Quasi-Optimal Channel Assignment for Real-Time Video in OFDM Wireless Systems

Jun Xu, Xuemin (Sherman) Shen, *Senior Member, IEEE,* Jon W. Mark, *Life Fellow, IEEE,* and Jun Cai, *Member, IEEE*

***Abstract*— In this paper, real-time video transmission with quality of service (QoS) assurance over orthogonal frequency division multiplexing (OFDM) wireless systems is studied. Three quasi-optimal subcarrier allocation schemes, namely *regular, delay tolerant* and *adaptive*, are proposed. In the proposed schemes, effective throughput per subcarrier employing adaptive forward error correction (FEC) coding in a Rayleigh fading channel is evaluated, and a capacity matrix that governs all active users and all assignable subcarriers is formulated. The Munkres algorithm is used in the schemes to achieve optimal subcarrier assignment. Performance analyses in terms of spectral efficiency and packet loss probability are presented. Multi-user multi-channel diversity is investigated to exploit the innate capacity gain in terms of fading variation and delay tolerance. The proposed schemes are scalable to both single-cell and multi-cell circumstances. Numerical results demonstrate that the proposed schemes can effectively achieve quasi-optimal spectral utilization and significantly improve system throughput.**

***Index Terms*— OFDM, optimal subcarrier allocation, QoS provisioning, Rayleigh fading, real-time video.**

I. INTRODUCTION

WITH the increasing real-time Internet video applications and the fast evolution of wireless networks, wireless Internet video services are expected to be extensively deployed. Video transmission consumes large bandwidth and requires stringent quality of service (QoS) requirements in terms of delay, bit error rate (BER) and video quality. However, the inherent scarce resource and hostile environment in a wireless network may limit the throughput of video applications. The newly proposed WiMax/IEEE 802.16 standard adopts orthogonal frequency division multiplexing (OFDM) as its prominent air interface to provide adequate transmission rates for multimedia applications [1]. OFDM is also considered as a major candidate for the 3G long term evolution (3G-LTE) to enable the cellular network to support 10 times higher rate than currently offered by HSDPA/HSUPA. In OFDM, the whole bandwidth is split into a number of orthogonal subcarriers (also referred to as *channels* hereafter), each of which carries information over the wireless channel. The orthogonality between the subcarriers circumvents interchannel interference (ICI). In addition, due to frequency selective fading, it is likely that a user can be selectively assigned a set of subcarriers with good channel condition all the time. Therefore, OFDM provides great flexibility for subcarrier assignment to support variable rate traffic and, at the same time, achieve high spectral efficiency.

Extensive research on optimal channel assignment to achieve maximum frequency utilization from both theory and implementation aspects has been reported in the literature, e.g., [2]–[9]. In [2], a theoretical framework is presented to achieve cross-layer optimization for average utility of all active users by channel assignment and power allocation. An approach using Nash Bargaining solution is proposed in [3] for fair multi-user channel allocation. Optimal resource management schemes for subcarrier and power allocation with fairness are proposed in [4],[5], and adaptive subcarrier and bit allocation in a multi-cell environment is investigated in [6]. Since optimal subcarrier assignment involves binary integer programming model, the non-polynomial complexity prohibits real-time application, and thus suboptimal solutions have been proposed in [4], [6]. Subcarrier allocation and bit loading algorithm is discussed in [7]. Solution approaches for optimal and suboptimal problems are also explored in [8], [9]. While great achievement has been made for optimal channel assignment, there are limitations in the existing approaches. First, due to the intrinsic complexity of integer programming, heuristic approaches are proposed instead. However, heuristic methods can not assure optimality and are only applicable in certain conditions. Second, although solutions for optimal assignment based on Hungarian Algorithm are discussed in [8], [9], they are short of complete exploration for the transmission and QoS evaluation of real-time video traffic. Furthermore, since video applications may represent a major part of applications in OFDM systems, it is important to investigate channel assignment, spectral efficiency and QoS provisioning for real-time video traffic.

QoS provisioning can be static or statistical. Providing statistical QoS for real-time video streams is more important in wireless networks since otherwise the resource will be underutilized [10]. A wireless effective capacity (EC) scheme has been proposed in [15] to provide QoS for variable rate videos, while multi-layered scalable video encoding has been

Manuscript received December 5, 2006; revised April 3, 2007; accepted May 11, 2007. The associate editor coordinating the review of this paper and approving it for publication was Y.-B. Lin. This work was supported by research grants from Bell University Laboratories (BUL) under the sponsorship of Bell Canada, the Natural Science and Engineering Research Council (NSERC) of Canada under the Strategic Project program, and the NSERC of Industrial Research Chair (IRC) program.

J. Xu, X. Shen, and J. W. Mark are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1 (email: {j5xu, xshen, jwmark}@bbcr.uwaterloo.ca).

J. Cai is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 3N8 (email: jcai@ee.umanitoba.ca).

Digital Object Identifier 10.1109/TWC.2008.060991.

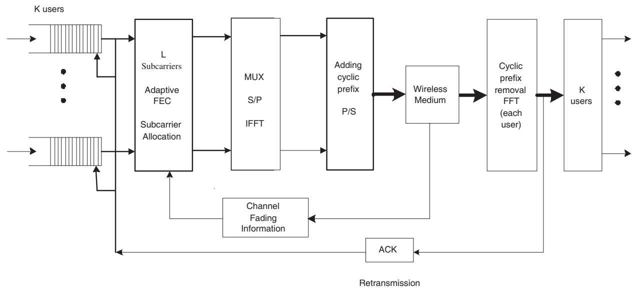

Fig. 1. Real-time video transmission over OFDM wireless channels.

used at the application layer to fast adapt a video source to the time-variant system capacity [10], [11]. Although optimal cross-layer design has been extensively investigated for packet video streams in 3G systems [13], [14], it is seldom addressed in the OFDM domain, especially for transmitting video with satisfactory QoS and efficient spectral utilization.

In this paper, the transmission of the real-time video with QoS provisioning over OFDM wireless channels is studied. Three quasi-optimal subcarrier allocation schemes, namely *regular, delay tolerant* and *adaptive*, are proposed, in which the Munkres algorithm [20] is used to assure the optimality of subcarrier assignment. Multi-user multi-channel diversity is investigated to exploit the innate capacity gain from fading variation and delay tolerance. Spectral efficiency pertaining to optimal channel assignment is analyzed. A closed form packet loss rate due to delay and channel error is derived to evaluate the QoS in video streaming. Scalability for single-cell and multi-cell situations is also investigated. The major contribution of this work is the proposal of quasi-optimal approaches for channel assignment to efficiently transmit video streams in OFDM systems by incorporating multi-user multi-channel diversity in a multipath delay fading environment.

The remainder of the paper is organized as follows. Section II presents the related models which capture the features of the OFDM system, adaptive channel coding, and the propagation channel. In Section III, optimal subcarrier allocations are formulated for both variable rate and multi-layered video streams. Three quasi-optimal channel assignment schemes are proposed in Section IV. Performance analysis for channel efficiency[1] and QoS evaluation are presented in Section V. Numerical results are given in Section VI, followed by conclusions in Section VII.

