SCHEDULING FOR MASSIVE MIMO UNDER POWER AND QOS CONSTRAINTS

Hung Pham*, **Bac Dang Hoai**†, **Ban Nguyen Tien**†
* The IT faculty of NAEM, Hanoi, Vietnam
† Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

Abstract—**Massive MIMO networks will support Quality of Service with the new sublayer Service Data Adaption Protocol to map QoS flows into data radio bearers. Different services in Massive MIMO networks will have different requirements of QoS. One of the most important requirements is the data rate of service. This leads to a considerable difference in transmit power for various classes of traffic. In this paper, we proposed a scheduling for Massive MIMO which supports QoS with the consideration of the priority as well as the minimum data rate termed as QoS-Assurance. To guarantee the minimum data rate of each type of service, a minimum transmit power is assigned for each class of traffic per service. Hence, with the other information such as user's used rate in the past and the priority of traffic, the probability of occupying channels is determined. The simulation results of QoS-Assurance scheduling are compared with that of Maximum Rate and QoS Scheduler. The results show that the QoS-Assurance scheduling can guarantee the minimum rate of service and get a higher useful throughput than Maximum Rate and QoS Scheduler.**

Keywords—**Massive MIMO, Scheduling, QoS, Power Control**

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is one of key enabler for 5G networks. The ideal of Massive MIMO is utilizing a large-scale antenna arrays at the base station (BS) [1]. The Massive MIMO system achieves higher multiplexing gain by increasing the number of antennas at the base station and users [2]. Furthermore, the Massive MIMO systems use simple linear precoders of maximum ratio transmission (MRT) or zero-forcing (ZF) to get the optimal performance [3], [4]. To exploit the advantages of Massive MIMO, it still faces some obstacles to get the optimal throughput.

To achieve the optimal throughput of Massive MIMO systems, the BS has to acquire an exact channel state information (CSI) for precoding data. The frequency division duplex (FDD) operation requires more training overhead than the time division duplex (TDD) operation. Therefore, many researches have adopted TDD operation in Massive MIMO systems to get the optimal performance [5], [1], [6], [7]. In the TDD operation, the BS estimates the CSI in the pilot period. In the downlink stage, the BS will choose a subset of users in the cell to transmit the data. This is called a scheduling process.

The boom of the Internet affected on the development of wireless networks. The IP-based services are transmitted over wireless transmission. So, wireless networks have to support the characteristics of IP traffic. The IP traffic is mapped into many service classes with different requirements of Quality of service (QoS) per class. QoS determines the priority, the delay, and the rate of each service. In 5G, wireless networks will support QoS from the core network to mobile terminals for the first time [8].

To get the high throughput of the system, only users with the high channel gains are selected to transmit the data for the Maximum Rate (MR) method [9], [10], [11]. Therefore, it is unacceptable for users with the bad channel gains. Another approach is using power control which increases the power for users in bad channel conditions and vice versa [12]. In [13], the authors focus on the energy and spectral efficiency of the whole system rather than the power of each user. To tackle the problem of MR, the historical average data of users is considered in Proportion Fairness (PF) to provide a same rate for all users. The purpose of PF is to create a fairness for all users [14]. However, from our point of view the fairness should be considered in term of classes of traffic not only for users because nowadays one user can use many applications at the same time on the mobile device.

In the line of research about the QoS, the delay is attracted a lot of research. The delay of traffic is considered in the Maximum-Largest Weighted Delay First (M-LWDF) when making scheduling decision [15]. The priority and guarantee bit rate are considered in [16]. However their algorithm on relative priority part still depends on the delay but from our point of view the priority of traffic only depends on the type of service instead of the delay. Moreover, their proposal transmits all the Guaranteed Bit Rate (GBR) traffics as the customer's demand to make sure that there is no drop for GBR class. It may leads some GBR users to take most the bandwidth for others non-GBR users. Power control is used in [17] to decide the amount of data transmission to improve the time-average delay performance. A combination of scheduling and power control is also studied in [18]. The minimum

Correspondence: Pham Hung, email: hungp@niem.edu.vn

Manuscript communication: received:2/2020, revised: 4/2020, accepted: 4/2020.

rate to support QoS constraint for Massive MIMO systems is researched in [19]. To guarantee only one minimum rate for all users, a joint subset of antennas and users are selected using convex relaxation.

To the best of our knowledge, in published literature's on QoS scheduling for Massive MIMO, the authors mainly focused on delay or concern only one minimum rate for all users. They have not concerned on the different minimum rates per class of traffic yet. In this paper, QoS-Assurance scheduling is proposed with a method to calculate the minimum power of each user to guarantee their minimum requirement rates. It is help to save a lot of power to transmit the data for other users. Our algorithm also distinguishes the priorities of classes so the more important classes can have more chance to be scheduled. The effect of QoS-Assurance is studied in terms of throughput, average rates and useful throughput. Our results show that the throughput has improved very much comparing to QoS Scheduler in [16] while it can guarantee the different minimum rates and the priority for users.

