Decentralized Heterogeneous Statistical QoS Provisioning for Uplinks Over 5G Wireless Networks

Wenchi Cheng†, Xi Zhang‡, and Hailin Zhang†

†State Key Laboratory of Integrated Services Networks, Xidian University, Xi’an, China
‡Networking and Information Systems Laboratory

Abstract—The newly imposed heterogeneous statistical delay-bounded quality of service (QoS) provisioning, which refers to the different/variable delay-bounded QoS guarantees among different wireless links, for the fifth-generation (5G) mobile multimedia wireless networks has received much research attention recently. The heterogeneous statistical delay-bounded QoS provisioning can be classified into two categories: centralized and decentralized heterogeneous statistical delay-bounded QoS provisioning where the delay-bounded QoS requirements are centralized at the base station (BS) and distributed at different mobile user equipments (UEs), respectively. The decentralized heterogeneous statistical delay-bounded QoS provisioning is more challenging than the centralized heterogeneous statistical delay-bounded QoS provisioning. In this paper, we build up the system model for decentralized heterogeneous statistical delay-bounded QoS provisioning in terms of aggregate effective capacity and heterogeneous QoS exponents. Based on this model, we develop the optimal joint bandwidth and power allocations scheme to maximize the aggregate effective capacity of uplink transmissions supporting the heterogeneous statistical delay-bounded QoS provisioning over 5G mobile multimedia wireless networks. The BS allocates the bandwidth based on the QoS exponents of all uplinks while the mobile UEs dynamically allocate the power based on the corresponding instantaneous channel state information (CSI), the corresponding QoS exponent, and the bandwidth allocated by the BS. We conduct the extensive simulations to validate and evaluate our developed optimal joint bandwidth and power allocations schemes, showing that our proposed heterogeneous statistical delay-bounded QoS provisioning scheme can significantly increase the aggregate effective capacity as compared with the homogeneous statistical delay-bounded QoS provisioning schemes.

Index Terms—Fifth-generation (5G) mobile wireless networks, quality of service (QoS), cross-layer design/optimization, decentralized heterogeneous statistical delay-bounded QoS provisioning, joint bandwidth and power allocations scheme.

I. INTRODUCTION

The statistical delay-bounded quality of service (QoS) provisioning, which represents one of key components of fourth-generation (4G) and pre-4G mobile multimedia wireless networks, has been shown to be a powerful technique to characterize and implement the delay-bounded QoS guarantee for wireless real-time traffics [1]–[5]. For the fifth-generation (5G) mobile multimedia wireless networks [6], [7], because the time-sensitive services may vary dramatically in both the large range from millisecond to a few seconds and the diversity from uniform/constant delay-bounded to different/variable delay-bounded guarantees among different wireless links, the delay-bounded QoS requirements for different types of services promote the new heterogeneous statistical delay-bounded QoS provisioning, which refers to the different/variable delay-bounded QoS guarantees among different 5G wireless links.

According to the distribution (centralization or decentralization) of delay-bounded QoS requirements in 5G mobile multimedia wireless networks, the heterogeneous statistical delay-bounded QoS provisioning can be classified into two categories: centralized and decentralized heterogeneous statistical delay-bounded QoS provisioning. For the centralized heterogeneous statistical delay-bounded QoS provisioning where the typical scenario is the downlink transmission, the delay-bounded QoS requirements are centralized at the base station (BS) [8]–[10]. For the decentralized heterogeneous statistical delay-bounded QoS provisioning where the typical scenario is the uplink transmission, the delay-bounded QoS requirements are distributed at different mobile user equipments (UEs) [9].

To maximize the system throughput for 5G mobile multimedia wireless networks, the bandwidth and power allocations for the 5G mobile multimedia wireless networks need to be adaptive to the variations of channel state information (CSI) and delay-bounded QoS requirements. For downlink transmissions, the delay-bounded QoS requirements for all downlinks are located at the BS. It is more efficient for the BS to jointly allocate the power and bandwidth in supporting the centralized heterogeneous statistical delay-bounded QoS provisioning. For uplink transmissions, the bandwidth allocation scheme is determined by the BS based on all delay-bounded QoS requirements distributed among different mobile UEs while the power allocation scheme is determined by each UE. Therefore, providing decentralized heterogeneous statistical...
delay-bound QoS guarantees for uplink transmissions is more challenging than guaranteeing centralized heterogeneous statistical delay-bound QoS for downlink transmissions.