II. System Model

In orthogonal frequency division multiple access (OFDMA) mode (e.g., WiMAX), subcarriers can be divided into subsets. Each subset may be allocated to different groups of users and contains properly spaced subcarriers such that the frequency separation between neighboring subcarriers is larger and the bandwidth of each subband is smaller than the coherence bandwidth. Consider a subset with L physically separated but logically adjacent subcarriers supporting K users (connections) ($L \geq K$) in the downlink of a wireless OFDM system as shown in Fig. 1. The arriving data streams are stored in transmit buffers at the base station before transmission or retransmission. The subcarriers are to be allocated to some or all active users in each radio frame (time slot). The OFDM is performed on a *block-by-block* basis. At the base station transmitter, the serial data sequence of user i, $i = 1, ..., K$, is first serial-to-parallel (S/P) converted into L_i ($L_i \leq L$) low rate parallel streams for L_i subcarriers. If all subcarriers have been assigned, the low rate streams from all or some of the K users are multiplexed over the L subcarriers. The orthogonal waveform modulation is carried out using an inverse fast Fourier transform (IFFT) and a parallel-to-serial converter. Following the convertor, a cyclic prefix, which equals the channel delay spread, is added to each symbol to eliminate intersymbol interference (ISI). The resultant sample sequences are then transmitted over the wireless medium. Each transmitted block is referred to as an *OFDM symbol*. The receiver reverses the process using an FFT operation after removing the cyclic prefix. The outputs of FFT are the symbols modulated onto the L subcarriers, each multiplied by a complex channel gain. These symbols are decoded to recover the information bits. The transmitted signal may incur errors in traversing the wireless fading channel. The fading conditions among subcarriers are approximately independent for each user due to frequency separations. Depending on the availability of channel state information, adaptive coding/modulation schemes can be used. Subcarrier allocation at the link layer schedules the traffic to meet the throughput and QoS requirements. Moreover, the scheduler can match a user with the best fit subcarrier(s) to achieve higher channel throughput.

The wireless channel is subject to Rayleigh fading which can be represented by a finite state Markov channel (FSMC) model [18]. In FSMC, the received signal-to-noise ratio (SNR) is partitioned into R states. Transitions can exist between adjacent states or spread over the whole state space depending on the partition scheme and the mobile speed. The steady-state probability of the ith state is $\Omega_i = \int_{\gamma_i}^{\gamma_{i+1}} \rho(\gamma) d\gamma$, $i = 1, ..., R$, where γ_i is the SNR threshold for the ith state and $\rho(\gamma)$ denotes the probability density function of SNR.

Adaptive FEC is applied when a subcarrier is assigned. The number of error correction bits is adaptively chosen to match the channel BER so that the throughput is maximized [19]. In this paper, without loss of generality, an (n, k) BCH (Bose, Ray-Chaudhuri, and Hocquenghem) linear block code with sufficient interleaving is considered, where n is the block length, k is the number of information bits, and k/n is the *code rate*. When BCH is strictly used for error correction on a channel with BER p, the probability of decoding error is upper bounded by [23]

$$P_E \leq \sum_{i=t+1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{1}$$

where t is the random-error-correction capability of the code. Let each link layer packet consist of n_1 code blocks. The link layer packet error rate (PER) becomes

$$PER_{link} = 1-(1-P_E)^{n_1} \leq 1-(\sum_{i=0}^{t} \binom{n}{i} p^i (1-p)^{n-i})^{n_1} \tag{2}$$

[1]The terms channel efficiency, frequency efficiency, and spectral efficiency are used interchangeably throughout the paper.

and the information throughput can be calculated as

$$U_{info} = \frac{k}{n}(1-P_E)^{n_1}. \quad (3)$$

Intuitively as the channel BER increases, the throughput decreases for fixed (n, k). However, if k can be switched at certain BER values, such as the crossover points, namely *cross-BERs*, or the corner-down points, namely *corner-BER*, as denoted in [16], the maximum throughput can be obtained subject to the error rate.

If quadrature amplitude modulation (QAM) is adopted in OFDM, subject to mapping, the approximate BER with constellation size of $M = 2^m$ is [24]

$$BER = P_e/m \quad (4)$$

where the symbol error rate P_e is upper bounded by

$$P_e \leq 2Q(\sqrt{2E_b/N_0}), m = 2, \quad (5)$$

$$P_e \leq 2Q(\sqrt{2m(E_b/N_0)} \cdot \sin(\pi/M)), m > 2 \quad (6)$$

where E_b/N_0 is the SNR per bit, and $Q(.)$ is the complementary function of Normal distribution.

In this paper, two traffic models are considered: (1) multi-layered video model; and (2) variable rate video model. The former is developed to facilitate adaption to network bandwidth fluctuation by increasing or reducing the number of layers of a video stream. The latter is caused by compression. Both models will be discussed in the corresponding parts of Sections III and V.

III. Optimal Subcarrier Allocation

In this section, optimal subcarrier allocation models are formulated for real-time video transmissions with efficient channel utilization and QoS assurance. Let x_{ij} be a binary decision variable representing the assignment of channel j to user i, $i = 1, ..., K$; $j = 1, ..., L$. $x_{i,j} = 1$ $(= 0)$ means channel j is (not) assigned to user i. Let $c_{ij}(t)$ represent the instantaneous subcarrier capacity, which is the information rate that a subcarrier can carry, when $x_{ij} = 1$ at time slot t with adaptive FEC. $c_{ij}(t)$ is affected by slow Rayleigh fading, i.e., the fading remains the same over a block or during a time-slot interval, and can carry the amount of information given by $\frac{k}{n} \cdot C_0$, where C_0 is the maximum information rate that a subcarrier can provide. At each time slot t, an instantaneous $K \times L$ subcarrier capacity matrix of $C(t) = \{c_{ij}(t)\}$ as a snap shot can be obtained via channel estimation. In what follows, we propose optimal assignment models for two types of real-time video streams: multi-layered video and variable rate video.

A. Channel Allocation for Multi-layered Video

With MPEG-4 codec, a video source can be encoded into a base layer (BL) and an enhanced layer (EL). BL contains the most important information, and EL provides additional information for better video quality [11]. The EL can be further split into several sub-layers. This structure is conducive to follow the variable capacity due to network fluctuation such as channel fading, mobility, etc. In general, the BL has a more stringent QoS requirement than the EL since an error in BL can be propagated to EL even if the EL is correctly received. Therefore, a higher priority and more stringent QoS requirements are set for the BL. Resource allocation is to first satisfy the BL requirements, and then to seek for the maximum throughput of EL sub-layers. Optimal channel allocation can be obtained by solving the following minimization problem:

$$\text{Min} \quad B = \sum_{i=1}^{K}\sum_{j=1}^{L} x_{ij} \quad (7)$$

$s.t.$

$$\sum_{j=1}^{L} x_{ij}c_{ij}(t) \geq \lambda_i(t), i = 1, ..., K \quad (8)$$

$$\sum_{i=1}^{K}\sum_{j=1}^{L} x_{ij} \leq L \quad (9)$$

where B is the bandwidth consumption of BL, $\lambda_i(t)$ is the instantaneous BL rate of user i, and $\sum_{i=1}^{K} x_{ij} \leq 1, j = 1, ..., L$. Given sufficient number of channels, constraint (9) can be nonbinding and $B < L$. Let $E = (L - B)^+$, where $(x)^+ = \max(0, x)$, which is the number of available channels for ELs. The optimization problem for EL can be obtained as follows:

$$\text{Max} \quad \sum_{i=1}^{K}\sum_{j=1}^{E} x_{ij}c_{ij}(t) \quad (10)$$

$s.t.$

$$P_{Li}(t_d > T_d) \leq \varepsilon, i = 1, ..., K \quad (11)$$

$$\sum_{i=1}^{K}\sum_{j=1}^{E} x_{ij} \leq E \quad (12)$$

where $\sum_{i=1}^{K} x_{ij} \leq 1, j = 1, ..., E$. In (11), $P_{Li}(.)$ is the packet loss probability of user i, t_d is the packet delay which is the time duration from the moment when a packet arrives at the transmit buffer to the moment when the packet is successfully received at the receiver, T_d is the delay tolerance, and ε is a predefined threshold. Constraint (11) indicates that the delay of an EL packet can not exceed the delay bound T_d with an outage ratio of ε. Comparing (8) and (11), BL requires adequate resource to avoid packet loss due to delay and error subject to FEC, while EL is delay tolerable. Delay tolerance provides the possibility of further improvement of the multi-user diversity gain which will be discussed in Section V. To solve the optimization problem, (11) has to be relaxed, the procedure of which will be discussed in Section IV-B. The above models demonstrate a two-step optimal resource allocation for the layered videos. If an EL has more sub-layers with different QoS, these models can be extended accordingly.

