Enhancement of Wireless Bandwidth Utilization through User's QoE

Harsha Chenji*
School of Electrical Engineering and Computer Science
Ohio University
chenji@ohio.edu

Zygmunt J. Haas
Wireless Networks Lab
Cornell University
haas@ece.cornell.edu

***Abstract*—Quality of Experience (QoE) measures a user's satisfaction with a service delivery. However QoE is a very subjective measure and is context dependent, making it difficult for a service provider to estimate and optimize user's QoE. In this paper, we look at how the provider can maximize QoE by optimizing wireless bandwidth allocation, especially for mobile cloud applications. The multi-stimuli version of the "IQX" hypothesis is used to model the QoE of a user, and this model is used in formulation of a nonlinear optimization problem, which is solved using NSGA-II. Simulations using realistic parameters based on 802.11n demonstrate a reduction in the required bandwidth by as much as 33% (i.e., more users can be accommodated by the system), while maintaining the same level of QoE. Our evolutionary-algorithm-based approach is able to discover the optimal bandwidth allocation. The problem of equalizing user QoE is explored and a tradeoff between QoE and fairness is studied, while being characterized using a Pareto front.**

I. Introduction

Widespread access to mobile wireless platforms has caused a paradigm change in multimedia consumption. With the introduction of high-capacity standards (e.g., LTE), an increasing amount of content is being streamed over last hop wireless networks (LHWNs). However, the quality of these wireless links are constantly changing, affecting the quality of service (QoS), and consequently, the QoE[1]. Low QoE could result in the user's dissatisfaction with the provided service. As a result, both content and network providers nowadays are looking at how QoE can be maximized for users. Unfortunately, QoE of a user is a very subjective measure due to the heterogeneity in user perception, the context in which the service is delivered, as well as the type of content being delivered - making a generic solution difficult to formulate. With the growth of cloud radio access networks (C-RANs), the cloud provider has the ability to control the user's LHWN device, and therefore, the network QoS. Challenges to this approach include finite spectral resources like bandwidth in the LHWN, as well as rapidly changing channel conditions. It is in such a resource constrained environment that the cloud provider has to operate, while maximizing user's QoE.

A survey of recent research in scheduling and resource allocation for optimal content delivery shows that more often than not, the network does not take into account the QoE requirements of each user. Indeed, for two scenarios with the same network conditions: 1) user QoE varies with the type of content being served, and 2) human factors involved in multimedia perception vary from user to user. Thus, the same type of content could be perceived differently by different users. For example, in an audio delivery system, depending on the language of the content, a user may prefer a lower level of background noise while another user may prefer higher speech intelligibility. QoS requirements for a sports video stream (e.g., low delay) is different than that for a movie (e.g., high bit rate). The LTE standard specifies that the resources should be allocated in decreasing order of connection priority - and if no resources remain, the connection waits. In contrast, we adopt the approach of allocating resources based on the user/application's needs (and not based on priority), such that QoE is optimized across all users. Therefore, the highly subjective nature of multimedia content necessitates a new channel resource allocation scheme at the LHWN device that is aware of the QoS requirements of a cloud user.

This paper straddles two researched areas: modeling the QoE of a user using network QoS metrics, and optimal channel resource allocation. We address two main research problems in this paper. How can QoE of a user be modeled using network QoS metrics? How can channel resources be allocated, such that user QoE is optimized? We propose a general model that utilizes utility functions to model a user's QoE. By way of an example, our model is based on three network QoS metrics: delay, packet loss ratio, and data rate. A feasibility region comprises the set of network QoS metrics achieved at feasible physical layer operating points; this region changes based on allocated bandwidth, and the modulation and coding scheme that is used. It is over this region that optimization is performed and bandwidth is allocated to each user, such that QoE is maximized. We consider multiple objectives: maximizing average QoE, maximizing the minimum QoE, and equalizing user QoE. An example that uses a realistic 802.11n feasibility region is provided. Two QoE-QoS models are built using results from recent research: the QoE-QoS relationship can be treated as that of a response-stimulus, and the QoE decreases exponentially with decrease in QoS. The optimal QoS value can be achieved by configuring the PHY layer appropriately. The contributions of this paper are as follows: 1) A spectral resource allocation scheme that is aware of users' QoS needs; 2) A problem formulation that is able to incorporate multiple single objectives, as well as dual objectives; and 3) An example scenario that proposes QoE-QoS models using results from recent research, that operates in a realistic 802.11n environment.

II. Related Work

In this section we review some recent research in modeling user QoE, as well as research that addresses channel resource allocation. Finally we place our work in the context of other research and motivate our approach. QoE is defined as "the

* This research was performed while at the University of Texas at Dallas.

