Low Complexity QoE-aware Bandwidth Allocation for Wireless Content Delivery

Harsha Chenji[†], Zygmunt J. Haas[*], Panfeng Xue[‡]
[†] School of Electrical Engineering and Computer Science, Ohio University, USA, chenji@ohio.edu
[*] School of Electrical and Computer Engineering, Cornell University, USA, zhaas@cornell.edu
[‡] Department of Computer Science, University of Texas at Dallas, USA, pxx101020@utdallas.edu

***Abstract*—Bandwidth allocation in spectrum-congested wireless content delivery networks should be performed based on the user's Quality of Experience (QoE). The relationship between QoE and network QoS is context- and user-dependent. We show that modeling the user's QoE introduces additional complexity to the bandwidth allocation problem. While recent research has proposed various ways in which complexity can be reduced, we have observed that the optimality (w.r.t. QoE) of the system is also reduced along with complexity. In this paper, the tradeoff between complexity and optimality is investigated and methods to control the tradeoff are proposed. A bandwidth allocation scheme for three different system objectives, including fairness, is formulated.**

I. Introduction

In today's wireless networks, limited bandwidth is shared between an increasingly large number of users. These users are consuming content, and a certain level of user satisfaction (QoE) is associated with the network conditions at the time of consumption. Some of the factors affecting QoE are the delay, data rate (not to be confused with "bandwidth") and the packet loss rate. These metrics are called the application layer QoS metrics, QoS_{app}. Note that QoE also depends on content specific metrics, such as video framerate and resolution, but they cannot be modified by the network. Given the same network link (characterized by QoS_{link}), each application can experience a different delay, based on the size of the data requested and the size of each packet. QoS_{link} is in turn affected by a set of PHY layer quantities QoS_{PHY}, such as the spectral efficiency, bandwidth, and the modulation/coding schemes. The amount of bandwidth allocated to the user, as well as the modulation scheme, can be controlled at the central base station or access point - as opposed to the SINR (of a mobile user) for example.

It is essential to allocate bandwidth among users based on their needs. Content is delivered over a wireless network, and users consume this content on heterogeneous devices with different screen sizes, for example. The content itself is heterogeneous, e.g. varying from real time sport video to movies to data downloads. Thus, a user perceives content differently based on the viewing device and the type of the content. That is, the equation relating QoE to QoS_{app} has coefficients that differ among users, and between types of content for the same user. This paper answers several key questions: how can the QoE of a user be modeled in terms of QoS_{app} metrics? If there are many users present, how should each user's QoE be optimized? How can fairness be ensured? From a systems perspective, what are the tradeoffs that need to be controlled? How feasible will an implementation of the solution be? Can computational complexity be reduced, in exchange for a reduction in QoE optimality?

The layout of this paper is as follows. First, we present the system model and a motivating scenario, where bandwidth is to be shared among two users. The QoE optimization problem is formulated, and it is shown this problem can be decoupled into two simpler subproblems - one of which is channel independent and the other is user dependent. These two subproblems are linked using a feasibility region Ψ that represents constraints on QoS_{app} according to channel conditions. In Section III-A, the decoupling technique is formalized, and the Rate Allocation and Bandwidth Allocation problems are formulated. Section IV shows how these two problems can be solved under different conditions: an error free regime and an error tolerating regime. Finally, Section V presents the performance evaluation.

II. State of Art

In this section, we present a survey of state of the art methods for QoE optimization in wireless networks. Typical spectral resources are bandwidth and power; most works propose and analyze joint power, subcarrier, and bitrate allocation algorithms [1]. For OFDMA systems, the bits per OFDM symbol is used in lieu of the bit rate. The phrase "subcarrier allocation" in literature sometimes refers to the allocation of a particular subcarrier to a user (based on frequency selective fading characteristics), rather than determine the total number of subcarriers to be assigned to a user (i.e., bandwidth allocation).

The authors of [2] propose a scheduling as well as resource allocation method for a CDMA system. The system utility function is the sum of users' utility functions, which are assumed to be concave w.r.t. the per-user throughput; at each scheduling instant, a rate vector is selected, so that its projection onto the gradient of the system utility function is maximized. [3] optimizes user QoE in wireless broadband networks by adjusting the DL/UL subframe ratio, as well as a priority based scheme for different traffic classes. Users are assumed to have a minimum data rate requirement. In [4], energy is saved and QoE is maximized in an OFDM system with group based mobile users by shutting down sub-channels; user QoE is related to throughput logarithmically. In [5], a unifying optimization framework for the subcarrier allocation problem (each subcarrier corresponds to a certain amount of bandwidth) is presented and solved using a Nash bargaining solution. [6] presents a theoretical framework for utility based subcarrier assignment in an OFDM based wireless network - it is shown that utility is eventually maximized if the aggregate marginal utility is maximized at each epoch ([5]).

This work differs from the above corpus as follows: 1) we consider bandwidth allocation and not power allocation; 2) the

978-1-5090-0073-9/15/$31.00 ©2015 IEEE

main focus of this paper is the computational complexity of bandwidth allocation, and how it can be reduced; 3) QoE is optimized instead of QoS, because each user perceives a QoS vector differently; 4) the adopted user utility function (the multi stimuli version of the IQX hypothesis) in our work depends not only on the rate/throughput, but also on the link BER; 5) we analyze three different system objectives; maximizing the average QoE, fairness, and equal QoE degradation. We show that by tolerating a small BER (instead of forcing a low BER by allocating more bandwidth/power or using a complex code), spectral efficiency can be significantly improved, since less bandwidth can be allocated at the same data rate to a user. Conversely, computational complexity can be reduced by allocating more spectral resources to ensure a low BER.

In our previous work [7], we proposed a genetic algorithms-based solution for wireless systems, where the set of modulation and coding schemes are pre-defined. A single optimization problem was solved, unlike the problem splitting approach adopted here. In this paper, the focus is on computational complexity and how it can be reduced. The optimization variables here are continuous, as opposed to discrete in our previous work.

