Evaluation of a Cross Layer Scheduling Algorithm for LTE Downlink

Andrijana Popovska Avramova, Member, IEEE, Ying Yan, Member, IEEE, and Lars Dittmann, Member, IEEE

Abstract—The LTE standard is a leading standard in the wireless broadband market. The Radio Resource Management at the base station plays a major role in satisfying users demand for high data rates and quality of service. This paper evaluates a cross layer scheduling algorithm that aims at minimizing the resource utilization. The algorithm makes decisions based on channel conditions, the size of transmission buffers and different quality of service demands. Simulation results show that the new algorithm improves the resource utilization and provides better guarantees for service quality.

Keywords — buffer-aware, LTE, QoS, subcarrier allocation.

I. INTRODUCTION

NOT only because of the technology but also because it fulfills defined requirements for a pure 4G generation and backward compatibility with its previous generations (GSM, UMTS…), 3GPP Long Term Evolution (LTE) is leading the wireless mobile world [1]. The air interface of this all-IP-based network architecture utilizes Single Carrier Frequency Division Multiple Access (SC-FDMA) in the uplink and Orthogonal Frequency Division Multiple Access (OFDMA) in the downlink transmission supporting both the time (TDD) and frequency (FDD) division duplex modes. The capacity can be enhanced with multiple antenna systems, while in LTE-Advanced techniques such as Carrier Aggregation (CA) and Coordinated Multipoint (CoMP) can be used as well in order to increase the air interface throughput.

Due to the all-IP based architecture, the air interface needs to accommodate a mixture of real and non-real time services. This means that all traffic including delay sensitive services needs to be scheduled [2]. The main purpose of the LTE scheduling system in the Base Station (BS) is to prioritize and allocate the available frequency-time resources to specific single user equipment (UE). The scheduling which is performed at the MAC (Medium Access Control) layer is not standardized and it is an implementation specific mechanism. The overall system performance and reusability of resources are mainly influenced by the scheduling algorithm [3]. The design of a downlink scheduling algorithm is a complex procedure and presents a number of design challenges, such as maximization of system capacity and spectral efficiency, bit error performances, fairness approach, etc.

So far, there has been a lot of research in modeling the scheduler in order to achieve the highest performance while avoiding latency and starvation problems. In this paper we evaluate a cross layer scheduler scheme that considers the physical layer (channel conditions), application layer (service type) and RLC layer (transmission buffer size). The results are compared with the maximum capacity algorithm with Quality of Service (QoS) service classification. The algorithms are evaluated on different criteria, such as achieved air interface throughput, corresponding delay metric for each type of service and fairness index.

This paper is organized as follows. In section II, we give some general insights on scheduling algorithms present in the literature. Section III gives an overview of the main challenges and considerations in the design of a MAC scheduler. Section IV describes a generalized packet scheduler model and introduces the proposed scheduling algorithm. Section V presents simulation parameters, while Section VI provides a discussion of simulation results. Concluding remarks are offered in the last Section.

II. OVERVIEW OF PACKET SCHEDULING SCHEMES

This section offers a brief overview of packet scheduling algorithms present in the literature based on [1]-[20]. Packet scheduling is an optimization problem where the scheduling function can be constructed such as to define the way the objectives of optimization are reached. In accordance with the manner the optimization is achieved, a simple classification of this function can be made: rate-adaptive, delay-adaptive and utility based schemes.

The rate-adaptive process can be instantaneous, proportional-fair and rate-adaptive with a margin. The instantaneous (maximum capacity) algorithm exploits multiuser diversity by assigning the subcarriers to those users that experience the best channel conditions. This algorithm tends to maximize the overall system capacity but fails to ensure fairness or any QoS guarantee. Additionally this algorithm does not improve cell edge performance and it can lead to starvation for users that...
have bad channel conditions. The proportional fair algorithm provides fairness among users such that the minimum data rate enforced on each user is maximized. This algorithm tends to achieve maximum capacity with a constraint of data rate fairness to all users. The rate adaptive with a margin additionally considers QoS requirements such that it tends to guarantee a minimum (threshold) data rate to each user.