[1] In this paper, we refer to network-level parameters as QoS parameters, and to application-level parameters as QoE parameters.

overall acceptability of an application or service, as perceived subjectively by the end-user" (ITU-T Rec. P.10/G.100). In addition to metrics like network QoS ([1]), QoE is influenced by economic, environmental, and sociological aspects ([2]) - collectively referred to as the user's context. The relationship between traditional network QoS parameters and QoE has been studied extensively ([3]). A popular QoE measure is the Mean Opinion Score (MOS) method (ITU-T Rec. P.10), which requires users to rate their experience on a five-point scale. However MOS is a subjective measure, is cost intensive, and difficult to perform in real time. Recently, user engagement ([4]) has emerged as a QoE measure.

Correlation-based approaches predict QoE metrics such as MOS, using objective, measurable metrics like the peak signal to noise ratio (PSNR) ([5]), application level QoS metrics (AQoS) like the average video bit rate ([6]), network level QoS metrics (NQoS) like jitter ([7]), or a combination ([8]). The end result in these approaches is a model ([9]), prediction framework ([10]) or an analytical formula ([11]) that calculates QoE. Causality-based approaches considers QoS-QoE as a stimulus-response relationship in humans. Psychophysics research suggests that the nature of this relationship is logarithmic ([12]) or power law ([13]). The authors of [14] derive an exponential relationship between QoS and QoE. An interesting observation ([15]) is that sometimes, both exponential and logarithmic relationships show strong correlation based on the choice of QoS metric. Correlation based approaches are sometimes specific to the data used to perform curve fitting ([16]). Methods which map PSNR to MOS have been shown to be inaccurate in terms of judging perceived visual quality ([17]). Some papers ([18], [4]) have looked into how the correlation changes with context (live vs. recorded video, free vs. paid).

In [19], a multi-application cross layer rate allocation scheme is proposed, MOS models for different types of applications are derived, and an optimization problem that allocates a transmission policy to each user is proposed. [20] discusses OFDM systems where bandwidth and power are allocated to maximize a system utility function which represents user QoS. However, the utility function is a function of throughput and queue length only; other objectives like minimum QoS and equalizing QoS are not considered. In [21], an OFDM system that delivers MPEG-4 streams is optimized through video packet management. [22] considers joint subcarrier and power allocation in multi-user OFDM systems, but the objective of the formulated problem is to minimize power consumption while providing a minimum MOS to users.

We model the user utility (QoE) as a function of *both* data rate and bandwidth, and not as a function of data rate only ([23]). Based on the modulation scheme chosen, the BER is calculated, which in turn affects the throughput and delay. As a consequence, sometimes a low rate/low BER policy can translate to a *higher* user utility than a high rate/high BER policy, based on the content. Power constraints are left as future work, as are channels with different fading properties. Our scheme is applicable to both cellular and WiFi networks. The work closest to our paper, in the sense of a cross-layer resource allocation scheme is [19]. However, the goal in [19] is to maximize the average MOS or throughput (but not ensure fairness) through rate allocation (not bandwidth allocation).

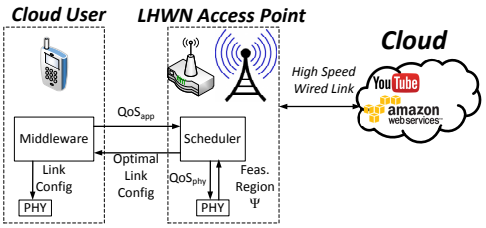


Fig. 1: An LHWN access point is connected to the cloud through a high speed link. The Middleware runs on the cloud user's device and builds the QoE-QoS model of the user as QoS_{app}. The Scheduler allocates bandwidth to each user based on QoS_{app}, and computes the optimal PHY operating point for PHY interfaces.

III. Problem Formulation and Example

In this section we introduce the terminology and system model. Functions of the Middleware and Scheduler are defined, as well as the utility functions used to model user's QoE, and the PHY feasibility region. The problem is formulated and a solution is proposed. A working example is provided for clarity.

