Downlink Resource Allocation in LTE

Aditya Thakkar
The University of Texas at Dallas
Email: aditya.thakkar@utdallas.edu

Rashmi Gottipatti
The University of Texas at Dallas
Email: rashmi.gottipati@utdallas.edu

Dr. Ravi Prakash
The University of Texas at Dallas
Email: http://www.utdallas.edu/ ravip/contact.html

***Abstract*—Advances in Internet technology is giving way to the development of a new generation of communication network such as cloud computing where content providers can store information and host applications on cloud. Applications hosted on cloud can be accessed by a user over the Internet using personal computing devices such as personal computer and laptop, and mobile devices such as smartphones and tablets. Personal computing devices are connected to a modem using wired or wireless link (in case of WiFi), which is in turn connected to the Internet through a wired link; whereas mobile devices are connected to the Internet via wireless mobile network. In the latter scenario where connection between a mobile device and mobile network is wireless, mobile network has to transmit data from Internet to the mobile device through wireless channel using radio frequencies, also known as radio resources. Radio resources are limited in nature, and need to be efficiently managed by the mobile network in order to serve maximum number of users accessing Internet based applications. Moreover, next generation mobile network, termed as Long Term Evolution (LTE) supports diverse traffic types such as VoIP, video streaming, real-time gaming, and web-browsing. Each traffic type has different Quality of Service (QOS) requirement. For example, VoIP traffic requires tight delay bound compare to web (HTTP) traffic. In addition, HTTP traffic has smaller packet loss rate requirement compare to VoIP traffic. This paper conducts detailed survey on downlink (downward transmission from mobile network to user) scheduling algorithms proposed in literature that schedules and allocates resources to users ensuring that the LTE standards and QoS requirements are obeyed.**

TERMINOLOGY

Following terminologies and abbreviations are widely used in this paper:

- *Channel Quality Indicator (CQI)* - Value sent by user device to the base station. It describes the quality of channel between base station and user device.
- *QoS Class Indicator (QCI)* - Describes the Quality of Service (QoS) specification as standardized in LTE specifications.
- *eNodeB* - Base station in LTE is known as eNodeB.
- *User Equipment (UE)* - User Device such as a smartphone.
- *QoS* - Quality of Service.
- *Radio Resource Manager (RRM)* - Component in LTE, that is responsible for radio resource management. Scheduler is part of the RRM.
- *Hybrid Automatic Repeat Request (HARQ)* - Component in RRM that is responsible of managing re-transmission of packets.
- *Link Adaptation (LA)* - Component in RRM that uses CQI to estimate instantaneous throughput of the user.
- *Forward Error Correction Code (FEC)* - Sequence of bits used to recover sub-set of errorneous bits in transmitted data.

I. INTRODUCTION

A Mobile network is composed of Radio Access Network (RAN) and the Core Network as illustrated in Figure 1. It can be observed from the figure, Core Network is connected to PSTN network and Internet backbone that provides telephony and Internet services respectively to the user; RAN, on the other hand, interfaces with the user. In addition, connection between RAN and core network is wired, and interface between the user and RAN is wireless. That is, user and RAN communicate over wireless radio frequencies - also referred to as wireless transmission channels, using radio waves. Simplified architecture of mobile network displayed in Figure 1 can be correlated to the computing system which encompasses hardware, operating system, and applications, which in turn corresponds to radio resources, RAN and users respectively in a mobile network system. In a computing system with a single processor the applications cannot execute concurrently. Thus, an operating system employs a scheduler to perform context switching between applications, assign required hardware resources and allow the application to execute for a pre-defined time interval. Similarly, due to limited availability of radio frequencies, RAN uses a scheduler to schedule and allocate radio resources to users for communication over a wireless channel. For example, when user wants to send an HTTP request to an Internet based application, the scheduler in RAN needs to allocate radio resources to the user so that it can transmit the request in upward direction to RAN, which will then be delivered to an Internet based application over the wired backbone. Similarly, if a user wants to receive data from Internet based application, for example, accessing a web-page for the request it previously sent, RAN receives the data from a web-application over the wired network, allocates resources for the user destined to receive web-page data and transmits it in a downward direction to the user.

A. Resource Allocation in Mobile Network

Resource allocation and scheduling decisions depend on the resource access scheme utilized and architecture of RAN. For example, in a 2G network system, RAN consists of base stations that are connected to a central entity called the Base Station Controller (BSC). In addition 2G mobile network utilizes Time Domain Multiple Access (TDMA) resource

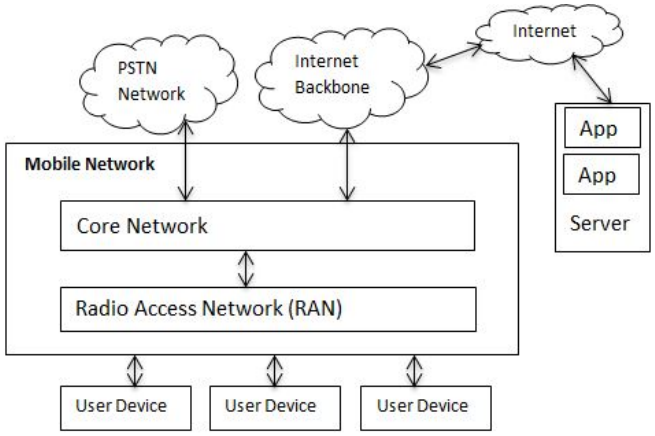

Fig. 1. Simplified view of interaction between user and mobile network and its services

access scheme wherein time is divided into slots and only one user is allowed to transmit during each time slot [1]. BSC in 2G system is responsible for managing radio resources by efficiently allocating time slots to the users. Similar to 2G, RAN in 3G system consists of base stations and a central entity called Radio Network Controller (RNC). Like BSC, RNC is responsible for managing radio resources on behalf of the base stations. However, unlike 2G system, which uses TDMA access scheme, 3G system utilizes Code Division Multiple Access (CDMA) access scheme wherein, in a given wide frequency channel, each user is assigned a different code such that each transmission results in the usage of a different portion of the radio frequency [2]. In other words, CDMA access scheme enables simultaneous transmission over single communication channel or radio frequency. The goal of RNC then becomes to choose codes that maximizes the number of simultaneous transmissions over single radio frequency. With the respective resource access scheme and RAN architecture, 2G and 3G networks are capable of providing 9.6 kbps and 2 mbps of data transmission rates respectively. In addition, simultaneous transmissions over single radio frequency enables 3G networks to provide a 6 fold increase in system capacity compared to 2G mobile network [2]. However, due to a surge in cloud applications, demand for cellular data is on the rise, especially data demand in downward direction due to variety of data driven multimedia applications such as live/buffered video streaming and gaming. As per studies conducted by Cisco, the mobile data traffic is expected to grow to 11.2 exabytes per month in 2017, with mobile cloud traffic driving 84% of the total mobile traffic compared to 0.9 exabytes of mobile traffic in 2012 with 74% coming from cloud applications [3]. In other words, resource management schemes used in 2G and 3G technology cannot meet the ongoing growth in demand for cellular data. With this in mind, next evolution of mobile network called Long Term Evolution (LTE) is standardized in 3GPP Release Specification Version 8 that aims to achieve a high data rate of 100 Mbps in downward direction and 50 Mbps in upward direction to meet future data communication demands [4].

B. Long Term Evolution (LTE) Overview

RAN in LTE network consists of only base stations, known as eNodeB. Unlike base stations in 2G and 3G networks, eNodeBs are connected to each other in a mesh and are responsible for radio resource management. The distributed nature of RAN architecture eliminates the need for process-intensive and high-availability of a centralized controller, which in turn reduces cost and avoids single-point of failure issues that persists in a centralized system. In addition, LTE uses Orthogonal Frequency Division Multiplexing (OFDM) resource access scheme to schedule multiple simultaneous transmissions over the same radio frequency [4]. After purchasing the radio frequency, LTE allows network operators to split the frequency into allowed channel bandwidth of size in the range of 1.4 MHz, 3MHz, 10 MHz, 15 MHz and 20 MHz. OFDM further splits channel bandwidth into 12 sub-carriers, each 15 KHz wide. OFDM also splits the time into frames of size 10 ms consisting of 10 sub-frames of size 1 ms - sub-frames are also referred to as Transmission Time Interval (TTI). Scheduler in the radio resource component at eNodeB then allocates resources at every TTI for upward and downward transmission in the form of Resource Block (RB). RB is defined as 180 Khz $(15KHzsubcarrier * 12subcarriers)$ frequency chunk that lasts for 0.5 ms slot duration. Short RB sizes allow the scheduler to schedule multiple simultaneous transmissions in a distributed manner i.e. it is free to assign RB from different frequency range to each data packet. Moreover, short RB sizes also increase reliability. For example, assume voice traffic was transmitted over single radio frequency, if there is interference in the frequency bandwidth used for transmission of voice data, packet loss incurred would be higher, reducing the reliability of transmission. On the other hand, if different RBs were assigned to voice data i.e. voice packets were transmitted on a different frequency, if one portion of the frequency range is experiencing interference, only the packets transmitted over RBs in that frequency range may be delivered with error or lost. This in turn increases reliability as the number of re-transmissions for the lost or erroneous packets are decreased. Moreover, due to diversity in traffic types, Quality of Service (QoS) requirements for each traffic is different. For example, a VoIP application needs tight delay bound of 100 ms compared to web-browsing or an FTP file transfer, which tolerate upto 300 ms of delay. On the other hand, to ensure quality of experience to the end-user, web-browsing can bear maximum packet loss rate of 0.00001% compared to VoIP traffic that can bear maximum packet loss upto 0.1%. Hence, LTE has standardized QoS requirements for various traffic types as described in QCI Class table in Appendix.

Because of changes in traffic diversity from 2G and 3G network to LTE, the scheduler in LTE needs to ensure it fulfills LTE service requirements and QoS requirement of

various traffic types as defined in QCI Classes is met. This paper conducts detailed survey on the scheduling algorithms proposed in literature that schedules and allocates resources to users in downlink or downward direction to ensure LTE requirements and QoS standards are met.

II. Motivation

One of the key objectives of LTE is to use common packet based infrastructure for all services i.e. voice and data [4]. In addition, because of diversity of traffic types in LTE network, each traffic type has different QoS requirements. A user in LTE is associated with set of virtual queues known as bearers, where each bearer holds traffic based on QoS characteristics of the traffic. In other words, bearers are used to differentiate traffic types based on QoS characteristics. For example, VoIP traffic and HTTP traffic have different QoS characteristics (as defined in QCI Class table in Appendix) - VoIP has a stringent delay requirement and HTTP traffic requires lower packet loss. Thus, VoIP traffic and HTTP traffic is associated with a different bearer. On the other hand, traffic from the HTTP protocol and FTP protocol have the same QoS requirements and are thus associated with the same bearers.

Figure 2 illustrates the life-cycle of the IP packet as it enters the mobile network, in this case the LTE network. When IP traffic enters a Core Network (CN) component in

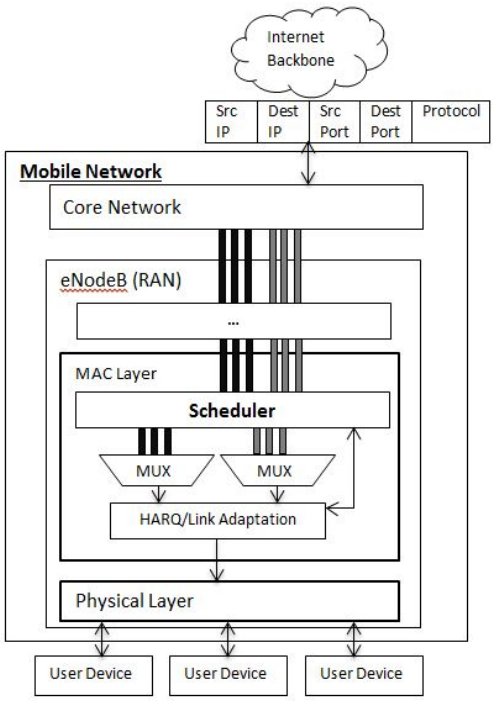

Fig. 2. Illustrates life-cycle of packet in LTE

LTE, CN filters out traffic for different users based on the destination IP; it differentiates the traffic type based on the source IP, source port and protocol type; inserts the packet in the bearer whose QoS characteristics are appropriate for the incoming traffic type; if there are no bearers whose QoS characteristics are pertinent with the incoming traffic type, it establishes a new bearer for the incoming traffic and assigns QoS characteristics to the bearer as defined in QCI Class. In other words, CN separates incoming traffic based on its QoS characteristics. For example, consider VoIP, HTTP and FTP to be the three incoming traffic types that arrive to the LTE network sequentially. When VoIP traffic first arrives, a new bearer is established and assigned QoS characteristics associated with VoIP traffic. When HTTP traffic arrives, it compares the QoS characteristics of the bearer associated with VoIP traffic to that of HTTP traffic. Since there is no match, it establishes a new bearer. When FTP traffic arrives, its QoS characteristics will match with that of bearer associated with HTTP traffic, hence, FTP traffic will be associated with the existing bearer associated with HTTP traffic. In addition, it can be observed in the figure, each user has a set of bearers associated with it - the set of black bearers belong to one user whereas the grey belongs to another user. The bearers from the Core Network enter eNodeB and before they enter the MAC layer for scheduling, they pass through various set of layers that provide services such as packet header compression, security, sequence numbering of packets, re-ordering of packets and duplication detection. The main goal of the MAC layer is to manage scheduling priorities among multiple bearers from multiple users, and multiplex bearers into single transport stream. To achieve this, MAC layer employs Radio Resource Manager (RRM) that consists of Scheduler, HARQ manager and a Link Adaptation (LA) module that closely co-operate with each other [5]. Scheduler is an entity that is responsible for ranking the bearers as per their QoS characteristics and assigning Resource Blocks or RBs to the bearers as per the calculated rank. HARQ manager manages re-transmission of past errorneous deliveries by providing format of the pending re-transmission. HARQ manager learns about the failure of a delivery through link layer acknowledgments. If it receives a Nack, it knows that delivery of packet was unsuccessful. On the other hand, Link Adaptation module is divided into an Outer-loop LA and a Inner-loop LA. Inner-loop LA receives a transmission channel quality report from the user in the form of Channel Quality Index (CQI) and uses this to provide physical layer with an estimate of the supported data rate for the RB assigned to the user. Based on this estimate, physical layer translates the assigned RB to a user into signals with appropriate strength. The scheduler also uses this estimated throughput in its ranking calculation. This allows the scheduler to assign RB to a user that can transmit the maximum amount of data. Outer-loop LA stabilizes any error in the CQI reports from the user. It can be observed from the figure, there are separate multiplexers for each user that controls the multiplexing of a set of bearers associated with the user into a single output stream. As it can be inferred, a scheduler in LTE network needs to ensure that it meets the goals standardized in the LTE specifications, provide differentiated services to the traffic based on its QoS characteristics, and closely-cooperate with other RRM modules such as HARQ and LA.

A. Related Work

Plenty of scheduling algorithms for LTE network have been proposed in literature that incorporate QoS characteristics standardized in QCI Classes, and ensure that the LTE service requirements such as efficient spectrum utilization, low latency and high throughput are met. Laneri authored a thesis [6] that measures performance of scheduling algorithms in a Super 3G network. Super 3G is the final evolution of 3G network that uses the same RAN architecture as 3G but instead of CDMA access scheme, it uses the OFDM resource access scheme, which is also used in LTE. In addition, it enforces QoS by differentiating between the flows requiring Guaranteed Bit Rate (GBR) and best-effort traffic. However, the thesis presents scheduling algorithms that are only concerned with VoIP traffic, fairness, and delay intolerance. This is not enough in LTE network, as it has stringent QoS requirements that needs to be adhered to. The most complete survey on downlink scheduling algorithms in LTE network was done by Capozzi, Piro, Grieco, Boggia and Camarda in [7]. In their survey, they have explored scheduling algorithms catering to GBR QoS metric, delay intolerance metric, and packet loss rate metric. In addition, they have presented schedulers designed specifically for VoIP traffic as voice is one of the key service offering in a mobile network. Though their work is extensive, the target audience of their survey is someone who has specific knowledge about LTE network and wants to use the survey as a reference guide. In addition, the survey fails to compare the presented solutions that would otherwise allow the reader to compare strengths and weakness of the algorithms.

The goal of this survey is to categorize the scheduling algorithms based on the QoS characteristics, describe the working of the algorithms, and analyze the feasibility of the algorithms in real LTE network by questioning the robustness of the assumptions made in the proposed solutions. In addition, the survey does not require the reader to have prior knowledge about mobile networks. Therefore, this paper is organized into five sections: taxonomy, description of algorithms, critical analysis of the algorithms, challenges and future work, and conclusion. In Section 3 we discuss the taxonomy phase that classifies the scheduling algorithm in terms of QoS metrics defined in LTE; Section 4 offers detailed explanation of the algorithms classified as part of taxonomy; Section 5 critically compares the solutions to understand the trade-offs associated with each solution; Section 6 discusses the challenges the current algorithms can face when deployed in a real LTE network and new requirements defined in future mobile network that will affect forthcoming scheduler design; and Section 7 concludes the survey.

