Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інфокомунікацій
(повна назва)
Кафедра Інфокомунікаційної інженерії ім.В.В.Поповського
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка
другий (магістерський)
(рівень вищої освіти)

Контроль перевантаження M2M зв'язку в LTE мережах (Control Overload M2M Communication in LTE Networks)
(тема)

Виконав:студент 2 курсу, групи ТСМім-20-2
спеціальності 172 "Телекомунікації та радіотехніка”
(код і повна назва спеціальності)
освітньої програми "Телекомунікаційні системи та мережі ”
(повна назва освітньої програми
тип програми освітньо- наукова
(освітньо-професійна або освітньо-наукова)
Маді Рабіх Луай Мунтаха
(прізвище, ініціали)
Керівник доц. Кадацька О.Й.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри Лемешко О.В.
(підпис) (прізвище, ініціали)

2022 р

Кваліфікаційна робота не містить відомостей, що заборонені до відкритого друку

Студент 2 курсу, групи ТСМім-20-2  Маді РабІх Луай Мунтаха

Керівник  доц. Кадацька О.Й.

Харківський національний університет радіоелектроніки

Факультет Інфокомунікацій
Кафедра Інфокомунікаційної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 172 "Телекомунікації та радіотехніка’’
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма "Телекомунікаційні системи та мережі "
(повна назва)

ЗАТВЕРДЖУЮ
Зав. кафедри_____
(підпис)
«____»_____2022 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Маді РабІх Луай Мунтаха
(прізвище, і'мя, по батькові)

1.Тема роботи Контроль перевантаження M2M зв'язку в LTE мережах (Control Overload M2M Communication in LTE Networks) затверджена наказом по університету від 14 .03 2022 р №377 Ст

2.Термін подання студентом роботи до екзаменаційної комісії: 20.06.2022р

3.Вихідні дані до роботи: Стандарти 3GPP release 10 . Сценарії зв'язку M2M в 3GPP. Системи масового обслуговування та марковський аналіз математичних моделей СМО

4.Перелік питань, які потрібно опрацювати в роботі. Аналіз особливостей M2M зв'язку в LTE мережах . Виявлення місць перевантаження M2M в LTE мережах Встановлення метрики перевантаження трафіку МТС. Створення математичної моделі алгоритму управління розподілом трафіку M2M та чисельний результат. Показати можливості імітаційного моделювання алгоритму. Висновки.

5.Перелік графічного матеріалу з точним зазначенням креслень,схем, плакатів, комп’ютерних ілюстрацій 1. Мета дослідження 2. Мережа зв'язку M2M 3.Загальна архітектура M2M додатків 4.Сценарій зв'язку з пристроями МТС, які спілкуються в МТС через сервер та без проміжних серверів МТС. 5.Області перевантаження в M2M через мережі LTE 6.Мережа LTE з більш ніж одним MME 7.Розподіл навантаження MME.8. Алгоритм управління розподілом трафіку M2M базових станцій мережі LTE на вузли MME 9. Компоненти SimEvents 10. Структура моделі СМО типу M/M/1/K для реалізації алгоритму управління розподілом трафіку M2M в Simulink Matlab R2015b.11. Висновки

5. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	доц. Кадацька О.Й.		18.06.22

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Збір матеріалів для дослідження	25.03.2022р.	Виконано
2	Розробка розділу1	05.04.2022 р.	Виконано
3	Розробка розділу 2	25.04.2022 р.	Виконано
4	Розробка розділу 3	12.05.2022р.	Виконано
5	Розробка розділу 4	28.05.2022 р.	Виконано
6	Оформлення роботи	20.06.2022 р.	Виконано

Дата видачі завдання 05.01.2022р. _____

Студент  Маді РабІх Луай Мунтаха
(підпис) (прізвище та ініціали)

Керівник роботи  доц. Кадацька О.Й.
(підпис) (посада, прізвище, ініціали)

ABSTRACT

Thesis contains: 81 pages, 25 figures, 3 tables, 29 references.

M2M ,TRAFFIC, QS THEORY, MME, OVERLOAD, NETWORK

The goal of the work is to control network overload based on modeling methods of generating machine-to-machine traffic.

The paper considers the segment of devices of the Internet of Things, which is actively developing, which are all kinds of sensors and exchange information automatically. For M2M traffic, when a large number of sensors of different network elements are instantly triggered, congestion may occur. In the work, areas of congestion in the applications of the M2M LTE network were found. The paper considers load control that occurs at different nodes of the LTE network. load management in the core network part, namely MME. Congestion metrics have been established that analytically assess MTC traffic congestion. A probabilistic strategy for traffic transmission from each base station connected to several MME nodes and an overload threshold are used. A model of an algorithm for managing the distribution of M2M traffic using the QS theory has also been developed. For a network model that serves machine-to-machine traffic, the probabilities of rejection and overload are calculated. The evaluation of the result confirms the need to apply the M2M traffic distribution management algorithm model.

The process of simulation modeling of load metrics in Matlab Simulink environment is considered.

РЕФЕРАТ

Пояснювальна записка містить: 81 сторінка, 25 рисунків, 3 таблиці, 29 джерел.

М2М, ТРАФІК, ТЕОРІЯ СМО, ММЕ, ПЕРЕВАНТАЖЕННЯ, МЕРЕЖА

Метою роботи є контроль перевантаження мережі на основі моделювання методів формування міжмашинного трафіку.

У роботі розглядається сегмент пристроїв Інтернету речей, що активно розвивається, які є всілякими датчиками і обмінюються інформацією автоматично. Для трафіку М2М при миттєвому спрацьовуванні великої кількості датчиків різних елементів мережі можуть виникнути навантаження. У роботі виявлено області навантаження у додатках М2М LTE мережі. У роботі розглядається контроль навантаження, що виникає на різних вузлах мережі LTE. управління навантаженням у частині базової мережі, а саме ММЕ. Встановлено метрики перевантаження, що аналітично оцінюють навантаження трафіку МТС. Використана ймовірна стратегія передачі трафіку від кожної базової станції, підключеної до кількох вузлів ММЕ та поріг перевантаження. Також розроблена модель алгоритму управління розподілом М2М-трафіку з використанням теорії СМО. Для моделі мережі, яка обслуговує міжмашинний трафік, розраховані ймовірності відхилення та перевантаження. Оцінка результату підтверджує необхідність застосування моделі алгоритму керування розподілом М2М-трафіку.

Розглядається процес імітаційного моделювання метрик навантаження в середовищі Матлаб Simulink.

CONTENTS

LIST OF ABBREVIATIONS

ACB - Access Class Barring

CN – Core Network

CTMC- Continuous-Time Markov chain

IoT - Internet of Things

FIFO- First-in-First-Out

HSS- Home Subscriber Server

M2M -Machine-to-Machine

MTC - Machine-Type Communications

MME- Mobility Management Entity

PRACH- Phisical Random Access Channel

PUSCH- Phisical Uplink Shared Channel

QS - Queueing Theory

RA –Random Access

RAN-Radio Access Network

RACH- Random Access Channel

RRM - Radio Resource Management

INTRODUCTION

The last few decades have been in full computing hardware double its capacity every two years. Size of devices and the increase in computing power has decreased by many folds during this time, communication networks have grown toward wireless. Wireless networks have begun to replaced the major part of wired networks. Wireless LANs, WANs and cellular networks have been used vastly to support wide variety of needs in areas. Cellular networks have grown at rapid speed, covering most parts of the world and they to grow more to serve the ever increasing demands. Wireless used in cellular communication was optimized to meet the human type communication (HTC) or voice data. The benefits of wireless cellular networks - availability in diverse geographical areas, cost, which may help their deployment in a variety of applications.

Today the current world usage of cellular networks is vast and the future potentials of its untapped power are tremendous. The remarkable growth in capabilities of computing machines has made it possible for device to interact with each other directly or with minimal or no human intervention is machine - to- machine communications (M2M) [1]. MTC refers to allowing direct communications between MTC devices or from MTC devices to one or more central MTC-based servers, using wired or wireless networks [3]. In a small span of time, MTC devices found their usage in smart meters, intelligent transportation systems, tracking and tracing gadgets, health sector (via telehealth services), security sector (via the use of automated audio-visual monitoring) industrial wireless automation, ambient assisted living, A rapid growth is observed in the design of MTC and its applications. It is projected that by the end of this decade, there will be over millions of MTC devices and connections in service. The major component of the MTC is the communication method being used to transfer the information from one device to another. With such a number of applications and growth, the medium of communication between MTC enabled devices takes the big role.

In order to take advantage of the huge opportunities raised M2M 3GPP is in process of establishing requirements for 3GPP network system improvements that support MTC in the Evolved Packet System [2]. Related transport services for MTC as provided by the 3GPP system and the related optimizations are being considered as well needed to ensure that MTC devices, MTC servers, MTC applications do not cause network congestion or system overload. But, incorporating the MTC devices into the existing cellular networks will bring its own set of challenges, However, in doing so, some challenging issues such as resource management and QoS degradation have to be addressed. The complexity of using the existing wireless for MTC in an environment that was optimized for HTC usage has opened up the vay for a wide area of investigations.

Machine-to-machine communications emerge to achieve communications among all devices. For high layer connections 3GPP have been the solution facilitating M2M communicaions. It is which is being standardized as an application to be supported by LTE. M2M communications has distinct features . To understand M2M communications in 3GPP, we an overview of the network architecture and features of M2M communications in 3GPP and define issues in physical layer transmissions and radio resources allocation.

The major component of the MTC landscape is the communication method being used to transfer the information from one device to another. There is a vast scope of study and improvement concealed in the MTC framework . With such a potential of applications and growth, the medium of communication between MTC-enabled devices takes the central stage. The characteristics of MTC include smaller packet sizes and frequent transmission. One of the most prominent ways to create a communication channel for MTC devices is to setup a completely optimized wireless network that will service the MTC traffic. But designing such a network may necessitate a huge amount of cost and resources. Another alternative way is to use the existing wired networks. But this approach will definitely limit the usage to compact geographical areas only. Thus, existing wireless cellular networks appear as the best possible solution. It has been advocated in [3] existing wireless networks can be used for combined HTC and MTC traffic. But, incorporating the

MTC devices into the existing HTC oriented cellular networks will bring its own set of challenges, for instance, in terms of adjusting the Third Generation Partnership Project (3GPP) standards [4]. However, in doing so, some challenging issues such as resource management and QoS degradation have to be addressed. The complexity of using the existing wireless infrastructures for MTC in an environment that was optimized for HTC usage has opened up the door for a wide area of investigations.

Task

we focus on the problem of congestion when deploying M2M devices over LTE networks;

propose mathematical modelling of algorithm the control distribution of M2M traffic ;

in resource planning problems use the classical queueing theory ;

evaluate of results.

1 MOBILE CELLULAR-BASED M2M DEPLOYMENTS

1.1 Machine to machine (M2M) communication

Machine-to-Machine (M2M) type communications it is automated applications. M2M involve machines or devices communication through a network without human intervention. The M2M devices can be embedded in different environments such as cars, machines and etc. They must transmit information over networks over a large area. M2M deployments use short-range or proprietary radio links. Mobile cellular-based M2M solutions are preferred, what have the advantage of easier installation and provisioning, especially for short-term deployments. Cellular mobile networks offer different network technologies for M2M communications. Many literature sources show a large increase in the number of MTC devices and also in the areas of MTC connection . Annual growth is more than 25% [5-6]. According to these forecasts more of machines or industrial devices will be potentially able to benefit from MTC. To support M2M communication, network operators must use their network to support multiple M2L devices. This creates network congestion for both data and management. Overload may occur due to simultaneous messages from many MTS devices.If many devices detect an event at the same time, they send their alerts toward a central server at the same time. The result of this will be congestion at the network nodes for the entire transmission path.

The Internet of Things uses communication between machines (M2M) [7]. At the same time, a large number of M2M devices communicate with MTC devices or servers to transfer information. The development of M2M applications [8] continues to be relevant. 3GPP and IEEE are implementing standardization processes to achieve these goals.

Mobile networks are a good candidate for suitable machine-to-machine communications. Mobile networks are used for communication in cities, regions, rural

areas. This creates advantages, since it is not necessary to connect new additional base stations for M2M communication. All resources of such networks can be used for inter-person (human to human H2H) and inter-M2M connections. In [9], discussed the importance of categorizing the QoS of MTC and HTC traffic in cellular networks. In [10] focused on the study of the resource management for combined HTC and MTC traffic, with the goal to provide desired QoS to both traffic types. In [11] studied the problem of radio access networks overload, resulting from a mass access to the network by MTC devices, which may degrade the service quality of HTC. Mobile networks are designed for H2H communications, so they need to be adapted for M2M communications.

According to [5] and [6] shown, that even for a single cell, the number of M2M devices can be significant. This number can be more than thousands and tens of thousands. If a large number of M2M devices want to access the eNodeB at the same time, then the radio access network (RAN) will experience congestion.

To request an uplink connection, The MTC device transmits the preamble on the uplink random access channel and if only one device has chosen, then the eNodeB identifies this request. There are various ways to reduce congestion, . There is [12] an Access Class Deny (ACB) mechanism to control radio access network congestion. This allows M2M devices to use a probabilistic strategy for transmitting requests. These studies and the results obtained show good use of the established ACB coefficients in [13]. In [14] it is shown how the Random Access Channel can be divided into H2H and M2M traffic, which provides access to H2H users.

Adaptation of the ACB factor is also used to reduce RAN congestion. It counts the number of successful transfers and the number of collisions for the same interval. If the number of transmitted preambles increases, so does the number of simultaneous transmissions of the preamble, which occurs with an increase in the ACB coefficient. A threshold value is used to control the ACB ratio and update it.

The characteristics of MTC include smaller packet sizes and frequent

transmission In MTS device applications, quality of service (QoS) must also be considered. One of the most prominent ways to create a communication channel for MTC devices is to setup a completely optimized wireless network that will service the MTC traffic. But designing such a network may necessitate a huge amount of cost and resources. Another alternative way is to use the existing wired networks. Designing such a network may necessitate a huge amount of cost and resources. Existing wireless networks can be used for combined HTC and MTC traffic.

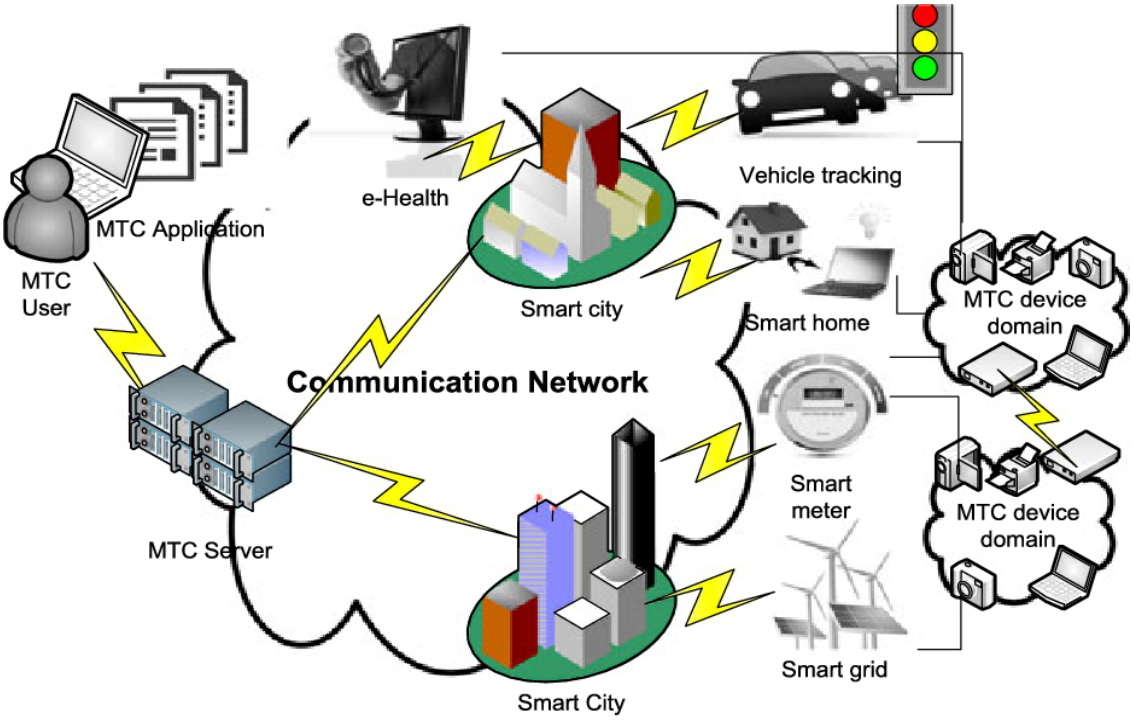

Figure 1.1- Machine-to-machine communication network

1.2 Cellular networks improvement for technologies MTC

Categorizing the QoS of MTC and HTC traffic in cellular networks the importance, focused on the study of the resource management for combined HTC and MTC traffic, with the goal to provide desired QoS to both traffic types. In [8] studied the problem of radio access networks overload, resulting from a mass access to the network by MTC devices It is may degrade the service quality of HTC.

The usage of MTC over the existing HTC traffic in cellular networks has resulted to new

challenges and complexity in terms of radio resource management (RRM). In the current era, various types of wireless networks are being used to transfer the data between different types of devices in wireless communication networks. Examples of such networks include wireless LAN, Wi-Fi, Wi-Max, ZigBee, TransferJet, Bluetooth, Ham radio network, cellular networks, to name a few. Among these, cellular networks are the most widely used networks because they have been proven to cover a wide range of geographical areas. The technology of these networks has evolved from GSM to 2G, 3G and nowadays 4G or LTE [9]. This technology evolution has been made possible due to noticeable ad- vancements in human-centric computing, microarchitecture design, power-performance of multi-threaded and multi-core processors, improved battery/power life, digital signal processing, transmitters design. Throughout its evolving cycles, cellular networks have continuously provided better data rates. In the initial starting phase of 2G-based cellular networks, a data rate of 14.4 kbps was achieved, and improved over time to 171 kpbs by using the General Packet Radio Service (GPRS) technology, to 384 kbps by using the Enhanced Data Rates for GSM Evolution (EDGE) technology, to 2 Mbps during the 3G phase using the Wideband Code Division Multiple Access (WCDMA) technology, to 14.4 Mbps using the High Speed Packet Access (HSPA) technology, and nowadays to the range 50-100 Mbps along with guaranteed QoS, improved spectrum efficiency and larger coverage using the 4G (or LTE) technology .

