Journal of Electrical Engineering, Electronics, Control and Computer Science – JEEECCS, Volume 6, Issue 20, pages 21-30, 2020

Survey of Scheduling Schemes in 5G Mobile Communication Systems

Maryam Imran Sheik Mamode
Department of Electrical and Electronic Engineering
University of Mauritius
Réduit, Mauritius
maryamsheikmamode@hotmail.com

Tulsi Pawan Fowdur
Department of Electrical and Electronic Engineering
University of Mauritius
Réduit, Mauritius
p.fowdur@uom.ac.mu

***Abstract* – 5G mobile communication systems are guaranteed to provide a major upgrade on the preceding technologies in terms of connectivity, mobility, speed and latency. They will enable automation in several industries and vertical markets by offering a variety of services and business models. 5G has brought about many changes with respect to scheduling concerning the E-UTRAN (Evolved-Universal Terrestrial Radio Access Network). The objective of this paper is twofold. First, a review of the 5G network architecture, scheduling mechanism and some existing scheduling algorithms is provided including some new features introduced by 5G concerning scheduling. Secondly, emerging scheduling techniques are examined and directions for future works and possible enhancements are proposed.**

Keywords-5G, Mobile Communications, Network Architecture, Scheduling techniques, algorithms, QoS

I. INTRODUCTION

This template, 5G can be interpreted as 5th Generation Mobile technology. This revolutionary mobile technology has altered the ability to use mobile phones with very high bandwidth [1]. Furthermore, 5G technology is capable of handling tremendous amount of data and is able to process unlimited call volumes and data broadcast using the latest mobile operating system. Some features of 5G comprise of a one-millisecond latency, hundred percent coverage, ninety-nine percent of network availability and a bandwidth of 1,000 per unit area. The data speed for 5G will reach up to 10 GB per second and the energy consumption in the network will be reduced by up to ninety percent by 5G technology. The amount of connected devices will increase by hundred which represents an estimate of 50,000 million connected devices at the same time. Japan and Korea had started working on 5G requirements since 2013 and NTT Docomo experimented on the first 5G launch in 2014. Prototype development was embarked upon by Samsung, Huawei and Ericsson in 2013. Japan is planning to kick-start 5G on the occasion of the 2020 Tokyo summer Olympics [2].

The first 5G specifications were introduced in Release 15 which mainly addressed the building up of the New Radio (NR) technical framework and the network architecture comprising of uplink and downlink decoupling, Central Unit (CU) – Distribution Unit (DU) high level segmentation and Stand-Alone or Non-Stand-Alone. Release 16 is also called "5G phase 2" and it will be completed in December 2019, which will be the first finalized 3GPP 5G system in an IMT-2020 submission [3]. The first 5G (NR) standard for commercial deployment emerged in 21 December 2017. Three new concepts have been introduced in 5G with regards to LTE/4G, namely Control Plane (CP)/ User Plane (UP) split, Network Slicing and Service Based Architecture (SBA). The table below shows the mapping for different entities in the network architecture.

TABLE I. MAPPING OF NODES 4G AND 5G [4]

Entity in 4G	Entity in 5G
Home Subscriber Server (HSS)	Authentication Server Function (AUSF) and User Data Management (UDF)
Policy and Charging Rules Function (PCRF)	Policy Control Function (PCF)
Mobility Management Entity (MME)	Access & Mobility management
MME, Serving Gateway (SGW) CP and Packet-Data-Network Gateway (PGW) CP	Session Management Function (SMF)
PGW UP, SGW UP	User Plane Function (UPF)

Some shortcomings of the 4G system with regards to Scheduling include physical layer latency of 1 ms per subframe duration, Synchronous TDD causing inability to adapt to differential loads in each Base Station and Coordinated multipoint transmission which requires huge backhaul capacity. The enhancements proposed by 5G systems include a reduced subframe duration of 0.1 ms, Dynamic TDD which can adapt to instantaneous loads in the networks, and Bidirectional training (BiT) through over-the-air (OTA) to reduce the load on backhaul network [5]. However, there are various challenges for the 5G scheduler design namely the large number of users, more spatial degrees of freedom in terms of MU-MIMO transmission, QoS requirements including latency and reliability, and real-time requirements among others.

This paper provides a detailed review of existing techniques used for scheduling as well as novel techniques which have been introduced in 5G communication systems. This paper also analyses the different scheduling methods proposed by researchers and the shortcomings involved. Relevant areas to pursue further works are presented and possible enhancements to existing systems are suggested.

This paper is organized as follows. Section 2 deals with 5G RAN. Section 3 is concerned with the scheduling mechanism in 5G and scheduling algorithms. Section 4 elaborates on the works performed by researchers on scheduling for 5G networks. Section 5 provides some directions for future works and concludes the paper.

II. 5G NETWORK ARCHITECTURE

As mentioned previously, the 5G communication systems differ from LTE/4G with respect to the CP/UP split, Network Slicing and SBA.

3GPP has proposed a flat architecture where the Control Plane (CP) functions are separated from the User Plane (UP). Operators can use the functional split to dimension and deploy the network as per their requirements [6]. Furthermore, 5G has been designed to reduce dependencies between the Access Network (AN) and the Core Network (CN). This is achieved using a converged access-agnostic core network with a common AN – CN interface that will include different 3GPP and non-3GPP access types.

Network slices are made up of several virtual networks that are built up using virtualization methods with the existing physical infrastructure and spectrum. There are various applications that can represent a network slice. Mobile broadband can constitute a network slice while massive-MTC applications can represent another network slice. A network slice contains its own resources and competencies specific to the application it represents, which seemingly enables it to be an independent network capable of operating on its own.

The 5G network architecture is defined as service-based and the interaction between network functions can be represented in two ways. Reference point representation shows the interaction that exist between the network function nodes described by point-to-point reference point between any two networks. Service-based representation is shown, where network functions within the control plane enables other authorized network functions to access their services. Service-Based Interfaces (SBI) are used by network functions in the 5G Core Control Plane (CP) for interactions. A CP NF (Network Function) can provide one or more NF Services. Both types of representations are shown in figures 1 and 2 [6].

