QoS Guaranteed Resource Block Allocation Algorithm in LTE Downlink

Liqun Zhao*, Yang Qin*[1], Maode Ma[+], Xiaoxiong Zhong*, Li Li*
*Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[+]Nanyang Technological University, Singapore
*[1]Corresponding author, email: yqinsg@gmail.com

***Abstract*—Long term evolution (LTE) is the next generation wireless communication network which is an all-IP based system. Nowadays, more and more kinds of services are being transmitted on wireless communication network with different requirements. So, to satisfy the quality of service (QoS) of multi-service requirements is one of the key challenges that need to be dealt in the LTE system. In this paper, we propose a cross layer solution called token bucket scheduler (TBS) for real time service that allocates resource block (RB) for different services in order to meet their QoS requirements. TBS utilizes instantaneous downlink channel signal to interference plus noise ratio (SINR) and service QoS information when determining whether a RB allocates to a real time service. Through extensive simulations, the results show that TBS algorithm greatly improves throughput, delay and packet loss ratio compared with the maximum-largest weighted delay first (M-LWDF) scheduling algorithm and exponential/proportional fair (EXP/PF) scheduling algorithm in real time service.**

Keywords-LTE; QoS

I. INTRODUCTION

With the rapid development of mobile communication technologies, the Third Generation Partnership Project (3GPP) has introduced LTE specifications as the next cellular networks to satisfy the ever-growing demand for mobile communication requirements. In 3GPP release 8, LTE provides high peak data rates of 100Mb/s on the downlink and 50Mb/s on the uplink for 20MHZ bandwidth. In order to improve spectral efficiency, LTE adopted Orthogonal Frequency Division Multiple Access (OFDMA) as the downlink access technology and adopted Single-Carrier Frequency Division Multiple Access (SC-FDMA) for uplinks [1].

With the growing usage of the mobile equipment, it is clear that more and more services will be transmitted through wireless networks, such as video application with different resolution, conversational voice, real time game and FTP. These growing applications bring some problems for wireless networks since those services generally have different QoS requirements. For example, besides the huge bandwidth requirement, the real-time video needs some other QoS requirements: such as delay, jitter and data rate. If a video packet does not arrive within delay constraint, it will be considered lost. Video packet loss incurs distortion. Usually, FTP service has no strictly QoS constraints.

LTE scheduling is one of the key parts that affects system throughput and user experience. There are many publications that widely address this topic. OFDM (Orthogonal Frequency Division Multiplexing) technology divides broad spectrum into multiple narrow bands and transmit data on these narrow channels. The main problem of RB allocation is to meet different service QoS requirements and maximize system throughput. To solve this problem, three basic algorithms have been proposed, Max C/I, Proportional Fair [2] (PF) and Round Robin (RR). In spite of Max C/I algorithm makes the highest system throughput, RR algorithm can guarantee service fairness, PF algorithm makes a trade-off between throughput and fairness. But these algorithms do not take QoS requirements for real time service into consideration. To guarantee the QoS requirements for real time service, several real time scheduling algorithms have been proposed. In [3], the author presented an algorithm called maximum-largest weighted delay first (M-LWDF) which supports different services and takes service delay requirements into account. The exponential/proportional fair (EXP/PF) scheduling algorithm [4, 5] is designed to increase the priority of real-time service which the head of line delay is close to delay threshold. M-LWDF and EXP/PF are more popular ways in LTE downlink schedule approach. Above mentioned algorithms have three drawbacks. First, If a service packet get a small allocation priority whose waiting time is close to deadline when channel quality decrease suddenly, packet violation will occur due to the unpredictable channel quality of wireless links. Second, all above discussed approaches can not guarantee real time flow data rate. These algorithms based PF adopt multi-user diversity to improve system throughput. For a certain service, multi-user diversity allocates different number of RBs according user's channel condition in each dispatching cycle. So, multi-user diversity is not suitable for rate sensitive service since data rate is varying according their channel condition. Third, all of above mentioned real time scheduling algorithm only differentiate service into real time and non-real time service, experiment shows that these algorithms cannot provide QoS strictly.