Cellular networks work and expand with the basic building blocks referred to as cells. Cells define the coverage area of a cellular network. These cells are served by an infrastructure known as Base Station (BS) or collection of BSs. The physical devices such as antenna, power backups, signal transmitters, processing devices, to name a few, are located in these BSs. The area covered by the cell depends on the capacity of the BSs, which are themselves connected to the core network and are assigned a group of radio frequency bands or channels. Each BS can support multiple users or devices which are connected to it by means of a range of radio frequencies or channels. Due to the rules and regulations that were put in place by governments and technology standard bodies, only a specific set of frequencies can be used to transmit the mobile signals. This limitation really limits the number of radio channels that can be used in the

BS for signaling. Because the channels are limited and there are multiple users and cells in the network, setting parameters such as data rate, user distribution, transmit power and receive power, modulation scheme, handover criteria, etc. is also important in resource management. Due to advances in wireless technologies, wireless cellular networks, which were initially optimized to transfer the voice data, are now transitioning from the mobile phones age to the wireless computing age, resulting to an increase in data transfer rates. This allows cellular networks to provide services such as video streaming. With this new development, all the communications that were originally initiated with human intervention via mobile phones or similar devices using HTC still prevail, in addition to allowing the devices to enable communication between each other without human intervention thanks to advances in hardware technology, computational power of devices, and artificial intelligence. This improvement has ope ned the door for future technologies such MTC to extend the features of the established cellular networks. Cellular networks have tremendous untapped potentials which can change the lifestyle of our next generations by providing far more services like wireless TV, security monitoring, tele-medicine, tracking and tracing, to name a few.

1.3 Machine-to-machine communication as a form of data transfer

Machine-to-machine communication (M2M) or MTC is a form of data transfer. Two or more entities interact independently with each other without any human interactions or supervision. This communication can either happen using the wired or wireless systems. The main idea of MTC is to reduce the dependency of devices over human actions. Need making them self-sufficient to initiate the actions based on the available network information. In order to replace the decision making intelligence of human with that of machines, it is required that some information be gathered from the devices, including the devices processing power.

1.3.1 Some applications based on MTC

Typically, a large number of MTC devices are involved in MTC applications, and in

most cases, MTC devices support the uplink transmission of data. MTC applications include but are not limited to transportation, health care, safety, security, tracking, home automation, to name a few, and cellular networks are suited for these types of applications. Some important benefits of using MTC in cellular networks include:

1) When M2M is used in mobile networks, services can be provided everywhere. At the same time, security, high speed, mobility, and high throughput are ensured. This does not require major changes in the standards;

2) Various technologies such as femtocells can be used to improve the quality of service for mission-critical applications.

A rapid growth in the use of MTC devices in cellular networks is expected to occur in the next decade, with an annual rate of more than 20 % . According to GSMA , a leading pan-european organization, MTC connections have reached more 350 million .

Salient features of MTC applications different from those of HTC applications have been described in the 3rd Generation Partnership (3GPP) project . It is not necessary for an MTC application to follow strictly all the features of HTC applications. These features can be activated individually in a system. The features of MTC as defined by 3GPP Release 10 can be summarized as follows[14].

• Small data transmissions: small data packets can be exchanged in MTC traffic. MTC devices can send the recorded data such as temperature, meter readings, GPS coordinate, to name a few. Large number of devices: MTC traffic associated with a large number of devices .They connected to a network at the same time.

• Low mobility: the movement of MTC devices is very limited.They is restricted to a certain predefined area only.

• Time controlled: the transmission and receipt of data by MTC devices are restrictedto particular time intervals (slots).

• Time tolerance: MTC devices can sense the traffic and can delay their data transmission.

• Priority alarm: MTC devices can send priority alarm messages such as fire

alert, to name a few.

• Packet switched only: packet switched services are provided to MTC devices with or without the need to allocate a mobile subscriber integrated services digital network number.

• Secure connection: a secure connection is required between MTC devices and servers.

• MTC monitoring: this feature is used by MTC applications that require the monitoring of the events related to MTC devices.

• Location specific trigger: MTC devices are triggered by using their location information.

• Infrequent transmission: random transmission and long intervals between consecutive transmissions from MTC devices can be implemented.

• Mobile originated communication only: mobile originated communication can be implemented for mobile MTC applications that require this feature.

• Infrequent mobility termination: This feature is used to reduce the mobility management frequency of MTC devices that support mobile originated communications.

• Network provided destination for uplink data: this feature can be used for the purpose of uplink transmission of data to the network.

• Group-based MTC features: MTC devices can be managed as a group in case the samemessage needs to be transmitted or a combined QoS policy needs to be enforced on multiple MTC devices.

MTC-based applications have been some of which are captured in table 2.2.

Table 1.1 -MTC applications

MTC applications	Examples
Tracking and tracing	Emergency call Fleet management Theft Tracking Traffic InformationNavigation Pay as you drive (PAYD)
Smart Meters	ElectricityGas Water
Health	Remote patient monitoring Assisted living Personal fitness
Security	Access controlAlarm Systems Surveillance systems
Home Automation	Thermostat control Lighting control Appliance control
Remote Maintenance and Control	Vehicle diagnostics Vending machine control

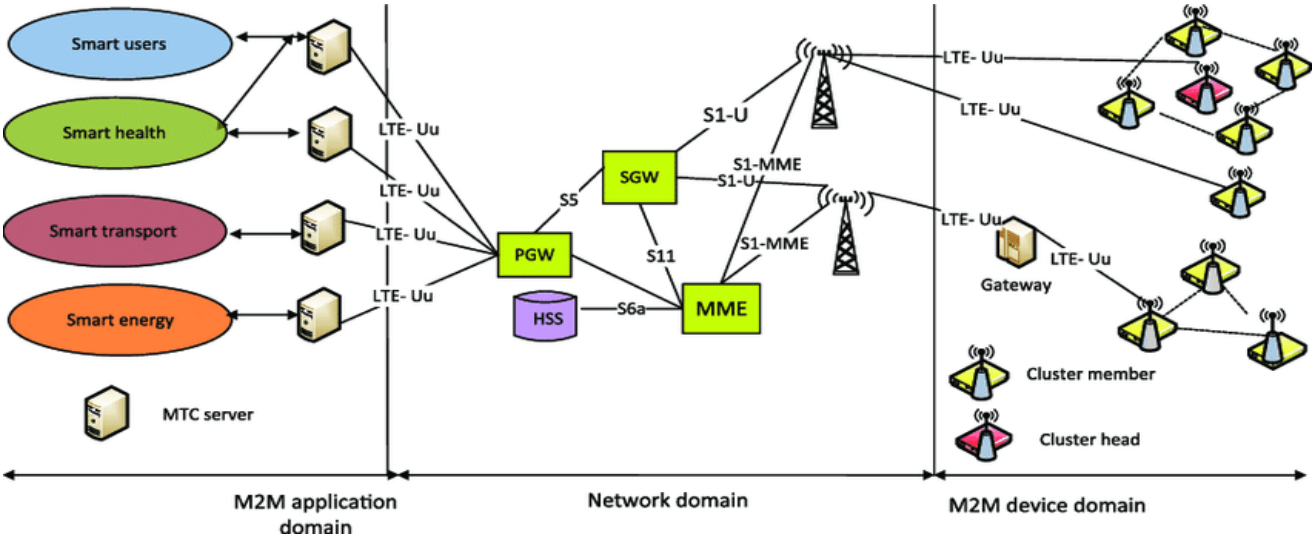

Figure 1.2- Generic architecture of an M2M application

1.3.2 The studies on MTC by 3GPP

In its Release 8 the initial study on MTC by 3GPP was introduced . In 3GPP release

10 , the support for MTC traffic along with the service requirements for MTC traffic were introduced. These include the subscription options, the process of sending and receiving the data based on triggers, the addressing schemes, the charging, security, and remote management requirements, to name a few. The system architecture Working Group 2 (SA2) of 3GPP defined the architectural requirements and models to support MTC in 3GPP networks . In the future 3GPP release , it is expected that significant efforts will be dedicated to analyzing and optimizing the network architecture to reduce the impact of MTS on traditional traffic in cellular networks. Due to the expected use of a large number of MTC devices, key issues such as IP addressing, signaling, congestion, communication overload. It is to name a few are required to be improved. One way to optimize the system in order to deal with these issues con- sists in using IPv6 addresses and grouping similar MTC devices for management purpose.

• MTC device domain: this domain is composed of all MTC devices that are installed for autonomous data collection and transmission. Smoke detectors, theft control devices, fire alarms, smart meters, fitness or health monitoring devices, data collector sensors, tracking devices, traffic sensors, to name a few, are examples of physical MTC-based devices belonging to this domain. These devices transmit data to MTC servers or among each other .

• Network domain: This domain is the backbone of the whole MTC . Its goal is to provide communication between MTC devices and MTC servers or among MTC devices, through a wired or wireless network. 3GPP cellular networks such as UMTS or LTE to be used as network domain for MTC applications.

• MTC application domain: he consists of MTC servers that serve as destination for the data transmitted by the MTC devices over the network. Based on various usage scenarios, MTC servers can be controlled and managed by mobile network operators or third party service providers [5]. MTC servers provide end users with an interface to access the assigned MTC applications.According to 3GPP the communication scenarios of MTC traffic can be segregated into two models based on various different requirements:

• Direct communication model: In this model, there is a direct communication among MTC devices which is provided by the 3GPP operator. MTC devices within the same network domain or different network domains can communicate to each other directly, in such a way as to establish a peer-to-peer connection . Figure 1.3 shows the communication scenario between MTC devices.

Figure 1.3 - MTC devices communicating directly with each other

• Indirect communication model: This model depicts a client-server model, where MTC devices (clients) transmit the data to one or more MTC servers. This scenario can find applications in smart metering, traffic controls, monitoring applications, to name a few . In this communication model, the MTC server can reside inside (or boutside) the network domain, thereby, it can be controlled by the 3GPP network provider (or a third party service provider). When the MTC server is inside the network domain, the network provider offers an API to the MTC users for accessing the server. Figure 1.4 shows the communication scenarios between the MTC devices and the MTC servers.

Wireless personal communications have been widely applied to exchange voice, audio, video, emails, photos, and more among individuals. Such demands of ubiquitous communications among humans thus drive the development of abundant advanced wireless technologies and systems such as the cognitive radio network (CRN) and Long Term Evolution-Advanced (LTE-Advanced). In addition to human-to-human (H2H) communications, an emerging technology

empowering full mechanical automation e.g., the Internet of Things and the smart grid is vigorously being developed.

(a)

(b)

Figure 1.4- M2M devices communicating with M2M servers
(a) MTC Serverlocated inside the network domain and (b) MTC Server located outside the network domain

In advanced long-term evolution (LTE-Advanced). Along with communication between people, technologies of complete mechanical automation are developing. These are machine-to-machine communications [11].

The following communication classes are used for this.

1) Communication between the device and the person who controls the operation of the device.

2) Communication between several parts of the calculator. In this case, intermediate calculations are used to exchange between parts of the overall computing process. Both cloud computing and distributed computing are used.

3) Communication between the person managing and carrying out the actions.

The sensor network is the simplest type of M2M communication. Sensor measure physical quantities and transmit this data to control them. An example would be a data acquisition and control system. In it, the control center makes a decision, for which it is necessary to periodically poll the sensors and ensure reliable communication. . However, the structure of communication in a sensor network For intelligent networks, each device can be both a sensor and a controlling person and an executor.

In M2M communications, it is necessary for such networks to provide complex connections between all smart devices.

For implementations of M2M communications, it is proposed to use Bluetooth (IEEE 802.15.1), Zigbee (IEEE 802.15.4) or WiFi (IEEE 802.11b). There is no consensus for building a network architecture of the general scenario of M2M communications, but in general, it is considered as a heterogeneous mobile specialized network, since M2M communications create multifunctional connections between all devices, M2M communications have the same difficulties as heterogeneous mobile specialized networks.

Therefore, research in this area for connections, routing congestion control, etc. can be used. Of course, there are limitations imposed by the hardware complexity of MTS devices, so the connection management of a large number of MTS devices scattered geographically must be adapted.

As a consequence, the 3GPP has begun to standardize the communication schemes to support M2M communications.

3GPP provides wired connections between stations, which in LTE-Advanced can be Universal Terrestrial Radio Access eNBs or home eNBs. Through this it is possible to provide connections between all higher level MTC devices. But the use of M2M communications in 3GPP will not necessarily be successful, and one of the difficult problems is the air interface. For IMT-Advanced in the fourth generation (4G) wireless system, the air interface in LTE-Advanced) is designed to meet the high peak data rate

for H2H communication, and for high-speed transmission, the air interface structure may not effectively support M2M communication. In M2M communication, small data volume transmitted by a very large number of devices. In the standardization process, the impact of M2M communication in LTE-\Advanced is currently being studied in 3GPP.

3GPP started with the M2M communications air interface standardization process. When it is necessary to apply LTE-Advanced transmission schemes to MTC devices, it is first necessary to identify the impact on system performance. Let's review M2M communication in 3GPP LTE-Advanced, and show what tasks arise with the LTE-Advanced air interface to support M2M communication.

1.4 Overview of M2M communications

According to the 3GPP definition, two M2M communication structures are defined, namely the communication of MTC devices with one or more MTC servers (shown in fig. 1.5).

Here, the user of such a connection - a sensor or an individual, can manage a large number of MTS devices through the MTS servers. The operator domains A and B in figure 1.5 can be the same, i.e. and the M2M servers and the entire LTE-Advanced infrastructure are in the same carrier domain. In LTE-Advanced, there are eNB macro cells, eNB pico cells, and HeNB femto cells.

Figure 1.5 shows the S1 interface between the Mobility Management Entity (MME), to serve the S-GW. The Packet Data Network Gateway P-GW and eNB are shown. The HeNB node can also communicate with the MME/SGW/P-GW using the S1 interface. Thus, MTS devices, when connected to these LTE-Advanced stations, are controlled by the M2M user through the MTS servers.

It is also possible that the MTS server is located in the same way as the MTS user is not in the operator's domain. In figure 1.5, the lines are designated as physical connections; so logical connections.

Another way to connect is the structure shown in figure 1.6. In such a structure M2M devices interact with other M2M devices without an intermediate MTS server. Here the lines show the physical connections; and logical connections.

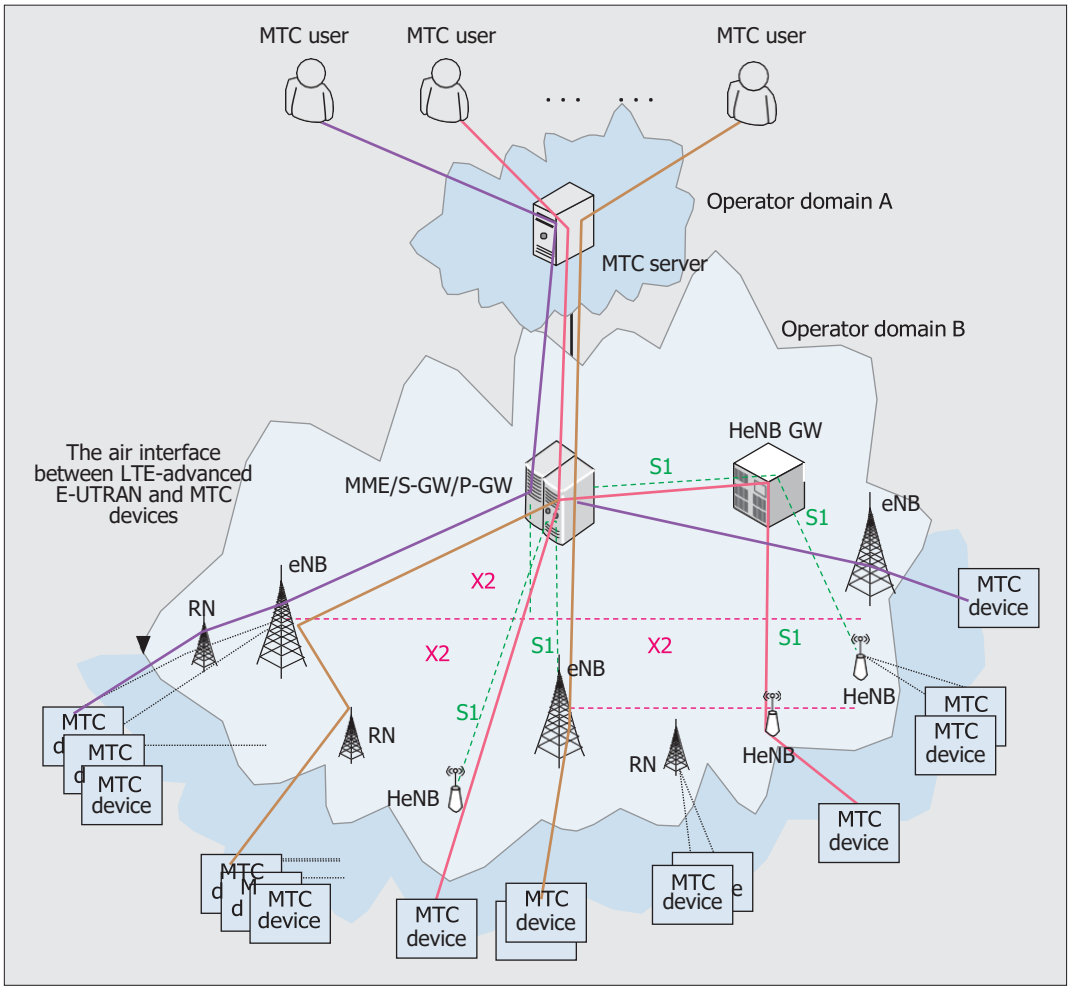

Figure 1.5- The communication structure with MTC devices communicating with the MTC server

Communications among MTC devices can happen within the same operator domain or among different ones. In both cases, MTC devices shall attach to LTE-Advanced stations, and packets are forwarded by the LTE-Advanced infrastructure. To enable communications between MTC devices and MTC server(s), the public land

mobile network (PLMN) shall allow transactions between an MTC device and an MTC server, initiated by either the MTC device or MTC server.