As finding the optimal solution for proportional fairness or maximum capacity is a complex task, the research community has exploited alternative ways in acquiring suboptimal solutions. Examples of algorithms that provide a suboptimal solution are given in [13] and [14]. In these papers the problem of maximum capacity under the constraint of limited power and end user’s rate conditions is considered. In [13] the Hungarian method is used for solving the optimization problem, while the authors in [14] utilize the Lagrange multiplier (water filling rule). Modifications of the rate adaptive problem can be found in [5], [15] and [16]. In [5] and [15] for the proportional fair algorithm and maximum capacity algorithm respectively, differentiation between guaranteed and non-guaranteed bit rate services has been introduced through weight coefficients. In [16] the authors deal with the maximum capacity under a constraint of minimum fairness by performing subcarrier exchange. They derive a method for finding the most appropriate subcarriers to be exchanged such that the capacity loss is minimal, while the fairness index is increased.

The delay based algorithms tend to optimize the overall system delay under certain constraints. The delay based function can be instantaneous where users that would experience a minimum delay are prioritized. It can also be proportional where the maximum delay enforced on each user is minimized. The rate adaptive and delay adaptive algorithms can be considered in combination. Additionally the objective function can involve a certain degree for cell edge performance, and/or cell load balancing. Examples for delay based algorithms can be found in [4]. The algorithms in [4] are compared under different criteria, such as the utilization of multi-user diversity property, fairness and delay requirements for real-time traffic. In [17] the proportional fair algorithm is modified such that a prioritization factor is calculated based on the delay requirements for each service.

Utility theory can be applied in the design of the packet scheduler. The utility function can be derived from surveys, user’s profiles and behavior, differentiation between concurrent services executing at one user and overall network traffic. Utility theory in OFDM scheduling has been studied in [18]-[20]. In [18] the authors propose a formal model for two scheduling algorithms using utility theory. They derive the necessary conditions for reaching a maximum of the utility function. In [20] the problem of network utility maximization is considered such that it is composed of the rate control of higher layers for elastic traffic and a scheduling problem at the transport layer. The work of [19] and [20] is also an example of a cross-layer optimization problem. Their results show that the cross layer framework in a mobile wireless network outperforms the traditional layered approach.

III. PACKET SCHEDULING: CHALLENGES AND CONSIDERATIONS

The previous section gives an overview of the most studied OFDM resource allocation schemes from the research literature. These algorithms deal with the optimization of capacity, fairness among users, cell edge performance, delay requirements and QoS achievement in general. This section elaborates the main challenges in the design and implementation of MAC scheduler for the air interface in LTE and LTE-Advanced. Furthermore, multi-layer heterogeneous networks and different types of radio access architectures such as Distributed Antenna Systems (DAS) and Cloud (Cooperative or Clean) Radio Access Networks (C-RAN) are considered.

Due to the all IP-based architecture in LTE, both circuit switched services (such as Voice), and packet switched service (example file transfer) need to be scheduled with priority differentiation in order to achieve satisfying users expectations for quality. In order to enhance the QoS, two different classes have been defined in LTE: guaranteed bit rate and non-guaranteed bit rate services. For each service type, priority, packet error loss rate and packet delay budget has been defined by 3GPP.

Since the transmission time interval (TTI) is 1ms, the scheduler needs to make decisions on every 1ms. This implies the need for high processing power in order to find an optimal solution every TTI. Another challenge for the scheduler in LTE is coordination between uplink (UL) and downlink (DL) transmissions. Hybrid automatic repeat request (HARQ) retransmissions, for example, require allocation alignment between the resources for UL and DL transmission.

The control channel used to signal the assignments to the users equipment (UE) has a limited number of OFDM symbols. Therefore scheduling has to be done in such a way that the control information is limited, while the data channels are fully utilized. This can be crucial when there are many users that transmit packets with a small size and have requirements for a low packet delay (example voice). In this case the control channel becomes fully occupied while the shared data channel is underutilized. One way to avoid this problem is to use semi-persistent scheduling [21] instead of dynamic scheduling. This would mean that some of the resources are reserved to certain users, and the update of the resource assignment is done over a couple of TTI intervals. With semi-persistent scheduling the decision period is increased and control signals are reduced. The challenge is finding the right decision interval as retransmission and silence period need to be taken into account.

