)}{h} = 0$$

$$P_{n-1}(t)\lambda + P_{n+1}(t)\mu - P_{n-1}(t)(\lambda + \mu) = 0$$

$$(\lambda + \mu)P_n = P_{n+1}\mu + P_{n-1}\lambda(n \ge 1)$$
(1)

$$P_0(t+h) = P_1(t)(1-\lambda h)\mu h + P_0(t)(1-\lambda h)$$

$$\mu P_0 = \lambda P_1 \tag{2}$$

From (1) and (2),

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

Sum of probabilities of steady states

$$P_0 + P_1 + P_2$$
 ... $P_{\infty} = 1$

$$P_0(1 + (\lambda/\mu) + (\lambda/\mu)^2 + \dots + (\lambda/\mu)^{\infty}) = 1$$

An important parameter of the queueing system is **traffic rate** $\rho = \frac{\lambda}{\mu}$

$$P_0 (1 + \rho + \rho^2 + \dots + \rho^{\infty}) = 1$$

$$P_0\left(\frac{1}{1-\rho}\right) = 1$$

$$P_0 = 1 - \rho$$

$$P_n = (1-\rho)^n \rho$$

To analyze a queuing model we study the following parameters

Average number of entities in the system (L_s)

Average number of entities waiting in the queue (L_q)

Average time spent by an entity in the system (W_s)

Average time spent by an entity waiting in the queue (W_q)

 L_s is the average number of customers in the system either waiting in the queue or being serviced.

$$L_{s} = \sum_{i=0}^{\infty} iP_{i}$$

$$= \sum_{i=0}^{\infty} i\rho^{i}(1-\rho)$$

$$= \sum_{i=0}^{\infty} i(\rho^{i} - \rho^{i+1})$$

$$= 1(\rho - \rho^{2}) + 2(\rho^{2} - \rho^{3}) + \dots$$

$$= \rho + \rho^{2} + \rho^{3} + \dots$$

$$= \rho(1+\rho+\rho^{2}+\rho^{3}+\dots)$$

$$= \frac{\rho}{1-\rho}$$

$$\implies L_s = \frac{\lambda}{\mu - \lambda} \tag{1.1}$$

 \mathcal{L}_q is the average number of customers in the system waiting in the queue

$$L_{q} = \sum_{i=1}^{\infty} (i-1)P_{i}$$

$$= \sum_{i=0}^{\infty} iP_{n} - \sum_{i=1}^{\infty} P_{n}$$

$$= L_{s} - (1 - P_{0})$$

$$= \frac{\rho}{1 - \rho} - (1 - (1 - \rho)) = \frac{\rho}{1 - \rho} - \rho$$

$$= \frac{\rho^{2}}{1 - \rho}$$

$$\implies L_{q} = L_{s}\rho \qquad (1.2)$$

 W_s is the average time spent by a customer in the queue system waiting or being serviced.

From Little's law can be written as

$$W_s = \frac{L_s}{\lambda} = \left(\frac{\lambda}{\mu - \lambda}\right) \frac{1}{\lambda} = \frac{1}{\mu - \lambda} \tag{1.3}$$

 W_q is the average waiting time spent by a customer in the queue.

From Little's law can be written as

$$W_q = \frac{L_q}{\lambda} = \left(\frac{\rho}{1-\rho}\right)\frac{\rho}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{1.4}$$

From the above equations,

$$W_q = W_s \rho \tag{1.5}$$

Now, we study queuing systems for LTE networks.

Chapter 2

N-Burst Model for LTE

Networks

After many years of research it has become clear that usual models of queuing theory are not sufficient for the modelling many telecommunication services. Maybe one of the reasons for this is that data is not transmitted continuously rather it is transmitted in bursts of packets. But this is not at all enough for explaining the unpredictable performance of routers. This traffic must also be self similar along with bursty. The self-similar property can be explained and modeled by using power-tailed distributions of burst-sizes.

2.1 Some Key Words

2.1.1 Telecommunication Network

Group of nodes which are joined by links which are used to exchange messages between them. To exchange messages, these links may use packet switching, circuit switching, and many other methods.

2.1.2 Self Similarity

The bursty traffic can be described using the notion of Self-Similarity. This property is often used to describe the infinite complex patterns. However in our case ,this term is used to describe the distribution of object i.e. The distribution of object remains similar even if it is viewed at varying scales.