We consider a scenario where a central device at the edge of the cloud ("Scheduler" in Figure 1) is serving multiple cloud users ("Middleware" in Figure 1) which are consuming different types of content. The user's anticipated satisfaction with each content is estimated using the mean opinion score (MOS) function, which is a number between 1 and 5. These scores are highly subjective and are content as well as user dependent. The set of parameters which characterize the wireless channel, like delay (d), packet error rate (p) and data rate (r), are collectively denoted as QoS_{phy}. A set of *utility functions* U_i model the MOS for each user in terms of all the QoS_{phy} metrics: $MOS_i = U_i(d, p, r)$, i=1...N. For each user, for each type of content (news, sports etc.), the coefficients in the utility function U_i are different. The Middleware profiles a user and calculates the coefficients in U_i, denoted as QoS_{app} (Figure 1). The PHY (Figure 1) is responsible for assessing the channel state and constructing a feasibility region Ψ, which represents the combinations of d, p, r that can be realized given current channel conditions. Once the Scheduler receives Ψ as well as QoS_{app} from different users' Middlewares, it calculates an optimal bandwidth allocation as well as PHY modulation/coding scheme such that $\sum MOS_i$ is maximized (this objective can be modified, as discussed below). Each PHY is then configured to operate at this operating point. This process is repeated whenever required, to compute new allocations.

In order to compute $U_i(d, p, r)$ we adopt the multi-stimuli version ([24]) of the IQX hypothesis ([14]) as the QoE-QoS model. The IQX hypothesis states that the change in QoE (in this case, measured as the MOS), for a change in QoS, depends on the current level of QoE:

$$\frac{\partial QoE}{\partial QoS} \propto -QoE \implies QoE = \alpha e^{-\beta QoS} + \gamma \qquad (1)$$

The authors of [24] extend the IQX hypothesis to include multiple QoS parameters, and show its applicability to video traffic, using multiple linear regression. Equation 1 can be linearized as $log(QoE) = log(\alpha) - \beta QoS$, because γ can

be omitted since it is a scaling factor. As shown in [24], for multiple QoS variables, $log(QoE) = a_0 + a_1 QoS_1 + \cdots + a_n QoS_n$ so that $QoE = e^{a_0} e^{a_1 QoS_1 + \cdots + a_n QoS_n}$. As in [15], we adopt delay, packet error probability and data rate as QoS parameters. Since QoE is measured as the MOS,$MOS_i = e^{a_0^i} e^{a_1^i d + a_2^i p + a_3^i r}$ for each user i. The coefficients $a_0^i \ldots a_3^i$ are part of QoS_{app} and can be obtained through experimentation by the Middleware, as discussed in the next section. We now relate the QoS_{phy} parameters to the spectral resource, namely the bandwidth W measured in Hz. The feasibility region Ψ can be constructed (i.e., d, p, r can be determined) given W and channel conditions. For a wireless channel of width WHz, the (coded) link data rate r depends on the modulation scheme, the number of spatial streams as well as the coding rate. Modulations schemes with higher data rates are, in general, more sensitive to channel conditions. The bit error probability p_b, the packet error rate p in terms of the SNR-per-bit, and the delay, for a modulation scheme MOD is:

$$p_b = f_{MOD}\left(\frac{E_b}{N_0}\right) = f_{MOD}\left(SNR \cdot \frac{W}{r}\right)$$
$$p = 1 - (1 - p_b)^B \text{ and } d = s/(r(1-p)) \quad (2)$$

for sufficiently long packets, where B is the number of bits in a network packet and s is the size of the data requested by the application in a transaction (henceforth referred to as "app layer maximum transmission unit (MTU)"). We approximate the lower BER of a coding scheme by using the coded data rate in the calculation of E_b/N_0. Thus, the feasibility $\Psi = \mathbb{R}^3(d, p, r)$ can be created by considering various modulation and coding schemes. The typical use case for our scheme involves multiple users connected, or attempting to connect to, a central access point. The available channel bandwidth W is fixed. The objective is to minimize the used bandwidth, while maximizing an MOS related objective. A fraction of the available bandwidth W, denoted as W_i needs to be assigned to each user. Based on the SNR and choice of modulation, a data rate can be achieved. The QoE of user i can then be calculated using the d, p, r values for that link. Therefore, the problem can be cast as an optimization problem: compute the per-user bandwidth allocation W_i and the per-user modulation/coding scheme such that the average QoE across all users is maximized, and where p, d, r can be determined as above:

$$\max_{(d,p,r)\in\Psi} \quad \frac{1}{N}\sum_{i=1}^{N} e^{a_0^i} e^{a_1^i d + a_2^i p + a_3^i r} \quad (3)$$
$$\text{s.t.} \quad \sum_{i=1}^{N} W_i \leq W$$