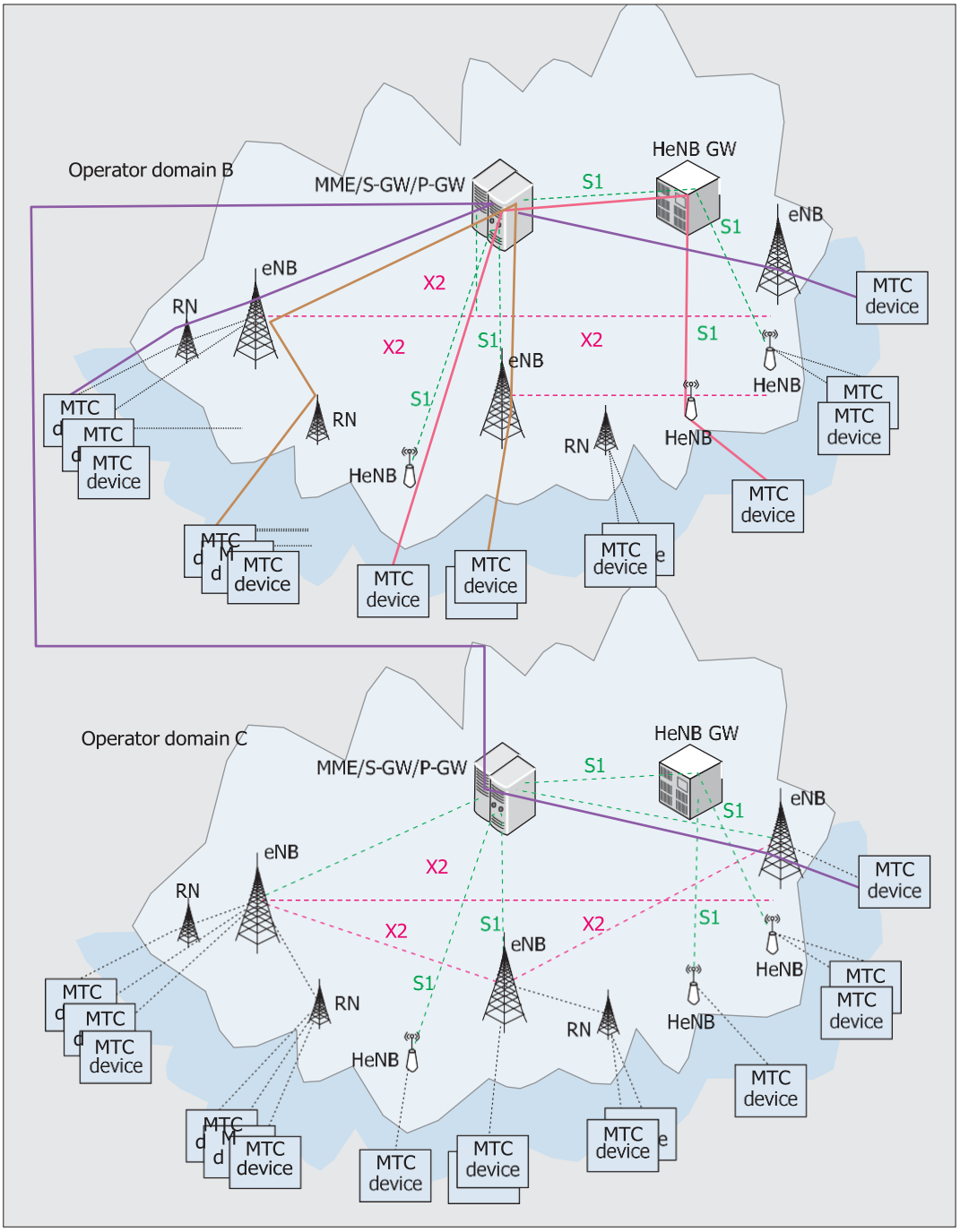

Figure 1.6- The communication structure of MTC devices communicating with each other without intermediate MTC servers

The PLMN shall also be able to authenticate and authorize an MTC device before

the MTC device can communicate with the MTC server.

1.5 Features of M2M communications

M2M communications use different applications. 14 functions in M2M communications are defined to characterize applications in 3GPP. Some communication characteristics in M2M are very different from those for H2H.

The common characteristics of communication between M2M and H2H are mobility, packet switching, secure connections. Distinctive characteristics are rare transmissions, transmissions of small amounts of data. In addition, MTS devices send and receive data only during a certain time interval. This is the time to grant access. and if the interval is prohibited, then the network rejects requests for access, for sending and receiving data and signaling of MTC devices. For some control tasks, as well as for resource allocation, the system should provide the ability to associate an MTC device with one or more MTC groups [10, 11]. The literature shows some control schemes for MTS devices with such functions in [11, 12]. Such schemes are not compatible with LTE-Advanced and therefore cannot be applied to LTE-Advanced without further research. Therefore, MTC device control schemes with such functions are under research and development in 3GPP.

Considering the features of M2M communication, which are designated in 3GPP, it is clear that the 3GPP network defines the basic rules in the coordination of communication between MTC devices, in communication between MTC devices and the MTC server. This architecture and features lead to a number of differences between 3GPP M2M communications and general M2M communications. M2M communication can be shared, it is a difficult task to solve a correct connection between any M2M devices that are distributed around the world. Therefore, the disadvantage of general M2M communication is loss of end-to-end connection, transmission errors, etc. On the

other hand, in 3GPP, failures must be avoided, and the network architecture in 3GPP supports reliable connections and communication. Due to the congested spectrum on a global scale, there may not be acceptable spectrum for general M2M communications in addition to the industrial, scientific and medical band. As a result, in addition to using the high frequency band (60 GHz), a future solution for the general M2M communications air interface could be the use of cognitive radio to reuse licensed spectrum.

1.6 Random access channel congestions

1.6.1 Channel adaptation methods required for M2M

The LTE-Advanced air interface is based on the communication standards IEEE 802.15.1, IEEE 802.15.4, etc., and also uses other advanced technologies, such as cognitive radio. 3GPP has a specific infrastructure that manages complex resource allocation mechanisms. The problem of detecting technical problems at the radio interface remains relevant. Therefore, radio interface design problems, assumptions and adaptations are being studied, tv 3GPP M2M communications.

To standardize M2M physical layer communications and apply existing physical layer mechanisms in LTE-Advanced, research is needed to study the mechanisms and features of M2M communications. It is necessary to study how to apply existing physical layer mechanisms in LTE-Advanced for M2M communication.To achieve the high peak data rate requirement for the 4G wireless system (1 Gb/s for static and 100 Mb/s for high speed mobility), sophisticated link adaptation techniques, such as single-user multiple-input multiple-output (SU-MIMO), multi-user MIMO (MU-MIMO), adaptive modulation, hybrid automatic repeat request (HARQ), beamforming and coordinated multiple-point (CoMP) transmission, are included as mandatory functions in 3GPP Release 11 and further versions.

Existing channel adaptation techniques require complex procedures to evaluate the

channel and report such results. It is known that due to the transfer of small amounts of data, each transfer between M2M devices may not require a high speed data transfer scheme. Instead, reliable transmission with low bit error rate, BER and low delay is important. The channel adaptation techniques in LTE-Advanced can lead to significant power consumption in MTC devices small delay.

In LTE-Advanced, the physical downlink control channel (PDCCH) area is the first 1, 2, or 3 orthogonal frequency division multiplex (OFDM) symbols of a subframe, which comprises several control channel elements (CCEs) as the minimum unit carrying radio resources allocation and control information. The number of CCEs allocated to the PDCCH is 20. PDCC Format 1 assumes that control information for one M2M device is transmitted when combining 2 consecutive CCEs. One transmission subframe supports 10 M2M devices. Thus, such transmission does not support a large number of MTC devices and a number of UEs. Therefore, studies are also needed on whether the PDCCH should be separated for M2M devices and UEs and additional resources for the PDCCH, A discussion of the problems and trade-offs in radio interface design is needed. In addition, further studies of 3GPP M2M communication are required and how to apply existing physical layer mechanisms in LTE-Advanced to M2M communications.

1.6.2 Possible solutions continuous collisions in RACH

In LTE-Advanced, random access works as follows. If the UE performs a handover from one eNB to another eNB and the number of UEs is limited. or when uplink synchronization is lost. But the number of M2M devices can be much larger than the number of UEs. Therefore, M2M devices and UEs have random access channel (RACH) collisions if so. This is one of the critical issues. The existing solutions are as follows.

1) Scheme based on delay. For such a solution, the backoff time of the UE is set to eg 25 ms and the backoff time of the M2M devices is set to a large value eg 970 ms. This situation can mitigate collisions and increase the backoff time and facilitate collision

resolution. Applying such a scheme can improve performance if there is no congestion in the RACH, but a backoff based scheme. does not solve the congestion problem when the RACH is congested.

2) Scheme based on access class prohibition (ACB). Such a scheme has been designed for UE access control. The ACB contains the following 16 Access Classes (ACs). AC 0-9 represents normal UE, AC 10 for emergency call representation. Access classes AC 11-15 for certain high priority services.

The access probability p and the AC prohibition time are transmitted by the eNB for the UEs corresponding to AC 0-9. The transmission is broadcast to reach the ACB. The UE randomly selects a value of q, $0<q<1$ and if $q>p$, then the UE starts the random access procedure. If it is not, then the UE is blocked for the duration of the AC ban. In [9] the same scheme is proposed for random access control of MTS devices. The use of the ACB access class prohibition scheme allows the high levels of congestion in the RACH to be mitigated by using a small p value. But such a small value generates too much delay in the delivery of the request. For the case when the overload does not last long, such a circuit may not have time to correct p in time.

3) Separation of RACH resources. The RACH resource sharing scheme separates preambles and time-frequency resources for UEs and M2M devices. This is to ensure that a common preamble for shared time-frequency resources of a large number of UEs and M2M devices is not used. The impact on the UE is reduced, but the performance is not improved at a very high level of overload of M2M devices.

4) Dynamic allocation of RACH resources. The eNB dynamically allocates additional resources to the RACH. The decision is made based on knowledge of the congestion level and knowledge of the total traffic load. Although dynamic RACH resource allocation This scheme is effective, but if you want to improve performance, then you need to apply additional resources.

At the moment none of these schemes has tangible advantages in terms of acceptable performance. Thus, the elimination of congestion in RACH is the subject of

research and requires further study.

1.7 Diverse QoS requirements in M2M communications

For MTC devices, some applications require time limits for critical applications. Disasters will occur if this time is not observed. Such signals can be the values of measurements from various counters, the transmission of navigation signals, in medicine, etc. Therefore, in M2M communications, a very important requirement is to provide QoS guarantees.

For H2H communications, many schemes and methods have been developed to provide QoS. These studies may not be directly applicable to M2M communications for the following reasons.

1) Extremely varied QoS requirements H2H communications with time constraints are multimedia where packet arrival periods are between 10 and 40 ms. Due to infrequent transmission, periods of packet arrival in M2M communications range from 10 ms to several minutes. Therefore, there are extremely diverse QoS requirements that cause difficulties in the development of radio resource allocation algorithms in M2M communications. For these, jitter characteristics must be provided. Jitter in this case completely defines the timing characteristics of periodic traffic, as the difference between the time of departure of two consecutive packets and the time of arrival of two consecutive packets.

2) Each M2M device can occupy only a few orthogonal frequency bands with small data volume, bulk transmissions and usually burst arrival of packets.

2.DETECTION OVERLOAD AREAS IN M2M OVER LTE NETWORKS

2.1 Description nodes LTE networks

To send messages, LTE networks use unidirectional channels in M2M communications. The bearer itself is installed in the MME, which is used to bi-directionally route IP traffic between the devices and the P-GW with QoS set for data only, not for signaling.

For different streams, multiple bearers can be configured for a UE and provide QoS or connectivity to different PDNs. Different interfaces are used for delivery to the P-GW.

1) eNodeB-Evolved RAN (radio access network), an advanced version of the LTE base station.

2) Packet data node gateway P-GW. It allocates IP addresses to UEs. It is an interface function between the LTE network and external packet data networks. The UE communicates with many P-GWs at the same time to access multiple PDNs. In addition, P-GW manages quality, inspects packets, and does packet filtering for each user. P-GW provides sharing, packet inspection, policy enforcement. In addition, P-GW provides mobility between 3GPP and non-3GPP technologies - WiMax ,CDMA1X , EvDo.

3) S-GW-Serving Gateway. in transit and forwards user data packets, as well as a mobile communication point for serving messages between eNodeBs and a point of exchange for replacement between 3G and LTE sets. In such a case, manage and save the information context of the UE, as well as save the forwarding data of the channels that are in communication when the UE is in the idle mode and when the MME is once in a rechannel with the UE.This is the main hub for the LTE access network. At the same time, it is determined, authenticated (mutually modified with the HSS) and the correspondence between them is maintained. They also take part in the wear activation/deactivation process and take over the functions associated with obtaining such data (authentication and security establishment between the network and the UE). In addition, it helps to

change the bill of lading in the radio coverage, accepting information about the UE. These functions are handled by peer-to-peer non-access (NAS) session.

4) MME-Mobility Management Entity. This is the main hub for the LTE access network. At the same time, it is determined, authenticated (mutually modified with the HSS) and the correspondence between them is maintained. They also take part in the wear activation/deactivation process and take over the functions associated with obtaining such data (authentication and security establishment between the network and the UE). In addition, it helps to change the bill of lading in the radio coverage, accepting information about the UE.These functions are handled by peer-to-peer non-access (NAS) session [16].

5) HSS-the Home Subscriber Server. It contains subscription data users. It also holds the PDN to which user can connect as well as dynamic information such as MME to which the user is currently attached or registered.

6) M2M-IWF-M2M Interworking Function hides the internal PLMN (Public Land Mobile Network) topology and relays or translates signaling proto-cols used over MTCsp to invoke specific functionality in the PLMN.

2.2 Main types of LTE network overload in the context of MTC applications

As shown above, one of the problems of the cellular network is the congestion that occurs due to the large number of MTS applications. A large number of M2M devices. located in a small area may rarely or often exchange data. These are small amounts of data and you need to send data at the same time. In this case, the same network nodes are used - eNodeB, MME, S-GW, P-GW, HSS. Of course, these actions lead to congestion, which reduces the possibility of traffic transmission for devices that do not support M2M.

In the context of MTC applications, two main types of LTE network congestion can occur depending on where it occurs [9, 19, 5].

• Radio network congestion: This often happens in the eNodeB. A large number of M2M devices simultaneously try to connect to the network, activate, change or deactivate

connection. These devices use the same channels to be able to connect to the same eNodeB.

• Core network congestion: this occurs when messages are transmitted simultaneously from a large group of M2M devices to different cells. This congestion occurs in the EPC and is mapped to the MME , S-GW and P-GW as well as to the HSS when many devices are registered to the same HSS.

These types of overloads are shown in the figure 2.1. overload due to user data. Currently, there is interest in managing congestion caused by a large number of signaling messages coming from several M2M devices.

Due to factors such as advanced LTE radio technology, LTE's wide bandwidth research on packet transmission is less important.

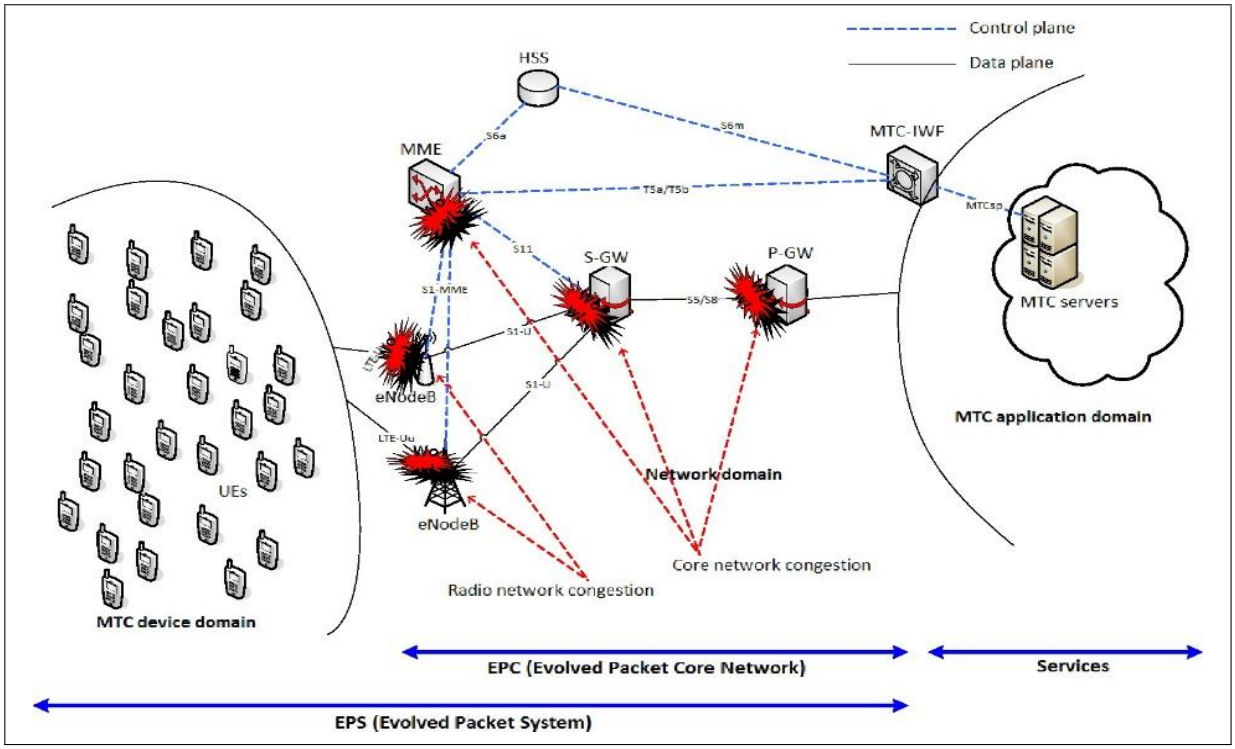

Figure 2.1 - Overload areas in M2M applications over LTE networks

2.3 Analysis of traffic failure paths in MTS over LTE networks

Let's analyze some ways to prevent congestion in LTE networks in the context of MTC and consider two categories. Category of soft mechanisms. This path consists in the mobile operator trying to minimize the number of attempts by M2M devices to connect without having to suppress them.

