An Analysis on Quality of Service Enhancement in Long Term Evolution Networks: Past, Present and Future

Nurma Larasati, Wee Kuok Kwee, Siew Chin Chong and Yityin Wee
Faculty of Information Science & Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

Abstract: The standard for wireless mobile communication system has been set higher since the introduction of Long Term Evolution (LTE) system by Third Generation Partnership Project (3GPP). Marketed as part of the 4G wireless technology, LTE offers an outstanding connectivity with high throughput, low latency and improved spectral efficiency. LTE possesses a Quality of Service (QoS) framework to ensure satisfying user-perceived end-to-end performance according to demand. QoS parameters are highly variable according to application, user and service differentiation. Therefore, to satisfy QoS requirements, various studies have proposed algorithms fitting for their specific traffic scenarios. This paper covers a survey of proposed algorithms in QoS framework of LTE system.

Key words: LTE • Scheduling • QoS

INTRODUCTION

The 3GPP started the study to a system evolution of its Universal Mobile Telecommunication System (UMTS) in 2004 aiming to keep its mobile communication system competitive by offering the higher throughput, lower latency and improved spectral efficiency the future multimedia-heavy users will require. Figure 1 depicts the architecture evolution from UMTS into the evolved architecture. UMTS’ core network and Radio Access Network (RAN) has been evolved in the new architecture. The packet-switched domain in UMTS core network has been directly replaced with Evolved Packet Core (EPC). Circuit-switched domain in older architecture is excluded as the voice calls will be transmitted by Voice over IP (VoIP) instead. Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) replaces the older system’s radio access network architecture and handles the radio communication between EPC and User Equipment (UEs). The new architecture is well-known as Long Term Evolution [1].

The whole new architecture is officially known as Evolved Packet System (EPS). It consists of two 3GPP work items; the first is System Architecture Evolution (SAE) project and another is the LTE project. SAE project evolves the core network meanwhile LTE project actually evolves the RAN, air interface and UEs. Nevertheless, LTE has turned into the representative title of the whole system instead of EPS. In a bandwidth of 20 MHz, LTE is expected to support downlink and uplink data rate up to 100 Mbps and 50 Mbps respectively. A set of minimum requirements for Fourth Generation (4G) mobile wireless technology is published by International Telecommunication Union (ITU) under the name IMT-Advanced. The specification states that 4G technologies are the systems with downlink and uplink data rate support of 600 Mbps and 270 Mbps respectively in 40 MHz bandwidth [2]. Unable to meet the 4G requirement, LTE is not supposed to fall under the category. Yet, marketing community kept their pursuit to brand LTE as a 4G telecommunication system with reasoning that there is indeed a distinct improvement of the new system compared to previous 3G system.
An enhanced LTE system is later introduced by the title of LTE-Advanced (LTE-A). It delivers 1000 Mbps peak downlink data rate and 500 Mbps peak uplink data rate. LTE-A is backwards-compatible with LTE. LTE sets performance targets to be fulfilled as following [3].

- Data rates. LTE should support downlink and uplink data rate up to 100 Mbps and 50 Mbps respectively.
- Scalable bandwidth. The bandwidth allocation in LTE ranges from 1.4 MHz to 20 GHz. The higher bandwidth is used to acquire higher data rate.
- Spectral efficiency. Spectral efficiency values of 5 Bps/Hz (downlink) and 2.5 Bps/Hz (uplink) are expected.
- Mobility. LTE is optimized for mobility less than 15 km/h. It is still able to retain high performance until 120 km/h and preserving connection at 300 km/h.
- Coverage. Within 5 km radius from base stations, the performance targets should be achieved. A slight degradation is acceptable within 30 km radius.

This paper will firstly introduce the architecture evolution of LTE in Section II. Then it will be followed by another section to discuss the QoS framework enhancement over period of time. In section IV, future challenges of QoS in LTE will be examined.

**System Architecture:** Figure 2 shows the LTE high-level architecture which consists of UE, E-UTRAN and EPC. The components are connected through interfaces namely UU, S1 and SGi. EPC then communicates with packet data networks (PDN), for examples, to Internet, IP multimedia system or private corporate networks.

The internal architecture of E-UTRAN and EPC is shown in Figure 3. E-UTRAN consists of several base stations named as evolved Node B (eNB or eNodeB). The eNB sends downlink transmission to the UEs and receives uplink transmission from the UEs. A single UE strictly only connects to a single eNB in a cell at a time. In cases of handover, which is a situation where UE changes its connection from one eNB to another, eNB sends signaling commands that relate to the radio transmission. X2 interface connects one eNB to another and it functions as optional interface for signaling and packet forwarding during handover. S1 interface, which connects the eNB and the EPC, is able to do the same function as X2 interface albeit slower and indirectly.
The EPC consists of Serving Gateway (S-GW), PDN Gateway (P-GW), Mobility Management Entity (MME) and Home Subscriber Server (HSS). The traffic data from E-UTRAN is forwarded to S-GW which thus route and forward the packet between the nodes and P-GW. P-GW connects the LTE network with the outside world such as Internet. Signaling messages from E-UTRAN is delivered to MME instead. MME manages the high level operation of the UEs such as intra-LTE handover matters via signaling messages. MME communicates with HSS, the central database that stores all of the network operator’s subscriber, to do its tasks.

Radio Transmission and Reception: The radio transmission and reception in LTE uses a technique called as Orthogonal Frequency Division Multiple Access (OFDMA). OFDMA not only allows the eNB to communicate with several UEs at the same time but also minimize the problems of fading and inter-symbol interference. A modified radio transmission technique called as Single Carrier Frequency Division Multiple Access (SC-FDMA) is used for LTE uplink.