III. System Model and Problem Formulation

The objective of this system is to optimize the users' QoE according to an objective defined by the network operator, but parsimoniously (w.r.t. computation resources). Reducing computational complexity is the key, especially in dynamic operating environments, where there is mobility or where the channel conditions change often. In order to achieve this goal, we first need to know how a user's QoE is modeled and whether it depends on content, context, or network related metrics. Once the QoE is modeled, we define the QoE optimization problem and provide insight into how the computational complexity can be reduced.

Network model: There are N users in the system (Figure 1). Each user i connects to an access point (AP). The bandwidth at the AP is limited to W Hz, and user i is allocated W_i Hz. The number of bits per PHY symbol of the link between user i and the AP is k_i, and E_b/N_0 is denoted as $\gamma_i = SNR_i/k_i$. Define the set of PHY layer QoS metrics as $QoS_{PHY} = \{W_i, k_i\}$. The BER of the link e_i is a function Γ of k_i and γ_i. Based on QoS_{PHY}, the link between user i and the AP can sustain a PHY layer data rate r_i with a Bit Error Rate (BER) e_i. Define a set of link QoS metrics $QoS_{link} = \{r_i, e_i\}$. The application run by user i uses packets of size B_i bits, such that the Packet Error Rate (PER) is $p_i = 1-(1-e_i)^{B_i}$. The effective data rate seen by the user at the application layer, inclusive of the PER, is $r_i(1-p_i)$. This application regularly requests data of size S_i, so the delay experienced by the user i is $d_i = \frac{S_i}{r_i(1-p_i)}$. To summarize:

$$e_i = \Gamma(k_i, \gamma_i); p_i = 1-(1-e_i)^{B_i}; d_i = \frac{S_i}{r_i(1-p_i)} \quad (1)$$

QoE-QoS model: QoE and QoS are two distinct but related quantities. A key difference is that QoS is measured using technical, network-centric terms such as delay and jitter, but QoE is measured using non-technical, user-centric terms, such as acceptability and satisfaction. The IQX hypothesis ([8]) proposes a generic relationship between QoE and QoS. It states that the change in QoE, for a change in QoS, depends on the current level of QoE (akin to a stimulus-response analogy):

$$\frac{\partial QoE}{\partial QoS} \propto -QoE \implies QoE = \alpha e^{-\beta QoS} + \Delta \quad (2)$$

Note that $\beta > 0$ for QoS metrics such as delay and packet loss ratio (smaller is better), and $\beta < 0$ for data rate (i.e., bigger is better). We adopt the multi-stimuli version ([9]) of the IQX hypothesis as the QoE-QoS model. The authors of [9] extend the IQX hypothesis to include multiple QoS parameters and show its applicability to video traffic using multiple linear regression. Equation 2 can be linearized as $log(QoE) = log(\alpha) - \beta QoS$ (Δ is omitted since it is a scaling factor). For multiple QoS variables, we have $log(QoE) = a_0 + a_1 QoS_1 + \cdots + a_n QoS_n$, so that $QoE = e^{a_0} e^{a_1 QoS_1 + \cdots + a_n QoS_n}$. We define the set of application level QoS metrics as $QoS_{app} = \{QoS_1, QoS_2, \ldots, QoS_n\}$.

These QoS stimuli could be content related (e.g., video frame rate) or network related (e.g., delay, jitter). Based on experiments conducted in [10], [11], in this paper, we assume that the QoS stimuli are network delay, packet error probability, and data rate (i.e., $QoS_{app} = \{p_i, d_i, r_i\}$). Note that QoS_{app} can be defined differently for different users and different types of content. Similarly, the QoE can be measured using a variety of metrics, but in this paper, we adopt the widely used and well known Mean Opinion Score (MOS). Therefore, for each user i, $MOS_i = e^{b_{i0}+b_{i1}p_i+b_{i2}d_i+b_{i3}r_i}$.

Define the set of constants $QoE_{coefs} = [b_{i0}, b_{i1}, b_{i2}, b_{i3}]$. We now provide two examples of how QoE_{coefs} can be obtained. The authors of [10] provide equations that relate MOS to d, p, r, for a file download: $MOS = 4.836 \cdot exp(-0.15d)$, $MOS = 5.5 \cdot exp(-20p)$, $MOS = 1.2 \cdot ln(1 \times 10^{-6}r)$. Link data rate varied from 0-10Mbps; however, 802.11n data rates range from 0-200Mbps. To overcome this mismatch, we increased the upper limit of the data rate to 200Mbps by changing the coefficients. This step can be justified by thinking of the "user" as an aggregating device to which thousands of users are connected. The resulting equation is $MOS = 1.2 \cdot ln(5 \times 10^{-8}r)$. Data points were extrapolated using this set of equations and re-fit onto the multi-stimuli IQX model using multiple linear regression. The resulting equation with $R^2 = 0.9799$ is:

$$MOS = e^{-6.8643p-0.10799d+1.1\times10^{-8}r} \quad (3)$$

A second QoE-QoS model can be found in [11]. Packet loss rates were varied for four audio codecs: G.723.1.B which has a capacity requirement of 6.4kbit/s, iLBC (15.2kbit/s), Speex (24.6kbit/s), and G.711 (64kbit/s). As in the previous case, these bit rates are much smaller than the 802.11n rates, so we increased the bit rates a thousandfold. After extrapolating the data, the following equation was obtained using multiple linear regression with $R^2 = 0.94681$:

$$MOS = e^{1.3629-1.5068p-0.10461d+3.5238\times10^{-10}r} \quad (4)$$

Architecture: The network model and the QoE-QoS model, explained above, can be integrated into a single architecture as follows. The Middleware (Figure 1) is a module that runs on the user's device. It is responsible for mapping the user QoE to network QoS metrics by determining QoE_{coefs}. The Scheduler (Figure 1) runs on the AP, and is responsible for bandwidth allocation. First, it receives QoE_{coefs} from

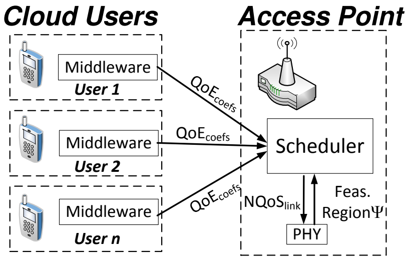


Fig. 1: The Middleware runs on the cloud user's device and builds the QoE-QoS model of the user as QoS_{app}. The Scheduler calculates the optimal rate allocation vector QoS_{link} and sends it to PHY, which computes the optimal bandwidth allocation QoS_{PHY}.