Solutions are as follows.

1) Pull model.Through the network, the MTS server sends alerts to the desired devices to start the MTS devices. Devices will be allowed to connect to the network. This may be the case when the location of the device is known to the application.

2) TAU is an abbreviation for tracking area update signaling. With TAU, a device notifies the network of its current location on the network. If M2M devices have little mobility, then you can not do this. This minimizes signaling and increases the TAU period. For static M2M devices, it can be disabled.

Category of rigid mechanisms. This way is that, in the event of a conflict, the mobile operator disables M2M devices that are trying to connect to the network at the same time. Solutions are as follows.

1) Creation of M2M device groups. Characteristics are taken into account low priority, low mobility, small data transfer, etc. Using these characteristics, groups of devices are created. The core network receives information about the groups and M2M devices of these groups. For a subscription, the group appears as one. Signaling is optimized and device management is simplified.

2) Distribution of time. The times are determined - permission, prohibition, communication windows, when M2M devices are allowed to connect to the network. Periods are determined based on the device's subscription to HSS. For the forbidden interval, M2M devices are not connected. For some applications, you can save the time slot for providing M2M device connectivity. Therefore, you do not need to be connected to the network all this time. This is a short communication window and sufficient to

achieve the goal.

3) Randomization. For such a path, randomization of the access time of M2M devices is used according to the communication window. The start times of the various communication windows of M2M devices are randomized to reduce signal and data traffic bursts during the short communication window.

4) Rejecting the connection request. This is a way to reduce signal overload. The M2M signaling traffic is rejected in the radio access network or in the MME. Rejections occur only for M2M traffic of MTS applications causing congestion. Non-MTS traffic is not affected. Let's define two sources of failure.

* Deviation by RAN.

The MME sends a notification to the RAN nodes to inform them of the MTC ban information. This occurs in accordance with the congestion status feedback from the P-GW and S-GW to establish access to the MTC control. The goal is to overcome the overload caused by MTS applications. Parameters such as inhibit ratio, MTC group to block inhibit time, etc.) are used. Random access channel resources can only support low and medium traffic load. They cannot serve M2M traffic for a large amount of data transferred, increasing the risk of collision for MTS traffic and for non-MTS traffic. In [11], various solutions are proposed for managing access to the MTS at the RAN level.

* Separation of RACH resources for MTC devices and non-MTC devices. This is used to provide network access for M2M traffic and H2H traffic and limit the amount of MTC traffic.

* Dynamic RACH resource allocation is also enabled. If the network knows the period of time during which M2M devices transmit information and it is necessary to increase RACH resources for devices

*Using a large delay window. Used for M2M devices to distribute all access attempts from M2M devices into large backoff windows. These are actions to reduce contention for RACH resources. It also increases the probability of access for higher priority H2H

traffic.

∗ ACB - providing access class. The method reduces collision probability for multiple transmissions simultaneously in the same RACH resource. It conveys the access probability to the UE/MTC devices and is compared with a random number generated by each UE. After that, it is decided whether it is possible to move to the random access channel and access is denied for the average access ban time.

* CAAC - congestion-aware access control. [9]. In CAAC, the rejection probability is determined by the MME, S-GW and P-GW nodes under congestion conditions. Groups of MTS devices are also created, priorities are assigned according to priority classes, and probabilities are assigned to groups. eNodeBs accept or reject M2M traffic.

∗ Rejection by MME. The HSS informs the MME about the allowed and prohibited timeslots and transmits them through the MTC server to the MTC devices. Congestion may occur at the allowed time, in which case the M2M devices are given a delay time for later access. Or, a congestion control notification message is sent to them to reduce data transfer.

2.4 Issues in resource allocation

Resource allocation and congestion control are complex issues that have been the subject of much study ever since the first network was designed. They are still active areas of research. One factor that makes these issues complex is that they are not isolated to one single level of a protocol hierarchy. Resource allocation is partially implemented in the routers, switches, and links inside the network and partially in the transport protocol running on the end hosts. End systems may use signalling protocols to convey their resource requirements to network nodes, which respond with information about resource availability. Resource allocation is mean the process by which network elements try to meet the competing demands that applications have for network resources—primarily link bandwidth and buffer space in routers or switches. Of course, it will often not be

possible to meet all the demands, meaning that some users or applications may receive fewer network resources than they want . We use the congestion control to describe the efforts made by network nodes to prevent or respond to overload conditions. Since congestion is generally bad for everyone, the first order of business is making congestion subside, or preventing it in the first place. This might be achieved simply by persuading a few hosts to stop sending, thus improving the situation for everyone else. However, it is more common for congestion-control mechanisms to have some aspect of fairness—that is, they try to share the pain among all users, rather than causing great pain to a few. Thus, we see that many congestion-control mechanisms have some sort of resource allocation built into them. It is also important to understand the difference between flow control and congestion control. Flow control involves keeping a fast sender from overrunning a slow receiver. Congestion control, by contrast, is intended to keep a set of senders from sending too much data into the network because of lack of resources at some point.

2.5 Packet-switched network

We consider resource allocation in a packet-switched network consisting of multiple links and switches . In such an environment, a given source may have more than enough capacity on the immediate outgoing link to send a packet, but somewhere in the middle of a network its packets encounter a link that is being used by many different traffic sources. Figure 2.2 illustrates this situation—two high-speed links are feeding a low-speed link. This is in contrast to shared-access networks like Ethernet and wireless networks, where the source can directly observe the traffic on the network and decide accordingly whether or not to send a packet. The datagrams are certainly switched independently, but it is usually the case that a stream of datagrams between a particular pair of hosts flows through a particular set of routers. This idea of a flow—a sequence of packets sent between a source/destination pair and following the same route through the network—is an important abstraction in the context of resource allocation.

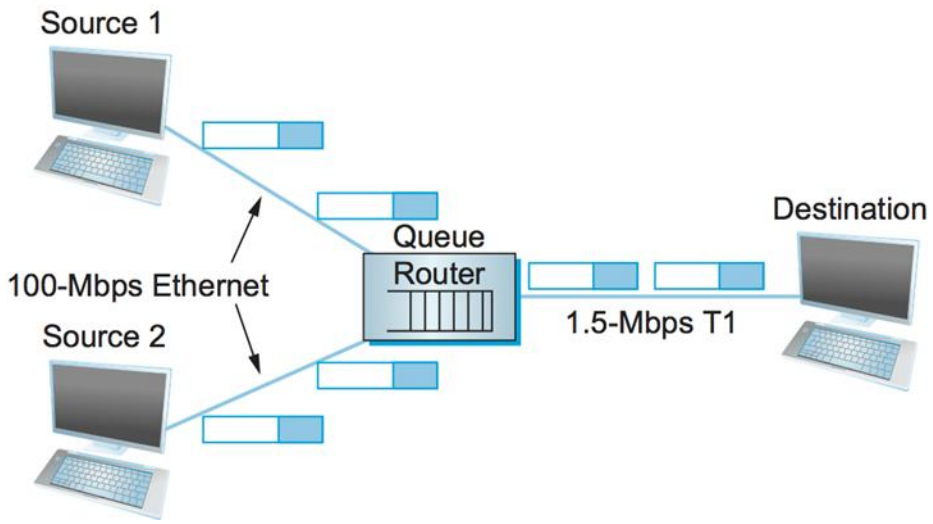

Figure 2.2 - Queueing system two high-speed links are feeding a low-speed link

One of the powers of the flow abstraction is that flows can be defined at different granularities. For example, a flow can be host-to-host (i.e., have the same source/destination host addresses) or process-to-process (i.e., have the same source/destination host/port pairs). In the latter case, a flow is essentially the same as a channel,

With best-effort service, all packets are given essentially equal treatment, with end hosts given no opportunity to ask the network that some packets or flows be given certain guarantees or preferential service.

Effective resource allocation. A good starting point for evaluating the effectiveness of a resource allocation scheme is to consider the two principal metrics of networking: throughput and delay. Clearly, we want as much throughput and as little delay as possible. Unfortunately, these goals are often somewhat at odds with each other. One sure way for a resource allocation algorithm to increase throughput is to allow as many packets into the network as possible, so as to drive the utilization of all the links up to 100%. We would do this to avoid the possibility of a link becoming idle because an idle link necessarily hurts throughput. The problem with this strategy is that increasing the number of packets in the network also increases the length of the queues at each router. Longer queues, in turn, mean packets are delayed longer in the network. To

describe this relationship, some network designers have proposed using the ratio of throughput to delay as a metric for evaluating the effectiveness of a resource allocation scheme. This ratio is sometimes referred to as the power of the network power = throughput / delay.Note that it is not obvious that power is the right metric for judging resource allocation effectiveness. For one thing, the theory behind power is based on an M/M/1 queuing network that assumes infinite queues;[16] real networks have finite buffers and sometimes have to drop packets. For another, power is typically defined relative to a single connection (flow); it is not clear how it extends to multiple, competing connections. Despite these rather severe limitations, however, no alternatives have gained wide acceptance, and so power continues to be used. We provide only this brief description of an M/M/1 queue. The 1 means it has a single server, and the Ms mean that the distribution of both packet arrival and service times is Markovian, that is, exponential. The objective is to maximize this ratio, which is a function of how much load you place on the network. The load, in turn, is set by the resource allocation mechanism.

2.6 The distribution service requests across a group of servers

Load balancing is an important element for the implementation of services brought to grow(fig. 2.3). Its basic principle is to distribute service requests across a group of servers, in intelligent manner. For this, a process of redirecting the tasks depending on the occupancy state of servers is required. Commonly, load balancing systems includes popular web sites, large Internet relay chat networks, high-bandwidth file transfer protocol sites, Network News Transfer Protocol (NNTP) servers, Domain Name System (DNS) servers [and also evolved to support databases. Indeed, network overloads as well as server and application failures often threaten the availability of these applications.

Whereas, they are expected to provide hight performance, hight availability, secure

and scalable solutions to support all applications. Ressource utilization is often out of balance, resulting in the hight-performance ressources remain idle while the low-performance ressources being overloaded with requests. Hence, for overload, performance and availability problems, a load balancing mechanism is a powerfull technique and a widely adopted solution.LTE network is a promising candidate for next generation wireless networks. But like GSM and WCDMA, it still has the problem of load unbalance . Much research has been done to deal with the load unbalance problem.

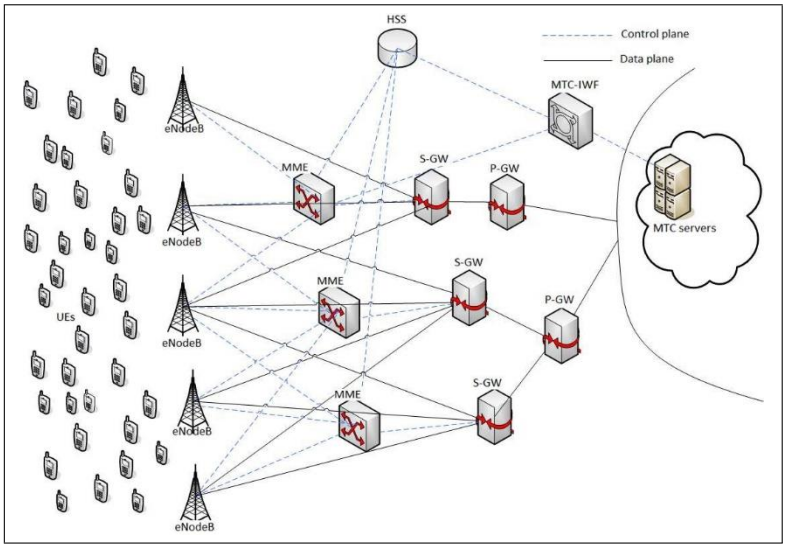

Figure 2.3 - LTE network with more than one MME

2.6.1 Load balancing in networks

Load balancing has many benefits and deals with various requirements that are becomng increasingly important in networks. It is particularly essential for networks that are very busy, it is in fact difficult to determine the number of requests that will be issued to a server. Therefore, the gains are significant:

• Increased scalability, High performance, High availability and disaster recovery.

• Having multiple servers handling many requests, and using a mechanism of load balancing to detect and identify the server that has sufficient availability to receive the traffic improves response time services.

• Load balancing permits to continue ensuring the service which remains available to users even if a server experiences downtime, because the traffic will be routed to an other server, depending on its load, proximity, or health.

• Optimal utilization of servers.

•Ensures that no single server is overwhelmed.

The key feature of a load balancing process is its abaility to direct service requests intelligently to the most appropriate server. We will present various load balancing algorithms, based on different parameters. Load balancing algorithms are classified into two typical approaches, we have static load balancing algorithms and dynamic ones.

2.6.2 Dynamic load balancing algorithms

With dynamic load balancing, changes are made to the distribution of work among servers (or processors) at runtime, they use recent load informations when making distribution decision [18]. Dynamic load balancing algorithms differ from the static ones in the fact that they allocate processes dynamically when one of the processors becomes under loaded .

There are several ways to do load balancing in LTE networks, In [19], they develop a practical algorithm for load balancing among multi-cells in 3GPP LTE networks with heterogeneous services. Its purposes are the load balancing of index of services with QoS requirements and network utility of other services. 3GPP LTE down- link multi-cell network deals with heterogeneous QoS users requirements, namely CBR (Constant Bit Rate) and BE (Best Effort) services.

Here, load balancing which is realized through enforced handover, aims to achieve maximum load balance index for CBR as well as utility function for BE users [18]. This is proposes a heuristic and practical realtime algorithm which could be executed in a distributed manner with low overhead, and could solve this multi-objective optimization problem in a sequential manner. Its objective is that, first, in response to varying network

conditions, each eNodeB in the network makes handover decision quickly and independently and second, to minimize the overhead of user status information exchange for decision making at each eNodeB. They propose a framework that consists of three aspects: QoS-guarantee hybrid scheduling, QoS-aware handover and call admission control.

The first one consists of allocating ressources according to the rate requirements for CBR users before scheduling the remaining resources for BE users to maximize the network utility, because BE users have less QoS requirements than CBR ones. They could use opportunistic scheduling among all CBR users to achieve less ressource occupation for each CBP user, then the ressource allocation depending on the average bandwidth efficiency is conservative. In the case of BE users, they use the proportional fair scheduling in which all BE users have the same log utility function. For the QoS-aware handover, they define a CBR user load balancing gain that permits for the CBR user to switch from cell i to cell j. When many CBR users are about to change their serving cells at the same time (which may result in oscillation of handover), a cell i chooses the best CBR user that achieves the largest benefit by changing its serving cell.Similarly, for a BE user, they define a load balancing gain of BE users, and the BE user that the cell selects is the one that achieves the largest gain because of changing its serving cell.Finally, for the call admission control, when a new CBR user enters the network, the condition of its admission to access a cell i is the availability of enough time-frequency resource to satisfy its QoS demand. However, for new BE user entering the network, there is no constraint on cell access. Other researches deal with load balancing problem in LTE-linked packet switched net- works. They often do not take into account QoS requirements, and use only proportional fairness as the scheduling metric among competing users . Hence, other works include the QoS requirements by proposing a weighted proportional fairness scheduling schemes [20] to reflect the network reality where QoS is required. However, users' QoS requirements cannot be strictly guaranteed by the weighting method. A self-optimizing load balancing algorithm in LTE mobile communication system is proposed.

Its objective is to further improve network efficiency by delivering additional performance gain. This is possible with the use of load balancing in LTE Self- Optimizing Networks (SON), where the parameter tuning is done automatically based on measurements. Here, the basic principle is to adjust the network control parameters in such a way that overloaded cells can offload the excess traffic to lowloaded adjacent cells, whenever available, with the purpose of achieving load balancing. This load balancing algorithm reacts to peaks in load and distributes the load among neighbouring cells to achieve better performance [20]. It aims to find the optimum handover offset value between the overloaded cell and a possible target cell. This aims to reduce the load in the targeted cell by ensuring that the handed over users to the target cell will not be returned to the source cell. In this algorithm for load balancing handover and collect RSRP (Reference Signal Received Power) measurements from UEs to potential TeNB. They group users corresponding to the best TeNB for load balancing in accordance with the difference between TeNB and SeNB (Serving eNB: the eNodeB that serves the previous cell) and obtain information from TeNB on available ressources. Then, they estimate a number of required PRBs (Physical Resource Blocks) after load balancing handover for each user in the load balancing handover group. This is before applying the load balancing procedure.