To overcome the aforementioned problems, in this paper we propose the decentralized heterogeneous statistical QoS provisioning framework for the uplink transmissions in 5G mobile multimedia wireless networks. Applying the effective capacity theory and its corresponding QoS exponents controlling parameter, we build up the system model for the uplink transmissions in terms of aggregate effective capacity and heterogeneous QoS exponents to support heterogeneous statistical delay-bound QoS guarantees for 5G mobile multimedia wireless networks. Based on our system model, we formulate the optimization problems to maximize the aggregate effective capacity for uplink transmissions with heterogeneous statistical delay-bound QoS requirements. To maximize the aggregate effective capacity, the BS allocates the bandwidth based on the QoS exponents of all uplinks while the mobile UEs dynamically allocate the power based on the corresponding instantaneous CSI, the corresponding QoS exponent, and the bandwidth allocated by the BS, thus forming the optimal joint bandwidth and power allocations scheme. We conduct the extensive simulations experiments to validate and evaluate our developed schemes, verify our analyses for the decentralized heterogeneous statistical delay-bound QoS provisioning over 5G mobile wireless networks, and compare the aggregate effective capacity when using our developed optimal joint bandwidth and power allocations scheme with that when using the traditional homogeneous delay-bound QoS provisioning based schemes.

The rest of this paper is organized as follows. Section II builds up the system model for the decentralized statistical delay-bound QoS provisioning over uplink transmissions in 5G mobile wireless networks. Section III formulates the optimization problem to maximize the aggregate effective capacity of uplink transmissions supporting the decentralized heterogeneous statistical delay-bound QoS requirements. Solving this optimization problem, we develop the optimal joint bandwidth and power allocations schemes. Section IV simulates and evaluates our developed optimal joint bandwidth and power allocations scheme for 5G mobile wireless networks. The paper concludes with Section V.

II. THE SYSTEM MODEL

We consider the 5G mobile wireless network model for uplink transmission under cross-layer design architecture, an example of which is shown in Fig. 1, where $N$ mobile UEs are distributively located in the 5G mobile wireless networks. As illustrated in Fig. 1, each UE transmits the uplink traffic using different orthogonal frequency-division multiple access (OFDMA) subchannels to the BS. Each link needs to satisfy its corresponding delay-bound QoS requirement, which is required by the service at each UE. The data-link layer packets corresponding to each uplink are buffered in first-in-first-out (FIFO) queues of each UE to be transmitted to the BS. The packets are divided into frames at the data-link layer and then split into bit-streams at physical (PHY) layer. Based on the service-determined QoS constraints corresponding to the real-time traffic of the UE and the CSI fed back from the BS, each UE needs to find an optimal power allocation scheme that can maximize the uplink throughput subject to its given statistical delay-bound QoS requirements. On the other hand, the BS allocates the system bandwidth to coordinate the uplink throughput corresponding to each uplink, thus maximizing the entire uplink throughput of the 5G mobile wireless networks. The channel power gains follow the stationary block fading model, where they remain unchanged during a time frame with the fixed length $T$, but vary independently across different time frames. The frame duration $T$ is assumed to be less than the fading coherence time, but sufficiently long so that the information-theoretic assumption of infinite code-block length is meaningful. Without loss of generality, we denote by $B$ the entire bandwidth of the 5G mobile wireless networks. We use the Nakagami-$m$ channel model which is very generic and often best fits land-mobile and indoor-mobile multiple propagations [11]. The probability density function (PDF) for the Nakagami-$m$ channel model, denoted by $p_T(\gamma)$, can be

![Fig. 1. The uplink mobile wireless network model under cross-layer design architecture.](image-url)
expressed as follows:

$$p_r(\gamma) = \frac{\gamma^{m-1}}{\Gamma(m)} \left(\frac{m}{\bar{\gamma}}\right)^m \exp\left(-\frac{m}{\bar{\gamma}}\gamma\right), \quad \gamma \geq 0,$$

(1)

where $\Gamma(\cdot)$ denotes the Gamma function, $m$ represents the fading parameter of Nakagami-$m$ distribution, $\gamma$ denotes the instantaneous channel signal-to-noise ratio (SNR), and $\bar{\gamma}$ is the average received SNR at the mobile receiver.

