High Spectrum Efficiency Delay Tolerant Scheduling and Resource Allocation of Diverse Traffic in LTE Networks

Tengfei Xing, Xiaoming Tao and Jianhua Lu State Key Laboratory on Microwave and Digital Communications Tsinghua National Laboratory for Information Science and Technology Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, China Email: {xingtf, taoxm, lujh}@wmc.ee.tsinghua.edu.cn

Abstract-In this paper, we propose a multiuser scheduling scheme for traffic with diverse delay constraints in the downlink of 3GPP UMTS/LTE. Traditional scheduling algorithms applied in Long Term Evolution (LTE) do not take much consideration of various delay tolerances of data packets, and most of them are on a slot-to-slot basis, which limits the ability to share spectrum and power resources among time. This would cause the network failing to deliver some packets or declining certain requests with longer delay tolerances, thereby lowering the efficiency of limited spectrum resources. Our proposed scheme schedules packets from multiple users by gathering information including service QoS, channel conditions and available resources in a preset time window, whose length equals the typical delay tolerance of multimedia data packets. The gathering of those information could be aided by channel/traffic estimations and predictions. By doing so, the algorithm achieves notably higher effective throughput than conventional schemes, thereby boosting spectrum efficiency. Simulation results show that our scheduling and resource allocation strategy can achieve 200% to 400% times of spectrum efficiency under typical system parameters.

Index Terms—delay tolerant scheduling, resource allocation, LTE, spectrum efficiency

I. INTRODUCTION

In the next decade, rapid growth of cellular communication service demands are expected to come. Applications with diverse Quality of Service (QoS), including high data rates, different real-time and interactive features, etc., will take up most of the traffic loads in cellular networks. Thereby, scheduling and allocation of radio resources is an area that deserve much attention in cellular systems such as LTE, since it is widely recognized as an element which can greatly affect the performance and spectrum efficiency of the network.

Due to the important role of scheduler in determining the overall system performance, there have been many studies on LTE scheduling in open literatures. The fundamental idea of a scheduler is to allocate each resource block to the user who can best make use of it according to some utility, and the scheduling problem is to determine the allocation of all the resource blocks to a subset of users in order to maximize some objective function, such as network throughput. [1]-[5] proposed different scheduling schemes considering heterogeneous traffic, especially their delay constraints. The delay constraints are often transformed into various instantaneous rate constraints. Moreover, [6] focused on energy efficiency when dealing with delay constrained traffic. However, to the best of the authors' knowledge, the existing studies rarely take delay tolerance of scheduling into consideration when designing algorithms since they are on a slot-to-slot basis. Thereby they are ineffective in dealing with heterogeneity/bursty of services. Our paper gives a possible solution to this problem, trying to make better use of spectrum resource to support various traffic by designing a delay-tolerant scheduling method.

We will first introduce the framework of delay tolerant scheduling, and formulate an optimization problem to represent the scheduler. To efficiently solve the problem, we will provide a two-stage heuristic algorithm consists of packet selection/subchannel allocation and power allocation with low complexity. The scheduler fully exploits the delay tolerance and jointly processes QoS and channel quality information within a certain period to get the optimal scheduling decision. Simulation results will be given to demonstrate the better performance with respect to spectrum efficiency. The advantage is that our algorithm prioritize the transmissions of the right packets, not the early packets, thereby achieves the overall optimal effect and efficiency of resource allocation.

In this paper, we first present the system model and formulate the delay tolerant sum effective rate maximization scheduling and resource allocation problem in Section II. This leads to a nonlinear hybrid-binary integer program. We then use a heuristic method to solve the problem in Section III to solve the resource allocation problem efficiently. Next we give simulation results in Section IV that compare our proposed delay tolerant scheduler with existing ones in terms of throughput. Concluding remarks will be provided in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Fig. 1 illustrates the downlink of an OFDMA single antenna (SISO) multiuser LTE network. A delay tolerant scheduling server (DTSS) is attached to the eNB, and it carries out RB scheduling and power allocation in a centralized manner through information exchange with the eNB. Let N be the



Fig. 1. System Architecture of Delay Tolerant Scheduling in the Downlink of 3GPP LTE

number of RBs and K be the number of UEs in a sector. The CSI is sent back from each UE to the eNB through a delay-free and error-free feedback channel, in order to let the DTSS do adaptive RB/power allocation and select suitable UE to serve. For the sake of analyzing spectrum efficiency, this paper does not involve specific modulation and coding. Suppose that all the users are pedestrians, so the factor of hand-over is not included here since the time scale of scheduling period is hundreds of milliseconds.