Subsequently, each SeNB sorts the list of the potential TeNB for each adjusted values of the handover offset, with respect to the number of possible load balancing handovers. Then, a predicted (virtual) cell load after handover is estimated for a given handover offset T and cell C from the list of the potential TeNB. If this predicted load is lower than a certain acceptable threshold at TeNB, load at SeNB is reduced by the amount generated by the users handed over with this offset and handover offset to this cell is adjusted to the T value. This is repeated until the handover offset T is smaller than the maximum alowed value, and the SeNB load is higher than acceptable value threshold.
In short, the main goal of this algorithm is to find the optimum HO offset that allows the maximum number of users to change cell without any rejections by admission control

mechanism at TeNB side. This is done by evaluating the load condition in a cell and the neighbouring cells, and then estimates the impact of changing the handover parameters in order to imporve the overall performance of the network.

2.6.3 The distribution the traffic to the MMEs in LTE networks

The MME load balancing is a functionnality that permits to direct the attach requests of UEs to an appropriate MME, it aims to distribute the traffic to the MMEs according to their respective capacities, so as to perform load balancing, particularly as LTE networks are planned for large deployments, like M2M applications deployment. Very few works deals with functionnality in the case of LTE networks. The only works that exist are those of 3GPP. We have, first, the load balancing between MMEs, where each MME have a Weight Factor (WF) configured, which is also kown as relative MME capacity since it is typically set according to the capacity of the MME itself relative to other MME nodes within the same MME pool [4]. This WF is conveyed to eNodeBs associated with the MME via S1-AP (Application Protocol) messages (see [3]) during initial S1 setup. An eNodeB can communicate with multiple MMEs in a pool, and it decides, based on the WF, to select the MME that can be loaded with attach requests. In fact, the probability of the eNodeB selecting an MME is proportianal to its WF. As illustated in figure 2.4 . Second, the load rebalancing may be needed if an MME needs to be taken out of sevice, or if it feels overloaded. Load rebalancing is a way to simply move UE attaches that are registred to a particular MME to another MME within the MME pool. Indeed, when an MME has been overloaded and cannot handle anymore attach requests, it frees up some ressources, then it releases the S1 and RRC (Radio Ressource Control) connections of the UEs (in ECM-CONNECTED state: a session with an active S1 connection) towards the eNodeBs, while asking UE to perform a "load balancing TAU (Tracking Area Update)". This is transmitted to UEs by eNodeBs in a RRC message.One a UE gets this message, it sends a TAU message to the eNodeB, which in turn routes

the TAU message to another active MME (selected via the MME selection function).

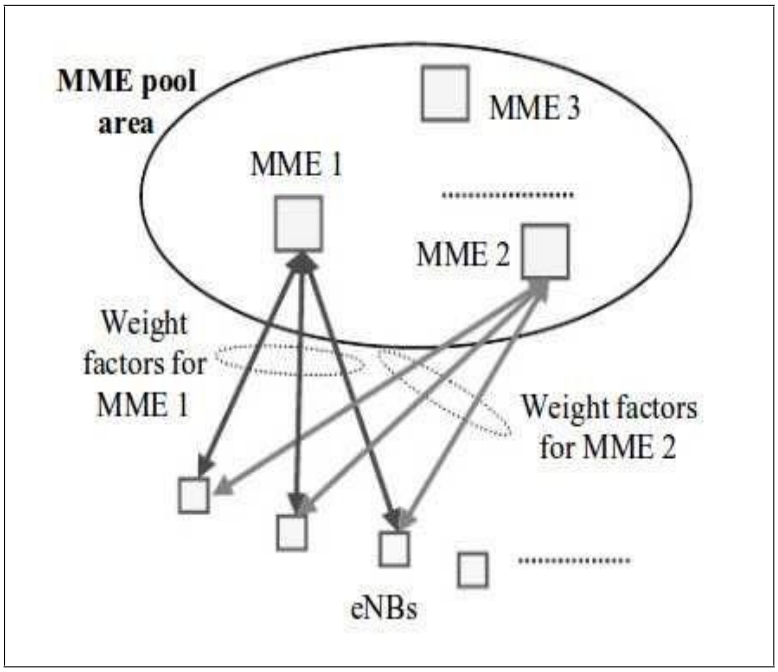

Figure 2.4 - MME load balancing

Therefore, the MME that is overloaded can move calls to an active MME that is not overloaded, and this, after pulling the UE context of the overloaded MME (via the S10 interface that connects the two MMEs) [22].

Using DNS at eNodeBs is another way to realize load balancing. In fact, the UE will populate GUMMEI 3 (Globally Unique Mobility Management Entity Identifier) to the eNodeB in an RRC message. Then, the eNodeB sends a DNS query to obtain MME information, based on the GUMMEI, and forward the UE message to the selected MME. If this last is not responding, then the eNodeB may forward the call to the next available MME in the pool.

3. THE OVERLOAD METRICS FOR ANALYTICALLY EVALUATION OVERLOAD OF M2M TRAFFIC

3.1 Queueing theory in resource planning problems . Formulation of queueing models

Queues are common in a communication networksystems. Are queues of inquiries waiting to be processed by an interactive computer system, queue of data base requests, queues of requests, etc. Typically a queue (or queuing system) has one service facility, although there may be more than one server in the service facility, and a waiting room (or buffer) of finite or infinite capacity. Customers from source enter a queuing system to receive some service. Here customer is thus maybe a packet in a communication network, a job or a program in a computer system, a request or an inquiry in a database system, etc. Upon arrival a customer joins the waiting buffer if all servers in the service center are busy. When a customer has been served, he leaves the queuing system. A special notation, called Kendall's notation, is used to describe a queuing system.

The notation has the form A/B/c/K, where:

- A describes the interarrival time distribution
- B the service time distribution
- c the number of servers
- K the size of the system capacity (including the servers).

The symbols traditionally used for A and B are:

- M for exponential distribution (M stands for Markov)
- D for deterministic distribution
- G (or GI) for general distribution.

When the system capacity is infinite ($K = \infty$) one simply uses the symbol A/B/c. M/M/1, M/M/c, M/G/1 and G/M/1 are very common queueing systems. For a single server queueing system, ρ denotes the traffic intensity [23] which is defined by ρ - mean service

time/ mean inter-arrival times mean service time mean inter-arrival times Assuming an infinite population system with arrival intensity λ, which is reciprocal of the mean inter-arrival time, let the mean service time be denoted by 1/μ, then we have

$$\rho = \text{arrival intensity} \times \text{mean service time} = \lambda/\mu$$

If ρ >1, then the system is overloaded since the requests arrive faster than they are served. It shows that more servers are needed. Let χ(A) denotes the characteristic function of the event A, that is

$$\chi(A) = 1 \quad \text{if A occurs}$$
$$\chi(A) = 0 \quad \text{if A does not}$$

Furthermore, let N(t) = 0 denotes the event that at time T the server is idle having no customers in the system. Therefore, utilization of the server during time T is defined by

$$1/T \int_{T\,0} \chi(N(t) \neq 0)\, dt\,,$$

where: T - a long interval of time. As $T \rightarrow \infty$, we get the utilization of the server denoted by Us and the following relations holds with probability

$$Us = \lim_{T\rightarrow\infty} 1/T \int_{o}^{T} \chi(N(t) \neq 0)dt = 1 - P_0 = E_\delta /(E_\delta + E_i) \qquad (3.1)$$

where: P_0 - the steady-state probability that the server is idle. Eδ and Ei denote the mean busy period and mean idle period of the server respectively.

3.2 Brief description of M/M/1 queue model

In this queueing system the customers arrive according to a Poisson process with rate λ. The time it takes to serve every customer is an exponential r.v. with parameter μ.

For it is model we say that the customers have exponential service times. The service times are supposed to be mutually independent and further independent of the interarrival times. When a customer enters an empty system his service starts at once; if the system is nonempty the incoming customer joins the queue. When a service completion occurs, a customer from the queue if any, enters the service facility at once to get served.

Let X(t) be the number of customers in the system at time t.

The process (X(t), $t \geq 0$) is a birth and death process with birth rate $\lambda i = \lambda$ for all $i \geq 0$ and with death rate $\mu i = \mu$ for all $i \geq 1$. (X(t), $t \geq 0$) is a Markov process. the probability of having two events (departures, arrivals) in the interval of time (t,t + h) is o(h)

$P(X(t + h) = i + 1 \mid X(t) = i) = \lambda h + o(h)\ \forall i \geq 0$

$P(X(t + h) = i - 1 \mid X(t) = i) = \mu h + o(h)\ \forall i \geq 1$

$P(X(t + h) = i \mid X(t) = i) = 1 - (\lambda + \mu) h + o(h)\ \forall i \geq 1$

$P(X(t + h) = i \mid X(t) = i) = 1 - \lambda h + o(h)$ for $i = 0$

$P(X(t + h) = j \mid X(t) = i) = o(h)$ for $|j - i| \geq 2$.

This shows that (X(t), $t \geq 0$) is a birth and death process. Let $\pi(i)$, $i \geq 0$, be the d.f. of the number of customers in the system in steady-state. The balance equations for this birth and death process read

$$\lambda\, \pi(0) = \mu\, \pi(1)$$

$$(\lambda + \mu)\, \pi(i) = \lambda\, \pi(i - 1) + \mu\, \pi(i + 1)\ \forall i \geq 1.$$

$$\rho = \lambda / \mu \tag{3.2}$$

The quantity ρ is referred to as the traffic intensity since it gives the mean quantity of work brought to the system per unit of time.

Stationary queue-length d.f. of an M/M/1 queue. If $\rho < 1$ then

$$\pi(i) = (1 - \rho)\, \rho\, i \tag{3.3}$$

for all $i \geq 0$.

Therefore, the stability condition $\rho < 1$ simply says that the system is stable if the work that is brought to the system per unit of time is strictly smaller than the processing rate (which is 1 here since there is only one server).

From now on ρ will always be defined as λ/μ unless otherwise mentione. Result therefore says that the d.f. of the queue-length in steady-state is a geometric distribution. From (3.3) we can compute (in particular) the mean number of customers E[X] (still in steady-state). We find

$$E[X] = \rho\,(1 - \rho) \tag{3.4}$$

Observe that $E[X] \to \infty$ when $\rho \to 1$, so that, in pratice if the system is not stable, then the queue will explode. It is also worth observing that the queue will empty infinitely many times when the system is stable since

$$\pi(0) = (1 - \rho) > 0.$$

We may also be interested in the probability that the queue exceeds, say, K customers, in steady-state. From (3.1-3.2) we have

$$P(X \geq K) = \rho\, K \tag{3.5}$$

The throughput T of an M/M/1 in equilibrium is $T = \lambda$.

3.3 Queuing theory for studying networks

Definition of queueing networks A queueing network is a system composed of several interconnected stations, each with a queue. Customers, upon the completion of their service

at a station, moves to another station for additional service or leave the system according some routing rules (deterministic or probabilitic) [22].

In many applications, an arrival has to pass through a series of queues arranged in a network structure (fig. 3.1-3.3).

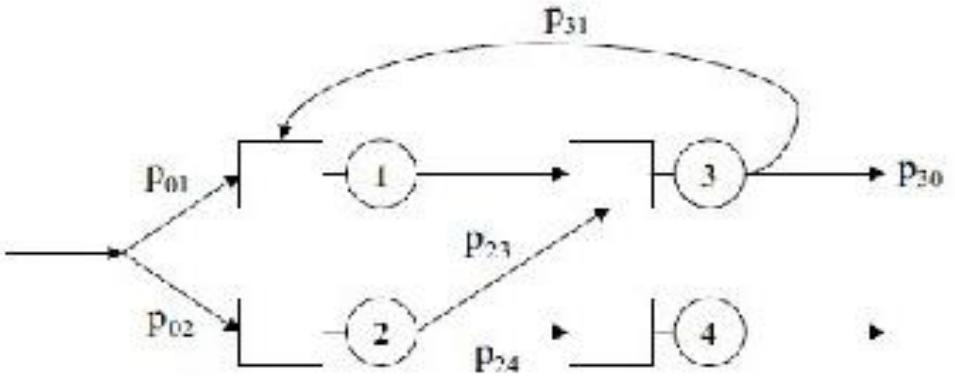

Figure 3.1- Queuing networks structure

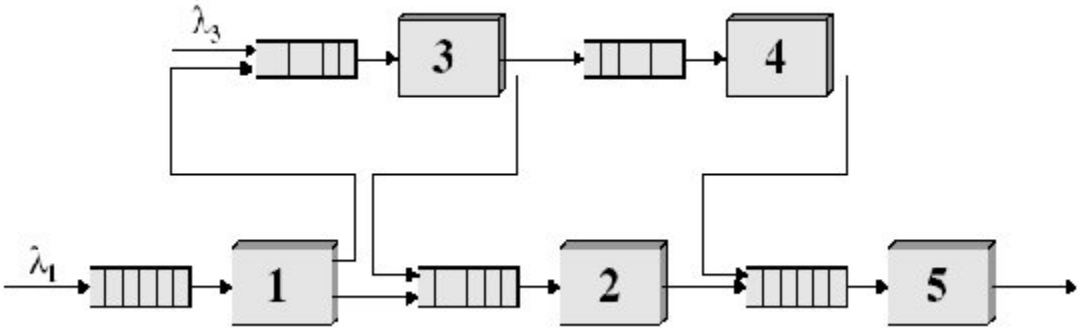

Figure 3.2- The queuing model of queuing system queue n server A

Queue system is characterized by

-Queue (buffer): with a finite or infinite size the state of the system is described by the queue size

-Server: with a given processing speed

- Events: arrival (birth) or departure (death) with given rates. For open queueing networks number of jobs in the system varies with time throughput = arrival rate .

Goal - to characterize the distribution of number of jobs in the system.

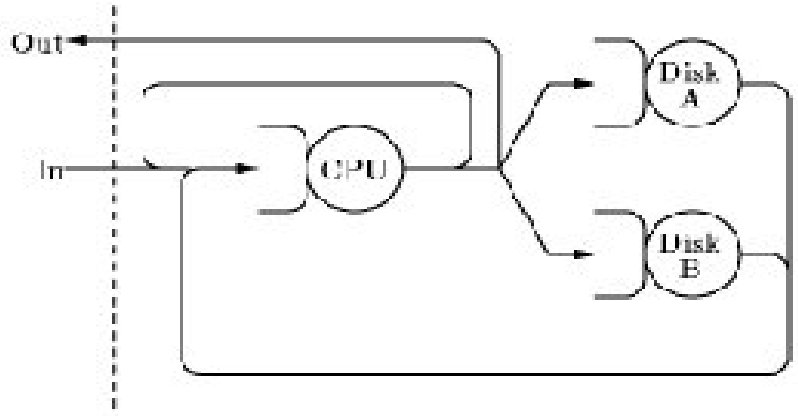

Figure 3.3- Open queueing network: external arrivals and departures

Characteristics of queuing systems:

Arrival process - the distribution that determines how the tasks arrives in the system.

Service process- the distribution that determines the task processing time .

Number of servers - total number of servers available to process the tasks.

Storage capacity- are buffer finite or infinite.

Specification of queueing systems operating policies:

Customer class differentiation - are all customers treated the same or do some have priority over others.

Scheduling/queueing policies - which customer is served next,

Admission policies - which/when customers are admitted.

A continuous-time Markov chain (CTMC) model is utilized to formulate the RRM scheme that supports two different types of service requests, namely HTC and MTC. The proposed RRM scheme allows to analyze the performance of MTC and HTC integration over the air interface. In its design, the radio resources are distributed between MTC traffic, HTC traffic, and an area shared by both the MTC and HTC traffic. Two thresholds are setup to distribute the radio channels in a predefined way. The system will accept the HTC or MTC service request as long as their dedicated radio channels are available. Once the HTC or MTC service requests that are already in the system have reached their

predefined thresholds, the system will forward any new incoming HTC or MTC request to the shared area if it has some resources available. Otherwise, the incoming HTC or MTC incoming service request will be rejected.

The value characterizing the congestion will be the length of the queue at the IP level. To estimate overload and failure rate, we will find overload probabilities and failure probabilities on the MME. To calculate such probabilities, we will use the tools of queuing theory. We will investigate the service of requests arriving at the node in the form of a system of queues M/M/1/K, which corresponds to the behavior of MME in our system. The QS has one server and a buffer (k-1 is the maximum number of waiting positions in the MME queue) with a FIFO service discipline [24,25]. On fig. 3.4 shows the state diagram for the M/M/1/K queue model, and fig. Figure 3.5 shows a simple M/M/1/K queue network.

In this queuing system, packets arrive in a Poisson process at a rate of λ (the average number of packets arriving per unit time), and the service time depends exponentially on the parameter μ. A new incoming packet is lost when it detects that the system is full. In fact, at the beginning, the traffic that the UE sends to the eNodeB does not follow a Poisson distribution. However, the aggregation of all traffic sent by the UE is distributed according to a Poisson process at a rate of λ.Its analytical performance evaluation provides a mathematical framework that can help us understand and predict the behavior of our system, as well as obtain its characteristics to describe its performance.

For the first-in-first-out (FIFO) link scheduling discipline packets arriving to the link output queue are queued for transmission if the link is currently busy transmitting another packet.

The state diagram has exactly K states provided that $c<K$.

The state diagram (c service technicians and N machines)

λ -arrival intensity per operating machine

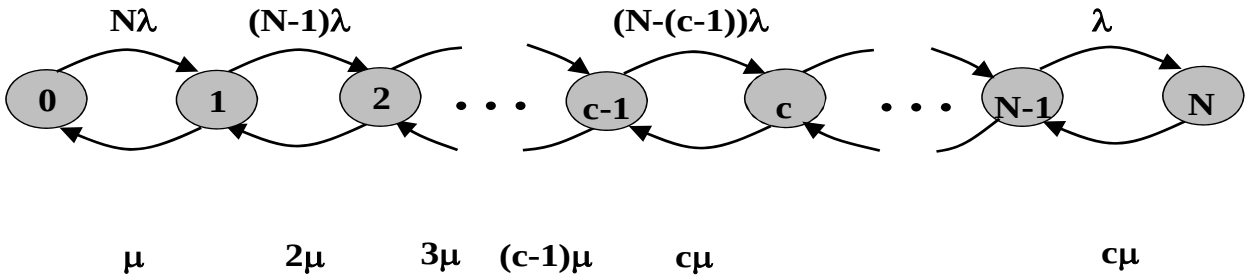

Figure 3.4 - M/M/c/K,N state transition diagram

μ - the service intensity for a service technician

An M/M/c model with limited calling population, i.e., N clients/Acommon application: machine maintenance;

-c service technicians is responsible for keeping N service stations (machines) running, that is, to repair them as soon as they break;

-customer/job arrivals = machine breakdowns.

