Title

Resource Allocation for Semi-Elastic Applications With Outage Constraints in Cellular Networks

Permalink
https://escholarship.org/uc/item/7zx1f9jn

Journal
IEEE Transactions on Vehicular Technology, 64(4)

ISSN
0018-9545

Authors
Yang, Chao
Jordan, Scott

Publication Date
2015-04-01

DOI
10.1109/tvt.2014.2331213

Peer reviewed

Resource allocation for semi-elastic applications with outage constraints in cellular networks

Chao Yang and Scott Jordan University of California, Irvine
Email: Chao.Yang@uci.edu, sjordan@uci.edu

***Abstract*—We consider resource allocation for semi-elastic applications, such as mobile video-conferencing, which require a maximum outage. We represent the performance of a session by a sigmoid utility function of the average bit rate over a time window consisting of many slots. The goal is to maximize the total expected utility of all active users. A principal challenge is the difference in time scales: outage is measured over sessions while average bit rate is measured over a time window such as a group of pictures. We propose that these differences in time scale can be elegantly addressed by shadow prices: a price per unit average rate over each time window and a price per unit outage. We further propose that the price per unit average rate depends on combined pathloss and shadowing, but not on fast fading. We show that resources can be efficiently allocated if the base station chooses prices based on total demand and the users respond by choosing average rates. The performance of our algorithm is illustrated by simulation results.**

Keywords—mobile video-conferencing, sigmoid utility, outage, resource allocation

I. Introduction

Video applications constitute a rapidly increasing portion of the total traffic on cellular networks. It is now estimated that video comprises one third of downstream North American mobile Internet access peak period traffic [1]. Most of this video traffic is streaming encoded using either Adobe Flash or MPEG. Some of this video traffic is video conferencing[1], e.g. Apple's FaceTime for iPhones and Skype's video conferencing application for smartphones. Fourth generation (4G) cellular networks will integrate voice, video, and data applications using packet switching.

Resource allocation for video applications has received great attention in recent years. For such applications, variable bit rate video encoding algorithms are typically used, e.g. MPEG. As a consequence, resource allocation should take into account application characteristics when best-effort packet switching would not result in acceptable performance. A few papers have proposed modeling the performance of video applications using a sigmoid utility function of the rate of the connection, i.e. utility is convex at rates less than a threshold and concave at rates above that threshold [2]. A number of papers model utility as a function of *instantaneous* rate, and thus propose resource allocation algorithms based on instantaneous rate [3][4][5][6][7][8][9][10]. For instance, in [7], we considered the case in which utility is a semi-elastic function of the rate achieved in each time slot. We proposed a near-optimal algorithm that iteratively finds optimal shadow prices for power and rates, and uses these shadow prices to allocate power and subcarriers.

However, we have argued that for video encoded using group-of-picture structures, e.g. MPEG, utility should be represented as a sigmoid function of the *average rate over each group-of-pictures* [11]. Thus the algorithms in [3][4][5][6][7] are not suitable to video conferencing. Since a group-of-pictures comprises many time slots, resource allocation based on average rate over each group-of-pictures can exercise considerably greater flexibility than resource allocation based on instantaneous rate, and thereby achieve better results. Below we will propose that utility be formulated as a sigmoid function of the average rate over a preceding time window consisting of many slots. This dependence on average rate over a time window, however, creates a challenge. Since users are mobile, channels are highly variable. There is thus significant uncertainty about a user's channel within the next time window, and hence about achievable average rate over the time window. A key resource allocation design problem consists of how to consider a user's current channel and likely future channels. In [11], we considered the case in which utility is a sigmoid function of the average bit rate over multiple time slots, motivated by video conferencing, and showed that greedy allocation to maximize incremental utility in the current time slot can be implemented in a distributed fashion by an exchange of price and demand amongst users, the network, and an intermediate power allocation module. We proposed resource allocation that considers both the average rate achieved so far and the future expected rate, and showed how the future expected rate can be estimated by modeling the probability that a user will be allocated a subcarrier in a future time slot.

However, video conferencing applications have a second key performance metric that has not been considered. Users view connections as unacceptable if the average rate drops below a minimum threshold too often. Such a notion of *outage* is common when modeling voice applications, and thus resource allocation for voice often obeys constraints on the probability that the instantaneous rate falls below a minimum threshold [12][13][14][15][16]. We argue here that for video encoded using group-of-picture structures, e.g. MPEG, outage occurs when the average rate over the group-of-pictures, not the instantaneous rate, falls below a minimum threshold[2]. This

[1] In "video conferencing", we include conversions using both audio and video amongst two or more parties.

[2] For video streaming using MPEG, the utility shape is still sigmoid. However, further study is required to determine the length of the time window.

calls for an approach to resource allocation that considers both utility and outage based on the average rate over the group-of-pictures. None of the prior literature on resource allocation for video using sigmoid utility functions, including our prior papers [7] and [11], considered outage.

The existence of an outage constraint poses a challenge to resource allocation. For best-effort applications, commonly represented by concave utility functions, outage is not an issue and total utility is maximized by allocating few resources to users with poor channels. For inelastic applications such as voice, commonly represented by step utility functions, outage constraints are satisfied by allocating resources to maintain a strict minimum threshold rate for each user whenever possible. For semi-elastic applications such as video conferencing, however, neither approach is optimal. The minimum rate requirement isn't as strict for semi-elastic applications as for inelastic applications. There is a benefit to allocating fewer resources to users with poor channels, but not so few as to violate the outage constraint. The key question is when and how much to compensate for a user's poor channel.

The major contributions of this paper are:

1) Formulating power and subcarrier allocation as an optimization problem with a metric of total user utility as a sigmoid function of the *average rate over each group-of-pictures*, and with *outage constraints*. Evaluation of utility over sliding windows is novel. The introduction of an outage constraint in such a setting is also novel.
2) Solution of a non-causal version of the optimization problem, and illustration of the structure of the solution. We illustrate how the usual shadow prices for rate and power now depend on the length of the time window over which utility is evaluated, and how shadow costs can be associated with outage constraints.
3) Transformation of the optimization problem into a causal problem that can be efficiently solved. We illustrate how the shadow prices for average rate and outage can be combined to elegantly determine power and subcarrier allocation, by transforming the objective functions using statistical averages and by determining the combined price as a function of a user combined pathloss and shadowing, which is nearly constant during the time window over which utility is evaluated.
4) We propose an iterative algorithm that determines power and subcarrier allocations based on quantization of combined pathloss and shadowing. We illustrate how this approach can maximize utility subject to long-term average outage constraints, how the use of a time window impacts the solution, and how these policies differ from alternate resource allocation policies.

The rest of this paper is as follows. In section II, we define a user's channel, rate, and utility. We also provide some background theory and knowledge. In section III, we consider a non-causal version of the problem. In section IV, we pose a simpler causal problem and propose a new optimal solution. Finally, in section V the performance of our algorithm is illustrated by numerical simulation results.

II. System and Utility Model

Our approach is to construct a link layer model, and to abstract the key elements of other layers. The physical layer is abstracted in the first subsection, and the network through application layers are abstracted in the next subsection. The goal is to obtain a model simple enough to lead to an understandable optimization problem, and to allow us to formulate reasonable link layer resource allocation algorithms.

A. System Model

We consider a single cell downlink Orthogonal Frequency-Division Multiple Access (OFDMA) system serving K users with N subcarriers. The bandwidth B of each subcarrier is assumed to be less than the coherence bandwidth of the channel so that the channel response can be considered flat.

The physical layer is abstracted using a common model for the relationship between channel, power, and instantaneous rate, see e.g. [17][18][19][20]. The instantaneous rate of user k on subcarrier n in time slot t is:

$$r_{k,n,t}(p_{k,n,t}) = B\log_2\left(1 + p_{k,n,t}\frac{|H_{k,n,t}|^2}{\sigma^2 + I}\right) \tag{1}$$

where $p_{k,n,t}$ is the power allocated, $|H_{k,n,t}|^2$ is the composite channel gain, σ^2 is the noise power, and I is the average inter-cell interference power[3]. The channel gain $|H_{k,n,t}|^2 = \alpha^2_{k,n,t}\gamma_{k,t}PL_{k,t}$ is composed of fast fading $\alpha^2_{k,n,t}$ which changes significantly in sequential time slots, slow fading and shadowing $\gamma_{k,t}$ which changes little in sequential time slots but may change significantly during a few seconds, and pathloss $PL_{k,t}$ which depends on user position and changes significantly during tens of seconds. Fast fading on different subcarriers is assumed to be independent to each other. The total instantaneous rate of user k in time slot t is the sum of the user's instantaneous rate over all subcarriers:

$$R_{k,t} = \sum_{n=1}^{N} r_{k,n,t} \tag{2}$$

Our notation is summarized in Table I.

B. Utility Model

There are several metrics used to evaluate the performance of real-time video. Many papers have proposed that network resources should be allocated to maximize the average peak signal to noise ratio (PSNR), see e.g. [25][26], sometimes subject to long term average rate constraints, see e.g. [27][28]. However, real-time voice and video are often evaluated by the delay and delay jitter of the stream. Some researchers have proposed replacing long term average rate constraints with some type of delay deadline, see e.g. [29] which minimizes wireless resource usage subject to statistical delay and loss

[3] See e.g. [21][22] for the dependence of average inter-cell interference upon the interference power spectral density. Alternatively, one may model time-dependent inter-cell interference based on activity in neighboring cells, see e.g. [23][24]. In a real system, $\frac{|H_{k,n,t}|^2}{\sigma^2+I}$ would be replaced by the measured signal to interference plus noise ratio (SINR).

TABLE I. NOTATION

Notation	Description
K	number of users
N	number of subcarriers
W	number of time slots in a time window
t	current slot number
M	number of slices in the partition of combined pathloss and shadowing
B	subcarrier bandwidth
P	downlink power of base station
$\overline{Pr}$	outage constraint
$S_k^{'}$	average rate of user k at maximum average utility
U_k	utility function of user k
$\gamma_{k,t}$	shadowing of user k
$PL_{k,t}$	pathloss of user k
$\psi_{k,t}$, ψ_k, ψ_k^m	combined pathloss and shadowing of user k
$\alpha_{k,n,t}^2$, $\alpha_{k,n}^2$	fast fading of user k
$\|H_{k,n,t}\|^2$, $\|H_{k,n}^m\|^2$	composite channel fading of user k
$p_{k,n,t}$, $p_{k,n}^m$	allocated power to user k on subcarrier n
$r_{k,n,t}$, $r_{k,n}^m$	instantaneous rate of user k on subcarrier n
$R_{k,t}$, R_k^m	instantaneous rate of user k in one slot
$S_{k,t}$, S_k, S_k^m	average rate of user k within preceding time window
μ, μ_t	power price
$\overline{\beta}_{k,t}$, β_k^m	average rate price
I and σ^2	interference power and noise power

constraints, [30] which maximizes concave utility subject to delay constraints, [31] which minimizes the expected end-to-end distortion subject to delay constraints, and [32] which minimizes the error propagation of a group of pictures subject to delay constraints.

However, video encoding algorithms often use a *group of pictures* as a central concept, and intentionally vary the bit rate over different frames within this group. As a consequence, delay and delay jitter within a group of pictures becomes less important, as the decoder will often intentionally delay packet processing until a frame is received. The most important performance metric becomes the number of packets received within a group of pictures. Some papers have thus adopted throughput as the primary performance metric for video, and often proposed that utility of multimedia applications be modeled as a sigmoid function of the throughput [2][33][3][4][34][35]. A sigmoid utility function also reflects the layered coding structure of MPEG video. The initial convex portion reflects the average rate required to transmit the base video layer. The concave portion reflects the use of incremental average rate to transmit enhancement layers; every additional enhancement layer increases user satisfaction, but with decreasing returns.