Let us consider that each UE generates data traffic in a Poisson manner. The average arrival interval is T_S TTIs. Each TTI is a basic subchannel scheduling block, and its length is configured due to the fast fading property of the propagation channel, commonly 1 ms. However, for the sake of reducing algorithm complexity, we use a larger TTI value. Here we use a non-causal hypothesis, which gathers all the service QoS, CSI and available resources (including RBs and power in all TTIs) information within the scheduling period for the scheduler. Let T_D TTIs (the unit TTI to describe time will be omitted from here on for simplicity) be the delay tolerance (or scheduling period) of the scheduler, which starts from an arbitrary packet transmission request. Thereby, we would get all the packet arrival time, packet sizes, delay tolerances of each packet, CSI, available RBs and total available power in T_D . Then we do RB scheduling and power allocation in the scheduling period T_D .

Suppose the duration of an RB is T_{RB} , and $1TTI = N_{RB} \cdot T_{RB}$. We suppose that all the RBs within a scheduling block is allocated to a single user. This is reasonable since the channel state does not vary much with in a TTI, and it also lowers the complexity of the algorithm. Then we put all the RBs on the same frequency into groups of N_{RB} , and the number of groups is T_D . Denote the number of RBs on the frequency dimension N_F . Thereby, the number of RB scheduling units within a scheduling period is $T_D \cdot N_F$. Let A_{kj} and S_{kj} be the arrival TTI index and size of the *j* th data packet of the *k* th UE (*j* = $1, \dots, J_k$) within the scheduling period. Certain distribution among a finite set of values for S_{kj} is used to model different types of traffic. Also, the delay bounds of all packets is set to be the end of the scheduling period, which is the instant at T_D . Then reshape all the A_{kj} and S_{kj} to be a column vector \hat{A} and \hat{S} with elements \hat{A}_l and \hat{S}_l $(l = 1, \dots, \sum_{k=1}^{K} J_k)$, with ascending orders of user first and then data packets. Let B_l be a binary matrix indicating which RB groups are allocated to the l th data packet,

$$B_l(m,n) = 1, \ (m = 1, \cdots, N_F, \ n = 1, \cdots, T_D) \Leftrightarrow$$

RB at (m,n) position is allocated to packet l
(1)

with $B_l(m, n)$ denoting to the (m, n) th value of B_l . Reshape B_l into a column vector b_l in a column-by-column manner, i.e. $b_l(mN_F+n) = B_l(m, n)$. In a SISO system, each RB at a certain time can only be allocated to a single UE's single packet, in order to avoid interference. Thereby the RB allocations are not overlapping, meaning that

$$b_s^T \cdot b_t = 0 \quad \forall s \neq t \tag{2}$$

which is not a necessary assumption in a multiuser MIMO system. The power allocated onto RB group m in TTI n is denoted by P_{mn} . We define an effective packet transmission as the packet is successfully transmitted with its full length and without going over its delay limit. The design goal of our scheduling algorithm is to maximize the throughput of effective packet transmissions given the delay tolerance of the scheduler and available system resources.

Let $\alpha(l)$ be the UE index of packet *l*. The achievable transmission rate of packet *l* in RB group *m* and TTI *n* can be expressed as:

$$r_{mn}^{l} = B_{l}(m,n) \left[W_{RB} \log_{2} \left(1 + \frac{P_{mn} H_{\alpha(l)}(m,n)}{N_{0} W_{RB}} \right) \right]$$
(3)

where $H_{\alpha(l)}(m, n)$ denotes the channel gain for user α_l in RB group *m* and TTI *n*, W_{RB} denotes the bandwidth of an RB and N_0 denotes the power spectrum density of noise.