III. Taxonomy of Solutions

Predecessors of LTE utilized spectrum efficiency, measurements from user describing quality of transmission channel between base station and user, and service differentiation metrics comprising of throughput and fairness as mechanisms to make scheduling decisions. However, with the evolution of wireless technology to LTE, the design of the resource scheduler has also evolved to meet the QoS specifications and new service requirements in LTE. Plethora of scheduler designs have been proposed in literature that meet LTE goals and provide differentiated service to diverse traffic. However, each proposed algorithm uses different QoS mechanisms as part of the design strategy due to which the algorithm may not be suitable to every multimedia application. In other words, the literature lacks a snapshot that would otherwise allow service providers and researchers to compare proposed algorithms, determine which applications are suitable for each type of algorithm, and get a sense of open areas that need to be addressed. This section discusses design factors employed by the LTE schedulers proposed in literature and provides classification of algorithms that provide service providers and researchers with a road-map of design strategies used by the algorithms.

Design Factors

The scheduler design needs to address service offerings of LTE, co-operate with other components in the LTE architecture, and develop service differentiation mechanisms that offer QoS to diverse traffic. In other words, the design factors can be categorized into LTE Focused factors and QoS Focused factors.

LTE Focused:

- *HARQ Re-transmissions* - In LTE, Hybrid Automatic Re-transmission Request (HARQ) component in RRM provides reliable communication over noisy channels. Every packet contains data and error correction codes, also known as Forward Error Correction (FEC) code that can be used to recover subset of erroneous data. Upon successfully receiving the packet, UE sends Acknowledgment (Ack) to the HARQ manager; on the other hand, if the data bits received are erroneous, UE tries to recover it using the FEC code in the packet. If the data bits cannot be successfully recovered, it sends a Negative-Acknowledgment (Nack) to the HARQ manager indicating request for re-transmission of erroneous packet. When the HARQ manager receives Nack, it re-transmits the packet. Re-transmission of packets needs the same number of RBs as the original transmission. Hence, the scheduler needs to closely work with the HARQ manager to allocate appropriate resources for re-transmissions.
- *Spectral Efficiency* - In LTE, communication between the user and the base station occurs over the air via radio frequencies or transmission channels comprised of radio waves. Moreover, transmission rate depends on the quality of radio waves used for transmission. In other words, spectral efficiency can be defined as utilizing radio waves to its fullest ability. This is achieved by dynamically adapting transmission rate to the quality of radio waves. To this aim, currently proposed algorithms account for quality of transmission channel - also referred

Algorithm		LTE Design Factors		
Metric Specific		CQI	Scheduling Decision	HARQ Re-transmissions
GBR Provisioned	Packet Set Scheduler (PSS)	Yes	per-PRB	
	Multi-QoS Aware Fair Scheduler (MQFS)	Yes	per-PRB	
	Dynamic Hybrid Scheduler (DHS)	Yes	per-PRB	
Delay Provisioned	Modified Largest Weighted Delay First (M-LWDF)	Yes	per-PRB	
	HARQ Aware Scheduling Algorithm (HASA)	Yes	per-RB	Yes
	Exponential/ Proportional Fair (Exp/PF)	Yes	per-RB	
	Two-Level Scheduler (TLS)	Yes	per-RB	
	Delay-Prioritized Scheduling (DPS)	Yes	per-RB	
	Log Rule	Yes	per-RB	
	Exp Rule	Yes	per-RB	
	Co-Operative Game Theory Based Scheduling (CGT)	Yes	per-RB	
System Centric	Overload-State Downlink Resource Allocation (OSDRA)	Yes	per-PRB	
Application Specific				
VoIP	VoIP Priority Mode (VPM)	Yes	per-PRB	
	RAD-DS	Yes	per-PRB	Yes
Video Streaming	DASH	Yes	per-PRB	
	Video Streaming	Yes	per-PRB	
	Channel Aware and Buffer Aware (CABA)	Yes	per-PRB	

TABLE I: Classifies Schedulers based on LTE Factors

to as Channel Quality Index (CQI) - between the base station and the user in order to determine maximum possible transmission rate for the user.

- *Computational Complexity* - The packet scheduler allocates resources every TTI, which is 1 ms in duration. This requires the scheduler to be computationally optimal. Therefore, the algorithms use linear time as the upper bound when computing number of iterations needed to schedule resources to users.
- *Scalability* - Short TTI of 1 ms enables the resource allocation schemes to distribute resources to multiple users in one 10 ms LTE frame; thus, increasing system capacity. In order to achieve this, scheduling algorithms take greedy approach by making scheduling decisions per-RB. That is, for every RB, scheduler determines best suitable user based on channel quality.

QoS Focused:

- *Fairness* - The scheduler needs to ensure that every traffic flow is given the opportunity to transmit. In order to achieve this, the scheduler needs to take into consideration past throughput of the user or the traffic flow. Past throughput describes amount of data a traffic flow has transmitted till now compared to other traffic flows. If one flow's transmission rate is less than other flows, then it should be given a priority for transmission.
- *Metrics* - In LTE, QoS is characterized by metrics defined in QoS Class Identifier (QCI) Classes *(refer to table in appendix)* standardized in LTE specification by 3GPP. Following is the description of the key metrics defined in QCI classes:
  - *Guaranteed Bit-Rate (GBR)*- GBR value indicates that the average throughput of the traffic flow should be atleast equal to GBR value and maximum of Maximum Bit Rate (MBR) value. In other words, at any point of time, the scheduler needs to make sure that it assigns enough resources to the flow such that minimum throughput requirement (GBR) is met. On the other hand, if GBR is not assigned to the traffic flow (referred to as Non-GBR flow), the scheduler is not bound to assign minimum resources to the flow and can assign maximum of Aggregated Maximum Bit-Rate (AMBR). GBR, MBR, and AMBR are negotiated between the service provider and user at the time of subscription. GBR is mainly used to distinguish between real-time traffic such as a VoIP call, video streaming and real-time gaming from best-effort traffic such as file transfer and web-browsing.
  - *Target Delay* - Target delay signifies maximum amount of time that the packet can be delayed. This is determined by measuring Head of Line delay (HOL). HOL is the amount of time for which the head of queue packet has been waiting for transmission. In order to preserve QoS, the HOL delay should not exceed the target delay. For example, consider a flow for a VoIP conversation that can tolerate a maximum delay of 100 ms before it is

Algorithm		QoS Design Factors						
Metric Specific		GBR	Target Delay	HOL Delay	Packet Loss Rate	Queue Length	UE Buffer Length	Priority
GBR Provisioned	Packet Set Scheduler (PSS)	Yes						
	Multi-QoS Aware Fair Scheduler (MQFS)	Yes						
	Dynamic Hybrid Scheduler (DHS)	Yes	Yes	Yes				
Delay Provisioned	Modified Largest Weighted Delay First (M-LWDF)		Yes	Yes				
	HARQ Aware Scheduling Algorithm (HASA)	Yes	Yes	Yes				
	Exponential/ Proportional Fair (Exp/PF)		Yes	Yes				
	Two-Level Scheduler (TLS)				Yes	Yes		
	Delay-Prioritized Scheduling (DPS)		Yes	Yes				
	Log Rule		Yes	Yes				
	Exp Rule		Yes	Yes				
	Co-Operative Game Theory Based Scheduling (CGT)		Yes	Yes				
System Centric	Overload-State Downlink Resource Allocation (OSDRA)	Yes	Yes		Yes	Yes		Yes
Application Specific								
VoIP	VoIP Priority Mode (VPM)		Yes		Yes	Yes		Yes
	RAD-DS	Yes	Yes	Yes			Yes	
Video Streaming	DASH	Yes	Yes					
	Video Streaming	Yes	Yes	Yes	Yes		Yes	
	Channel Aware and Buffer Aware (CABA)	Yes	Yes		Yes		Yes	

TABLE II: Classifies Schedulers based on QoS Factors

noticeable by the user. Assume that one end of the VoIP conversation is served with a delay within the target delay of 100 ms - let's name it Side A - and the transmission at other end - let's name this as Side B - exceeds the target delay of 100 ms. In this scenario, Side A will receive voice projected by Side B within delay unnoticeable by the user's perception; however, voice projected by Side A will be received by Side B after a noticeable delay. This delay will result in an undesirable out of sync conversation between Side A and Side B. In a nutshell, target delay provides a tight bound that guarantees quality of service within human perception.

– *Packet Loss Rate* - This metric describes maximum tolerable packet loss acceptable for a traffic flow. In addition, if the packet loss exceeds the budget, re-transmission may be triggered by the user. For example, if the application at the user's end was running on top of TCP protocol, packet loss will prevent application from sending acknowledgement, which will trigger re-transmission after the time-out interval. Continuing with the example discussed previously, VoIP can tolerate maximum packet loss of 10% before it is noticeable by the user. If Side

A from previous example experiences packet loss within the tolerable budget of 10% and packet loss at Side B exceeds the tolerable budget, the conversation between Side A and Side B will be out of context.

- *Buffer Overflow* - The traffic queues at the base station are limited in size. The scheduler needs to ensure that the queue does not overflow. If the queue overflows, the packet loss rate may exceed the tolerable rate and may trigger re-transmission, which may further cause a delay in service. In addition, the data received by the user is stored in buffer in the user's device before it is retrieved by the application. This buffer is of limited size as well. If the scheduler does not consider the length of the buffer size at user's end, the buffer will overflow, triggering possible re-transmission, which may again induce delays.

Classification

As discussed a LTE scheduler not only need to address goals characterized by LTE architecture but also need to meet the new QoS requirements acquired as part of the evolution to LTE. Based on the two responsibilities of the scheduler - address LTE goals and meet new QoS requirements - and the design factors defined above, we classify algorithms into two classes: metric based and application based. Metric based algorithms can be further classified as GBR provisioned, delay driven and system centric algorithms. GBR provisioned algorithms' objective is to ensure that the traffic is bestowed required GBR transmission service; delay driven algorithms makes sure that traffic flows do not violate their required target delay as the violation may result in triggering of re-transmissions; system centric algorithms take rational approach by designing scheduler with real network system in mind. For example, the algorithm discussed in this paper - Overload-State Downlink Resource Allocation, considers the fact that network may sometimes operate to its maximum capacity i.e. in overload state; hence, the scheduler needs to robustly schedule resources in overload state such that QoS of the application is preserved. In other words, metric based algorithms strive to generalize the solution for multiple traffic types by considering metrics that are common between different traffic types. Trade-off of this approach is that many traffic types may share same target value for a given metric; for example, conversational voice and conversational video, both require GBR. However, we would not be able to distinguish between these two traffic types unless we considered other metrics such as delay budget, which is 100 ms and 150 ms for conversational voice and video respectively. Thus, metric based algorithms need to consider multiple QoS metrics in order to ensure the scheduler is able to differentiate one traffic from another. On the other hand, application specific algorithms tend to consider metrics desired by the specific application. Considering the fact that VoIP and video streaming are one of the most used multimedia applications, application based algorithms can be further divided into VoIP based and video streaming based. Table I and II maps the algorithm classifications to LTE design factors and QoS design factors that the respective algorithm satisfies.

IV. Description of Algorithms

Majority of the proposed QoS schedulers described below differ in the method used to prioritize the traffic flows and the QoS metrics employed. Most of them use traditional Proportional Fair (PF) scheduler from predecessor wireless broadband system in some form or the other. Hence, it is important to first understand the design of PF scheduler and why it is extensively used by the LTE packet schedulers.

PF scheduler [8] first calculates PF priority for each user and then schedules the users based on the calculated priority - highest is preferred. PF priority is calculated by dividing current instantaneous throughput of the user by past average throughput of the user (1). As discussed previously, transmission rate or throughput of the user depends on the quality of transmission channel between the base station and user. In other words, PF uses CQI from user to determine optimal transmission rate for the user.

$$PF(n) = \frac{R_{curr}(n)}{R_{past}(n)} \tag{1}$$

where $R_{curr}(n)$ and $R_{past}(n)$ corresponds to current and past average throughput of the user respectively. It can be observed that as the average throughput of the user decreases - possibly due to continuous bad channel quality, its priority increases. This ensures that a user with continuous bad channel quality is not starved. In other words, this property of ensuring fairness among users, makes PF an attractive component in LTE packet schedulers. In addition, it should be noted that although PF scheduler ensures fairness, it ignores QoS restrictions such as GBR, target delay, and packet loss rate (PLR) of the user defined in LTE. Hence, PF solution is not the preferred standalone solution in LTE and is usually coupled with other QoS based scheduling technique as discussed further in this section.

Guaranteed Bit-Rate (GBR) based algorithms

Priority Set Scheduler (PSS):

PSS [9] is designed to make sure traffic flows that require GBR are served with higher priority. PSS accomplishes this in three steps: first, the flows go through scheduler admission control that decides if a flow is eligible to enter the scheduler, second, eligible flows are ranked by the ranking function, and lastly, resources or RBs are assigned to the traffic flows based on their rank.

Admission Control - The admission controller considers a flow to be eligible for scheduling if the flow has any pending re-transmission from HARQ, or if the flow has enough buffered data or the HOL packet delay is nearing the target delay. First condition ensures that the scheduler does not neglect re-transmissions, and it co-operates with other components of Radio Resource Manager (RRM) such as HARQ. By checking the amount of buffered data against a threshold value, scheduler intends to ensure that the resources

allocated are being utilized fully and hence, the spectrum is not wasted when transmitting data with highest possible bit rate. For example, lets assume the scheduler does not check for the amount of buffered data and there are 2 flows, one with the buffered data below the threshold - let's call this Flow A and other with data equal to or above the threshold - let's call this Flow B. In addition, assume rank of Flow A calculated by the ranking function is higher than rank of Flow B, and both A and B have the same channel quality, i.e. same transmission rate. This means that Flow A will be scheduled ahead of Flow B. From this scenario it can be inferred that if the amount of data in buffer is less than the calculated transmission rate and if admission controller does not check for the second condition, throughput of Flow A will be less than calculated rate, leading to wastage of resources. On the other hand, if the admission controller prevented Flow A from being scheduled, Flow B would have been scheduled with highest possible throughput, resulting in spectrum efficiency. Lastly, checking HOL delay ensures that transmission does not violate target delay, which would otherwise cause packet loss and trigger re-transmission as discussed in previous section.

Ranking Function - The ranking function, also referred to as Time Domain (TD) Scheduler, aims to rank users based on GBR. First, it separates users into two sets - Set 1 consists of users that have not achieved their respected GBR goal and other set consists of users that have achieved their GBR goal and users that do not require GBR i.e. Non-GBR users. The scheduler then assigns rank or priority to each user in the respective sets in following way: users in Set 2 are ranked using PF scheduler (1) to ensure fairness. Rank of users in Set 1 is calculated by taking reciprocal of past average throughput. This ratio gives an estimation of whether the user is close to achieving its GBR value or far from achieving its GBR value. For example, higher past average throughput indicates that the user has been served at a higher rate and may be close to achieving its GBR, in which case the user will be given lower rank. On the other hand, lower past average throughput indicates that user has been served as often and that it may be far from achieving its GBR value, in which case it will be given higher priority. The ranking function then sorts the users in both sets in descending order and selects N_{Mux} users based on their ranking, which becomes input of the resource assignment function, which we will discuss below. The complexity of the ranking function depends on computation time of division as it considers ratios to compute the rank. Complexity of division operation is of the same order as that of multiplication, which can be computed in linear time [10]. Hence, overall complexity of the ranking function is $O(M.N)$, where M is number of flows or input to ranking function and N is time complexity of the division operation.

Resource Assignment Function - This function, which is also referred to as Frequency Domain (FD) scheduler assigns resources or RBs to N_{Mux} users received as input from the ranking function. Resource allocation function computes its own ranking on the received input to prioritize RB allocation. It proposes three strategies to rank the N_{Mux} users. First strategy uses PF solution to rank users in a fair manner. This strategy calculates average only when the user is scheduled. Disadvantage of this approach is that sample size for calculating average might be small, and if it contains extreme values i.e. very high and very low values, the average will be distorted. For example, consider a simple example with 2 elements in the sample size with value 3 and 11. In this case the average will be 7. This distortion gives inaccurate picture of user's past average throughput and may result in imprecise rank. Second strategy calculates rank by dividing current instantaneous throughput by estimate of throughput if the user was scheduled every interval i.e. running average of the throughput. Compared to first strategy, this strategy considers large sample for calculating the average, which refines the average providing accurate ranking. Third strategy on the other hand calculates rank by dividing current channel quality by past average quality. This strategy makes sure that a user who was experiencing bad channel quality in the past is given high priority if its current channel quality has improved. In case of GBR flow this ratio allows scheduler to compensate user for the past low transmission rate due to bad channel quality. The computation complexity incurred in this step is $O(N_{mux}.N)$, where N is time complexity of division and N_{mux} is number of inputs to resource assignment function.