**A. The Homogeneous Delay-Bounded QoS Provisioning**

Based on large deviation principle, the author of [12] showed that with sufficient conditions, the queue length process $Q(t)$ converges in distribution to a random variable $Q(\infty)$ such that

$$\lim_{Q_{th} \to \infty} \frac{-\log(P\{Q(\infty) > Q_{th}\})}{Q_{th}} = \theta \quad (2)$$

where $Q_{th}$ is the queue length bound and the parameter $\theta > 0$ is a real-valued number. The parameter $\theta$, which is called the QoS exponent, measures the exponential decay rate of the delay-bounded QoS violation probabilities. A larger $\theta$ corresponds to a faster decay rate, which implies that the system can provide a more stringent QoS requirement. A smaller $\theta$ leads to a slower decay rate, which indicates a looser QoS requirement. Asymptotically, when $\theta \to \infty$, this implies that the system cannot tolerate any delay, which corresponds to the very stringent statistical delay-bounded QoS constraint.

On the other hand, when $\theta \to 0$, the system can tolerate an arbitrarily long delay, which corresponds to the very loose statistical delay-bounded QoS constraint.

The sequence $\{R[k], k = 1, 2, \ldots\}$ is defined as the data service-rate, which is a discrete-time stationary and ergodic stochastic process. The parameter $k$ represents the time frame index with a fixed time-duration equal to $T$. The $R[k]$ changes from frame to frame and $S[t] = \sum_{k=1}^{\infty} R[k]$ represents the partial sum of the service process. The Gartner-Ellis limit of $S[t]$, expressed as $A_C(\theta) = \lim_{t \to \infty} (1/t) \log(\mathbb{E}\{e^{\theta S[t]}\})$, is a convex function differentiable for all real-valued $\theta$, where $\mathbb{E}\{\cdot\}$ denotes the expectation. The authors in [1] defined the effective capacity as the maximum constant arrival rate which can be supported by the service rate to guarantee the specified QoS exponent $\theta$. If the service-rate sequence $R[k]$ is stationary and time-uncorrelated, the effective capacity can be written as

$$C(\theta) = -\frac{1}{\theta} \log\left(\mathbb{E}\{e^{-\theta R[k]}\}\right),$$

(3)

where $\mathbb{E}\{\cdot\}$ is the expectation operation.

**B. The Decentralized Heterogeneous Statistical Delay-Bounded QoS Provisioning**

For uplink transmissions supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning, we aim to maximize the entire uplink throughput of the 5G mobile wireless networks. To maximize the uplink throughput for the 5G mobile wireless networks, in principle all UEs and the BS need to get all QoS exponents. Although the BS can get the information of QoS exponents of each UE through each subchannel, it is impractical and inefficient for the BS to send all QoS exponents to each UE using each subchannel. Alternatively, each UE dynamically allocates its power based on its delay-bounded QoS requirement, the instantaneous CSI fed back from the BS, and the bandwidth allocated by the BS.

We denote by $\theta_i$ ($1 \leq i \leq N$) the QoS exponent for the $i$th uplink corresponding to the $i$th UE. Then, we define $\{\theta_1, \theta_2, \ldots, \theta_N\}$ as the heterogeneous QoS exponents for the 5G mobile wireless networks. Based on Eq. (3), we can write the effective capacity for the $i$th uplink as follows:

$$C_i(\theta_i) = -\frac{1}{\theta_i} \log\left(\mathbb{E}_{\theta_i}\{e^{-\theta_i TB_i \log_2(1+P_i \gamma_i)}\}\right),$$

(4)

where $\gamma_i$ is the instantaneous channel SNR corresponding to the $i$th uplink, $P_i$ is the instantaneous transmit power of the $i$th UE, $B_i$ is the bandwidth allocated to the $i$th uplink, and $\mathbb{E}_{\theta_i}\{\cdot\}$ is the expectation over $\gamma_i$.