Here we adopt throughput as the primary performance metric. However, since performance depends on the number of packets received *within a group of pictures*, we have previously proposed to model user utility as a sigmoid function of the average rate *over a time window consisting of W time slots* [11], as pictured in Fig 1. For video, the time window is likely to be chosen to be one group of pictures.

As our approach is to construct a link layer model, the network through application layers including the video sequence, the video encoder and decoder, RTP, TCP or UDP, packet transmission, routing, and packet fragmentation are abstracted using a utility function that depends only on the average throughput over a time window of W time slots. The power of this model is that it can lead to optimization problems at the link layer that in turn result in reasonable link layer resource allocation algorithms. This abstraction presumes that the video encoder/decoder is capable of prioritizing packets within a time window, and that either it signals packet priority to the link layer or dynamically responds to packet receptions and losses. In the former case, a video encoder/decoder may use layered coding and set the priority of a packet based on the video layer, and the link layer may transmit packets within a time window in priority order rather than temporal order. In the latter case, a video encoder may itself transmit packets within a time window in priority order, based on feedback from the video decoder or from RTP about received and dropped packets. In either case, we posit that the resulting quality can thus be abstracted into a single variable – the number of packets received within a time window – rather than on which packets are received, packet loss, packet delay, packet delay jitter, etc. The utility function thus describes the perceived performance of the video stream, and its shape depends on how effectively the video encoder/decoder encode information and determine packet priority.

In time slot t, denote the average rate of user k during the previous W time slots by $S_{k,t} = \sum_{\tau=t-W+1}^{t} R_{k,\tau}/W$. The utility of user k is assumed to be a function $U_k(S_{k,t})$ which maps the average throughput achieved in W time slots to the level of the satisfaction perceived by the application [11]. There exists an inflection point S_k^f such that U_k is convex for $S_{k,t} < S_k^f$ and concave for $S_{k,t} > S_k^f$. We denote the average rate at the maximum average utility by $S_k^{'}$, namely $S_k^{'} = \arg\max_{S_{k,t}} U_k(S_{k,t})/S_{k,t}$. This shape is thought to reflect the nature of the compression techniques used in semi-elastic applications, which are designed to adjust to fluctuations provided that short-term throughput remains above a threshold, but which do not fail gracefully when short-term throughput falls below that threshold.

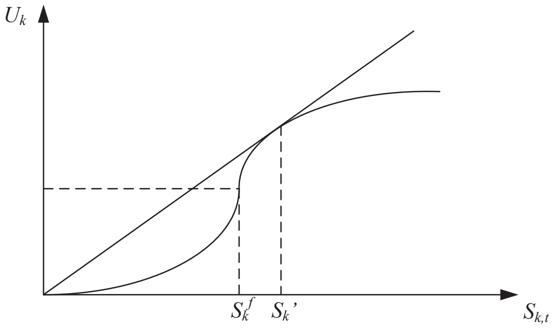

Fig. 1. A sigmoid utility function

III. NON-CAUSAL PROBLEM AND SOLUTION

We first consider a non-causal system in which at the beginning of $T >> W$ time slots the network knows the future channel gains of each user in each time slot. While knowledge of future information is clearly unrealistic, it will provide guidance to design algorithms for causal systems.

A. Optimization problem

Denote the power allocation by $\mathbf{p} = \{p_{k,n,t}, \forall k, n, t\}$. Each subcarrier can be allocated to at most one user, thus denote the feasible set of power and subcarrier allocations by $\mathbf{A} = \{\mathbf{p}$ s.t. $\forall t, n,\ p_{k,n,t} > 0$ for at most one user $k\}$. The utility generated per unit average rate is maximum at $S_k^{'}$. Thus, it is reasonable to assume that a compression algorithm would be designed on the assumption that the average rate is maintained above this threshold[4]. A user is thereby considered to be in outage during time slot t if and only if $S_{k,t} < S_k^{'}$. The proportion of time that user k is in outage is the number of time slots that user k obtains an average rate lower than the target average rate divided by the total number of slots [36], and is denoted by

$$Pr(S_{k,t} < S_k^{'}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left(S_{k,t} < S_k^{'}\right) \tag{3}$$

where $\mathbf{1}()$ is an indicator function.

The system objective is presumed to be to maximize the total user utility within T slots[5] under constraints that the total transmitted power in each time slot not exceed the available P and that the outage probability is within a certain threshold $\overline{Pr}$:

$$\begin{aligned} &\max_{\mathbf{p}\in\mathbf{A}} \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} U_k(S_{k,t}) \qquad (4)\\ \text{s.t.}\quad &\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n,t} \le P\ \forall t;\ p_{k,n,t} \ge 0\ \forall k,n,t\\ &Pr(S_{k,t} < S_k^{'}) \le \overline{Pr}\ \forall k \end{aligned}$$

B. Optimal resource allocation

Because future channel gains are assumed known in this non-causal problem, allocations of power and subcarriers can be made jointly for all users and all time slots in a single decision. This allows full consideration of the variation of channel for each user from time slot to time slot and of the average bit rate during each group-of-pictures achieved as a result of allocations made in each time slot.

However, the direct solution of problem (4) requires solving KNT fixed point equations. We are thus motivated to solve a dual problem. The idea, used previously for strictly concave utility functions [7][37][38], is to separate the determination of each user's average rate $(S_{k,t})$ and the allocation of instantaneous power $(\mathbf{p})$ using a set of intermediate variables $\mathbf{d} = \{d_{k,t}, \forall k, t\}$ as bounds on the achieved rates. This decomposition can be used to allow the instantaneous power to be determined on a faster time scale than average rate. The allocated instantaneous power can then depend on the current channel and recent channels, allowing the user's rate over longer time scales to depend on the channel distribution; see e.g. [38] for more information about the general approach and the resulting duality gap and see e.g. [7] for one example of such a decomposition. A similar decomposition can be applied to sigmoid utility functions. The problem (4) can be transformed into:

$$\begin{aligned} &\max_{\mathbf{p}\in\mathbf{A},\mathbf{d}} \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} U_k\left(d_{k,t}\right) \qquad (5)\\ \text{s.t.}\quad &S_{k,t} \ge d_{k,t},\ \forall k,t\\ &\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n,t} \le P,\ p_{k,n,t} \ge 0,\ \forall k,n,t\\ &Pr(S_{k,t} < S_k^{'}) \le \overline{Pr}\ \forall k, \end{aligned}$$

It is easier to search for the optimal shadow prices and to let them determine the optimal resource allocation than to directly search for the optimal power and subcarrier allocations. This can be done by posing a dual problem, see e.g. [39]. The Lagrange of (5) is given by:

$$\begin{aligned} &J(\mathbf{d},\mathbf{p},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu})\\ =\ &\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} U_k(d_{k,t}) + \sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_{k,t}\left(S_{k,t} - d_{k,t}\right)\\ +\ &\sum_{t=1}^{T}\mu_t\left(P - \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n,t}\right)\\ +\ &\sum_{k=1}^{K}\nu_k\left(\overline{Pr} - Pr(S_{k,t} < S_k^{'})\right) \qquad (6) \end{aligned}$$

where $\boldsymbol{\lambda} = \{\lambda_{k,t}, \forall k, 1 \le t \le T\}$ are the Lagrangian multipliers associated with the average rate constraints, $\boldsymbol{\mu} = \{\mu_t, 1 \le t \le T\}$ are the Lagrangian multipliers associated with the power constraints, and $\boldsymbol{\nu} = \{\nu_k, \forall k\}$ are the Lagrangian multipliers associated with the outage constraints.

The dual function is then given by:

$$\overline{J}(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\mathbf{p}\in\mathbf{A},\mathbf{d}} J(\mathbf{d},\mathbf{p},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}) \tag{7}$$

and the dual problem is to optimally choose the Lagrangian multipliers:

$$\overline{J}^{*} = \min_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}} \overline{J}(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu})\ \text{ s.t. } \boldsymbol{\lambda} \succeq 0, \boldsymbol{\mu} \ge 0, \boldsymbol{\nu} \succeq 0 \tag{8}$$

First-order necessary conditions for solution of problem (7) depend on the derivatives $\partial J/\partial\mathbf{p}$ [40], which in turn depend on the derivatives $\partial Pr(S_{k,t} < S_k^{'})/\partial\mathbf{p}$. However, $Pr(S_{k,t} < S_k^{'})$ is the sum of several indicator functions, and thus these partial derivatives are not always defined. Alternatively, one could exhaustively search amongst all possible $\mathbf{p}$; however, the computational complexity of this approach is very high. Thus we propose replacing the outage probability constraint $Pr(S_{k,t} < S_k^{'}) \le \overline{Pr}\ \forall k$ in (5) by a constraint on the interpolation of the indicator function:

$$\frac{1}{T}\sum_{t=1}^{T}\max\left(0, 1 - S_{k,t}/S_k^{'}\right) \le \overline{Pr},\ \forall k \tag{9}$$

[4] Alternatively, one might assume that S_k^f is a reasonable threshold.

[5] For simplicity of notation, we presume that the user has been in the system since time $t = -(W-2)$, so that $S_{k,t}$ is defined at $t = 1$.

Denote the time slots that user k is in outage by $O_k = \{t | S_{k,t} < S_k^{'}\}$. Then the revised version of the Lagrangian equation (6) is:

$$\begin{aligned} & J(\mathbf{d}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}) \qquad (10) \\ \approx\ & \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} U_k(d_{k,t}) + \sum_{t \notin O_k}^{T} \sum_{k=1}^{K} \lambda_{k,t} \left(S_{k,t} - d_{k,t} \right) \\ & + \sum_{t \in O_k}^{T} \sum_{k=1}^{K} \left[\left(\lambda_{k,t} + \frac{\nu_k}{T S_k^{'}} \right) S_{k,t} - \lambda_{k,t} d_{k,t} \right] \\ & + \sum_{t=1}^{T} \mu_t \left(P - \sum_{k=1}^{K} \sum_{n=1}^{N} p_{k,n,t} \right) + \sum_{k=1}^{K} \nu_k \left(\overline{Pr} - \frac{|O_k|}{T} \right) \end{aligned}$$

The number of time slots that user k is in outage, $|O_k|$, is a discontinuous function of the set of allocated powers for user k, $\{p_{k,n,t}, \forall n, t\}$. At the points at which $J(\mathbf{d}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu})$ is continuous, i.e. $\partial |O_k| / \partial \mathbf{p} = 0$, the first order conditions can be used to show that the optimal power allocation is:

$$p_{k,n,t} = \left(\frac{B \overline{\beta}_{k,t}}{\mu_t \ln 2} - \frac{\sigma^2 + I}{|H_{k,n,t}|^2} \right)^+ \qquad (11)$$

and that subcarrier n should be allocated to the user

$$\arg\max_k \Phi_{k,n,t} \qquad (12)$$

where

$$\begin{aligned} \Phi_{k,n,t} = & \overline{\beta}_{k,t} B \left[\log_2 \left(\frac{B \overline{\beta}_{k,t}}{\mu_t \ln 2} \frac{|H_{k,n,t}|^2}{\sigma^2 + I} \right) \right]^+ \\ & - \mu_t \left(\frac{B \overline{\beta}_{k,t}}{\mu_t \ln 2} - \frac{\sigma^2 + I}{|H_{k,n,t}|^2} \right)^+ \qquad (13) \end{aligned}$$

where $\beta_{k,t} = \lambda_{k,t} + \mathbf{1}(t \in O_k) \nu_k / (T S_k^{'})$ and $\overline{\beta}_{k,t} = \sum_{\tau=t}^{t+W-1} \beta_{k,\tau} / W$.