Multi-QoS Aware Fair Scheduler (MQFS):

Goal of MQFS [11] is to prioritize transmission of GBR traffic flows and also compute priority factor for Non-GBR flows based on priority of QCI Classes *(refer to appendix for QCI class criteria)*. Similar to PSS, MQFS first ranks users based on the QoS characteristics and then the resource allocation function allocates RBs to the flows.

Ranking Function - Ranking function first assigns weights to the QCI classes with weights decreasing as QCI class priority increases; it then separates GBR based flow and Non-GBR flows into two user sets. GBR flows in Set 1 are organized in such a way that flows with higher weight are above flows with lower weights. Set 2 on the other hand is organized in descending order of priority. Priority is calculated by dividing the weight associated with the QCI class by past accumulated throughput that is normalized using a constant between 0 and 1. This provides some degree of prioritization among best effort traffic flows. However, depending on the value of the weights, the ratio also tends to starve traffic flows assigned to higher QCI Class. For example, consider the parameters considered in the simulation where QCI7 is assigned weight of 10, QCI8 is assigned weight of 5 and QCI9 is assigned weight 1. Calculating the ratio with respect to these weights, it can be observed that traffic belonging to QCI Class 9 will almost never be scheduled unless there are no traffic flows present from other QCI classes. This defeats the fairness design factor discussed in the previous section. The ranked Set 1 and Set 2 are then provided to Resource Allocation Function for scheduling. Since the ranking function uses

ratio to compute rank of Non-GBR flows, the computation complexity of the single ranking function operation is same as complexity of division operation, which is linear. Overall complexity of the ranking function is equivalent to $O(M.N)$, where M is number of Non-GBR flows and N is complexity of division operation.

Resource Allocation Function - Resource Allocation Function takes the greedy approach to assign the resources first to users in GBR list and then to users in Non-GBR list. It iteratively assigns best RB to each user based on priority. In case of GBR flows, the scheduler performs a check at the end of each iteration to see if the flow has achieved its GBR criteria and if it has enough data buffered. It will not assign further resources if any of these criteria are satisfied. In terms of runtime of the iterative algorithm, computational complexity is linear, specifically equal to number of RBs to be scheduled as it iterates over each RB. The computational complexity of this function depends on number of available RBs. Hence, net computational complexity is equivalent to $O(M.N)$, where M is number of flows and N corresponds to number of RBs. In addition, this allocation performs further checks at each iteration that may increase overall decision making process. This is an important factor to consider as the scheduler (ranking function and resource allocation function together) need to make decisions every TTI - 1 ms.

Dynamic Hybrid Scheduler (DHS):

Consider a system with a VoIP and video streaming where both VoIP and video streaming are GBR based flows. In addition, assume the HOL packet delay of the flows to be 50 ms, and 200 ms respectively; in addition, as per standardized QCI Class Identifiers *(refer to Appendix)* target delay of VoIP and video streaming flows is 100 ms and 300 ms respectively. In this scenario it can be observed that video streaming is close to its target delay and should be prioritized for scheduling. In other words, when prioritizing GBR flows over Non-GBR flows, the scheduler also needs to rank flows within the GBR class to make sure the flows do not violate its respective target delay budget.

DHS [12] uses this notion to provide delay bound GBR service to traffic flows. DHS first computes priority of each user by dividing HOL delay by the target delay. This ratio gives precise estimate on how close the packet is to its target delay. Closer the flow is to its target delay, higher is the priority. After calculating the priority, it sorts the flows in descending order of its priority. The scheduler then iterates over the sorted list and assigns required GBR amount of resources to the flows if the flow has enough data buffered. The residual resources obtained after serving GBR flows are shared among all the traffic flows using the same iterative approach. In addition, DHS spreads out GBR of a flow across multiple scheduling intervals if the system is operating in its full capacity i.e. in overload state. It does this by dividing GBR value by a constant that can be adjusted depending on the capacity at which the system is operating. At normal load, the constant is set to 1 indicating that the user should be served the GBR value in its entirety. In addition, the constant can be used to provide fairness between GBR traffic and Non-GBR traffic. For example, assume the constant to be 2. This will halve the GBR value of a user for a scheduling time interval. In addition, total amount of resources needed to serve all GBR flows will also be halved leaving the scheduler with more residual resources than it would have if the constant value was set to 1. Having more residual resources means that the scheduler can iterate over the priority list and serve traffic flows - which can be best-effort flow or real-time flow - close to its target delay. Moreover, the worst case computation complexity of DHS depends on time it takes to sort the priorities, and maximum of the runtime performance of calculating priority for N users and iteratively assigning RBs to traffic flows. Comparison based sorting algorithms such as Merge Sort or Quick Sort takes $O(NlogN)$ in general where N refers to number of flows. In addition, time complexity of calculating priority for N flows and iteratively assigning RBs is equal to $O(M.N)$ where M is time it takes to perform division operation or number of RBs to be scheduled, and N corresponds to number of flows. On worst case, the performance of this algorithm is $O(N^2)$

Delay Provisioning Algorithms

Modified Largest Weighted Delay First (M-LWDF):

M-LWDF [13] tends to schedule users close to its target delay in a fair manner. It achieves this by scheduling users per-RB based on priority that is product of HOL delay to target delay ratio, instantaneous throughput to past average throughput (same as PF solution) ratio, and logarithm of odds that HOL packet delay of the flow violates the target delay as described in the equation below:

$$Priority = (-log(\delta_i)) \cdot \frac{d_i}{D_i} \cdot \frac{r_i}{R_i} \tag{2}$$

where δ_i is the maximum acceptable probability that HOL delay (d_i) of user exceeds the delay threshold (D_i). r_i and R_i refers to the current throughput and past average throughput respectively. The delay ratio and δ_i embodies the QoS requirement to ensure that flow close to its target delay and with lower tolerance to exceed its respective target delay, is received the highest priority. The throughput ratio, which is the same as the one used in PF solution ensures fairness among users. In addition, multiple traffics including real-time and best-effort have same target delay defined. For example, both video streaming traffic which is real-time, and HTTP traffic which is best-effort, have same target delay of 300 ms. Hence, the scheduler needs to make sure that it is able to distinguish between real-time traffic from best-effort traffic by choosing appropriate δ_i. The computational complexity of M-LWDF depends on complexity of division, multiplication, and logarithm operations. Division is of the same order as of multiplication and complexity of 2 integer multiplication is linear [10]. In addition, logarithm function for real number can be computed using Taylor series in $O(n^{1.5})$, where n is number of digits of precision at which the function is to be

evaluated [14]. Worst case complexity of computing priority for N users is $O(N.n^{1.5})$.

HARQ Aware Scheduling Algorithm (HASA):
In practical scenario, because of interference in transmission channel or degradation in channel quality, user may not receive all the data correctly. Some data might be corrupted, which needs to be re-transmitted. As discussed previously, HARQ manager in Radio Resource Management (RRM) component of LTE is responsible for re-transmissions of errored data. In addition, the scheduler is responsible of closely co-operating with HARQ manager to prioritize scheduling of re-transmission to ensure traffic's target delay is not violated, traffic's packet loss rate does not exceed the packet loss rate standardized in QCI Class and maximize effective user throughput. With this aim HASA [15] modifies M-LWDF packet scheduler to accommodate for re-transmission from HARQ. Like M-LWDF, HASA makes scheduling decision per-RB. When a traffic flow has packets to be re-transmitted, the scheduler prioritizes re-transmissions based on following expression:

$$Priority = \frac{-log(\delta_i)}{D_i} \cdot exp(\frac{D_i.\alpha}{D_i - d_i}) \cdot \frac{r_i}{R_i} \quad (3)$$

Where D_i is the threshold delay of the traffic as standardized in QCI Class and α is a constant. Rest of the parameters hold same significance as in M-LWDF. The exponential excerpt is the core of the expression that embodies QoS criterion. The target delay to time before expiry (denominator) ratio in exponential ensures that users close to their respective target delay receive higher priority. In addition, the α is tuned to make sure GBR service of the flow is satisfied. In addition, when there are no re-transmission packets queued, the scheduler serves traffic flows as per M-LWDF, except if the buffered data in the queue is less than the instantaneous throughput achievable on the RB to be scheduled, it replaces $\frac{r_i}{R_i}$ ratio with $\frac{Q_i}{R_i}$, where Q_i is the queue length or amount of data buffered in queue at an instance. Otherwise it uses same instantaneous throughput to average throughput ratio as in M-LWDF. This modification allows the scheduler to select user with maximum amount of buffered data for the RB to be scheduled to ensure spectrum efficiency. Computational complexity of HASA depends on complexity of logarithm function and exponential operation. Logarithm operation can be computed using Taylor series in $O(n^{1.5})$ as discussed above. Computational complexity of exponential is $O(M(n)log(n))$ [16], where n corresponds to n-digit number and $M(n)$ is complexity of multiplication, which in theory can be computed in linear time. Worst case complexity for N users results in $O(N.n^{1.5})$.

Exponential/Proportional Fair (EXP/PF):
Similar to M-LWDF, EXP/PF [13][17] aims to provide delay driven scheduling service to users. To achieve this, it first defines a product consisting of first two terms in (2), let's call this A. Then, it calculates average of A among the flows present in the scheduler. It then determines priority by using exponential function as described below:

$$Priority = exp(\frac{A - A_{avg}}{1 + \sqrt{A_{avg}}}) \cdot \frac{r_i}{R_i} \quad (4)$$

The average in exponential function helps normalize the delay of the user. In addition, exponential function grows at faster rate; this means that small change in input causes variation in the result. The variation increases accuracy of the scheduler by reducing number of ties. Also, the throughput ratio in (3) helps preserve the fairness among users. In order to distinguish between real-time traffic and best-effort traffic, Exp/PF prioritizes best-effort based on following product:

$$\frac{Max\ HOL\ packet\ delay\ of\ all\ real\ time\ flows}{Number\ of\ real\ time\ flows} \cdot \frac{r_i}{R_i} \quad (5)$$

The first ratio in (4) ensures that all the real-time flows are served prior to the best-effort traffic, and the throughput ratio ensures fairness among users. Computation of real-time equation (3) involves exponential; computational complexity of exponential is $O(M(n)log(n))$ [16], where $M(n)$ is complexity of multiplication. Theoretically multiplication can be performed in linear time as described in [10]. In addition, computation complexity of best-effort product (4) is linear as it involves division, which can be computed in linear time. Hence, worst case computation for N numbers is $O(N.M(n).log(n))$, where N is number of flows and n corresponds to n-digit number.

Two-Level Scheduler (TLS):
Algorithms discussed so far first rank users based on the QoS requirement and then assigns resources as per the need and supply to highest ranked users. TLS [18] takes a different approach by pre-allocating quota of data that a real-time traffic flow can transmit in a given 10 ms long LTE frame. It then allocates RBs to users every scheduling interval i.e. every TTI to meet the assigned quota. Once the quota is met, the traffic flow is not allowed to transmit until the next LTE frame when the quota for the real-time flow will be re-calculated. To achieve this, TLS isolates quota calculation functionality and resource assignment functionality by fragmenting scheduler into Upper Level Scheduler and Lower Level Scheduler.
Upper Level Scheduler - Upper level scheduler, which is also referred to as the frame level scheduler borrows concepts of control loop and discrete time convolution to calculate delay-bound quota for real-time flows that the flow should transmit in the given LTE frame. The quota for every real-time flow is independently calculated at the beginning of each LTE frame. In convolution theory, the system receives input function - typically a signal. The convolution operation then combines the received input signal with the system's own signal, also referred to as impulse response signal to produce the output signal. The output signal is a blend of the two input signals and expresses amount of area overlap between the two input signals. In addition, it can be inferred that the system impulse response signal shapes the external input

signal to produce the output signal. Upper level scheduler treats the queue length or estimate on amount of data to be buffered in an LTE frame as an input signal signal. The response function in the upper level scheduler is delay bounded such that output produced is delay bounded. In other words quota to be transmitted in the given frame is delay bounded. Maximum upper bound queuing delay produced by the upper layer scheduler is $(M_i + 1) \times T_f$, where M_i is length of impulse response function and T_f is length of LTE frame, which is 10 ms. The upper level scheduler considers worst case scenario where lower level scheduler is not able to transmit the quota of data in the given time frame. In this case, the upper level scheduler defines bigger quota in the next frame to accommodate the pending quota from previous frame. This way, the upper level scheduler ensures that the packet does not miss its delay bound. For example, consider scenario in Figure 3 with real-time gaming traffic, which has target delay of 50 ms. Assume the length of impulse response to be 3, that is upper bound on delay is 30 ms. In addition, for the sake of simplicity, assume the quota of data to be transmitted during every frame is 10; as per the delay bound the deadline of the quota of packet in frame 1 i.e. T=0 is 30. It can be observed in the figure that at the end of first frame, T=10, the lower level scheduler was only able to transmit 5 packets of the quota (lightly shaded) and 5 are pending. The upper level scheduler increases the quota for this flow to 15 in the second frame, to make sure pending packets are scheduled before it misses the deadline. In addition, it can be observed at T=20 i.e. starting of third frame, when there are no pending packets from previous frame, the quota calculation returns to normal. Another advantage of is that upper level scheduler co-operates with HARQ manager to ensure past transmissions were successful. In case of re-transmissions, the queue length is adapted to accommodate for re-transmission of packets. Worst case computational complexity of the upper level scheduler is $O(NM_{i\_max})$, where N is total number of real-time flows, and M_{i_max} is maximum impulse response length among the real-time traffic flows [18].

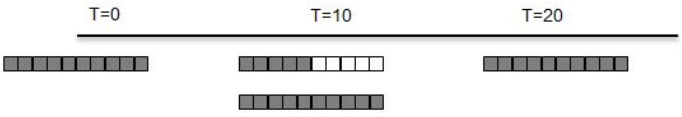

Fig. 3. Illustrates Dynamic Adaptation of Quota to Satisfy Delay Bound

Lower Level Scheduler - Goal of the lower level scheduler is to allocate resources to real-time flows every scheduling interval i.e. 1 ms such that the respective quotas are met. In addition, surplus of RBs left after serving the quotas of real-time flows is used to schedule best-effort traffic. Lower level scheduler uses PF solution (1) to allocate resources to real-time and then best-effort traffic. Computational complexity of lower level scheduler depends on the complexity of division operation, which is linear.

Delay-Prioritized Scheduling (DPS):
DPS [19] is a simple algorithm that iteratively assigns RB to users based on their deadline until the resources or RBs are fully used up. The operation of of the scheduler can be divided into three steps:

1) *Step 1* - Deadline for each flow is calculated in following manner: $d_i(t) = T_i - W_i, i \ \forall \ flows$, where T_i is the delay threshold for the flow and W_i is the HOL packet delay of the flow.
2) *Step 2* - Flow with earliest deadline i.e. minimum $d_i(t)$ among all flows is selected.
3) *Step 3* - This step employs greedy solution i.e. best PRB based on selected user's channel condition is assigned to the selected flow.

The above steps are repeated until PRB's are exhausted. Advantage of this approach is that depending on type of traffic, the scheduler can assign delay threshold depending on the target delay and acceptable packet loss rate for the flow. For example, VoIP traffic has target delay of 100 ms and acceptable packet loss rate of 10%; real-time gaming on the other hand has target delay of 50 ms and acceptable packet loss rate of 1%. In other words, real-time gaming traffic should be assigned much smaller delay threshold compared to VoIP traffic to make sure that real-time traffic scheduling is prioritized. Computation complexity of this algorithm depends on number of iterations it has to make. The algorithm first iterates over all the flows to calculate deadline for each flow in $O(N)$ time where N is number of flows. It then iterates through the list to select user closest to its deadline and iterates over the list of PRB to select best possible PRB for the user depending on channel quality. The algorithm runs in $O(NM)$ time where N is number of flows and M is number of RBs available for scheduling. As it can be observed, simplicity comes at an expense of computation complexity.

Exp Rule and Log Rule:
Log rule and Exp rule aim to provide delay bounded scheduling service to the user [20]. In this effort, both Log Rule and Exp Rule make use of queue status and channel status to calculate priority. Using queue status as a factor prevents the queue, which is finite in size from overflowing. Both the rules consider following constants to calculate priority of user:
$c = 1.1,\ b_i = \frac{1}{\mathbb{E}[K_i]}, and\ a_i = \frac{6}{0.99 \times TargetDelay}\ or\ a_i = \frac{10}{0.99 \times TargetDelay}$, where K_i is spectral efficiency of the user.

- *Exp Rule* - Exp Rule is modified version of Exp/PF algorithm that schedules user with highest priority, which is calculated as follows:

$$i \in arg\ max_{1 \le i \le N} b_i . exp\left(\frac{a_i w_i(t)}{1 + \sqrt{(1/N)\sum_j w_j(t)}}\right) . K_i(t) \quad (6)$$

Where w_i is Head of Queue (also known as Head of Line) packet delay. Exp rule grows faster compared to log rule;

this increases variation in the result from Exp Rule among users. The variation increases accuracy of the scheduler by reducing number of ties. Exp Rule differ from Exp/PF in 2 ways. First, Exp/PF distinguishes between real-time and best-effort flows by including δ_i factor, which describes the maximum acceptable probability that HOL delay of user can exceed the target delay. For example, video streaming which is real-time flow and FTP which is best-effort flow, have same target delay of 300 ms. In this case, imposing δ_i factor, which would be much smaller for video streaming traffic compare to FTP traffic due to its real-time nature helps distinguish between real-time and best-effort flows. Second, Exp/PF considers instantaneous throughput to past average throughput to ensure fairness. On the other hand, Exp Rule considers instantaneous spectrum efficiency to average spectrum efficiency ratio ($\frac{K_i}{b_i}$) to ensure fairness. Moreover, computational complexity of Exp Rule is same as Exp/PF, which is equal to $O(N.M(n).log(n))$, where $M(n)$ is complexity of multiplication, N is number of flows and n corresponds to n-bit integer.