Example:- Two users are using a shared 802.11n wireless link, each of whom is using a different application. In order to construct the feasibility region Ψ, p_b should be calculated. For BPSK/QPSK and M-QAM ([25]) modulation schemes:

$$\text{(BPSK/QPSK) } p_b = 0.5 \cdot erfc(\sqrt{SNR \cdot W/r}) \quad (4)$$

$$\text{(M-QAM) } p_b = \frac{\sqrt{M}-1}{\sqrt{M}\log_2\sqrt{M}} \cdot erfc\left(\sqrt{\frac{3\log_2 M}{2(M-1)}\frac{E_b}{N_0}}\right)$$
$$+ \frac{\sqrt{M}-2}{\sqrt{M}\log_2\sqrt{M}} \cdot erfc\left(3\sqrt{\frac{3\log_2 M}{2(M-1)}\frac{E_b}{N_0}}\right) \quad (5)$$

Note that these formulae are a reasonably good approximation of the actual BER. A list of modulation and coding schemes (MCS), along with the corresponding data rates, can be found in the 802.11 standard. For example, MCS index 42 specifies that three spatial streams are to be used, with 64-QAM (6 bps), 16-QAM (4 bps) and QPSK (2 bps). The coding rate is 1/2, thus there are $0.5 * (6+4+2) = 6$ data bits per symbol. There are 52 sub-carriers (20MHz); for an OFDM symbol rate of $4\mu s$, the data rate r is calculated as $6 * 52/4\mu s = 78Mbps$. Overall p_b is calculated as the average of BERs for 64-QAM (Equation 5 with $M = 64$), 16-QAM (Equation 5 with $M = 16$) and BPSK/QPSK (Equation4). Once p_b is calculated, p and d can be found using system parameters B and s respectively. Thus, the feasibility region Ψ can be constructed for each user.

Utility functions: The authors of [15] 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 the mismatch we artificially increased the upper limit of the data rate to 200Mbps by changing the coefficients: $MOS = 1.2 \cdot ln(5 \times 10^{-8} r)$. 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:

$$MOS = e^{-6.8643p - 0.10799d + 1.1\times10^{-8} r} \quad (6)$$

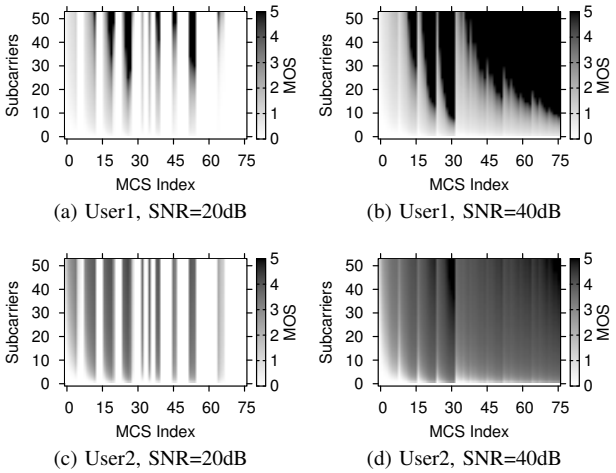
A second QoE-QoS model can be found in [19]. Packet loss rates were varied for three audio codecs: G.723.1.B which has a capacity requirement of 6.4kbit/s, iLBC (15.2kbit/s), Speex (24.6kbit/s), and G.711 (64kbit/s). Again, these bit rates are much smaller than the 802.11n rates so we artificially increased the bit rates a thousandfold - this step can be justified by thinking of the "user" as a VoIP aggregating device to which thousands of users are connected. After extrapolating the data, the following equation was obtained using multiple linear regression with $R^2 = 0.94681$ (note that $a_0^1 = 1.3629$):

$$MOS = e^{1.3629} e^{-1.5068p - 0.10461d + 3.5238\times10^{-10} r} \quad (7)$$

Optimization: Bandwidth allocation becomes a subcarrier allocation problem. The input vector V to the optimization solver is of length $2N$. For each user i, there are two elements $V(i)$ and $V(i+1)$ in V: the MCS index and the number of subcarriers to be allocated respectively. This latter value is representative of the allocated bandwidth W_i - the standard defines 52 subcarriers for 20MHz bandwidth, and so 26 subcarriers would represent 10MHz of bandwidth. Suppose that the number of subcarriers to be allocated is S (< 52 for 802.11n 20MHz). There are 76 possible MCS indices for 802.11n 20MHz. When the objective is to maximize the average MOS of the two users, the optimization problem (Equation 3) now becomes:

$$\max_{V} \quad 0.5 * (e^{-6.8643p - 0.10799d + 1.1\times10^{-8} r}$$
$$+ e^{-1.5068p - 0.10461d + 3.5238\times10^{-10} r}) \quad (8)$$
$$\text{s.t.} \quad V(1) + V(3) \leq S \quad (9)$$
$$0 \leq V(0), V(2) \leq 75 \quad (10)$$

where d, p, r are obtained as above. This problem is an integer nonlinear optimization problem, with linear constraints. The objective function is non-smooth, and thus, a global optimization technique is suitable. We choose evolutionary algorithms