B. Channel Allocation for Variable Rate Video

The characteristic of a variable rate video stream, which can be modelled as a Markov modulated rate process (MMRP), has been well studied in the literature, e.g., [22]. The variable traffic rate can be quantized into several levels, and each level can be represented by a state of the MMRP. Transitions between the states are governed by the underlying continuous-time Markov chain. The MMRP model is suitable for analysis, which will be shown in Section V. The optimization problem can be formulated as seeking maximum throughput or achieving maximum frequency efficiency. We consider the second one since the maximum throughput problem is similar to (10). Let $\lambda = [\lambda_1, \lambda_2, ..., \lambda_N]$ represent the state dependent video rate, and $\pi = [\pi_1, \pi_2, \cdots, \pi_N]$ represent the stationary probability, where N is the number of states of the MMRP. Denote the nominal capacity as the product of the number of assigned subcarriers and the maximum capacity per subcarrier. The channel efficiency η is defined as

$$\eta = \frac{\lambda \cdot \pi^T (1 - P_L(t_d > T_d))}{\text{number of subcarriers} \cdot C_0}$$

where $P_L(t_d > T_d)$ is the mean per-flow packet loss probability due to delay and channel error. The optimal channel efficiency problem can be formulated as follows:

$$\text{Max} \quad \frac{\lambda \cdot \pi^T (1 - P_L(T_d))}{U_E \cdot C_0} \tag{13}$$

$s.t.$

$$P_{Li}(t_d > T_d) \le \varepsilon, i = 1, ..., K \tag{14}$$

$$U(t) = \sum_{i=1}^{K} \sum_{j=1}^{L} x_{ij}(t) \le L \tag{15}$$

where $U(t)$ is the instantaneous number of assigned channels for each snap shot, U_E is the average number of assigned channels per unit time for each user, i.e., $U_E = \sum_{t=1}^{T} U(t)/(K \cdot T)$, T is the duration of a flow, and $\sum_{i=1}^{K} x_{ij} \le 1, j = 1, ..., L$.

The proposed optimization models can not be directly solved because firstly, solving integer programming has the exponential order of complexity [6], and secondly, stochastic constraints have to be relaxed. In the following section, we present viable solutions for the proposed models.

IV. Quasi-Optimal Solutions

In this section, we briefly explain the procedure of the Munkres Assignment Algorithm [20], and then propose three quasi-optimal[2] subcarrier assignment schemes by using the Munkres algorithm.

A. Munkres or Hungarian Algorithm

Consider a pair-match problem with a cost matrix. If a constant is added to or subtracted from all entries in a row or column of the assignment cost matrix, the optimal assignment is unchanged. Based on this result, continually subtracting constants from rows and columns in the cost matrix until a zero cost assignment is made will result in an optimal solution to the original problem. The solution procedure for a square cost matrix is as follows.

Assignment Procedure

1) Locate the smallest number in each row and subtract it from all the entries in this row.
2) Locate the smallest number in each column and subtract it from all the entries in this column.

[2]Since the formulated optimization problems are not directly solved, the proposed alternative solution is referred to as *quasi-optimal*.

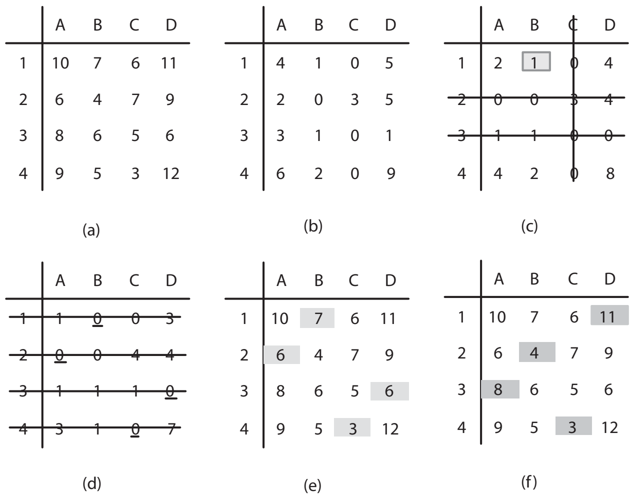

Fig. 2. Optimal assignment with Munkres Algorithm.

3) Cross out all the zeros using the smallest number of lines. If the number of lines equals the number of rows (or columns), optimal (zero) assignment has been found, stop. Otherwise, go to (4).
4) Choose the smallest uncovered number and, (i) subtract it from all other uncovered numbers, (ii) add it to the numbers where the lines are crossed out, (iii) go to (3).

Fig. 2 shows the assignment procedure by the Munkres algorithm. Given the original cost matrix in Fig. 2(a), Fig. 2(b) is obtained by subtracting the smallest number of each row. A minimum of three lines are used to cross out all zeros as shown in Fig. 2(c), and the smallest uncovered number is the shaded '1'. By doing step 4, Fig. 2(d) is obtained with four zero assignments marked by underscores. The optimal cost is 6+7+3+6=22, as shown in Fig. 2(e), compared with heuristic smallest assignment, 3+4+8+11=26, illustrated in Fig. 2(f). This algorithm can be extended to rectangular matrix by adding dummy rows or columns with zero costs [21]. The complexity is of low polynomial order[3]. For the maximization problem, a similar procedure can be carried out by using the largest element of the matrix to subtract all the entries.

B. Subcarrier Assignment

Because the QoS requirements and traffic characteristics may vary, three channel assignment schemes, namely *regular, delay tolerant*, and *adaptive*, are proposed by properly using the Munkres algorithm to feasibly work out the optimal solutions for the optimization problems defined in Section III.

1) Regular Scheme: Let L subcarriers be assigned to K active users (connections) at the downlink. An active user is defined as one with nonempty buffer at the moment of channel assignment, with each having a queue length of q_i, $i = 1, ..., K$. The original assignment matrix is K by L with the element c_{ij}, which is the instantaneous subcarrier capacity. The channel assignment at the base station is done by iteratively assigning available channels to active users. In

[3]It is believed that the complexity can be of $O(n^3)$ subject to implementation.