- *Log Rule* - Log Rule uses logarithm of HOL packet delay to target delay ratio to calculate priority of the user. In addition, it uses instantaneous spectrum efficiency to average spectrum efficiency as a means to guarantee fairness to users. Log Rule can be expressed by following equation:

$$i \in arg\ max_{1 \leq i \leq N} b_i \times log(c + a_i w_i(t)).K_i(t) \quad (7)$$

Where w_i is HOL packet delay of user. Compare to Exp Rule, Log rule grows slowly. Hence, the output priority of log rule may result in ties among users. However, instantaneous spectrum efficiency to average spectrum efficiency ratio helps break the ties as probability of spectrum efficiency being same for multiple users is slip. A drawback to Log Rule is that it is not able to distinguish between real-time and best-effort traffic. It uses HOL packet delay to target delay ratio to prioritize the user, but that is not enough as multiple traffic types have same target delay. For example, buffered video (real-time traffic) and TCP based applications such as FTP, P2P, and HTTP have delay budget of 300 ms. Computational complexity of log with real number can be computed using Taylor series in $O(n^1.5)$, where n is number of precision at which the function is evaluated. On worst case, for N users it is equivalent to $(N.n^{1.5})$, where N is number of flows.

$$i \in arg\ max_{1 \leq i \leq N} b_i \times log(c + a_i w_i(t)) \times K_i(t) \quad (8)$$

Cooperative Game Theory Based Scheduler:

In Cooperative Game Theory (CGT), group of players form coalitions. The coalitions then compete; the winning coalition shares the prize or profit among its players. LTE supports different types of traffics such as VoIP, buffered video streaming, real-time gaming, live streaming, conversational video, and other best effort traffic types. Each of these traffic types have different QoS requirements as specified in the QCI Classes table in Appendix. In other words, each traffic flow is classified into one of the traffic types. In CGT base scheduler [21] the traffic types are treated as players and resources claimed by each player is equal to $g_i = k_i.b_i$, $1 \leq i \geq N$, where N is number of players or traffic class, k_i is number of flows associated with each player, and b_i is amount of resources claimed by each flow associated with the player. In addition, with N players there can be 2^N coalitions. Also each coalition has characteristic function associated with it that describes surplus resources left after satisfying resource requirements of players not part of the coalition; this surplus resource is nothing but the dividend acquired by the coalition after distributing resources to non-coalition players. In the CGT based schedulers, Resource Distribution functions initiates and plays the cooperative game among the coalitions to find optimal resource distribution among the players, then Resource Allocation function allocates the resources to the flows associated with each player. This is described below. *Resource Distribution Function* - Resource distribution is found by playing the cooperative game as follows:

1) Calculate characteristic function for 2^N possible coalitions.
2) Use Shapely Value function to determine average payoff to a player if the player was to be part of a particular coalition.
3) The payoff or dividend calculated by Shapely Value is the optimal resource distribution for the players.

Shapely Value function ensures that resource distribution does not depend on the order in which the player enters coalition. This guarantees fairness. Computation complexity of Shapely Value is NP Complete [22]. However, there are linear time approximation methods that can be used to estimate the average payoff per player. One such approximation method is Owen's multi-linear extensions [23].

Resource Allocation Function - This function provides delay bound resource allocation service to the flows. This function uses Exp rule as described above, except HOL packet delay for real-time flows is calculated using virtual token mechanism. Each flow is assigned a virtual queue with token arriving to this virtual queue at a constant rate. The constant rate at which tokens arrive is equivalent to minimum required throughput for the flow. The HOL packet delay is then defined as: $V_i(t) = \frac{Q_i(t)}{r_i}$, where $Q_i(t)$ is the length of the virtual queue and r_i is the constant rate at which the tokens arrive, which is equivalent to minimum required throughput of the flow. Once the flow is served, number of tokens in the virtual queue is reduced by amount of data transmitted in terms of bits. Advantage of associating HOL delay with minimum required throughput is that it ensures that the flow is guaranteed minimum throughput. Computational complexity of this resource allocation function is same as Exp rule.

System Centric Schedulers

Overload-State Downlink Resource Allocation (OSDRA):
In practical scenario, capacity of operator's LTE network is limited. When the network is operating in its full capacity, it is known to be in overload state. This algorithm [24] takes a novel approach by allowing network operators to utilize QoS values in LTE metadata associated with a traffic flow to rank the flows and assign resources. A key factor that differentiates this algorithm from others is that it takes into account network overload state when making scheduling decisions. The scheduler works as follows:

1) *Assign GBR Resources* - The first step to the algorithm is to assign required GBR worth of resources to the traffic flows.
2) *Ranking Function* - This step ranks traffic flows by normalizing the QoS values embedded in the LTE metadata of the associated flows and then sorting the ranking in descending order. Ranking function normalizes the QoS values by using techniques in artificial neural networks. In artificial neural networks, activation function is responsible for normalizing the inputs. The ranking function uses hyperbolic tangent activation function from artificial neural networks to normalize the inputs i.e. QoS values. It can be defined as $f(x_i, p_i) = p_i.\tanh(x_i)$, where x_i is the input and p_i is the weight of the QoS metric defined by the network operator. In addition, the input x_i is proportion of the measured QoS value of the traffic relative to the QoS metric defined in QCI Class for the respective traffic. p_i allows the operator the flexibility to place more emphasis on a particular QoS metric. For example, if the operator finds out that Quality of Experience (QoE) for a particular application, let's say VoIP application, is more negatively impacted by packet delay, then the operator can increase the weight of the delay QoS metric to ensure that VoIP application is within the required tight delay bound. Moreover, $\tanh$ can be computed using Taylor series in $O(n^{\frac{1}{2}}.M(n))$, where M(n) is computational time of multiplication operation, which is linear; n refers to number of digits of precision desired for evaluation. This means that for N flows, computational complexity of the ranking function is $O(N.n^{1.5})$.
3) *Resource Assignment Function* - Resource Assignment Function takes a greedy approach to assign the residual resources left after assigning GBR worth of resources to all the GBR based traffic flows. It is worth noting that each traffic flow belonging to a user can transmit at most Maximum Bit-Rate (MBR) worth of data in case of GBR based traffic and Aggregated Maximum Bit-Rate (AMBR) worth of data in case of Non-GBR based traffic. Resource Assignment Function ensures that the maximum transmission limit of the traffic flow is not violated. In addition, this function uses Fractional Knapsack technique to allocate resources. Fractional knapsack technique is a approach where a thief with finite sized knapsack strives to steal items with different value and quantity such that worth of the knapsack is maximum. It does this by taking greedy approach, i.e. it iteratively selects maximum quantity of highest value items until the knapsack is filled. In our case, resources left after assigning GBR worth of RBs to GBR based traffic flows act as the size of the knapsack and the ranks calculated by the Ranking Function act as the value of the traffic flows. The resource assigning function then iterates over the ranks and assigns maximum amount of resources until either the resources run out i.e. knapsack is filled or the flow meets MBR or AMBR value, in which case it will move to the next rank flow and assign it the resources. This can be performed in one pass i.e. computational complexity of this function is $O(N)$, where N is number of RBs available for assignment.

VoIP Focused Algorithms:
Required Activity Detection with Delay Sensitivity (RAD-DS):
RAD-DS handles and optimizes the performance of scheduling for traffic mixes of VoIP and best effort users. The idea is that strict prioritizing of VoIP traffic over best effort traffic will lead to degradation in the overall cell throughput. To maximize the overall cell throughput, we need to guarantee a fair timely sharing of resources between VoIP traffic and best effort traffic. For this goal to be achieved, a concept named Packet Bundling is used. Packet Bundling combines many VoIP packets into a single one with the same payload [25]. By such a mechanism, the number of control resources needed by VoIP reduces thereby making more of the PDCCH channel available for the best effort users. And when the channel capacity increases for the best effort users, we can maximize the overall cell throughput as well.

So, RAD-DS is derived from RAD and typically the RAD algorithms estimate the required scheduling activity of every user based on the users traffic profile and this required scheduling activity is used to make scheduling decisions. The required scheduling activity indicates the required time share by a user for data transmission. To allocate resources to VoIP users, RAD-DS performs scheduling in two stages ranking the users followed by resource allocation. RAD-DS specialty is that it inherits RAD and adds a delay sensitivity concept, which we will see below [26].
Ranking Function - VoIP users are assured these two QoS parameters - Guaranteed Bit Rate (GBR) and delay bound. VoIP traffic has strict delay bound of 50 ms and assures Guaranteed Bit Rate (GBR) of 16 kbps. In fact, RAD-DS scheduler differentiates the VoIP traffic from best effort traffic by the delay bound and GBR requirements (as the traffic type is not known at the base station). In the first step of the Ranking Function, the users are selected based on the available PDCCH channel capacity and among these users, the schedulable users are determined based on whether they have data buffered at the base station ready for transmission; and atleast one HARQ channel available for retransmission and

a pending retransmission. After determining the schedulable users, the Ranking Function ranks each user and higher the rank, the more the chances of that user being transmitted to the Resource Allocation Function (FD Scheduler) for further scheduling. The ranking is performed based on the below functions.

Required Activity Function - This function estimates the time share required by a user for data transmission and it varies based on the type of traffic i.e whether best effort traffic or VoIP traffic. The required time share of the VoIP users is directly proportional to the GBR. More time share is allocated to a user which has high GBR requirement, however the maximum time share that can be allocated to a user is 1. Expected throughput is also taken into consideration and as it decreases for a user, the required time share for that user is increased and more time share is allocated for that user giving such users more priority to ensure fairness. If you were wondering about the required time share for the best effort users, any excess time share will be shared equally among the best effort users.

Delay Sensitivity Function - VoIP has strict delay bounds and RAD-DS, for that purpose, takes into consideration the delay sensitivity concept that imposes certain time constraints to users with a delay bound. To prioritize VoIP traffic, the best effort users are assigned a flat delay sensitivity function i.e assigning this delay sensitivity function to be 1. The ranking function, referred to as TD scheduler, ranks every user based on these two functions and transmits the N_{mux} users with the highest metric to the Resource Allocation function.

Resource Allocation Function - The Resource Allocation Function, also referred to as FD scheduler, is responsible for assigning PRBs to the N_{mux} users received as input from the Ranking Function. It ranks users independently from the Ranking Function by using the PF scheduled metric. The metric is calculated to be the ratio of the instant throughout to the expected throughput of a user. This way, if the instant throughput is reduced, that user is assigned higher rank and is given more priority, which leads to allocation of fair amount of resources to different users. Also the Resource Allocation function takes into consideration the buffer size of the UE and allocates accordingly. If the buffer size is limited for a user and the scheduler assigns PRBs to that user, the chances are that the user might not be able to utilize the PRBs fully and this would lead to wastage of the spectrum. So to make sure that the limited buffer users aren't allocated too many PRBs beyond their capacity, candidate lists are maintained to map the association of the PRBs with the users respectively. This mapping will give scheduler the information regarding the buffer state of the users so it can consider the buffer size of the user and allocate PRBs. So at every TTI, the Resource Allocation function iterates over the metric values for each user in decreasing order and allocates PRB to user if user is not being allocated yet and if PRB is available, then removes the PRB from the candidate list so that the mapping is consistent and in the last step, remove user from the candidate list if the allocated PRBs can serve the user so that the data buffered at the base station can be transmitted.

Another advantage of RAD-DS is that it also considers retransmission requests from HARQ component so the number of HARQ channels allocated is 6 with the maximum number of retransmissions being 4, which means that corrupted packets can be retransmitted. Its important for the scheduler to cooperate with HARQ manager to prioritize scheduling of re-transmissions to ensure that the traffics target delay is not affected. Also, to support the delay sensitivity mechanism, HOL delay is taken into consideration, which means that the older packets in the buffer are given higher priority. HOL delay makes sure that the packet from a flow is processed within its delay constraints. Channel conditions are also taken into consideration for link adaption, which would ensure fairness. Different delay functions are used by the RAD-DS scheduler based on the type of traffic whether VoIP or best effort. If its VoIP, more aggressive delay functions are used such as the exponential delay function, which offers the best VoIP capacity as it preserves the delay sensitivity. But for best effort traffic, flat delay function is used. Moreover, in the simulations, the assumption is that if one user leaves the network, another user of the same traffic type is replaced, which is not a very realistic assumption when it comes to a real LTE Network. Moreover, RAD-DS doesnt take into account the Packet Loss Rate and the Queue Size, in which case if a buffer of a flow is close to its maximum limits, this flow needs to be scheduled right away before overflows. The computation complexity depends on the Resource Allocation function for maintaining and iterating over the candidate lists each time the scheduler wants to assigns resources.

MAC Scheduling Scheme for VoIP :

Taking into account delay and loss sensitive VoIP characteristics, not to neglect the fading effect, voice service implementation becomes particularly challenging. There are several schemes proposed for voice processing in addition to providing QoS, such as the smart blanking and anti-jitter playback buffering. But these schemes do not closely consider the fading effect and so the performance of voice service could be degraded. To overcome this issue and to increase the performance of voice services besides preserving the delay and loss sensitive characteristics, we need to prioritize VoIP traffic over other traffic. But by doing so, we are neglecting the QoS of other multimedia calls of different traffic type, which will lead to the overall system performance and effect the downlink throughput which is undesirable. So the idea is to support QoS of VoIP traffic but also to make sure there is no side effect on the downlink throughput and the overall system performance. To achieve this, we need a more dynamic approach for prioritizing VoIP traffic. Before delving into the details of this MAC layer scheduling scheme, it is also important to understand the resource management infrastructure that is defined.

Resource Management Infrastructure -The underlying infrastructure consists of three functions - Radio Admission Control (RAC), MAC scheduling, State Information

Management. RAC function is to decide whether to accept any new calls to pass them to MAC Scheduler for allocating resources. In the second stage, the MAC scheduler is responsible for assigning PRBs to ongoing calls. And in this scheme, each user call has separate transmission queues; so essentially, the MAC scheduler would be assigning PRBs to the transmission queues. This way, it helps in differentiating traffic types and also in independently handling downlink calls. So the allocated PRBs will carry data bits from the corresponding calls only and this is a good way to simplify scheduling. The state management function provides the first two functions, which are the RAC and MAC schedulers, the necessary information such as wireless conditions between the base stations and each user which is CQI, packet drop rates and average length of each transmission queue. Let us closely look at the MAC Scheduling function [27].

MAC Scheduling Function - The idea is to serve VoIP traffic by giving it more priority over other types of traffic so this function determines whether VoIP priority mode needs to be activated or not, and if it is activated, how long should it be activated before it negatively hits the other multimedia calls of different traffic types. To determine that, one important factor for consideration is the VoIP packet drop ratio. Once the VoIP priority mode is activated, only VoIP traffic is being served to maintain the QoS of VoIP, which means that all the resources will be allocated to ongoing VoIP calls. So essentially, the algorithm tackles three challenges to prioritize resource allocation to VoIP users - first, at every TTI, it determines if the algorithm needs to operate in Priority mode or Normal mode, second it defines a strategy to prioritize allocation of resources to VoIP users when in VPM mode (i.e if any ongoing VoIP calls are present), third, it characterizes a way to dynamically adapt the length of the VPM mode based on certain criteria, which we will see.

VoIP priority mode activation - At every TTI, the algorithm checks two things: 1. If there are any ongoing VoIP calls 2. The duration of the VoIP priority mode that is employed before exceeding the threshold. If the duration has exceeded the limit, the algorithm operates in normal mode where the resources are allocated to traffic flows based on queue length and channel quality. Otherwise, it operates in priority mode where VoIP users are given priority. But operating in the priority mode, left over PRBs are allocated to best effort traffic in the same way as the normal mode (considering the queue length and channel quality) to avoid wastage of resources.

Resource Allocation - This is done in two steps. First, it uses Channel Adaptive Fair Queuing (CAFQ) technique to rank the users based on the buffer queue length and the channel quality. In other words, a particular user will be prioritized if it has more buffered data in the queue and better channel quality compared to other users. Once the ranking of the user is completed, the scheduler uses Round Robin (RR) technique to allocate one PRB at a time to the user. The scheduler could assign more than one PRB to a call if the size of the packet is larger than the capacity of the PRB. The RR technique is repeated until there are no available PRBs or no call has data to send.