Finally, the delay tolerant scheduler can be formulated as an sum rate maximization problem:

$$\max_{b_l, P_{mn}} \left\{ \sum_{l=1}^{L} C_l \right\}$$
where $C_l = \left\{ \begin{array}{c} \hat{S}_l, if\left(\sum_{m,n} r_{mn}^l\right) \cdot TTI \ge \hat{S}_l \\ 0, else \end{array} \right.$ (4)

s.t.
$$b_l(n) \in \{0, 1\} \quad \forall l, n$$
 (5)

$$b_l(mN_F + n) = 0 \quad \forall n = 1, \cdots, \hat{A}_l - 1, m$$
 (6)

$$b_s^T \cdot b_t = 0 \quad \forall s \neq t \tag{7}$$

$$\sum_{m,n} P_{mn} \le P_{\max} \cdot T_D \tag{8}$$

where P_{max} is the total transmit power limit of the eNB. (6) means that no RBs is allowed to be allocated to a packet before its arrival. Clearly this is a highly non-linear hybrid-binary integer program, for which no efficient solution exists.

As an NP-hard problem, exhaustive searching algorithm with high complexity can solve the problem. In the next section, a low-complexity allocation scheme based on heuristic methods is proposed.

It is worth mentioning that the heterogeneity and bursty of services is embodied through the above modeling of packet arrival and QoS (size and delay tolerance). Specifically, with independent arrivals, the possibility of bursty packet arrivals is increased with the number of users. Also with different arrival and same deadline, various levels of delay tolerance is presented for each packet. Moreover, the difference in packet sizes represents heterogeneity of services.

III. MAXIMUM EFFECTIVE RATE SCHEDULER AND RESOURCE Allocation

We propose a heuristic scheme that achieves sub-optimal solution to the proposed delay tolerant scheduling problem. The scheme is divided into iterations of the following two stages. In the first stage, the algorithm selects a packet to serve, and the RBs are assigned to this packet under the assumption that the eNB's total transmission power left (initially P_{max}) is equally distributed among every RB left in both frequency and time dimensions, i.e. $P_{mn} = \frac{P_{max} \cdot T_D}{N_F \cdot T_D}$ initially. This stage only implements RB selection and allocation. In the second stage, power are allocated to the RBs assigned in the first step in order to save as much transmission power as possible and potentially lowers the number of RBs required. The allocated RBs and power are excluded from the resource left for the next iteration. The exit condition of the iterations is that none of the packets can be served with the RBs and power left. The step-by-step iterations to determine the packets to serve and the separation of subchannel allocation and power allocation enable a suboptimal algorithm; however, it makes the complexity significantly lower than the exhaustive search.

A. Throughput-oriented Packet Selection and RB Allocation

Due to the target of maximizing the sum rate of effective packet transmissions and our assumption that each packet have different arrival time and the same deadline, it is reasonable to use the following criteria to select which packets should be scheduled with a higher priority:

- 1) For packets with the same sizes, the one with a longer delay tolerance should be scheduled first.
- 2) For packets with the same delay tolerances, the one with a smaller size should be scheduled first.
- For any packets, the one with a smaller average rate requirement (its size divided by its delay tolerance) should be scheduled first.

The main idea behind these criteria is to consume as few resources (including RBs and power) per unit of data as possible, in order to serve more packets with the same total resources. Firstly, due to the principle of diversity, the possibility to have an RB with good channel quality within a longer period of time is higher. Therefore, the packet with a longer delay tolerance may consume less resources, making it a favorable choice. Also, within the same period, the packet

	TAB	LE I		
PACKET SELECTION	RB Allocatio	N OF DELAY	TOLERANT	SCHEDULER

0. Preliminary Process (only execute once at the beginning of a scheduling period):

Calculate the average rate requirements for every packet, $R_l = \frac{\hat{S}_l}{T_D - (\hat{A}_l - 1)}$.

Set $b_l(mN_F + n) = 0$ $\forall m, n, l. P_0 = P_{\max} \cdot T_D$.

Create an empty queues Q_1 , storing the indexes of packets selected to serve.