Notation: We denote normal letters (e.g., a) for scalars, column vectors and matrices are lowercase and uppercase boldface letters (e.g., $\mathbf{h}$ and $\mathbf{H}$). $\mathbf{I}_N$ is the identity matrix. $\mathbf{0}_N$ is all-zero matrices of size $N \times N$. $\mathbf{A}^T$ is the transpose matrix of a matrix $\mathbf{A}$, $\mathbf{A}^*$ is the conjugate transpose, and $\mathrm{tr}(\mathbf{A})$ is the trace. $\mathbb{E}[\cdot]$ is the statistical expectation operator.

II. SYSTEM MODEL

We study a single-cell massive MIMO including of one BS and K_a users. The BS has M antennas and the users has only a single antenna, where $M \geq K_a$. All users share the same time and frequency resource. The system works in TDD mode and the perfect channel reciprocity is assumed. Let T be the frame time in terms of symbols. In every frame the first τ_p symbols are used to estimate the CSI then the remaining ($T - \tau_p$) symbols are used to transmit the data.

A. Uplink Training

To estimate the channel matrix between the BS and users, the BS assigns orthogonal pilot sequences $\mathbf{V} \in \mathbb{C}^{K_t \times \tau_p}$ with the length of τ_p symbols to K_t users ($\tau_p \geq K_t$) selected from associated user set K_a. Let $\mathbf{H} \in \mathbb{C}^{M \times K_t}$ be the channel matrix between the BS antenna array and K_t users. We consider a block fading channel model where the channel coefficients keep unchanged during each frame. Let $\mathbf{h}_k$ be the $M \times 1$ channel vector for the k-th user, which is a column of matrix $\mathbf{H}$ and is given by

$$\mathbf{h}_k = \mathbf{g}_k \sqrt{\beta_k} \tag{1}$$

where the elements of $\mathbf{g}_k$ are i.i.d. Gaussian distributed with zero mean and unit variance, β_k is the large-scale fading coefficient that counts for path-loss and log-normal shadowing which is a constant for many frames. The BS obtains the $M \times \tau_p$ observation matrix:

$$\mathbf{Y}_r = \sqrt{\tau_p p_p}\mathbf{H}\mathbf{V} + \mathbf{N} \tag{2}$$

where p_p is the transmit power per user, $\mathbf{N}$ is a noise Gaussian matrix with independent and identically distributed (i.i.d) entries $\mathcal{CN}(0, \sigma^2 \mathbf{I}_M)$. Then, the minimum mean-square error (MMSE) estimate of $\mathbf{H}$ is given by [20]:

$$\begin{aligned}
\hat{\mathbf{H}} &= \frac{\sqrt{\tau_p p_p}}{\sigma^2 + \tau_p p_p}\mathbf{Y}_r \mathbf{V}^H \\
\hat{\mathbf{h}}_k &= \frac{\sqrt{\tau_p p_p \beta_k}}{\sigma^2 + \tau_p p_p \beta_k}\hat{\mathbf{y}}_{r,k} \\
\mathbb{E}\{(\mathbf{h}_k - \hat{\mathbf{h}_k})(\mathbf{h}_k - \hat{\mathbf{h}_k})^H\} &= (\beta_k - \frac{p_p \tau_p \beta_k^2}{p_p \tau_p \beta_k + \sigma^2})\mathbf{I}_M \\
\xi_k &= \frac{p_p \tau_p \beta_k^2}{p_p \tau_p \beta_k + \sigma^2}
\end{aligned} \tag{3}$$

B. Downlink Transmission

After pilot stage, the BS chooses a scheduling user set $\mathbb{K}_\mathrm{s} = \{1, 2, ..., K_s\}$ from the pilot user set $\mathbb{K}_\mathrm{t} = \{1, 2..., K_t\}$ with $K_s \leq K_t$ to serve in the downlink transmission. Let $\mathbf{x} \in \mathbb{C}^{K_s \times 1}$ is the data for K_s users and $\mathbb{E}\left\{||\mathbf{x}||^2\right\} = 1$.