OFDMA is a multiple access scheme based on digital modulation known as Orthogonal Frequency Division Multiplexing (OFDM). In OFDM transmitter, the information bits are split into parallel sub-streams which then disperse the bits into several different frequencies known as sub-carriers. OFDMA shares resources to all UEs dynamically as improvement from the OFDM. Besides, OFDMA also utilizes two extra processes, namely cyclic prefix insertion and channel estimation and equalization. Cyclic prefix insertion is useful in getting rid of inter-symbol interference problem entirely. Meanwhile, channel estimation and equalization is used to detect and remove any phase shift issue from the transmitted data so that the correct information bits can be recovered.
One LTE frame consists of one equal sub-frames with Transmission Time Interval (TTI) of 1 ms each.

Each sub-frame can be divided into two time slots. RB, as the the smallest radio resource unit assignable to a UE for data transmission, refers to a unit of one time slot (0.5 ms) in one sub-channel as seen in Figure 4. On the other hand, two consecutive time slots (1 ms) in a sub-carrier is called as Scheduling Block (SB).

Radio Bearer: At a high level, data delivery in an LTE system is done through radio bearer, specifically the EPS bearer. It is a bi-directional pipe between UE and P-GW which uniquely classifies packet flows into common QoS treatments. When a packet start to flow from UE to EPC, it will face a tunnel header which classify the packet into specific bearer based on the packet’s QoS requirement specified by the UE’s service and subscriber differentiation. QoS itself defines how the data in transmission should be treated, utilizing parameters such as scheduling policy, rate shaping policy, queue management policy, etc.

Two types of EPS bearers are Guaranteed Bit-Rate (GBR) bearer and Non-Guaranteed Bit-Rate (Non-GBR) bearer. A GBR bearer sets a definite minimum bit-rate values to the traffic. Thus, packet flows through GBR bearer is shielded from congestion-caused packet loss, causing it suitable for real-time services. Non-GBR bearer offers no MBR guarantee. EPS bearer is also classified into default bearer and dedicated bearer. Default bearer is an automatically set-up non-GBR bearer once a UE is registered in an EPC. This bearer provides always-on connectivity to the outside PDN such as Internet. Instead, dedicated bearer can be any of the bearer types. Typically it provides better QoS than the default bearer.

QoS Parameters and Protocol Stack: The rapid growth of mobile broadband subscribers cause the cellular operators to offer multi-services (e.g. mobile TV, VoIP) and subscriber differentiation (e.g. post vs pre-paid subscription, corporate vs private subscription). QoS mechanism is used to enable network operators configuring service and subscriber differentiation to control the packet traffic performance of a specific subscriber and service group [4]. In general, QoS is a term to describe the user-perceived “collective effect of service”.

QoS requirements are varied across applications, services and users; yet the common requirements discussed are the throughput, delay, packet error rate and jitter-associated aspects of a given application. Application-based QoS deals with putting the right QoS requirements for the specific mobile application, such as data, video and multimedia. On the other hand, policy-based QoS is to control service and subscriber differentiation [5].

To note, the perceived quality seen from user perspective is based on end-to-end performance. Therefore, effective delivery of QoS has to be delivered across the network. Network architecture, higher-layer protocols and network elements interaction between the UEs and eNBs are the three components to guarantee the end-to-end QoS. Both control plane and data plane mechanisms are essential in QoS delivery [5]. The control plane allows users and network negotiating and agreeing on QoS specifications while the data plane enforces the agreed QoS specifications by controlling the network resource amounts that every application/user can consume.

<table>
<thead>
<tr>
<th>QCI</th>
<th>Resource Type</th>
<th>Packet Delay Budget (ms)</th>
<th>Packet Loss Rate</th>
<th>Priority</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GBR</td>
<td>100</td>
<td>10^{-2}</td>
<td>2</td>
<td>Voice call</td>
</tr>
<tr>
<td>2</td>
<td>GBR</td>
<td>150</td>
<td>10^{-4}</td>
<td>4</td>
<td>Video live streaming</td>
</tr>
<tr>
<td>3</td>
<td>GBR</td>
<td>300</td>
<td>10^{-4}</td>
<td>5</td>
<td>Buffered streaming</td>
</tr>
<tr>
<td>4</td>
<td>GBR</td>
<td>50</td>
<td>10^{-3}</td>
<td>3</td>
<td>Real-time gaming</td>
</tr>
<tr>
<td>5</td>
<td>Non-GBR</td>
<td>100</td>
<td>10^{-6}</td>
<td>1</td>
<td>IMS signaling</td>
</tr>
<tr>
<td>6</td>
<td>Non-GBR</td>
<td>100</td>
<td>10^{-2}</td>
<td>7</td>
<td>Voice call, video streaming, interactive gaming</td>
</tr>
<tr>
<td>7</td>
<td>Non-GBR</td>
<td>300</td>
<td>10^{-6}</td>
<td>6</td>
<td>Buffered streaming</td>
</tr>
<tr>
<td>8</td>
<td>Non-GBR</td>
<td>300</td>
<td>10^{-6}</td>
<td>8</td>
<td>WWW, e-mail, FTP</td>
</tr>
<tr>
<td>9</td>
<td>Non-GBR</td>
<td>300</td>
<td>10^{-6}</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>
Each bearer is associated with a set of QoS parameters to allow flow differentiation and thus identified with a QoS Class Identifier (QCI) each. QCI is a scalar used within the network to refer the parameters controlling the packet-forwarding treatment. A standardized QCI classification is shown in TABLE I. Radio Resource Management (RRM) translates those identifiers into admission policies, scheduling parameters, link layer protocol configuration, queue management threshold, etc. The goal of standardizing these QoS classes is to ensure the services and applications mapped in a QCI should receive common minimum QoS level in the case of roaming and deployments in multi-vendor network.