In this paper, we propose a new approach which allocates appropriate number of RBs to real time service to guarantee QoS strictly. Our goal is to provide strictly QoS for real time service, maximizing system throughput and controlling traffic flow data rate. The proposed algorithm not only differentiates service to real time service and non-real time service, but also differentiates all real time service by using different token generation rates. We use token bucket depth to control the

978-1-4673-2699-5/12/$31.00 © 2012 IEEE

amount of traffic transmitted in one real time traffic flow in a cycle. In each dispatching, the total number of RBs that the system allocates to a real time flow is determined by traffic flow QoS information. We present numerical results to illustrate the performance under a variety of scenarios and show the benefits of the proposed algorithm compared with M-LWDF and EXP/PF, especially in real time video service.

The remainder of this paper is organized as follows. Section II introduces the system background. In Section III, Proposed algorithm is described in detail, which includes flow classifier, buffer manager and flow scheduler. We present comprehensive simulation results about the system performance in Section IV. Section V concludes the paper.

II. SYSTEM BACKGROUND

In 3GPP release 8, we can see the evolved packet system (EPS). The EPS consists of evolved packet core (EPC) and evolved universal terrestrial radio access (E-UTRAN). It is also referred to as LTE [6]. The EPC consists of mobile management entity (MME)/serving gateway (S-GW). The only node in LTE is the evolved node B (eNB), so called base station. The eNBs are connected with each other through X2 interface distributed the network coverage area, and with EPC through S1 interface. An eNB may be served by more than one MME.

LTE has improved and enhanced 3G air interface. In LTE air interface, user plane protocol stack is in charge for LTE downlink data transmission. Fig. 1 shows the LTE user plane architecture. It consists of packet data convergence protocol (PDCP), radio link control (RLC), medium access control (MAC) layer and physical layer. The eNB receives the IP packets from packet data network gateway (PDN-GW) and delivers them to PDCP with or without robust header compression (ROHC). Then, PDCP protocol data unit delivers to RLC. The RLC protocol is responsible for segmenting and concatenation. It also performs error correction with the automatic repeat request (ARQ) method. Three modes used in RLC for different service: unacknowledged mode, acknowledge mode, and transparent mode. (e.g. Unacknowledged mode used to reduce the delay for video telephone). After that, MAC layer allocates wireless resource for physical layer to transmit. At the physical layer, data comes from MAC layer are turbo coded and mapped using one of the following schemes: quadrature phase shift keying (QPSK), 16-QAM, or 64-QAM, the resulting symbols are mapped in the subcarriers.

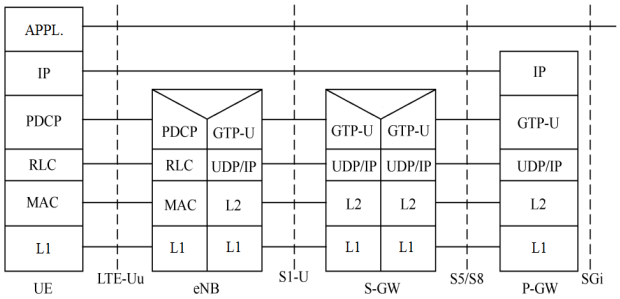

Figure 1. LTE user plane

EPS adopts the concept of bearer as QoS level of granularity (i.e., different QoS treatment operates on per-bearer level.) Bearer is a packet flow established between PDN-GW and UE. The service can be differentiated into separated data flows and then mapped to a bearer. Each bearer associates with a scalar value called QoS class identifier (QCI) configured by operator. Bearers which have same QCI receive common QoS treatment [7]. (e.g., scheduler policy, queue management policy, admission threshold and RLC configuration). The QCI standard is given in Fig. 2. From the table, we can see that each QCI belongs to either guaranteed bit rate (GBR) or non-guaranteed bit rate (Non-GBR) and associates with priority, packet delay bucket and packet error loss rate.

QCI	Resource type	Priority	Packet delay budget (ms)	Packet error loss rate	Example services
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	5	300	10^{-6}	Non-conversational video (buffered streaming)
4	GBR	3	50	10^{-3}	Real time gaming
5	Non-GBR	1	100	10^{-6}	IMS signalling
6	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
7	Non-GBR	6	300	10^{-6}	Video (buffered streaming)
8	Non-GBR	8	300	10^{-6}	TCP-based (e.g. WWW, e-mail) chat, FTP, p2p file sharing, progressive video, etc.
9	Non-GBR	9	300	10^{-6}	

Figure 2. QCI standard

Besides QCI, some other QoS attributes associated with the LTE bearer as follows:

- Allocation and retention priority (ARP): ARP differentiates the control plane treatment related to call admission control or overload control.
- Maximum bit rate (MBR):the upper bound of the bearer, only valid for the GBR bearers.
- Aggregate MBR (AMBR):used to limit total bit rate consumed by a single subscriber.