The target average rate is

$$d_{k,t} = \max \left\{ S_k^{'}, \arg\max_{d_{k,t}} [U_k(d_{k,t})/T - \beta_{k,t} d_{k,t}] \right\} \qquad (14)$$

The Lagrangian multipliers $\boldsymbol{\mu}$ can be interpreted as shadow costs for power, $\boldsymbol{\lambda}$ as shadow costs for average rate for users not in outage, and $\boldsymbol{\nu}$ as shadow costs associated with the outage constraints. Thus $\beta_{k,t}$ can be interpreted as a shadow cost for average rate for user k in slot t; it is comprised of $\lambda_{k,t}$ plus an outage price ν_k when a user is in outage. Equation (14) states that the system allocates rates so as to maximize total user surplus, defined as total user utility minus total user cost.

If user k is charged a price of $\beta_{k,t}$ per unit average rate, then equation (11) states that power should be allocated according to a water-filling algorithm using the average shadow cost for average rate over the next W time slots. Equation (12) states that subcarriers should be allocated so as to maximize the system's profit from subcarrier n, defined as the revenue from selling average rate minus the cost of power.

The replacement of the outage probability constraint in (5) by the constraint (9) on the interpolation of the indicator function will likely result in some outage probabilities that exceed the outage threshold. This can be addressed by iteratively increasing the shadow costs $\boldsymbol{\nu}$ until each user's outage satisfies the outage constraint.

IV. Causal problem and solution

With an understanding of the structure of the optimal resource allocation, we now turn to the causal resource allocation problem. The major challenge to constructing a causal resource allocation policy is that the resource allocation policy described by (11)-(14) requires determination of the average rate prices $\{\overline{\beta}_{k,t}\}$, which depends on the outage prices $\{\nu_k\}$. However, in the non-causal formulation, these prices requires knowledge of future channel information which is clearly impossible in real systems with mobile users. The challenge is thus how to replace knowledge of future channel information by knowledge solely of current channel information and a distribution of the channel over time. This should be done in a manner that attempts to compensate for a user's poor channel and thereby guarantees outage performance. In this section, we propose basing the average rate price $\overline{\beta}_{k,t}$ on a user's combined pathloss and shadowing, which is nearly constant during time window W, thus avoiding the need to consider future fast fading. We then transform the problem (4) to consider an infinite time horizon using statistical averages, and propose to reserve an average rate margin for users with poor combined pathloss and shadowing to guarantee outage performance.

A. *Components of a user's channel*

Before crafting a causal optimization problem, it is worthwhile to consider the various components of a user's channel, and how resource allocation should depend upon each of them. Recall that the channel gain $|H_{k,n,t}|^2 = \alpha_{k,n,t}^2 \gamma_{k,t} PL_{k,t}$ is composed of fast fading $\alpha_{k,n,t}^2$, slow fading and shadowing $\gamma_{k,t}$, and pathloss $PL_{k,t}$. We classify fading as "fast" if and only if the correlation between fading at times t and $t + W$ is small enough so that it can be considered independent, and we classify the fading as "slow" otherwise.

This classification helps in two manners. First, using these definitions, the fluctuations in fast fading will largely average out during a time window, whereas fluctuations in slow fading, shadowing, and pathloss will not. This is useful from an analytical perspective. Second, consider the available knowledge of these components: fast fading, slow fading, shadowing, and pathloss. It is reasonable to presume that the base station will know the joint distribution of each of these components for each user conditioned on mobile speed, and will also know the distribution of user locations and speeds. Knowledge by the base station of the *instantaneous* channel requires feedback from the user's handset. Commonly, feedback will communicate the instantaneous channel gain, and further analysis may estimate individual components, see e.g. [41]. However, because the purpose of the average rate price is to allocate resources during a time window to maximize utility as a function of throughput achieved during a time window, it is worthwhile to distinguish fast fading (which will largely average out during a time window) and other components (which will not).

We turn to the question of how resource allocation should depend upon each of these channel components. One approach is to compensate for a user's entire channel gain. There are two reasons for this approach. First, if the system objective is to equalize SINR amongst users, then compensation achieves this objective. Second, if outage is defined solely a function of instantaneous rate, then compensation may be required to satisfy the outage constraints. However, if the system objective is not to equalize SINR and if outage is based on the average rate achieved over a longer time period, then compensation is neither optimal nor required.

For instance, many papers consider an objective of maximizing total user throughput, and propose allocating power according to a water-filling algorithm using a power price. Water-filling is based on instantaneous channel gain, but it achieves a higher total user throughput than equalization of SINR, since it uses power more effectively. Our goal is maximizing total user utility. We thus will propose a water-filling algorithm using the average shadow cost for average rate over the next W time slots, rather than based on instantaneous channel gain. As a consequence, it will make sense to distinguish between fast fading and other channel components for purposes of utility maximization.

Furthermore, in our model, a user is considered to be in outage during time slot t if and only if $S_{k,t} < S_k^{'}$. This is a different type of constraint than considered in papers that model other applications. When outage is defined solely as a function of instantaneous rate, then instantaneous fast fading is a contributor to outage. In contrast, if outage depends on the average rate achieved over a time window, then since the fluctuations in fast fading will largely average out during a time window, outage depends only on the distribution of fast fading, not on instantaneous values. As a consequence, it will also make sense to distinguish between fast fading and other channel components for purposes of satisfying outage constraints.

Our approach is thus to take into account the entire channel gain when determining power and subcarrier allocation, presuming that feedback provides this information. However, we will not attempt to provide compensation for fast fading, because compensation does not maximize total utility and is not necessary to satisfy long-term outage constraints. Instead, we will propose that the power and subcarrier allocation should be determined by the power price μ_t and a version of the average rate price $\overline{\beta}_{k,t}$. The power price μ_t will depend on the instantaneous channel gain, because it is used to ensure that allocated power does not exceed available power in each time slot. In contrast, although the average rate price will depend on *instantaneous* slow fading, shadowing, and pathloss, it will only depend on the *distribution* of fast fading. The average rate price will be used to allocate resources during a time window to maximize utility as a function of throughput achieved during a time window. The average rate price will also be sufficient to ensure that outage constraints are satisfied, since outage also only depends on the *distribution* of fast fading.

B. A causal problem using statistical averages

To formulate a causal resource allocation policy, we propose to consider each $\overline{\beta}_{k,t}$ as a single decision variable, rather than using separate decision variables $\lambda_{k,t}$ and ν_k. We then propose to use (11)-(12) to determine power and subcarrier allocations. Since average rate should be allocated on a slower timescale than power and subcarriers, we similarly propose to use (14) to determine target rates for each user, but with $\beta_{k,t}$ replaced by $\overline{\beta}_{k,t}$, i.e.

$$d_{k,t} = \max\left\{S_k^{'}, \arg\max_{d_{k,t}}[U_k(d_{k,t})/T - \overline{\beta}_{k,t}d_{k,t}]\right\} \quad (15)$$

As discussed above, we will require knowledge of the distribution of users' channels. Define a random variable $\alpha_{k,n}^2$ representing fast fading for user k on subcarrier n; denote $\boldsymbol{\alpha^2} = \{\alpha_{k,n}^2, \forall k, n\}$. It often assumed the fading distribution is known, see e.g. [42][43][44], often following a Rayleigh distribution [45]. If the fading distribution is unknown, it can be estimated in real time, see e.g. [46][47], although this adds additional complexity. Errors introduced by estimation errors are important, but are outside the scope of this paper.

Combine slow fading, shadowing, and pathloss for user k into a single random variable ψ_k, whose distribution is presumed known; denote $\boldsymbol{\psi} = \{\psi_k, \forall k\}$. Similarly, it is often assumed that these distributions are known.

Correspondingly, it will be helpful to think of the power allocations as random variables $p_{k,n}$ and the achieved average rates as random variables S_k that are functions of the random variables $\boldsymbol{\alpha^2}$ and $\boldsymbol{\psi}$ and of the resource allocation policy.

A key question is what to base the decision variables $\{\overline{\beta}_{k,t}, \forall k, t\}$ upon. The fluctuations in fast fading will largely average out during a group-of-pictures, whereas fluctuations in slow fading, shadowing, and pathloss will not. Thus the average rate price is principally influenced by combined path loss and shadowing, and not by fast fading. We propose that $\overline{\beta}_{k,t}$ should be a function of ψ_k, denoted $\beta_k(\psi_k)$. The decision variables $\{\overline{\beta}_{k,t}, \forall k, t\}$ are thus determined by a choice of a set of functions $\{\beta_k(\psi_k), \forall k\}$. Thus, within a group of pictures, a user can in some sense wait for a good channel on the basis of fast fading, but not on the basis of slow fading, shadowing, and pathloss. However, combined slow fading, shadowing and pathloss can be considered by the resource allocation algorithm in a manner that substantially affects the resulting outage.

It remains to specify the optimization metric and constraints to use for determination of these functions. The optimization metric can be written as a statistical average over an infinite time horizon:

$$\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T} U_k(S_{k,t}) \rightarrow \sum_{k=1}^{K} E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}} U_k(S_k) \quad (16)$$

The resource allocation policy is determined by $\{\beta_k(\psi_k), \forall k\}$. Denote the set of shadowing and pathlosses for all users except user k by $\boldsymbol{\psi}_{-k} = \{\psi_{\hat{k}}, \forall \hat{k} \neq k\}$. User's k's achieved average rate S_k depends not only on its ψ_k

but also upon other users' shadowing and pathlosses and on all users' fast fading. Since fast fading will largely average out over a group-of-pictures, the expectation over $\boldsymbol{\alpha^2}$ can be brought inside the utility function without loss of accuracy. Other users' shadowing and pathlosses affect the resource allocation through determination of the power price $\mu_{k,t}$, but consideration of them in determination of the resource allocation policy $\{\beta_k(\psi_k), \forall k\}$ is too complex; thus we also propose to bring this portion of the expectation inside the utility function:

$$E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}} U_k(S_k) \approx E_{\psi_k} U_k(E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} S_k) \tag{17}$$

Similarly, the total power can be written as a statistical average over an infinite time horizon:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n,t} \to \sum_{k=1}^{K} E_{\psi_k} \sum_{n=1}^{N} E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(p_{k,n}) \tag{18}$$

and the outage probability can be written as a statistical average over an infinite time horizon:

$$Pr(S_{k,t} < S_k^{'}) \to E_{\psi_k} Pr_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(S_k < S_k^{'}) \tag{19}$$

The optimization problem using statistical averages is thus to determine the set of functions $\{\beta_k(\psi_k), \forall k\}$ that maximize the average utility subject to an average power constraint and to probability of outage constraints on each user:

$$\begin{aligned} &\max_{\{\beta_k(\psi_k),\forall k\}} \sum_{k=1}^{K} E_{\psi_k} U_k(E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} S_k) \\ &\text{s.t.} \sum_{k=1}^{K} E_{\psi_k} \sum_{n=1}^{N} E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(p_{k,n}) \le P \\ &E_{\psi_k} Pr_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(S_k < S_k^{'}) \le \overline{Pr}, \forall k \end{aligned} \tag{20}$$

C. Quantization

The optimization problem in (20) is causal. Offline, it requires determination of the set of average rate price functions $\{\beta_k(\psi_k), \forall k\}$. Then online, these functions are used in conjunction with (11)-(12), with $\overline{\beta}_{k,t}$ replaced by $\beta_k(\psi_k)$, to determine power and subcarrier allocations. The remaining issue is that $\{\beta_k(\psi_k), \forall k\}$ are continuous functions, and there is no straightforward manner to optimize them. In this subsection, we thus propose that ψ_k be quantized.