Figure 1. 5G System Service-based architecture [6]

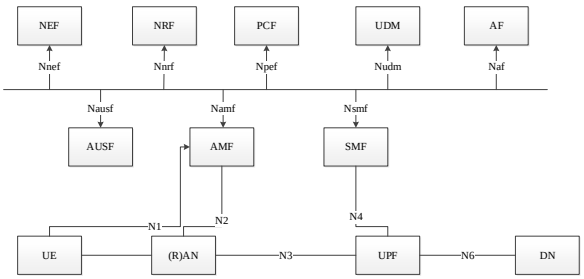
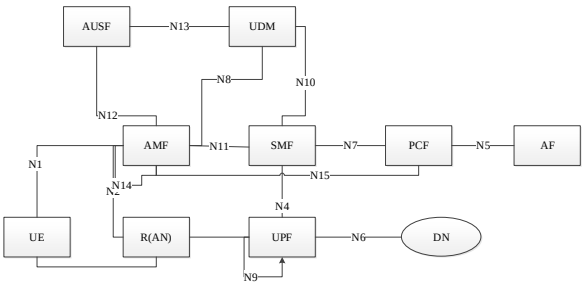


Figure 2. Reference point representation of 5G system



The different entities are defined in table below

TABLE II. ENTITIES IN 5G NETWORK

Entity	Definition
AF	Application Function
AMF	Access and Mobility Management Function
AUSF	Authentication Server Function
NEF	Network Exposure Function
NRF	Network Resource Function
PCF	Policy Control Function
SMF	Session Management Function
UDM	Unified Data Management
UPF	User Plane Function

A point to point interface connects two specific entities (for example N13 is the interface only between the AUSF and the UDM) while in a service-based architecture, the interface originating from any entity represents an Application Programming Interface (API) and it can be used by any other entity [7]. In service-based architecture, the network is more flexible and can adapt easily to unanticipated requirements. In point to point architecture, in case a new network entity needs to be defined, several new interfaces and related protocols will have to be standardized to connect to other existing entities. This often leads to a rigid and complex network. For adding a new network entity in the service-based architecture, only the API of the entity needs to be standardized

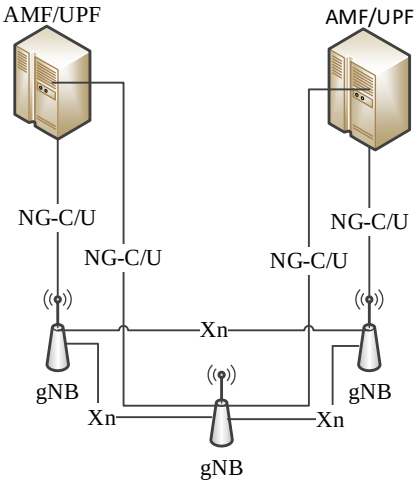
The upper part of Figure 1 indicates the group of network functions that constitute the 5G control plane. They all have service-based interfaces and are

represented as being linked by a network bus rather than point to point connections. The interface name contains the entity name with an "N" as prefix. In this configuration, a NF will query a NRF to establish communication with other NFs. The addition of any new function will only generate a new record in the NRF database. In the lower part of Figure 1, there are point to point interfaces, identified by "N" and a number. Figure is fully reference point representation [7].

A NRF stores information about the network functions and the list of available services. Some services can be accessed through the NEF which is a central point for service exposure and has the main role of authorizing all requests for access which come from sources external to the system. The AUSF is responsible for authentication. The Network Slice Selection Function (NSSF) chooses appropriate network instances for uses and the necessary AMF which handles mobility related processes. The Session Management Function (SMF) provides session management, allocates and manages IP address, DHCP services among other functions. The PCF handles policies and rules in the 5G system and the AF requests the latter for services for impacting on traffic steering rules. The UDM manages user identification related information, access authorization. It also stores data about the user serving NFs and supports Lawful Interception procedures [8].

To elaborate on the 5G RAN, a simplified 5G architecture is shown in figure 3 where the NGC denotes the Next Generation Core and 5G Radio network interfaces include Xn, NG-C, NG-U and Uu (Radio interface) [9].

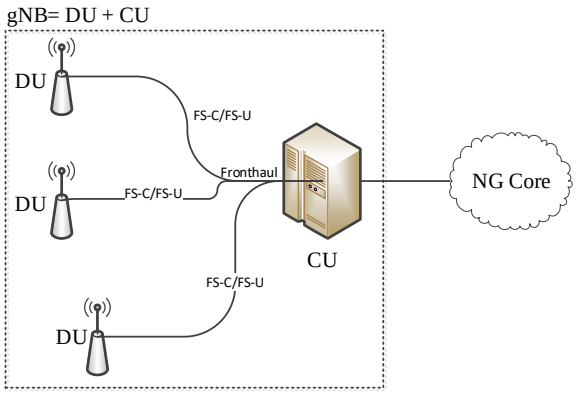
Figure 3. Simplified 5G architecture



The gNB denotes the 5G base station. The gNB node provides NR user plane and control plane protocol links to the UE, and it is liked through the NG interface to the 5GC. The 5G NR (New Radio) gNB is connected to AMF and UPF in 5GC (5G Core Network). Two units represent the protocol layers namely DU (Distributed Unit) and CU (Central Unit).

Figure 4 shows the logical architecture of the gNB with Central Unit (CU) and Distributed Unit (DU). Fs-C and Fs-U deliver control plane and user plane connectivity over Fs interface [10].

Figure 4. Logical architecture of gNB



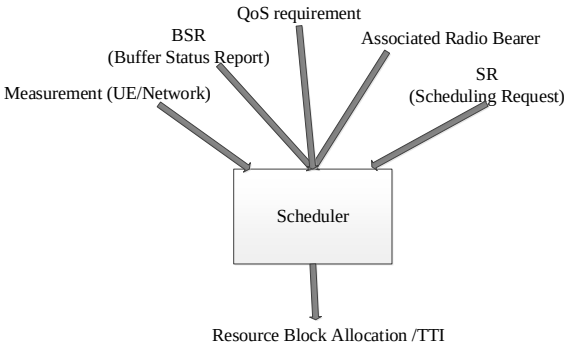
The Central Unit (CU) is a logical node which caters for gNB functions such as Transfer of user data, Control of Mobility, Radio access network sharing, Positioning and Session Management among others. The CU manages the operation of DUs through the Fs interface. A central unit (CU) can be represented by a BBU/REC/RCC/C-RAN/V-RAN.