Bearer provides differential treatment for service with different QoS requirements. The QoS parameters of bearer associated with control functions (e.g. eNB assumes packet loss by admission control function). One bearer exists with QoS parameters and IP address. A UE may have multiple IP addresses if the subscriber connects to more than one data networks. In this case, UE can have more than one separated bearers.

In LTE, radio resources are allocated both in time/frequency domain. In time domain, they are distributed by transmission time interval (TTI), each lasts 1 ms. One TTI consists of two time slots of 0.5 ms, corresponding to 7 OFDM symbols in the default configuration with short cyclic prefix. Ten consecutive TTIs form a LTE frame. In the frequency domain, the whole bandwidth is divided into 180kHz subwidth. RB is the minimal allocable unit which spanning over one time slot in time domain and over 180kHz bandwidth [8]. The eNB makes the scheduling decision per TTI based on the channel

quality indicator (CQI) feedback from user equipment (UE).

III. CROSS LAYER ALLOCATION ALGORITHM-TBS

In this section, we introduce a new approach of resource management in LTE downlink to satisfy real time service requirements, especially for GBR bearers. Analyzing the drawbacks of EXP/PF and M-LWDF algorithm as section I mentioned, we can find that which service will be scheduled and how many RBs will be allocated to a service is completely determined by the head of line packet delay and channel condition rather than by service QoS information. However, head of line delay and channel quality cannot completely represents service QoS requirements. Herein we use token bucket mechanism to maintain a certain amount of data in every dispatching cycle. Our goal is to provide appropriate RBs for real time service which is determined by QoS information of traffic flow. The proposed scheduler is shown in Fig. 3, which consists of three parts: flow classifier, buffer manager and flow scheduler. Flow classifier can divide flows into two categories: real time service and non-real time service. When RLC layer data arrives, buffer manager stores data and controls the data rate. Flow scheduler allocates RBs to flows. In order to simplify the scheduling, two steps have been defined: TBS scheduler and non-real time scheduler. TBS scheduler is used to allocate appropriate number RBs to real time flows while non-real time scheduler is used to allocate the rest of RBs with PF algorithm.

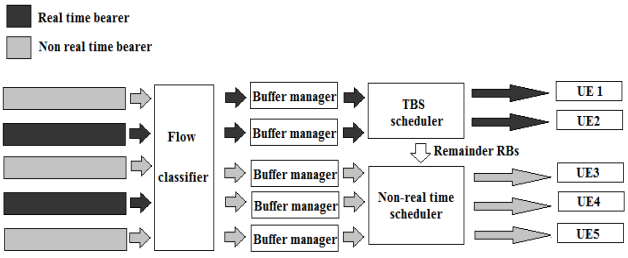

Figure 3. The proposed algorithm

A. *TBS scheduling*

We exploit token bucket depth to control the amount of traffic transmitted in one real time traffic flow in a cycle and use token generation rate to differentiate real time flows. For real time flows, buffer manager not only stores data at MAC layer, but also uses token bucket mechanism to control the amount of data transmitted for each real time flow in each cycle. Buffer manager must utilize service application layer information then calculates how many data should be transmitted in one dispatching cycle. A real time flow i can be characterized by the following parameters, as Fig 4 shows.

- R_i: token generation rate, each P-byte packet consumes P tokens,
- L_i: token bucket depth of flow i, flow i should transmit L_i byte, in the beginning of each TTI, update L_i by plus R_i, at the end of TTI, L_i reduces by actual amount of data transmitted.

R_i can represent different kinds of flow QoS requirements, such as data rate and delay. Usually, different kinds of flow should set different token generation rates. For example, video flow needs higher data rate than VoIP flow, so video flow token generation rate is higher than that of VoIP flow.