2.1.3 Reliability Function

It is also called Complementary cumulative distribution function. If we have a non negative variable Z having mean E(Z). Then it can be described as

$$R_Z(x) = Pr(Z > x) \tag{2.1}$$

2.1.4 Bursts

Burst is a group of consecutive packets which have shorter inter-packet gaps than the packets which arrive after or before the burst of packets.

2.1.5 Packet Delay

Time taken by packets to move from source to destination, or rather from one source to another.

2.1.6 Buffer Overflow

When unexpected number of entities arrive at the router the requests are lost, this is called buffer overflow.

2.1.7 Semi Markov process

If we have a process that can be in N states 1,2...N and the system stays in i^{th} state for a random interval of time having mean μ_i and makes a transition to state j with a probability P_{ij} then such a process is called semi-Markov process. Also abbreviated as SM.

2.1.8 Power tail distribution

Power-Tail distributions with exponent β follow the properties

$$E[X^{l}] = \int_{0}^{\infty} x^{l} f(x) dx \begin{cases} = \infty & \text{for } l \ge \beta \\ < \infty & \text{for } l < \beta \end{cases}$$

Similarly,

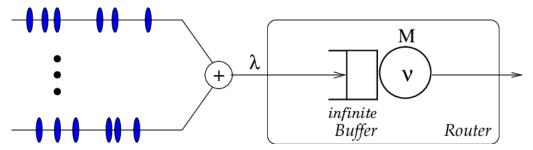
$$\lim_{x \to \infty} x^{l} R(x) = \begin{cases} = \infty \text{ for } l > \beta \\ = 0 \text{ for } l < \beta \end{cases}$$

Density function c/x^{β} is a simple form having the above property.

Truncated Power-Tail distribution is when $x^{\beta}R(x)$ remains almost a constant for large order of magnitudes and drops to 0 eventually

2.2 LTE Network model

A simple LTE network consists of a evolved nodeB, which provides service to a set of N wireless devices. These devices use the LTE technology for downloading data (down-link) and upload and request data (up-link). The mobile devices use several applications to access information using down-link. This



In this figure the packet generated by each user arrive at a single queue which is maintained by a single router or server

Figure 2.1: NBurst/M/1 queue

results in variable traffic for requests in the LTE system. Down-link communication will be our main focus.

The number of packets requested by a mobile device is considered as a random variable. The properties and distribution of this random variable is purely based on the mobile application used to request the data. Which is why standard queuing models are not appropriate to understand telecommunication systems.

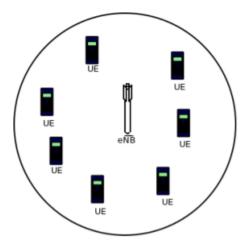


Figure 2.2: LTE network topology

2.2.1 Requirements for an appropriate network model

The main difference from the standard models is the requests in telecommunication networks come in bursts rather than continuously. The unpredictable traffic must be both bursty and self-similar. The self similarity can be achieved if the burst sizes are made to follow power-tailed distribution which makes bursts differ in orders of magnitude.

A model settling for self similarity and burstiness for analysis purpose is not easy. Many models like M/G/1 queue with changing parameters, batch-arrival model, continuous burst flow models failed to mimic the traffic in telecommunication networks. A novel queue model N-burst/M/1 is introduced for this purpose. Independent entities each send forth bursts of packets to a queue on an ON-OFF fashion(bursty model).

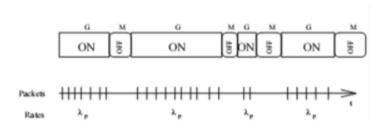


Figure 2.3: On-OFF queue model

2.3 Description of Model

N Burst Model is one of the variants of ON-OFF models. The arrival process of N-Burst is the superposition of N ON-OFF type identical and independently distributed source's traffic streams. Let (λ_p) be the peak transmission rate during a burst i.e. at ON time. No packets are transmitted during OFF period.

Given below are the essential parameters of the model:-

 κ : average rate of packet arrival for each source.

 λ : The overall arrival rate that is produced by N sources i.e. $\lambda = \kappa N$

 N_b : Average number of packets during a burst (at ON period)

 λ_p : The peak transmission rate of a node sending packets during a burst.

 λ_b : Mean burst arrival rate = λ/N_b

 \overline{ON} : the average time during which the node is active

 \overline{OFF} : the average time during which the node transmission is OFF.