(a) User1, SNR=20dB (b) User1, SNR=40dB

(c) User2, SNR=20dB (d) User2, SNR=40dB

Fig. 2: Heatmap showing the MOS (higher MOS is better) of User 1 (data download workload) at (a) low SNR and (b) high SNR; MOS of User 2 (VoIP workload) at (c) low SNR and (d) high SNR. On the X axes is the MCS index, and on the Y axis is the allocated bandwidth represented as the number of allocated subcarriers. Darker areas indicate larger MOS.

to solve this problem, because of 1) their stochastic nature and 2) their ability to accommodate multiple objectives. For example, when S is not specified, the above problem can be transformed into a multi-objective problem where the second objective is to minimize the used bandwidth $V(1) + V(3)$. Genetic algorithms, and the NSGA-II algorithm in particular, are popular choices for evolutionary algorithms. The input, which is the integer valued vector V, is called a "chromosome". Through multiple operations like crossover, selection and mutation, new candidate chromosomes (i.e., solutions) are generated. In case of multiple objective functions, a Pareto front is obtained.

IV. Performance Evaluation

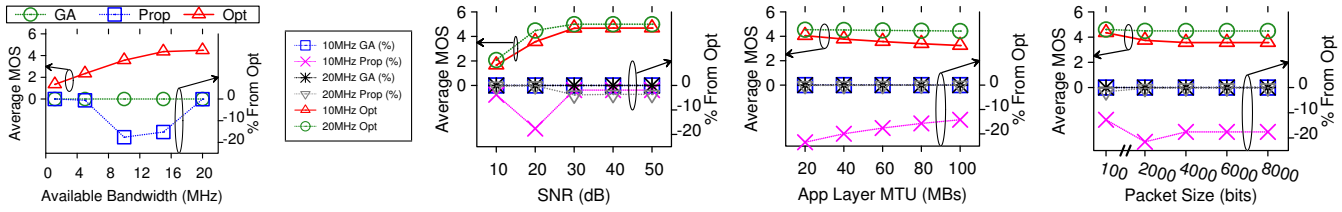
In this section we present the performance evaluation of our scheme. First, we discuss the MOS profiles of the two users, and then evaluate the performance of the bandwidth allocation schemes. The users are assumed to be connected to an 802.11n based LHWN access point. The Middleware, installed on both of the users' devices, has profiled the application's coefficients as a_0, a_1, a_2, a_3 (Equations 6 and 7) and sent it to the Scheduler running on the LHWN device. It should be noted that profiling an application and determining the coefficients is not a trivial task and a research problem in itself. The cloud service provider has tasked the Scheduler with reducing bandwidth usage while ensuring certain user Quality of Experience objectives, so that more users can be accommodated on the same spectrum. Available channel bandwidth is 20MHz (52 subcarriers), and channel conditions change frequently. There are a total of 76 possible modulation and coding schemes for 20MHz 802.11n, using a maximum of 4 spatial streams.

User Profile: The MOS profile of each user at different SNRs can be seen in Figure 2 (higher MOS is better). During unfavorable channel conditions with a SNR of 100, the MOS profile of user 1 is shown in Figure 2a, and user 2 in Figure 2c. As the number of subcarriers increases, so does the allocated bandwidth - thus increasing the throughput and MOS in general. The MCS index determines the actual data rate. The large white spaces between the "strips" are an artifact of the MCS index layout. The ranges of indices are broadly divided based on spatial streams (SS): indices 0-7 for 1 SS, 8-15 and 32-37 for 2 SS, 16-23 and 38-51 for 3 SS, and 24-31 and 52-75 for 4 SS. Within each sub-range, all possible modulation schemes are available. However, some of these modulation schemes have a high BER at low SNR, while all modulation schemes have low BER at high SNR. This is what causes the white "strips" in Figure 2. One can see that these strips disappear at a SNR of 10000 (Figures 2b and 2d), where *all* modulation schemes have low BER. Because the data rate increases and BER/PER decreases as better modulation schemes are chosen at high SNRs, for a user, the throughput increases, thus decreasing the delay and increasing the MOS in general. At low SNRs, the data rate can be multiplied up to four-fold by using multiple spatial streams and a modulation scheme with a low symbol rate. This is how the MOS can approach its maximum value even at low SNRs; however, only very few MCS indices cause high MOS (as compared to high SNR conditions).