The Distributed Unit (DU) contains some of the gNB functions, depending on the functional split option. It is controlled by the CU and can be represented by a RRH/RRU/RE/RU [10].

III. SCHEDULING MECHANISM IN 5G AND EXISTING SCHEDULING TECHNIQUES

Scheduling can be defined as the process of assigning resources for data transmission [11]. There are many factors which determine when and what resources are allocated for a specific user. A schema of scheduler illustrating some of the factors considered for scheduling is depicted in figure 5.

Figure 1: 5G scheduler [11]

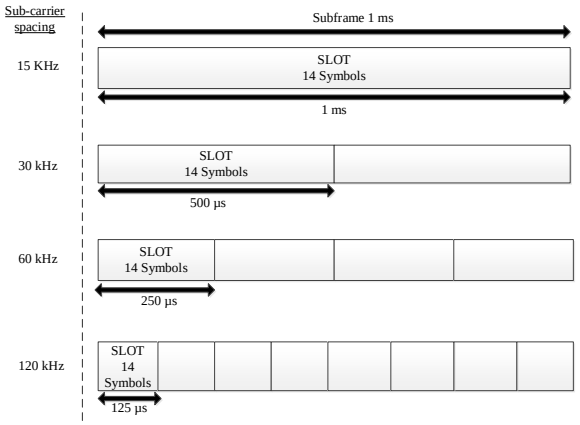


For the scheduler operation, the UE buffer status and the QoS requirements of each UE and associated radio bearers, are taken into account to allocate resources between UEs. The scheduler can also allocate resources based on radio conditions at the UE which are known via measurements made at the gNB and/or communicated by the UE. Radio resources are assigned in a unit of slot (for example one mini-slot, one slot, or multiple slots) and the radio resources are made up of resource blocks. Following a scheduling

request, the UE will take account of resources assigned by receiving a scheduling channel. Among the measurements used to determine scheduler operation, the uplink buffer status reports (evaluating the data buffered in the UE's logical channel queues) are used for providing support for QoS-aware packet scheduling. Furthermore, Power headroom reports (evaluating the disparity between the maximum transmit power of the UE and the approximated power for uplink transmission) are used in power aware packet scheduling [12].

There are basically two broad categories of scheduling namely frequency domain scheduling and time domain scheduling, which is similar to LTE TDD scheduling omitting some time domain factors [11]. A resource element can be defined as the basic time-frequency resource unit that can be utilized for downlink or uplink transmission. It can also be interpreted as one sub-carrier over one OFDM symbol [13]. A Resource Block (RB) is a group of twelve sub-carriers which are contiguous in frequency, over one slot in time. It is the smallest unit of radio resource that can be assigned to a user. Radio resources are classified into radio frames, subframes, slots and mini-slot. The radio frame has a duration of 10ms and constitutes 10 subframes with each subframe of a duration of 1 ms. Each subframe contains one or more adjacent slots containing 14 OFDM symbols. A mini-slot in Release 15 contains 2, 4, and 7 OFDM symbols and the time duration of a slot depends on the sub-carrier spacing as illustrated in Figure 6 [14].

Figure 2: Frame structure 5G



The resource element mapping takes place in the Physical Downlink Shared Channel (xPDSCH) for downlink transmission and in the Physical Uplink Shared Channel (xPUSCH) for uplink transmission, before OFDM signal generation.

The Physical Downlink Control Channel (xPDCCH) transmits the user information, RB allocation and the selected Modulation and Coding Scheme to the UE. The UE interprets the PDCCH payload and verifies if it is scheduled in order to access the correct PDSCH payload. This procedure is repeated at each TTI [15].

5G has introduced several new concepts with regards to scheduling, in order to cater for the significantly large number of users and advanced functionalities. One of the key features introduced by 5G is massive MIMO. It consists using a huge number of antennas and terminals. Massive MIMO relies on Multi-User (MU) MIMO [16]. With a single antenna, transmitting several data streams would result in interference while in MU MIMO, the signals are transmitted through different paths and with the right encoding, the receiving antenna will be able to construct the original signal. Another feature that has improved in feasibility is dynamic TDD which primarily modifies the cell's frame configuration to adjust to the varying traffic in order to ameliorate the system throughput. The application of dynamic TDD has become more plausible in small cell scenarios where a particular user's data rate could be improved by adjusting the TDD pattern to the user's uplink or downlink transmission [17]. Furthermore, 5G was designed to reduce end-to-end latency by ten times as compared to LTE. This includes the Transmit Time Interval, HARQ processing time, Frame Size, Round Trip Time and Discontinuous reception [18].

Some existing time domain packet scheduling algorithms were selected and elaborated upon. The Maximum Rate algorithm capitalizes on high capacity and maximum throughput by evaluating fluctuating channel conditions. It prioritizes users with a more favourable channel condition while UEs with severe channel degradation are not scheduled. Thus, there is no fair assignment of resources among users. In each TTI, the algorithm selects a UE maximizing the following algorithm:

$$x_i = y_i(t), \tag{1}$$

Where $y_i(t)$ represents the instantaneous data rate of user i using the total bandwidth at time interval t.

The Round Robin (RR) algorithm was developed in order to assign equitable resources among users in LTE mobile systems. Contrary to the Maximum Rate algorithm, the RR algorithm allows users opportunity in turn to transfer packets. Thus, RR algorithm greatly improves fairness but also causes throughput degradation due to the fact that channel quality is not considered [19]. The Proportional Fair (PF) algorithm was developed for CDMA networks catering to Non-Guaranteed Bit Rate (GBR) services. Its purpose was to reach a decent trade-off between fairness and throughput by increasing throughput of UEs which have better instantaneous achievable data rate than mean throughput. However, the PF algorithm is not optimally designed for real time services as it does not take into consideration the buffer status of the UE. The scheduling formula for PF is represented by equations below.

$$a_i(t) = \frac{b_i(t)}{B_i(t)}, \tag{2}$$

$$B_i(t+1) = \left(1 - \frac{1}{t_c}\right) B_i(t)$$

$$+C_i(t+1) * \frac{1}{t_c} * b_i(t+1) \quad (3)$$

$$C_i(t+1) = \begin{cases} 1 & when\ packets\ of\ user\ i\ are\ allocated\ at\ interval\ t+1 \\ 0 & when\ packets\ of\ user\ i\ are\ not\ allocated\ at\ interval\ t+1 \end{cases} \quad (4)$$

Where: $b_i(t)$ represents the momentary data rate for user i calculated during time interval t,

$B_i(t)$ indicates the mean throughput of user i during time interval t,

$C_i(t+1)$ indicates the selection of the packet for transmission during time interval t+1,

t_c denotes a time constant which can be used to capitalize on throughput and fairness with the PF algorithm.