**QoS Framework in LTE:** QoS, in general, is a term to describe the user-perceived “collective effect of service”. QoS metrics cover the performance of throughput, spectral efficiency, delay, packet loss rate, energy cost, etc. The QoS requirements varies over applications, users and services. RRM is used to translate the QoS parameters into radio resource-related decision. Several strategic requirements are mentioned in [7-15] to achieve the LTE QoS requirement. First, network operator should have the ability to control the policy for service and user differentiation at ease. Second is to have UE has minimum involvement in QoS and policy control as possible. Besides UE being prone to unauthorized access, this strategy guarantees consistent exception handling among multi-vendor UEs. Third, QoS for Access Agnostic Client Application for both UE and non-UE should be supported. Fourth strategy is to have a fast session setup with delay as minimum as possible. Lastly, LTE-based services should be backwards compatible. A comprehensive study of existing proposed algorithms of QoS framework in LTE is to be discussed in this chapter.

RRM is the procedure which translate QoS parameters into efficient radio resource-related decision, such as call admission policies and packet scheduling algorithm. Adoption of advanced RRM is essential in order to achieve performance level up to the Shannon limit—a theoretical maximum transfer rate of a communication channel. The following discusses various algorithms and protocols proposed to solve their specific RRM policies (i.e. Call Admission Control policy). The algorithms have the same general objective which is to satisfy UE’s QoS requirements.

**Call Admission Control (CAC):** CAC aims to prevent service degradation and network congestion of already accepted users and therefore important to QoS provisioning [7]. CAC restricts the amount of ongoing connections, known as calls and either accept or reject each of them based on the resource availability to meet the new calls QoS requirement while maintaining the QoS of already accepted calls [8]. CAC algorithm requires good signal strength and resource availability in the selected eNB to be satisfied in order to accept new call to the network [9].

CAC regulates the traffic volume of real-time services, particularly in wireless mobile network and VoIP, by considering the total calls, total utilized bandwidth, or the total passing bits or packets over time. Once the defined limit is reached, channel degradation or call blocking will be done. Thus, to avoid this channel degradation, call requests are identified into either New Call (NC) or Handoff Call (HC) and the service types are split into either VoIP or video. The VoIP services are allocated according to traffic density whereas non-real time and non-VoIP services are according to marginal utility function-based channel condition [8].

CAC consists of static CAC and dynamic CAC [10]. The earlier reserves several resources for HCs while the latter does admission control based on estimation of radio channel status and available resources. Static CAC algorithms include guard channel algorithm, queueing principle algorithm and fractional guard channel algorithm. Guard channel algorithm reserves several channel for handoff calls through simple and easy implementation [11]. However, it is hard to determine the amount of channel need to be reserved due to the changes in user movement. Fractional guard channel algorithm probabilistically agrees to either accept or reject NCs based on the channel quality in order to alleviate congestion [12]. Both channel reservation algorithms are still inefficiently allocate resources as they can reject new calls when the resources are available and may excessively blocking new calls over handoff calls [13]. Queueing principle algorithm is proposed to overcome the issue. It accepts call requests when the resources are available. If otherwise, the calls are put into queue until some idle channels appears after handoff or call release [14].

Dynamic CAC involves local predictive, distributed and shadow cluster algorithms. Local predictive algorithm estimates resources amount to be reserved based on Wiener process, estimates HCs arrival times and do resource reservation for HCs [15]. It has lower dropping probability of HCs yet possible less-accurate estimation. Estimation accuracy is offered by distributed CAC procedure which calculates both adjacent cells and local resources [16]. However, this algorithm assumes every
call has same service type and QoS requirements. To solve the issue, a shadow cluster CAC algorithm is made by considering the user mobility, i.e. movement direction, velocity, current position [17]. The adjacent cells of base station which user resides are marked as shadow cluster that user has the probability to move to. A portion of shadow cluster’s resource is reserved for handoff calls. This algorithm provides higher resource utilization but may cause overhead as all base stations are expected to possess information about all user devices’ mobility.

Paper [10] proposed a delay aware CAC to ensure delay-related QoS requirements in LTE network. It accepts or rejects NCs according to every service type’s packet delay and current RB utilization of the eNB. It causes lower RB utilization and data rate than non-CAC, but effectively bounds average packet delay for each service type.

In a heterogeneous network, which refers to a condition in which two or more wireless communication networks (e.g. Cellular LTE and WLAN/WiFi networks) coexist, a mobile device can only access to one network at a time; hence the importance of the admission strategy. A simple admission control unconditionally prefers WLAN due to cheaper cost and higher bandwidth than other early 3G networks [18]. Yet, the WLAN-first scheme may cause the WLAN getting overcrowded and future attempts to access WLAN can result to failure.

On the other hand, [19] believes that an upward VHO (upward vertical handover/connection transfer to WLAN from cellular network) will be only for a short period of time and it will soon after return back to cellular network (through downward VHO). It proposes an algorithm to probabilistically rejecting VHOs attempt of highly mobile users in order to maintain reasonable WLAN throughput.

Observing the results from [18,19] as well as incorporating the ring-based cell model from [20], an efficient CAC by adaptively altering the preference settings in accordance to the traffic changes of heterogeneous network [21]. A fixed probability of user preferring WLAN regardless of the user location and the network current load is deemed inappropriate. The paper incorporate the usage of ring-based cell model in which the user located further from BS get higher resource allocation than those closer to the BS. The proposed traffic management policy calculates the WLAN migration probability according to traffic load and the user position in the ring model. The WLAN migration preference values include functions of handoff and blocking failure probabilities as well as NC/HC attempt failure probabilities. In general, this policy decides to migrate users far from the eNB to access WLAN, causing more resources released to cellular network.

However, [21] did not consider the user mobility difference in double coverage and single coverage network. WLAN is generally set up for indoor environment and it is unlikely to have the same user mobility in double coverage. An improvement is added in [22] by adding differences in mobility as a parameter in the ring-based cell. Its analysis shows that in a 3 ring based model, it is wise to increase preference for inner ring, but not innermost, than the outer one as the remaining resources in WLAN increases further.