Simulation Setup: The jMetal library provided an implementation of constrained NSGA-II in our Java based simulator. The default values of the parameters are as follows: $SNR = 100 = 20dB$, $B = 4000$ and $s^1 = s^2 = 60MB$. The performance of our scheme (GA) is compared to two other schemes for two different objectives: maximizing the average MOS and maximizing the minimum MOS. Each data point for GA is averaged over 20 random runs. The two other schemes are: Opt - which uses brute force search to determine the bandwidth allocation, and Prop - which divides bandwidth equally among the users. For a given number of subcarriers, both these schemes choose the MCS index that maximizes MOS by iterating over all MCS indices. Note that Opt is feasible to implement only when the number of users, as well as the available bandwidth, are small; as the number of subcarriers increases, the number of ways to divide it uniquely among many users increases exponentially. While Prop has the least computational overhead, Opt has the highest, owing to the exhaustive search. Finally, the problem of equalizing user MOS is discussed, and evaluated. A note on how results are visualized: we compare GA and Prop against Opt not by using the absolute MOS values, *but as a percentage relative to Opt*. This is because a difference in MOS of 0.1 at the MOS value of 1 (10%) has a higher impact on the user's already low QoE more than a difference of 0.1 at the MOS value of 3 (3.33%). We assume that user SNRs are the same, for comparison purposes only.

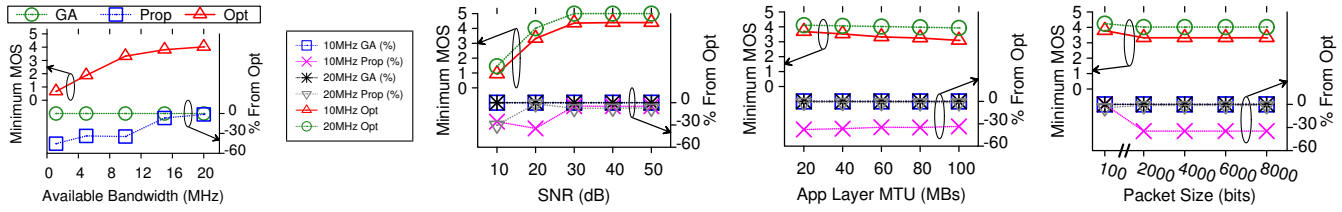
A. Maximizing the Average MOS

Figure 3 shows the performance of our scheme GA when the objective is to maximize the average user MOS (Equation 3). The available bandwidth is constrained at various values ($S = 3, 13, 26, 39, 52$ in Equation 9), to generate each data point. There is no constraint on the maximum spatial streams that can be used, i.e., no constraint on the MCS index. The result is shown in Figure 3a. As available bandwidth increases, so does the MOS, due to higher data rates. The performance of GA is *identical* to Opt. On the other hand, Prop behaves inconsistently - sometimes as much as 17.6%



(a) Comparison (b) Legend (c) SNR (d) MTU (e) Packet Size

Fig. 3: Maximizing the average MOS of users. (a) Performance of our scheme (GA) as compared to optimal (Opt) and proportional (Prop) for allocation of 20MHz (52 subcarriers), SNR = 20dB; (b) legend for figures (c)-(e). Effect of: (c) SNR, (d) app layer MTU, and (e) packet size on the performance of GA, Opt, and Prop, at allocations of 10 and 20MHz.



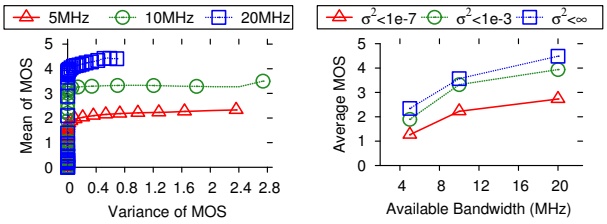
(a) Comparison (b) Legend (c) SNR (d) MTU (e) Packet Size

Fig. 4: Maximizing the minimum MOS of users. (a) Performance of our scheme (GA) as compared to optimal (Opt) and proportional (Prop) for allocation of 20MHz (52 subcarriers), SNR = 20dB; (b) legend for figures (c)-(e). Effect of: (c) SNR, (d) app layer MTU, and (e) packet size on the performance of GA, Opt, and Prop, at allocations of 10 and 20MHz.