Set	Members	User	Channel
QoE_{coefs}	$b_{i0}, b_{i1}, b_{i2}, b_{i3}$	Dependent	Independent
QoS_{app}	p_i, d_i, r_i	Dependent	Dependent
QoS_{link}	r_i, e_i	Independent	Dependent
QoS_{PHY}	W_i, k_i	Independent	Dependent

TABLE I: The sets of quantities as defined in the system model, and their dependence on the user and the channel.

all N users. Next, it queries the PHY module of the CBS for a *feasibility region* Ψ, which represents system limitation (e.g., the Shannon limit on channel capacity) given total bandwidth W Hz and current channel conditions. Once the Scheduler receives Ψ, it calculates an optimal set of network QoS metrics $QoS^*_{app} = \{p_i, d_i, r_i\}$ that maximize a given system objective (Equation 5), subject to Ψ. Note that QoS_{app} is user dependent (since it depends on B_i and S_i), but for a *given user*, they can be converted to user independent network link QoS metrics QoS_{link}. So, QoS_{app}, which is both user and channel dependent (Table I), is converted to a set of link QoS metrics $QoS_{link} = \{r_i, e_i\}$, which are user independent but channel dependent (Table I). The PHY module receives QoS_{link} from the scheduler and proceeds to calculate $QoS_{PHY} = \{W_i, k_i\}$, where rate $r_i = W_i k_i$ and BER e_i can be achieved with bandwidth W_i and current channel conditions, such that $\sum W_i = W$ (Equation 6).

A. Problem Formulation and Decoupling

The main task of the AP is to allocate spectral resources (i.e., calculate $QoS_{PHY} = \{W_i, k_i\}$), such that the QoE of all the users is optimized in some way. This is called the QoE Optimization Problem. Assuming that the symbol rate for user i relates to the bandwidth W_i such that $r_i = W_i k_i$,

Problem 1. *The QoE Optimization Problem*

$$\underset{QoS_{PHY}}{\mathit{maximize}} \quad f_{OBJ}(QoE_1, QoE_2, \ldots, QoE_N) \tag{5}$$

$$\mathit{subject\ to} \quad \sum W_i = W \tag{6}$$

$$k_i < f_{CAP}(k_i) \tag{7}$$

$$\mathit{where} \quad r_i = W_i k_i \tag{8}$$

Here, f_{CAP} in Equation 7 is a capacity constraint, which limits the data rate based on channel conditions or system limitations or both. An example for f_{CAP} is the Shannon limit on channel capacity $f_{CAP}(k_i) = \log_2(1 + k_i \gamma_i)$.

We notice that the optimization variable QoS_{PHY} is channel dependent and user independent (Table I), while the objective function involves the quantities QoE_i, which are channel independent but user dependent. Therefore, every time the channel conditions change, bandwidth allocation will have to be performed. However, QoS_{app} can be expressed in terms of the network link QoS metrics $QoS_{link} = \{r_i, e_i\}$, which are *user independent* (Table I) - i.e., r_i, e_i do not involve B_i, S_i or QoE_{coefs} but are *channel dependent* (meaning, r_i, e_i depend on QoS_{PHY}, which in turn depends on W and SNR_i). This means that the optimal values of QoS_{app} are independent of the channel conditions. When the channel conditions change, *only* QoS_{PHY} need to be re-calculated, thus reducing the computational complexity.

In this spirit, we propose to split Problem 1 into two sub-problems. The two problems are "linked" using the feasibility region Ψ, which represents some system limitation on QoS_{app} given channel conditions W_i and SNR_i. The size of the feasibility region controls the "gap" of optimality that is traded off for reduced computational complexity. Whenever channel conditions change, Problem 1 can be solved with high complexity, or only QoS_{PHY} can be calculated (as one of the two sub-problems) for a small reduction in optimality. Another way complexity can be reduced is by eliminating variables. For example, more spectral resources can be allocated to ensure that $e_i \approxeq 0$, thus eliminating the e_i variable from the optimization problem, and avoiding the complex modeling between BER and SNR. This is the key idea in this paper and is illustrated in the next few paragraphs.

The Rate Allocation (RA) Problem, the first of the two sub-problems, which is solved by the Scheduler, is defined as follows:

Problem 2. *The Rate Allocation (RA) Problem*

$$\underset{QoS_{link}}{\mathit{maximize}} \quad f_{OBJ}(QoE_1, QoE_2, \ldots, QoE_N) \tag{9}$$

$$\mathit{subject\ to} \quad \sum g_j(QoS_{app}) \leq 0 \tag{10}$$

Here, instead of QoS_{PHY}, $QoS_{link} = \{r_i, e_i\}$ is the optimization variable. The search space for QoS_{link} is limited by Ψ. By specifying Ψ analytically, exhaustively listing all possible combinations of $\{r_i, e_i\}$ is avoided; this set can be very large depending on N and W. The objective (Equation 9) is designated by the network administrator. Constraints $g_j(QoS_{app})$ represent the feasibility region Ψ.