In today’s mobile network, energy consumption becomes one of the major considerations. There are many green efforts in reducing the power consumption at base stations and mobile nodes which seek a maximum possible minimization. Discontinuous reception (DRX) enables the mobile node to send/receive in burst and power down in between the cycles for transmission. The DRX technology imposes complexity and constraints in the design of the scheduler as it needs to take into account the DRX cycle which is active at the UE when deciding on resource assignments. Control of the transmission power at the eNB
(as well as the low power nodes of small cells) is of high importance in the minimization of inter-cell interference. This is of importance for multiple-cell when there is 1:1 frequency reuse as well as in multi-layer heterogeneous networks. The interference from macro cells to the underlying small cells can be high and can even cause coverage holes. Therefore joint scheduling and/or synchronization of the scheduler and different nodes are required.

Multiple Input Multiple Output (MIMO) antenna technique was introduced in order to increase the spectrum efficiency and increase the overall capacity. MIMO additionally impacts the complexity of the scheduler as the search space is further expanded. Carrier Aggregation and Coordinated Multipoint Transmission and Reception have been introduced by 3GPP as enhancements of the air interface in LTE-Advanced. These technologies together with beam-forming imply additional considerations in the scheduler design. CA allows aggregation of two or more LTE carriers in order to allow a larger bandwidth than 20MHz. CoMP allows simultaneous transmission and reception from multiple points which aims at improving the overall quality of the received signal. CoMP is useful in improving the cell-edge performance in which case the signal strength is very low and requires a higher amount of resource blocks from the shared data channel. Beam-forming can also be used for improving the signal quality at cell edge users, and reducing the interference towards users that are not served by the antenna that is forming the beam. The formation of the beam is determined by the beam forming vector which is created at the scheduler. Both CoMP and beam-forming require high cooperation among the schedulers and increase the amount of data and signaling required and can be acquired in a Distributed Antenna System as well as in C-RAN.

In C-RAN [22] the base station is composed of baseband Unit (BBU) and Remote Radio Heads (RRH). The C-RAN architecture is composed of three main elements: distributed RRHs, a high bandwidth low latency optical transport network and BBU pool. The RRH does not belong to a particular BBU and the BBU pool implementation is based on virtualization technologies. Coordinated scheduling is simplified with the C-RAN, such that several cells can be grouped into a cluster which can be formed in a dynamic fashion. The formation of clusters as well as the dynamic nature of assignment of baseband resources to a group of RRH need to be considered in the scheduler design.

In this section the main challenges in the design of a MAC scheduler such as multi-cell, multi-layer and advanced technologies introduced in LTE-Advanced have been covered. Furthermore different radio access architectures have been considered. In the following section we generalize the components of the scheduler in order to present the proposed scheduler.

IV. PACKET SCHEDULER DESIGN

A. General Scheduler Design

The efficient utilization of the air interface in a scheduling system such as LTE, is achieved as a combination of subcarrier allocation, adaptive modulation and coding and power allotment. Due to the limited radio resources such as available spectrum, power and equipment utilities (for ex. number of antennas), the resource allocation needs to be done such that these resources are utilized in a most optimal way while the system performances are maximized. In order to simplify the resource management, the decision process can be divided into three components [5], [9]:

Time Domain: The goal is to generate a prioritized list of UEs (their radio bearers). It can be based on different parameters such as HARQ retransmission, QoS requirements, users’ profiles, etc. The number of users that can be scheduled at each TTI is limited by the size of the Physical Downlink Control Channel (PDCCH).

Frequency domain, subcarrier allocation: Here the goal is to decide on the ordering and amount of resources (subcarriers) assigned to each user. Usually it is based on channel quality indicators, but can also include the QoS requirements, (transmit and/or receive) buffer status, etc. The size of the Physical Downlink/Uplink Share Channel (PDSCH and PUSCH) limits the total number of resource elements assigned.

Frequency domain, bit selection (or power allocation): The power assigned to each subcarrier defines the amount of bits to be transmitted. Total transmission power and cell interference govern the power assignment.