The BS calculates a linear precoding matrix $\mathbf{F} \in \mathbb{C}^{M \times K_s}$ from channel estimation $\hat{\mathbf{H}}$ to map the data $\mathbf{x}$ to antennas. The power of k-th user p_k is under power constraint $\sum\limits_{k=1}^{K_s} |\mathbf{f}_k|^2 p_k \leq P$. The received signal at the k-th user can be written as:

$$\begin{aligned}
\mathbf{y}_k &= \mathbf{h}_k^T \mathbf{f}_k \sqrt{p_k} x_k + \sum_{l=1, l \neq k}^{K_s} \mathbf{h}_k^T \mathbf{f}_l \sqrt{p_l} x_l + \mathbf{n}_k \\
&= \mathbb{E}\{\mathbf{h}_k^T \mathbf{f}_k\}\sqrt{p_k} x_k \\
&+ \sum_{l=1}^{K_s} \mathbf{h}_k^T \mathbf{f}_l \sqrt{p_l} x_l - \mathbb{E}\{\mathbf{h}_k^T \mathbf{f}_k\}\sqrt{p_k} x_k + \mathbf{n}_k
\end{aligned} \tag{4}$$

where $\mathbf{h}_k$ is the channel vector for the k-th user, and $\mathbf{f}_k$ is the k-th column of matrix $\mathbf{F}$.

The instantaneous signal to interference noise ratio (SINR) for the k-th user can be written as:

$$\gamma_k = \frac{p_k \left|\mathbb{E}\{\mathbf{h}_k^T \mathbf{f}_k\}\right|^2}{\sum\limits_{l=1}^{K_s} p_l \mathbb{E}\{\left|\mathbf{h}_k^T \mathbf{f}_l\right|^2\} - p_k \left|\mathbb{E}\{\mathbf{h}_k^T \mathbf{f}_k\}\right|^2 + \sigma^2} \tag{5}$$

The achievable rate of the k-th user is

$$\mathbf{R}_k = \log_2(1 + \gamma_k) \tag{6}$$

The sum-rate of the system is:

$$\mathbf{R}_{sum} = \sum_{l=1}^{K_s} \log_2(1 + \gamma_k) \tag{7}$$

III. PROBLEM FORMULATION AND PRIOR SOLUTIONS

A. Problem Formulation

In the user plane protocol stack of 5G, a new layer Service Data Adaption Protocol (SDAP) are added on the top of PDCP layer to map between a QoS flow and a data radio bearer [8]. It is the first time, the QoS is concerned fully at the radio bearer. The connection rate is one of the most important factors for the QoS requirements. Scheduling algorithm plays an important role to guarantee the QoS for these applications.

The channel state information $\hat{\mathbf{H}}$, the total transmit power P and the data rate requirement T_k of the user k-th are gathered for the BS. From this information, the best user subset $\mathbb{K}_{\mathrm{s}}$ are chosen from the pilot user set $\mathbb{K}_{\mathrm{t}}$ to maximize the sum-rate of the system in each time frame.

We will find the minimum transmit power to meet the rate requirement for each user and use the remain power for more users therefore the sum-rate of the whole system will increases:

$$\begin{array}{c} \max\limits_{\{K_s\}} \sum\limits_{k=1}^{K_s} \log_2(1+\gamma_k) \\ s.t \quad \sum\limits_{k=1}^{K_s} |\mathbf{f}_k|^2 p_k \le P \\ \log_2(1+\gamma_k) \ge T_k \end{array} \tag{8}$$

B. Prior Solutions

1) Maximum Rate: The traditional algorithm to maximize the sum rate of the system is Maximum Rate. The sum rate of K users in the system is $\max \sum\limits_{k=1}^{K} R_k$, where R_k is the rate allocated to user k. The Maximum Rate is archived with a transmit power given by [9], [21]:

$$p_k(m,f) = \begin{cases} [\frac{1}{\lambda_k} - \frac{\mathbf{N_0}}{\|\mathbf{H}_k(m,f)\|^2}]^+, \\ \text{if}\|\mathbf{H}_k(m,f)\|^2 \ge \frac{\lambda_k}{\lambda_l}\|\mathbf{H}_l(m,f)\|^2 \\ 0 \qquad \text{otherwise} \end{cases} \tag{9}$$

where $[x]^+ = \max(0,x)$, $\mathbf{H}_k(m,f)$ is the channel gain of user k in resource block (RB) m of sub-frame f and λ_k is constant which is chosen to satisfy power constraint

$$\sum_{k=1}^{K} p_k \le P \tag{10}$$

The result in (9) shows that only users with the best channel gains are scheduled. A variant of this resource allocation strategy with no power control is called 'Maximum-rate constant-power' scheduling where the only users with the best channel gains is scheduled but no adaptation of the transmit power [9]:

$$p_k = \frac{P}{K} \tag{11}$$

2) Proportional Fairness: For the customer's perspective, the MR is not preferred because the bad channel gain users never receive any traffic. Proportion Fairness solves this issue by taking into account the user's historical data to provide the same rate for all users. It chooses a user whose metric M is highest. The priority metric of the i-th user, M_i is given in the following equation [14].

$$M_i = \arg\max \frac{R_i(t)}{\bar{R}_i(t)} \tag{12}$$

$$\bar{R}_i(t) = (1-\frac{1}{t_c}) * \bar{R}_i(t-1) + \frac{1}{t_c} * R_i(t-1) \tag{13}$$

where,

$R_i(t)$ is the instantaneously achievable transmission rate.