Note, the maximum number of clients in the system = N.

Assume that (N-n) machines are operating and the time until breakdown for each machine i, T_i, is exponentially distributed $T_i \in \exp(\lambda)$. If U = the time until the next breakdown $\Rightarrow U = \min\{T_1, T_2, \ldots, T_{N-n}\} \Rightarrow U \in \exp((N-n)\lambda))$ [26].

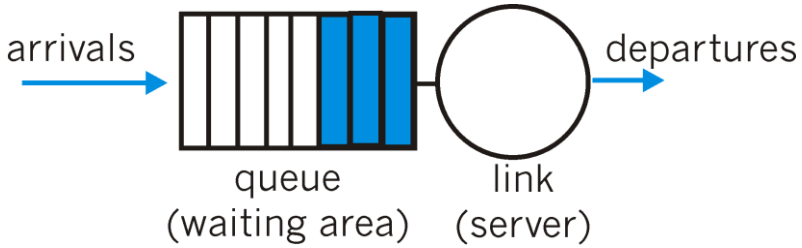

Figure 3.5- Simple M/M/1/K queueing node

If there is not sufficient buffering space to hold the arriving packet, the queue's packet discarding policy then determines whether the packet will be dropped ("lost") or whether other packets will be removed from the queue to make space for the arriving packet. Figure 3.6 shows an example of the FIFO queue

in operation.

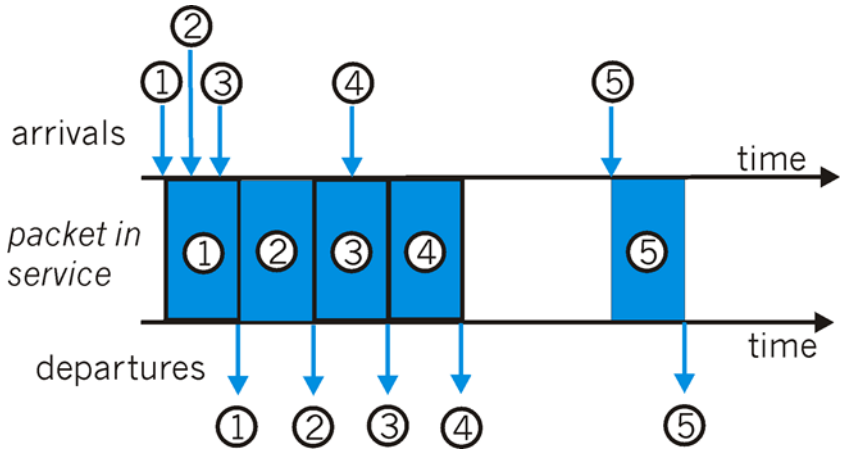

Figure 3.6- The first-in-first-out (FIFO) model queue in operation

Packet arrivals are indicated by numbered arrows above the upper timeline, with the number indicating the order in which the packet arrived. Individual packet departures are shown below the lower timeline.The time that a packet spends in service (being transmitted) is indicated by the shaded rectangle between the two timelines. Because of the FIFO discipline, packets leave in the same order in which they arrived. Note that after the departure of packet 4, the link remains idle (since packets 1 through 4 have been transmitted and removed from the queue) until the arrival of packet 5.

The overall functioning of dynamic MME load balancing and admission control algorithm At first, we aim to balance the load among all MMEs. This is achieved by performing the load balancing at the level of each eNodeB separately. The load balancing is handled by a probabilistic routing strategy at each eNodeB that is connected to more than one MME, following the inverse congestion probabilities of MMEs (after calculating the proprtions and the cumulative probabilities). The eNodeB retrieves these congestion probabilities from its MMEs. Also is defined a congestion probability threshold which is the limit that indicates us that the MME will be congested if more packets are forwarded to it. When all

MMEs are about to be congested, which means that there is no way to avoid congestion by balancing the load among MMEs, we move to the second step, which consists of the traffic rejection. Rejecting signaling traffic occurs when MME load balancing is no longer sufficient to deal with congesiton. It is done following the reject probabilities. It rejects the amout of traffic that can cause congestion, in order to maintain the number of packets in the MME queue much smaller than the maximum MME queue length.

Rejecting traffic is done at the eNodeB that have triggered the rejection((the one who was making the load balancing among its MMEs) as well as at the eNodeBs that share their single MMEs with the eNodeB that triggered the rejection, if they are believed to be the cause of congestion. in other terms, if an enodeb have its reject probability higher than the reject probability of the enodeb that triggered the rejection.

Analysis of queues requires defining some performance measures, and there are many possible measures of performance for queueing systems, and some of them are: probability of the number of customers in the system, probability of waiting for service, average quantity ρ equal λ gives the system load, it is referred to as traffic intensity. For M/M/c systems, there is a stability condistion $\rho < c$ that simply shows that the system is stable if the work that is brought to the system is strictl smaller than the processing rate. However, in M/M/1/k the system is always stable, even if $\rho \geq 1$, therefore, there is no stability condition .

Pn = The probability that there are exactly n customers/jobs in the system (in steady state, i.e., when t)

L= Expected number of customers in the system (in steady state)

L_q = Expected number of customers in the queue (in steady state)

W = Expected time a job spends in the system

W_q= Expected time a job spends in the queue

An M/M/c model with a maximum of K customers/jobs allowed in the system

If the system is full when a job arrives it is denied entrance to the system and the queue.

4 MATHEMATICAL MODELLING OF ALGORITHM THE CONTROL DISTRIBUTION OF M2M TRAFFIC

4.1 Customer's servicing arriving at the MME node

Real networks have finite buffers and sometimes have to drop packets. The load, in turn, is set by the resource allocation mechanism.

4.1.1 The M/M/1/K model

In practice, queues are always finite. In that case, a new customer is lost when he finds the system full (e.g., telephone calls). The M/M/1/K may accomodate at most K customers, including the customer in the service facility, if any. To describe queue models, we use the QS theory [24-25]. Let λ and μ be the rate of the Poisson process for the arrivals and the parameter of the exponential distribution for the service times, respectively.

Let $\pi(i)$, $i = 0, 1, \ldots, K$, be the d.f. of the queue-length in steady-state. The balance equations for this birth and death process read

$$\lambda\, \pi(0) = \mu\, \pi(1)$$
$$(\lambda + \mu)\, \pi(i) = \lambda\, \pi(i-1) + \mu\, \pi(i+1)$$
$$\text{for } i = 1, 2 \ldots, K-1 \quad \lambda\, \pi(K-1) = \mu\, \pi(K).$$

Stationary queue-length d.f. in an M/M/1/K queue .

If $\rho = 1$ then

$$\pi(i) = (1-\rho)\, \rho\, i\, (1-\rho)(K+1) \tag{4.1}$$

for $i = 0, 1, \ldots, K$, $\pi(i) = 0$ for $i > K$.

If $\rho = 1$ then

$$\pi(i) = 1/(K+1) \tag{4.2}$$

for $i = 0, 1, \ldots, K$, and $\pi(i) = 0$ for $i > K$.

$(X(t), t \geq 0)$,

where: X(t) is the number of customers in the system at time t, can be modeled as a birth and death process with birth rates $\lambda i = \lambda$ for $i = 0, 1, \ldots, K - 1$ and $\lambda i = 0$ for $i \geq K$. In particular, the probability that an incoming customer is rejected is $\pi(K)$.

4.1 2 The overload probability and reject probabilities

In our case, the congestion metric is the queue length. It means that the probability of congestion is the probability of having the number of packets (attach request) in the queue greater or equal a certain H - the overload detection threshold. Hence, we define the congestion probability as follows.

M2M traffic after service at the base station arrives at one of the MME nodes attached to the base station.

Let's examine a part of the LTE network with *J* base stations and *K* mobility management nodes. To describe the model, we introduce 2 sets - a set of base stations *J={eNB1,..., eNBj}* and a set of mobility management nodes *K={MME1,...,MMEk}*
We define additional sets *Jk-set* of base stations from which the load on the MMEk node comes *k=1,...,K* and *Kj* - a set of mobility management nodes that receive traffic from base stations *eNBj, j=1,...,J.*

Optimization of traffic management, which takes from the original template at the mobility management nodes, is possible only when a non-universal mobility management console is connected to the base station, i.e. $Kj \geq 2$.

The distribution of traffic coming from base stations to mobility management nodes occurs with a probability strategy. Each MME node at given intervals measures its load level and sends this data to all base stations from the set *Jk, k=1*, ..., *K.* Each base station eNBj , receiving data on the load of mobility control nodes *Kj j=1,...,J* redistributes traffic flows going for further service to one of the nodes of the set *Kj*.

Let us determine the influence of time intervals through which the network load data is corrected.

For a preliminary assessment and the possibility of applying (the QS model), we

study a simplified case when a Poisson flow M2M arrives at the MME node from *eNBj* the base station with intensity λ_k , the service time at the *MMEjk* node is distributed exponentially with the parameter μ_k . The probability of overloading *Pov[k]* the *MMEk* node can be found using the methods of the mathematical Queueing theory. The service of requests arriving at the *MMEk* node is examined in the form of Queueing theory M/M/1/r,H , $r<\infty$, H-threshold value in the buffer.

Let *X(t)* - the number of customers in the QS at time t, $t \geq 0$. The space X of a random process *X(t)* has the form $X=\{0,\ldots,r+1\}$ .A random process $\{X(t),t\geq 0\}$. is a Markov process.

Let there be n customers in the QS, $n \subset X$ at the moment t .

Let p_n - the probability that there are exactly *n* requests $n \subset X$ in the system, the total load on the MME node, is equal, $\rho = \lambda_k / \mu_k$, λ_k -the total intensity of requests arriving at the *MMEk* node, formula

$$\text{If } \rho \neq 1,\ \text{Pov}[k]= \rho^n(1- \rho)/ (1- \rho^{r+2})$$
$$\text{If } \rho=1,\ \text{Pov}[k]=1/(r+2) . \qquad (4.3)$$

Assume that all MMEk nodes are identical. We define the probability of overloading the MMEk *Pov[k]* node as formula

$$\text{Pov}[k]= P\{n\geq H\} =\sum p_n ,\ \ n=H\div r+1,\quad 1\leq H\leq r, \qquad (4.4)$$

where: H - the overload detection threshold.

Using formulas (4.3) and (4.4), we obtain the probability of overload in the form .

$$\text{Pov[k] for}\quad \text{If } \rho \neq 1,\ \text{Pov}[k]= \rho(1- \rho^{r-H+2})/ (1- \rho^{r+2})$$
$$\text{If } \rho=1,\ \text{Pov}[k]=(r+2-H)/(r+2) \qquad (4.5)$$

Let us introduce the probability of overloading the MME_k node *Pov[j,k]* with the

load that came only from *j* the base station *eNBj*, *j=1,…,J, k=1,…,K.*

$$\text{Pov[j,k] for} \quad \text{If} \quad \rho \neq 1, \ \text{Pov[j,k]} = \rho^{n}(1-\rho^{r-H+2})/(1-\rho^{r+2})$$
$$\text{If} \quad \rho=1, \quad \text{Pov[j.k]} = (r+2-H)/(r+2) \qquad (4.6)$$

where: $\rho = \lambda_{jk} / \mu_k$ takes into account only the load coming only from the base station *j, eNBj, j=1,…,J.*

P_{rej} -probabilitie reject to service the customers is defined as follows

$$P_{rej} = \rho^{n} \cdot (1-\rho)/(1-\rho^{r+1})$$

The congestion of the MME mobility control node is determined by the condition *Pov thres.*

$$\text{Pov[k]} \geq \text{Pov thres} \qquad (4.7)$$

comparing the overload probability with a predetermined threshold value (4.7).

4.2 Controlling algorithm the distribution of M2M traffic base stations network LTE to MME nodes

1. We study a case when a Poisson flow M2M arrives at the MME node from eNBj the base station with intensity λk , the service time at the MMEjk node is distributed exponentially. The distribution of traffic coming from base stations to mobility management nodes occurs with a probability strategy.

Regardless of how many requests enter the input of the serving system, this system (queue and clients being served) cannot accommodate more than N-requirements (requests), i.e., clients that are not waiting are forced to be served elsewhere. The source that generates service requests has an unlimited (infinitely large) capacity.

2. M2M traffic after service at the base station arrives at one of the MME nodes attached to the base station. The probability of congestion is the probability of having the

number of packets (attach request) in the queue greater or equal a certain the overload detection threshold. We examine a part of the LTE network with J base stations and K mobility management nodes .We have 2 sets - a set of base stations J={eNB1,…, eNBj} and a set of mobility management nodes K={MME1,…,MMEk}. Traffic between base stations is not balanced. A threshold value for the probability of overload has been introduced *Pov thres.*

3. The control distribution of M2M traffic which takes from base station to at the mobility management nodes, is possible only when a nodes mobility management is connected to the base station for Kj ≥ 2.

4. Each base station eNB_1 sends a node load request set at *Kj , j=1,…,J* . For K ≥2 go to step 2, otherwise, for base station eNB_1, optimization of traffic flows is impossible and go to the other base station.

5. Each MME node at given intervals measures its load level and sends this data to all base stations from the set Jk, k=1, ..., K. Each base station eNBj , receiving data on the load of mobility control nodes Kj j=1,…,J redistributes traffic flows going for further service to one of the nodes of the set Kj. The base station eNB_1 receives data on the workload of each node of the set *Kj , j=1,…,J.* redistributes traffic flows going for further service to one of the nodes of the set Kj.

6. Redistribution of traffic directed from the base station eNB_1 to the mobility management nodes getes from the set K *Kj , j=1,…,J.*according to the proportional probability, *Pprop[j,k].*

$$Pprop[j,k]= (1- Pov[j,k])/ \sum (1- Pov[j,k]), \text{ for set } j \subset Kj. \quad (4.8)$$

i.e. until the next correction time, the base station eNB_1 will receive *M2Mk* traffic with intensity λ_j x *Pprop[j,k]* ,

where λ_j is the intensity of the M2M traffic leaving the base station eNB_1.

7. If all nodes from the set Kj , j=1,…,J. are overloaded, then the maximum value of the busiest MME node is found. *max Pov[j,k]* for the busiest MMEk node.

The probability of overloading *Pov[k]* the MMEk node can be found using the methods of the mathematical Queueing theory. The busiest line between the base stations from the set J and the MMEk node is found by probability *Pov[j,k] , j⊂Jk* .The maximum value *max Pov[j,k]* is selected and a message is sent to the base station eNB1 about the refusal to receive traffic for the current time period until the next load redistribution.

Average number of customers in the queue (queue length):

$$L_q = \lambda(1-P_{rej})\,[Ls/(\lambda(1-P_{rej}))-1/\mu]$$

Average number of customers in the system

If $\rho \neq 1$, $Ls = \rho[1-(N+1)\,\rho^N + N\rho^{N+2}]/(1-\rho)(1-\rho^{N+2})$

If $\rho=1$, $Ls=N/2$,

for If $\rho \neq 1$, $Pov[k]=1-[(1-\rho)(\rho^N)/(1-\rho^{N+2})$

If $\rho=1$, $Pov[k] = 1/(N+2)$

If all nodes from the set *Kj , j=1,…,J.* are overloaded, then the maximum value of the busiest MME node is found. *max Pov[j,k]* for the busiest *MMEk* node. The busiest line between the base stations from the set*J* and the *MMEk* node is found by probability *Pov[j,k] , j⊂J_k* .The maximum value *max Pov[j,k]* is selected and a message is sent to the base station *eNB_1* about the refusal to receive traffic for the current time period until the next load redistribution.

4.3 An analysis of the numerical results

In this model *M/M/1/r,H* of queuing system the arrival distribution of customers follows Poisson distribution and the distribution for service time follows exponential distribution with number of parallel servers. number of population and queuing capacity is limited to K. This situatin ofher happens in queuing for machine repair system where the number of population is equal to the number of machine is K.

The performance for M/M/c/*K/h* queuing system are given by below.

Arrival Rate (number of customers/ unit time) – λ.

Service rate (number of customers/ unit time) -μ.

Number of servers -c.

Capacity of the system equal capacity is limited population - K.

Maximum queue size equal capacity without number of servers.

Arrival Rate (number of customers/ unit time) – λ=210 .

Service rate (number of customers/ unit time) -μ=30.

Capacity of the system equal capacity is limited population - K=500.

Maximum queue size K-c.

Table 4.1 – The performance parameters

Characteristic	c=7	c=8	c=14
Number of servers	7	8	14
Queuing intensity	7000	7000	7000
Queuing utilization	99.799 %	87.5 %	50 %
Queue length in queue	245.003	4.447	0.014
Queue length in system	251.989	11.447	7.014
Delay in queue	1.169	0.021	0.000
Delay in system	1.202	0.055	0.033
Probability of idle server	0.001 %	0.056%	0.091%

We will show the example of computation on the offered algorithm.

For the network segment, we will calculate and evaluate the workload of the MME nodes. Consider machine-to-machine communication traffic coming from the base stations eNBs of the LTE network to the MME mobility management entity. We calculate a case when machine-to-machine communication traffic, after being serviced at the base station, arrives at one of the 5 MME nodes attached to the base station. K=5 W=7. The condition

$Kj \geq 2$ has been met. The flow of claims arriving at the MME is distributed according to the Poisson law and has an intensity $\lambda = 0.88$ (claims per unit of time). The operating time of the MME is distributed according to the exponential law and on average is equal to 1.02 units time. r=6, H=3 buffer threshold.