Based on Eq. (4), we define the aggregate effective capacity, denoted by $C_U(\theta_1, \theta_2, \ldots, \theta_N)$, as the sum of effective capacities corresponding to all uplinks supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning in the 5G mobile wireless networks as follows:

$$C_U(\theta_1, \theta_2, \ldots, \theta_N) = \sum_{i=1}^{N} C_i(\theta_i) = \sum_{i=1}^{N} \frac{1}{\theta_i} \log\left(\mathbb{E}_{\theta_i}\{(1+P_i \gamma_i)^{-\beta_i}\}\right),$$

(5)

where we define $\beta_i \triangleq (\theta_i TB_i / \log 2)$ as the normalized QoS exponent corresponding to $\theta_i$ and $B_i$.

Mathematically, there exists the unique real-valued number $\theta_o \in [\theta_{min}, \theta_{max}]$, where $\theta_{min} \triangleq \min\{\theta_1, \theta_2, \ldots, \theta_N\}$ and $\theta_{max} \triangleq \max\{\theta_1, \theta_2, \ldots, \theta_N\}$, such that the following equation holds [13]:

$$C_U(\theta_1, \theta_2, \ldots, \theta_N) = \sum_{i=1}^{N} \frac{1}{\theta_i} \log\left(\mathbb{E}_{\theta_i}\{(1+P_i \gamma_i)^{-\beta_i}\}\right) = -\frac{1}{\theta_o} \sum_{i=1}^{N} \log\left(\mathbb{E}_{\theta_i}\{(1+P_i \gamma_i)^{-\beta_i}\}\right),$$

(6)

which is a simpler expression of aggregate effective capacity than that shown in Eq. (5). In the following, we develop the optimal joint bandwidth and power allocations scheme to maximize the aggregate effective capacity shown in Eq. (6).

**III. OPTIMAL JOINT BANDWIDTH AND POWER ALLOCATIONS FOR UPLINK TRANSMISSIONS SUPPORTING THE DECENTRALIZED HETEROGENEOUS STATISTICAL DELAY-BOUNDED QoS PROVISIONING**

We formulate the optimization problem, denoted by $P1$, to maximize the aggregate effective capacity of uplink transmissions supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning
delay-bounded QoS provisioning, as follows:

**P1:**

\[
\arg \max_{(B_i, P_i), 1 \leq i \leq N} \left\{ -\frac{1}{\theta_0} \sum_{i=1}^{N} \log \left( \mathbb{E}_{\gamma_i} \left\{ \left( 1 + P_i \gamma_i \right)^{-\beta_i} \right\} \right) \right\} \\
= \arg \min_{(B_i, P_i), 1 \leq i \leq N} \left\{ \mathbb{E}_{\gamma} \left[ \prod_{i=1}^{N} \left( 1 + P_i \gamma_i \right)^{-\beta_i} \right] \right\}
\]

\[
\text{s.t. : } 1. \sum_{i=1}^{N} B_i \leq B; \quad 2. P_i \geq 0, 1 \leq i \leq N; \quad 3. \mathbb{E}_{\gamma_i} \{ P_i \} \leq \mathcal{P}_i, 1 \leq i \leq N,
\]

where \( \mathcal{P}_i \) is the average power constraint for the \( i \)th UE and \( \gamma \triangleq (\gamma_1, \gamma_2, ..., \gamma_N) \) is the instantaneous channel SNR for the 5G mobile wireless networks. In Eq. (7), (a) holds because \( \log(\cdot) \) is a monotonically increasing function and all uplinks are independent with each other.

Although problem **P1** is a strictly convex optimization problem, it is very difficult to derive the closed-form optimal solution for problem **P1** by directly solving **P1**. Thus, we need to find another alternative, but more efficient way, to solve problem **P1**. We notice that the solution of problem **P1** consists of two parts: the UE-determined optimal power allocation and the BS-determined optimal bandwidth allocation. This implies that the UEs determine the transmit power while the BS allocates the bandwidth. To obtain the maximum aggregate effective capacity for all uplinks in the 5G mobile wireless networks, all UEs and the BS need to share the QoS exponents corresponding to all UEs. However, it is impractical and inefficient for the BS to directly send the QoS exponents corresponding to all UEs to each UE using all subchannels. Alternatively, each UE dynamically allocates its power based on its delay-bounded QoS requirement and the instantaneous CSI fed back from the BS. Thanks to the feedback channels from the BS to each UE, the BS can send the information regarding the allocated bandwidth to each UE through all feedback subchannels. In fact, the bandwidth is allocated based on the QoS exponents of all UEs, and thus contains the information of QoS exponents corresponding to all UEs. Then, each UE can allocate its power based on its delay-bounded QoS requirement, the instantaneous CSI, and the bandwidth allocated by the BS.