$\bar{R}_i(t)$ is the average data of i-th user at time t.

t_c is the update window size.

$\bar{R}_i(t-1) = 0$ if the user i is not selected for the transmission at the time $t-1$.

It can be realized that the PF providers a higher priority not only to the users with good channel gain but also to the users with low average data rate.

3) QoS Scheduler: In [16], Ameigeiras et al. proposed a QoS Scheduler that considers the priority, the delay of traffic classes as well as tries to fulfill the required Guaranteed Bit Rate for GBR traffic. Their algorithm improved the performance of system when the delay of an user reaches the upper bound D by using a sigmoid function for the part of metric regarding to delay

$$P_k[n,s] = (1+f(w_k))\tfrac{R_k[n,s]}{[r_k[n]]^\alpha} \tag{14}$$

where, $f(w_k) = \frac{1}{1+e^{-a_k(w_k-D)}}$.

α is a factor that controls the degree of fairness.

w_k represents the Head of Line Delay of user k.

The parameter a_k adjust the slope of sigmoid function.

The parameter c establishes its upper bound.

Then it multiplies with the relative priority of traffic $F_k^{QCI_m}$

$$P_k^{QCI_m}[n,s] = P_k[n,s]F_k^{QCI_m} \tag{15}$$

The algorithm to determine $F_k^{QCI_m}$ depends on a quality performance indicator Q_k for each bearer. Q_k is a function of the delay d_k of user k for GBR class, and a function of the transmitted data rate of user k for non-GBR class.

$$Q_k[n] = \begin{cases} d_k[n] = (1-\rho_d)d_k[n-1] + \rho_d\frac{q_k[n]}{\lambda_k} \\ QCI_m = 1,2,3,4 \\ r_k[n] = (1-\rho_r)r_k[n-1] + \rho_r r_k[n] \\ QCI_m = 5,6,7,8,9 \end{cases} \tag{16}$$

where, $q_k[n]$ denotes the number of bits in the queue of bearer k in TTI n.

λ_k is an estimator of the average arrival bit rate on the bearer k.

$r_k[n]$ represents the transmitted data rate in TTI n by the bearer k.

ρ_d and ρ_r are time averaging constants. The function in (16) does not check about the minimum requirement rate of GBR users, it may lead to one user with high volume of traffic will use most of the bandwidth which should be shared with other non-GBR users.

IV. PROPOSED SOLUTION

At beginning, we estimate the power that can meet the requirement of rate per user in the case the maximum ratio transmission (MRT) based precoding is used. We also assume that all users have the same data rate requirement T_k:

$$\begin{aligned} R_k = \log_2(1+\gamma_k) \geq T_k \\ \gamma_k \geq 2^{T_k}-1 \\ \frac{p_k\left|\mathbb{E}\{\mathbf{h}_k^T\mathbf{f}_k\}\right|^2}{\sum\limits_{l=1}^{K_s} p_l\mathbb{E}\{\left|\mathbf{h}_k^T\mathbf{f}_l\right|^2\} - p_k\left|\mathbb{E}\{\mathbf{h}_k^T\mathbf{f}_k\}\right|^2+\sigma^2} \geq 2^{T_k}-1 \end{aligned} \tag{17}$$

The p_k in (8) can be found by channel inversion method [22]

$$p_k = \tfrac{P}{K_s|\mathbf{f}_k|^2} \tag{18}$$

The precoding matrices are

$$\mathbf{f}_k^{MRT} = \hat{\mathbf{h}}_k^* \tag{19}$$

Therefore, we have

$$\mathbb{E}\{\mathbf{h}_k^T\mathbf{f}_k\} = \xi_k M \tag{20}$$

$$\mathbb{E}\{\|\mathbf{h}_k^T\mathbf{f}_k\|^2\} = \xi_k^2 M^2 + \xi_k\beta_k M \tag{21}$$

$$\mathbb{E}\{\|\mathbf{h}_k^T\mathbf{f}_l\|^2\} = \xi_l\beta_k M \tag{22}$$

From (17) to (22), we have

$$\gamma_k = \frac{p_k\xi_k^2M^2}{\sum\limits_{l=1}^{K_s} p_l\xi_l\beta_k M+\sigma^2} \tag{23}$$

$$\geq \frac{p_k\xi_k^2M^2}{\sum\limits_{l=1}^{K_s} p_l\beta_l\beta_k M+\sigma^2} \tag{24}$$