Determine the probabilistic characteristics of the MME operating in the single server mode, using the formulas of paragraphs 3.1-3.2 and 4.1.- 4.2.

1.Average service intensity for the system when there are n customers/jobs in it. (the total service intensity for all occupied servers)

$\mu = 1/t = 0.98$

2.The utilization factor for the service facility(the expected fraction of the time that the service facility is being used)

$\rho = \lambda / \mu = 0.898$

3. The probability that at time t, there are n customers/jobs in the system $\rho^k \cdot (1- \rho)/ (1- \rho^{K+1})$

$P_0 = (1- \rho)/ (1- \rho^{K+1}) = 0.103$

$P_1 = \rho \cdot P_0 = 0.898 \cdot 0.103 = 0.092$

$P_2 = \rho^2 \cdot P_0 = 0.898^2 \cdot 0.103 = 0.083$

$P_3 = \rho^3 \cdot P_0 = 0.898^3 \cdot 0.103 = 0.066$

$P_4 = \rho^4 \cdot P_0 = 0.898^4 \cdot 0.103 = 0.059$

$P_5 = \rho^5 \cdot P_0 = 0.898^5 \cdot 0.103 = 0.054$

$P_6 = \rho^6 \cdot P_0 = 0.898^6 \cdot 0.103 = 0.048$

4. Probabilitie reject to service the customers

$P_{rej} = P_6 * P_0$

$P_{rej} = 0.048 * 0.103 = 0.0049$

5. Relative and absolute bandwidth

$q = 1- P_{rej} = 1- 0.0049 = 0.995$, $A = \lambda \cdot q = 0.875$

6.Average number of customers spends in the system (in waiting line and being served)

$Ls = \rho[1- (K+1) \rho^K + K \rho^{K+1}] / [(1- \rho)(1- \rho^{K+1}] = 0.898[1-7*0.522+6*0.469]/[(1-0.898)(1-$

0.469)] = 0.152/0.055=2.76

7. Ws= Ls / [λ (1- P_N)= 2.76/0.88(1-0.048) =2.76/0.837=3.29 unit time

8. $W_q=W_S-1/\mu$ = 2.473-1/0.98 =2.27

9. Lq- average number of customer in waiting line for service(queue length):

$L_q=\lambda(1-P_6)\cdot W_q$ = 0.88 (1-0.048) 2.27=1.9

10. The probability of overloading Pov[k] the MME_k node

Pov[k]= Pov =ρ^H (1- ρ^{r-H+2})/(1- ρ^{r+2}) = 0.898^3 ($1-0.898^{5}$) /$1-0.898^8$ = 0.72X(1- 0.578)/1-0.41=0.72 X 0.422/0.59=0.303/0.59 =0.515= 51.5 %

11.According to the step 6 of the M2M traffic control algorithm, the traffic is redistributed according to the formula (4.8). If all K MME nodes are overloaded, the step 7 is performed.

The work of the considered MME node can not be considered satisfactory, since the MME does not serve applications on average in 51.5% of cases. Therefore, it is necessary to apply the algorithm for redistributing M2M traffic between MME nodes. Average number of customers spends in the system (in waiting line and being served) 2.76 . Average number of customer in waiting line for service is 1.9 . Average time a customers spends in the system (in waiting line and being served) 3.29 unit time.

The solution for this type of QS can be obtained both in analytical form and with the help of a simulation model, which will be implemented using sets of visual blocks SimEvents. To determine the congestion of the MME mobility control node, which is determined by the condition *Pov[k]≥ Pov thres.* a comparison of the overload probability with a given threshold value Pov thres has been introduced. If this condition is met, the MME node will be overloaded and it is necessary to redistribute traffic between the MME nodes.

4.4 The model description in discret-event simulation

4.4.1 A queueing systems with mathematical models

Mathematical modeling of the queueing theory is one of the branches of applied mathematics which studies and models the waiting lines.Is used statistical approach for

distributions which can be applied to the situationswhere excessive demands are to be fulfilled on a limited resource. Poisson's formula was meant only for the repeated callers presented a paper that opened a general review of some points in congestion theory to enhance the study for a single server queue where input is Poisson and service time is generally distributed. He further extended the study of the stochastic processes for the theory of queues and their analysis by the method of the imbedded Markov chain. The study was carried out first reviewing on single-server queues and using the similar technique to the analysis of many server queuing system.

Stochastic process is a key factor to specify in queuing systems because it describes the arrival pattern as well as the structure and the discipline of the service facility [27]. Queueing system deals with queue length and waiting times. The concept of queue is applied not only in the waiting system by the human beings but also in modern technology of computer and other service providers by the devices. In general, it is not necessary that service will be immediately available to address the demand of all the customers, so that they are forced to line up. In the queueing system, the one who demands the service is referred as customer, which may be a person, a task or a commodity. The other element of the queueing system is the one who provides the service with some defined discipline, called the server. It may be people, machine or objects. Some of the service disciplines are first come first served (FCFS), last come first served (LCFS), service in random order (SIRO), priority, processor sharing (PS), round-robin (RR). We study performance measures of a queueing system where only the limited number of customers are served and arrival or service or both occur in a batch. If any of the customers come after the prescribed quota has already been served, the server does not provide the service to the new comer.. The models we investigate have important applications in the study oftele-traffic, computer and flexible manufacturing systems, production processes, transportation, monitoring, controlling and managing complex engineering systems that have finite buffer system.

4.4.2 The description the components of systems queueing and used models in queueing

To build simulation models in the Matlab + Simulink environment, the SimEvents library is provided, with the help of which tools you can design and simulate random dynamic systems with continuous and discrete components with discrete events and discrete time, which include distributed control systems, hardware configurations, collection networks and information transfer, etc[28,29].

To implement complex models, sometimes it becomes necessary to use other major libraries of the Simulink graphical language, such as Math Operations (mathematical operations), Signals (signals), Ports & Subsystems (ports and subsystems) and others. we will consider the main components of the SimEvents library, which are basic for the implementation of a simulation model of queuing systems .

The main concept of discrete event-based simulation:

- entity – entity (order);
- event - an instant discrete event that changes the state and / or cause of other events.

The SimEvents Matlab library includes the following block tool libraries (fig. 4.1-4.2)

- Attributes - definition of attributes (parameters) of entities;
- Event Translation - converting an event signal into one or more functions;
- Generators - library of order generators;
- Queues - library of queues;
- Servers - library of services (channels);
- SimEvents Ports and Subsystems - a library of ports and subsystems;
- SimEvents User Defined Funct - definition of entity attributes using a function;
- Entity Management - flow management (unification, distribution) of entities;
- Gates - flow control depending on the conditions imposed on the incoming component;
- Probes – counter of initial entities with a record to the port and/or attributes;
- Routing - blocks of switches and flow control;

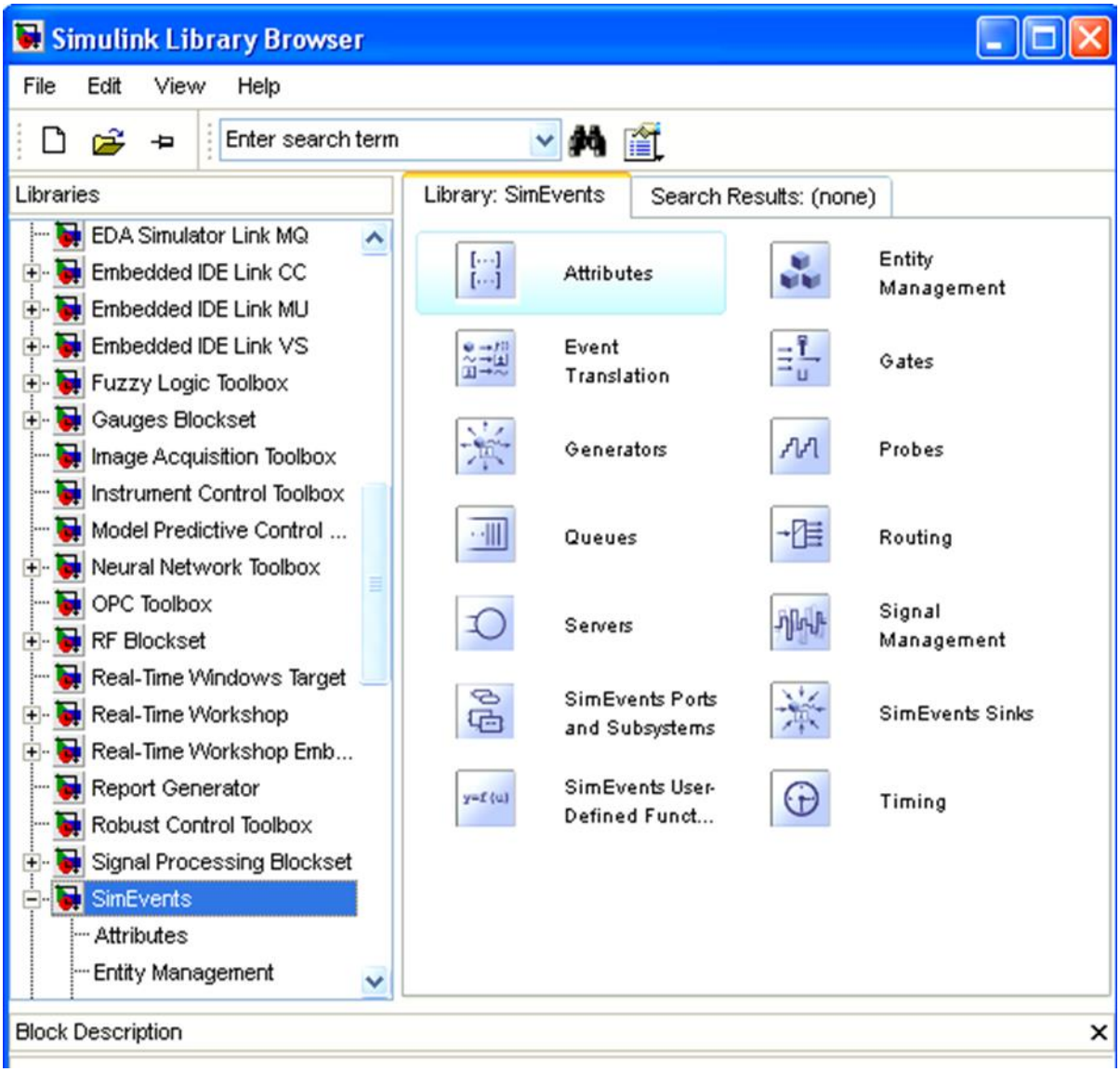

Figure 4.1- Components of the SimEvents library

- Signal Management - management of signals that determine events;
- SimEvents Sinks - blocks of absorption of orders and graphical presentation of results;
- Timing - time control blocks.

When modeling discrete events, entities can move through networks of queues (queues), servers (servers) and switches (switches). The graphic blocks of the SimEvents library represent a set of components that process entities, but the entities themselves do not have a graphical representation.

Modern versions of the Mftlab environment include an extended set of SimEvents library components, which allows you to create complex simulation models of queueing

networks, but often only such basic blocks as generators, queues and services are enough to develop a simple model. To model the queue, the blocks shown in figure 4.3 are used differing in the order (discipline) of queue processing.

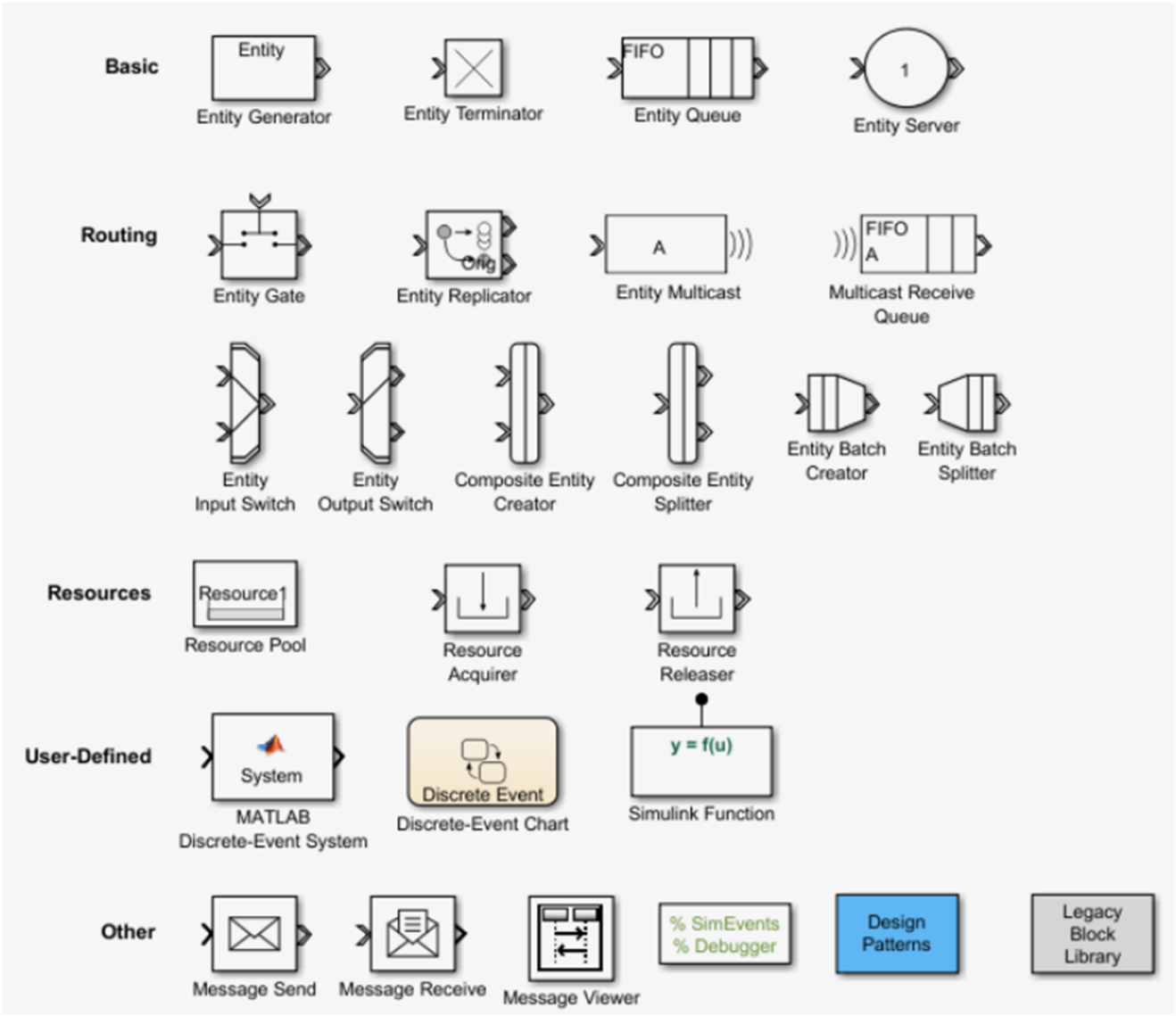

Figure 4.2- SimEvents library 5.1

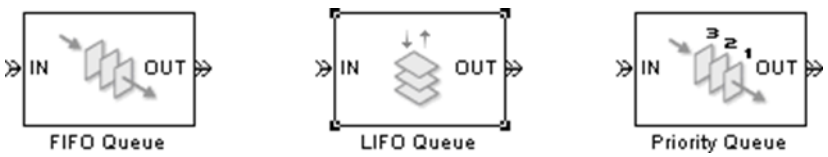

Figure 4.3- Order of queue processing

- FIFO (first input first ounrun) - a queue serviced on a first-in-first-out basis;
- LIFO (last input first output) - a queue serviced by the stack principle - first in - last out;
- Priority Queue – priority queue.

The blocks have the same structure and the same setting fields. Only the Priority Queue block contains two additional fields on the first tab, where the priority discipline and sorting order are determined (fig. 4.4).

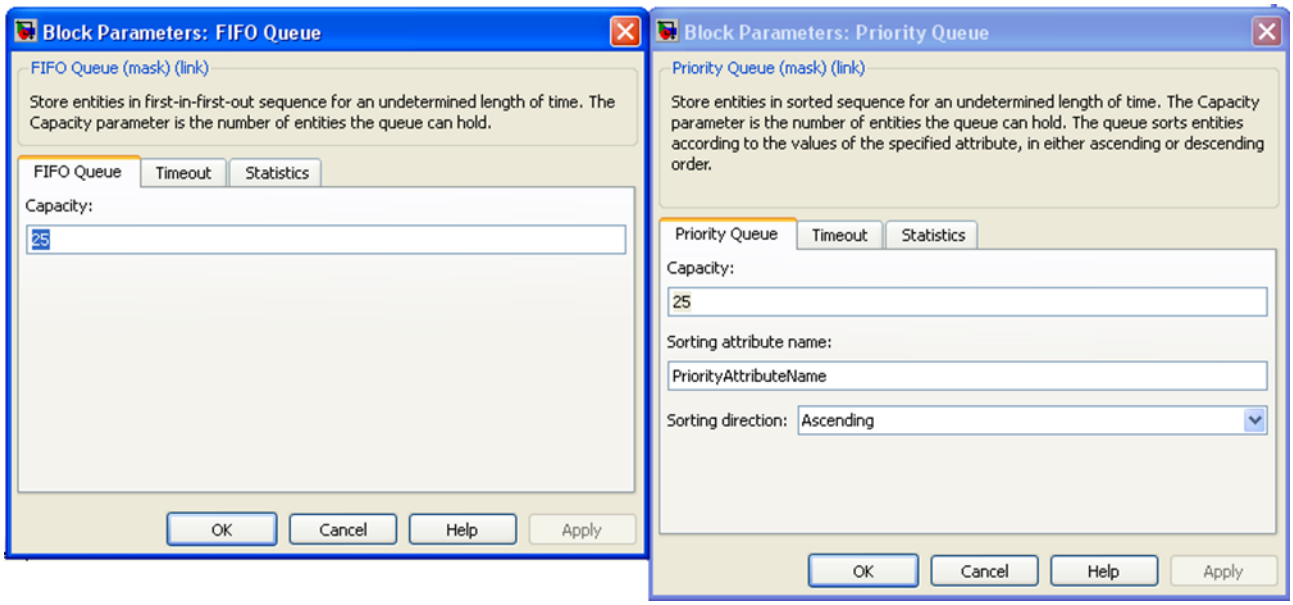

Figure 4.4- Main tab of the block queue dialog box

The statistics settings tab is shown in figure 4.5 and containing the following Fields. To configure blocks, you need to set the length of the capacity queue - a natural number.