We assume that the distribution of ψ_k is independent of k. Partition the domain of ψ_k into M slices, with the lower bound of slice m denoted by ψ^m. This will allow the determination of $\{\beta_k(\psi_k), \forall k\}$ to be reduced to determination of a finite set $\{\beta_k^m, \forall k, m\}$, where β_k^m substitutes for $\beta_k(\psi^m)$.

Denote the probability of slice m by $q_m = Pr(\psi_k = \psi^m)$. When in slice m, denote the corresponding channel by $|H_{k,n}^m|^2 = \alpha_{k,n}^2 \psi_k^m$.

With this quantization, the equation that determines user rates (15) is transformed into:

$$d_k^m = \max\{S_k^{'}, \arg\max_{d_k^m}[U_k(d_k^m)q_m - \beta_k^m d_k^m]\} \tag{21}$$

and the equation that determines power allocation (11) is transformed into:

$$p_{k,n}^m = \frac{B\beta_k^m}{\mu \ln 2} - \frac{\sigma^2 + I}{|H_{k,n}^m|^2} \tag{22}$$

where μ is the Lagrangian multiplier associated with the average power constraint in (20). The equation that determines subcarrier allocations (12) is transformed into:

$$\arg\max_{k} \Phi_{k,n}^m \tag{23}$$

where

$$\begin{aligned} \Phi_{k,n}^m &= \beta_k^m B \left[\log_2\left(\frac{B\beta_k^m}{\mu \ln 2}\frac{|H_{k,n}^m|^2}{\sigma^2 + I}\right)\right]^+ \\ &- \mu\left(\frac{B\beta_k^m}{\mu \ln 2} - \frac{\sigma^2 + I}{|H_{k,n}^m|^2}\right)^+ \end{aligned} \tag{24}$$

Denote the resulting average rate for user k while in slice m by S_k^m. The optimization problem thus becomes:

$$\begin{aligned} &\max_{\{\beta_k^m,\forall k,m\}} \sum_{k=1}^{K}\sum_{m=1}^{M} U_k(E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} S_k^m) q_m \\ &\text{s.t.} \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{n=1}^{N} E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}\left(p_{k,n}^m\right) q_m \le P \\ &\sum_{m=1}^{M} Pr_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(S_k^m < S_k^{'}) q_m \le \overline{Pr}, \forall k \end{aligned} \tag{25}$$

D. Algorithm

In this subsection, we outline an algorithm that may be used to iteratively solve problem (25). We start with calculation of the expected average rate $E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} S_k^m$. Denote the instantaneous rate of user k on subcarrier n while in slice m by $r_{k,n}^m$. Then:

$$\begin{aligned} E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} S_k^m = E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} R_k^m = \sum_{n=1}^{N} E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}\left(r_{k,n}^m\right) \quad (26) \\ = \sum_{n=1}^{N} E_{\boldsymbol{\alpha^2}} B \log_2\left(1 + p_{k,n}^m \frac{\alpha_{k,n}^2 \psi^m}{\sigma^2 + I}\right) Pr\{n \Rightarrow k, m\} \end{aligned}$$

where $Pr\{n \Rightarrow k, m\}$ is the probability that subcarrier n is assigned to user k given that user k is in slice m. For a given β_k^m and μ, the equation that calculates the probability is derived in Appendix 1.

Similarly, the average power in (25) can be expressed as:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{n=1}^{N} E_{\boldsymbol{\alpha^2}}(p_{k,n}^m Pr\{n \Rightarrow k, m\}) q_m \tag{27}$$

The average rate of user k while in slice m, S_k^m, is a sum over multiple time slots of the sum over multiple subcarriers of $r_{k,n}^m$. In Appendix 2, we use the Central Limit Theorem to approximate this sum, conditioned on $\boldsymbol{\psi}_{-k}$, by a Gaussian distribution: $S_{k,\boldsymbol{\psi}_{-k}}^m \sim \mathcal{N}(N E_{\boldsymbol{\alpha^2}}(r_{k,n}^m|\boldsymbol{\psi}_{-k}), N \cdot (\delta_{k,\boldsymbol{\psi}_{-k}}^m)^2/W)$

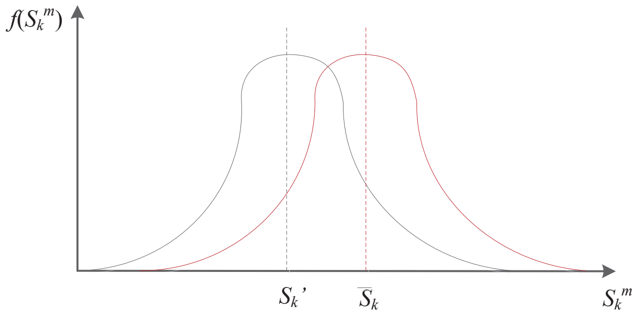

Fig. 2. Use of an average rate margin shifts the distribution of average rate S_k^m

where $S^m_{k,\boldsymbol{\psi}_{-k}}$ is the average of user k in slice m for a given of $\boldsymbol{\psi}_{-k}$ and $N \cdot (\delta^m_{k,\boldsymbol{\psi}_{-k}})^2/W$ is the variance. Removing the conditioning on $\boldsymbol{\psi}_{-k}$ thus gives that S_k^m is a mixture Gaussian distribution, with the number of terms equal to the number of possible values for $\boldsymbol{\psi}_{-k}$, i.e. M^{K-1}. The calculation of the mean value of S_k^m only requires knowledge of $Pr\{n \Rightarrow k, m\}$, whose expansion includes M^{k-1} terms. However, the calculation of the outage probability requires the tail probability of each the terms. If M^{K-1} is too large, then the calculation of this tail probability is cumbersome; in this case, we propose increasing the target average rate from $S_k^{'}$ to $\bar{S}_k = \eta S_k^{'}$ where $\eta \geq 1$ is a margin sufficient to ensure acceptable outage. By varying η, the probability density function of S_k^m can be adjusted as shown in Fig.2, allowing the outage requirement to be satisfied.

The offline portion of the iterative algorithm needs to choose $\{\beta_k^m, \forall k, m\}$ based on distribution of combined shadowing and pathloss ψ_k and the distribution of fast fading $\boldsymbol{\alpha^2}$ so as to maximize total average utility. Let i denote the iteration number. The average rate price for user k in slice m, β_k^m, should be set in conjunction with a corresponding target average rate $d_k^{m,i}$ so that user k's expected average rate is equal to its target average rate. This requires an offline estimation of the power price μ.

We propose a subgradient method with bounds to iteratively update $\{\beta_k^m, \forall k, m\}$:

$$\beta_k^{m,i+1} = \max[\min(\beta_k^{m,i} + s_P^i(d_k^{m,i} - E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(R_k^{m,i})), \overline{\beta}_k^m), \underline{\beta}] \tag{28}$$

where s_P^i is a suitable step size, $d_k^{m,i}$ is the target average rate at iteration i, the lower bound $\underline{\beta}$ can be set to a small suitable constant, and the upper bound $\overline{\beta}_k^m$ can be set to a small multiple of $\overline{\lambda}_k^m$ where $\overline{\lambda}_k^m$ can be derived from (21) as

$$\overline{\lambda}_k^m = q_m dU_k(S_{k,t})/dS_{k,t}|(S_{k,t} = S_k^{'}) \tag{29}$$

The outage price is included in β_k^m. If $\beta_k^m > \overline{\lambda}_k^m$, then the outage price of user k given slice m is $\nu_k^m = \beta_k^m - \overline{\lambda}_k^m$. The target average rate $d_k^{m,i}$ is determined by

$$d_k^{m,i} = \max\left\{\overline{S}_k, \arg\max_{d_k^{m,i}}[U_k(d_k^{m,i})q_m - \beta_k^m d_k^m]\right\}, \tag{30}$$

The iteration is terminated when:

$$\begin{aligned} &\beta_k^{m,i+1} - \beta_k^{m,i} < \rho \; \forall k \\ &\text{or } E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(R_k^{m,i+1}) = E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(R_k^{m,i}) \; \forall k \end{aligned} \tag{31}$$

where ρ is a small constant.

For the update of μ, we propose a bisection algorithm:

$$\begin{aligned} &\text{If } \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{n=1}^{N} E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}\left(p_{k,n}^m\right) q_m - P > 0, \\ &\text{then } \mu^{i+1} = (\mu^i + \overline{\mu}^i)/2, \underline{\mu}^{i+1} = \mu^i, \overline{\mu}^{i+1} = \overline{\mu}^i \\ &\text{else } \mu^{i+1} = (\mu^i + \underline{\mu}^i)/2, \underline{\mu}^{i+1} = \underline{\mu}^i, \overline{\mu}^{i+1} = \mu^i \end{aligned} \tag{32}$$

where the initial lower bound μ^0 can be set to a small suitable constant and upper bound $\overline{\mu}^0$ can be set to a big suitable constant. The iteration for μ is terminated when:

$$\mu^{i+1} - \mu^i < \epsilon \tag{33}$$

where ϵ is a small constant. The offline algorithm is outlined in Table II.

The average rate margin can be set according to the outage performance. It includes two steps. The first step can be completed off-line. We propose to approximate the mixture distribution of S_k^m using a skewed gaussian distribution [48]. This produces a reasonably good fit, since although the number of terms in the mixture gaussian distribution is very large, the overall distribution remains fairly smooth. (The estimation method of the variance of S_k^m, i.e. of δ_k^m, is shown in Appendix 3.) We start with $\eta = 1$. Then the outage of S_k^m is estimated as follows:

$$\begin{aligned} Pr_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(S_k^m < \bar{S}_k) \;=\; & F\left(\frac{\bar{S}_k - E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} S_k^m}{\sqrt{2}\delta_k^m}\right) \\ & - \frac{1}{6} f\left(\frac{\bar{S}_k - E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} S_k^m}{\sqrt{2}\delta_k^m}\right) \cdot skew \end{aligned} \tag{34}$$

where $skew$ is the skewness [48], i.e. the third standardized moment, of S_k^m. The outage is then increased in steps of $\Delta\eta$ and the outage estimates of S_k^m updated using (34) until the estimated average outage $\sum_{m=1}^{M} Pr_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}(S_k^m < S_k^{'}) q_m$ is below $\overline{Pr}$. The second step is online adjustment. We observed the actual outage, and if the estimated η is insufficient to meet the desired threshold, then it can be further increased in steps of $\Delta\eta$ until the threshold is met.

At the end of the offline algorithm, we have determined $\{\beta_k^m, \forall k, m\}$. We discard μ, since it will be recalculated in the online algorithm based on actual demand. The complexity of subgradient updates is polynomial in the dimension of the dual problem, and thus the complexity of the offline algorithm is polynomial in the number of users K.