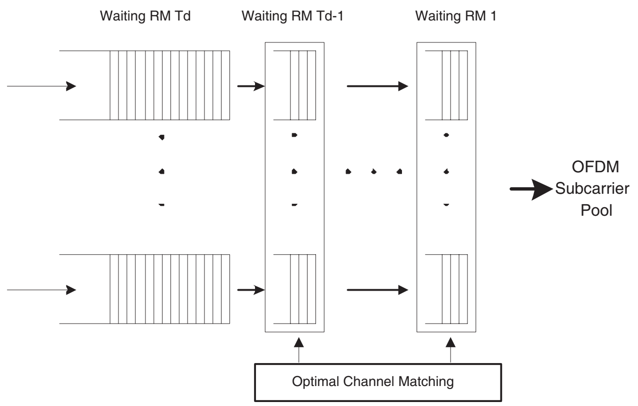

Fig. 3. Prioritized assignment based on packet delay.

each round, one channel is assigned to one user based on the channel matching result by the Munkres algorithm. After each round, when a channel is assigned or a user's queue becomes empty, the cost matrix is updated by removing the assigned channels or the *non-active* users (the users with empty buffer). This procedure continues until all the channels are used up or all the queues are empty. This method is referred to as *Regular scheme*.

2) Delay Tolerant Scheme: A small delay tolerance is allowed for video packets. This tolerance is important to obtain multiplexing gain for aggregate traffic in the network. If a packet is imminently exceeding its delay tolerance (it is sometimes referred to as most urgent packet), it has a priority to obtain the resource. In this way, priority groups are formed considering the delay tolerance of packets. Let T_d be the delay bound and t_e be the duration from the current time to the due time when the delay tolerance expires. t_e is also referred to as *delay remain*. The most urgent packet has $t_e = 1$ while a new arrival has $t_e = T_d$. We group the queued packets according to their delay remains. If $t_e = 1$, the packet is grouped in Group 1. If $t_e = T_d$, the packet is grouped in Group T_d. In each group, packets are indexed with each user. This can be imagined as a scenario of a multi-queue system with several waiting rooms, as shown in Fig. 3. Packets in the same waiting room, i.e., the same group, has the same delay remain. Transmission priorities are given by the ascending order of groups. Thus, packet scheduling is done by adaptively assigning channels to active users according to the delay remains of their packets. Scheduling in each group is done in the same way as the regular scheme. When the capacity of an assigned channel is larger than the number of packets in one room, the difference will be filled by the packets in the subsequent room. The grouping is flexible in the sense that some waiting rooms can be merged. If only the most urgent packet is of concern, all rooms except room 1 can be merged for implementation convenience.

3) Adaptive Scheme: Due to the dynamics of the video traffic and fluctuation of the wireless channel, it is possible that there is no sufficient network capacity to accommodate all the traffic immediately. When the instantaneous arriving traffic is larger than the available capacity, the system is temporally *overloaded*. The peak traffic can be queued and served when the instantaneous traffic are smaller than the available capacity, i.e., in the *underload* state. As long as the average traffic load does not exceed the average system capacity, given a delay tolerance, one can hold the overload traffic to the underload state to achieve better channel utilization.

Consider an N-state MMRP model for a variable rate video source with state dependent arrival rate $\lambda_i,\ i = 1, ..., N$. Suppose adaptive FEC is designed in a way that the packet error rate is random and very small. Let v_i be the mean retransmission rate at state i. We design N_c controllable service rates, each of which may support one or more states of arriving traffic. In each state of the MMRP, the arrival rate plus the retransmission rate may be larger or smaller than the corresponding service rate. If $N_c \leq N$, the MMRP can be partitioned into N_c sections, which is referred to as a partitioned Markov process [17].

For the kth section with service rate μ_k, $k = 1, ..., N_c$, the service rate falls between two consecutive boundaries, states L_{k-1} and L_k, i.e., $\lambda_{L_{k-1}} + v_{L_{k-1}} < \mu_k < \lambda_{L_k} + v_{L_k}$. The buffer occupancy tends to increase when the states in the kth section have arrival rates larger than μ_k, and these states are called *overload* states; otherwise, the buffer occupancy tends to decrease when the arrival rates of the states are less than μ_k, and these states are called *underload* states [22]. Hence, the kth section of the Markov chain is further partitioned into two sets, $\mathcal{Z}_F^k = \{i : L_{k-1} < i \leq L_k | \lambda_i + v_i > \mu_k\}$ for overload states and $\mathcal{Z}_E^k = \{i : L_{k-1} < i \leq L_k | \lambda_i + v_i < \mu_k\}$ for underload states, where $k = 1, 2, ..., N_c$. As a result, the Markov chain with N_c sections is constructed by interleaving *overload* and *underload* states. Given a properly designed partitioning scheme, the packet loss rate can be evaluated and satisfied with improved channel utilization, which will be shown in Section V and Section VI, respectively.

Channel Assignment Algorithm

1) Form the candidate user set and available subcarrier set.
2) Assign a required controllable service rate for a candidate user based on the arriving rate.
3) Assign one subcarrier to each candidate by the Munkres algorithm. If a candidate's assigned cumulative subcarrier capacity meets the required service rate, remove this user from the candidate set. Remove all assigned subcarriers from the available subcarrier set.
4) If all required service rates are met or the available subcarriers have run out, stop; otherwise, update the candidate set and available subcarrier set, and go to (3).

Next we compare the three schemes with respect to the optimal models presented in Section III. The regular scheme can satisfy constraint (8) subject to SNR satisfaction and total available channels. It can also satisfy constraints (11) and (14), subject to traffic pattern, SNR and total available channels, but may not provide the best frequency efficiency. It is a viable procedure for optimality because the multi-user gain in terms of channel fading can be fully obtained. The delay tolerant scheme can relax constraints (11) and (14), and make complete exploration of multiplexing gain with regard to the channel fading and delay tolerance, thus achieve much better channel efficiency than the regular scheme. However, with heterogeneous multimedia transmission environment, at the time when channel resource is abundant for video traffic, this scheme may lead to over-provisioning of QoS, which could potentially affect other types of services. The adaptive scheme can relax constraint (14), and constraint (11) if the EL of a layered source is of variable rate. Capacity gain can be fully obtained from multi-user diversity and delay tolerance. For instance, if a user experiences the worst fading for all channels, the information can be held during the delay tolerance until a good channel appears. There is no buffer counting with this scheme (this could save one frame of delay), so the arriving traffic is virtually circuit-switched to the channels. In addition, the QoS requirement is adequately provided to avoid the overuse of the resource. In Section V, an analytical approach is derived to evaluate the QoS performance, given an adaptive channel allocation scheme.

C. Multi-Cell Diversity

OFDM can be used in grid networks where the coverage of cells overlap. A user in the overlapped area is able to access more than one base station. Due to the uncertainty of user mobility and location distribution, it is possible that the loads in the cells are unbalanced. To achieve better spectral efficiency, it is desirable to balance the load even though the user distribution is nonuniform. This can be done by merging the channels of the home cell and one or more of its neighbor base stations into a large channel pool, and merging the candidate user sets of all corresponding cells into a large candidate set. The users of these merged cells can share all the resource of the channel pool (assuming there is no conflict in using the frequencies). All the proposed channel assignment schemes can be directly used in the way of expanding the cost (capacity) matrix by adding the channel estimation information of neighboring cells to the local users, as shown in Section VI. The performance of channel efficiency is improved at the cost of more channel estimations.

V. Performance Analysis

In this section, we first analyze the performance of spectral efficiency in terms of the multi-user multi-channel diversity in OFDM, and then present an analytical approach to evaluate the packet loss probability due to delay and channel error.