**Downlink Scheduling Algorithms in LTE:** Besides CAC, one of the most significant RRM functions is the packet scheduling. Packet scheduler allocates SBs to selected active UEs and transmit their packets in a way to utilize the most out of the radio resource and satisfy the QoS requirements [23, 24]. Most schedulers work in two phases: Time Division Packet Scheduler (TDPS) and Frequency Division Packet Scheduler (FDPS). In terms of transmission direction, packet scheduling is divided into uplink (packet scheduling from UE to eNB) and downlink (packet scheduling from eNB to UE) scheduling. Scheduling decision occurs in the eNB.

In order to achieve a good scheduling decisions, several aspects related to the channel are to be evaluated [6].

- Status of transmission queues to drive decisions minimizing delays.
- Resource allocation history as benchmark to improve fairness of future transmission.
- Buffer status which exploits receiver-side buffer condition to avoid buffer overflows.
- Channel quality is acquired through Channel Quality Identifier (CQI) reports. CQI reports estimate the downlink channel quality based on Signal-to-Interference plus Noise Ratio (SINR).
- QoS requirements is met by deciding policies according to the flow’s QCI values.

**Scheduling Strategies Classification:** All LTE downlink schedulers which have been proposed play with different approaches of trading off a certain computational complexity with decision optimality. The factors commonly taken into account in defining LTE allocation policy are the network complexity/scalability, spectral efficiency, fairness and QoS provisioning. According to paper [6], the LTE downlink scheduling strategies can be classified into five groups: channel-unaware; channel-aware/QoS-unaware; channel-aware/QoS-aware; semi-persistent for VoIP support; and energy-aware scheduling strategies.
Channel-Unaware Scheduling Strategies. The strategies are based on assumption that time is of no importance and transmission media is error-free. Direct implementation is unrealistic and thus it is mostly integrated with other channel-aware approaches. First-In First-Out (FIFO), Maximum Rate (MR), Round Robin (RR), Blind Equal Throughput (BET), weighted fair queueing and guaranteed delay are among the strategies fall under this group.

Channel-Aware/QoS-Unaware Scheduling Strategies. CQI feedbacks help the scheduler estimating the channel quality experienced by each UE. Maximum Throughput (MT) in the current TTI is one of the strategies achievable through the channel awareness approach. However, it causes unfair resource sharing as users in poor channel quality such as cell-edge users will get lesser percentage of available resources, or worse, they will suffer in starvation. The schemes fall into this strategy are Proportional Fair (PF) scheduling, Buffer-aware schedulers and Joint Time and Frequency domain schedulers.

Channel-Aware/QoS-Aware Scheduling Strategies. Schedulers in this category are schedulers for guaranteed data rate, schedulers for guaranteed delay requirement and dynamic schedulers for VoIP support. Delay sensitive flows such as in VoIP and video conferencing need not only sufficient network capacity but an effective QoS scheduler as well. In certain situation, it is essential to schedule users with expanding delays/queues and whose current channel not favorable. A queue and channel-aware QoS scheduler for LTE downlink transmission, especially for delay sensitive flows. The schemes fall under this strategy are Schedulers for Guaranteed Data-Rate, Schedulers for Guaranteed Delay Requirements and Dynamic Schedulers for VoIP support.

Semi-persistent for VoIP support Scheduling Strategies. Semi-persistent scheduling has the aim to increase the capacity of VoIP network. It is not designed specifically for improving spectral efficiency nor reducing packet delay.

Energy-Aware Scheduling Strategies. Energy saving can be done in both eNB and UE. Energy aware schedulers are beneficial to have the system able to run for long period of time with the least energy consumption.

Conventional Packet Scheduling: In overview, a downlink packet scheduler works as shown in Figure 5. UE send CQI report to the serving eNB for every 1 TTI. Every UE is assigned with a buffer in eNB in which every packet inside is time-stamped and queued for delivery in First-In First-Out (FIFO) basis. On every TTI, scheduling decision is performed by packet scheduler and one or more RBs can be allocated for every UE. After a UE with highest metric is selected to be transmitted, the number of bits allocated per RB depends on assigned Modulation and Coding Scheme (MCS) which is selected on the basis of reported SINR value [23].

Several conventional downlink packet scheduling algorithms are as listed in the following. Since the birth of LTE, various new or modified algorithms have been proposed to improve the QoS of LTE even better.

a) Maximum Carrier-to-Interference Ratio (Max C/I) Scheduling

Max C/I scheduling allocates resources to users experiencing the best instantaneously radio links condition, as estimated from carrier-to-interference ratio. But this algorithm may cause users with bad channel condition (i.e. cell edge users) starve from resources [25].

b) Round Robin (RR) Scheduling

RR scheduling emphasizes on user fairness in which every UE take turns in being assigned resources. Being fair as it is, the most resource-hungry UE (i.e. UE with bad channel condition) will face drawback.

c) PF Scheduling

PF scheduling algorithm was proposed to provide a trade-off between the fairness and system throughput. This algorithm allocates RBs by considering the UE’s past average throughput and channel quality experience. The scheme examines throughput history and assumes a weighing factor of expected data rate to ensure any futures users with bad channel quality can still be served.
within a period of time. In other words, PF algorithm favours UEs with low average throughputs and UEs with good channel quality. PF scheduling does not consider QoS requirement, making it unsuitable for real-time applications [23].