below optimal at 10MHz. This is because less bandwidth needs to be allocated to user 2 (Figure 2c), and not in equal proportion. To ensure an average MOS of about 3.5, the reduction in required bandwidth, when using GA, as compared to Prop, is about 33%. This means that more users can be accommodated since bandwidth usage is reduced. The effect of changing channel conditions for allocations of 10 and 20MHz is shown in Figure 3c. As the SNR increases from 10dB to 50dB, channel conditions improve, decreasing the BER/PER and increasing the throughput, thereby increasing the MOS. Not much improvement can be noticed between 40 and 50dB, since high modulation rates like 128-QAM are not available in the 802.11n standard (unlike the upcoming PHY standards such as 802.11ac). We note that improving SNR provides for much higher MOS than increasing the bandwidth. At both 10 and 20MHz, GA performs identical to Opt, while Prop deviates by at most 17.6% and about 3% on average. The effect of changing application layer MTU (s^j) for users can be seen in Figure 3d. With a smaller MTU, the delay is reduced, thus increasing the MOS; for video applications this delay can be interpreted as the initial video buffering time. At a high MTU, the delay is higher - but since Prop allocates bandwidth independent of MTU, MOS increases because of the user profile characteristics. However, depending on the application encoding, using lower MTU sizes can incur higher overhead. At the same time, using higher MTUs require larger buffers and processing capabilities. Depending on the application, the MTU size should be chosen and relayed to the Middleware, so that an optimal amount of bandwidth can be allocated. GA's performance is identical to Opt, while Prop deviates by at most 23.4% at 10MHz and 0.11% at 20MHz. Note that as MTU changes, all algorithms adjust the MCS index accordingly; this explains the somewhat constant performance across MTUs. Effect of increasing packet size (B) is shown in Figure 3e. The PER increases with increasing B for constant BER (which is determined by the bandwidth and MCS index). The penalty in MOS is not that high - only about 10%, as packet size increases from 100 to 8000 bits. The advantage of large packet sizes is that multiple frames can be aggregated at the data link layer, as proposed in 802.11n as well as 802.11ac. Multiple TCP packets could fit inside a single frame. Thus, our scheme is able to take advantage of frame aggregation by using the packet size to generate the feasibility region, which is used by the optimizer. Again, GA performs identical to Opt, while Prop deviates by at most 21.5% at 10MHz and 1.6% at 20MHz.

B. *Maximizing the Minimum MOS*

GA performs identical to Opt (Figure 4a) and outperforms Prop, which performs inconsistently. GA is able to save 25% bandwidth when ensuring a minimum MOS of 3.9. At a low SNR of 10dB (Figure 4c), the minimum MOS is close to zero. This can be contrasted with Figure 3c, where the objective was to maximize the average MOS. In order to boost the minimum MOS, a lot of bandwidth has to be allocated to user 1, whose MOS profile is not as "flat" as user 2. Since this is not an ideal operating condition, we address the problem of equalizing user MOS in the next section. GA performs identical to Opt, while Prop deviates by at most 34.66%. Effect of increasing MTU is seen in Figure 4d. For a five fold increase in MTU size from 20 to 100MB, the reduction in MOS is about 21.5%. Performance of Prop is fairly constant across MTUs, since there are no constraints on the MCS index. In Figure 4e, the result of increasing packet size over two order of magnitudes is seen. The reduction in MOS in this case is about 20%. In both cases, GA performs identical to Opt, while Prop performs within 37%. We conclude that, irrespective of the objective, the performance of GA is identical to Opt; however, this result comes at the cost of increased computation. In a cloud service scenario, the LHWN can easily use cloud computation resources to compute this optimal bandwidth allocation, instead of performing it on the device itself. Prop performs inconsistently, based on the available bandwidth.



(a) (b)

Fig. 5: At SNR of 20dB: (a) Pareto front for GA scheme between variance and mean of users' MOS, for bandwidth allocations of 5, 10, and 20 MHz. (b) performance of GA scheme for three different tolerances of variance in user MOS.

C. Equalizing MOS and Fairness

It is possible that while maximizing the average MOS in a system, some users are unfairly penalized with a low MOS or small bandwidth allocation. Fairness can be quantified using a measure of dispersion - in this case, the variance of users' MOS. Ideally the variance should be zero, but in practice a very low variance can be tolerated, if it means an increase in average MOS. This inherent trade off can be characterized as a Pareto front. It can be obtained using our optimization scheme as follows. The objective of the optimization problem (Equation 8) is changed to dual objectives: minimizing a measure of dispersion (variance) of the users' MOS, while maximizing the average MOS. Using a chromosome pool size of 500, a Pareto front was obtained and can be seen in Figure 5a. Clearly, the variance in user MOS can be reduced to zero, but at the cost of decreased average MOS. This is because there is no tuple of (d, p, r) such that $U^1(d, p, r) = U^2(d, p, r)$. However, this observation is specific to the QoE-QoS profile of the two users, as well as the channel conditions. The network operator may choose to equalize users' MOS within a specified tolerance. Our system can easily accommodate such a demand, by solving the dual objective problem for a given available bandwidth, and then choosing the solution that maximizes performance while satisfying the tolerance bound. The results for such a scheme can be seen in Figure 5b for three different values of variance: $10^{-7}, 10^{-3}$, and ∞. We see that MOS can be equalized more favorably (i.e. with higher average MOS) at higher bandwidths, while a tight tolerance results in very low QoE at low bandwidths. Tolerating a very small MOS variance of 10^{-3} allows the average MOS to double. The key takeaway here is that MOS can be equalized, but at the cost of average MOS, and that this tradeoff can be controlled by the network operator.