A. Spectral Efficiency

Assume user i, $i = 1, ..., K$, can be assigned one or more of the L subcarriers. From user i's point of view, if each subcarrier is an independent FSMC process with identical transition matrix A, he is virtually associated with a situation of L channels characterized by a joint Markov process with transition matrix $A^{(L)} = \underbrace{A \otimes ... \otimes A}_{L}$, where $\otimes$ denotes Kronecker product. Denote $\Omega^{(L)}$ the stationary probability of $A^{(L)}$. The probability that the capacity of assigned channels exceeds the payload of user i, $\Pr(\sum_{j=1}^{L} c_{ij}x_{ij} \geq C_i)$, is the summation of all the states in $\Omega^{(L)}$ that satisfies $\sum_{j=1}^{L} c_{ij}x_{ij} \geq C_i$, where c_{ij} is the capacity of subcarrier j for user i, C_i is the payload, x_{ij} is a binary variable, and $\sum_{j=1}^{L} x_{ij} \leq H_i$, where H_i is a predefined limit on assignable channels of user i. This probability will increase with L since the size of the space is larger. Given A has R states, the total number of states of $A^{(L)}$ will be R^L and the size of $A^{(L)}$ is $R^L \times R^L$. This is the same for each of the K users. Thus, a multi-user multi-channel diversity gain can be obtained in a pool of K users and L subcarriers, namely a (K, L) system.

Let $c_{ij} \in \{c_1, ..., c_R\}$ be a random variable representing the capacity of subcarrier j for user i, where $c_i, i = 1, ..., R$, is the capacity of state i in an FSMC and $\{c_i\}$ is in an ascending order. The instantaneous throughput of the pooled system, $\sum_i^K \sum_j^L c_{ij}x_{ij}$, subject to $\sum_j^K x_{ij} \leq 1$, will be ranged from $c_1 L$ to $c_R L$ in discrete values. Let $C^{(K,L)}$ be the set of summations, $c_i^{(K,L)}$, $i = 1, ..., N^{(K,L)}$, be a value in $C^{(K,L)}$, and $N^{(K,L)}$ be the size of $C^{(K,L)}$. The optimal spectral efficiency for the (K, L) system is

$$\eta_{opt} = \frac{\sum_i^{N^{(K,L)}} c_i^{(K,L)} \Pr[\mathrm{Max}(\sum_i^K \sum_j^L c_{ij}x_{ij}) = c_i^{(K,L)}]}{c_R L} \tag{16}$$

where $\Pr[\mathrm{Max}(\sum_i^K \sum_j^L c_{ij}x_{ij}) = c_i^{(K,L)}]$ is the marginal probability when the instantaneous throughput has a maximum value of $c_i^{(K,L)}, i \in N^{(K,L)}$, and can be obtained from stationary probabilities of each user's joint FSMC $A^{(L)}$.

Consider a two-user, two-channel system with each channel having two states, denoted as (*high, low*) and associated with capacity *(h, l)*, respectively. A joint FSMC with matrix $A^{(2)}$ for each user has four states: *(hh, hl, lh, ll)*. Let $X_i, i = 1, 2$, denote the instantaneous state of $A^{(2)}$ for user i. The summation of capacities, $C_{X_1+X_2}$, has three possible values: *(2h, h+l, 2l)*. The marginal probabilities of $\mathrm{Max}(C_{X_1+X_2})$ can be found as follows:

$$\begin{aligned} &P[Max(C_{X_1+X_2}) = 2h] \\ &= P(X_1 = hh, X_2 = hh) + P(X_1 = hh, X_2 = lh) \\ &+ P(X_1 = hh, X_2 = hl) + P(X_1 = hl, X_2 = hh) \\ &+ P(X_1 = hl, X_2 = lh) + P(X_1 = lh, X_2 = hh) \\ &+ P(X_1 = lh, X_2 = hl), \end{aligned} \tag{17}$$

$$\begin{aligned} &P[Max(C_{X_1+X_2}) = h + l] \\ &= P(X_1 = hh, X_2 = ll) + P(X_1 = hl, X_2 = hl) \\ &+ P(X_1 = hl, X_2 = ll) + P(X_1 = lh, X_2 = lh) \\ &+ P(X_1 = lh, X_2 = ll) + P(X_1 = ll, X_2 = hh) \\ &+ P(X_1 = ll, X_2 = lh) + P(X_1 = ll, X_2 = hl), \end{aligned} \tag{18}$$

$$P[Max(C_{X_1+X_2}) = 2l] = P(X_1 = ll, X_2 = ll). \tag{19}$$

Since X_1 and X_2 are independent, the joint probabilities in (17), (18), and (19) can be obtained by the product of corresponding stationary probabilities, i.e., $P(X_1, X_2) = P(X_1)P(X_2)$.

In general, for a K-user, L-channel system with a R-state FSMC model, the size of joint spaces will be R^{KL}. Thus, for small values of K and L, (16) is applicable. However, when K or L becomes large, there will be a space explosion which makes the performance analysis impractical. To the best of our knowledge, there is no analytical work available in the literature for assessing the optimal spectral efficiency in an OFDM system with multi-user multi-channel diversity. Nonetheless, based on the approaches proposed in Section IV, the optimal spectral efficiency can be achieved subject to a certain subcarrier allocation, which will be shown in Section VI.

B. QoS Performance Analysis

We further evaluate the packet loss rate due to delay and channel error considering the MMRP model for a variable rate video stream based on the fluid model analysis [22]. The focus is to find a controllable optimal subcarrier allocation in the sense that QoS requirement is satisfied without overusing the resource. Following the discussion in Section III-B, we consider a variable rate video with transition probability matrix $\mathbf{P} = \{p_{ij}\}, i, j = 1, ..., N$, where p_{ij} is the one-step transition probability from state i to j, and N is the number of states. The infinitesimal generator matrix $\mathbf{M}$ can be obtained by $\mathbf{M} = \mathbf{I} - \mathbf{P}$, where $\mathbf{I}$ is an identity matrix. The stationary probability of the MMRP is $\boldsymbol{\pi}$.

At the transmitter, the arriving packets are buffered before they are transmitted or retransmitted. Let ϵ be the packet error rate due to the residual channel error after adaptive FEC coding. Taking into account retransmissions, the effective rate at state i is $\lambda_i + v_i$, where an upper bound of the mean retransmission rate for a given ϵ is

$$v_i \le \frac{(1-\epsilon)\epsilon}{1-2\epsilon}\left(\sum_{j\in\{j:\lambda_j<\mu_j\}} p_{ji}\lambda_j + \sum_{j\in\{j:\lambda_j>\mu_j\}} p_{ji}\mu_j\right),\ \ i, j = 1, ..., N. \tag{20}$$

Define $\mathbf{F}(x) = [F_1(x), F_2(x), \cdots, F_N(x)]$, where $F_i(x) = \Pr[X \le x, S = i]$, X and S are random variables denoting the buffer occupancy and the state of the underlying MMRP, respectively. We use the fluid-flow approach to analyze the queuing behavior of the transmit buffer by solving the following linear differential equations

$$\frac{d\mathbf{F}(x)}{dx} = \mathbf{F}(x)\mathbf{M}\mathbf{D}^{-1} \tag{21}$$

where $\mathbf{D} = diag.[u_i + v_i - \mu_i]$, $1 \le i \le N$.