B. Proposed Scheduler Design

This section presents the proposed scheduling algorithm based on the channel quality indicator and the transmission buffer size (TX CAP). Fig. 1 illustrates this algorithm.

In the time domain the users (or radio bearers) are sorted according to the service priority. The Voice over IP (VoIP) service has the highest priority, then the Video streaming service followed by the Web browsing service and the file transfer (FTP) service at last. In the frequency domain, the resource block (RB) assignment is done according to the transmission buffer size. The users with the same service priority are sorted according to the transmission buffer size and the channel quality indicators. The users with the larger buffer size are assigned more RBs. Additionally if the data that is available for transmission does not fill more than one resource block, the user is not considered for scheduling. Regarding the power allocation, equal power is assumed for each subcarrier.
In order to confirm the performance improvement of the proposed scheduler, the results are compared with the maximum capacity algorithm where the UE are again sorted according to the channel quality indicators (referred to as MAX CAP algorithm). The MAX CAP algorithm is illustrated in Fig. 2.

Fig. 2. MAX CAP Algorithm illustration.

V. SIMULATION PARAMETERS

A. Traffic Parameters

Four different service types have been considered: VoIP, Video service, Web browsing and FTP sessions. The purpose of choosing mixed data traffic is to show the scheduler design impact on the QoS and different services. Based on the traffic parameters for different services in [24]-[26], the traffic parameters for the four services considered during simulations are shown in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoIP</td>
<td></td>
</tr>
<tr>
<td>Silence length</td>
<td>Exponential(0.65 sec)</td>
</tr>
<tr>
<td>Talk length</td>
<td>Exponential(0.352 sec)</td>
</tr>
<tr>
<td>Call Duration</td>
<td>Constant (90 sec)</td>
</tr>
<tr>
<td>Inter-call duration</td>
<td>Exponential(30 sec)</td>
</tr>
<tr>
<td>Encoding scheme</td>
<td>GMS EFR</td>
</tr>
<tr>
<td>Video</td>
<td></td>
</tr>
<tr>
<td>Inter-frame time</td>
<td>Constant(0.1 sec)</td>
</tr>
<tr>
<td>Nr. pkts per frame</td>
<td>Constant (8pkts)</td>
</tr>
<tr>
<td>Packet size (bytes)</td>
<td>Truncated Pareto (Mean= 100, Max= 250) location = 40 shape = 1.2</td>
</tr>
<tr>
<td>HTTP</td>
<td></td>
</tr>
<tr>
<td>Main Object Size</td>
<td>Truncated Lognormal min = 100, max = 2Mb (mean = 10710, std. Dev. 20532)</td>
</tr>
<tr>
<td>Embedded Object Size (bytes)</td>
<td>Truncated Lognormal min = 50, max = 2Mb (mean = 7758, std. Dev. 126168)</td>
</tr>
<tr>
<td>Nr. embedded objects per page (rand nr - location)</td>
<td>Truncated Pareto (shape=1.1, location=2) max = 53</td>
</tr>
<tr>
<td>Reading Duration</td>
<td>Exponential (30sec.)</td>
</tr>
<tr>
<td>FTP</td>
<td></td>
</tr>
<tr>
<td>File size (Mbytes)</td>
<td>Truncated Lognormal max= 5 (mean = 2, Std Dev. = 0.722)</td>
</tr>
<tr>
<td>Reading time</td>
<td>Exponential (180 sec)</td>
</tr>
</tbody>
</table>

B. Simulation Tool

OPNET Modeler Tool has been used in order to perform evaluation of the proposed scheduling scheme. Built-in LTE models were used in order to ensure confidence and correctness of the evaluation.

Current OPNET version (v16.0.A) [27] does not support adaptive modulation and coding in LTE. Therefore the model was enhanced with this feature by using the Effective Exponential SINR Mapping as described in [28].

C. Network Simulation Parameters

Simulations consider a single-cell multi-user scenario. A scenario of 16 UEs, such that there are 4 UEs for each service class, and an average distance to the eNB of 450 meters is considered.