According to the knowledge of traditional statistical delay-bounded QoS provisioning theory [2], to maximize the effective capacity of the \( i \)th uplink, the relationship between the power allocation \( (P_i) \) for the \( i \)th user and the bandwidth \( (B_i) \) for the \( i \)th uplink needs to satisfy

\[
P_i = \begin{cases} 
\frac{1}{\gamma_i} - \frac{1}{\gamma_{i(0,i)}}, & \text{if } \gamma_i \geq \gamma_{i(0,i)}; \\
0, & \text{if } \gamma_i < \gamma_{i(0,i)},
\end{cases}
\]

where \( \gamma_{i(0,i)} \) is the cut-off SNR threshold corresponding to the \( i \)th uplink and can be determined by substituting Eq. (11) into \( \mathbb{E}_{\gamma_i} \{ P_i \} = \mathcal{P}_i \). Based on the analytical relationship between the power allocation and the bandwidth allocation specified by Eq. (11), although each UE independently determines its power allocation, the BS can coordinate the bandwidth allocations for all UEs to maximize the aggregate effective capacity supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning. Also, based on Eq. (11) and \( \mathbb{E}_{\gamma_i} \{ P_i \} = \mathcal{P}_i \), \( \gamma_{i(0,i)} \) is almost independent with \( B_i \), and thus can be treated as constant for different values of \( B_i \).

Then, plugging Eq. (11) into the objective function of problem **P1** given in Eq. (7), we obtain:

\[
\arg \min_{B, 1 \leq i \leq N} \left\{ \mathbb{E}_{\gamma} \left[ \prod_{i=1}^{N} \left( 1 + P_i \gamma_i \right)^{-\beta_i} \right] \right\}
\]

\[
= \arg \min_{B, 1 \leq i \leq N} \left\{ \mathbb{E}_{\gamma} \left[ \prod_{i=1}^{N} \left( \frac{\gamma_i}{\gamma_{i(0,i)}} \right)^{-\beta_i} \right] \right\}
\]

where we have \( \gamma_i \geq \gamma_{i(0,i)} \). Because \( [-\beta_i/(\beta_i + 1)] \) monotonically decreases as \( B_i \) increases, the optimal bandwidth allocation corresponding to the solutions of Eq. (12) need to satisfy the constraint of \( \sum_{i=1}^{N} B_i = B \).

Thus, we can convert problem **P1** into its equivalent optimization problem **P2**, which is formulated as follows [14]:

**P2:**

\[
\arg \min_{B, 1 \leq i \leq N} \left\{ \mathbb{E}_{\gamma} \left[ \prod_{i=1}^{N} \left( \alpha_i \right)^{-\beta_i} \right] \right\}
\]

\[
\text{s.t. : } \sum_{i=1}^{N} B_i = B,
\]

where

\[
\left\{ \begin{array}{l}
\alpha_i = \frac{\gamma_i}{\gamma_{i(0,i)}}; \\
A_i = \log_{\frac{\theta_{i(0,i)}}{\theta_{T,i}}} \end{array} \right.
\]

are two new variables introduced to simplify the following mathematical derivations.

Solving problem **P2**, we can derive the optimal joint bandwidth and power allocations scheme, which is given by the following Theorem 1, for uplink transmissions supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning in 5G mobile wireless networks.

**Theorem 1:** The optimal joint bandwidth and power allocations scheme, which is the solution to optimization problem **P2**, for uplink transmissions supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning is given by Eq. (16) (see the next page), where \( 1 \leq i \leq N \) and \( L \) is the degree of the Taylor polynomial for \( \log \alpha_i / \log \alpha_k \) with \( 1 \leq j, k \leq N \).