The Blind Equal Throughput (BET) algorithm has been applied in LTE systems and as the name implies, it does not take channel conditions into consideration for resource assignment. It records the past instance mean throughput of each UE to come to a fair allocation among UEs. The BET algorithm seeks to maximize $d_i(t)$ as follows:

$$d_i(t) = \frac{1}{B_i(t)}, \quad (5)$$

Where $d_i(t)$ represents the preference of a UE at a specific time interval and $B_i(t)$ still indicates the mean throughput during time interval t for user i. The BET algorithm is not suitable for getting high throughput when compared with PF and Maximum Rate as it fails to taken into account channel conditions. The Delay Prioritized Scheduling (DPS) algorithm includes packet delay information. In order to satisfy the QoS requirements for GBR services, the DPS algorithm gives priority to UEs that have delays above a threshold in downlink LTE networks.

$$x_i(t) = T_i - F_i(t), \quad (6)$$

Where $F_i(t)$ represents the Head of Line (HoL) packet delay during time interval t for user i,

T_i indicates the buffer's delay threshold depending on the service category

$x_i(t)$ is the real time of the HoL packet during time interval t for user i.

The Modified-Largest Weighted Delay First (M-LWDF) algorithm has for aim to increase the QoS of real-time UEs. Several parameters, namely packet delay, mean throughput, momentary data rate and bandwidth are taken into account for the M-LWDF algorithm. This algorithm has been applied in CDMA-High Data Rate (HDR) systems. The equation used for M-LWDF algorithm is shown below.

$$g_i(t) = h_i * F_i(t) * \frac{b_i(t)}{B_i(t)}, \quad (7)$$

$$h_i = -\frac{log\delta_i}{T_i}, \quad (8)$$

Where g_i indicates user i's QoS requirement,

$F_i(t)$ represents the HoL packet delay during time interval t for user i,

$b_i(t)$ indicates the momentary data rate

$B_i(t)$ is the mean throughput for user i in time interval t.

δ_i indicates the Packet Loss Ratio and and T_i indicates the buffer delay threshold for user i.

The main application of Exponential Rule (EXP) Algorithm was in HDR/CDMA framework for real-time and non-real time services. It is represented by the following formula.

$$k_i(t) = \alpha_i * F_i(t) * \frac{b_i(t)}{B_i(t)} * exp\left(\frac{\alpha_i * F_i(t) - \alpha F_{-avg}}{1 + \sqrt{\alpha F_{-avg}}}\right), \quad (9)$$

$$\alpha F_{-avg} = \frac{1}{N} \sum_{i=1}^{i=N} \alpha_i * F_i(t), \quad (10)$$

Where $k_i(t)$ represents the precedence for user i to receive packets during time interval t,

α_i indicates the QoS requirement for user i,

$F_i(t)$ stands for the HoL packet delay for user i during time interval t,

$b(t)$ represents the momentary data rate,

$B_i(t)$ is the average throughput for user i during time interval t

N indicates the total number of users.

Channel-Dependent Earliest Due Deadline (CD-EDD) was developed to cater for sensitive traffic in mobile systems. Similar to the M-LWDF and EXP algorithms, the CD-EDD algorithm takes into consideration mean throughput, momentary data rate and information concerning packet delay when assigning resources. In case the mean throughput and momentary data rate of a particular user are similar, the CD-EDD will consider the user with the more urgent HoL delay as a priority for transmission while M-LWDF and EXP algorithms consider the longest buffer delay of the base station as priority. The formula used for CD-EDD is shown below.

$$m(t) = \alpha_i * \frac{b_i(t)}{B_i(t)} * \frac{F_i(t)}{T_i - F_i(t)}, \quad (11)$$

Where $m_i(t)$ indicates the precedence for user i during time interval t,

α_i shows the QoS requirement for user i,

$b_i(t)$ is the instantaneous data rate,

$B_i(t)$ represents the mean throughput for user I during time interval t,

$F_i(t)$ is HoL packet delay for user i at time interval t

T_i indicates the buffer delay threshold for user i.

IV. OVERVIEW OF STATE OF THE ART RESEARCH IN SCHEDULING

In this section a review of state-of-the-art papers on scheduling schemes for 5G is given. Gaps are also identified for potential future works from the current research.

A. *Downlink scheduling and Resource allocation for 5G MIMO-Multicarrier: OFDM vs FBMC/OQAM [20]*

Femenias, G. et al., have developed a cross-layer downlink Scheduling and Resource allocation (SRA) algorithm using a system that models the queuing process at the data-link control layer with either OFDM or Filter Bank Multi-Carrier (FBMC)/ Offset QAM (OQAM) waveforms [30]. The latter boasts of an advantage of increased spectral efficiency over subcarrier orthogonality in frequency selective channels. Time and frequency synchronization at the receiver can be achieved easily only in the downlink of OFDM systems. This fact, combined with the use of cyclic prefix (CP) for controlling Inter-Symbol Interference (ISI), restricts the maximum spectral efficiency of the system and hence limits use of OFDM in systems. The authors have compared systems using both modulation waveforms in terms of goodput, delay, fairness, and service coverage. The proposed cross-layer SRA was evaluated using a LTE link level mode. A downlink single cell MIMO-multicarrier system which has a main base station serving few mobile stations, uniformly distributed over the whole coverage area is considered. The macro cell propagation model for urban area was used to simulate path losses. A base station antenna height of 30 m and a log-normally distributed shadow fading with a standard deviation of 10 dB was assumed. Furthermore, a MIMO configuration of 2x2 was considered. Results showed that the omission of CP in FBMC/OQAM based system and the fact that it is possible to operating without large guard bands, is a huge benefit despite the new sources of interference generated, which were not present in OFDM systems. The theoretical gain of 18% which the FBMC/OQAM system has over OFDM system is fully realized in practice.