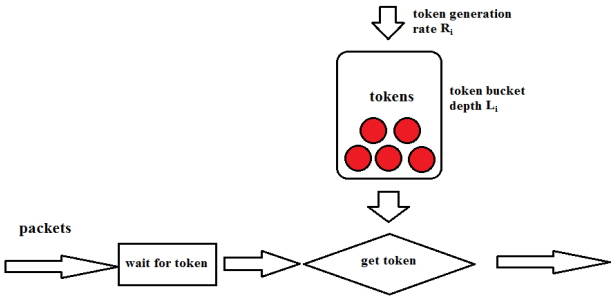

Figure 4. Token bucket mechanism

In order to utilize channel resource efficiently, We allocate RBs to real time flows according to their channel quality. So, metric $m_{i,j}$ for flow i, in RB_j is computed as follows:

$$m_{i,j} = SINR \tag{1}$$

The TBS scheduler starts with the real time flows selected by flow classifier. RB allocation is done iteratively, each iteration allocates one RB to a flow with highest $m_{i,j}$, denoted as $m_{h,k}$ (if there are more than one service have the same metric in a RB, we will start from the service which arrives first):

$$m_{h,k} = \arg\max\{m_{i,j}\} \tag{2}$$

The next step is allocating appropriate number of RBs to real time flows. We check whether the amount of data meets the requirements when RB_k is allocated to flow h. we assume RBa_1, RBa_2, ...RBa_n have been allocated to flow h. Let $d_{h,j}$ be the maximum amount of data for flow h in RB_j, The amount of data can be calculated through SINR by using the exponential effective SINR mapping (EESM) [9]. So the total amount of data D_h can be calculated as:

$$D_h = \sum_{t=a_1}^{a_n} d_{h,t} + d_{h,k} \tag{3}$$

After that, check whether the service is allocated enough RBs to meet QoS requirements according to (4).

$$\begin{cases} \text{Allocate } RB_k \text{ to user h and remove user h} & \text{if } D_h \geq L_h \\ \text{Allocate } RB_k \text{ to user h} & \text{if } D_h < L_h \end{cases} \tag{4}$$

Then, an iteration is over. TBS scheduler terminates when no real time bearer or no RB left. After servicing a real flow, the number of tokens of the corresponding token bucket is reduced by the actual amount of data transmitted. An advantage of using token bucket is to compensate a certain amount of data when channel quality decreases suddenly. For example, when channel quality decreases, the actual amount of data transmitted is less than it required. In the end of TTI, we update L_i by actual amount of data. So in the next TTI transmission, system will compensate the insufficient amount of data in the previous TTI.

B. *Non-real time scheduling*

Once the TBS is over, non-real time scheduler using PF algorithm allocates the rest RBs to non-real flows to achieve a trade-off between fairness and system throughput. In every TTI, MAC layer computes the instantaneous achievable data rate $r_{i,j}$ for flow i, in sub channel j. PF algorithm allocates RBs to a flow with the $m_{i,j}$ for bearer i in RB j, Which is computed as follows:

$$m_{i,j} = \frac{r_{i,j}}{\overline{r_i}} \quad (5)$$

where $\overline{r_i}$ represents average data rate. In each TTI, the $\overline{r_i}$ is given by

$$\overline{r_i}(t) = (1-\alpha) \bullet \overline{r_i}(t-1) + \alpha \bullet r_i(t-1) \quad (6)$$

where t and t-1 are the current sub frame and the previous sub frame. PF assigns the RB with the highest $m_{i,j}$ to corresponding flow and then removes the RB. Until no RB left, the algorithm is over. The computational complexity is $O(N_{RB} \times N)$, where N_{RB} is the total number of RBs, N is the number of active services.

IV. SIMULATION RESULTS AND ANALYSIS

In this section, we compare the performance of the M-LWDF and EXP/PF with TBS algorithm we proposed. Here, we assume that one user has only one service. We have developed a scenario (with 40% of users using video flows, 40% of users using VoIP flows and the rest of users using best effort flows) to perform our experiments. Simulation parameters are shown in table I.