If all places in the queue are occupied, the input port IN is unavailable and the order is rejected, if the output port is blocked (for example, all services are busy), then the order remains in the block. The block skips the order according to the discipline of the queue.

The second tab Timtout contains one field Enable TO port for timed-out entities, the selection of which allows to make the TO (timed-out) port available. This field is relevant in the case when it is supposed to limit the time the order is in the queue. If the order's time in the queue expires, the order exits the block.

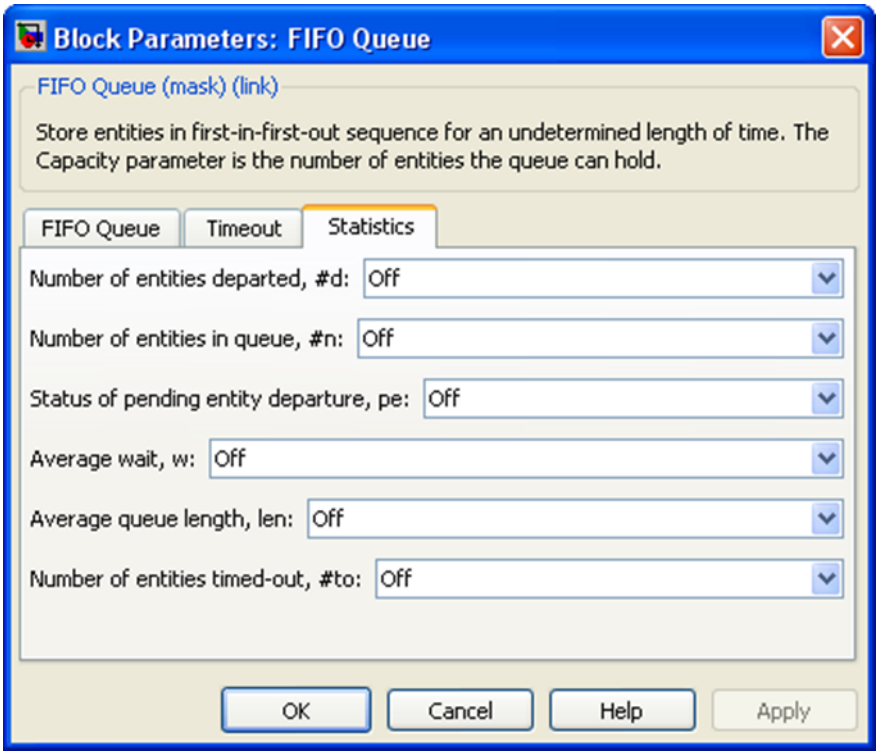

Figure 4.5- Setting the queue block statistics

- Number of entities departed, #d – number of orders that left through the output port OUT after the start of simulation;

- number of objects in queue, ; #n is the number of orders in the queue;

- Status of pending entity departure pe - checks for the presence of a signal output port labeled pe.

- Average wait, w – average waiting time in the queue;

- Average queue length, len – average queue length;

- Number of entities timed out, #to - controls the presence and behavior of the output port.

4.4.3 Attributes of to entities

The use of attributes allows you to manage priority queues, time-limited events (timeout), and entity flows.

The library contains two graphic blocks (fig. 4.6):

- Get Attribute (take attributes) - the block allows you to determine the attributes of

the entity and their values for further use, leaving them unchanged (fig. 4.7);

- Set Attribute (set attributes) - the block allows you to assign attributes to entities (fig.4.8). Each attribute has its own name and value.

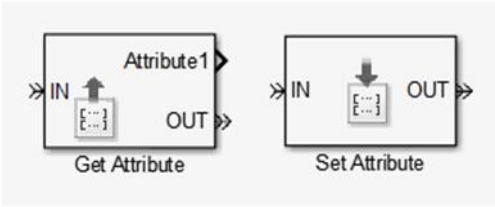

Figure 4.6 - Attribute control blocks

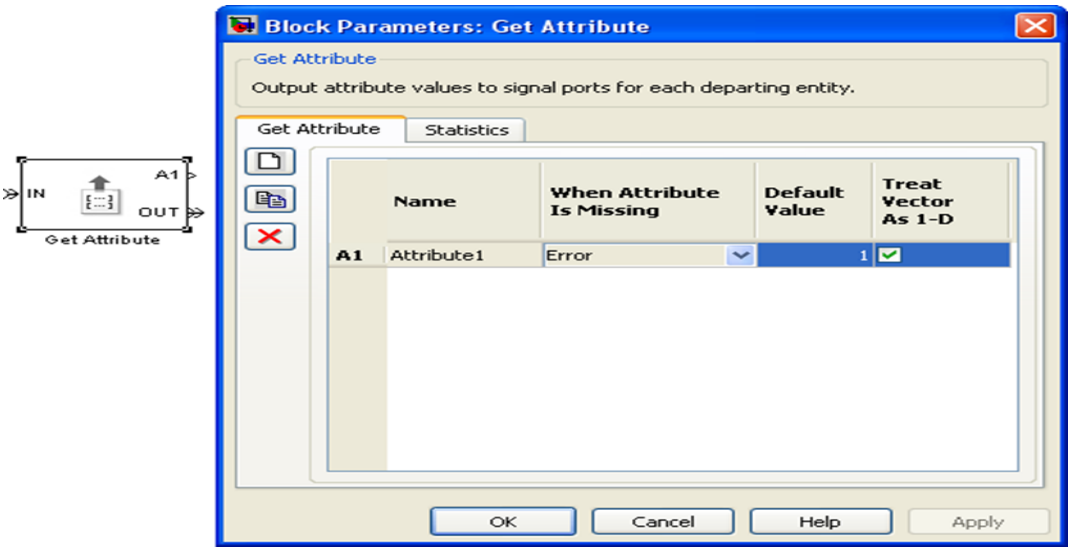

Figure 4.7 - Get Attribute block and its dialog box

The blocks panel has the following setting keys:

– create an attribute field; – copy attribute field; – delete attribute field and two tabs:

- statistics.

- setting attribute parameters;

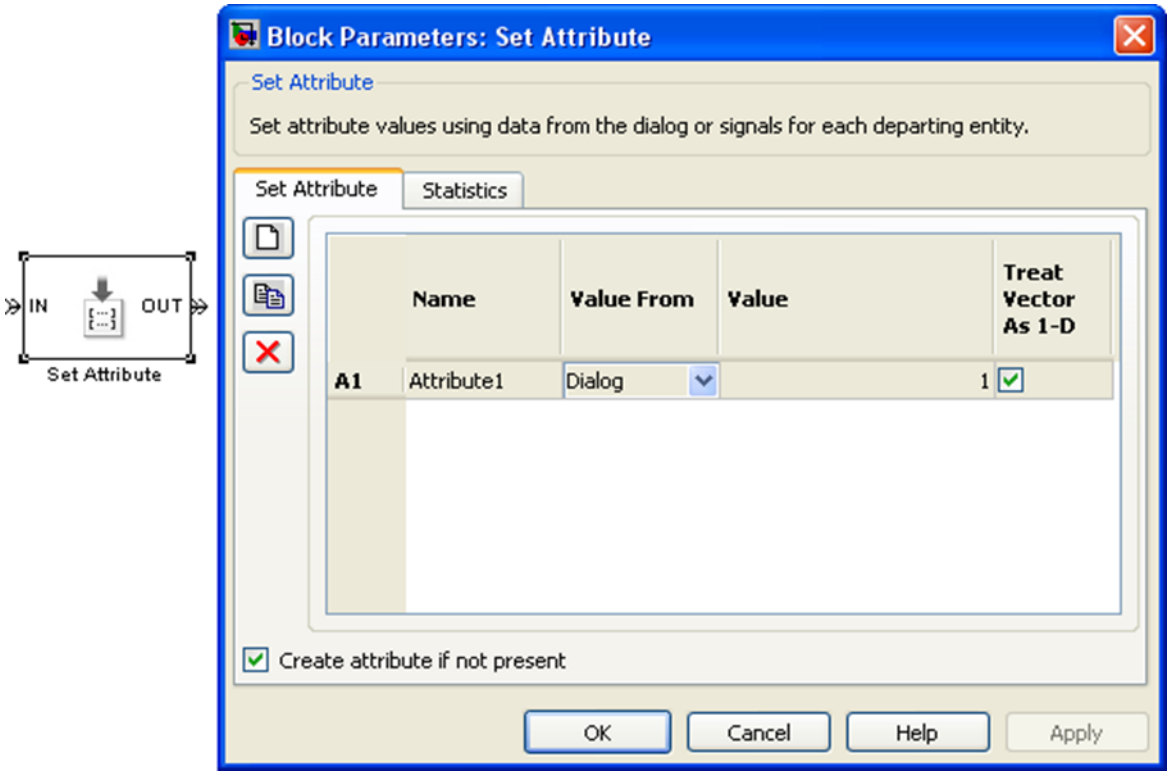

Figure 4.8- Set Attribute block and its dialog box

4.4.4 Servers blocks

The library contains three blocks that simulate QS services.

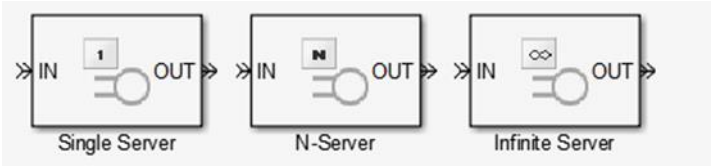

Figure 4.9- Server blocks

- Single Server - a single service;
- N-Server - N-services, where the number N is determined by the user;
- Infinite Server - an endless number of services.

The blocks have a similar structure, the Single Server block, and for the remaining

blocks, we note only the features of their implementation and options.

The structure of the M/M/1/K type queuing system model for implementing the model using blocks SimEvents Matlab R2015b is given in the Appendix.

The developed model of controlling algorithm the distribution of M2M traffic can be used on the stage of planning. During modernization of existent network or planning of new it can be required to estimate its functioning still to the purchase of equipment, software . The modeling enables to the developer of the system to experiment with the system (existing or supposed) in those case, when to do it on the real object inexpedient or it is impossible.

CONCLUSIONS

In the thesis, devices connected to the Internet are considered, which are all kinds of sensors that exchange information automatically without human intervention. In this is an actively developing segment of the Internet of Things (M2M, Machine-to-Machine to refer to this type of communication.) cases of emergency situations are described. and instantaneous operation of a large number of sensors, in which there is an explosive growth of traffic entering the service. In this case, overloads may occur in various elements of the network. The paper considers the maintenance of machine-to-machine traffic, which starts at base stations associated with mobility management entities (MMEs). When the mobility management nodes are overloaded, it is proposed to redirect traffic to other MME nodes in a timely manner, thereby unloading the corresponding network elements.

An algorithm for distributing traffic between mobility control nodes is given for a simplified model of a network segment serving machine-to-machine traffic, . Shows traffic control when network elements are overloaded. using a probabilistic traffic transmission strategy from each base station connected to multiple MME nodes and an overload threshold indicating the workload of the mobility management node.

The QS theory is used, which made it possible to obtain a mathematical model for describing the algorithm for the distribution of machine-to-machine traffic in an analytical form, numerical examples are given.

Analyzing the simulation capabilities of SimEvents toolbox in accordance to the requirements to model the presented application we can conclude that this type of modeling is suitable and has all the capabilities to simulate this tasks.

On results the model dependences between the parameters of networks and functional characterizing its properties can be built . Are explored the tasks related to the choice of optimum parameters and construction of rational strategy of management are decided and etc. A block diagram of the simulation model is given, which will be implemented using sets of visual blocks SimEvents.

The developed model of controlling algorithm is also the distribution of M2M traffic can be easy subject of programming in any comfortable for an user environment.

REFERENCES

1. Chen and S. Lien, "Machine-to-Machine Communications: Technologies and Challenges," *Ad Hoc Networks* , vol. 18, July 2014, pp. 3-23.
2. A. Malm and T. Ryberg, “Wireless M2M and Mobile Broadband Services”, Berg Insight, Feb. 2007.
3. 3GPP, “System Improvements for Machine Type Communications”, TS 22.368 V10.1.0, Jun. 2010.
4. 3GPP, “Service Requirements for Machine-Type Communications”, TS 22.368 V10.1.0, Jun. 2010.
5. F. Ghavimi and H. H. Chen, “M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications,” IEEE Commun. Surv. Tutorials, vol. 17, no. 2, pp. 525–549, 2015
6. Lucero,H “Maximizing Mobile Operator Opportunities in M2M: The Benefits of an M2M-Optimized Network”, ABI research, 1Q 2010
7. Ding, J., Nemati, M., Ranaweera, C. and Choi, J. (2020) IoT Connectivity Technologies and Applications: A Survey. IEEE Access, 8, 67646-67673. https://doi.org/10.1109/ACCESS.2020.2985932
8. T. Taleb and A. Kunz, “Machine Type Communications in 3GPP Networks: Potential, Challenges, and Solutions”, to appear in IEEE Communications Magazine.
9. 5. T. Kim, K. S. Ko, and D. K. Sung, “Prioritized Random Access for Accommodating M2M and H2H Communications in Cellular Networks,” in 2015 IEEE Globecom Workshops (GC Wkshps), pp. 1–6, Dec 2015
10. 3GPP TS 22.368 v10.1.0 (2010-06). Service requirements for machine-type communications (mtc), June 2010.
11. Ahmadian A., Galinina O.S., Gudkova I.A., Andreev S.D., Shorgin S.Ya., and Samouylov K.E. On capturing spatial diversity of joint M2M/H2H dynamic uplink transmissions in 3GPP LTE cellular system // Lecture Notes in Computer Science. – 2015. – Vol. 9247. – P. 407–421.

12. X. Jian, X.-P. Zeng, Y.-J. Jia, J.-Y. Yang, and Y. He, “Traffic modeling for machine type communication and its overload control,” Tongxin Xuebao/Journal Commun., vol. 34, no. 9, pp. 123–131, 2013

13. Jihun Moon Yujin Lim Adaptive Access Class Barring for Machine-Type Communications in LTE-A Conference: 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN) July 2016 https:// DOI:10.1109/ICUFN.2016.7537058

14. Edemacu, Kennedy, Bulega, Tonny Resource sharing between M2M and H2H traffic under time-controlled scheduling scheme in LTE networks 2014 8th International Conference on Telecommunication Systems Services and Applications (TSSA) Identifiers https:// DOI 10.1109/TSSA.2014.7065909

15. 3GPP TS 23.401 V8.12.0 (2010-12). General packet radio service (gprs) enhancements for evolved universal terrestrial radio access network (e-utran) access, December 2010

16. Samouylov K., Sopin E., Vikhrova O. Analyzing Blocking Probability in LTE Wireless Network via Queuing System with Finite Amount of Resources // Communications in Computer and Information .

17. Naumov V., Samouylov K., Yarkina N., Sopin E., Andreev S., and Samuylov A. LTE performance analysis using queuing systems with finite resources and random requirements // Proc. of the 7th International Congress on Ultra Modern Telecommunications and Control Systems ICUMT-2015 (October 6–8, 2015, Brno, Czech Republic). – USA, New Jersey, Piscataway, IEEE. – 2015. – P. 100–103.

18. Suman Das, Harish Viswanathan, and Gee Rittenhouse. Dynamic load balancing through coordinated scheduling in packet data systems. In INFOCOM, 2003.

19. Sherihan Abu Elenin and Masato Kitakami. Performance analysis of static load balancing in grid. International Journal of Electrical and Computer Sciences IJECSIJENS, 11(3):170–177, June 2011.

20. Zhang Lin1, Li Xiao-ping2, and Su Yuan2. A content-based dynamic load-balancing

algorithm for heterogeneous web server cluster. ComSIS, 7(1), February 2010.

21. William Leinberger, George Karypis, and Vipin Kumar. Load balancing across nearhomogeneous multi-resource servers. In Heterogeneous Computing Workshop, pages 60–71, 2000.

22. Taleb T. and K. Samdanis, "Ensuring Service Resilience in the EPS: MME Failure Restoration Case", in Proc. IEEE Globecom 2011, Houston, USA, Dec. 2010.

23 Ivo Adan and Jacques Resing. Queueing Theory. Department of Mathematics and Computing Science Eindhoven University of Technology P.O. Box 513, 5600 MB Eindhoven, The Netherlands, February 2002.

24 Philippe NAIN. BASIC ELEMENTS OF QUEUEING THEORY Application to the Modelling of Computer Systems. University of Massachusetts, Amherst, MA, January 1998.

25 Eitan Altman and Alain Jean-Marie. The loss process of messages in an m/m/1/k queue. In INFOCOM, pages 1191–1198, 1994

26.Shanmugasundaram, S. and Banumathi, P. (2016). A simulation study on M/M/C queueing models, International Journal for Research in Mathematics and Mathematical Sciences, 2(2), 52-61

27.Banks, Jerry, John Carlson, and Barry Nelson. Discrete-Event System Simulation, Second Ed. Upper Saddle River, N.J.: Prentice-Hall, 1996.

28.Cassandras, Christos G., and Stéphane Lafortune. Introduction to Discrete Event Systems. Boston: Kluwer Academic Publishers, 1999.

29. Discrete-Event Simulation: Modeling, Programming, and Analysis (Springer Series in Operations Research and Financial Engineering) Sep 21, 2011.