A candidate solution QoS^*_{link} is calculated by the Scheduler as a result of solving the Rate Allocation Problem. This result is sent to the PHY module (Figure 1), which is then able to solve the Bandwidth Allocation (BA) Problem:

Problem 3. *The Bandwidth Allocation (BA) Problem*

$$\underset{QoS_{PHY}}{\mathit{minimize}} \quad \sum \left[(b_{i3}(r_i - r^*_i))^2 + \left(\frac{b_{i1}}{B_i}(e_i - e^*_i) \right)^2 \right] \tag{11}$$

$$\mathit{subject\ to} \quad \sum W_i = W \tag{12}$$

$$\mathit{where} \quad e_i = \Gamma(k_i, \gamma_i) = \Gamma(k_i, SNR_i/k_i) \tag{13}$$

To recap, the system bandwidth is limited to W Hz (Equation 12). By solving Equation 11, the PHY module obtains a set of W_i, k_i. The hardware is then configured appropriately

$M=2^k$	γ (dB)	Min. SNR (dB)	Min. γ (dB)
4	12.5495	4.7712	1.7609
16	16.4608	11.7609	5.7403
64	20.8719	17.9934	10.2119
256	25.6412	24.0654	15.0345

TABLE II: The minimum required γ for a BER of 10^{-9} is shown in column 2, and the minimum required γ at a spectral efficiency of $\eta = k$ is shown in column 4.

such that user i is allotted W_i using modulation scheme k_i. We now discuss how these problems can be solved, and under what conditions.

IV. Solution

In this section, solutions to the RA and BA problems are presented. The effect of using a channel code upon γ_i is modeled for M-QAM modulation. Then, the solutions in the error free and error-tolerating regimes are presented. The key idea is that by forcing $e_i \approxeq 0$, some of the optimization variables can be eliminated, leading to reduced computational complexity in return for a small reduction in optimality. An expression is derived for k_i when $e_i \approxeq 0$, and this expression is used as a constraint in the BA problem. Three different objectives are analyzed for RA: maximizing the average MOS of all users, fair resource allocation among users such that MOS are equalized, and enforcing equal MOS degradation among users when the available bandwidth decreases.

Channel coding: We assume that (as long as $r_i < W_i \log_2(1+SNR_i)$) there exists a channel code that, irrespective of the modulation scheme, provides a coding gain that is inversely proportional to the number of bits per symbol: i.e., at higher k, the channel code provides a smaller coding gain. The system is forced to reduce k in order to reduce e_i - but doing so also reduces r_i when W_i is limited. Thus a realistic tradeoff is setup. Consider square M-QAM (QPSK for $M = 4$) where $k = \log_2 M$ is even:

$$e_i = \frac{4(\sqrt{2^{k_i}}-1)}{k_i\sqrt{2^{k_i}}} \cdot Q\left(\sqrt{\gamma_i \cdot \frac{3k_i}{(2^{k_i}-1)}}\right)$$

The required E_b/N_0 values to achieve BER of 10^{-9} are shown in column 2 of Table II. Column 3 shows the minimum SNR required to be able to use the MQAM modulation scheme, and column 4 shows the minimum required E_b/N_0 at a max spectral efficiency of k. We can see that the required coding gain is about $11dB = 12.5893$ in each case. Therefore, the coding gain (as a ratio) provided by the channel code, when used in conjunction with 2^k-QAM, is assumed to be $12.5893/k$. The new equation for the BER of M-QAM, assuming the existence of the above code is:

$$e_i = \frac{4(\sqrt{2^{k_i}}-1)}{k_i\sqrt{2^{k_i}}} \cdot Q\left(\sqrt{\frac{12.5893}{k} \times \gamma_i \times \frac{3k}{(2^k-1)}}\right) \tag{14}$$

A. Error-free Regime

Problem 1 can be decoupled into two sub-problems in the error free regime, where $p_i = e_i \approxeq 0$. Thus,

$$\ln(MOS_i) = b_{i0} + b_{i1} \cdot 0 + b_{i2}\frac{S_i}{r_i(1-0)} + b_{i3}r_i \tag{15}$$

$$\ln(MOS_i) = a_i - \frac{b_i}{r_i} + c_i r_i \ , \ b_i, c_i, r_i > 0 \tag{16}$$

where $a_i = b_{i0}, b_i = -b_{i2}S_i, c_i = b_{i3}$. Typically, MOS decreases with increasing delay and decreasing data rate, and $S_i > 0$. Therefore, $b_i, c_i, r_i > 0$. The Scheduler determines $QoS_{link} = \mathbf{r} = [r_1, r_2, \ldots, r_n]$. The Rate Allocation Problem now becomes:

$$\begin{aligned} &\underset{\mathbf{r}}{\text{maximize}} && f_{OBJ}(\mathbf{r}) && (17)\\ &\text{subject to} && \sum \delta_i r_i = W && (18)\\ &\text{where} && \delta_i = \frac{1}{\log_2(1+SNR_i)} && (19) \end{aligned}$$

Equation 19 represents the feasibility region Ψ. The solution $\mathbf{r}^*$ is sent to the PHY module, which allocates spectral resources in order to effect the required data rate. In the Bandwidth Allocation Problem, the PHY module determines a feasible data rate vector $\hat{\mathbf{r}}$:

$$\begin{aligned} &\underset{\hat{\mathbf{r}}}{\text{minimize}} && \|\hat{\mathbf{r}} - \mathbf{r}^*\|^2 && (20)\\ &\text{subject to} && \sum W_i = W && (21)\\ & && e_i = 0 && (22) \end{aligned}$$

The first and second derivatives of MOS_i can be obtained in closed form: $MOS_i' = MOS_i\left(c_i + \frac{b_i}{r_i^2}\right)$ and $MOS_i'' = MOS_i\left(c_i + \frac{b_i}{r_i^2}\right)^2 - MOS_i\frac{2b_i}{r_i^3}$. A test of convexity for MOS_i: $MOS_i'' > 0$ over $(0, \infty) \implies c_i^2 r^4 + 2b_i c_i r^2 - 2b_i r + b_i^2 > 0$. If the function QoE_i is convex, then f_{OBJ} may be convex (e.g., $f_{OBJ} = \sum QoE_i/N$) - this fact can be used to reduce complexity by obtaining closed form expressions for Hessian elements during optimization.