Then online, these functions are used in conjunction with (11)-(12) to determine power and subcarrier allocations. This online algorithm allocates resources slot by slot based on instantaneous $\psi_{k,n,t}$ and fast fading $\alpha_{k,t}$. While the quantization means that only a limited set of average rate prices are determined in the offline algorithm, we can improve performance

TABLE II. OFFLINE ALGORITHM TO SELECT THE AVERAGE RATE PRICES $\{\beta_k^m\}$

```
Initialize $\mu^0 = \underline{\mu}^0$, and $\beta_k^m = \underline{\beta}\ \ \forall k$
Repeat
  Repeat
    Calculate total expected power by (27)
    Update $\mu$ using (32)
  Until (33)
  Calculate $d_k^{m,i}$by (30)
  Update $\boldsymbol{\beta}$ using (28)
Until (31)
$\{\beta_k^m\}$ are stored.
```

TABLE III. ONLINE ALGORITHM TO ALLOCATE RESOURCES

```
Every slot, initialize $\mu_t^0 = \underline{\mu}_t^0$
Update user's pathloss and get the approximation of $\bar{\beta}_{k,t}$ $\forall k$ by equation (35)
Repeat
  Allocate subcarrier and power by (11) and (12)
  Update $\mu_t$ using (32), but with $\mu$ changed to $\mu_t$
Until (33)
```

using a user's actual $\psi_{k,t}$ by interpolating between neighboring β_k^m. At current slot $\psi_{k,t} = \gamma_{k,t} PL_{k,t}$, then an interpolated average rate price can be set to:

$$\overline{\beta}_{k,t} \approx \begin{cases} \beta_k^m - (\beta_k^m - \beta_k^{m-1}) \cdot \dfrac{\psi_{k,t} - \psi^m}{\psi^{m-1} - \psi^m} \\ \quad , \text{ if } \ \psi^m < \psi_{k,t} < \psi^{m-1}, \text{ or } \psi_{k,t} < \psi^M \\ \beta_k^m - (\beta_k^{m+1} - \beta_k^m) \cdot \dfrac{\psi_{k,t} - \psi^m}{\psi^m - \psi^{m+1}} \quad , \text{ if } \ \psi_{k,t} > \psi^1 \end{cases} \tag{35}$$

if $\psi_{k,t}$ and ψ^m are expressed in dB.

The power price μ_t can be iteratively updated to satisfy the power constraint in each slot using a similar bisection algorithm as in equation (32), but with $\sum_{m=1}^M \sum_{k=1}^K \sum_{n=1}^N E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}} \left(p_{k,n}^m\right) q_m$ changed to $\sum_{k=1}^K \sum_{n=1}^N p_{k,n,t}$ and μ changed to μ_t. The online algorithm is outlined in Table III. The online algorithm is a standard water filling algorithm, and hence its complexity is linear in the total number of subcarriers N [49]. The updating of the power price μ_t is done at the base station. If $N = 1000$ and $K = 40$, we have tested that for a given set of average rate prices $\{\overline{\beta}_{k,t}, \forall k\}$, the online algorithm typically needs 20-30 iterations to converge. Thus the computational cost of the online algorithm may be acceptable in OFDMA systems.

V. SIMULATION RESULTS

In this section, we examine the performance of the proposed iterative algorithms via simulation. We adopt the parameters of an OFDMA scenario as described in the 3GPP standard [41]. The system bandwidth is 10MHz and total number of subcarriers is 1000. The base station transmission power is 46dBm with an antenna gain of 15 dbi. The inter-cell distance is 750m. We assume 40% of the total resources are assigned to video users, with the remaining 60% assigned to data and voice users. Thus the total number of subcarriers for video users is $N = 400$ and the power constraint $P = 42$dbm. All users move at a constant speed of 10km/h, with direction determined by a random walk [50]. The pathloss is determined by:

$$PL_k = 128.1 + 37.6 \log_{10}(l_k) + 21 \log_{10}(f_c/2.0) \ \ \text{dB} \tag{36}$$

where l_k is the distance from the user to the base station in kilometers and $f_c = 2$ is the central frequency in Ghz. If users stay in the cell for a long period of time and/or if users' speed is high enough, these users are approximately uniformly distributed in the cell. Thus the cumulative distribution function of l_k is

$$F(l_k) = \frac{l_k^2 - l_{min}^2}{l_{max}^2 - l_{min}^2} \tag{37}$$

where $l_{min} = 0.01$km is the minimum distance to base station and $l_{max} = 0.25$km is the radius of the cell. The shadowing follows a lognormal distribution with mean value 0dB and variance 10dB [41]. Based on these distributions, one can derive the distribution of ψ_k. The domain of $\log(\psi_k)$ is partitioned into $M = 45$ slices. The step size of the first 30 slices is 1.9dB and that of the last 15 slices is 1.2dB.

Fast fading is assumed to follow a Rayleigh distribution, and thus the gain of fast fading follows an Exponential distribution with mean value 1 [45]. The length of one time slot is 1 ms [51]. If we consider a significant phase change to be π/4, then the coherence time is 3.4ms (see e.g. [52], pg. 31). Thus in the simulation, we generate fast fading every 3ms independent of previous fading.

We model the interference by:

$$I = 10^{P/10}/N \cdot 10 \cdot 10^{-(PL_{edge} + Loss_{penetration} - Gain_{antenna})/10} \cdot 0.8 \tag{38}$$

where PL_{edge} is the pathloss at the edge of one cell, $Loss_{penetration} = 10$db is loss from penetration, and $Gain_{antenna} = 15$db is the antenna gain. The resulting interference I on each subcarrier is $2.8495 * 10^{-9}$mW.

The thermal noise on each subcarrier is

$$\sigma^2 = 10^{N_0/10} \cdot B = 3.9811 * 10^{-14}\text{mW} \tag{39}$$

where $N_0 = -174 dbm/Hz$ is the thermal noise density, and $B = 10000Hz$ is the bandwidth of each subcarrier. Because $I \gg \sigma^2$, we set the sum of interference and thermal noise on each subcarrier $I + \sigma^2 = 2.8495 * 10^{-9}$mW.

All users have the same utility function, given by:

$$U_k(S_{k,t}) = \begin{cases} a(S_{k,t}/4)^2, & \text{if } S_{k,t} < 240\text{kbps} \\ c(S_{k,t}/4 + b)^{1/3}, & \text{else} \end{cases} \tag{40}$$

where $a = 4/5 * (2/5)^{1/3}/(12/5)^2$, $b = -2$, $c = 4/5$ and $S_{k,t}$ is expressed in units of 100kbps. The average rate at the maximum average utility $S_k^{'} = 300$ kbps. We allow 3% outage, i.e. $\overline{Pr} = 0.03$. We set the moving speed of each user as 10 km/h and simulate 45 minutes of real time.

We adopt common MPEG parameters: a video encoded at 30 frames per second using a group-of-pictures consisting of 12 frames [53]. We set the length of the time window equal to one group-of-pictures, resulting in W = 133. We set the average rate margin step size $\Delta\eta = 0.03$.

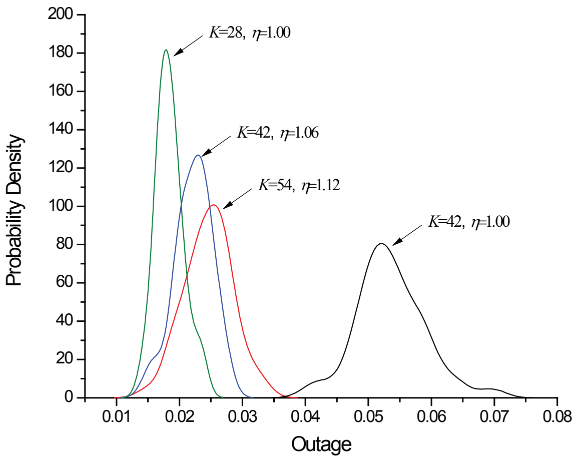

Fig. 3. Distribution of Outage

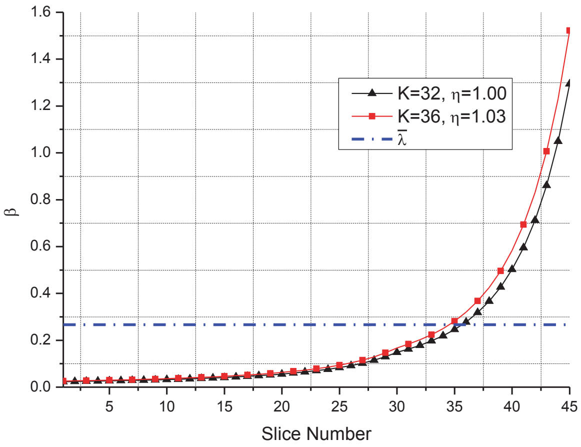

Fig. 4. Average rate price β^m versus slice number

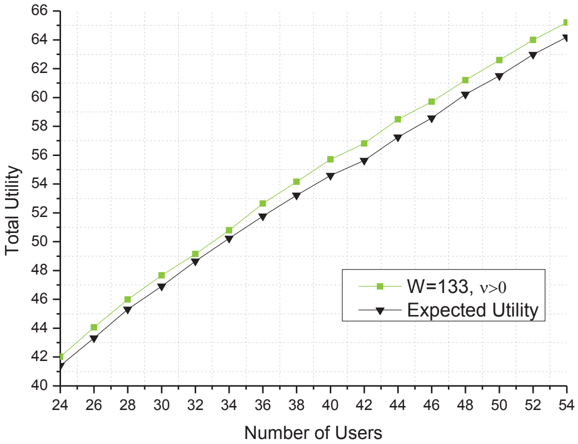

Fig. 5. Total utility versus the number of users

The simulation parameters result in a system that is near capacity. To illustrate the competition between users for resources, consider the following particular situation: (1) 40 users are uniformly distributed in the cell; (2) they have mean and constant shadowing, slow fading, and fast fading; and (3) each subcarrier is allocated the same power. Suppose that the goal is to allocate subcarriers to maximize the minimum rate. Then the user with the minimum rate achieves $S_k = 298$ kbps, and most users achieve a rate near 320kbps. However, when mobility, fading, and shadowing are incorporated, the average rates vary in time, and thus satisfying outage requirements is more difficult.

Because all users have the same utility function, we consider the average outage over all users. The estimated outage and actual outage performance is listed in Table IV. For a fixed average rate margin η, as the number of users increases, the average outage increases. Thus as the number of users increases, η needs to increase to satisfy the average outage requirement. For all cases except for $K = 52$, the off-line estimated η is sufficient to satisfy the average outage constraint. However, when $K = 52$, the system approaches its capacity limit and the off-line estimated η is not sufficient to guarantee the average outage requirement; in this case, the online adjustment increases η to bring the observed average outage below the required threshold.

Recall that the outage constraint is a bound on the expected outage for each user. Individual users will experience outage over the duration of their connections that vary from their expected outage dependent upon the user's actual shadowing and pathloss during their connection. In Fig. 3, we show the distribution of user's outage for three combinations of number of users K and rate margin η: (28,1.00), (42,1.06), and (54,1.12). As discussed above, the distribution of S_k^m can be adjusted using the rate margin, and thus the average outage correspondingly depends on the rate margin. For comparison, the figure shows the distribution of user's outage for $K = 42$ when the rate margin $\eta = 1.00$, which is not sufficient to meet the average outage constraint.

The right column of Table IV also shows the 3^{rd} percentile and 97^{th} percentile of outage. The variation in outage amongst users is decreasing with the connection duration, and would converge to zero as the connection duration increases to infinity, since the outage would converge to its expected value. However, for finite connection durations, the 97^{th} percentile outage often slightly exceeds the 3% bound on the expected outage for each user. If one wanted to further reduce the percentage of users who experience greater than 3% outage during their connection, one might be able to accomplish this by further increasing the rate margin (by an amount that is decreasing with the connection duration); however, this will come at the cost of a reduction in the average utility.

The average rate prices β^m are shown in Fig. 4 as a function of the slice number m for two scenarios. When $\beta^m > \bar{\lambda}$, the outage price is greater than 0. At higher slices, ψ^m decreases due to combined slow fading, shadowing, and pathloss, and the resulting outage price increases.