The gap in this research work can be attributed largely to the use of LTE/LTE-Advanced parameters as well as using a LTE model for simulation. For example, a default MIMO configuration of 2x2 and system bandwidth of 10 Mhz has been used. 5G contains massive MIMO which is a configuration of 16x16 and beyond. The minimum bandwidth used for 5G is 50MHz. Furthermore, three scheduling algorithms have been investigated in this research namely PF, Exponential and M-LWDF. Other algorithms, especially channel-aware ones can be investigated and the performance of the different scheduling techniques compared. Sparse Code Multiple Access (SCMA) and Non-Orthogonal Multiple Access (NOMA) are other multiple access schemes that can be used and evaluated against legacy OFDM systems and factors affecting their performances determined.

B. *Effective 5G Wireless Downlink Scheduling and Resource Allocation in Cyber-Physical Systems [21]*

For efficient cross-layer downlink Scheduling and Resource Allocation (SRA), Vora, A. and Kang K. have sought to devise a dynamic programming algorithm which has a time complexity that is polynomial. The proposed algorithm considered the channel and queue state and supported fairness. It was formulated based on the available bandwidth and required RBs, while allocating resources to maximize the total utility. Some 5G cases where the algorithm could be used include Machine Type Communication (eMTC), Ultra-Reliable Low Latency Communication (URLLC) and enhanced Mobile BroadBand (eMBB). The proposed algorithm was evaluated against another cross-layer greedy algorithm which was used for eMTC, URLLC and LTE, developed in [20]. A greedy algorithm usually takes a decision based on certain criterion disregarding choices that occurred previously or that will happen in the future, thus yielding a sub-optimal solution. For simulation, a base station that has an infinite traffic queue was assumed. MATLAB LTE toolbox was used for performance evaluation and the lteDLResourceGrid function was altered so that both the greedy SRA and dynamic programing algorithms could be implemented. The goodput and fairness level was measured for two main 5G waveforms namely OFDM and FBMC. The delay spread channel models chosen included the Extended Pedestrian A (EPA), the Extended Vehicular A (EVA) and the Extended Typical Urban (ETU) models. It was observed that for eMTC and URLLC, the proposed SRA algorithm outdid the greedy algorithm by up to 17.24%, 18.1%, 2.5% and 1.5% for average goodput, correlation impact, goodput fairness and delay fairness, respectively. For LTE, the new SRA algorithm exceeded in performance by 60%, 2.6% and 1.6% for goodput, goodput fairness and delay fairness.

The main gap that has been identified in this research work is the use of LTE toolbox to simulate a 5G system. Moreover, the authors have developed their novel scheduling algorithm using the algorithm in [20] as baseline and also compared only those two algorithms. Other cross-layer SRA algorithms could also have been investigated and compared. Some other factors like QoS and Buffer status reports can be taken into consideration to devise an optimal algorithm. As the authors mentioned in their conclusion, other channel waveforms for 5G can also be used and the performance evaluated with common waveforms.

C. *Towards 5G: A Reinforcement Learning based scheduling solution for data traffic management [22]*

Comsa I. et al., have come forth with a scheduling program that has the ability to choose from various

scheduling rules based on the instantaneous scheduler states. It is used in cases with stringent QoS requirements where packets delays and packet drop rates are minimized. In effect, a flexible RRM (Radio Resource Management) packet scheduler is devised, which has the ability to adapt to dynamic scheduling environments. RRM functions include power control, resource management, interference management and packet scheduling among others.

The proposed technique uses different scheduling rules over each TTI instead of a single scheduling rule, based on instantaneous conditions like dynamic traffic load and QoS parameters. Machine Learning (ML) is use as basis for a Reinforcement Learning (RL) principle which learns the scheduling rule at each instantaneous scheduler state for improved delay and Packet Drop Rate (PDR) of users. Each scheduling rule is represented by a function and the RL algorithm is used to update those functions for each TTI until the learning process is complete. Five such RL algorithms were evaluated and compared in terms of dynamic network conditions and traffic types among others. The main aim of the proposed RL program is to enhance the heterogeneous delay and Packet Drop Rate (PDR) requirements for Constant Bit Rate (CBR) and Variable Bit Rate (VBR) traffic models. For simulation, an OFDMA downlink transmission was considered and a RRM scheduler C/C++ object oriented tool having LTESim simulator was used for performance evaluation of the proposed system [23].

Four classic scheduling rules were considered namely Logarithmic (LOG), EDF, and two EXP algorithms. A system bandwidth of 20 MHz and ARQ scheme of maximum 5 retransmissions were used. For the delay objective, the novel program surpasses the classic scheduling rules in performance for both cases of VBR and CBR traffic. For the latter, there is a gain of more than 10% of feasible TTIs. The proposed framework notes a gain of 15% as compared to scheduling rules for CBR traffic by choosing suitable scheduling rules corresponding to traffic loads, network conditions and QoS requirements. In case of VBR traffic, there is a gain of 10% due to the fact there are larger-sized packets.

The main gap identified in this research work is the use of LTE systems to simulate research about 5G. The minimum bandwidth in 5G is 50 Mhz. The idea of using a learning entity to yield better performance can be developed and deep learning algorithms can be used for that effect. Deep learning technique is part of ML in Artificial Intelligence (AI) which is able to learn independently from unstructured data. Furthermore, other waveforms apart from OFDM and other scheduling rules can be considered and evaluated.

D. Agile 5G scheduler for Improved E2E performance and flexibility for different network implementations [24]

A multi-user scheduling framework has been presented in [34] by Pedersen K., et al. from an End-to-End (E2E) perspective. The authors mainly sought to present an extensive survey of packet scheduling enhancements brought about by 5G. A novel E2E QoS framework was studied and it offered improved scheduling functions for satisfactory QoS. The QoS framework worked in accordance with MAC scheduler which contained flexible characteristics of the 5G system namely dynamic TTI sizes, flexible timings and punctured scheduling, among others. Extensive system level simulations have been generated in order to confirm the superiority of 5G scheduling enhancements. A typical three sector macro site, operating at 2 GHz with a bandwidth of 10 MHz and 2x2 MIMO, was considered. First the E2E eMBB performance is demonstrated through file download over TCP. A 2ms delay and homogeneous traffic following a Poisson distribution was assumed. It was observed that longer TTIs yielded higher average spectral efficiency while excessive queuing delays were noticed with shorter TTI sizes.