TABLE I. LTE DOWNLINK SIMULATION PARAMETERS

Parameters	Values
Number of eNBs	1 eNB with radius 1 km
Spectrum	10 MHz
Mobility model	Random walk
Number of users	10,20,30,40,50,60
Frame structure	FDD
Slot duration	0.5 ms
TTI	1 ms
Max delay	0.08 s
UE velocity	3 km/h
Simulation duration	150 s
Flow duration	120 s

The video service is a trace based application which original sequence is 25 frames per second. Video sequences have been compressed using H.264 standard with average code rate of 242 kbps. The VoIP application generates G.729 voice flows. In particular, the voice flow has been modeled with an ON/OFF Markov chain, where the ON period is exponentially distributed with mean value of 3 s (source data rate is 8 kb/s), and the OFF period has a truncated exponential probability density function with an upper limit of 6.9 s and average value of 3 s (source data rate is 0 kb/s).

For video flows, Figure 5-7 illustrate the performance of each video user under M-LWDF, EXP/PF and the TBS algorithm. The aggregate throughput increases by 43% for 60 users in TBS algorithm. Comparing with other two algorithms, delay and packet loss ratio both decrease in TBS algorithm. The reason is that TBS can guarantee QoS for real time service in advance. Therefore, the proposed algorithm shows a good performance for video flows.