The total utility as a function of the number of users K is shown in Fig. 5. Total utility is an increasing concave function of the number of users within the considered range, indicating that users are generally in the concave portion of their utility curves. We also plot the expected utility in the optimization metric in (25) that results from the offline algorithm, labelled in the figure as "Expected Utility". The real utility is slightly above the expected utility, which shows

TABLE IV. AVERAGE RATE MARGIN AND OUTAGE

Number of Users	Estimated η	Estimated Average Outage	Online adjusted η	Real Average Outage	3rd Percentile	97th Percentile
24	1.00	0.014	1.00	0.018	0.015	0.022
26	1.00	0.016	1.00	0.021	0.016	0.024
28	1.00	0.019	1.00	0.022	0.018	0.026
30	1.00	0.021	1.00	0.025	0.020	0.028
32	1.00	0.025	1.00	0.028	0.023	0.032
34	1.00	0.029	1.00	0.029	0.024	0.033
36	1.03	0.018	1.03	0.022	0.018	0.028
38	1.03	0.022	1.03	0.026	0.021	0.032
40	1.03	0.027	1.03	0.028	0.023	0.033
42	1.06	0.021	1.06	0.024	0.019	0.028
44	1.06	0.024	1.06	0.025	0.020	0.029
46	1.06	0.027	1.09	0.026	0.020	0.030
48	1.09	0.028	1.09	0.026	0.022	0.030
50	1.09	0.030	1.09	0.028	0.023	0.031
52	1.09	0.029	1.12	0.025	0.019	0.031
54	1.12	0.029	1.12	0.029	0.024	0.034

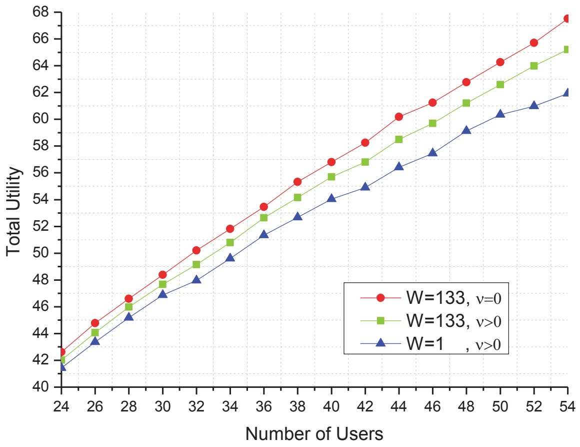

Fig. 6. Total utility under various policies

that only small errors are introduced by the statistical average model using quantization. The difference can be explained as follows. When we estimate total utility in the offline algorithm, we use a fixed estimated power price μ. However, the online algorithm updates the power price μ_t every time slot, thereby more efficiently utilizing knowledge of the channel fading of each user and providing a performance gain.

For comparison, in Fig. 6, we also plot the utility that would be earned from a policy that allocates resources without regard to the time window, i.e. it attempts to maximize average utility of instantaneous rate. This is labelled in the figure by $W = 1$. The difference between the $W = 1$ and $W = 133$ curves represents the increase in utility gained by allocating resources in a more flexible manner within a group-of-pictures. When $W = 1$, a user evaluates application performance based on the instantaneous rate of one time slot. Thus, when the channel in a time slot is poor, the system must decide whether to allocate a very large amount of power and subcarriers. In contrast when $W = 133$, when a channel is poor the system may examine past achieved rates and likely future rates within the time window, and often will allocate power and subcarriers in a more moderate and efficient manner. As the number of users increases, the advantage of allocating resources over a longer time window increases, since the larger number of users results in a wider variety of channel qualities.

In Fig. 6, we also compare the utility for $W = 133$ with outage pricing to the utility that would be earned from a policy that allocates resources using a window $W = 133$ but without outage pricing, i.e. $\nu = 0$. The existence of an average rate margin decreases the total utility across the entire range of number of users shown. The average rate margin is used to allocate additional power and subcarriers to users who would otherwise violate outage constraints. The result of this policy is that users with poor channels who are in danger of achieving rates less than $S_k^{'}$ during a time window will see improved performance, but this comes at the cost of decreased performance for users with rates above $S_k^{'}$. Together, this results in lower utility but decreased outage. With the increasing of total user number, the gap between two curves also increases. This is because more users will be in outage and more resources need to be allocated to users with poor channels. The performance loss of users with rates above $S_k^{'}$ also increases.

To further understand the effect of the choice of window size, in Fig. 7 we plot the total user utility versus the number of users for a variety of window sizes. We would expect that total user utility should be increasing with the window size. We might also expect that total user utility should be concave in the window size, i.e. the marginal benefit decreasing as the window size increases, at least when the window size exceeds the correlation time of fast fading. The simulation results confirm that utility is increasing with window size; however, the differences between the utility resulting from window sizes of 33, 99, and 133 are too small (compared to confidence intervals) to judge the prediction of decreasing returns.

In Fig. 8, we illustrate the average utility per user versus the number of users, as well as the 3^{rd} and 97^{th} percentile

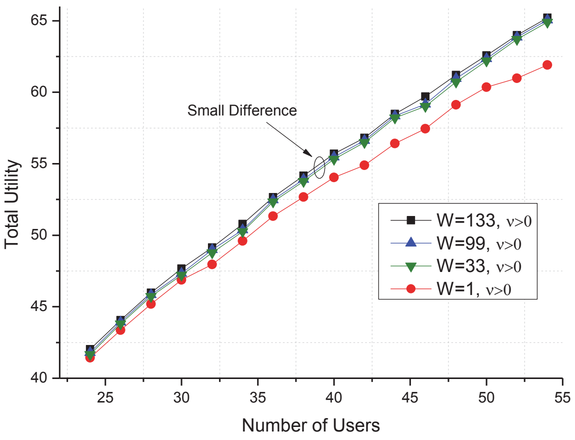

Fig. 7. Total utility with different window sizes

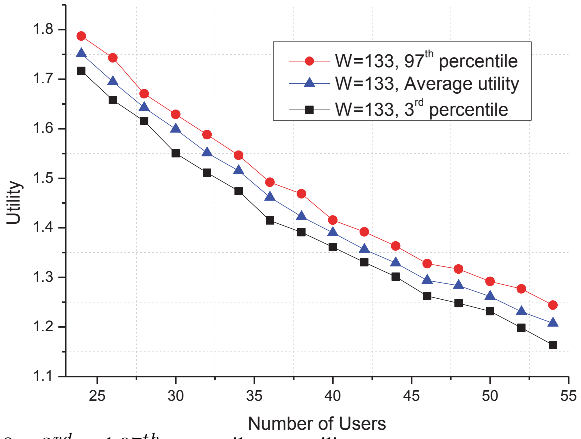

Fig. 8. 3^{rd} and 97^{th} percentile user utility

of user utility. As the number of users increases, the resources allocated to each user decreases and thus these three values all decrease. The gaps between these three curves mostly depends on the length of time a user resides in the system, rather than on the number of users.

The resulting outage under each policy is shown in Table V, also as a function of the number of users. As expected, the probability of outage increases with the number of users. If the average rate margin were removed, i.e. $W = 133$ and $\nu = 0$, we found above that total utility increases. This occurs, however, at the cost of increased outage. Indeed, when $K = 32$ the average outage has exceeded the 3% threshold. For comparison, the table also shows the outage that would result from a policy that allocates resources without regard to the time window, i.e. $W = 1$. Above we found that this approach decreases utility due to the lack of flexibility of allocating resources within a time window. Here we find that it severely violates the 3% outage constraint even when $K = 32$.

Finally, we compare our algorithm with other two policies commonly used in the literature: (1) allocating the same average resources to each user and (2) allocating resources so that each user achieves the same average rate. (These two polices are described in detail in Appendix 4.) We set $K = 32$ and show the average rate in each slice achieved under each policy in Fig 9. The average outage under the policy using an average rate margin is 0.028, as previously shown in Table V. The policy of allocating the same average resources to each user results in a lower average rate both for users in low numbered and in high number slices, and also results in an average outage of 0.052. The policy of allocating resources so that each user achieves the same average rate results in a nearly steady average rate among different slices, and also results in an average outage of 0.051. Both of these commonly used policies thus result in an average outage that exceeds the desired outage threshold. In addition, both achieve lower total user utility than our algorithm. The average resource allocation policy underachieves because average power allocation cannot fully utilize channel knowledge, and the average rate resource allocation policy underachieves because it attempts to be fair between users in different slices rather than to maximize total utility.

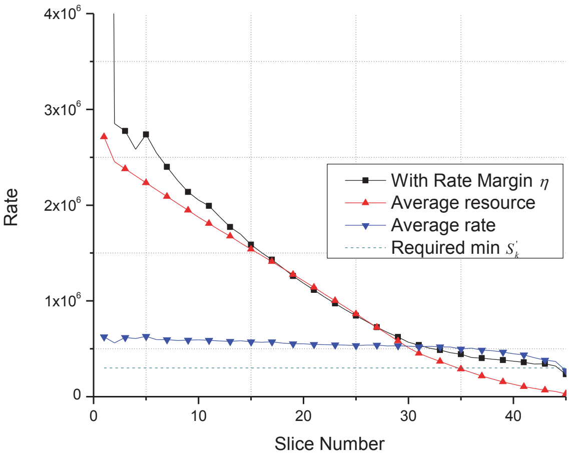

Fig. 9. Average rate under various policies

VI. Conclusion

In this paper, we consider resource allocation for mobile video-conferencing applications. We model the utility of such applications as a function of the average rate within a group-of-pictures. The goal is to maximize the total expected user utility subject to outage constraints. The key challenge is how to a design resource allocation algorithm that satisfies the outage constraints under user mobility.

First, we pose and solve a non-causal problem in which all future channel information is known. We solve the dual optimization problem, and obtain the optimal power and subcarrier allocations. We find that these allocations could be implemented if the network charged the user a price per unit average rate and the user selected the average rate that maximized its surplus. However, the optimal price per unit average rate is the sum of a shadow cost for average rate for users not in outage and a shadow cost associated with the outage constraints. The problem is that the latter outage price can only be determined with knowledge of all future channel information.

We then turn to posing and solving a causal problem. To formulate a causal resource allocation policy, we propose

TABLE V. OUTAGE UNDER VARIOUS POLICIES

Number of Users	W=133		W=1		W=133, $\nu = 0$
	Average Outage	η	Average Outage	η	Average Outage
24	0.018	1.00	0.044	1.00	0.023
26	0.021	1.00	0.048	1.00	0.025
28	0.022	1.00	0.052	1.00	0.027
30	0.025	1.00	0.055	1.00	0.032
32	0.028	1.00	0.060	1.00	0.037
34	0.029	1.00	0.062	1.00	0.041
36	0.022	1.03	0.060	1.03	0.044
38	0.026	1.03	0.070	1.03	0.048
40	0.028	1.03	0.078	1.03	0.059
42	0.024	1.06	0.089	1.06	0.057
44	0.025	1.06	0.109	1.06	0.060
46	0.026	1.06	0.121	1.06	0.067
48	0.026	1.09	0.130	1.09	0.072
50	0.028	1.09	0.140	1.09	0.077
52	0.025	1.09	0.155	1.09	0.081
54	0.029	1.12	0.170	1.12	0.088

basing the average rate price entirely on a user's channel and transforming the problem using statistical averages over an infinite time horizon. We illustrate how the average rate prices can be set as a function of the combined shadowing and pathloss, and how to quantize these variables to make the computation feasible. An average rate margin is reserved to satisfy the outage constraint. We give an example of online and offline algorithms that can iteratively determine near-optimal resource allocations.

The performance of our algorithm is illustrated by simulation results that show that the proposed algorithm efficiently determines resource allocations that satisfy average outage probability constraints, and that only small errors are introduced by the statistical average model using quantization. We find that a significant increase in total utility can be achieved by allocating resources in a more flexible manner within a group-of-pictures. The existence of outage constraints limits the capacity of the system. Simulation results quantify both these limits and the reduction in total utility sacrificed in order to satisfy outage constraints.