**Proof:** The proof is provided in Appendix A.

**Remark on Theorem 1:** Using our developed optimal joint bandwidth and power allocations scheme specified by Eq. (16) in Theorem 1, the BS allocates the bandwidth to each link based on the QoS exponents and the cut-off SNR thresholds of all uplinks. Then, the BS sends the bandwidth allocation scheme \( (B_i) \) to the \( i \)th UE using the corresponding subchannels \( (1 \leq i \leq N) \). The bandwidth allocation scheme contains the
\[
B_i = \frac{\sqrt{\pi T} \left( B + \log_2 \left( \frac{\sum_{i=1}^{N} 1}{\theta_i} \right) \right)}{\sum_{i=1}^{N} \left( \theta_i \gamma_{(0,i)} \right)} - \frac{\log 2}{\theta_i T}, \\
P_i = \begin{cases} 
\sqrt{\frac{\gamma_{(0,i)} \sum_{i=1}^{N} \left( \theta_i \gamma_{(0,i)} \right)}{B + \sum_{i=1}^{N} \frac{1}{\theta_i}}} & \gamma_i \geq \gamma_{(0,i)} \\
0, & \gamma_i < \gamma_{(0,i)} 
\end{cases}
\]

IV. PERFORMANCE EVALUATIONS

We conduct extensive simulation experiments to evaluate the performance of our developed optimal joint power and bandwidth allocation scheme for uplink transmissions supporting the decentralized heterogeneous delay-bounded QoS provisioning in 5G mobile wireless networks. Throughout our simulations, we set the entire bandwidth with \( B = 10 \text{ MHz} \), the time frame length with \( T = 1 \text{ ms} \), the fading parameter of Nakagami-\( m \) distribution with \( m = 2 \), the average transmit power constraint with \( P_i = 1 \text{ W} \) where \( 1 \leq i \leq N \), and the degree of the Taylor polynomial with \( L = 10 \).

Figure 2 depicts the optimal bandwidth allocation scheme allocated by the BS supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning versus the QoS exponents \( \theta_1 \) and \( \theta_2 \) for \( N = 2 \), \( N = 3 \), and \( N = 4 \), respectively. However, in Fig. 2(b), \( \theta_3 \) is fixed to \( 10^{-3} \); in Fig. 2(c), \( \theta_3 \) and \( \theta_4 \) are fixed to \( 10^{-3} \) and \( 10^{-2} \), respectively. As shown in Figs. 2(a), 2(b), and 2(c), the allocated bandwidth for the first UE \( (B_1) \) increases as \( \theta_1 \) increases and \( \theta_2 \) decreases. The allocated bandwidth for second UE \( (B_2) \) increases as \( \theta_2 \) increases and \( \theta_1 \) decreases. This implies that if the statistical delay-bounded QoS requirements for the uplink is very stringent, the BS allocates more bandwidth for this uplink to coordinate the effective capacities of all uplink transmissions, thus achieving the maximum aggregate effective capacity for the 5G mobile wireless networks. Figs. 2(b) and 2(c) also show that if the QoS exponents are fixed \( (\theta_3 = 10^{-3} \) and \( \theta_4 = 10^{-2} \)) the allocated bandwidth \( (B_3 \) and \( B_4) \) for the UEs (the third UE and the fourth UE) increase as QoS exponents \( (\theta_1 \) and \( \theta_2) \) of other UEs increase. Analyzing the curves of \( B_1 \) and \( B_2 \), we can observe that the BS allocates equal bandwidth \( (B_1 = B_2) \) for UEs when the UEs have the same statistical delay-bounded QoS requirements \( (\theta_1 = \theta_2) \). This verifies that the homogeneous statistical delay-bounded QoS provisioning is the special case of the heterogeneous statistical delay-bounded QoS provisioning.
optimal bandwidth allocation scheme can be used for both the heterogeneous and homogeneous statistical delay-bounded QoS provisioning.