 μ : the mean service rate of the router

 $\rho = \frac{\lambda}{\mu}$: Load utilisation of the router

We have another parameter which is of prime importance called burstiness parameter b defined below:

$$b = \frac{\overline{OFF}}{\overline{ON} + \overline{OFF}} = 1 - \frac{\kappa}{\lambda_p} = \frac{\lambda}{N\lambda_p}$$
 (2.2)

Since ρ (the amount of data sent per unit time), and ρ can be held constant as b is varied over 0 to 1. So the parameter b can be considered as a shape parameter.

Case 1: b=0

When b=0, it means that average time during which the node transmission is OFF is 0. In this case all the bursts come after one another and ON/OFF process reduces to a simple one.

Case 2: b=1

In this case, the average time during which burst comes is zero. It also means that all the packets in the burst arrive simultaneously i.e. They arrive as a bulk process.

All other cases of b are not simple and require a careful analysis.

2.4 Distribution of Sub Processes

For simplicity let us consider N=1 i.e 1-Burst model.

This model depends on four separate distributions, with random variables denoted by X_{ON} , X_{OFF} , X_{SV} and X_{INP} .

ON: On time distribution have mean \overline{ON} . Therefore the mean number of packets during a burst is given by $N_b = \overline{ON} \cdot \lambda_p$

OFF: **OFF** time distribution have mean \overline{OFF} (It will depend on how often the bursts are generated)

SV: Service time distribution has mean $1/\mu$ (It will depend on distribution of packet sizes, service rate μ depends on packet size and router speed) INP: Inter packet time distribution is the distribution during a burst whose mean is $1/\lambda_p$

The parameters to be varied must have to be selected carefully in order to have comparison study of the model useful. We can have two distributions whose means are different but have same shape, whose higher moments will be scaled proportionately. The description of this can be given as:

Let X and E(X) be the random variable and its mean respectively, then its Complementary Cumulative Distributive Function be given by

$$R_X(x) = Probability(X > x)$$
 (2.3)

Let Z is the non negative random variable with mean E(Z) and let its distribution has same shape as $R_X(.)$. Then,

$$R_Z(x) = R_X(rx)$$
, where $r = E(X)/E(Z)$, for

$$E(Z) = \int_0^\infty R_Z(x) dx$$

$$= \int_0^\infty R_X(rx) dx$$

$$= \int_0^\infty R_Z(u) du/r$$

$$= E(X)/r$$

$$= E(Z)$$

Remark 2.4.1. Moments of functions which have same shape scale according to the below formula:

$$E(Z^n) = r^n E(X^n) \tag{2.4}$$

or

$$\frac{E(Z^n)}{[E(Z)]^n} = \frac{E(X^n)}{[E(X)]^n}$$
 (2.5)

Using this we can vary the distributions of all the above methods and keeping their shapes same.

2.5 Limiting cases for $b \to 0$ and $b \to 1$

We consider this for 1-Burst process i.e. for N=1.

From (2.2) we can conclude that the transmission of packets can occur at the slowest rate at \overline{ON} times when $\overline{OFF} = 0$ or thus b=0 ($\lambda_p = \kappa$). In

this case the distribution of \overline{OFF} time is irrelevant.

It can be said that since there is no halt in between two bursts, \overline{ON} distribution also does not have an impact on the system.

Remark 2.5.1. Let G_{INP} represents a general distribution which have Inter arrival times as governed by X_{INP} . Interpret the similar symbols accordingly.SM denotes Semi Markov.

Thus for $b \to 0$ the 1-Burst process reduces from $SM/G_{SV}/1$ queue (X_{OFF}, X_{INP}) and X_{ON} are the distributions on which SM depends) to a $G_{INP}/G_{SV}/1$ queue.

On the other hand if we increase λ_p indefinitely, then according to (2.2), we get $b \to 1$. This is the case when \overline{ON} tends to 0. It also means that all the packets in the burst arrive almost simultaneously i.e. They arrive as a bulk process The bulk size is distributed according to ON-time distribution whose mean is N_b . The time in between two bursts is distributed according to \overline{OFF} . Now the 1-Burst process reduces to a $G_{(OFF)}^{[ON]}/G_{SV}/1$ queue.