1. Initialization:

a) Let Φ be the set of the index of packets left un-selected, Θ_{RB} be the set of the index of RBs left un-allocated, and P_0 the power left un-allocated.

b) Do a sorting of R_l ($l \in \Phi$) in ascending order, and store the index of the results in a queue Q.

c) Set $P_{mn} = \frac{P_0}{\|\Theta_{RB}\|}$. $\hat{C}_l = \hat{S}_l \ (l \in \Phi)$. 2. Packet Selection and RB allocation:

a) Select a packet l_1 with the smallest R_l from Q. Exclude l_1 from Q. Create an empty queue Q_S to store the indexes of RBs that will be allocated to packet l_1 .

b) Do a sorting of $H_{\alpha(l_1)}(m, n)$ $(u = mN_F + n \in \Theta_{RB}, n \ge \hat{A}_{l_1})$ in descending order, and store the index of the results u in a queue Q_H . c) Select an RB whose index is $u_0 = m_0N_F + n_0$ with the largest $H_{\alpha(l_1)}(m, n)$ value. This is equivalent to finding $u = \arg \max_{u \in \Theta_{RB}} r_{mn}^{l_1}$ $(n \ge \hat{A}_{l_1})$. Exclude u_0 from Θ_{RB} . Add u_0 to Q_S .

d) Allocate RB u_0 to packet l_1 , and calculate the unserved data size, $\hat{C}_{l_1} = \hat{C}_{l_1} - r_{m_0 n_0}^{l_1} \cdot TTI$.

e) If $\hat{C}_{l_1} \leq 0$, exclude packet l_1 from Φ and add it to Q_1 , exclude all elements in Q_S from Θ_{RB} , mark the corresponding elements of b_{l_1} to 1, and finish allocation. Go to power allocation.

f) If $\hat{C}_{l_1} > 0$ and Q_H is not empty, go to step c).

g) If $\hat{C}_{l_1} > 0$ and Q_H is empty, which means that this packet cannot be served with the resource left, exclude packet l_1 from Φ , and go to step a).

with a larger size generally requires more RBs or power. However, the possibility to have more RBs with better channel quality is less for a fixed user. This also leads to a smaller efficiency of resource utilization. Hence, the packet with a smaller size is favorable among the ones with the same delay tolerance. Last but not least, the third criterion is a combination of the first two.

We propose a throughput-oriented packet selection and RB allocation scheme in Table I, based on the criteria above. First, we calculate the average rate requirements of all the packets. Then we select the packet with the lowest rate requirements, l_1 , as a candidate to be scheduled in this scheduling period. Among all the RBs that are valid to be allocated to l_1 (meaning that they are not allocated to other packets, and their time index has to be larger or the same as \hat{A}_{l_1}), the ones with better channel quality regarding UE $\alpha(l_1)$ are allocated to l_1 one by one. Once the total data size provided by the allocated RBs exceed \hat{C}_{l_1} , l_1 is successfully scheduled, and the algorithm goes to the next stage of power allocation. The elements of the indicator vector b_{l_1} are also marked to 1 correspondingly, and the allocated RBs from the set of available RBs are excluded for further scheduling. If all the RBs are allocated to l_1 and the total data size still cannot exceed \hat{C}_{l_1} , then l_1 cannot be served. This will lead to another packet selection with the next lowest rate requirements.

B. Resource Efficient-oriented Power Allocation

After packet l_1 is selected in the first stage, we can further optimize the consumption of transmit power and number of RBs. In the RB allocation process, equal power allocation is assumed. Now we use an inverse-waterfilling (IWF) method to minimizing power subject to a rate constraint (meaning that the packet needs to be delivered at its full size), which is a dual problem of conventional waterfilling [6].

Suppose the indexes in Q_S is $u_i = m_i N_F + n_i$ $(i = 1, \dots, d)$, where *d* is the number of allocated RBs. The IWF method can be formulated as an convex optimization problem:

$$\min_{\substack{r_{u_1}^{l_1}, \dots, r_{u_d}^{l_1} \\ where P_{u_i} = N_0 W_{RB} \cdot \frac{2^{\binom{l_1}{u_i} W_{RB}}}{H_{\alpha(l_1)}(m_i, n_i)}}$$
(9)

s.t.
$$\left(\sum_{i=1}^{d} r_{u_i}^{l_1}\right) \cdot TTI = \hat{S}_{l_1}$$
 (10)

$$r_{u_i}^{l_1} \ge 0 \quad \forall i = 1, \cdots, d \tag{11}$$

where $r_{u_i}^{l_1} = r_{m_i n_i}^{l_1}$ is the achievable rate on RB *u* and P_{u_i} is the power allocated on RB *u*. After the optimal r_{u_i} is solved, P_{u_i} can be simply calculated. This problem can be easily solved using the Lagrangian method:

$$r_{u_i}^{l_1} = W_{RB} \cdot \left\langle \log_2 \left(\frac{H_{\alpha(l_1)}(m_i, n_i)}{H_{th}} \right) \right\rangle_0^{\infty}$$
(12)

where H_{th} is the water level and the solution to

$$\sum_{i=1}^{d} \left\langle \log_2 \left(\frac{H_{\alpha(l_1)}(m_i, n_i)}{H_{th}} \right) \right\rangle_0^{\infty} = \frac{\hat{S}_{l_1}}{W_{RB} \cdot TTI}$$
(13)

We can see that an RB is utilized only if a positive energy is scheduled on it, i.e., $r_{u_i}^{l_1} \ge 0$ or equivalently $H_{\alpha(l_1)}(m_i, n_i) > H_{th}$. Then we calculate the total energy consumed, and subtract it from P_0 , the total energy left for the other un-scheduled packets. For all u_i that $r_{u_i} \le 0$, meaning that these RBs are not needed for transmitting packet l_1 and thereby can be reallocated to other packets, add them back to Θ_{RB} . It is obvious that this power allocation step not only optimizes the power consumption for the packet transmission, but also potentially saves some spectrum resource blocks for further scheduling of more packets.

C. Brief Summary and Performance Metric of Delay Tolerant Scheduling

Combing the above two stages, we summarize the solution to the delay tolerant scheduling problem in Table II. After the scheduling is done for as many packets as possible, the maximized sum data size served is $\sum_{l \in Q_1} \hat{S}_{l_1}$. It is easy to see that the complexity of the above method is much lower than exhaustive search.

TABLE II Solution to Delay Tolerant Scheduling

While Φ is not empty, where Φ is the set of the index of packets left un-selected:

1) Follow the process in Table I, select a packet l_1 to serve, and allocates RBs in Q_S to l_1 .

2) Do power allocation among RBs in Q_S by solving the convex optimization problem in (9); go back to step 1).

IV. SIMULATION RESULTS AND PERFORMANCE ANALYSIS

In order to compare the proposed scheduling scheme with existing ones, we use a standard compliant LTE system level simulator [7] that is publicly available, based on which necessary modules are further developed.

The simulation parameters are summarized in Table III. We use a microscale fading channel model with channel gain constant during a 20ms TTI and independent among all TTIs. Due to the delay-free and error-free CSI feedback assumption, the DTSS gets the CSI information before transmissions. The packets have sizes and delay properties as typical multimedia traffic (including audio, video streaming, etc.), and the scheduling period is set to be 10 or 20 TTIs. The delay tolerance of packets span from 1 TTI (20ms) to 20 TTIs (400ms), which corresponds to the QoS demands of most typical services. All the packets can be viewed as non-realtime (NRT) service requests, which have average rate constraints within their own valid periods. Thereby, the heterogeneity of services are embodied through the difference of packet sizes and delay tolerances.

First, we consider the case with a fixed number of five users and compare the performance of different schedulers. Fig. 2 shows the sum throughput of all the users versus their average signal-to-noise ratio (SNR) when applying different schedulers. Our proposed DTS with scheduling period (delay tolerance) of 10 and 20 TTIs are shown with three typical scheduling algorithms, including Round Robin (RR), MaxMin (MM) and Proportional Fair (PF) scheduling. A best effort (BE) upper bound, which is the capacity limit of five users with all full buffer Best Effort (BE) traffic without considering packet arrivals and delay constraints, is simulated and shown for comparison.