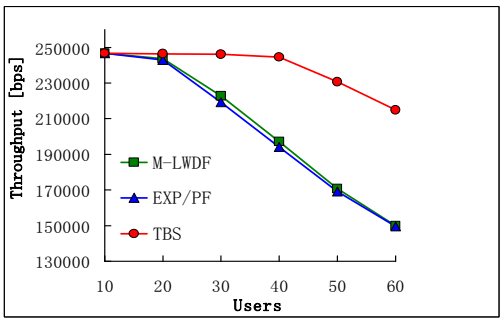

Figure 5. Average video throughput per user

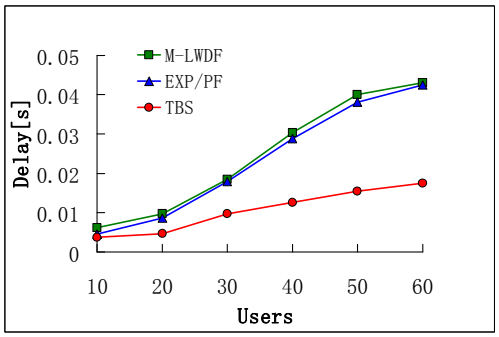

Figure 6. Average video delay

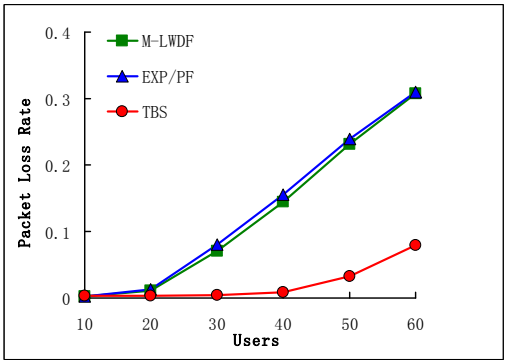

Figure 7. Video packet loss ratio

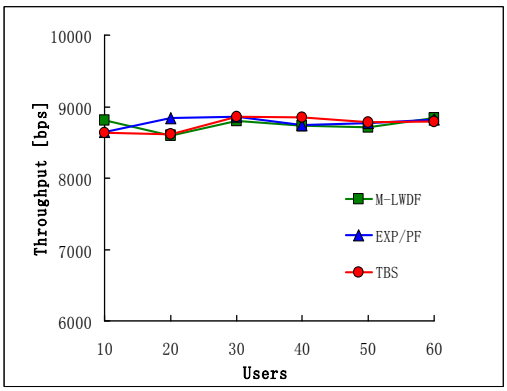

Figure 8. Average VoIP throughput per user

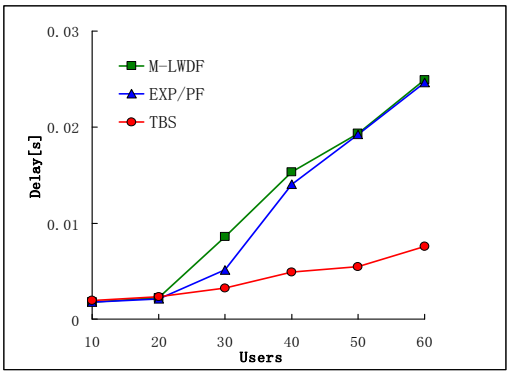

Figure 9. Average VoIP delay

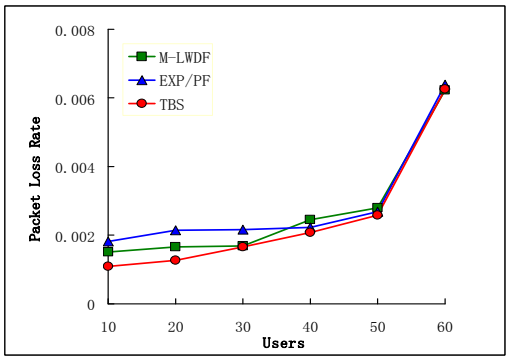

Figure 10. VoIP packet loss ratio

For VoIP flows, there is no significant improvement when TBS algorithm is used. Figure 8-10 show the performance of VoIP flows. Three algorithms present a stable throughput and a nearly packet loss ratio when the number of users changes from 10 to 60. TBS algorithm has lower delay since VoIP RB allocation is before best effort flow allocation.

For best effort flows, we only compare average throughput since there are no QoS guarantees for these flows. Fig. 11 shows the comparison of the throughput per user of three algorithms. We can see that the TBS algorithm has a lower throughput than other two algorithms. The reason is the TBS algorithm can guarantee the amount of data for real time flows TTI by TTI. More over, the TBS algorithm always allocates RBs to real time flow with a good channel condition, and then allocates the rest RBs to best effort flow.

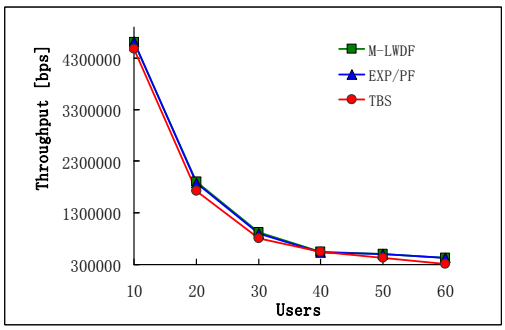

Figure 11. Best effort throughput

V.CONCLUSION

In this paper, we propose a new algorithm TBS in LTE downlink with QoS provisioning. For real time service, TBS algorithm can differentiate real time service by using token bucket, and determine the number of RBs that the system allocates to a flow with QoS information. Simulation results show that performance measures in terms of real time service throughput, packet loss ratio and delay are better than those of a system adopted M-LWDF or EXP/PF algorithm under the scenarios with videos flows, VoIP flows and best effort flows. Future work will consider the uplink scheduling and system call admission control.

REFERENCES

[1] 3GPP,Tech.Specif. Group Radio Access Network Requirements for Evolved UTRAN and evolved UTRAN, 3GPP TS 22.278.

[2] R.kwan, C.leung and J,Zhang, "Proportional fair multiuser scheduling in LTE," IEEE Signal Processing letters, vol.16, no.6, pp.461-464, Jun 2009.

[3] M.Andrews, K.kumaran , K.ramanan, A.stolyar, P.whiting and R.Vijayakumar, "Providing Quality of Service over a shared wireless link," IEEE Commun. Mag., vol. 39, pp. 150-154, Feb. 2001.

[4] J.-H. Rhee, J.M. Holtzman and D.K. Kim, " Scheduling of Real/Non-real Time Service: Adaptive EXP/PF Algorithm," 57th IEEE semiannual Vehicular Technology conference, vol. 1, pp.462-466, 2003.

[5] J.-H. Rhee, J.M. Holtzman and D.K. Kim, "Performance Analysis of of the Adaptive EXP/PF channel Scheduler in an AMC/TDM system, "IEEE Communications letters, vol.8 ,pp. 4978-4980, Aug. 2004.

[6] Ronit Nossenson, "Long-Term Evolution Network Architecture," IEEE International conference, PP. 1-4. 2009

[7] Mehdi Alasti and Behnam Neekzad, Jie Hua and Rath Vannitham, "Quality of Service in WIMAX and LTE Networks," IEEE Communications magazine, pp.104-111, May 2010

[8] 3GPP.Tech.Specif. Group Radio Access Network;Physical Channel and Modulation, 3Gpp TS 36.211.

[9] Y. Blankship, P.Sartori, B. Classon and K.Baum, "Link Error Prediction Methods for Multicarrier System," IEEE Vehicular Technology conference FALL, LosAngels, 2004