V. Future Work

The authors are currently designing a testbed where the above scheme is to be implemented. QoS_{app} will be gathered for multiple users using extensive experimentation. Incorporating limits on the number of spatial streams used (i.e., number of antennas) as well as a finite number of encoders is ongoing. For multiple users, 802.11ac will replace 802.11n because of the increase in supported data rates. Finally, aspects such as computation complexity, other QoS metrics, non-uniform subcarrier fading as well as optimal power allocation will be considered.

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

References

[1] W. C. Hardy. *QoS Measurement and Evaluation of Telecommunications Quality of Service*, July 2001.
[2] R. Stankiewicz, P. Cholda, and A. Jajszczyk. Qox: What is it really? *Communications Magazine, IEEE*, April 2011.
[3] M. Alreshoodi and J. Woods. Survey on qoe/qos correlation models for multimedia services. *CoRR*, 2013.
[4] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang. Understanding the impact of video quality on user engagement. SIGCOMM '11.
[5] J. Klaue, B. Rathke, and A. Wolisz. EvalvidA framework for video transmission and quality evaluation. In *Computer Performance Evaluation. Modelling Techniques and Tools*. Springer, 2003.
[6] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience for internet video. SIGCOMM '13.
[7] T. Wang, A. Pervez, and H. Zou. Vqm-based qos/qoe mapping for streaming video. IC-BNMT '10.
[8] H. J. Kim and S. G. Choi. A study on a QoS/QoE correlation model for QoE evaluation on IPTV service. ICACT '10.
[9] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor. Quality of experience-driven adaptation scheme for video applications over wireless networks. *Communications, IET*, July 2010.
[10] M. Venkataraman and M. Chatterjee. Inferring video QoE in real time. *Network, IEEE*, 2011.
[11] R. G. Cole and J. H. Rosenbluth. Voice over IP performance monitoring. *SIGCOMM Comput. Commun. Rev.*, April 2001.
[12] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo. The logarithmic nature of QoE and the role of the weber-fechner law in QoE assessment. ICC '10.
[13] S. Khorsandroo, R. Md Noor, and S. Khorsandroo. A generic quantitative relationship to assess interdependency of QoE and QoS. *KSII Transactions on Internet & Information Systems*, 2013.
[14] M. Fiedler, T. Hossfeld, and P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *Network, IEEE*, March 2010.
[15] J. Shaikh, M. Fiedler, and D. Collange. Quality of experience from user and network perspectives. *annals of telecommunications - annales des tlcommunications*, February 2010.
[16] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler. QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems. *Computer Communications*, March 2010.
[17] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, April 2004.
[18] F. Agboma and A. Liotta. QoE-aware QoS management. MoMM '08.
[19] 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.
[20] J. Huang, V. Subramanian, R. Agrawal, and R. Berry. Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks. *IEEE Journal on Selected Areas in Communications*, February 2009.
[21] J. Gross, J. Klaue, H. Karl, and A. Wolisz. Cross-layer optimization of OFDM transmission systems for MPEG-4 video streaming. *Computer Communications*, July 2004.
[22] B. Li, S. Li, C. Xing, Z. Fei, and J. Kuang. A QoE-based OFDM resource allocation scheme for energy efficiency and quality guarantee in multiuser-multiservice system. Globecom Workshops '12.
[23] G. Song and Y. Li. Cross-layer optimization for OFDM wireless networks-part i: theoretical framework. *IEEE Transactions on Wireless Communications*, March 2005.
[24] S. Aroussi, T. Bouabana-Tebibel, and A. Mellouk. Empirical QoE/QoS correlation model based on multiple parameters for VoD flows. Globecom '12.
[25] K. Cho and D. Yoon. On the general BER expression of one- and two-dimensional amplitude modulations. *IEEE Transactions on Communications*, July 2002.