1) Maximizing the Average MOS: In this scenario, f_{OBJ} returns the average MOS:

$$\underset{\mathbf{r}}{\text{maximize}} \quad f_{OBJ}(r) = \frac{1}{N}\sum_{i=1}^{N} e^{a_i - \frac{b_i}{r_i} + c_i r_i} \tag{23}$$

$$\text{subject to} \quad c(r) = \sum_{i=1}^{N} \delta_i r_i = W$$

If f_{OBJ} is concave, then the above problem becomes a concave maximization problem. It is useful to examine whether the Jacobian and the Hessian are sparse, since linear algebra computations can be sped up.

$$\nabla_i f(r) = \frac{MOS_i'}{N} \qquad \nabla_i c(r) = \delta_i$$

$$\nabla_{ij}^2 f(r) = \begin{cases} \frac{MOS_i''}{N} & \text{if } i = j \\ 0 \text{ otherwise} \end{cases} \qquad \nabla_{ij}^2 c(r) = 0$$

$$H_{ij}(r, \lambda) = MOS_i''/N \text{ if } i = j, \qquad 0 \text{ otherwise}$$

Thus, there are only N non-zero elements in the Hessian.

2) Fair Rate Allocation: The ratio of standard deviation to mean is minimized, to ensure equal MOS among all users.

$$\min_{\mathbf{r}} \quad f(r) = \frac{\sigma(MOS_1, MOS_2, \ldots, MOS_n)}{\mu(MOS_1, MOS_2, \ldots, MOS_n)} \tag{24}$$

$$\text{s.t.} \quad c(r)\sum_{i=1}^{N} \delta_i r_i = W$$

The derivatives are as follows:

$$\sigma^2 = \frac{1}{N-1}\sum MOS_i^2 - \frac{1}{N(N-1)}\left(\sum MOS_i\right)^2$$

$$f(r) = \frac{\sqrt{(N-1)^{-1}\sum MOS_i^2 - N^{-1}(N-1)^{-1}(\sum MOS_i)^2}}{N^{-1}\sum MOS_i}$$

Let $X = \sum MOS_i$ and $Y = \sum MOS_i^2$

$$f(r) = \sqrt{\frac{N^2 Y}{(N-1)X^2} - \frac{N}{N-1}}$$

$$\nabla_i f = \frac{1}{2f(r)}\frac{N^2}{N-1}\frac{2X \cdot MOS_i \cdot MOS_i' - 2Y \cdot MOS_i'}{X^3}$$

The rest of the equations are omitted for brevity - equations concerning $c(r)$ are identical to the previous section.

3) Equal MOS Degradation: Suppose that the channel conditions have degraded, and the system can no longer provide users with the required data rate. Total available bandwidth has reduced to $\epsilon W, 0 < \epsilon < 1$, and each user's data rate has to be reduced such that the MOS is reduced equally. Let the previous MOS for user i be $\overline{MOS_i}$. Our task is to calculate MOS_i such that the decrease in MOS is equal for all users, while still adhering to the bandwidth constraint:

$$\min_{\mathbf{r}} \quad f(r) = \frac{\sigma\{MOS_i - \overline{MOS_i}\}}{\mu\{MOS_i\}} \tag{25}$$

$$\text{s.t.} \quad c(r) = \sum_{i=1}^{N}\delta_i r_i = \epsilon W$$

The analysis for this case is very similar to the previous case (save for $c(r) = \epsilon W$ instead of $c(r) = W$), and is thus omitted for brevity.

4) Bandwidth Allocation: In order to solve Problem 20, we need an equation that relates the data rate to the allocated bandwidth. Typically these are two independent variables, but a constraint in the high SNR regime is that the BER is zero. Using Equation 14:

$$e_i = \frac{4(\sqrt{2^{k_i}}-1)}{k_i\sqrt{2^{k_i}}} \cdot Q\left(\sqrt{\frac{12.5893 \times \gamma_i \times 3}{(2^{k_i}-1)}}\right)$$

$$Q(x) = \frac{1}{12}e^{-\frac{1}{2}x^2} + \frac{1}{4}e^{-\frac{2}{3}x^2} \quad \text{Let } y = e^{\frac{12.5893\gamma_i \times 3}{(2^{k_i}-1)}}$$

$$\implies e_i = \frac{4(\sqrt{2^{k_i}}-1)}{k_i\sqrt{2^{k_i}}}\left(\frac{1}{12}y^{-1/2} + \frac{1}{4}y^{-2/3}\right)$$

In the error free regime, BER should be zero for practical purposes:

$$e_i \approxeq 10^{-9} \implies (\frac{1}{12\sqrt{y}} + \frac{1}{4\sqrt[3]{y^2}}) \approxeq 10^{-9} \tag{26}$$