$$= \frac{M^2(\frac{\tau_p p_p\beta_k}{\sigma^2+\tau_p p_p\beta_k})\beta_k\frac{P}{K_s}}{\sum\limits_{l=1}^{K_s} M\beta_k\frac{P}{K_s}+\sigma^2} \tag{25}$$

$$= \frac{(\frac{\tau_p p_p\beta_k}{\sigma^2+\tau_p p_p\beta_k})\beta_k M^2 P}{K_s(\beta_k MP+\sigma^2)} \geq 2^{T_k}-1 \tag{26}$$

$$\frac{(\frac{\tau_p p_p\beta_k}{\sigma^2+\tau_p p_p\beta_k})\beta_k M^2 P}{(2^{T_k}-1)(\beta_k MP+\sigma^2)} \geq K_s \tag{27}$$

It is obvious that the minimum rate of the user k only depends on the parameters β_k, K_s and P, not related to other users. So we can extend the result in 23 for the case that users have different minimum rates. Here, we can calculate the p_k of user k by adjusting the parameter K_s in 18 to achieve the target rate T_k.

The list K_s is built by selecting the user having the highest metric. The metric for the k-th user is a function of the demanded transmit power of user p_k , the channel quality h_k, and the priority of traffic ψ_k at frame n:

$$\rho_k[n] = (1+f_w(k))\frac{|\mathbf{h}_k|^2}{{\psi_k}^{\alpha}p_k}\frac{T_k[n]}{\bar{R}_k[n]} \tag{28}$$

where, α is a factor for adjusting the importance of class.

$T_k[n]$ is the target transmission rate at frame n.

$\bar{R}_k[n]$ is the past average data of k-th user at frame n.

For example, after the user k-th is chosen the remain power will be decreased by $f_k^2p_k$. This process will be repeated until the remain power cannot meet for the demanded power of any remain user. The algorithm is described in the Algorithm 1

Algorithm 1 QoS-Assurance Scheduling

```
1: The BS initializes $\mathcal{S}(1:K_t)=0, \mathbb{K}_\mathrm{t}=\{1,2,...,K_t\}, i=1$
2: Compute $f_w(k), K_s, p_k, \rho_k[n]$ for $\forall k \in \mathbb{K}_\mathrm{t}$ in 14, 27, 18, 28
3: Select the user i:
       $stop = 0$;
       $\rho_i[n] = \arg\max \rho_k[n], \forall k \in \mathbb{K}_\mathrm{t}$
4: if $P \geq \|\mathbf{f}_i\|^2 p_i$ then
       $\mathbb{K}_\mathrm{t} = \mathbb{K}_\mathrm{t} \setminus i$;
       $stop = 1$;
       $P = P - \|\mathbf{f}_i\|^2 p_i$;
       $\mathcal{S}(i) = 1$;
5: end if
6: if $stop = 1$ then
       Come back to step 3
7: else
       Algorithm finishes.
8: end if
```

V. SIMULATION RESULT

To investigate the performance of QoS-Assurance scheduling, various case studies have been simulated to compare following scheduling policies:

- MR scheduling
- QoS Scheduler in [16]
- QoS-Assurance scheduler

The following Table 1 lists the main simulation parameters.

Fig. 1 shows the rate per user in the case of QoS-Assurance method as the number of antennas increases. In this case, we assumed that all users in the cell are trained and then served to see how much the average rate each user will get. We study two cases: one is $K_t = K_s = 39$, the other is $K_t = K_s = 119$. We can see from the figure,

Table I

SIMULATION PARAMETERS	
PARAMETERS	ASSUMPTIONS
Number of classes C	2
Minimum rate of class 1 users	2 (bit/s/Hz)
Minimum rate of class 2 users	1 (bit/s/Hz)

with $K_s = 39$ the averaged rate per user will go to about 2 (bit/s/Hz). With $K_s = 119$ the averaged rate per user will go to about 1 (bit/s/Hz). So it is confirmed that our power control in QoS-Assurance method can guarantee any minimum data rate for users by changing the number of serving users K_s.

Fig. 2 shows the sum rate of the system for the QoS-Assurance method. We can see from the figure that the sum rate of system will increase as the number of serving users K_s goes up. With the number of scheduled users $K_s = 119$, the sum rate gets about 120 (bit/s/Hz). It decreases 30 percent down to about 80 (bit/s/Hz) for the $K_s = 39$. So serving more users will get a higher sum rate but the rate per user will go down.