According to the PF scheduling policy, the \( i \)th UE is prioritized for transmission in time slot \( k \) based on the following equation.

\[
Q(k) = \arg \max_{1 \leq j \leq N} \frac{r_j(t)}{R_i(t)}
\]  

(1)

and:

\[
R_i(t) = \left( 1 - \frac{1}{t} \right) R_i(t-1) + \frac{1}{t} \cdot c_i(t-1)
\]  

(2)

where, \( \theta \) is the instantaneous achievable data rate, \( \bar{\theta} \) is the average achievable data rate of \( i \)th at time \( t \) and \( \epsilon \) is a constant value greater than 1 representing the update window size. When \( \epsilon \) is considerably large, the UE with strongest CQI is scheduled for transmission. Yet, if the \( \epsilon \) is closer to approaching value 1 (\( \epsilon \)), PF algorithm disregards the CQI and instead perform Round Robin (RR) scheduling to the packet. Therefore, \( \epsilon \) can be manipulated to shift the tradeoff between shorter packet delay and cell throughput [26].

d) Maximum-Largest Weighted Delay First (M-LWDF) Scheduling

M-LWDF algorithm schedules multiple real-time data users with varying QoS requirements by considering the states of queue and CQI. The algorithm prioritize UE by the following equation.

\[
Q(k) = \arg \max_{1 \leq j \leq N} D_j(k) \cdot \frac{r_j(t)}{R_i(t)}
\]  

(3)

where, \( \epsilon \) is the Head of Line (HOL) packet delay, \( \bar{\epsilon} \) is the maximum achievable data rate at transmission interval \( k \) and \( \bar{\epsilon} \) is the QoS parameter of \( i \)th UE. It should be noted that HOL packet delay refers to the time difference between the packet’s arrival time and the current time. Because the M-LWDF considers HOL packet delay along with PF properties (PF priority metric in Equation 1 is incorporated in M-LWDF priority metric as seen in Equation 3), the algorithm has a good throughput and fairness performance as well as relatively low Packet Loss Ratio (PLR) [27]. M-LWDF outperforms PF and EXP/PF the most in terms of PLR, average throughput, packet delay, spectral efficiency and fairness [23].

**EXP/PF Scheduling:** EXP/PF algorithm is a composite of two schedulers, namely the EXP rule and PF rule. EXP rule guarantees delay bound of real time services while PF rule maximizes the system throughput while maintaining proportional fairness between UEs. This algorithm prioritize real time services over non-real time services if the HOL delays are closing to the delay deadline.

LTE-A has a promising feature known as Carrier Aggregation (CA) to support wider bandwidth in packet transmission [28–30]. Classic PF scheduling in LTE-A can be integrated with CA to create an efficient packet scheduling with timely delivery of real time traffic and delay-free non-real time traffic [29].

**Enhanced Algorithms for Downlink Scheduler:** Various scheduling algorithms has been proposed enhancing the conventional and basic scheduling schemes. In general, the proposed schemes offer high throughput with user fairness balance by modifying existing algorithm, merging multiple utility function at once, etc. It is common for the proposed algorithms to aim in solving a certain network scenario’s issues, such as video streaming playback, hand-off service degradation and RB optimization issues. The following is a comprehensive study of latest proposed scheduling algorithms for better downlink QoS satisfaction in LTE network for Single-In Single-Out (SISO).

Multiple conventional scheduling algorithms previously mentioned can be integrated together to counteract one’s weakness with another’s advantages. A scheduling algorithm that operates in between Best CQI and RR scheduling is proposed in [31] while a combination of M-LWDF, EXP/PF and virtual token mechanism is proposed in [32]. In [31], Best CQI rule contributes to throughput optimization whereas RR rule contributes to user fairness. The first time slot of granular RB level is allocated for UE with highest CQI, while the second time slots is allocated for other UEs in RR manner, cyclically in turn.

Paper [32] proposed a modified algorithm out of M-LWDF and EXP/PF rules for real time flows. In real time flows, queue length is tightly related to QoS provisioning. So virtual token mechanism is adopted to consider the queue length, benefiting the Virtual Token M-LWDF (VT-M-LWDF) algorithm in acquiring high throughput with low PLR. VT-M-LWDF [32] and M-LWDF is integrated to propose an algorithm with more balanced performance metrics of inter-class flows (i.e. average throughput, PLR, user fairness, spectral efficiency) [33].
M-LWDF is also modified into Adaptive Modified Largest Weighted Delay First (AMLWDF) scheduling algorithm in [34]. Multiple extra criteria is added into its scheduling formula: delay time, average date rate, instantaneous data rate, QCI priority, delay constraints and fairness.

The paper in [35] proposed an enhancement to conventional QoS-unaware RR and Best CQI scheduling in order to enable them supporting more resources for real-time flows during congestion. QoS-aware RR (QRR) scheduling is designed to prioritize real-time flow without any QCI consideration whereas Queue-aware Best CQI (QBC) scheduling is designed perceive all real-time flows as if they have the best channel quality. QRR and QBC scheduling provides higher throughput for real-time traffic flows with the cost of slight degradation in either throughput or fairness.

Another integrated scheduling schemes proposed is Modified Earliest Deadline First-Proportional Fair (M-EDF-PF) for real-time services [36]. Earliest Deadline First (EDF) scheme schedules packet with the closest deadline expiration. Yet, it is channel-unaware and unsuitable for cellular network. On the other hand, PF scheduler is a QoS-unaware algorithm guaranteeing balanced throughput and fairness. M-EDF-PF combines the bounded delay guaranteed characteristic of EDF and fairness characteristics of PF. It improves PLR, throughput, fairness index, packet delay and spectral efficiency.

Meanwhile, Generalized Largest Weighted Delay First (GLWDF) scheme from [37] provides packet-level service differentiation, unlike other general schedulers (i.e. PF) which uses flow-level service differentiation. This packet-level service differentiation is achievable through Error Propagation Length (EPL) which configures priorities among real-time video packets of a single video flow. GLWDF scheduler is reliable when the number of UEs grows. The prioritization scheme lets important video packets to face shorter HOL delay and causes more of the packets surviving real-time delay bound.