Dynamic Adaptation of priority mode - The length of priority mode is dynamically adjusted based on the VoIP packet ratio calculated at the eNodeB. The significance of VoIP packet drop ratio is that it indicates the number of ongoing VoIP calls based on the VoIP packet drop ratio. If the drop ratio is high, it is safe to assume that there are many ongoing VoIP calls, which are not being dedicated enough number of resources. On the contrary, low drop ratio indicates that all the VoIP calls are being satisfied with the amount of resources allocated. Using this strategy, the TTIs can be dynamically dedicated to the VoIP calls only. So with respect to the implementation part of it, we can pre-configure the minimum and maximum thresholds for the VoIP packet drop ratio. And if the ratio goes below the minimum threshold, we can increase the number of TTIs and vice-versa. So this scheme provides QoS for VoIP and neutralizes the negative effect on the overall system performance in case other types of traffic is present. The advantages of this scheme is that it considers the average queue length of the transmission queues, Channel Quality so that a user with continuous bad channel quality is not starved, size of the queue length, average throughput, VoIP packet delay and VoIP packet drop rate. But on the other side, this scheme doesnt provide Guranteed Bitrate (GBR) to the user. Moreover, it doesnt consider retransmissions from HARQ. Also, failure in consideration of the UE Buffer Length may result in packet drops and in turn result in retransmission of the dropped packets. Another negative effect is that the VoIP calls are short and have small amount of data to be transmitted, so the PRB capacity is not used up fully whereas there is higher PRB utilization for non-real time calls when operating in the normal mode.

Video Streaming Focused Algorithms

Dynamic Adaptive Streaming over HTTP (DASH):

Conventional scheduler designs like Proportional Fair (PF) and the MAX schedulers are based on average rate based functions which means that they either consider average throughput or the maximum throughput while making scheduling decisions, but this kind of approach is not suitable for video streaming applications as they fail to consider the playback buffer state of the users [28]. With the immensely growing video traffic over the Internet, it is challenging to ensure Quality of Experience (QoE), such as maintaining high quality, reliability, and minimal latency of about 10 ms, of rich multimedia applications like the video streaming applications. Also, there is a need for a novel video processing technique that would bypass the mobile network operator (MNO) and deliver the video content directly to the end user. MPEGs Dynamic Streaming over HTTP (DASH) is one such technique that transmits data from the Internet directly to the user by using HTTP data connections. So the goal of this scheme is to make the LTE network aware of HTTP metrics, which are being used in upcoming technologies such as DASH. So the underlying idea is that the traditional

schedulers like MAX, which serves the best users or the PF scheduler that serves users in a fairer manner, can be tweaked to consider the video characteristics at the users end. This would guarantee optimization of resource assignment for video transmission and also ensure fair resource sharing among multiple users. The algorithm proposes the following: - The DASH server contains media content with different versions in terms of resolutions or video qualities or bitrates. The server also maintains important information regarding the media content which is defined as media description representation (MPD). The MPD is very useful to the clients as they are responsible for managing the video sessions. The clients can use MPD to determine which video version could be suitable to download based on the channel quality of the user or the hardware capabilities supported by the user. Furthermore, the each video version is divided into small segments so that the client can dynamically adapt from one segment to another based on certain criteria, for instance, the bitrate requirements. The MPD at the DASH server also defines how to adaptively transfer from one video segment to another video segment so the client can derive this information provided by the MPD and allocate the necessary buffer sizes at its end in order to achieve interruption-free video content. The eNodeB should perform Deep Packet Inspection (DPI) to sniff the Media Description Packets (MPD) generated by DASH server and extract the required metric. - Use the extracted metric, which is the target bitrate, along with PF scheduler or Maximum Throughput scheduler to rank users based on channel quality and data rate requirement. If a user achieves maximum data requirement, the user is assigned low priority and weight and vice-versa. These values are taken into consideration for the next scheduling cycle, so this way, its guaranteed that users with certain target bitrate being met are penalized in resource assignment and also the users that do not meet the minimum requirement are given more priority and are favored. Considering Target Bitrate (TBR) in this model is advantageous because UEs that have maximum priority receive the top video quality. Throughput and also the Packet Delay Budget also are considered which is typically about 150 ms; and this model minimizes the playout interruptions by optimizing resource assignment in multi-user video streaming environment. On the downside, this scheme doesnt consider retransmissions from HARQ and also doesnt place a strict bound on the packet loss rate. Moreover, it doesnt consider HOL Delay and Queue Length. In addition to those, in real-time situations, the clients may not have infinite buffers to download data into, and considering infinite buffer lengths, the performance of video streaming might have variations. Complexity of this algorithm is effectively the Maximum Throughput ratio or the PF ratio that is used as a weighting factor to rank users and assign priorities. The PF ratio to make scheduling decisions per TTI can be accomplished in linear time so this mechanism is computationally robust. Another point is that this mechanism makes use of EPS (evolved packet system) messages to perform DPI that was discussed earlier. So the complexity also depends on the exchange on the number of messages exchanged with the BS in a highly mobile environment.

Real-Time (RT) Video Streaming Scheduler:
The real-time video streaming applications have stringent requirements, as the user shouldnt perceive any delay occurring in the video streaming [29]. So the packets need to be received by the user within a delay threshold, otherwise the packets are discarded and considered to be lost. For this to be achieved, the packet loss rate should be minimized. Typically, the packet scheduling algorithms use channel quality, average throughput and minimum average throughput and allocate resources to the users, and if there are any remaining resources, these schedulers allocate the remaining available resources to the users that being allocated the less number of resources. This process continues till all are resources have been exhausted. This approach is not ideal because its neglecting the fact that some users might not have any data to receive but are still allocated resources which will lead to wastage of spectrum and also the buffer of each user is not taken into consideration while allocating resources. To overcome that limitation, the RT Video Streaming Scheduler allocates one PRB to a user that has data to receive and the remaining PRBs are allocated based on the buffer sizes of the users. So radio resources are allocated if the users have data to be transmitted and the buffer information of users is considered to make sure theres no under-utilization of the allocated RBs. The algorithm works in 3 stages:

1) *Step 1* - The number of PRBs to be allocated to active users is determined in this step. It could be that for some active users, there is limited data in the eNodeB buffer, which leads to wastage of PRBs. So to avoid this, Step 1 also makes sure to reallocate the PRBs for those active users with limited buffer data.
2) *Step 2* - This step allocates the leftover PRBs from Step 1 to users with high HOL delay. All active users are sorted in ascending order of priority based on earliest deadline. Next PRBs are assigned to sorted active flows as per their deadline in iterative fashion until all the PRBs replenish.
3) *Step 3* - This step is executed after Step 1 if and only if, the total number of allocated PRBs in Step 1 exceeds the capacity. Once again, it uses HOL packet delay to decide from whom to take back PRB. PRBs from users with HOL packet far from expiry are retracted.

HOL packet delay is considered which has a threshold set to 20 ms and also the Delay Threshold; and packets are considered to be lost if they dont reach within this threshold. Packet loss rate is also minimized by keeping it below a threshold. Delay sensitive applications can be supported by considering the average instantaneous downlink SNR and packet delay information in addition to the packet arrival information. This scheme also considers the instantaneous downlink SNR values, average throughput and minimum average throughput for each user. On the downside, this

algorithm doesnt consider retransmissions from HARQ and it needs to keep aside certain amount of resources to handle HARQ retransmissions. Resources are allocated based on the ratio of data in the buffer of each user over the sum of data in the buffer of all users. Priority is assigned to users that have more data in their respective buffers and these active users are sorted in descending order of priority, so with increasing number of active users in a cell, the complexity might vary.

Channel-Adapted and Buffer-Aware Packet Scheduling:
The importance of channel quality has been emphasized in almost all the algorithms [30]. However, some of the algorithms assume that the buffers at the user terminals are infinite and that every arriving packet can be stored and buffer overflow wouldnt occur, but this assumption is not feasible in a real LTE network. So the channel-adapted and buffer-aware algorithm has been proposed for the integrated consideration of the channel quality that would increase the throughput and ensure fairness, and also finite buffers that would minimize the buffer overflow and packet loss. So the CABA Scheduling algorithm proposes two important components:
Channel Quality Index (CQI) and Buffer Status Report (BSR)- Each user has one queue at the eNodeB with different length from the others. CQI is sent to eNodeB scheduler in MAC layer. If the CQI is good, then a higher modulation scheme such as 64QAM is used to support the wireless channel and increase the throughput. On the other hand, 16QAM is used if the CQI is low to ensure data transmission reliability. This approach is an effort of the PHY-MAC layer jointly. This way, the system throughput is maximized and fairness is also ensured. So we can conclude that that if the channel condition is perfect and the instantaneous data rate is high, the user is assigned more priority and this would indeed boost the system performance. And we also calculate the average data rate, which indicates that the user is scheduled more frequently, so we decrease the priority of this user to ensure fairness to other users. And, buffer overflow will result in packet loss and to avoid this, each user will send a Buffer Status Report (BSR) to the eNodeB scheduler. The BSR sent by the UEs with uplink provides information about the amount of data in UE buffer. Effectively, the users with less available buffer space will be prioritized and to minimize the packet loss which was one of the goals of this algorithm. And the CABA algorithm prioritizes RT services over Non-RT services, so a weighted factor is assigned to RT traffic considering both have the same channel condition. Another weighted factor that is a ratio of the Guaranteed bitrate to the average throughput can be used to assign priority to RT traffic, to satisfy its data rate requirement. And if the average data rate is lower than the GBR, the weighted factor can be used to increase the scheduling priority. This way, CABA maximizes the system throughput, minimizes the packet dropping rate, ensures fairness and also prioritizes RT traffic over Non-RT traffic.

V. Critical Analysis of Algorithms

Algorithms discussed in this paper, use different QoS parameters and decision making technique to design scheduler that provides service differentiation to diverse traffic types. In order to present comparative discussion, algorithms discussed in previous section are analyzed in terms of its system model and simulation model.

System Model Analysis
System model of the algorithms discussed in this paper are benchmarked against following attributes:

1) *Re-transmissions and Spectral Efficiency* - As discussed previously, the scheduler needs to co-operate with HARQ manager to ensure re-transmissions from HARQ manager are prioritized to prevent the traffic from exceeding its target delay. In addition, the transmission channel between the base station and the user may not be perfect. That is, interference in the channel may cause the transmitted data to get corrupted by the time it reaches the user. For this reason, the scheduler needs to adapt the transmission rate based on the CQI report from user. In other words, factors such as re-transmission and spectral efficiency define robustness of the scheduler in real LTE system. Thus, it is vital to compare algorithm design in terms of whether it prioritizes re-transmissions and ensures that the spectrum is efficiently utilized.
2) *Buffer Length* - In real LTE system, queue length associated with the traffic at base station is finite in size. In addition, the buffer length on user's device is also limited in size. Hence, in order to avoid queue from overflowing and make the scheduler deployable in real LTE system, the algorithm needs to give consideration to queue length at both base station and at user's device when making scheduling decisions.
3) *Computation Complexity* - Because the packet scheduler allocates resources every TTI - 1 ms, algorithm complexity needs to be low. Some of the factors that result in low complexity include, making decision per-PRB, decoupling the scheduler into different such as ranking module and resource allocation module, and linear computational problems.
4) *Types of Suitable Traffic* - As mentioned previously, each algorithm uses different metrics that address different QoS issues. In other words, an algorithm optimal for one traffic type may not be optimal for other traffic types. Traffic types suitable with the respective algorithm will provide high level overview of what QoS issue the algorithm addresses.

Simulation Model Analysis
Simulation is an important way of verifying functionality of the proposed algorithm. The factors considered in simulation plays a crucial role in defining effectiveness of the algorithm. We evaluate simulation model of the proposed algorithms in terms of following factors:

1) *Mobility Model* - LTE provides high performance for users mobile in range of 15 km/h - 120 km/h; in addition, it maintains connectivity for mobility of up to 350 km/h [28]. This necessitates the scheduling algorithm to operate efficiently in low-high mobility.
2) *Multiple Types of Traffic* - Supporting diverse traffic types is one of the key features of LTE network. In other words, a robust algorithm should make optimal decision when multiple traffic types are present. Thus, the algorithm should be simulated in an environment with multiple types of traffic.
3) *Packet Arrival Model* - Rate of traffic generation depends on the type of application, which can be modeled using probability distribution. For example, in VoIP application, arrival of calls can be modeled using Poisson distribution and traffic of an ongoing call or video streaming application can be modeled using Exponential distribution. In other words, type of traffic model used will characterize the type of traffic class suitable for the respective algorithm.
4) *Overload State Vs Non-Overload State* - Depending on number of users present, the system may operate to its full capacity i.e. overload state. Testing the algorithm in presence of high number of users and low number of users will help judge robustness and efficiency of the algorithm.

Guaranteed Bit-Rate (GBR) based algorithms

GBR based algorithms are comprised of Priority Set Scheduler (PSS), Multi-QoS Aware Fair Scheduler (MQFS), and Dynamic Hybrid Scheduler (DHS).

System Model Analysis

1) *Comparison* - It is important to note that although PSS and MQFS prioritize GBR traffic over non-GBR traffic, it does not prioritize different GBR traffic flows based on delay. DHS on the other hand, not only prioritizes GBR traffic over non-GBR traffic, it also prioritizes GBR traffic flows based delay. This ensures that traffic close to its target delay is given more importance. For example, consider a system with VoIP and real time gaming flows. Both VoIP and real-time gaming are GBR flows and have target delay of 100 ms and 50 ms respectively. Let's assume that real-time gaming flow is close to its target delay and needs to be transmitted before VoIP traffic. However, in case of PSS and MQFS that prioritizes GBR flows based on past throughput and weight associated with QCI Class respectively, rank of VoIP traffic might be higher compare to real-time gaming flow in which case real-time traffic transmission delay will exceed its respective target delay. This will in return result in noticeable delay to human perception that will negatively affect user experience.
2) *Re-transmissions and Spectral Efficiency* - GBR based algorithms discussed in the paper assume re-transmissions from HARQ manager to be part of the same traffic flow queue as it was before the original transmission. In other words, re-transmissions and new traffic share the same queue. However, GBR based algorithms do not prioritize re-transmissions. In case of PSS or MQFS that do not consider the delay sensitivity factor, there is a possibility that newer traffic will be scheduled before the re-transmissions in which case target delay deadline for the re-transmitted packet will be violated. In addition, all the GBR based algorithms consider CQI from user to adapt the transmission rate based on channel quality between user and base station..
3) *Buffer Length* - GBR algorithms assume the queue length at base station and user device to be infinite in length. This assumption may lead to buffer overflow when deployed in real networks as in real systems, queue is finite in size.
4) *Computational Complexity* - In the worst case, all the GBR based algorithms take about $O(N^2)$ as discussed in previous section.
5) *Suitable Applications* - Traffic types that require GBR service as mentioned in QCI Classes *(refer table in appendix,)* are suitable. However, PSS and MQFS may not be able to meet the delay requirements of the GBR traffic flows. This is because it does not prioritize GBR flows based on delay requirement. For instance, consider the example discussed above where real-time traffic flow might be closer to its target delay compared to VoIP traffic. However, as PSS and MQFS do not consider delay requirements when calculating rank, rank of VoIP traffic might be greater. This may cause the the real-time gaming flow to be transmitted with delay greater than the target delay. If the delay is large enough, it may become apparent to user's perception causing the user experience to be out of sync. In addition, applications generating traffic at high data rate such as video streaming applications, may experience performance degradation as the GBR algorithms considered neglect queue length at base station and UE.

Simulation Model

1) *Mobility Model* - PSS and DHS do not consider mobility of the users in the system. DHS does not consider mobility of the nodes because the simulation is performed in HSDPA network where performance is measured in terms of number of arriving calls supported in presence of varying load. MQFS on the other hand, considers random waypoint model as the mobility model. In random waypoint model, mobile nodes move in random direction with random velocity. This allows to investigate performance of the scheduler for users walking (at speed lower than 5 km/hr) and users traveling in car (at speed varying from 50 km/hr to 120 km/hr). As it can be observed, simulation of PSS and DHS do not provide clear estimate on operation of the respective scheduler in real mobile environment due to which it may not be deployable in real LTE system.
2) *Types of Traffic* - PSS considers VoIP calls and DHS

considers cellular calls. MQFS on the other hand, considers diverse traffic types with 2 real-time and 2 best-effort flows. Real-time flows include VoIP and buffered video stream; best-effort traffic include HTTP and FTP. Lack of usage of diverse traffic types in PSS and DHS prevents the simulation simulation from observing operation of scheduler in presence of diverse traffic types.