From now on, we will study the issue of guarantee the minimum rate for users with QoS-Assurance method and the Maximum Rate method. In this case, there are $K_t = 58$ users separating into two classes. The class 1 includes 29 users requiring a minimum rate 2 (bit/s/Hz) and the users in class 2 requiring a minimum rate 1 (bit/s/Hz). The Maximum Rate method will use an equal power for all users $p_k = \frac{P}{K_t}$. Fig. 3 shows the sum rate of Maximum Rate method is always a little higher than the one of QoS-Assurance method.

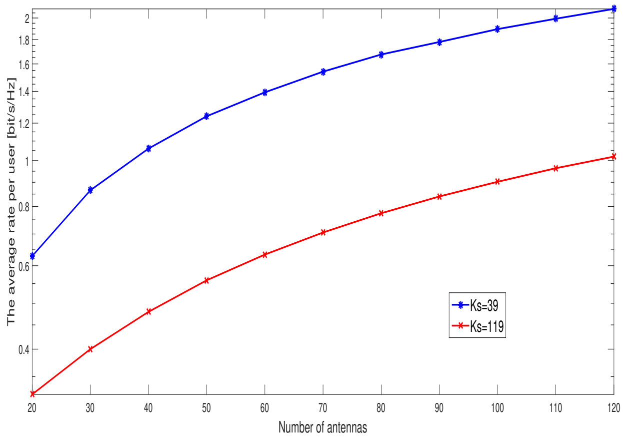

Figure 1. The rate per user for QoS-Assurance method.

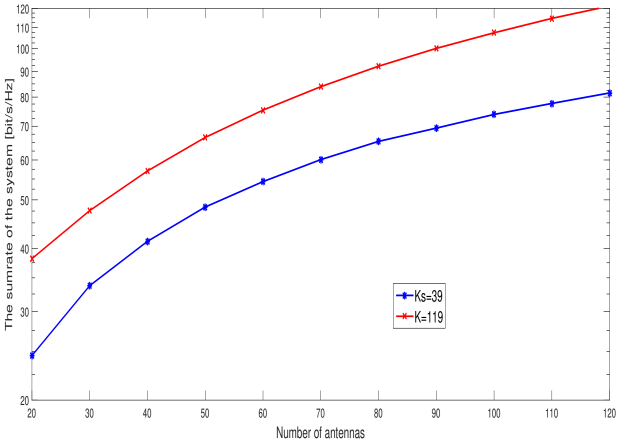

Figure 2. The sum rate of the whole system.

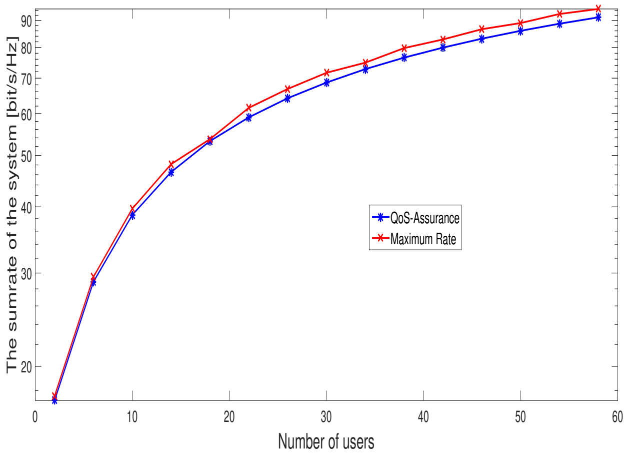

Figure 3. The comparison of sum rate.

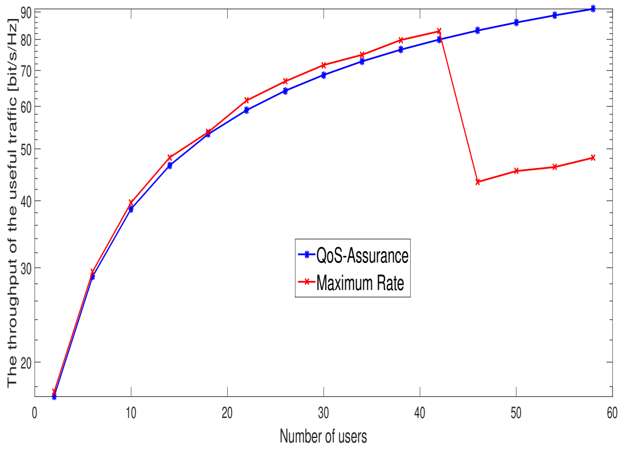

Figure 4. The comparison of useful throughput between two methods.