Solving, we obtain $ln(y^*) = 36.5$, using MATLAB's `vpasolve` for example. Now,

$$\ln y^* > 36.5 \implies SNR_i - 0.9664k_i(2^{k_i}-1) > 0 \tag{27}$$

We now cast the optimization problem as follows, with $N+1$ constraints, enabling us to calculate optimal k_i:

$$\min_{W_i,k_i} \quad f(w,k) = \sum_{i=1}^{N}(W_i k_i - r_i^*)^2 \tag{28}$$

$$\text{s.t.} \quad c_1(w,k) = \sum W_i = W \tag{29}$$

$$c_{i+1}(w,k) = SNR_i - 0.9664k_i(2^{k_i}-1) > 0 \tag{30}$$

$$\text{limit} \quad 1 \leq k_i < \log_2(1 + SNR_i) \tag{31}$$

We have:

$$\frac{\partial f}{\partial w_i} = 2k_i(W_i k_i - r_i^*) \text{ and } \frac{\partial f}{\partial k_i} = 2W_i(W_i k_i - r_i^*)$$

$$\frac{\partial c_1}{\partial w_i} = 1 \text{ and } \frac{\partial c_1}{\partial k_i} = 0 \quad \frac{\partial^2 f}{\partial w_i^2} = 2k_i^2 \text{ and } \frac{\partial^2 f}{\partial k_i^2} = 2W_i^2$$

$$\frac{\partial^2 f}{\partial k_i \partial w_i} = \frac{\partial^2 f}{\partial w_i \partial k_i} = 4W_i k_i - 2r_i^*$$

$$\frac{\partial c_{i+1}}{\partial k_i} = -0.9664(2^{k_i}-1) - 0.6698k_i 2^{k_i}$$

$$\frac{\partial^2 c_{i+1}}{\partial k_i^2} = -1.3397 \cdot 2^{k_i} - 0.4643k_i 2^{k_i}$$

All other terms are zero - making the Hessian and Jacobian sparse.

B. Error-Tolerating Regime

In this section we analyze the case when $p_i \neq 0, e_i \neq 0$. Problem 1 can be decoupled when k_i is fixed, but not otherwise.

Fixed k_i: If k_i is known, then e_i can be determined using Equation 14. Let e_i^* denote the value of e_i at $k_i = k_i^*$. The MOS equation becomes:

$$MOS_i = e^{b_{i0}+b_{i1}(1-(1-e_i^*)_i^B)+\frac{b_{i2}S_i}{r_i(1-e_i^*)_i^B}+b_{i3}r_i}$$

This equation is identical to the equation for the "error-free" regime, with $a_i = b_{i0} + b_{i1}(1-(1-e_i^*)_i^B)$, $b_i = \frac{-b_{i2}S_i}{(1-e_i^*)_i^B}$ and $c_i = b_{i3}$. Thus, the analysis from the previous section is applicable here, and $\mathbf{r}^*$ can be obtained. The bandwidth allocation problem becomes:

$$\underset{W_i}{\text{minimize}} \quad \sum_{i=1}^{N}(W_i k_i^* - r_i^*)^2 \tag{32}$$

$$\text{subject to} \quad \sum W_i = W \tag{33}$$

Variable k_i: The feasibility region Ψ represents constraints on r_i, e_i, and e_i depends on k_i. Therefore, Ψ cannot be expressed independent of k_i. Because the Rate Allocation problem is designed to be channel independent, decoupling Problem 1 is not possible in this scenario. The QoE Optimization problem is posed as follows:

$$\underset{W_i,k_i}{\text{maximize}} \quad f_{OBJ}(MOS_1, MOS_2, \ldots, MOS_N) \tag{34}$$

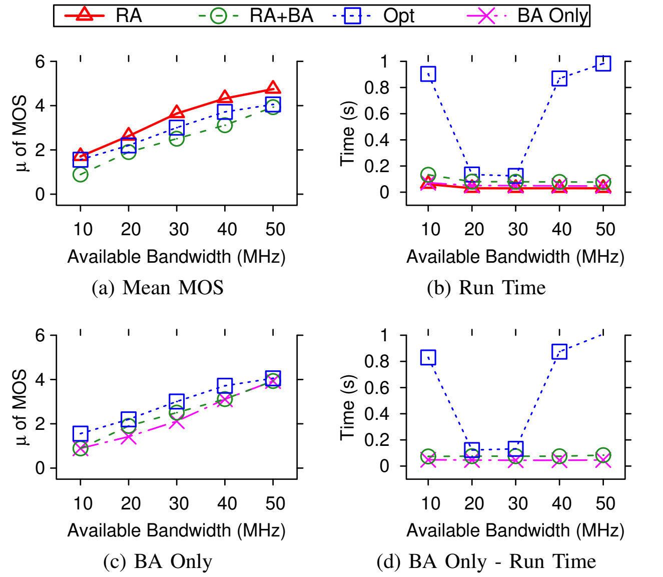
$$\text{subject to} \quad \sum W_i = W \tag{35}$$

$$\text{limit} \quad k_i < \log_2(1 + k_i\gamma_i) \tag{36}$$

where $r_i = W_i k_i$. Solving this problem requires the most computational resources, but is optimal.

V. Evaluation

In this section, we present the performance evaluation of the RA and BA problems in the error free (RA,BA) and error tolerating regimes (Opt). Results were obtained using simulation, with the KNitro suite of nonlinear optimization algorithms. While the first derivatives for all problems were provided analytically, the Hessians for only RA and Opt were estimated using a quasi-Newton BFGS method. For each optimization problem, results were obtained over 50 randomly generated initial start points. The default number of users is 4.



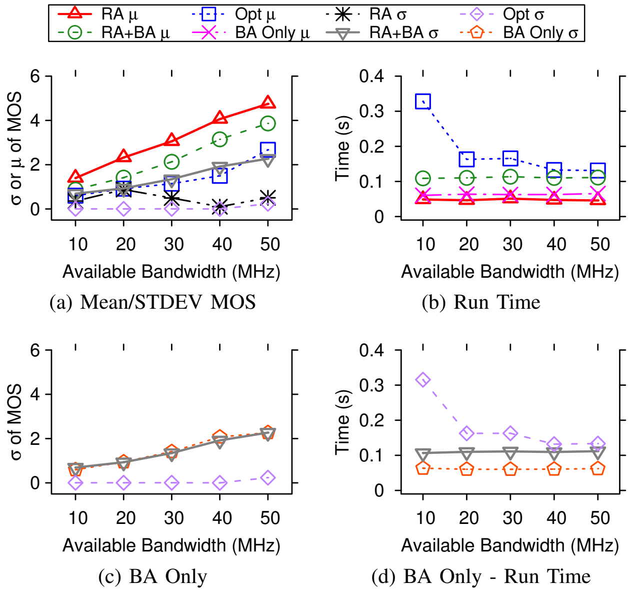
(a) Mean MOS (b) Run Time

(c) BA Only (d) BA Only - Run Time

Fig. 2: Maximizing the average MOS: (a) comparison of RA, BA, and OPT; (b) run time for (a); (c) comparing the results of BA based on r^* obtained at $W = 50MHz$ to RA+BA and Opt at $W = 10 \ldots 40$MHz; (d) run time for (c)