Figure 3 shows the optimal power allocation scheme, which is determined by the UE and its allocated bandwidth, versus the QoS exponent and the instantaneous SNR. As shown in Fig. 3, when the statistical delay-bounded QoS requirement is very loose, the power allocation increases as the SNR increases. When the statistical delay-bounded QoS requirement is very stringent, the power allocation first increases and then decreases as the SNR increases. As we know, the traditional QoS-driven power allocation scheme for homogeneous statistical delay-bounded QoS provisioning is the water-filling scheme for the very loose statistical delay-bounded QoS requirements and the channel-inversion scheme for the very stringent statistical delay-bounded QoS requirements. However, when we consider the optimal joint power and bandwidth allocations scheme supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning, the optimal power allocation scheme with very stringent statistical delay-bounded QoS provisioning is not the channel-inversion scheme. The optimal power allocation is strictly concave over the instantaneous SNR with very stringent statistical delay-bounded QoS provisioning.

Figure 4 compares the aggregate effective capacities for the heterogeneous statistical delay-bounded QoS provisioning with that for the homogeneous statistical delay-bounded QoS provisioning. As shown in Fig. 4, the aggregate effective capacity with heterogeneous statistical delay-bounded QoS provisioning is larger than the aggregate effective capacity with homogeneous statistical delay-bounded QoS provisioning. The homogeneous statistical delay-bounded QoS provisioning can achieve the maximum aggregate effective capacity as the same as the heterogeneous statistical delay-bounded QoS provisioning only when the UEs have the same statistical delay-bounded QoS requirements (θ_1 = θ_2 as shown in Fig. 4). Also, when the difference value between θ_1 and θ_2 gets very large, the aggregate effective capacity of homogeneous statistical delay-bounded QoS provisioning gets very small while our heterogeneous statistical delay-bounded QoS provisioning yields the very large aggregate effective capacity.

V. CONCLUSIONS

We proposed the decentralized heterogeneous statistical QoS provisioning framework for the uplink transmissions over 5G mobile multimedia wireless networks. We built up the system model for decentralized heterogeneous statistical delay-bounded QoS provisioning in terms of aggregate effective capacity and heterogeneous QoS exponents. Using our decentralized heterogeneous statistical delay-bounded QoS provisioning model, we formulated the aggregate effective capacity maximization problem, which is a strictly convex optimization problem, but very difficult to solve. To efficiently solve this optimization problem, we developed the optimal joint bandwidth and power allocations scheme, where the BS allocates the bandwidth based on the QoS exponents of all uplinks while the mobile UEs dynamically allocate the power based on the corresponding instantaneous CSI, the corresponding QoS exponent, and the bandwidth allocated by the BS. The obtained extensive simulation results show that our developed optimal joint bandwidth and power allocations scheme significantly increases the aggregate effective capacity supporting the decentralized heterogeneous statistical delay-bounded QoS provisioning as compared with the homogeneous statistical delay-bounded QoS provisioning schemes.

APPENDIX A

PROOF OF THEOREM 1

Proof: To derive the optimal solution for problem P2, we construct the Lagrangian function, denoted by J, for the optimization problem P2 as follows:

\[ J = \mathbb{E}_\gamma \left\{ \prod_{i=1}^{N} (\alpha_i - \frac{B_i}{\gamma_i + \lambda_i}) \right\} + \lambda \left( \sum_{i=1}^{N} B_i - B \right), \quad (17) \]
where \( \lambda \) is the Lagrangian multiplier associated with the constraint specified by Eq. (14). Taking the derivative of \( J \) with respect to \( B_j \) \((1 \leq j \leq N)\) and setting its results to zero, we can obtain

\[
\frac{\partial J}{\partial B_j} = -\frac{A_j}{(B_j + A_j)^2} \mathbb{E}_{\gamma_j} \left[ (\alpha_j - \frac{B_j}{\pi_j + \alpha_j}) \log \alpha_j \right] \cdot \prod_{i=1, i \neq j}^N \mathbb{E}_{\gamma_i} \left[ (\alpha_i - \frac{B_i}{\pi_i + \alpha_i}) \right] + \lambda = 0.
\]  

(18)

In the similar way, we can also take the derivative of \( J \) with respect to \( B_k \) \((1 \leq k \leq N)\) and setting its results to zero. Thus, we have

\[
\frac{\partial J}{\partial B_k} = -\frac{A_k}{(B_k + A_k)^2} \mathbb{E}_{\gamma_k} \left[ (\alpha_k - \frac{B_k}{\pi_k + \alpha_k}) \log \alpha_k \right] \cdot \prod_{i=1, i \neq k}^N \mathbb{E}_{\gamma_i} \left[ (\alpha_i - \frac{B_i}{\pi_i + \alpha_i}) \right] + \lambda = 0.
\]  