Simulations are performed for a minimum specified bandwidth for LTE and the ITU "Pedestrian A" channel model. It is assumed that a fixed 80% of the PDCCH is to be used for downlink assignements. Table 2 shows the rest of system configuration parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Bandwidth</td>
<td>1.4 MHz</td>
</tr>
<tr>
<td>Multiplexing</td>
<td>FDD</td>
</tr>
<tr>
<td>Subcarrier Spacing</td>
<td>15kHz</td>
</tr>
<tr>
<td>Cycle Prefix</td>
<td>Normal</td>
</tr>
<tr>
<td>Number of eNB</td>
<td>1 (no Sectors)</td>
</tr>
<tr>
<td>Number of UE</td>
<td>16 users</td>
</tr>
<tr>
<td>Channel Profile</td>
<td>ITU Pedestrian A</td>
</tr>
<tr>
<td>RLC mode</td>
<td>UN-ACK Mode</td>
</tr>
<tr>
<td>CQI measurements</td>
<td>Wideband</td>
</tr>
<tr>
<td>Time window for CQI</td>
<td>1ms</td>
</tr>
<tr>
<td>Reporting interval</td>
<td>5 ms</td>
</tr>
<tr>
<td>Simulation Duration</td>
<td>1700 sec</td>
</tr>
</tbody>
</table>

VI. SIMULATION RESULTS

This section presents the results from the simulation based analysis for the scheduling scheme and configuration presented in the previous sections. The comparison metrics that are presented are the average achieved throughput and delay for each class of service. Several simulations have been executed for different seed values and the results have been averaged.

![Fig. 3. PDCCH Utilization and fairness index.](image-url)
PDCCCH and the fairness index for the two algorithms. The average utilization with the TX CAP scheme is 85% compared to 100% with the MAX CAP scheme. This is a result of the fact that UEs with larger buffers are prioritized, and therefore the assigned number of RB per UE is higher. This leads to a lowered number of UEs scheduled in one TTI as all RB are already allocated. The size of the control channel can be reduced and the available OFDM symbols can be used for the data shared channel. Thus the available space in the PDCCCH can be further utilized for downlink data transmission and therefore the overall throughput can be increased.

Fig. 3 also shows the fairness index for the two algorithms. The fairness index shows that the two schemes provide a high fairness index and the two schemes consider differentiation among different services and balance the throughput requirements for each service.

VoIP service has a certain degradation in the throughput compared to other services. All other services have an increased throughput, especially the FTP services. The reason for this is that when HTTP and FTP users are scheduled, the number of UEs scheduled for that TTI is lowered as FTP users will be allocated a larger number of RB. Here the rest of the PDCCCH not used for assignments is utilized for downlink data transmissions.

VII. CONCLUSION

This paper elaborates on the LTE downlink scheduling schemes and gives an evaluation by simulation with a built-in LTE OPNET model. The results of the proposed scheme have been compared to the MAX CQI algorithm with QoS awareness.

The results have shown that when transmission buffer size is considered by the scheduler in combination with channel conditions and QoS classification, the delay requirements for real time services can be improved. Additionally the number of colons required by the PDCCCH can be reduced and used for downlink transmission. Therefore the data rate for non-real time services can be improved.

In this phase of the project, only wideband measurements have been considered. In the future, we plan to include sub-band measurements. Additionally the time domain scheduling in case of an increased number of VoIP users can lead to a degraded performance of other service types, especially the FTP services. Therefore other parameters need to be further considered in order to satisfy the minimum QoS demands for non-real type of traffic.

REFERENCES


Fig. 4 represents the average end-to-end delay for delay sensitive services (VoIP and Video Streaming), page response time for HTTP service and file download time for FTP service. As it can be seen from the figure, the delay sensitive services have a reduced average packet end-to-end delay with the TX CAP algorithm. Due to the QoS classification the real time services will have a higher priority and they will always be considered for transmission in the first place. Therefore the delay is reduced compared to the MAX CAP algorithm. Furthermore, the confidence intervals for the video service have been reduced to a high extent. On the other hand, the HTTP service and FTP service are slightly degraded.

Fig. 5 represents the average throughput measured at different UEs and averaged for each service type. Only the