Fig. 4 shows the throughput of useful traffic in that only the traffic of users who meet their requirement on minimum rate is summed. For the Maximum Rate method, it is obvious when the $K_t \geq 39$, almost of users will get the rates lower than 2 (bit/s/Hz) so most of the scheduled users belonging to the class 2 do not have useful traffic. Consequently, the useful throughput of Maximum Rate method falls down sharply before it continues to increase slightly due to the increase of the user number. On the other side, the users belong to class 1 will meet the requirement of 2 (bit/s/Hz) and the

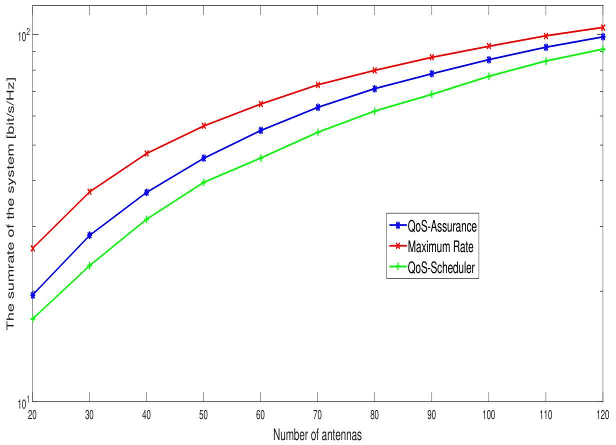

Figure 5. The comparison of throughput between three methods.

users belong to class 2 will satisfy the requirement of 1 (bit/s/Hz) for QoS-Assurance. So all the sum rate of QoS-Assurance method are useful throughput. It is clear that the Maximum Rate provides the highest sum rate of the system but our QoS-Assurance scheduling gets a better useful throughput, improves the fairness among classes of traffic, and especially guarantee the minimum rates for users.

Above case studies mainly emphasize the advantage of the power control scheme in our proposal. The last case study shows an effect of the priority metric on the throughput. Fig. 5 shows the comparison of throughput among three methods: QoS-Assurance, Maximum Rate and QoS Scheduler. The power control is applied to all three methods. So the main difference is that QoS-Assurance has the priority metric. From the set of $K_t = 100$ users with one half of the user set is class 1 users and the others is class 2 users, a subset of K_s users will be chosen to serve. It can be seen that the Maximum Rate gets the highest throughput, the QoS-Assurance takes the second place and the last one is QoS Scheduler. In the QoS-Assurance, the trade-off between the throughput and fairness depends on the priority metric. The throughput of QoS-Assurance will be closer to the one of Maximum Rate if the factor for adjusting the priority α is going to zero but the throughput of class 1 is going to smaller also. In other words, the fairness of classes decreases.

VI. CONCLUSIONS

QoS-Assurance scheduler guarantees a various of the minimum rates under power constraint, supports the priority of traffic and also improves the fairness among classes of traffic. It has improved the useful throughput of the system comparing with the traditional methods such as Maximum Rate and QoS Scheduler. The numerical results show that QoS-Assurance can deploy multi-rate services efficiently on Massive MIMO networks.

REFERENCES

[1] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, November 2010.

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up mimo: Opportunities and challenges with very large arrays,” *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, January 2013.

[3] J. Hoydis, S. ten Brink, and M. Debbah, “Massive mimo in the ul/dl of cellular networks: How many antennas do we need?” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, February 2013.

[4] C. Kong, C. Zhong, and Z. Zhang, “Performance of zf precoder in downlink massive mimo with non-uniform user distribution,” *Journal of Communications and Networks*, vol. 18, no. 5, pp. 688–698, October 2016.

[5] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Massive mu-mimo downlink tdd systems with linear precoding and downlink pilots,” in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, October 2013, pp. 293–298.

[6] H. Yang and T. L. Marzetta, “Performance of conjugate and zero-forcing beamforming in large-scale antenna systems,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 172–179, February 2013.

[7] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive mimo for next generation wireless systems,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[8] 3GPP, “Tr 38.804 v14.0.0,” *3GPP*, no. 03, 2017.

[9] M. Shariat, A. U. Quddus, S. A. Ghorashi, and R. Tafazolli, “Scheduling as an important cross-layer operation for emerging broadband wireless systems,” *IEEE Communications Surveys Tutorials*, vol. 11, no. 2, pp. 74–86, 2009.

[10] M. Bohge, J. Gross, A. Wolisz, and M. Meyer, “Dynamic resource allocation in ofdm systems: an overview of cross-layer optimization principles and techniques,” *IEEE Network*, vol. 21, no. 1, pp. 53–59, January 2007.

[11] M. Alkhaled, E. Alsusa, and W. Pramudito, “Adaptive user grouping algorithm for the downlink massive mimo systems,” in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[12] T. Yoo and A. Goldsmith, “On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, March 2006.

[13] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and spectral efficiency of very large multiuser mimo systems,” *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, April 2013.

[14] F. Afroz, R. Heidery, M. Shehab, K. Sandrasegaran, and S. Shompa, “Comparative analysis of downlink packet scheduling algorithms in 3gpp lte networks,” *International Journal of Wireless Mobile Networks*, vol. 7, pp. 1–21, October 2015.