There are several extensions of this work that would be interesting. Perhaps the most important may be an extension from the single cell analysis provided here to an analysis that considers multiple cells. There is a deep and rich research literature on resource allocation for circuit-switched wireless networks that proposes methods to move network capacity to where it is most needed. However, it remains much less clear how resource can be shifted in packet-switched networks from lightly loaded to more heavily loaded cells to support semi-elastic applications such as video conferencing. We expect that such an extension would involve two elements. First, a model is required for the effect of inter-cell interference on utility, through its effect on throughput over a group-of-pictures. Second, an algorithm is required based on the ability of a network to shift capacity on a time scale matched to utility, user mobility, and connection duration.

REFERENCES

[1] Sandvine, "Global Internet phenomena report," http://www.sandvine.com, Tech. Rep., 2011.

[2] S. Shenker, "Fundamental design issues for the future Internet," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1176–1188, Sep. 1995.

[3] J. Lee, R. Mazumdar, and N. Shroff, "Downlink power allocation for multi-class wireless systems," *IEEE/ACM Trans. Netw.*, vol. 13, no. 4, pp. 854–867, Aug. 2005.

[4] P. Hande, Z. Shengyu, and C. Mung, "Distributed rate allocation for inelastic flows," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1240–1253, Dec. 2007.

[5] G. Abbas, A. Nagar, and H. Tawfik, "On unified quality of service resource allocation scheme with fair and scalable traffic management for multiclass internet services," *Communications, IET*, vol. 5, no. 16, pp. 2371–2385, 2011.

[6] J. Jin, A. Sridharan, B. Krishnamachari, and M. Palaniswami, "Handling inelastic traffic in wireless sensor networks," *Selected Areas in Communications, IEEE Journal on*, vol. 28, no. 7, pp. 1105–1115, 2010.

[7] C. Yang and S. Jordan, "Downlink user selection and resource allocation for semi-elastic flows in an OFDM cell," *ACM Wireless Networks*, vol. 19, no. 6, pp. 1407–1421, 2013.

[8] A. Sehati, M. Talebi, and A. Khonsari, "NUM-based rate allocation for streaming traffic via sequential convex programming," in *Communications (ICC), 2012 IEEE International Conference on*, June 2012, pp. 1239–1243.

[9] F. Wang, X. Liao, S. Guo, H. Huang, and T. Huang, "Dynamic rate and power allocation in wireless ad hoc networks with elastic and inelastic traffic," *Wireless Personal Communications*, vol. 70, no. 1, pp. 435–457, 2013.

[10] P. Vo, S. Lee, and C. Hong, "The random access NUM with multiclass traffic," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, p. 242, 2012.

[11] C. Yang and S. Jordan, "Power and rate allocation for video conferencing in cellular networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, p. 31, 2013.

[12] S. Ni, Y. Liang, and S.-G. Häggman, "Outage probability in GSM-GPRS cellular systems with and without frequency hopping," *Wirel. Pers. Commun.*, vol. 14, no. 3, pp. 215–234, Sep. 2000. [Online]. Available: http://dx.doi.org/10.1023/A:1008951318876

[13] C. Fischione, M. Butussi, K. Johansson, and M. D'angelo, "Power and rate control with outage constraints in CDMA wireless networks,"

Communications, IEEE Transactions on, vol. 57, no. 8, pp. 2225 –2229, Aug. 2009.

[14] C. Zarakovitis, Q. Ni, D. Skordoulis, and M. Hadjinicolaou, "Power-efficient cross-layer design for OFDMA systems with heterogeneous QoS, imperfect CSI, and outage considerations," *Vehicular Technology, IEEE Transactions on*, vol. 61, no. 2, pp. 781 –798, Feb. 2012.

[15] J. Luo, R. Yates, and P. Spasojevic, "Service outage based power and rate allocation for parallel fading channels," *Information Theory, IEEE Transactions on*, vol. 51, no. 7, pp. 2594 –2611, July 2005.

[16] S. Ghazanfari-Rad, J.-F. Frigon, and B. Sanso, "Theoretical framework for quality of service analysis of differentiated traffic in 802.11 wireless local area networks," *Communications, IET*, vol. 6, no. 15, pp. 2326–2334, 2012.

[17] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171–178, Feb. 2003.

[18] Z. Shen, J. Andrews, and B. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[19] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[20] Y. M. Tsang and R. Cheng, "Optimal resource allocation in SDMA/multiinput-single-output/OFDM systems under QoS and power constraints," in *Wireless Communications and Networking Conference, 2004 IEEE*, vol. 3, 2004, pp. 1595–1600.

[21] A. Oborina, V. Koivunen, and T. Henttonen, "Effective SINR distribution in MIMO OFDM systems," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, 2010, pp. 511–515.

[22] R. Madan, S. Boyd, and S. Lall, "Fast algorithms for resource allocation in wireless cellular networks," *Networking, IEEE/ACM Transactions on*, vol. 18, no. 3, pp. 973–984, 2010.

[23] C. Seol, K. Cheun, and S. Hong, "A statistical inter-cell interference model for downlink cellular OFDMA networks under Log-Normal shadowing with Ricean fading," *Communications Letters, IEEE*, vol. 14, no. 11, pp. 1011–1013, 2010.

[24] J.-S. Sheu and W.-H. Sheen, "Characteristics and modelling of inter-cell interference for Orthogonal Frequency-Division Multiple access systems in multipath Rayleigh fading channels," *Communications, IET*, vol. 6, no. 17, pp. 3015–3025, 2012.

[25] H. Mansour, Y. Fallah, P. Nasiopoulos, and V. Krishnamurthy, "Dynamic resource allocation for MGS H.264/AVC video transmission over link-adaptive networks," *Multimedia, IEEE Transactions on*, vol. 11, no. 8, pp. 1478–1491, Dec. 2009.

[26] P. Li, H. Zhang, B. Zhao, and S. Rangarajan, "Scalable video multicast with adaptive modulation and coding in broadband wireless data systems," *Networking, IEEE/ACM Transactions on*, vol. 20, no. 1, pp. 57–68, Feb. 2012.

[27] H. Zhang, Y. Zheng, M. Khojastepour, and S. Rangarajan, "Cross-layer optimization for streaming scalable video over fading wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 28, no. 3, pp. 344–353, April 2010.

[28] I.-M. Kim and H.-M. Kim, "A new resource allocation scheme based on a PSNR criterion for wireless video transmission to stationary receivers over Gaussian channels," *Wireless Communications, IEEE Transactions on*, vol. 1, no. 3, pp. 393–401, Jul 2002.

[29] Q. Du and X. Zhang, "Statistical QoS provisionings for wireless unicast/multicast of multi-layer video streams," *Selected Areas in Communications, IEEE Journal on*, vol. 28, no. 3, pp. 420–433, April 2010.

[30] J. Huang, Z. Li, M. Chiang, and A. Katsaggelos, "Joint source adaptation and resource allocation for multi-user wireless video streaming," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 5, pp. 582–595, May 2008.

[31] D. Wu, S. Ci, H. Wang, and A. Katsaggelos, "Application-centric routing for video streaming over multihop wireless networks," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 12, pp. 1721–1734, Dec. 2010.

[32] C.-M. Chen, C.-W. Lin, and Y.-C. Chen, "Cross-layer packet retry limit adaptation for video transport over Wireless LANs," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 11, pp. 1448–1461, Nov. 2010.

[33] R. Trestian, A. Moldovan, C. Muntean, O. Ormond, and G. Muntean, "Quality Utility modelling for multimedia applications for Android Mobile devices," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium on*, June 2012, pp. 1–6.

[34] W.-H. Kuo and W. Liao, "Utility-based radio resource allocation for QoS traffic in wireless networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 7, pp. 2714–2722, July 2008.

[35] M. H. Cheung, A. Mohsenian-Rad, V. Wong, and R. Schober, "Random access for elastic and inelastic traffic in WLANs," *Wireless Communications, IEEE Transactions on*, vol. 9, no. 6, pp. 1861–1866, 2010.

[36] M. Tao, Y.-C. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2190–2201, Jun. 2008.

[37] T. C.-Y. Ng and W. Yu, "Joint optimization of relay strategies and resource allocations in cooperative cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 2, pp. 328–339, Feb. 2007.

[38] Z.-Q. Luo and S. Zhang, "Duality gap estimation and polynomial time approximation for optimal spectrum management," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2675–2689, 2009.

[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[40] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming Third Edition*. Springer, 2008.

[41] *E-UTRA Radio Frequency (RF) system scenarios*, 3GPP TR 36.942 v10.2.0 Std., May 2011.

[42] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *Selected Areas in Communications, IEEE Journal on*, vol. 21, no. 5, pp. 684–702, 2003.

[43] I. Wong and B. Evans, "Optimal downlink OFDMA resource allocation with linear complexity to maximize ergodic rates," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 3, pp. 962–971, 2008.

[44] G. Caire and K. Kumar, "Information theoretic foundations of adaptive coded modulation," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2274–2298, 2007.

[45] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[46] X. Wang and G. Giannakis, "Resource allocation for wireless multiuser OFDM networks," *Information Theory, IEEE Transactions on*, vol. 57, no. 7, pp. 4359 –4372, July 2011.

[47] Z. Li and X. Wang, "Utility maximization over ergodic capacity regions of fading ofdma channels," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 7, pp. 2478–2485, 2012.

[48] J. PITMAN, *Probability*. Springer-Verlag, 1993.

[49] D. Palomar and J. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *Signal Processing, IEEE Transactions on*, vol. 53, no. 2, pp. 686–695, 2005.

[50] *Selection procedures for the choice of radio transmission technologies of the UMTS*, 3GPP TR 30.03 Std., April 1998.

[51] *Evolved Universal Terrestrial Radio Access*, 3GPP TR 36.300 v10.2.0 Std., Jan. 2011.

[52] P. V. David Tse, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[53] *ITU-T recommendation H.264 : Advanced video coding for generic audiovisual services*, Std., February 2014.