Simulation results also showed that fewer radio resources for HARQ retransmissions have been punctured but the drawback is more latency for eMBB users because there is greater chance of generating another HARQ retransmission as compared with the case where the first retransmission contained the full transport block.

The gap in this research work is mainly the absence of realistic 5G parameters for simulation. A MIMO configuration of at least 16x16 and bandwidth of 50 MHz should have been considered for evaluation of 5G system. Moreover, this paper only deals with the enhancements that have been brought about by 5G but it does not take into consideration other factors affecting scheduling namely QoS requirement, buffer status and channel measurements.

E. Payload-size and Deadline-aware scheduling for upcoming 5G networks: Experimental Validation in High-load scenarios [25]

Monhof S. et al., have evaluated a Payload-size Deadline-aware (PayDA) scheduling algorithm by making use of a Software-Defined Radio (SDR) based eNodeB. The PayDA algorithm is Real-Time aware, based on the EDF scheduling rule with the improvement of taking into account the left-over packet size for each data flow. The novel technique was implemented and evaluated using a 3GPP Rel. 9 compatible program called the CommAgility SmallCellSTACK and RBs are allocated to custom-built users based on embedded PCs and Huawei ME909s-120 LTE modules. The PayDA algorithm works by calculating the scheduling formula for every Dedicated Radio Bearer (DRB), hence yielding several DRBs for one UE. The specific DRBs for each UE is processed and the scheduled rank calculated with the scheduling metric. The highest scheduling metric prevails for the scheduled DRB, to obtain the required amount of RBs. For performance analysis of the PayDA technique, the HOL delay, Deadline-Miss Ratio (DMR) and data rates, are obtained from the Radio Link Control (RLC) and scheduler modules in the eNodeB. In the authors' previous work [26], the PayDA algorithm was evaluated only through simulations using LTE-Sim while in this work, the algorithm is validated for various data traffic scenarios using extensive laboratory measurements. Experimental results showed that the mean HOL delay

increased as the number of users peaked for homogeneous traffic, owing to packet flooding of queues due to a high cell load. Furthermore, the PayDA algorithm caused a significant decrease in HOL delay compared to Maximum Rate, PF and RR algorithms. PayDA algorithm also managed to cause a considerable decrease in average DMRs for heterogeneous traffic as compared with the othe scheduling algorithms.

The main loophole is this research is the use of LTE based systems to simulate the proposed algorithm which has been projected for use on 5G system. A 5G system-based simulations and experimental set-up can be considered and results compared for better realistic analysis. Moreover, channel conditions and user mobility have not been taken into account and those will have a great impact on scheduling results in real-life situations. Other scheduling algorithms especially channel-aware ones can be used and compared with the PayDA algorithm.

F. Inter-cellular scheduler for 5G wireless networks [27]

Gueguen C., Ezzaouia M. and Yassin M. have devised an inter-cellular scheduler which has a dynamic cell bandwidth assignment to better serve overloaded cells and to mitigate issues in ensuring high QoS. It was based on Mean Cell Packet Delay Outage Ratio (MCPDOR) which took note of the cell emergency in order to get more radio resources and chose the targeted cell to help. The proposed scheduler was also referred to as Inter-cellular Bandwidth Fair Sharing scheduler (IBFS). It used a hybrid Inter-Cell Interference Coordination (ICIC) methodology of two parts. In the first part, a central controller managing a cluster with multiple cells, determined the bandwidth sharing between different cells. In a second part, UE scheduling was determined by the local cell as the role of resource assignment was given to the base stations. Thus, resources were dynamically adjusted among adjacent cells to achieve user satisfaction. Two versions of the scheduler was proposed namely IBFSload which drew out bandwidth from cells with less data to process, to allocate to cells with more load and IBFSMCPDOR that allocated more resources to cells with the highest delay outage by taking from other cells. For simulation, an OFDM system was used, free space path loss and multipath Rayleigh fading was considered. Frequency reuse-3 model has been used for comparison and users having high variable bit rate have been allocated cell 1 and all other cells have less variable traffic but with global traffic load 15% more than cell 1. It is observed that the reuse-3 model yielded a MCPDOR value of 9% in cell 1 and nearly 0% in other cells. Results showed that IBFSMCPDOR surpassed IBFSload and the reuse-3 model in performance. In the case of overloaded cells, IBFSload and IBFSMCPDOR greatly diminished the peak delay with a mean packet delay of 23ms and 22.75 respectively while the reuse-3 model provided an average packet delay of 67.25ms. In case of situations with varying radio conditions, it is noted that IBFSMCPDOR performs better than IBFSload.

As mentioned by the authors, the gap identified is to determine the minimum bandwidth that can be taken from another cell without affecting the donor cell's QoE. Furthermore, the authors have considered traffic load and queue delay as basis for the novel scheduler proposed. Other factors can also be considered including QoS requirement and measurement reports. The use of Reinforcement learning and deep learning algorithms can also be integrated to provide better scheduling operations. Instead of allocating bandwidth to help offload traffic, other solutions can also be devised.

G. Comparison of Data Traffic Scheduling Techniques for Classifying QoS over 5G mobile networks [28]

Dighriri M., et al. have investigated and analysed three scheduling algorithms namely Priority Queuing (PQ), First-In-First-Out (FIFO) and Weighted Fair Queuing (WFQ). A data traffic aggregation model is proposed to implement the three algorithms. It is based on aggregation of data from various Machine-to-Machine (M2M) devices at the Packet Data Convergence Protocol (PDCP) layer of a Relay Node (RN). At the PDCP layer, the payload has a minimum of headers. The easiest technique to be used for scheduling is FIFO, where the first packet in the queue is served first regardless of priority, protection or fairness. However, if the first packet is blocked, the rest of the queue is also blocked. Moreover, there is unfair bandwidth assignment among different flows and FIFO is also prone to jitters. PQ, on the hand caters for priority of data traffic. It serves the high priority traffic first with low-delay propagation. Nevertheless, starvation of lower priority traffic occurs occasionally. WFQ cater for priority as well as fairness. It assigns resources to traffic with high priority and then proceeds to divide the available bandwidth among high-bandwidth flows in a fair way. However, it does not support classification of flows and leads to multiple flows in one queue.