[15] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, “Providing quality of service over a shared wireless link,” *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, February 2001.

[16] P. Ameigeiras, J. Navarro-Ortiz, P. Andres-Maldonado, J. M. Lopez-Soler, J. Lorca, Q. Perez-Tarrero, and R. Garcia-Perez, “3gpp qos-based scheduling framework for lte,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, p. 78, March 2016.

[17] Z. Chen, E. Bjornson, and E. G. Larsson, “Dynamic scheduling and power control in uplink massive mimo with random data arrivals,” in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[18] J. Choi, N. Lee, S. Hong, and G. Caire, “Joint user scheduling, power allocation, and precoding design for massive mimo systems: A principal component analysis approach,” in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 396–400.

[19] J. Akhtar and K. Rajawat, “Quality-of-service constrained user and antenna selection in downlink massive-mimo systems,” in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, 2019, pp. 1–6.

[20] H. Lee, S. Park, and S. Bahk, "Enhancing spectral efficiency using aged csi in massive mimo systems," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

[21] S. Stefania, *LTE The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. Wiley, 2009.

[22] Bjornson, "The massive mimo paradigm fundamentals and state of the art," *GLOBECOM*, p. 82, 2016.

THUẬT TOÁN LẬP LỊCH CHO MẠNG VÔ TUYẾN NHIỀU ĂNG TEN CỠ RẤT LỚN DƯỚI ĐIỀU KIỆN GIỚI HẠN VỀ CÔNG SUẤT VÀ CHẤT LƯỢNG DỊCH VỤ

Tóm tắt: Mạng viễn thông nhiều ăng ten cỡ rất lớn hỗ trợ đảm bảo chất lượng dịch vụ bằng cách thêm vào một lớp con mới ở phía trên lớp Giao thức hội tụ gói dữ liệu để ghép các luồng dữ liệu vào các sóng mang vô tuyến phù hợp. Mỗi dịch vụ trong mạng vô tuyến nhiều ăng ten cỡ rất lớn sẽ có các yêu cầu khác nhau về chất lượng dịch vụ. Một trong những yêu cầu quan trọng nhất là tốc độ dữ liệu của dịch vụ. Điều này dẫn đến một sự khác biệt đáng kể trong công suất phát của các lớp dịch vụ khác nhau. Trong bài báo này, một thuật toán lập lịch lý thuyết, gọi là QoS-Assurance, được đề xuất cho mạng viễn thông nhiều ăng ten cỡ rất lớn đã xem xét các tham số về độ ưu tiên cũng như tốc độ tối thiểu của thuê bao. Để đảm bảo tốc độ tối thiểu của mỗi loại dịch vụ, một công suất phát tối thiểu được tính toán cho từng lưu lượng của mỗi loại dịch vụ. Vì thế, với các thông tin khác như lưu lượng đã sử dụng của thuê bao trong quá khứ và độ ưu tiên của lưu lượng, xác suất chiếm kênh của lưu lượng sẽ được xác định. Kết quả mô phỏng của thuật toán QoS-Assurance được so sánh với kết quả của phương pháp Maximum Rate và QoS Scheduler. Kết quả chứng minh rằng thuật toán QoS-Assurance có thể đảm bảo tốc độ tối thiểu của dịch vụ và có lưu lượng hữu ích cao hơn thuật toán Maximum Rate và QoS Scheduler.

Từ khóa- Hệ thống nhiều ăng ten cỡ rất lớn, Lập lịch, QoS, Điều khiển công suất.

Hung Pham graduated from Ha Noi University of Science and Technology (HUST) in 2004. He received the Master of Electronics Engineering from the Dongguk University in Korea in 2011 with the thesis of Scheduling for LTE network. He is currently a Ph.D student at Faculty of Telecommunications 1, Posts and Telecommunications Institute of Technology (PTIT). The main topic of current research is scheduling for Massive MIMO and mmWave networks.

Dang Hoai Bac received his master and doctorate at Posts and Telecommunications Institute of Technology (PTIT), in 2004 and 2010. From 2009 to 2010, he was a researcher at Orange Lab, France Telecom, Paris, France. Currently, he is an associate professor and vice president at Posts and Telecommunications Institute of Technology. His research areas are: Automatic Control, Signal Processing,Embedded Systems and System Integration.

Nguyen Tien Ban received his doctorate at Saint-Petersburg State University of Telecommunications (SUT), Russian Federation in 2003. Currently, he is an associate professor in Telecommunication Faculty 1, Posts and Telecommunications Institute of Technology. His research areas are: Network Performance Analysis and Design, Network Design and Optimization, Modeling and Simulation of Telecommunication Systems.