VII. Appendix 1

Denote

$$\zeta_k = \frac{B\beta_k^m}{\mu \ln 2}\frac{|H_{k,n}^m|^2}{\sigma^2 + I} \tag{41}$$

The subcarrier allocation is determined by (23), which can be reorganized into:

$$\begin{aligned}\Phi_{k,n}^m &= \frac{\beta_k^m B}{\ln 2}\left\{\left[\ln\left(\frac{B\beta_k^m}{\mu \ln 2}\frac{|H_{k,n}^m|^2}{\sigma^2+I}\right)\right]^+ + \frac{\mu \ln 2}{B\beta_k^m}\frac{\sigma^2+I}{|H_{k,n}^m|^2} - 1\right\}\\ &= \frac{\beta_k^m B}{\ln 2}\left\{[\ln(\zeta_k)]^+ + \frac{1}{\zeta_k} - 1\right\}\end{aligned} \tag{42}$$

Consider another user $\hat{k}$ in slice $\hat{m}$. If user k is allocated subcarrier n, then

$$\Phi_{k,n}^m \quad > \quad \Phi_{\hat{k},n}^{\hat{m}} = \frac{\beta_{\hat{k}}^{\hat{m}} B}{\ln 2}\left\{\left[\ln\left(\zeta_{\hat{k}}\right)\right]^+ + \frac{1}{\zeta_{\hat{k}}} - 1\right\} \tag{43}$$

or equivalently

$$\Phi_{k,n}^m \ln 2/B/\beta_{\hat{k}}^{\hat{m}} + 1 > \left[\ln\left(\zeta_{\hat{k}}\right)\right]^+ + \frac{1}{\zeta_{\hat{k}}} \tag{44}$$

Directly solving this equation is complex. Consider the following approximation. Define $y = \ln(\zeta_{\hat{k}})$. If $\Phi_{k,n}^m \ln 2/B/\beta_{\hat{k}}^{\hat{m}} + 1 < 1.267$, then a Taylor expansion gives:

$$\ln\left(\zeta_{\hat{k}}\right) + \frac{1}{\zeta_{\hat{k}}} = y + e^{-y} \approx 1 + \frac{y^2}{2} = 1 + \frac{\left[\ln(\zeta_{\hat{k}})\right]^2}{2} \tag{45}$$

else

$$\ln\left(\zeta_{\hat{k}}\right) + \frac{1}{\zeta_{\hat{k}}} \approx \ln(1+\zeta_{\hat{k}}) \tag{46}$$

Substituting (45) and (46) into (44), if user k is allocated subcarrier n, then:

$$g\left(|H_{k,n}^m|^2\right) > |H_{\hat{k},n}^{\hat{m}}|^2 \tag{47}$$

where the function $g(|H_{k,n}^m|^2)$ is given by:

$$\begin{cases} e^{\sqrt{2\phi_{k,n}^m \ln 2/B/\beta_{\hat{k}}^{\hat{m}}}}(\sigma^2+I)\mu \ln 2/B/\beta_{\hat{k}}^{\hat{m}} \\ \qquad\qquad\qquad , \text{if } \ \Phi_{k,n}^m \ln 2/B/\beta_{\hat{k}}^{\hat{m}} + 1 < 1.267 \\ (e^{\phi_{k,n}^m \ln 2/B/\beta_{\hat{k}}^{\hat{m}}+1} - 1)\cdot(\sigma^2+I)\mu \ln 2/B/\beta_{\hat{k}}^{\hat{m}}, \ \text{else} \end{cases}$$

Consider all the possible combined pathloss and shadowing of user $\hat{k}$:

$$\begin{aligned}&Pr(\Phi_{k,n}^m > \Phi_{k,n}^{\hat{m}})\\ &= \sum_{\hat{m}=1}^{M} Pr\left(\Phi_{k,n}^m > \Phi_{\hat{k},n}^{\hat{m}} | \psi_{\hat{k}} = \psi^{\hat{m}}\right) q_{\hat{m}}\\ &= \sum_{\hat{m}=1}^{M} Pr\left(g(|H_{k,n}^m|^2) > |H_{\hat{k},n}^{\hat{m}}|^2 | \psi_{\hat{k}} = \psi^{\hat{m}}\right) q_{\hat{m}}\end{aligned} \tag{48}$$

Because the fading of all users are independent, the probability user K is allocated subcarrier n is

$$\prod_{\hat{k}\neq k}^{K} Pr(\Phi_{k,n}^m > \Phi_{\hat{k},n}^{\hat{m}}) \tag{49}$$

Thus for a given β_k^m and μ:

$$\begin{aligned}&Pr\{n \Rightarrow k, m\}\\ =\ &\prod_{\hat{k}\neq k}^{K}\left(\sum_{\hat{m}=1}^{M} Pr\Big(g(|H_{k,n}^m|^2) > |H_{\hat{k},n}^{\hat{m}}|^2 | \psi_{\hat{k}} = \psi^{\hat{m}}\Big) q_{\hat{m}}\right)\end{aligned} \tag{50}$$

VIII. Appendix 2

Define $S_{k,\boldsymbol{\psi}_{-k}}^m$ as the average rate of user k in slice m for a given set of $\boldsymbol{\psi}_{-k}$. The probability that user k is allocated subcarrier n for a given $\boldsymbol{\psi}_{-k}$ is

$$Pr\{n \Rightarrow k, m, \boldsymbol{\psi}_{-k}\} = \prod_{\hat{k}\neq k}^{K}\left[Pr\Big(g(|H_{k,n}^m|^2) > |H_{\hat{k},n}^{\hat{m}}|^2 | \psi_{\hat{k}}\Big)\right] \tag{51}$$

The average rate of user k on subcarrier n for a given set of $\boldsymbol{\psi}_{-k}$ is

$$\begin{aligned}&E_{\boldsymbol{\alpha}^2}\left(r_{k,n}^m | \boldsymbol{\psi}_{-k}\right)\\ =\ &E_{\boldsymbol{\alpha}^2}\left\{\left[B\log_2\left(1 + p_{k,n}^m \frac{\alpha_{k,n}^2 \psi^m}{\sigma^2+I}\right)\right] Pr\{n \Rightarrow k, m, \boldsymbol{\psi}_{-k}\}\right\}\end{aligned} \tag{52}$$

Correspondingly its second moment is:

$$\begin{aligned}&E_{\boldsymbol{\alpha}^2}\left((r_{k,n}^m)^2 | \boldsymbol{\psi}_{-k}\right)\\ =\ &E_{\boldsymbol{\alpha}^2}\left\{\left[B\log_2\left(1 + p_{k,n}^m \frac{\alpha_{k,n}^2 \psi^m}{\sigma^2+I}\right)\right]^2 Pr\{n \Rightarrow k, m, \boldsymbol{\psi}_{-k}\}\right\}\end{aligned} \tag{53}$$

and thus its variance is

$$(\delta_{k,\boldsymbol{\psi}_{-k}}^m)^2 = E_{\boldsymbol{\alpha}^2}\left((r_{k,n}^m)^2 | \boldsymbol{\psi}_{-k}\right) - \left[E_{\boldsymbol{\alpha}^2}\left(r_{k,n}^m | \boldsymbol{\psi}_{-k}\right)\right]^2 \tag{54}$$

Because $R_k^m = \sum_{n=1}^N r_{k,n}^m$, by the Central Limit Theorem, the distribution of R_k^m for a given set of $\boldsymbol{\psi}_{-k}$ is approximately $\mathcal{N}\left(NE_{\boldsymbol{\alpha}^2}\left(r_{k,n}^m | \boldsymbol{\psi}_{-k}\right), N\cdot(\delta_{k,\boldsymbol{\psi}_{-k}}^m)^2\right)$. Because $S_{k,t} = \sum_{\tau=t-W+1}^{t} R_{k,\tau}/W$, by the Central Limit Theorem, the distribution of $S_{k,t}$ is approximately $\mathcal{N}(NE_{\boldsymbol{\alpha}^2}(r_{k,n}^m | \boldsymbol{\psi}_{-k}), N\cdot(\delta_{k,\boldsymbol{\psi}_{-k}}^m)^2/W)$. Removing the conditioning on $\boldsymbol{\psi}_{-k}$, S_k^m gives a mixture Gaussian distribution with the number of terms equal to the number of possible values for $\boldsymbol{\psi}_{-k}$, i.e. M^{K-1}.

IX. Appendix 3

The variance is estimated based on $\{\beta_k^m, \forall k, m\}$ and μ. Denote $\Omega = \{\boldsymbol{\psi}_{-k}\}$ as the set of all possible values of $\boldsymbol{\psi}_{-k}$. Then the variance of the mixture distribution is

$$\begin{aligned}&(\delta_k^m)^2\\ =&\sum_{\Omega}\left[(E_{\boldsymbol{\alpha}^2}(S_k^m | \boldsymbol{\psi}_{-k}) - E_{\boldsymbol{\alpha}^2,\boldsymbol{\psi}_{-k}} S_k^m)^2 + (\delta_{k,\boldsymbol{\psi}_{-k}}^m)^2\right] Pr(\boldsymbol{\psi}_{-k})\end{aligned} \tag{55}$$

where $Pr(\boldsymbol{\psi}_{-k})$ is the probability of a specific set of $\boldsymbol{\psi}_{-k}$.

If M^{K-1} is too large, the calculation of (55) is cumbersome. Equation (55) includes two parts. The second part can be approximated as follows:

$$\begin{aligned}&\sum_{\Omega} Pr(\boldsymbol{\psi}_{-k})(\delta^m_{k,\boldsymbol{\psi}_{-k}})^2 \\ \approx\ & \frac{N}{W}\left\{E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}\left((r^m_{k,n})^2\right) - [E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}\left(r^m_{k,n}\right)]^2\right\}\end{aligned} \tag{56}$$

where the average rate of user k on subcarrier n is given in (26), and its second moment is:

$$\begin{aligned}&E_{\boldsymbol{\alpha^2},\boldsymbol{\psi}_{-k}}\left((r^m_{k,n})^2\right) \\ =\ & E_{\boldsymbol{\alpha^2}}\left\{\left[B\log_2\left(1+p^m_{k,n}\frac{\alpha^2_{k,n}\psi^m}{\sigma^2+I}\right)\right]^2 Pr\{n\Rightarrow k,m\}\right\}\end{aligned} \tag{57}$$

The first part is the variance of the mean average rate. We propose to consider limited terms rather than all M^{K-1} terms. To reduce the complexity, here we consider two situations: all $K-1$ users are in the same slice and all $K-1$ users are equally distributed in two slices. Based on this sampling, the variance of the mean average rate can be roughly estimated. Combining with equation (56), the approximation of equation (55) can be calculated.

X. Appendix 4

Average resource allocation:

Step 1: $p_{k,n,t} = P/N,\ \forall k,n$.

Step 2: $\mathbf{C} = \{n|$ subcarrier n has not been assigned$\}$, $\mathbf{B} = \{k|$user k has not been assigned any subcarriers$\}$. $\bar{k} = \arg\min_{k\in\mathbf{B}}\psi_{k,t}$.

Step 3: Allocate subcarrier $\arg\max_{n\in\mathbf{C}} r_{\bar{k},n}$ to user $\bar{k}$. Denote the subcarriers allocated to user k as $\mathbf{Ch}_k$.

Step 4: Repeat step 3 until user $\bar{k}$ has been allocated $\lfloor K/N\rfloor$ subcarriers or $\mathbf{C}$ is empty.

Step 5: Repeat step 2-4, until $\mathbf{B}$ is empty.

Average rate allocation:

Step 1: Run average resource allocation algorithm, and get the subcarrier allocation result $\mathbf{Ch}_k$.

Step 2: Set ΔP to be a small step size, $R_{k,t} = 0$, $p_{k,n,t} = 0\ \forall k,n$, and $\bar{k} = \arg\min_k R_{k,t}$.

Step 3: Allocate power ΔP to user $\bar{k}$ on subcarrier $\arg\max_{n\in\mathbf{Ch}_{\bar{k}}}\Delta r_{\bar{k},n}$ where $\Delta\bar{r}_{k,n} = [r_{\bar{k},n}(p_{\bar{k},n}+\Delta P) - r_{\bar{k},n}(p_{\bar{k},n})]$.

Step 4: Update $P = P - \Delta P$ and $p_{\bar{k},n} = p_{\bar{k},n} + \Delta P$. Repeat steps 2 and 3 until $P = 0$.

PLACE
PHOTO
HERE

Chao Yang received the B.S and M.S. degrees in Electrical Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2006 and 2009 respectively, and received the PhD. Degree in Networked Systems from University of California, Irvine in 2014. His research interests include resource allocation, admission control and network optimization for both computer networks and cellular networks.

PLACE
PHOTO
HERE

Scott Jordan (S'86-M'90) received the A.B. in Applied Mathematics and the B.S., M.S. and Ph.D degrees in Electrical Engineering and Computer Science from the University of California, Berkeley, in 1985, 1987, and 1990, respectively. From 1990 until 1999, he served as a faculty member at Northwestern University. Since 1999, he has served as a faculty member at the University of California, Irvine. During 2006, he served as an IEEE Congressional Fellow, working in the United States Senate on Internet and telecommunications policy issues. His research interests currently include net neutrality, pricing and differentiated services in