The solution of (21) can be readily obtained as

$$F_i(x) = \pi_i + \sum_{j\in\{z_j<0\}} a_j\Phi_{ji}e^{z_jx} = \pi_i - \sum_{j\in\{z_j<0\}} w_{ji}e^{z_jx} \tag{22}$$

where $i = 1, \ldots, N$, $(z_j, \boldsymbol{\Phi}_j = [\Phi_{j1}, \Phi_{j2}, \cdots, \Phi_{jN}])$ is the (eigenvalue, eigenvector) pair satisfying the eigensystem $z_j\boldsymbol{\Phi}_j\mathbf{D}_N = \boldsymbol{\Phi}_j\mathbf{M}_N$, $1 \le j \le N$, a_j's are the coefficients, and $w_{ji} \ge 0$ is the weight to be determined.

Due to the stringent delay requirement of the real-time traffic, the packet will be dropped if its transmission delay exceeds a threshold, namely delay bound T_d (frames). Define a set of virtual buffer bounds $\{B_i(T_d), i = 1, 2, \ldots, N\}$ for each state of the MMRP associated with the delay bound T_d. Let $\mathcal{Z}_F = \{i \in N | \lambda_i + v_i > \mu_i\}$ and $\mathcal{Z}_E = \{i \in N | \lambda_i + v_i < \mu_i\}$ denote the set of overload and underload states, respectively. In the underload states, the buffer occupancy is low so that $\Pr[X > x, S = i | i \in \mathcal{Z}_E] = \pi_i - F_i(x | i \in \mathcal{Z}_E) = 0$, or $\sum_{j\in\{z_j<0\}} w_{ji}e^{z_jx} = 0$. As a result, for all the underload states, $w_{ji} = 0$. In the overload states, $\Pr[X > x, S = i | i \in \mathcal{Z}_F] = \pi_i - F_i(x | i \in \mathcal{Z}_F) = \sum_{j\in\{z_j<0\}} w_{ji}e^{z_jx}$, denoted by $G_i(x)$. The probability of buffer overflowing a virtual buffer bound in an overload state, $G_i(B_i(T_d))$, equals the weighted summation of exponentials with negative eigenvalues associated with the overload states, i.e.,

$$G_i(B_i(T_d)) = \sum_{j\in\{z_j<0\}} w_{ji}e^{z_jB_i(T_d)}. \tag{23}$$

Asymptotically, $G_i(x)$ can be governed by a few *significant* eigenvalues (the negative eigenvalues with small magnitude). The significant eigenvalues include the *dominant* (the largest negative) eigenvalue and several next largest negative eigenvalues, and are denoted by $\{z_j^*, j = 1, ..., J\}$, where J is the number of significant eigenvalues.

Assume all weights are equal, i.e., $w_{ji} = w_i$, $j = 1, ..., J$. w_i of an overload state i can be obtained by solving

$$G_i(0^+) = \sum_{j=1}^{J} w_i e^{z_j^*\cdot 0^+} = \frac{(\lambda_i + v_i - \mu_i)\pi_i^T}{\pi\lambda^T}, \tag{24}$$

i.e., $w_i = (\lambda_i + v_i - \mu_i)\cdot\pi_i^T/[J\cdot\pi\lambda^T]$, where $\lambda = \{\lambda_i, i = 1, 2, \ldots, N\}$.

Let $\mu_i(t = 0)$ be the service rate at time $t = 0$. Then $B_i(T_d) = \sum_{j=0}^{T_d-1}\mu_i(t = j)$ is a random variable depending on the service rates of the next T_d frames. The expected value of $B_i(T_d)$ can be obtained by

$$\begin{aligned}\overline{B}_i(T_d) &= E\left[\sum_{j=0}^{T_d-1}\mu_i(t = j)\right]\\ &\doteq \mu_i(0)\left(1 + \sum_{j=1}^{N}p_{ij}^{(1)} + \sum_{j=1}^{N}p_{ij}^{(2)} + \cdots + \sum_{j=1}^{N}p_{ij}^{(T_d-1)}\right)\end{aligned} \tag{25}$$

where $p_{ij}^{(m)}$ is the (i, j)th element of the transition probability matrix $\mathbf{P}^{(m)} = \mathbf{P}^m$ after m evolutions, $m = 1, ..., T_d - 1$. Given t_d denotes the time duration from the instant of a packet arriving at the transmit buffer until it is successfully received at the receiver, we have

$$\begin{aligned}\Pr[t_d > T_d, S = i] =& G_i[X > B_i(T_d)]\\ \doteq& \sum_{j=1}^{J} w_i e^{z_j^*\overline{B}_i(T_d)}\\ =& \sum_{j=1}^{J} w_i\cdot\exp[z_j\cdot\mu_i(0)(1 + \sum_{j=1}^{N}p_{ij}^{(1)}\\ &+ \sum_{j=1}^{N}p_{ij}^{(2)} + \cdots + \sum_{j=1}^{N}p_{ij}^{(T_d-1)})].\end{aligned} \tag{26}$$

The total packet dropping probability $P_d(T_d)$ due to delay exceeding the delay bound is given by

$$P_d(T_d) = \Pr[t_d > T_d] = \sum_{i\in\mathcal{Z}_F}\sum_{j=1}^{J} w_i e^{z_j^*\overline{B}_i(T_d)}. \tag{27}$$

If an erroneous packet is within its delay bound, it can be continuously retransmitted until it is correctly received

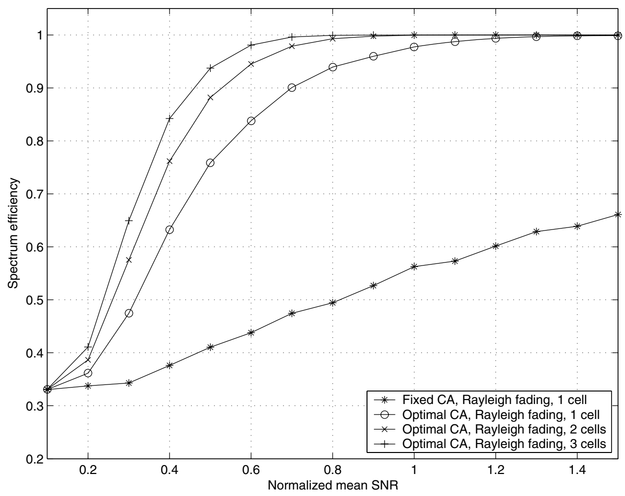

Fig. 4. Comparison of optimal assignment vs fixed assignment.

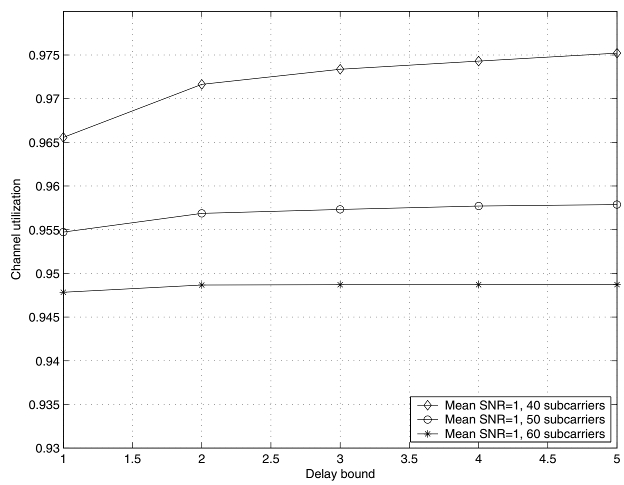

Fig. 5. Channel utilization with different number of assignable channels.

or exceeds the delay bound. Given that the delay bound is T_d frames, the round trip acknowledgement can be received within one frame, and the average channel error rate is ϵ, the total packet loss probability P_L due to delay and channel error can be obtained as

$$P_L(T_d) \doteq \sum_{i=0}^{T_d-1} \epsilon^i \cdot P_d(T_d - i). \tag{28}$$

Equation (28) is a conservative evaluation of the packet loss probability when ϵ is small.