The gap in this research in the investigation of data traffic of M2M devices and disregarding other data traffic flows which are mainly from mobile systems. There is a major difference between traffic and simulation through a relay node and through air interface. 5G systems using wireless communication can also be investigated with the three algorithms mentioned and compared to the current research. Moreover, other scheduling algorithms can also be considered.

H. QoS-Driven Scheduling in 5G Radio Access Networks - A Reinforcement Learning Approach [29]

Comsa I., De-Domenico A. and Ktenas D. have devised a scheduler capable of selecting a scheduling mechanism at each TTI, in order to improve user satisfaction in terms of QoS requirements. It involved the use of Neural Networks (NN) and Actor Critic (AC) Reinforcement Learning to learn from past occurrences. Given the complexity of the scheduling framework due to the ongoing and multi-dimensional state space, the AC solutions are approximated via

NNs. The scheduler state and system conditions at any instant are taken into consideration as well as the QoS requirements in terms of delay, Packet Loss Rate (PLR) and GBR. At TTI=t, the scheduler took note of its state and assigned a scheduling mechanism according and at TT=t+1, the scheduler is in the presence of a new state with an associated value from the previous action. Likewise, the latter undergoes a series of events until an optimal policy is achieved. The Actor in AC yielded the best action while the Critic evaluated each state to improve the scheduling strategy. For simulations, an OFDMA system is used and a cell cluster size of 7 is assumed with the central cell targeted for training process. The scheduling mechanisms were each trained for 107 ms and the user equipment were switched from IDLE to ACTIVE mode and back after every 103 ms. The user speed was assumed to be 30 kmph, the mobility of the UE was random and no handover procedure was done. Ten simulations of 50s were used for the exploitation stage and the corresponding results were averaged. An advanced version of LTE-Sim was used for simulation. Performance results showed that only 106 ms was required to learn the scheduling mechanisms for CBR, VBR and video traffic. At the exploitation stage, the scheduler was able to considerably increase the time for satisfying the QoS requirements.

The main gap in this research consisted the use of a LTE system to simulate 5G performance. A 5G system should instead be used and the results compared with findings of this research work. Moreover, it is mentioned that the RL controller's performance cannot be checked analytically before using the scheduler. Thus we cannot compare and evaluate the RL algorithm. Other deep learning algorithms can also be used and compared with the proposed scheduler where factors affecting QoS including the channel state and queue state can be considered. Future predictions of the scheduling state and network conditions can be generated to provide for a more robust scheduling strategy.

I. Performance Evaluation and Comparison of Scheduling Algorithms on 5G Networks using Network Simulator [30]

Perdana D., Sanyoto A. N. and Bisono Y. G. have compared the performance of Round Robin (RR) and Proportional Fair (PF) algorithms on 5G mmWave system. Several scenarios have been considered including voice and video traffic. The flexible TTI concept which form part of the 5G technology, has been implemented in the scheduler. Thus, the two scheduling algorithms are evaluated in terms of delay, throughput and fairness. RR is a channel independent algorithm which works on fairness by rotating the queue process with the same time assignment for each process. PF, on the other hand, takes note of the channel conditions and has the task of equitable allocation of resources without compromising throughput and fairness. Network Simulator 3.27 was used for simulations including an additional mmWave module. The number of nodes was increased gradually from 20 to 100, in increments of 20. The UEs were randomly positioned with no mobility. The value of the packet size and data rate was adapted to those on one of the VoIP codecs G.729 and video codec H.264. Simulation results showed that RR boasted of a throughput value which was 3.65% better than PF with the same fairness index. Thus, RR is the better choice for voice traffic. However, for video traffic, PF exceeded RR in throughput by 1.24, leading to PF being the better candidate for video traffic. RR algorithm had better fairness index for both voice and video.

The gap in this research work is the fact that user mobility and varying channel conditions have not been taken into account. As the authors mentioned in the conclusion section, other scheduling algorithms could also be implemented with the Network Simulator. Their performance could then be evaluated and compared with this research.

J. Scheduling Algorithms for 5G Networks with Mid-haul Capacity Constraints [31]

Sinha a., Andrewsy M. and Ananthy P. have used a virtual RAN architecture where Remote Units (RU) are linked to a Central Unit (CU) through a mid-haul which is made up of a Passive Optical Network (PON). The bandwidth requirement for a mid-haul configuration varies according to the user traffic volume and channel conditions. It has been observed that greedy methods could not give optimal results due to inability to cater for the air interface constraint and PON capacity constraint. The authors have sought to propose an optimal rate assignment algorithm for fixed RB allocation to a user and two other algorithms to cater for the PON capacity constraint. For simulations, only the downlink traffic is considered and it is assumed that the data buffers of users are invariably full in the CU. To reduce the latency, the scheduling process occurred such that no queue was built-up in the RUs and all scheduling decisions made by the CU. The two algorithms proposed were adapted from the PF algorithm. The first one, MAX-YIELD made optimal use of radio resources while the second one, MAX-VALUE, took note of the PON capacity constraints. For simulations, 1000 users were assumed using 100 RUs, with PON capacity of 1 Gbps and 1000 Gbps but having one PON limitation. A two-dimensional Poisson distribution was assumed for users and RUs over the service area with a channel bandwidth of 20 MHz. Simulation results showed that for 1000 Gbps PON capacity, the latter does not limit the system performance and hence the MAX-YIELD algorithm achieved an optimal value. For capacity of 1 Gbps, neither the MAX-YIELD not MAX-VALUE algorithms are optimal. It was observed that Dynamic Programming (DP) and Linear Programming (LP) algorithms performed better than the two aforementioned algorithms while being aware of both capacity constraint and channel conditions.

The gap in this research is the fact that the authors have only considered the PON capacity as limitation for performance of 5G systems. Other factors forming part of the air interface such as UE and network measurement and user mobility have not been taken into consideration as well as the QoS requirement. Moreover, PF algorithm has been used as basis to develop other algorithms. Other scheduling algorithms could be used, especially channel aware ones, and

their performances evaluated and compared with those in this research work.

CONCLUSION

In this paper, an extensive survey of 5G scheduling works has been undertaken and several state-of-the-art scheduling techniques have been studied. The 5G network and the scheduling mechanism in 5G is presented including the various factors affecting scheduling. Several research works have been analyzed and the gaps in these papers have been identified. In most cases, it has been observed that simulations have been achieved using LTE/LTE-Advanced parameters and system models. Other additional enhancements/ further investigation has also been identified in those research works, which could lead to improvement of scheduling operation. The use of reinforcement learning and the application of Artificial Intelligence (AI) in 5G schedulers appears to be a very promising prospect. Therefore, as future works, the proposed solutions will be implemented and the performances of the systems compared with results of the research papers mentioned in order to provide an optimal 5G scheduling framework.