Two level scheduling is also a way to utilize two or more scheduling algorithms after another, commonly the upper level is for UE selection purposes while the lower level is for RB allocation. In [38], Frame Level Scheduler (FLS) is used as the upper level on top of PF scheduling. FLS decides frame by frame the data amount to transmit from real time source to satisfy its delay bound. A discrete-time linear control loop is exploited by FLS to come up with a low computational complexity solution to FLS design. The data is transmitted to lower level scheduler to allocate RBs based on FLS bandwidth requirement, using PF algorithm. Paper [39] proposes a lesser delay by average of 2.46% than [38] by changing the lower level scheduler with Modified EXP (M-EXP).

Time Division Packet Scheduler (TDPS) is proposed as the upper level scheduler above Frequency Division Packet Scheduler (FDPS) in [40]. TDPS is a PF-based algorithm to select which UE to be served first. Then, FDPS receive the packet flow of selected UE and assigns RB with highest SINR values for it. The two-level downlink algorithm aims to optimize resource allocation during overbooking scenarios, the moment when number of UEs are far more than the resource capacity. The proposed scheduler in [41], namely Adaptive-Efficient Packet Scheduling Algorithm (A-EPSA), firstly identify traffic into GBR and non-GBR traffic before flowing the separate traffic into the two-level downlink scheduler described in [40]. A-EPSA improve QoS for both traffic while maintaining user fairness, good throughput, low delay and low PLR.

In OFDMA network, diversity schemes (i.e. frequency diversity, multi-user diversity/frequency selectivity) is a technique to combat fading and interference. Frequency Diversity Scheduling (FDS) uniformly distributes subcarriers under the same user over entire frequency band, to avoid frequency-selective fading experienced by high-mobility users. On the other hand, Frequency Selectivity Scheduling (FSS) selects UE with the best CQI at a time for scheduling, suitable for low-mobility users. In real scenario, low-mobility and high-mobility coexist in a system thus neither FSS nor FDS is adequate enough. Thus, paper [42] proposed a PF-based Frequency Diversity and Selectivity Scheduling (FDSS) to gain low-mobile users advantage while fulfilling frequency diversity required by high-mobility users.

In [42], classification of high-mobility and low-mobility is through estimating whether a user within or outside certain speed limit set. Paper [43], instead, offers a mobility classification algorithm insensitive to channel’s delay profile.

Actual traffic is comprised of Real Time (RT) traffic, i.e. UDP and Non-Real Time (Non-RT) traffic, i.e. TCP. Paper [44] introduces Strict Prioritization scheme as well as Reconfigurable Traffic Prioritization scheme. The objective of both schemes is to ensure both traffic are served according to the QoS needs. Strict Prioritization scheme strictly prioritize RT traffic over Non-RT. While RT services enjoy high throughput and low PLR, the
scheme causes significant drawbacks for Non-RT services. Reconfigurable Traffic Prioritization adds fairness to Non-RT services. Network administrators are given ability to control the trade-off between high performance of real-time services and degradation of non-real time services.

Heterogeneous traffic is separately served in the proposed resource allocation algorithm in [45]. Delay-sensitive RT services adopt minimum and maximum tolerable delay time. The preferences might lead to scheduling users with growing delay yet least favourable channel condition. Meanwhile, the Non-RT services are ensured of its user fairness.

To satisfy QoS for video streaming, paper [46-48] introduces their approaches in scheduling scheme. Paper [46] improves video streaming QoS by deciding amount of enhancement layer information, based on layered coding scheme, to be transmitted under different network load scenarios. On the other hand, video playout buffer level is the focus of paper [47, 48] to improve Quality of Experience (QoE) score at the receiver’s player. Paper [47] uses the playout buffer level information to reduce amount of resources allocated for user with enough buffered data. It results in lesser amount of pauses during video playback. Playout Buffer Aware (PBA) scheduler introduced in [48] identifies the urgency of the received flow, analyses its UE’s playout status and then define the appropriate minimum rate requirement for the UE to guarantee playout continuity.

Packet Prediction Mechanism (PPM) is another proposed scheduling algorithm for RT services [49]. It consists of three phases. The first phase is in frequency domain to select UE with the best CQIs in order to achieve good throughput. The second phase is in the time domain. It estimates packet delays, loss rates and remaining lifetime to ensure time requirements are met. Lastly, the prediction results help PPM in rearranging the transmission order, discarding packets with unsatisfied delay requirement and more accurately predicting future incoming packets.

QoE refers to measurement of performance satisfaction from user perspective. It is loosely termed, or sometimes interchangeably, with QoS. In favour of better QoE, an extension of scheduling algorithm is introduced in [50] called as Packet Intensity (PI). PI is made as a replacement of Peak Signal-to-Noise Ratio (PSNR) metric to evaluate a service quality experienced by user. PI quantifies discontinuity of the service experienced and can be extended into any chosen scheduling scheme. As a result, the extended algorithm will favour resource allocation with high PI value in order to reduce the service discontinuity.

QoE-based downlink scheduler is proposed in [51] for Voice over IP (VoIP) services. The scheduler prioritize UEs based on the QoE score and thus allocate the resources. The priorities are estimated using parameters of Mean Opinion Score (MOS), average throughput, channel state and GBR/non-GBR traffic. MOS values represents the overall effect of packet loss and delay important in defining the state of user perception.