Since QoE_{coefs} were obtained for only two users as discussed previously, additional users were created by duplicating the profile of a user chosen from the two. SNR_i was 40dB, S_i was 60 Mbits, and B_i was 4000. It is worthwhile to note that in a separate experiment, computing the Q function or its first/second derivatives, frequently seen in the expression for $H(x, \lambda)$, introduced large computational overhead, taking up almost 30% of the total computation time. This cost can be avoided for the RA and BA methods, since additional spectral resources are used to ensure $e_i = 1$E-9 and thus eliminating the Q function (and introducing a simpler constraint instead).

Maximizing Average MOS: The results for the objective of maximizing the average MOS of all users is shown in Figure 2. The performance is compared across three variants: Rate Allocation in the error free regime (RA), Bandwidth Allocation (BA) based on **r*** from RA, and optimization in the error-tolerating regime (Opt) with non-fixed k_i. As expected, mean MOS increases with increasing bandwidth (Figure 2a). The calculated mean MOS is high (RA μ), because the feasibility region is determined by the Shannon limit only. The realized mean MOS is lower (BA μ), since spectral resources are spent on ensuring that the BER is $< 10^{-9}$, and also because the bandwidth is limited. The optimal mean MOS (Opt μ) is in between the previous two values, because it allows a small BER in exchange for lesser bandwidth. σ of resulting MOS from these three methods is not shown, because only the mean is maximized in the objective. However, this performance of Opt comes at a cost (Figure 2b). Solving RA or BA alone takes about a tenth of the time as Opt - but note that both RA (which yields QoS^*_{link}) and BA (which uses QoS^*_{link} to yield W^*, k^*) need to be solved in order to obtain a W^* and k^*. We see that Opt takes lesser time at lower bandwidths, but takes about 33% more time at higher bandwidths. Therefore, it is preferable to use RA+BA at lower bandwidths (BA μ vs Opt in Figure 2a), but Opt has higher average MOS at higher bandwidths, with higher computation cost.



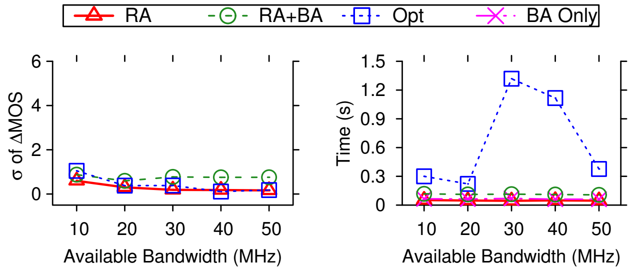
(a) Mean/STDEV MOS (b) Run Time

(c) BA Only (d) BA Only - Run Time

Fig. 3: Fair Resource Allocation (minimizing σ/μ): (a) comparison of RA, BA, and OPT; (b) run time for (a); (c) comparing the results of BA based on r^* obtained at $W = 50MHz$ to RA+BA and Opt at $W = 10 \ldots 40$MHz;(d) run time for (c)

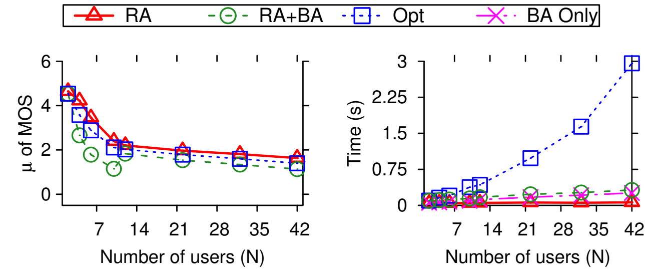
However, the RA problem need not be solved every time the bandwidth changes, thus saving on computation. Figure 2c shows the effect of solving only the BA problem, based on a **r*** obtained for a Ψ corresponding to $W = 50MHz$ ($\epsilon = 1$ on the X axis). The lines for Opt and RA+BA are identical to the corresponding lines in Figure 2a. We see that BA only incurs a small penalty at lower bandwidths and is identical to RA+BA at $\epsilon = 1$, as expected. Therefore, it is possible to solve only BA and use a previously solved RA at a different bandwidth. Additionally, the computation time for BA is very low (Figure 2d), when compared to RA+BA and Opt. Clearly, Opt always has the highest mean MOS, but also the highest computation time. Therefore, it is shown that decoupling the QoE Optimization problem is beneficial and leads to reduced computational complexity in exchange for a small reduction in optimality.

Fair Resource Allocation: The results for minimizing σ/μ of users' MOS are shown in Figure 3. RA generates a vector **r*** with the highest μ and a low σ (RA μ and RA σ in Figure 3a). However, the realized MOS values have a lower μ and higher σ (BA σ), since the BA problem uses a least squares approach. Compared to the previous section (maximizing μ), increasing r_i does not guarantee a better objective value (lower σ). Note that the objective here is the ratio σ/μ and not σ only. Opt has the lowest ratio - primarily because of a low σ rather than a high μ (Opt μ). Note that this value of μ is lesser than the value in Figure 2a. As before, RA $<$ Opt $<$ BA for performance when σ/μ is compared. However, Opt needs comparable resources (Figure 3b) when minimizing σ/μ as opposed to maximizing $\sum \mu$. In spite of the complexity of the objective function as well as the CPU cycles required to compute the Jacobian and the gradient at each iteration, the CPU time in practice also depends on the choice of the initial points at each iteration. Opt needs almost twice as much resources, when compared to RA+BA (Figure 2b).