REFERENCES

[1] FreeWimaxInfo.com, 2019, What is 5G Technology and Features [online]. Available at http://freewimaxinfo.com/5g-technology.html

[2] Gemalto, 2019, Introducing 5G networks – Characteristics and Usages [online]. Available at https://www.gemalto.com/mobile/inspired/5G

[3] 3GPP, Releases [online]. Available at https://www.3gpp.org/specifications/44-specifications/releases

[4] EventHelix, 2018, 5G Service-Based Architecture (SBA) [online]. Available at https://medium.com/5g-nr/5g-service-based-architecture-sba-47900b0ded0a

[5] Venkatraman G., Laddu P. and T• olli A., Duplexing and Scheduling for 5G Systems,Centre for Wireless Communications (CWC), Department of Communications Engineering (DCE), University of Oulu, Oulu, FI-90014

[6] Grandmetric, 2017, 5G Core Network – a short overview [online]. Available at: https://www.grandmetric.com/2017/06/05/5g-core-network-a-short-overview/

[7] Detti A.,2018, 5G Italy White Book: from Research to Market : Functional Architecture, ebook. Available at: https://www.5gitaly.eu/wp-content/uploads/2019/01/5G-Italy-White-eBook-Functional-architecture.pdf

[8] ETSI, 2018, MEC in 5G networks, ETSI White Paper No. 28 [online]. Available at: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf

[9] RF Wireless World, 5G NR network interfaces-Xn,NG,E1,F1,F2 interface types in 5G [online]. Available at https://www.rfwireless-world.com/Tutorials/5G-NR-network-interfaces.html

[10] Techplayon, 2017, 5G NR gNB Logical Architecture and It's Functional Split Options [online]. Available at http://www.techplayon.com/5g-nr-gnb-logical-architecture-functional-split-options/

[11] ShareTechnote, 5G/NR - Scheduling [online]. Available at: http://www.sharetechnote.com/html/5G/5G_Scheduling.html

[12] ETSI, 2018, ETSI TS 138 300 V15.3.1 (2018-10) [online]. Available at: https://www.etsi.org/deliver/etsi_ts/138300_138399/138300/15.03.01_60/ts_138300v150301p.pdf

[13] Sciencedirect, 2018, Physical Resource Block [online]. Available at: https://www.sciencedirect.com/topics/computer-science/physical-resource-block

[14] Intel, 2018, 5G NR-Driving Wireless Evolution into new vertical domains

[15] Aminu L.M. et al., Downlink scheduling algorithms in LTE Networks: A survey. In IOSR Journal of Mobile Computing & Application (IOSR-JMCA), Volume 4, Issue 3 (Jul. - Aug. 2017), PP 01-12

[16] IEEE, Massive MIMO for 5G, 2017 [online]. Available at https://futurenetworks.ieee.org/tech-focus/march-2017/massive-mimo-for-5g

[17] Pauli V., Li Y. and Seidel E., Dynamic TDD for LTE-A and 5G. Nomor Research GmbH, Munich, Germany, 2015

[18] YTD2525 – UPDATE ON TELECOM DEVELOPMENT AND INNOVATION UNTIL THE YEAR 2525, Latency in 5G, Legacy in 4G, 2019 [online]. Available at https://ytd2525.wordpress.com/2014/07/04/latency-in-5g-legacy-in-4g/

[19] Heidari R., Packet Scheduling Algorithms in LTE Systems, A Thesis Submitted to University of Technology Sydney Faculty of Engineering and Information Technology University of Technology Sydney New South Wales, Australia, October 2017

[20] Femenias, G.; Riera-Palou, F.; Mestre, X.; Olmos, J.J. Downlink Scheduling and Resource Allocation for 5G MIMO-Multicarrier: OFDM vs FBMC/OQAM. IEEE Access 2017, 5, 13770–13786.

[21] Vora A. and Kang K. Effective 5G Wireless Downlink Scheduling and Resource Allocation in Cyber-Physical Systems. In IEEE 5G World Forum (5GWF), Santa Clara, CA, USA, 9–11 July 2018.

[22] Comsa I., et al.,Towards 5G: A Reinforcement Learning-based Scheduling Solution for Data Traffic Management. In IEEE Transactions on Network and Service Management, Volume: 15 , Issue: 4 , Dec. 2018

[23] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE Cellular Systems: An Open-Source Framework,". In IEEE Transactions on Vehicular Networks, vol. 60, no. 2, pp. 498 – 513, 2011.

[24] Pedersen, K. et al., Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations. In I E E E Communications Magazine, 2018.

[25] Monhof, S., et al, Payload-size and Deadline-aware Scheduling for Upcoming 5G Networks: Experimental Validation in High-load Scenarios. In IEEE 88th IEEE Vehicular Technology Conference (VTC-Fall), 2018.

[26] Haferkamp M., Sliwa B., Ide C., and Wietfeld C., "Payload-size and deadline-aware scheduling for time-critical cyber physical systems," in Wireless Days 2017, Porto, Portugal, Mar 2017.

[27] Gueguen C., Ezzaouia M. and Yassin M., Inter-cellular scheduler for 5G wireless networks. In Physical Communication, 2016.

[28] Dighriri M., et al., Comparison Data Traffic Scheduling Techniques for Classifying QoS over 5G Mobile Networks. In 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA), 2017

[29] Comsa I., De-Domenico A. and Ktenas D., QoS-Driven Scheduling in 5G Radio Access Networks - A Reinforcement Learning Approach. In GLOBECOM IEEE Global Communications Conference, 2017

[30] Perdana D., Sanyoto A. N. and Bison Y. G., Performance Evaluation and Comparison of Scheduling Algorithms on 5G Networks using Network Simulator, INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL ISSN 1841-9836, e-ISSN 1841-9844, 14(4), 530-539, August 2019.

[31] Sinha A., Andrewsy M. and Ananthy P., Scheduling Algorithms for 5G Networks with Mid-haul Capacity Constraints. In arXiv:1903.11270, 2019