It is shown that DTS can achieve up to 2 to 4 times of throughput than existing schedulers at normally used common SNR region (0 to 10 dB). The throughput gain is larger for lower SNR values and diminishes gradually as the SNR increases. This gain is achieved since under low SNR, existing schedulers may not successfully serve packets at its full size or accept certain requests, due to both the tighter constraints of frequency and power resources within a shorter period and lack of utilization of QoS and CSI to make proper scheduling decisions. On the other hand, DTS is able to jointly utilize the resources within a longer period and optimize the scheduling decisions, by collecting the QoS and CSI information within the delay tolerance, thereby achieving more successful transmissions of packets and higher throughput. Other than

TABLE III Simulation Parameters

Parameter	Value		
System Bandwidth	1.4MHz		
Number of subcarriers	72		
Number of RBs N	12		
Number of user K	5 per sector		
Packet Arrival Interval (Poisson)	1 TTI		
Channel Model	ITU-T PedB [10]		
eNodeB Settings	distance 500m, tx power 20W		
Large-Scale Fading	3GPP TS25.814		
Shadowing	R9-Claussen, 10dB variance		



Fig. 2. Sum Throughput versus SNR with different schedulers

a first-in-first-out manner, DTS prioritize the transmissions of the right packets, not simply the early packets. With a higher throughput, the spectrum efficiency is also times higher. Moreover, by increasing delay tolerance, the throughput of DTS is further boosted and thereby the gap between it and the BE upper bound decreases.

Then, we show the benefits of multiuser diversity in the scheduling process. Fig. 3 shows the sum throughput of different schedulers versus different number of users under a fixed average SNR of 10dB. It is shown that as the number of users increases in the same cell, the throughput gradually increases. This is due to the effect of multiuser diversity. Similar application of this concept can be found in [8]. Also, we can observe that the increasing rate of throughput of DTS is larger than the ones of other schedulers. The reason is that our algorithm can better collaborate multiuser diversity with frequency and time diversity of channel fading.

V. CONCLUSIONS

In this paper, we proposed a delay tolerant scheduling scheme for real-time traffic of multiple users in an OFDMAbased LTE downlink network. We introduced the framework of delay tolerant scheduling, and formulate the target of maximizing spectrum utilization in supporting heterogenous traffic as an optimization problem. To efficiently solve the



Fig. 3. Sum Throughput versus number of UEs with different schedulers

problem, a two-stage heuristic algorithm consists of packet selection/subchannel allocation and power allocation with fair complexity is given. This algorithm embodies the essence of DTS, which is to utilize all the spectrum and power resources onto the most proper packets. This is done by exploiting the delay tolerance of the scheduler and jointly processing QoS and channel state information within a longer period to get the optimal scheduling decision. Simulation results show that our scheduler outperforms existing algorithms with respect to system throughput and spectrum efficiency.

ACKNOWLEDGMENT

This research work was supported by National Basic Research Program of China (No. 2007CB310601) and NSFC (No. 60972021).

References

- M. Tao, Y. Liang, and F. Zhang, Resource Allocation for Delay Differentiated Traffic in Multiuser OFDM Systems, ;IEEE Transactions on Wireless Coomunications, ;vol. 7, no. 6, pp. 2190 - 2201, Jun. 2008.
- [2] R. Zhang, Optimal Dynamic Resource Allocation for Multi-Antenna Broadcasting With Heterogeneous Delay-Constrained Traffic, IEEE Jou. on Sel in Sig. Proc, ;vol. 2, no. 2, pp. 243 - 255, Apr. 2008.
- [3] R. Zhang, Optimized Multi-Antenna Broadcasting for Heterogeneous Delay-Constrained Traffic, ;IEEE ICC 2008, May. 2008.
- [4] T. Girici, and A. Epl, Practical Resource Allocation Algorithms for QoS in OFDMA-based Wireless Systems, ;IEEE CCNC 2008, Jan. 2008.
- [5] T. Girici, and A. Epl, A channel-aware queue-aware resource allocation algorithm for OFDMA system with heterogenous delay requirements, ;ICPCSPA 2010, Jul. 2010.
- [6] J. Lee, and N. Jindal, Energy-efficient Scheduling of Delay Constrained Traffic over Fading Channels; IEEE Transactions on Wireless Coomunications, ;vol. 8, no. 4, pp. 1866-1875, Apr. 2009.
- [7] [Online]. Long Term Evolution (LTE) System Level simulator. Available: http://www.nt.tuwien.ac.at/ltesimulator/
- [8] B. Niu, O. Simeone, O. Somekh and A. M. Haimovich, Ergodic and Outage Sum-Rate of Fading Broadcast Channels with 1-Bit Feedback, ;IEEE Transactions on Vehicular Technology, ;vol. 59, no. 3, pp. 1282 -1293, Mar. 2010.