VI. Numerical Results

In this section, numerical results are presented to evaluate the proposed channel allocation schemes for real-time video flows. Single-cell and multi-cell situations are considered. Suitability and scalability issues are also discussed.

A. Simulation Parameters

Consider an OFDM wireless system with Rayleigh fading channels. The radio frame is set to 20 ms. The carrier frequency is 1G Hz. Mobile speed is 10km/hour. The modulation scheme is 4QAM/QPSK. The fading condition is assumed to be fixed in each time frame. The available (assigned) bandwidth of the subcarrier, the source rates and the service rates are measured in link layer packets per frame, where each packet contains $n_1 = 20$ code words. The BCH block code has a block size $n = 256$ and information length $k \in \mathbf{k} = \{239\ 223\ 199\ 179\ 155\ 131\ 107\ 79\}$. The nominal capacity of a subcarrier $C_0 = 5$ corresponding to the highest code rate. The Markov fading process is partitioned by corner-SNRs (the corresponding SNRs to corner-BERs). The parameters of FSMC transition matrix can be evaluated by monitoring the SNR for a certain amount of time. The stationary probability, Ω=[.2922 .0384 .0617 .0495 .0656 .1006 .1117 .2803], is independent of mobile speed. The average capacity of a subcarrier is $(\frac{\mathbf{k}}{\max(\mathbf{k})} \cdot \Omega^T \cdot C_0)$.

For video traffic, the size of BL for the layered video is the same for all users, while EL is assumed to be fully coded with variable rate. The throughput of EL is of interest here. The variable rate video flow, which is generated using an autoregressive model, has a mean of 0.50 bit/pixel (equivalent to the mean value of 19.3 packets/frame by proper conversion), standard deviation of 0.23 bit/pixel, and time constant equal to 3.9. The simulated video sequence is then represented by the MMRP model. The observed rate span of the video sequence is $[0, 50+)$ packets/frame, which is evenly quantized into eight rate levels (states) in an ascending order. The discrete transition probability $\mathbf{P}$ and the stationary probability π are obtained by monitoring the simulated sequence of 1000 frames, and $\pi = [.0501\ .1107\ .2206\ .2725\ .2032\ .1034\ .0333\ .0062]$.

B. Channel Efficiency and QoS Performance

We define the spectral efficiency as the ratio of total effective throughput to the total nominal capacity (the nominal capacity is the maximum capacity of assigned channels, as defined in Section III), and the channel utilization as the ratio of total effective throughput to the total service rates (a service rate is less than or equal to the nominal capacity subject to FEC). Spectral efficiency, channel utilization, QoS performance for single-cell and multi-cell situations are determined for the proposed channel allocation schemes. The general situation of 40 subcarriers/10 users per cell is considered.

Fig. 4 shows the spectral efficiency for channel assignment (CA) using regular scheme compared with fixed channel assignment when fully coded video stream is transmitted. It is shown that the proposed assignment scheme greatly increase the frequency efficiency. It also shows that the performance becomes better as SNR increases. In addition, when the size of channel pool is larger by combining two or three cells (user base is also increased by the same scale), the channel efficiency can be further improved. Fig. 5 shows that in the single-cell case, the performance of channel utilization increases as the available number of channels increases. This is because when the number of channels increases, a user has a larger chance to find an instantaneous good channel. For multi-cell diversity, consider two cells with unbalanced traffic: one cell has 15 users/40 channels, and the other 5 users/40 channels. SNR of home cell is referred to as *local SNR*, and SNR from the base station of neighbor cell is

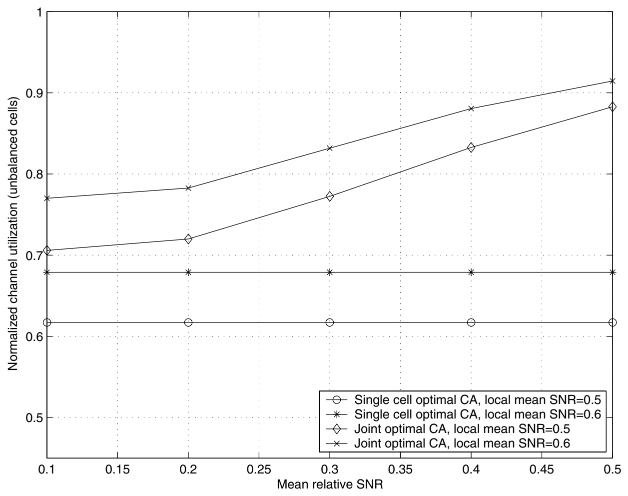

Fig. 6. Comparison of joint and single-cell channel assignments for unbalanced traffic.

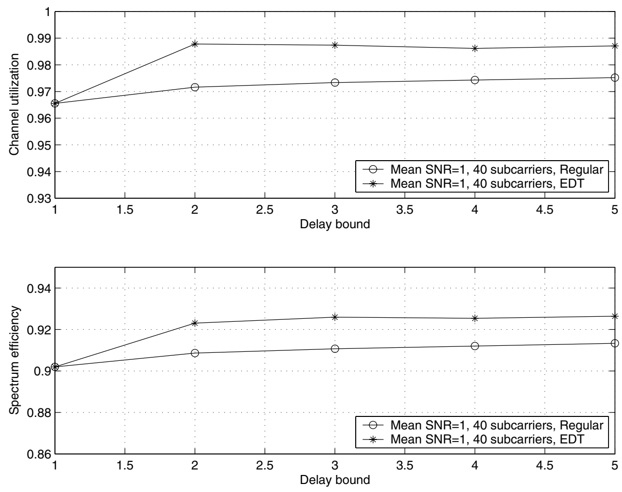

Fig. 8. Comparison of spectral efficiency/channel utilization for the regular and EDT schemes.

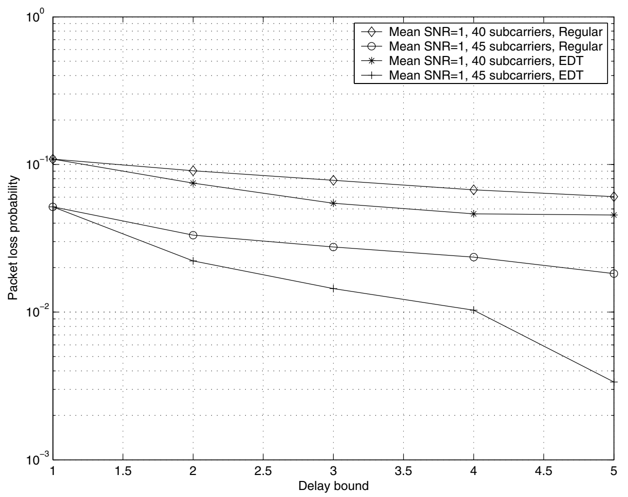

Fig. 7. Comparison of packet loss probability for the regular and EDT schemes.