Remark 2.5.2. The above notation of $G_{(OFF)}^{[ON]}$ represents the bulk arrival process whose Inter Arrival times X_{OFF} and bulk sizes proportional to X_{ON} .

Our base model should be simple, so we take X_{OFF} , X_{SV} and X_{INP} as exponential distributions. The ON distribution will be varied to obtain different results. Our model have parameters from each of these models. For these cases,

 $b \to 0$ the system becomes an $M_{\lambda}/M_{\mu}/1$ queue

 $b \to 1$ the system becomes an $M_{\lambda_b}^{[(ON)]}/M_{\mu}/1$ queue

Remark 2.5.3. M_{λ} stands for exponential inter arrival times having rate λ .

2.6 LTE Performance Metrics

Remark 2.6.1. Although, the analytic N-Burst model can be evaluated for all possible distributions, the calculations are not trivial. But it is possible to gain insight by looking at the limiting cases of b=0 and b=1. In the description below, we will hold ρ constant, while varying b for various ON time distributions. That is the average load is held constant, while the packets in a burst are bunched up or spread out as much as possible according to the value of b.

There are many modes to use this model. First one being when the idle time approaches zero which leads to continuous flow (This is described by b=0 and no bursts). So, in this case the model can be reduced to Poisson arrival which can also be represented as $(M_{\lambda}/M_{\mu}/1$ queuing model) and thus the mean delay

$$mdb0 = \frac{1/\mu}{1-\rho} \tag{2.6}$$

where ρ is as described above (as given in [1]).

The other mode of use is when the active time is reduced to almost 0 (This is described by b=1). Now in this case, there is a simultaneous arrival of all packets and thus mean delay can be written as

$$mdb1 = C.\left(\frac{1/\mu}{1-\rho}\right) \tag{2.7}$$

(as given in [1]), where $C = \frac{E(N(N+1)/2)}{E(N)}$ random variable N represent the count of packets in a single burst.

Remark 2.6.2. :b=1 is the bulk arrival process given by queue $M_{\lambda_b}^{[(ON)]}/M_{\mu}/1$. This behaviour is well known and can give the above result (2.7).

If we assume that load utilisation ρ of router is constant, then the best performance of nodes occurs in the case when b=0 and the worst performance for b=1. The size of the burst depends on the burstiness parameter b (each b value represents a unique type of traffic). Now we will write without proof that the queuing model can be represented as matrix exponentiation approach and steady state solution for the system can given as in [1]: The end to end packet delay using little's formula can be given as:

$$\overline{DEL} = \frac{1}{\kappa} R(I - R)^{-1} \overline{\epsilon}$$
 (2.8)

Block probability is given by:

$$Pr(Block) = \frac{1}{\kappa} \pi(R^B L) \overline{\epsilon}$$
 (2.9)

where R can be calculated by solving system as Quasi-Birth-Death process,I is the identity matrix, κ is the average arrival rate, and ϵ is the unity vector.

Remark 2.6.3. However, we have taken these results directly from [1], we will derive them analytically using simulations.

Chapter 3

Future Work

In the future we will focus on two tasks mainly:-

- 1) We will be simulating the M/M/1 queuing model in the context of simple LTE Networks. For this we will be using NS3 simulator which is a discrete-event network simulator for Internet systems. We will check various performance metrics such as packet loss, delay, throughput etc of a particular network topology and confirm if they are indeed following the above model characteristics.
- 2) We will take a more appropriate model called NBurst/M/1 model on a different LTE network topology. Firstly, We will define an appropriate network topology and then an ON time distribution X_{ON} , and simulate the network to check the performance metrics analytically. We will again use NS3 simulator for this taking appropriate assumptions wherever necessary to check the Network 's characteristic properties such as throughput, packet loss, packet delay, block probability, etc.

Bibliography

- [1] Adel Agamy and Ahmed Mohamed. Influence of various application types on the performance of lte mobile networks. *International Journal of Electrical and Computer Engineering*, 10:5420–5429, 10 2020.
- [2] John F. Shortle, James M. Thompson, Donald Gross, Carl M. Harris. Fundamentals of Queueing theory. Fourth Edition. Wiley Publications, 2008.
- [3] Lester Lipsky, Manfred Jobmann, Michael Greiner, and Hans-Peter Schwefel. Comparison of the analytic n-burst model with other approximations to telecommunications traffic. *International Journal of Electrical and Computer Engineering*, pages 122–132, 02 2001.