3) *Packet Arrival Model* - In PSS, simulation model considers Poisson call arrival for GBR traffic where a call arrives in the LTE system every regular interval, in this case, every 8 ms. The non-GBR traffic is bursty i.e. it arrives at random time, which is buffered in 2 Mbps buffer. DHS takes probabilistic approach in defining packet arrival model. Each call connection is in ON state with probability $P_{ON} = 0.3$; when in ON state, arrival data rate is 1.6 Mbps. In MQFS, VoIP traffic considers exponential traffic model for inter-call arrival and talk time; Video streaming considers constant arrival of 24 frames per second; HTTP traffic considers bursty traffic model; Lastly, FTP traffic considers uniform inter-request time.
4) *Overload State Vs Non-Overload State* - In PSS, the simulation defines overload and non-overload state for both GBR and non-GBR traffic. For GBR traffic, overload state is defined in terms of arrival rate of the data, average call rate and number of UE present. The simulation considers data rate of 128 kbps, average call rate of 24 UE/sec/cell and maximum of 84 UEs per cell in case of overload state; and 256 kbps data rate, 12 UE/sec/cell call rate and 42 UEs per cell for non-overload state. In non-GBR traffic, the overload and non overload states are defined in terms of average call rate, and number of UEs in the system. Overload state considers arrival of 8 UE/sec/cell and 25 maximum UEs per cell; non-overload state considers arrival of 6 UE/sec/cell and 10 UE per cell. MQFS considers 85 users comprising of 20 VoIP, 5 video streaming, 40 HTTP, and 20 FTP users. However, MQFS simulation does not define criteria for overloading and non-overloading state. DHS on the other hand tests performance of the algorithm in 59% traffic load.

Delay driven algorithms

delay driven algorithms are comprised of M-LWDF, HASA, Exp/PF, TLS, DPS, Log and Exp Rule, and CGT.

System Model Analysis

1) *Comparison* - From QCI Classes *(refer table in appendix)* it can be observed that multiple GBR and Non-GBR traffic have same target delay. In other words, making scheduling decisions simply on the basis of delay bound may cease GBR flows from receiving GBR service. For example, consider a system with GBR based video streaming traffic flow and non-GBR based traffic flow, both of which have target delay of 300 ms. Assume that packets from both these flows have been waiting in the queue for same amount of time. When the scheduler has the opportunity to pick one of these flows to schedule, it should schedule the GBR based video streaming flow considering it is below its GBR value. This decision ensures that GBR value of the flow is met, guaranteeing a pre-defined promised service to the traffic. As a simple analogy consider 2 queues at a ride in an amusement park. Let one queue be Fast Lane and the other a normal queue. Fast Lane is equivalent to GBR traffic flow and normal queue is equivalent to non-GBR flow. Let there be 2 people with same target delay - in this case expected time by which the person needs to be served - such that one has Fast Lane pass the other person does not. Person with Fast Lane pass enters the Fast Lane and the other enters normal queue at the same time. Assume that Fast Lane queue is longer and normal queue length is shorter such that the 2 people who enter the queue at the same time meet at the head of the line. That is, each have waited for equal amount of time to be served. Given the Fast Lane throughput is less than desired, person in the Fast Lane will be served prior to person in normal queue. In a nutshell, the algorithm needs to make sure it not only satisfies the delay bound of the traffic flow but also satisfies the GBR service. However, none of the algorithms discussed in this paper except for HASA considers GBR along with delay bound. In addition, even HASA only considers GBR for re-transmissions from HARQ as new traffic is served as per M-LWDF.
2) *Re-transmissions and Spectral Efficiency* - HASA is the only algorithm that prioritizes re-transmissions from HARQ. Upper layer component of TLS on the other hand, leaves aside resources for re-transmissions. However, it does not prioritize and schedule re-transmissions, it leaves this responsibility on lower layers. Other algorithms neglect the re-transmissions from HARQ. In addition, all the algorithms determine optimal transmission rate based on CQI from user.
3) *Buffer Length* - TLS is the only algorithm that considers queue length when making decision. Rest of the algorithms assume the queue length at the base station to be infinite in size. In addition, none of the delay driven algorithms consider buffer length at user device, which may cause queue overflow when deployed in real LTE network.
4) *Computational Complexity* - M-LWDF, HASA and Log rule take $O(N.n1.5)$ time where N is number of flows and n is number of digits of precision at which the function is to be evaluated. TLS takes, $O(N.M_{i\_max})$ where N is number of real-time flows and M_{i_max} is the maximum impulse length among the real-time flows. DPS on the other hand takes $O(N.M)$ where N is number of flows and M is number of resource blocks to be scheduled. Lastly, worst case performance of Exp Rule and CGT based algorithm is $O(N.M(n).log(n))$ where N is number of flows, M(n) is computational complexity

of multiplication operation, and n corresponds to n-bit integer.

5) *Suitable Applications* - Delay driven algorithms are suitable to both real-time and best-effort traffic mentioned in QCI Classes in Appendix. However, the algorithms may not be able to fulfill GBR requirements of the traffic flows as it neglects this factor in its design.

Simulation Model Analysis

1) *Mobility Model* - In M-LWD, Exp/PF and DPS, users constantly move in random direction with velocity in the range of 1-100 km/hr; in HASA users move at constant speed of 10 km/hr in random direction. In addition, TLS considers random waypoint model with velocity of 3 km/hr and 120 km/hr. Lastly, Log Rule, Exp Rule and CGT based schedulers consider constant movement in random direction at 3 km/hr. As it can be observed, most of the delay driven algorithms except for HASA, Log Rule, Exp Rule, and CGT based scheduler perform experiments in low to high mobility to ensure scheduler makes optimal decision to pedestrian user and vehicular user.
2) *Types of Traffic* - M-LWDF and Exp/PF simulates 128 kbps video streaming traffic; HASA simulates 256 kbps video streaming traffic. TLS simulates real-time and best-effort traffic. As part of the real-time traffic it utilizes VoIP and video streaming application; as part of best-effort traffic it considers infinite-buffer buffer being full always. DPS on the other hand only considers buffered streaming traffic. Log rule and Exp rule examine buffered video stream and live video stream traffic. Lastly, CGT base scheduler studies the performance of the scheduler with 50% VoIP traffic and 50% video streaming traffic. It can be seen that most of the algorithms consider two most demanding multimedia applications i.e. VoIP and video streaming traffic as part of the experiments. However, M-LWDF, Exp/PF, HASA, and DPS only consider buffered streaming traffic which may not be enough to measure the performance of the scheduler, as VoIP that provides calling capabilities over IP protocol is an integral traffic type in LTE that needs to be tested.
3) *Packet Arrival Model* - Buffered video stream traffic used in M-LWDF, Exp/PF, HASA, DPS, Log Rule and Exp rule is modeled by truncate pareto model. Live video traffic in Log Rule and Exp Rule, VoIP traffic in TLS and CGT based scheduler is modeled using ON and OFF markov process where the markov process spends equal amount of time in ON state and OFF state. When in ON state, traffic arrives at constant rate. In addition, video stream in TLS and CGT based scheduler arrives at constant rate.
4) *Overload State Vs Non-Overload State* - M-LWDF and Exp/PF considers 80-120 video streaming users in the simulated cell. Log Rule and Exp Rule considers 57 cells with 3 base stations and 20 users per cell; in total it considers 1140 users. DPS experiments with 20-100 users and CGT based scheduler examines the performance with 60 users - 50% VoIP and 50% video streaming. However, none of these algorithms define clear criteria for overloading and non-overloading state, which makes it difficult to test robustness of the scheduler. Lastly, TLS considers multi-cell scenario with 19 cells, each consisting of 1 base station. The algorithm is simulated for 10, 15 and 20 users per cell, with 10 implying non-overloading state and 20 implying overloading state.

System Centric Algorithms

System centric algorithms are comprised of OSDRA algorithm.

System Model Analysis

1) *Comparison* - OSDRA gives flexibility to the network operator by allowing them to utilize multiple QoS parameters when prioritizing user for scheduling and fine tune weights of the QoS parameters as per network usage. This enables network operator to provide QoS to users based on network usage. In addition, OSDRA's main objective is to just rank traffic flows and assign RBs based on the QoS parameters.
2) *Re-transmissions and Spectral Efficiency* - It does not prioritize re-transmissions from HARQ. It assumes that resources available are for newer transmissions. In addition, the algorithm neglects CQI from user when assigning resources. This may not result in optimal resource assignment. For example, consider 2 flows A and B. Let rank calculated by OSDRA for A be higher than rank of B. Because rank of A is higher, flow A is assigned as much resources as it needs upto maximum of MBR value if it is a GBR flow or upto AMBR value if it is a non-GBR flow. Without considering CQI from user, it is possible that scheduler assigns resource blocks to users that do not comply with CQI from user. In other words, the traffic flow will not be able to utilize the resource to its full capability. That is, if the CQI value is low, indicating bad channel quality and RB assigned can support high throughput, user will not be able to transmit up to maximum supported throughput. This will in turn reduce spectrum efficiency of the system.
3) *Buffer Length* - OSDRA considers queue length of the traffic at the base station. In addition, the weight associated with the queue length can be used to fine tune the priority depending on the rate at which data arrives at queue. This prevents the queue from overflowing.
4) *Computational Complexity* - Computation complexity of OSDRA depends on the ranking function, which uses hyperbolic tangent $(\tanh)$ to calculate priority, and resource assignment function that uses fractional knapsack approach to assign resources based on ranks. $\tanh$ can be computed using Taylor series in $O(n^{\frac{1}{2}}.M(n))$, where M(n) is computational time of multiplication operation, which is linear []. n refers to number of digits of precision desired for evaluation. This means that for N

flows, it will take about $O(N.n^{1.5})$ time. On the other hand, resource assignment using Fractional Knapsack can be performed in single pass i.e. $O(N)$ time.

5) *Suitable Application* - OSDRA design gives flexibility to the network operator to tune QoS parameters based on the network capacity. This makes it suitable for all types of traffic described in QCI Class *(refer to table in Appendix)*.

Simulation Model Analysis

1) *Mobility Model* - The simulation of OSDRA does not consider mobility of the users. It assumes that data for the respective traffic queue arrives at the queue based on desired traffic distribution.
2) *Types of Traffic* - To evaluate robustness of the algorithm in real LTE system, it experiments with all the traffic types defined in QCI Classes.
3) *Packet Arrival Model* - The VoIP traffic is modeled using exponential distribution. In addition, talking and silent state transitions were modeled as geometric distribution. On the other hand, different types of data traffic such as HTTP, buffered video stream and gaming were modeled as series of ON and OFF times. That is, the traffic arrives to the queue at constant rate when only in ON state. The simulation does not give details on ON and OFF probabilities used for different traffic types.
4) *Overload State Vs Non-Overload State* - The simulation of OSDRA defines 50 active connections as normal or non-overloading state and 100 active connections as overloading state.

VoIP based algorithms:

The VoIP based algorithms comprise of Required Activity Detection with Delay Sensitivity (RAD-DS) and MAC scheduling scheme for VoIP (VoIP priority mode).

System Model Analysis

1) *Comparison* - The maximum packet delay for a VoIP packet is 100 ms and both the RAD-DS scheduler and the Mac Scheduler take into account target delay. RAD-DS has the delay sensitivity mechanism and to preserve that, it also makes sure to give importance to HOL Delay, so in that way older packets in the buffer are given more priority. When the HOL delay is beyond a threshold, RAD-DS scheduler schedules the VoIP users over the best effort users. The maximum packet delay for the RAD-DS is 50 ms and for the Mac Scheduler it is 20 ms. But the Mac scheduler doesnt take into consideration the HOL Delay unlike the RAD-DS scheduler. For example, if theres a system with two VoIP flows, both of which having a packet delay of 100 ms. Considering that they have been waiting in the queue for the same amount of time and one of the flows has its buffer filling up fast, so if the scheduler doesnt consider HOL delay, buffer overflow will occur and will lead to re-transmission of the packets, which would effect the target delay as well. So its important for the scheduler to give consideration to HOL delay to preserve target delay. When it comes to packet loss, VoIP can tolerate upto maximum of 10 % before the user can perceive it. So the VoIP algorithms should consider the packet loss rate and minimize it. For example, in a voice conversation, if at one end the packet loss rate is maintained below 10% and at the other end it is more than 10%, the conversation will be out of context and result in poor user experience. However, RAD-DS doesnt consider packet loss rate and packets received after a delay threshold will be dropped and considered as lost, due to its stringent requirement on the target delay. However, the Mac Scheduler takes into account the packet loss rate and maintains it 1% or below and defines the threshold to be 5%, even with increasing number of VoIP calls. So the Mac Scheduler triumphs the technical difficulties in handling the VoIP flows, which are both delay and loss sensitive, unlike the RAD-DS scheduler that handles only delay. The Mac scheduler can handle both because the VoIP priority mode assigns the PRBs first to VoIP calls. In fact, the packet drop rate is very important for the Mac Scheduler as it adapts the duration of the VoIP priority mode based on the packet loss rate (as discussed in the Description section). VoIP users are assured GBR of 16 kbps by the RAD-DS scheduler but the Mac scheduling scheme doesnt provide GBR to the user. So the RAD-DS scheduler not only prioritizes GBR traffic flows over non-GBR based traffic, but also prioritizes the GBR traffic flows based on delay. This ensures that the flow close to its target delay is given more importance for given two GBR flows. But if there is a GBR VoIP flow and a non-GBR flow such as Live Video Streaming or Interactive Gaming, the packet delay value for both is the same which is about 100 ms and so the scheduler would pick the GBR VoIP flow and schedule it first which may result in poor experience for the Live Video Streaming flow.
2) *Re-transmissions and Spectral Efficiency* - RAD-DS considers the retransmission requests from the HARQ component and the number of HARQ channels allocated for retransmission is 6 with the maximum number of retransmission being 4. The advantage of allocating separate channel for retransmissions would be to give it priority, for example if it is mixed with other flows, chances are that the scheduler picks the one that is first, neglecting the delay factor, in which case the target delay for the retransmission packet is affected. However, the Mac Scheduling Scheme doesnt consider retransmission from HARQ. Also both the schedulers take into consideration the CQI from user to adapt the transmission rate based on the channel quality and also to ensure the users with bad channel quality arent starved. For example, if there are two flows and the scheduler picks the one according to GBR or target delay but fails to consider CQI, in which case it will end up assigning more PRBs to that user than it can support and all the resources will be wasted as the user will not be

able to utilize the PRBs fully which will result in poor spectrum efficiency.

3) *Buffer Length* - RAD-DS doesnt take into consideration the queue length that makes it crucial to schedule the flow, which has the buffer close to its maximum limit, before it overflows. The Mac Scheduling Scheme takes into consideration the size of queue length at eNodeB, in fact the scheduling order is calculated based on the queue length as one of the parameters, so longer the queue length, earlier the corresponding call will be scheduled; but fails to consider the buffer length at the UE, which will result in packet drops and in turn result in retransmission of the dropped packets. However, with respect to VoIP calls, the queue lengths will be short as the VoIP packet sizes are small and so VoIP calls have small amount of data to transmit. So this could be a negative effect as the calls may not be able to utilize the PRBs fully. So in the normal mode of operation (other than the VoIP priority mode), if there is a non-realtime call and a VoIP call, the non-realtime calls have longer queues so the scheduler schedules the non-realtime calls due to the higher queue length.
4) *Computational Complexity* - the Mac Scheduling algorithm, the complexity depends on the number of active flows followed by allocating the PRBs in a round robin way. For RAD-DS algorithm, the complexity depends on the number of users, delay function used and traversing the candidate lists that maintains the mapping (as described in description section).
5) *Suitable Applications* - These algorithms support GBR traffic and delay sensitive applications such as conversational voice; conversational video (live streaming), non-conversational video (buffered streaming) and real time gaming.

Simulation Model Analysis

For RAD-DS, the simulations are carried out to access VoIP capacity and are run with VoIP traffic only. And for the VoIP priority mode, the simulations are carried out both in VoIP priority mode and the normal mode on nodes operating using CAFQ algorithm. The ability to support exceeding VoIP calls without VoIP priority deteriorates as the packet loss rate increases. With the help of VPM, the packet drop rate is 1% as when compared to 5% in the previous case when the number of VoIP calls is 10. For VoIP calls = 50 as well, the packet drop rate is almost about 1%. Also, the throughput decrease ratio was less than 3% by adaptively adjusting the consecutive VoIP priority mode.

1) *Mobility Model* - For the RAD-DS, the UEs in the system can move in random direction at the rate of 3 km/hr, and for the Mac Scheduling scheme, the UEs in the system can move in random direction at the rate of 1 km/hr.
2) *Types of Traffic* - For RAD-DS, the simulation considers real time and best effort traffic. Real time traffic includes VoIP and best effort traffic includes HTTP and FTP. And for the Mac Scheduling scheme, This scheme serves only VoIP traffic and the limit of the TTIs dedicated to VoIP traffic is dynamically adjusted based on the number of the VoIP calls present. But operating in VPM mode, the left over PRBs are allocated to best effort traffic.
3) *Packet Arrival Model* - For RAD-DS, the simulation model considers GBR traffic with inter-arrival rate of 20 ms and SID packet inter-arrival time of 160ms. And for the MAC scheduling scheme, the simulation model considers GBR traffic with inter-arrival rate of 20 ms.
4) *Overload State Vs Non-Overload State* - For RAD-DS, as the total number of users per cell increases, the percentage of users that doesnt fulfill the satisfaction criterion also increases (the VoIP user satisfaction criterion is having a packet error rate less than 2%). And with respect to the MAC scheduling scheme, this simulation doesnt necessarily define the Overload and Non-Overload states.

Video Streaming based algorithms:

The VoIP based algorithms comprise of Dynamic Adaptive Streaming over HTTP (DASH), Real-Time Video Streaming Scheduler and Channel-Adapted and Buffer-Aware Packet Scheduling (CABA).