In relation to scheduling, users experiencing bad channel condition (i.e. cell edge users) tend to “starve” from resources. Paper [52] proposes MY_SCH_Not_Fair and MY_SCH_Fair scheduling algorithms to prevent resource starving from happening. MY_SCH_Not_Fair scheduling finds the maximum value from channel feedback information matrix and schedules a user in the RB whose CQI is the highest. Until the end of TTI, that user is not permitted to be scheduled. Yet, this algorithm does not consider previous history of which user has been allocated. In order to give better fairness, MY_SCH_Fair scheduling improves the previous algorithm by taking into accounts whether a user has been previously allocated with resources or not. Previously allocated UE will not be scheduled in the current TTI to give fair chance to other UEs. MY_SCH_Not_Fair [52] is modified by paper [53] due to its wastage of RBs whenever the number of UEs is smaller than amount of available RBs. It proposes a scheduling which allows every active UE assigned multiple RBs per TTI while available, instead of restricting one RB per single UE only.

QoS-based Fairness Aware Downlink Scheduling (QFS) offers a scheduling scheme to solve starvation and fairness issue based on CQI, user request amount, dynamic priority coordination and user fairness [54]. It identifies packets into three queuing types: urgent queue, GBR queue and Non-GBR queue. Respectively, urgent queue has higher priority over the later.

Besides resource starvation, Hard Hand-off (HO) used by LTE causes service degradation, especially to delay-sensitive services such as video streaming. In LTE, Hard HO procedure should break UE’s current cell connection before connecting to another during handover. General scheduling algorithms do not consider the service degradation caused by Hard HO, which is what paper [55] tries to otherwise propose. The proposed scheme divides the scheduler into two QoS-driven
operational control modules, namely the Transmission Deadline Control (TDC) module and the Hand-Off Control (HOC) module. It offers QoS guarantee to multimedia users and fairness to regular users even during hard HO situation.

Paper [56] utilizes the technique in [55]. LTE flows are categorized into VoIP, Best Effort (BE) and Background (BK) to be treated by the algorithm’s specific QoS-driven control modules. TDC and HOC modules compute multimedia traffic amount and inform the quota to another module called as Resource Block Allocation (RBA) module. RBA module checks the quota with available capacity. If quota does not exceed capacity, resource allocation is granted. Otherwise, RBA module allocates as many as possible under channel capacity constraint.

Utility function is a conception to measure user satisfaction of a wireless communication. Utility-Based Scheduling (UBS) algorithm in [57] tends to reduce algorithm complexity to some extent. It performs well in throughput, fairness and PLR but with the cost putting a part of data packets into longer delay. The algorithm uses three types of utility function. Firstly, constant data rate voice flow is treated in a way user is ensured end-to-end strict performance. Second, the streaming flows are allowed to have adaptive dynamic rate between specific thresholds. Lastly, the email flows will get access to larger value of utility as the data rate increases.

Paper [58] proposes Delay-Constrained Proportional Fairness (DCPF) scheduling. DCPF uses a utility function that considers difference between packet HOL delay and average delay, aiming to reduce cross-layer information dependence.

Network operators could also put their subscriber’s device characteristics as a criterion for better multimedia quality experience. Thus, paper [59] introduces Utility-based Priority Scheduling (UPS) algorithm. UPS algorithm decides flow prioritization based on device classification (i.e. display resolution), device’s energy consumption rate and the multimedia stream QoS.

Common scheduling strategies are generally either downlink-specific or uplink-specific. Nevertheless, the proposed algorithm in [60] tries to ensure utility functions that captures both QoS delay and channel quality for both uplink and downlink flows. The joint scheduling is opportunistic, exploits multi-user diversity and comes with trade-off cost between complexity and signaling overhead. There are three main stages involved: prioritization based on priority function of joint uplink/downlink, debt sharing stage between both links and power allocation stage.

Besides, general scheduling algorithms are only based on one layer perspective (i.e. MAC layer only). Three layers are involved into a single Cross-Layer Resource Allocator (CLRA) design: application layer, MAC layer and physical layer [61]. CLRA receives video sequence distortion information from application layer and UEs feedbacks from physical layer. The received metrics is processed for allocation decision.

Paper [62] introduces Optimized Resource Block Allocation and Scheduling for Real Time Services (ORAS). ORAS allocates reserved Rbs for RT services. The incoming RT traffic is accepted into RB allocation algorithm which considers both instantaneous data rate and average data rate. Moreover, if the resources are busy, the RT traffic is directed into lower level scheduling algorithm (i.e. PF) instead for a specific period of time.

In terms of optimized RB allocation, to find the best RB scheduling for multiple UEs can be identified as a combinatorial optimization problem. There are so many feasible scheduling among the exponentially-growing search space which causes exhaustive search to identify the best one. Metaheuristic approach and Operations Research in mathematics are exploited in [63, 64] to efficiently acquire optimized scheduling.

A Genetic Algorithm (GA) based metaheuristic approach is used in [63] to iteratively generate better scheduling decision. GA algorithm mimics the survival-of-the-fittest principle whereby an initial randomized scheduling passes through phases of evaluation, selection, crossover, mutation and termination. It returns high PSNR, low PLR and low latency performance while satisfying user fairness.

On the other hand, Assignment Model from Operations Research in mathematics is involved to optimize trade-off balance between throughput and fairness in [64]. Firstly, the algorithm maps the channel feedback information of different UEs to every RB of current TTI into CQI-feedback Matrix. The matrix is fed into the Assignment Model in which produces the most appropriate allocation of UE to RB, per TTI, which minimizes total resource cost the most. The proposed algorithm, New-SCH algorithm, outstands from RR and MY_SCH [52] scheduling algorithm in [52] in terms of throughput performance, only to be seconded by Best CQI scheduling algorithm. Yet, New-SCH scheduling offers better trade-off of throughput with user fairness than the Best CQI scheme.
Additional Techniques for Scheduling Strategies: 

In multi-cell environment, the downlink transmission is subject to Inter-Cell Interference (ICI). Long distance between an eNB and its neighbouring cells, among other strong interference, can attenuate the signal. The issue requires application of efficient radio resource management techniques reducing the ICI and thus applied a coordinated processing to convert the interference into useful signal [65]. Findings in [66] shows that interference often occurs on multi-cells with different sizes each. It can degrade the total performance of hierarchical multilevel network.