(a) STDEV of ΔMOS (b) Run Time

Fig. 4: Equal MOS degradation when available bandwidth is reduced from $W = 50MHz$ to $10 \ldots 40$MHz, based on $\overline{MOS}$ obtained at $W = 50MHz$ using Opt: (a) σ of ΔMOS for RA, BA, Opt; (b) run time for (a).



(a) Mean MOS (b) Run Time

Fig. 5: Effect of increasing the number of users when available bandwidth is $W = 50MHz$: (a) μ of MOS for RA, BA, Opt; (b) run time for (a)

As before, we examine the effect of performing only BA, based on $\mathbf{r}^*$ at a different bandwidth (50MHz as above). Figure 3c shows a large deviation of BA from the expected value ($\sigma = 0$). BA yields only a slightly higher σ as compared to RA+BA, while Opt has the best performance with $\sigma \approx 0$. But BA has the lowest complexity (Figure 3d), lower than RA+BA, and much lower than Opt. Given that BA increases σ only by a small amount, performing BA only is a feasible choice (comapred to RA+BA) when performing fair resource allocation with emphasis on low complexity.

Equal Degradation: The results for equal MOS degradation are shown in Figure 4a. Here, the performance metric is $\frac{\sigma\{\Delta MOS\}}{\mu\{MOS\}}$, where $\Delta MOS = MOS_i - \overline{MOS}_i$. The vector of previous MOS values $\overline{MOS}_i$ was obtained using Opt at $W = 50MHz$ with the objective of maximizing the average MOS. We can see that RA and Opt always ensure that the ratio σ/μ is the smallest, but BA introduces a gap from optimality. Since $\overline{MOS}_i$ was obtained using Opt and not RA+BA (at $\epsilon = 1$), the σ for RA+BA is non-zero at $\epsilon = 1$. The absolute MOS values are not shown in Figure 4a, but RA+BA was able to obtain a higher MOS (but also disproportionately higher $\sigma\{\Delta MOS\}$), making the ratio $\frac{\sigma\{\Delta MOS\}}{\mu\{MOS\}}$ smaller for Opt. As always, this performance comes at a price, as seen in Figure 4b. If the network operator is willing to tolerate a small non-negligible variance in user MOS degradation, instead of zero variance, then the RA+BA solver can be used since it has a higher average MOS, less uniform degradation, but also greatly reduced complexity.

Scalability: The results for increasing the number of users is shown in Figure 5, where the objective was to maximize the average MOS. As expected, average MOS decreases with an increase in the number of users (Figure 5a), but tends to flatten after a certain number of users (˜10 users). Opt always achieves a higher MOS than RA+BA. We can see that the computation time increases almost quadratically with N (Figure 5b), as compared to linear for RA and BA. This is because of the number of non-zero elements in the respective Hessians. RA has a $N \times N$ Hessian with N non-zero elements; BA has a $2N \times 2N$ Hessian with $2N$ non-zero elements; but Opt has a $2N \times 2N$ Hessian with computationally complex expressions involving the Q function. This complexity worsens as the number of users increases.

VI. Conclusions

In this paper, we have formulated the QoE optimization problem over a one hop wireless content delivery network, and shown that computation complexity can be traded for optimality if certain conditions are satisfied. The QoE of a user is mapped to network QoS metrics using the multi-stimuli IQX model. The problem is decoupled into two sub-problems linked together by a feasibility region Ψ. When the available bandwidth changes, only one of the two sub-problems needs to be solved for a slightly less optimal bandwidth allocation. Similarly, complexity can also be reduced by allocating additional spectral resources to ensure that the link BER is zero for practical purposes, thus avoiding complex theoretical modeling of the link BER for a given modulation scheme. Results are obtained using simulation and verified that computation can indeed be traded off for optimality.

Acknowledgment: This work has been supported in part by NSF under the following grant numbers: CNS-1040689, ECCS-1308208 and CNS-1352880.

References

[1] Y.-F. Liu and Y.-H. Dai. On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems. *IEEE Transactions on Signal Processing*, February 2014.

[2] V. G. Subramanian, R. A. Berry, and R. Agrawal. Joint scheduling and resource allocation in CDMA systems. *IEEE Transactions on Information Theory*, May 2010.

[3] T. M. Nguyen, T. Yim, Y. Jeon, Y. Kyung, and J. Park. QoS-aware dynamic resource allocation for wireless broadband access networks. *EURASIP Journal on Wireless Communications and Networking*, June 2014.

[4] Y. Zhang, H. Long, Y. Peng, A. V. Vasilakos, and W. Wang. QoE and energy efficiency aware resource allocation for OFDM systems in group mobility environments. *International Journal of Communication Systems*, April 2013.

[5] H. Xu and B. Li. Efficient resource allocation with flexible channel cooperation in OFDMA cognitive radio networks.

[6] G. Song and Y. Li. Cross-layer optimization for OFDM wireless networks-part i: theoretical framework. *IEEE Transactions on Wireless Communications*, March 2005.

[7] H. Chenji and Z. J. Haas. Enhancement of Wireless Bandwidth Utilization through User's QoE. WCNC '15.

[8] M. Fiedler, T. Hossfeld, and P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *Network, IEEE*, March 2010.

[9] S. Aroussi, T. Bouabana-Tebibel, and A. Mellouk. Empirical QoE/QoS correlation model based on multiple parameters for VoD flows. Globecom '12.

[10] J. Shaikh, M. Fiedler, and D. Collange. Quality of experience from user and network perspectives. *annals of telecommunications - annales des tlcommunications*, February 2010.

[11] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer. MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication. *Advances in Multimedia*, July 2007.