(19)

Combining Eqs. (18) and (19) and noting that \( \frac{\partial J}{\partial B_j} = \frac{\partial J}{\partial B_k} \), we can obtain

\[
\frac{A_j}{(B_j + A_j)^2} \mathbb{E}_{(\gamma_j, \gamma_k)} \left[ (\alpha_j - \frac{B_j}{\pi_j + \alpha_j}) (\alpha_k - \frac{B_k}{\pi_k + \alpha_k}) \log \alpha_j \right] = \frac{A_k}{(B_k + A_k)^2} \mathbb{E}_{(\gamma_j, \gamma_k)} \left[ (\alpha_k - \frac{B_k}{\pi_k + \alpha_k}) (\alpha_j - \frac{B_j}{\pi_j + \alpha_j}) \log \alpha_k \right],
\]  

(20)

where \( \mathbb{E}_{(\gamma_j, \gamma_k)} \) denotes the expectation over \( \{\gamma_j, \gamma_k\} \).

Applying the Taylor series expansion [15], we have

\[
\log (\alpha_j) \approx \left( \frac{\gamma_j - \gamma_{(0,j)}}{\gamma_{(0,j)}} \right) \sum_{l=1}^L (-1)^{l-1} \left( \frac{\gamma_j - \gamma_{(0,j)}}{l} \right)^{l-1};
\]

\[
\log (\alpha_k) \approx \left( \frac{\gamma_k - \gamma_{(0,k)}}{\gamma_{(0,k)}} \right) \sum_{l=1}^L (-1)^{l-1} \left( \frac{\gamma_k - \gamma_{(0,k)}}{l} \right)^{l-1},
\]  

(21)

where \( L \) is the degree of the Taylor polynomial. Then, we can obtain:

\[
\frac{\log \alpha_j}{\log \alpha_k} \approx \left( \frac{\gamma_{(0,k)}}{\gamma_{(0,j)}} \right)^L
\]  

(22)

which enables Eq. (20) to be reduced to:

\[
\frac{A_j}{(B_j + A_j)^2} \left( \frac{\gamma_{(0,k)}}{\gamma_{(0,j)}} \right)^L = \frac{A_k}{(B_k + A_k)^2}.
\]  

(23)

Solving Eq. (23), we can obtain the relationship between \( B_j \) and \( B_k \) as follows:

\[
B_k = B_j \sqrt{\frac{A_k}{A_j} \left( \frac{\gamma_{(0,k)}}{\gamma_{(0,j)}} \right)^L} + A_j \sqrt{\frac{A_k}{A_j} \left( \frac{\gamma_{(0,k)}}{\gamma_{(0,j)}} \right)^L} - A_k.
\]  

(24)

Then, using Eqs. (14) and (24), we can derive the optimal bandwidth allocation scheme as follows:

\[
B_i = \frac{\sqrt{A_i} \left( B + \sum_{i=1}^N A_i \right)}{\sqrt{\gamma_L \sum_{i=1}^N \frac{A_i}{\gamma_{(0,i)}}}} - A_i, \quad 1 \leq i \leq N.
\]  

(25)

Substituting Eq. (25) into Eq. (11) and also using Eq. (15), we can obtain Eq. (16), where \( \gamma_{(0,i)} \) is numerically determined by Eq. (26). Therefore, Theorem 1 follows.

\[
\int_{\gamma_{(0,i)}}^{\infty} \left[ -\frac{\sqrt{\pi} \gamma_{(0,i)}}{\sqrt{\gamma_{(0,i)}} \sum_{i=1}^N \sqrt{\gamma_{(0,i)}} \frac{1}{\gamma_{(0,i)}^2}} \right] \left[ \frac{\sqrt{\pi} \gamma_{(0,i)}}{\sqrt{\gamma_{(0,i)}} \sum_{i=1}^N \sqrt{\gamma_{(0,i)}} \frac{1}{\gamma_{(0,i)}^2}} \right] - \frac{1}{\gamma_{(0,i)}^2} \right] p_r(\gamma) d\gamma = 1.
\]  

(26)