System Model Analysis

1) *Comparison* - These algorithms have tight requirements on delay and they prioritize real time traffic over non real time traffic. As per the QCI table, the packet delay for conversational video and live video streaming is about 100 ms. All these three algorithms take into consideration the packet delay and strive to maintain it below 100 ms. Also these algorithms consider GBR when making scheduling decisions, making sure that the users that do not meet with the minimum requirement are given more priority. So if there are two flows one with GBR conversational video and another one with non-GBR video live streaming and both of them having a packet delay of about 100 ms, the scheduler would pick the GBR based flow neglecting that the other flow has the same packet delay and that it is reaching the threshold. So the algorithms need to prioritize GBR traffic without neglecting the target delay factor.
2) *Re-transmissions and Spectral Efficiency* - These algorithms dont consider retranmissions from HARQ. The video packets, if lost, generate a variable impact on the users perceived QoE. So these algorithms need to adapt HARQ for enhanced QoE. Both the DASH algorithm and CABA consider CQI from the user and adapt the transmission rate based on the channel quality between the user and the base station, but the RT-Video Streaming scheduler doesnt. So if there are two flows, one with bad channel quality, the RT video streaming scheduler allocates the PRBs to both the flows and that flow wont be able to utilize it fully, leading to poor spectrum efficiency.
3) *Buffer Length* - These algorithms consider the buffer

length in order to achieve interruption-free video content. However, DASH algorithm considers that the user has infinite buffers to download data into and that buffer overflow wouldnt occur, but this is not a realistic assumption when it comes to a real LTE network. The RT video streaming scheduler and CABA prioritize users having less buffer space as it means there is more data send, leading to minimization of packet loss.

4) *Computational Complexity* - Complexity of DASH is effectively the Maximum Throughput ratio or the PF ratio that is used as a weighting factor to rank users and assign priorities. The PF ratio to make scheduling decisions per TTI can be accomplished in linear time so this mechanism is computationally robust. For the RT video streaming scheduler, resources are allocated based on the ratio of data in the buffer of each user over the sum of data in the buffer of all users. Priority is assigned to users that have more data in their respective buffers and these active users are sorted in descending order of priority, so with increasing number of active users in a cell, the complexity might vary.
5) *Suitable Applications* - These algorithms has potential to support traffic flows requiring tight delay bounds and low packet loss rate. In other words, delay sensitive applications such as conversational voice, live streaming, buffered streaming and real time gaming can be supported.

Simulation Model Analysis

For DASH, the simulations are carried out on both the modified MAX and PF schedulers and they are shown to outperform the traditional schedulers for video streaming. Occurrences of video interruptions are being monitored during the DASH session.

1) *Mobility Model* - DASH simulation doesnt consider the mobility of the nodes. In the RT Video Streaming scheduler simulation, users constantly move at speeds between 1-100 km/hr in random directions. For CABA, the users move at 3km/hr.
2) *Types of Traffic* - For DASH, HTTP traffic is considered which is very bursty in nature. The best effort traffic flows have not been considered for simulation. In RT video streaming scheduler, the simulation only considers video streaming traffic. For CABA, the traffic type considered for simulation is 50% FTP and 50% Video.
3) *Packet Arrival Model* - For DASH, Video streaming considers a constant arrival time of 24 frames per second. For the RT Video Streaming Scheduler, the packets are streamed into users buffers from variable bit rate (VBR) source encoders running at 128 kbps in average and the inter- arrival time of each packet follows the Truncated Pareto distribution. For CABA, the video frame rate is 25 frames/sec
4) *Overload State Vs Non-Overload State* - In DASH, the number of cells is 19 sites with 3 sectors per site. And the number of UEs ranges from 10-60. With modified MAX scheduler for 30 UEs, the interruption probability increases more than 10% and for 60 UEs, not even 20% of the users receive the video without interruptions when the DASH session is 5 minutes. With the modified PF scheduler, the interruption probability only reaches 30%, which is slightly better than the modified MAX scheduler. In the RT Video Streaming Scheduler, the simulation considers a single hexagonal cell containing 20 to 100 RT video streaming users. The PLR performance decreases with increasing number of users as the RT video streaming packets get buffered and packets that arrive with HOL delay beyond the threshold are discarded. It is shown that approximately 80 and 50 users can be supported by this algorithm and ensuring the QoS requirement of the RT video streaming which is to maintain 1% PLR threshold. For CABA, there are 16 number of users and the number of cells is 3.

VI. Future Work

The algorithms discussed in this paper adopt design strategy with a goal of providing QoS to multimedia applications, ensuring fairness, increasing system throughput and spectral efficiency. However, the proposed algorithms make certain key assumptions and neglect factors that maybe challenging when deploying them in real LTE network. Moreover, the proposed algorithms adhere to LTE specifications in 3GPP-Release 8. 3GPP has initiated process of evolving LTE network to LTE-Advanced (LTE-A) network and has released LTE-A specifications in 3GPP Specification Release Version 10 and 11. This evolution comes in the wake of International Mobile Telecommunications-Advance (IMT-A) specifications drafted by International Telecommunication Union - Radiocommunication Sector (ITU-R) to meet future traffic requirements. Below we discuss challenges that could arise if current algorithms were deployed in real LTE network, we propose optimizations to System Model and Simulation Model of the proposed algorithms and discuss the key LTE-A design aspects that can effect the design of current and future scheduling algorithms.

Challenges

- *Inaccurate Channel Quality Index*- Downlink transmission in LTE occurs from base station to the user device - also known as User Equipment (UE) via transmission channel over the air. User periodically or aperiodically (upon request from base station) sends Channel Quality Index (CQI) to the base station to notify the base station about the quality of the transmission channel. In case of periodic CQI reporting, the interval period between each CQI report is determined based on configuration parameters such as System Frame Number (SFN), used to keep base station and UE in sync, and is exchanged between base station and UE during channel establishment.
  CQI is a important factor that is used to determine amount of RBs to be assigned to the user. For example,

consider a scenario where a user is sitting in a coffee shop and accessing video stream on an LTE enabled mobile device. Furthermore, assume that passing vehicles such as trucks degrade the quality of transmission channel between the base station and the user. This channel degradation, reduces ability of the transmission channel to transmit more data on a given RB. The scheduler uses the channel quality information in CQI to assign appropriate amount of RBs to the user. In the example discussed above, when the channel quality is bad due to passing trucks, CQI reported will be less and thus, scheduler will assign more RBs to the user to make up for the degradation in transmission rate. On the other hand, when channel quality improves with decrease in traffic, capacity of channel to transmit more data over an RB increases. Hence,the scheduler will assign less amount of RBs to the user.

However, the algorithms discussed in this paper assume that the CQI delivery from the user to base station to be instantaneous and accurate. However, in practical scenario, CQI reports are not delivered to the packet scheduler instantaneously. Factors such as interference from communication between neighboring devices and base station, and natural blockage of transmission due to a building, traffic or weather can cause the CQI report to get corrupted. Hence, the CQI reported from the user is first processed by the base station for any inconsistencies before it is accessible by the packet scheduler. The CQI processing delay incurred at the physical layer is about 3 ms [31]. Moreover, one such scenario is presented in Figure 4 where the darker area indicates good channel quality, and lighter area indicates bad channel quality. When user detects change in channel quality, it reports CQI to the base station, indicated by the arrow in the figure. However, because of factors such as interference, processing delay, and natural blockage, packet scheduler receives the CQI after some latency as depicted by the dotted line in Figure 4. In addition, if the channel quality between the user and base station is volatile, user will send CQI more often. As observed in the figure, often change in channel quality along with latency in delivery of CQI to packet scheduler can create period of inaccuracy. This period of inaccuracy can degrade performance of the scheduler. This is because scheduler determines optimal transmission data rate for user based on last known CQI, which may differ from the current actual channel quality. For example, if the delayed CQI delivery indicated good channel quality, scheduler will transmit data at higher rate in order to increase throughput. But if the current channel quality of the user has degraded, user will not be able to successfully receive all the data. In other words, packets will be discarded and re-transmission would be required. In contrast, the throughput will be low if the delayed CQI delivery indicated bad channel quality and actual channel quality improved during the latency period. Thus, in practical scenario, an optimal scheduling algorithm needs to consider channel estimation error or predict actual CQI in order to mitigate performance degradation caused by latency in CQI delivery. There are solutions already being proposed in [32][33][34] that utilize past CQIs from user along with other mathematical approximation methods such as stochastic approximation to precisely predict current CQI in linear time.

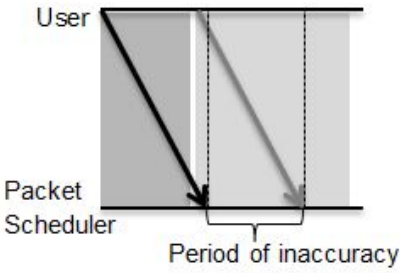

Fig. 4. Illustration of Inaccurate CQI Report

- *Infinite Queue Length at eNodeB* - In LTE, each multimedia application is assigned a different queue per user at the base station or eNodeB. For example, if a user is streaming a video and performing VoIP call in parallel, two queues will be assigned to this user, one for each type of application. The algorithms proposed in this paper, assume the traffic queues at eNodeB to be infinite in length, which is not practical. In reality where queue length is finite, assumption of infinite queue length may lead to buffer overflow if the arrival data rate to the queue is higher than the rate at which the scheduler serves this queue. The buffer overflow may in turn cause two problems; one, the user will make another request to get the data that was discarded; two, if the connection between base station and source that is generating data is reliable, for example TCP connection, re-transmission of data will take place. Nonetheless, buffer overflow will cause delay, degrading the QoS of the multimedia application. Thus, it would be ideal for the scheduler to consider queue length when making scheduling decisions.

- *Buffer Status at UE* - When user or UE receives data from base station, it stores it in UE buffer, which is then removed from the buffer after processing it. Majority of the proposed algorithms do not consider available buffer length at UE, which is limited in size. This means that if the buffer space at UE is full, incoming packets would be dropped and re-transmission between base station and UE would be triggered causing delay in service. This can have negative effect on QoS for certain traffic. If the incoming traffic happened to be Video Traffic or VoIP, the user may experience the delay or degrade in Quaity of Experience (QoE) because VoIP and video streaming traffic are sensitive to delay and packet loss. Hence, the

algorithms should consider Buffer Status Report (BSR) reported by UE to the base station before transmitting data [35].

- *Retransmissions of Packets* - As discussed previously, in LTE, Hybrid Automatic Re-transmission Request (HARQ) component in RRM provides reliable communication over noisy channels. Every packet contains data and error correction codes, also known as Forward Error Correction (FEC) code that can be used to recover subset of errored data. Upon successfully receiving the packet, UE sends Acknowledgment (Ack) to HARQ manager; on the other hand, if the data bits received are errorneous, UE tries to recover it using the FEC code in the packet. If the data bits cannot be successfully recovered, it sends Negative-Acknowledgment (Nack) to the HARQ manager indicating request for retransmission of the erroneous packet. When HARQ manager receives Nack, it has two options for re-transmission: it can duplicate the original packet and re-transmit it; or, it can re-compute FEC code and transmit new packet that consists data in original packet and newly generated FEC code [36]. Regardless of the option selected, re-transmissions require same number of PRBs as original transmission and needs to be prioritized to ensure QoS for the traffic flow is preserved [37]. However, majority of the algorithms discussed in this paper neglect re-transmissions from HARQ. Packet Set Scheduler (PSS) and Two-Level Scheduler (TLS) are the only schedulers discussed in this paper that consider re-transmissions from HARQ. In other words, the scheduler needs to closely co-operate with other components of RRM such as HARQ manager to ensure optimal operation of the system.

System Model Optimization

- *Scheduler Admission Control (SAC)* - In currently proposed solutions, scheduler considers all the traffic queues as input when making scheduling decisions. However, this can negatively affect the decision process of the scheduler. For example, consider 2 flows namely Flow A and Flow B. Assume amount of data in Flow A and its HOL packet delay to be less than certain threshold. On the other hand, assume the Flow B to have data or HOL packet delay above the predefined threshold. If both Flow A and Flow B were to enter the scheduler, scheduler will iterate over all the flows, in this case Flow A and Flow B to make scheduling decision. In the decision making process it is possible that scheduler would assign least priority to Flow A and thus, not schedule Flow A as it does not have enough data to transmit. At the same time, if the scheduler iteratively assigns RBs to the flows and if Flow A's priority is high enough, Flow A will be assigned some RBs. If the amount of data in Flow A happens to be less than the transmission rate supported by the assigned RBs, RBs would be underutilized. On the other hand, if only Flow B was allowed to enter the scheduler, the scheduler would have to iterate over less number of flows to make decision, and in the given example assign more resources to Flow B. Hence, the goal is to implement Scheduler Admission Controller that limits the number of traffic flows that enter the scheduler. This accelerates scheduler's decision process as now it only has to iterate over limited number of flows. The criteria for admission control should be Head of Line (HOL) packet delay, which indicates amount of time packet at head of the queue has been waiting, and queue length or amount of data buffered in the queue at base station. This is similar to Schedulability Check step in Priority Set Scheduler discussed in this paper. Some of the options that can be used for SAC are described below:
  1) *Delay and Queue Length Threshold* - Set a threshold value for queue length and delay based on the traffic type. That is, threshold value for VoIP traffic might be different and lower compared to threshold values of video streaming traffic because each traffic has different delay requirement. In addition, delay threshold value set is less than the target values defined in the QCI table. The different threshold values ensures that QoS of the traffic is not degraded before it enters the scheduler. If the HOL packet delay or queue length surpasses its respective threshold value, then it is eligible to enter the scheduler. This ensures that number of flows entering scheduler every scheduling interval or TTI is limited. This not only helps scheduler to speed up the decision making process, but also ensures that all the flows entering the scheduler have enough data such that RB's capacity is fully utilized when scheduled. However, there is a downside to this approach; if total amount of resources available is more than the total resources required by all the flows in scheduler, residual resources will be wasted. If on the other hand, flows that are ineligible to enter the scheduler as per SAC's criteria were present, the wastage of resources would be minimal. In other words, this approach is only fruitful if total number of available resources is less than or equal to total number of resources required by all the flows in scheduler
  2) *Queue Length to Target Bitrate Ratio* - In case of traffic requiring certain Guaranteed Bit-Rate (GBR), target bit-rate is equivalent to GBR, otherwise it is equal to Aggregated Maximum Bit Rate (AMBR), that is negotiated between user and vendor during subscription [38]. For each traffic flow, compute queue length to target bit rate ratio. The traffic flow is allowed to enter the scheduler if and only if the ratio is $\geq$1. The ratio ensures that only flows with enough data are permitted to the scheduler.

- *Carrier Aggregation* - During data transmission, scheduler selects bit rate for the user that results in highest

amount of throughput and is suitable to user's channel quality. In addition, the bit rate cannot exceed the channel bandwidth. When the system is not heavily loaded and a user requires higher bite rate, scheduler may aggregate multiple channels so that it can serve the user with higher bit rate. This is one of the techniques proposed as a requirement in the next evolution of communication standards called LTE-Advanced [39]. However, this may introduce additional requirements in antenna capabilities at the UE. Antenna design in UE depends on bandwidth, and frequency on which the device operates along with other factors. If for example, the scheduler transmitted data on frequency not supported by the UE, UE will not be able to successfully receive transmitted data. In addition, if the resulting channel bandwidth after aggregation is more than what the user can support, the UE will not be able to receive the data correctly. Hence, it becomes important for the scheduler to take into consideration antenna capabilities of UE when aggregating channel bandwidths.

Simulation Model Optimization

- *Multi-Cell Environment* - In reality, geographic region is sectored into hexagonal area called cells. Each cell encompasses base stations (eNodeB) that serve UEs. As the user moves from one cell to another, communication responsibilities associated with the base station in old cell is handed over to eNodeB that is nearest to the user in new cell. However, during the handover, the user might be in the process of receiving data from the base station. This means that the handover needs to make sure that the scheduler in the new eNodeB appropriately prioritizes resource scheduling to ongoing services of the newly joined user such that QoS is preserved. In other words, the performance of the scheduler should be tested in simulation environment that mimics real LTE network with multi-cell and handover operations. Unfortunately most of the algorithms presented in this paper except for Log Rule, Exp Rule and Two-Level Scheduler, would only test credibility of the proposed scheduler in single cell environment without any handover.

- *Performance in Overload State* - When the system is operating in its full capacity i.e. overload state, multiple users competing for the same resources may have same rank computed by the scheduler. The scheduler needs to break any scheduling ties and ensure fairness among users such that competing users with same priority do not starve for resources. For example, consider an overload state where there are 2 VoIP sessions - A and B, and 2 HTTP traffic flows. In addition, assume the scheduler has enough resources for every scheduling interval to schedule only 1 traffic flow; obviously VoIP traffic will be given priority over HTTP traffic flow which is non real-time traffic and considered as best effort traffic. Furthermore, let's say, 2 VoIP traffic flows have similar parameters such that the scheduling rank or priority computed by the scheduler is same for both VoIP calls. Now the question is, which VoIP call should the scheduler pick? If the scheduler breaks the tie randomly and picks traffic A in one scheduling interval, it needs to make sure VoIP flow B is given priority in the next scheduling period over A. The decision of scheduler in such scenario helps evaluate robustness of the proposed solution. Unfortunately, majority of the algorithms discussed in this paper fail to evaluate the solution in overload state. Thus, to test robustness of the proposed algorithms it is be important to test performance of the algorithm in overload state.