Coordinated scheduling emphasizing on interference-aware joint scheduling based on PF is proposed in [67]. The neighbouring BSs cooperate among one to another to jointly assign frequency resource to UEs via a fast backhaul network. The paper proposed another method of dynamic interference coordination but in which the scheme is not as reliable as the interference-aware joint scheduling algorithm. The scheme proposed in [68] applies joint scheduling, power allocation, MCS selection to handle the ICI and system throughput. Coordinated Multi-Point (CoMP) can be facilitated to employ faster and stricter ICI mitigation techniques. Low complexity is achieved by segregating optimization problem into two suboptimal ones.

Besides requiring efficient resource allocation, multimedia application is generally heavy-loaded and run for long period of time, proving the factor of power optimization as important. UE could be equipped with power saving scheme such as discontinuous reception (DRX). The opportunistic scheduler in [69] trades off QoS satisfaction and power conservation based on six dependencies: channel condition which is acquired from CQI reports, to give more resources to UE with lesser average throughput, to prioritize UE with more buffer space to avoid packet loss, to put higher priority to allocate resources to UE whose data rate lower than GBR, to implement DRX status to restrain delays within threshold defined in QCI and to prioritize UE with oldest packet delay.

Energy-aware strategies can be applied to both eNBs and UEs. Paper [70] discussed the needs of power saving scheme for UEs in LTE-Advanced networks for Internet of Things (IoT). IoT, in general, integrates various devices with the Internet, such as video surveillance and data streaming to central server over long period of time. IoT is expected to continuously transmitting data for long period of time, thus the energy efficiency concern arises. LTE-A standard defined the discontinuous reception/transmission (DRX/DTX) mechanism [7]. DRX/DTX mechanism collaborates with UE and eNB to periodically wake up the UE to receive/transmit data, otherwise let the UE falls into non-wake-up period to save energy. It adopts specific timer and any unexpected delayed data can still be transmitted/received after the regular wake-up period.

Thus, paper [70] addressed the joint optimization on power saving and QoS provision, analyzed the limitation of existing DRX system and proposed an efficient sleep scheduling to optimize the DRX. A similar scheme is introduced in [71]: eNB in sleep mode frequently while handling the high-speed data transmission during low time interval, based on Maximum Throughput scheme. In [72], compacting radio resource allocation as much as possible in the time domain to allow eNB turning off transmission equipment is proposed. This method saves 61% energy consumption yet it requires the tedious task of organizing allocation information in a very minimum control data amount.

The study from [73,74] about the environmental impact of increasing energy consumption due to ever-increasing traffic of mobile network. This cause the green technology becomes popular; power consumption is minimized, eco-sustainability ensured and the operative cost kept low [75]. Bandwidth Expansion Mode algorithm is deployed in [70] to do energy saving in low traffic load scenario. In order to save energy, the algorithm assigns a lower-rate coding scheme to each users and then subsequently expanding the spectrum occupation.

Paper [76] offers a Minimum Transmit Power-based (MP) scheduling algorithm capable to achieve power-efficient transmission without disregarding the throughput gain and fairness. It focuses on the ratio of transmit power per bit and then allocates RB to the least ratio amount. Only during low load traffic the modification of resource allocation policy may give big impact to power saving. Otherwise, maximization of spectral efficiency should be the best choice. Scheduling plays a big factor in power saving.

A proposition to avoid scheduling packets that will be eventually dropped in the destination is forwarded in [77]. Its objective is to reduce radio blocks’ wasting. The pre-emptive scheduler consists of four principal blocs: Approximate Reception Time Control (ARTC), Decision Block (DB), Resources Allocation Bloc (RAB) and the PF bloc. The proposed pre-emption strategy is by defining an “alert zone”, which is the right time to send packet out, for
every real time traffic. Once a packet enters the alert zone, scheduler must serve it. Otherwise, the packet will be dropped in the final receiver [78,79].

CONCLUSION

LTE offers high performing connectivity to support the high demand of multimedia users. In order to ensure LTE is up to the expected performance, QoS is introduced. QoS requirements are translated by RRM into effective radio resource-related decision such as call admission control policy and downlink scheduling algorithm. In general, to achieve high QoS provision, trade-off schemes between two scheduling rules occur, such as throughput-fairness trade-off scheme. QoS provisioning of LTE is so significant that novel downlink scheduling strategies, among many other aspect, has been researched and proposed by various parties. There are conventional downlink scheduling algorithms with its own advantages and weakness (i.e. PF, M-LWDF) which in turns have been modified or enhanced into various new algorithms performing better in actual network scenario.

There are many challenging conditions in real network scenario to be solved about. First, traffic flow is actually heterogeneous; RT and Non-RT services coexist with one another. In general, RT traffic is prioritized higher than Non-RT. The first challenge is about how to satisfy high-demanding RT service’s QoS requirement without neglecting fairness parameter for Non-RT traffic. Second, LTE’s usage of Hard HO causes service degradation to RT users. Hard HO will break current cell connectivity first before making a new one and thus the disruption is intolerable for delay-sensitive applications such as video streaming. Third, users in a network have varying mobility. Fourth, LTE architecture is subject to energy consumption and interference issue. Last but not least, users with bad channel condition (i.e. cell edge users) are suffering from resource starvation due to the common scheduling rule to allocate resources for users with better CQI.

The aforementioned challenges are common concern of recent studies. Various solutions have been proposed to tackle the issue. The current trend of LTE downlink algorithm studies is in regards of how to best-allocate the limited radio resources to satisfy the QoS of multimedia-heavy real-time users. As mobile traffic demands are estimated to grow a thousand-fold by 2020 compared to 2010 [78], it might not be wrong to assume that the scheduling studies trend is going the right way.

REFERENCES


77. Missing