- *Focused Simulation* - As observed, the algorithms discussed in this paper use different metrics, and techniques to prioritize users for scheduling. Therefore, in order to perform comparative analysis between proposed solutions, it is important to test performance of the solution in same simulation environment. The proposed algorithms can utilize open source LTE simulator called LTE-Sim [40] that simulates downlink and uplink scheduling solutions in multi-cell and multi-user environment.

LTE-A Technical Advancements

In order to support ambitious data rate of 1 Gbs in downlink and 500 Mbs in uplink, 3GPP has standardized innovative solutions such as Carrier Aggregation, Coordinated Multi-Point Transmission (CoMP), Multiple Antenna solutions and Co-operation with Heterogeneous Networks (HetNets) [39]. In addition, LTE-A is backward compatible with LTE technology. Effects of each of these technologies on scheduler design are described below.

- *Carrier Aggregation* - As discussed previously, in LTE, transmission of data from base station to UE occurs via communication channel over the air, which has limited capacity. LTE-A allows aggregation of upto 5 adjacent communication channels. This increases the capacity needed to achieve ambitious data rate of 1 Gbps in downlink [39]. From scheduler's perspective, it can serve user faster at higher data rate. However, because LTE-A is backward compatible, complexity of LTE scheduler increases as it will need to determine if the user is using LTE or LTE-A technology and make scheduling decisions accordingly.

- *Coordinated Multipoint Transmission (CoMP)* - In communication network, communication in neighboring cell can propagate to the user in current cell causing interference. LTE-A proposes close co-operation of base stations in neighboring cells to synchronize and coordinate transmissions so that interference can be avoided. In addition, it proposes that data can reside on the user's base station and neighboring cell's base station; the base stations can then coordinate to send data to the

user in distributed fashion [41]. This synchronization and coordination reduces the inter-cell interference, however it also introduces inter-base station synchronization issues such as ensuring the clocks of base stations are synchronized. Moreover, if the approach where data resides at multiple base stations which is then transmitted to the user in distributed fashion, is used, then the scheduler may need to take inter-cell scheduling decisions, which may increase the complexity of the scheduler.

- *Multiple Antennas* - Let's consider a scenario where people are talking to each other face to face. Depending on the noise in the environment, the speaker projects the voice accordingly. That is, if there is noise in the background the speaker would talk loudly so that it is clearly heard by the listener; on the other hand speaker would speak gently if there is no background noise. The spoken voice travels over the air in the form of sound waves, which is captured by listener's ears. Similarly, in wireless communications, data transmission is initiated by the antenna at base station, which then travels over communication channel setup between user and base station over the air, and received by the antenna at user's device. Like the speaker who dynamically adapts to the noise in background by projected voice at lower or higher rate, antenna also dynamically adapts to the noise in the communication channel by transmitting data with different power. In LTE, base station and user each can adopt upto 4 antennas to simultaneously transmit and receive multiple streams of data. This increases throughput efficiency. In LTE-A, base station and user can adopt upto 8 antennas. However, compare to LTE, which serves single user over 4 antennas, in LTE-A, base station can serve multiple users simultaneously using the 8 antennas. In addition, these multiple users can be served using same resources [42]. From scheduler's perspective, this brings in new set of challenges. Effectively, the consideration would be the problem of Downlink Scheduling wherein a base station with multiple antennas serves multiple users. Priority functions need to be determined for allocating each BS antenna to users, assuming that different users can be served by the transmitter antenna at each slot.

- *Heterogeneous Networks (HetNets)* - HetNets consists of base stations and small nodes such as pico cells, femto cells and micro cells. The small nodes can either act as range extenders or as a small scale base station. As a range extender, small nodes enhance coverage of base stations with a goal of improving channel quality between user and base station. It achieves this by forwarding transmissions from base station to the users in its close proximity. As small scale base stations, small nodes co-operate with the base stations and assume similar responsibilities as the base stations. In other words, small nodes are eligible of scheduling resources for the users. This brings challenges such as interference caused due to co-existence of base station and small nodes. Hence, apart from scheduling resources, the scheduler has an added responsibility of mitigating this co-existence interference.

VII. Conclusion

Vastness of literature surrounding scheduling in LTE shows that scheduling and resource allocation is an important aspect of LTE network. However, it becomes a challenge to compare different solutions and find tradeoffs of the proposed algorithms due to lack of proper classification. In this paper we presented a roadmap of design strategies employed by scheduling algorithms proposed in literature, analyzed feasibility of its deployment in real LTE system by questioning robustness of the design principles and assumptions made, and discussed trivial optimizations along with future challenges scheduler design will face in next generation wireless network.

Majority of the algorithms presented in this paper make certain assumptions and neglect factors that can serve as a challenge when deployed it in real LTE system. For example, most of the proposed algorithms assume the buffer length at base station and user to be infinite in size, and assume the Channel Quality Index (CQI) received from the user to be accurate. In addition, majority of the algorithms neglect retransmissions from HARQ component in LTE, which can in turn downgrade QoS of the traffic because of latency incurred during re-transmissions.

In addition, as an optimization, computational complexity of the algorithm can be reduced if the scheduler employed admission control mechanism that allows traffic flows to enter the scheduler if and only if it meets the eligibility criteria. Lastly, next generation wireless network named LTE-Advanced network is backward compatible with LTE and has standardized innovative solutions such as simultaneously supporting multiple users on multiple antennas, co-operating with Heterogeneous Networks, mitigating inter-cell interference, and aggregating transmission channels to increase transmission rate. These new innovative solutions will require next generation scheduler design to make decisions based on capabilities of the user. That is, if the user supports LTE-A, then the scheduler needs to make decision based on new service offerings defined by the standardized innovative solutions. On the other hand, if the user device only supports LTE, then the scheduler will need to make decisions based on the factors highlighted in this survey.

References

[1] 3GPP (1999), *Digital cellular telecommunication system (Phase 2+); Physical layer on the radio path; General Description,* TS 45.001, v11.20.0

[2] Jamalipour, A.; Wada, Tadahiro; Yamazato, T., *A tutorial on multiple access technologies for beyond 3G mobile networks,* Communications Magazine, IEEE , vol.43, no.2, pp.110,117, Feb. 2005

[3] (Cisco), *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20122017,* pp. 11, February 2013

[4] Ekstrom, H.; Furuskar, A.; Karlsson, J.; Meyer, M.; Parkvall, S.; Torsner, J.; Wahlqvist, M., *Technical solutions for the 3G long-term evolution*, Communications Magazine, IEEE , vol.44, no.3, pp.38,45, March 2006

[5] Pokhariyal, A.; Monghal, G.; Pedersen, K.I.; Mogensen, P.E.; Kovacs, I.Z.; Rosa, C.; Kolding, T.E., *Frequency Domain Packet Scheduling Under Fractional Load for the UTRAN LTE Downlink,* Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th , vol., no., pp.699,703, 22-25 April 2007

[6] J.C. Laneri, *Scheduling Algorithms for Super 3G,* Master's Degree Project, KTH, Stockholm, 2006.

[7] Capozzi, F.; Piro, G.; Grieco, L.A.; Boggia, G.; Camarda, P., *Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey,* Communications Surveys & Tutorials, IEEE , vol.15, no.2, pp.678,700, Second Quarter 2013

[8] Kela, P.; Puttonen, J.; Kolehmainen, N.; Ristaniemi, T.; Henttonen, T.; Moisio, Martti, *Dynamic packet scheduling performance in UTRA Long Term Evolution downlink*, Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on , vol., no., pp.308,313, 7-9 May 2008

[9] Monghal, G.; Pedersen, K.I.; Kovacs, I.Z.; Mogensen, P.E., *QoS Oriented Time and Frequency Domain Packet Schedulers for The UTRAN Long Term Evolution*, Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE , vol., no., pp.2532,2536, 11-14 May 2008

[10] Knuth, Donald E., *The Art of Computer Programming volume 2: Seminumerical algorithms,* Addison-Wesley 1997, p. 311

[11] Zaki, Y.; Weerawardane, T.; Gorg, C.; Timm-Giel, A., *Multi-QoS-Aware Fair Scheduling for LTE*, Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd , vol., no., pp.1,5, 15-18 May 2011

[12] Skoutas, D.N.; Rouskas, A.N., *Scheduling with QoS provisioning in Mobile Broadband Wireless Systems*, Wireless Conference (EW), 2010 European , vol., no., pp.422,428, 12-15 April 2010

[13] Ramli, H.A.M.; Basukala, R.; Sandrasegaran, K.; Patachaianand, R., *Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system* Communications (MICC), 2009 IEEE 9th Malaysia

[14] Hazewinkel, Michiel, *Taylor Series,* Encyclopedia of Mathematics, vol.6, 2001

[15] Ramli, H. A M; Sandrasegaran, K.; Basukala, R.; Afrin, T.S., *HARQ aware scheduling algorithm for the downlink LTE system,* Modeling, Simulation and Applied Optimization (ICMSAO), 2011 4th International Conference on , vol., no., pp.1,4, 19-21 April 2011

[16] R. P. Brent; P. Zimmermann, *Computer Arithmetic,* Cambridge Monographs on Computational and Applied Mathematics (No. 18), November 2010, pp.68-69

[17] Jong-Hun Rhee; Holtzman, J.M.; Dong-Ku Kim, *Scheduling of real/non-real time services: adaptive EXP/PF algorithm*, Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual , vol.1, no., pp.462,466 vol.1, 22-25 April 2003

[18] Piro, G.; Grieco, L.A.; Boggia, G.; Fortuna, R.; Camarda, P., *Two-Level Downlink Scheduling for Real-Time Multimedia Services in LTE Networks* Multimedia, IEEE Transactions on , vol.13, no.5, pp.1052,1065, Oct. 2011

[19] Sandrasegaran, K.; Ramli, H.A.M.; Basukala, R., *Delay-Prioritized Scheduling (DPS) for Real Time Traffic in 3GPP LTE System* Wireless Communications and Networking Conference (WCNC), 2010 IEEE , vol., no., pp.1,6, 18-21 April 2010

[20] B. Sadiq, R. Madan, and A. Sampath, *Downlink scheduling for multiclass trafc in lte*, EURASIP J. Wireless Communication Network, vol. 2009, pp. 99, 2009.

[21] Iturralde, M.; Wei, Anne; Ali Yahiya, T.; Beylot, A. -L, *Resource allocation for real time services using cooperative game theory and a virtual token mechanism in LTE networks*, Consumer Communications and Networking Conference (CCNC), 2012 IEEE , vol., no., pp.879,883, 14-17 Jan. 2012

[22] Conitzer, Vincent; Sandholm, Tuomas, *Computing shapley values, manipulating value division schemes, and checking core membership in multi-issue domains,* Proceedings of the 19th national conference on Artifical intelligence, 2004, pp 219-225

[23] Owen, Guillerm, *Multilinear extensions and the banzhaf value,* Naval Research Logistics Quarterly, vol. 22, no.4, pp 741-750, 1975

[24] Brehm, Michael (2012). *Resource allocation algorithms optimized for overload states in the LTE downlink* (doctoral dissertation). Retrieved from ProQuest Dissertations and Thesis.

[25] Puttonen, J.; Henttonen, T.; Kolehmainen, N.; Aschan, K.; Moisio, Martti; Kela, P., *Voice-Over-IP Performance in UTRA Long Term Evolution Downlink*, Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE , vol., no., pp.2502,2506, 11-14 May 2008

[26] Monghal, Guillaume; Laselva, D.; Michaelsen, Per-Henrik; Wigard, J., *Dynamic Packet Scheduling for Traffic Mixes of Best Effort and VoIP Users in E-UTRAN Downlink*, Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st , vol., no., pp.1,5, 16-19 May 2010

[27] Sunggu Choi; Kyungkoo Jun; Yeonseung Shin; Seokhoon Kang; Byoungjo Choi, *MAC Scheduling Scheme for VoIP Traffic Service in 3G LTE*, Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th , vol., no., pp.1441,1445, Sept. 30 2007-Oct. 3 2007

[28] Wirth, Thomas.; Sanchez de la Fuente, Yago; Holfeld, Bernd; Schierl, Thomas Schierl, *Advanced downlink LTE radio resource management for HTTP-streaming*, ACM Multimedia, page 1037-1040. ACM, (2012)

[29] Ramli, H.A.M.; Sandrasegaran, K.; Basukala, R.; Patachaianand, R.; Minjie Xue; Cheng-Chung Lin, *Resource allocation technique for video streaming applications in the LTE system*, Wireless and Optical Communications Conference (WOCC), 2010 19th Annual , vol., no., pp.1,5, 14-15 May 2010

[30] Yan Lin; Guangxin Yue, *Channel-Adapted and Buffer-Aware Packet Scheduling in LTE Wireless Communication System*, Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on, vol., no., pp.1,4, 12-14 Oct. 2008

[31] Basukala, R.; Ramli, H. A M; Sandrasegaran, K.; Chen, L.,, *Impact of CQI feedback rate/delay on scheduling video streaming services in LTE downlink,* Communication Technology (ICCT), 2010 12th IEEE International Conference on , vol., no., pp.1349,1352, 11-14 Nov. 2010

[32] Yang, Y.; Sandrasegaran, K.; Zhu, X.; Fei, J.; Kong, X.; Lin, C.C, *Frequency and time domain packet scheduling based on channel prediction with imperfect CQI in LTE*, International Journal of Wireless and Mobile Networks (IJWMN), vol. 5, no. 4, Aug. 2013

[33] M. R. Souryal and R. L. Pickholtz, *Adaptive modulation with imperfect channel information in OFDM,* in Communications, 2001. ICC 2001. IEEE International Conference on, 2001, pp. 1861-1865 vol.6.

[34] I. C. Wong and B. Evans, *Optimal resource allocation in the OFDMA downlink with imperfect channel knowledge,* Communications, IEEE Transactions on, vol. 57, pp. 232-241, 2009.

[35] 3GPP (2011a)., *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2* (TS 36.300 v10.3.0 ed.). 3GPP.

[36] Kian Chung Beh; Doufexi, A.; Armour, S., *Performance Evaluation of Hybrid ARQ Schemes of 3GPP LTE OFDMA System,* Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on , vol., no., pp.1,5, 3-7 Sept. 2007

[37] Pokhariyal, A.; Pedersen, K.I.; Monghal, G.; Kovacs, I. Z.; Rosa, C.; Kolding, T.E.; Mogensen, P.E., *HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution,* Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th , vol., no., pp.2761,2765, 22-25 April 2007

[38] Brehm, Michael (2012). *Resource allocation algorithms optimized for overload states in the LTE downlink* (doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. AAT 3547639)

[39] Parkvall, S.; Dahlman, E.; Furuskar, A.; Jading, Y.; Olsson, M.; Wanstedt, S.; Zangi, K., *LTE-Advanced - Evolving LTE towards IMT-Advanced,* Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th , vol., no., pp.1,5, 21-24 Sept. 2008

[40] Piro, G.; Grieco, L.A.; Boggia, G.; Capozzi, F.; Camarda, P., *Simulating LTE Cellular Systems: An Open-Source Framework,* Vehicular Technology, IEEE Transactions on , vol.60, no.2, pp.498,513, Feb. 2011

[41] Daewon Lee; Hanbyul Seo; Clerckx, B.; Hardouin, E.; Mazzarese, D.; Nagata, S.; Sayana, K., *Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges,* Communications Magazine, IEEE , vol.50, no.2, pp.148,155, February 2012

[42] Lingjia Liu; Runhua Chen; Geirhofer, S.; Sayana, K.; Zhihua Shi; Yongxing Zhou, *Downlink MIMO in LTE-advanced: SU-MIMO vs. MU-MIMO,* Communications Magazine, IEEE , vol.50, no.2, pp.140,147, February 2012

APPENDIX

QoS Class of Identifier (QCI) classes as illustrated in the table below defines IP level packet characteristics for different traffic types.

Resource Type	QCI	Priority	Packet Delay	Packet Error Loss Rate	Example Service
GBR	1	2	100 ms	10^{-2}	Conversational Voice
	2	4	150 ms	10^{-3}	Conversational Video
	3	3	50 ms	10^{-3}	Real Time Gaming
	4	5	300 ms	10^{-6}	Non-Conversational Video
Non-GBR	5	1	100 ms	10^{-6}	IMS Signalling
	6	6	300 ms	10^{-6}	Video (Buffered Streaming), TCP-based flows (eg, www, e-mail, chat, ftp, p2p file sharing, progressive video etc.)
	7	7	100 ms	10^{-3}	Video, Video (Live Streaming), Interactive Gaming
	8	8	300 ms	10^{-6}	Video (Buffered Streaming), TCP-based flows (eg, www, e-mail, chat, ftp, p2p file sharing, progressive video etc.)
	9	9			

TABLE III: Standardized QoS Class Identifiers for LTE