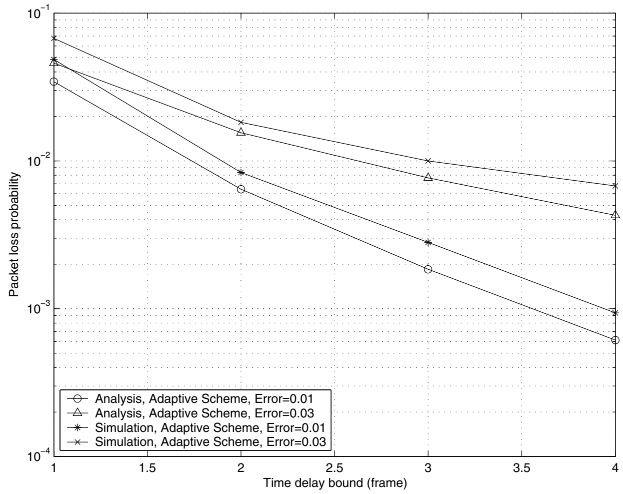

Fig. 9. Packet loss probability for the ADP scheme.

referred to as *relative SNR*. Assume local and relative SNRs for either cell are symmetrical with each other. Joint channel assignment is obtained by combining the channels and users of two cells. The channel utilization of single-cell is obtained by averaging the utilizations of each cell. From Fig. 6, it can be observed that channel utilization of joint assignment increases as the relative SNR increases. In addition, the improvement of channel utilization for multi-cell diversity is remarkable compared with that for single-cell.

We present the comparison of the regular scheme and the delay tolerant scheme for the variable video streams in Figs. 7 and 8. The delay tolerant scheme is formed by setting two waiting rooms: one for the most urgent packets (delay remain =1) and the other for all the others, namely earliest due time (EDT). It can be observed that the packet loss probability is smaller while the channel utilization/efficiency is larger for EDT compared to the regular scheme. This is because EDT achieves more interstate multiplexing gain in terms of delay tolerance.

For the adaptive scheme (ADP), the variable rate video streams are modelled as MMRP with state dependent rate $\lambda = [3\ 9\ 15\ 21\ 27\ 33\ 39\ 45]$. To adapt to the MMRP rate λ, the service rate is chosen as $\mu = [10\ 10\ 20\ 20\ 30\ 30\ 40\ 40]$ with four levels. Each level is a multiple of the subcarrier's nominal capacity. Underload and overload states are properly interleaved. Fig. 9 shows the packet loss probability with respect to delay tolerance. It can be seen that the analytical results agree well with the simulation in terms of different channel error rates. It is also observed in Fig. 10 that ADP scheme can achieve remarkable channel efficiency.

C. Discussion

The numerical results show that the proposed schemes achieve superior performance over fixed channel assignment in terms of spectral efficiency. The regular scheme is flexible but does not provide assured QoS. The delay tolerant scheme offers better performance than the regular one in terms of QoS and spectral efficiency. The adaptive scheme provides controllable rate adaptation to the video stream and can satisfy the QoS adequately without the overuse of resource. In addition, since there is no buffer counting for the ADP scheme, the

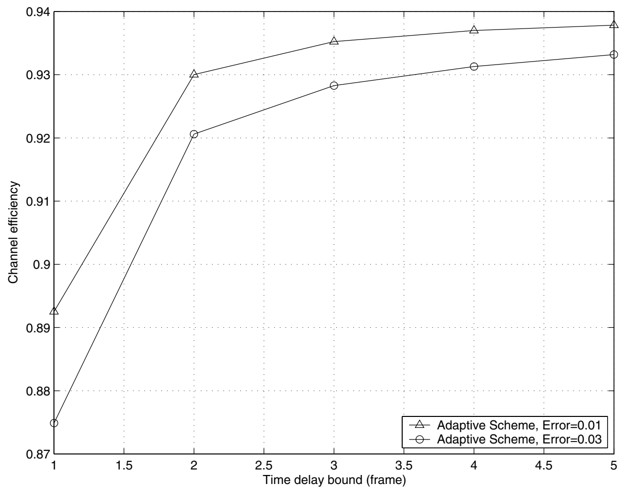

Fig. 10. Channel efficiency for the ADP scheme.

delay performance will be further improved. In general, these schemes are simple, elegant and easy to implement. A key characteristic is that the Munkres algorithm is subtly applied in the proposed schemes. This algorithm is scalable and has a guaranteed optimal solution, and can be incorporated with the proposed schemes to effectively achieve an viable alternative to the optimal spectral utilization subject to the resource and QoS constraints. The complexity of the proposed schemes is essentially similar considering that it mainly comes from the Munkres algorithm. The number of iterations for the regular and delay tolerant schemes is subject to the queue length of all users. Service rate based on the arrival pattern needs to be found for the adaptive scheme, which is, however, proportional to the queue length. There could be other algorithms to solve the optimal assignment problem, but the Munkres (Hungarian) algorithm is one of those that can solve the linear assignment problem within time bounded by the low polynomial order.

VII. Conclusions

In this paper, spectral efficiency for transmitting real-time videos with QoS provision over time-varying OFDM wireless channels has been explored. Viable quasi-optimal channel assignment schemes are proposed to satisfy QoS requirements and at the same time pursue maximum channel utilization. Multi-user multi-channel diversity is investigated to exploit the intrinsic multiplexing gain in terms of fading variation and delay tolerance. The proposed schemes are scalable for both single-cell and multi-cell situations. As the size of sub-carrier pool increases, more multiplexing gain can be obtained. Specifically, a controllable adaptive rate allocation is proposed to provide adequate QoS without overusing the resource for variable rate video streams. Analysis for spectral efficiency and QoS performance in terms of packet loss rate due to delay and channel error has been presented. Numerical results demonstrate that these proposed schemes can effectively maximize the spectral utilization with QoS satisfaction and significantly improve system throughput.

References

[1] A. Ghosh, D. R. Wolter, J. G. Andrews, and R. Chen, "Broadband wireless access with WiMax/802.16: Current performance benchmark and future potential," *IEEE Commun. Mag.*, pp. 129–136, Feb. 2006.

[2] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks–Part I: Theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 614–624, Mar. 2005.

[3] Z. Han, Z. Ji, and K. R. Liu, "Fair multiuser channel allocation for OFDMA networks using Nash bargaining solutions and coalitions," *IEEE Trans. Commun.*, vol. 53, pp. 1366–1376, Aug. 2005.

[4] J. Cai, X. Shen, and J. W. Mark, "Downlink resource management for packet transmission in OFDM wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1688–1703, July 2005.

[5] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in Multi-user OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2726–2737, Nov. 2005.

[6] Y. J. Zhang and K. B. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1566–1575, Sep. 2004.

[7] G. Kulkarni, S. Adlakha, and M. Srivastava, "Subcarrier allocation and bit loading algorithm for OFDM-based wireless networks," *IEEE Trans. Mobile Comput.*, vol. 4, pp. 652–662, Nov. 2005.

[8] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcast.*, vol. 49, pp. 362–370, Dec. 2003.

[9] Z. Zhang, Y. He, and K.P. Chong, "Oppotunistic downlink scheduling for multiuser OFDM systems," in *Proc. IEEE WCNC*, Mar. 2005, pp. 1206–1212.

[10] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "End-to-end QoS for video delivery over wireless Internet," *Proc. IEEE*, vol. 193, pp. 123–134, Jan. 2005.

[11] H. Radha, Y. Chen, K. Parthasarathy, and R. Cohen, "Scalable Internet video using MPEG-4," *ELSEVIER Signal Processing: Image Commun.*, vol. 15, pp. 95–126, 1999.

[12] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Channel-adaptive resource allocation for scalable video transmission over 3G wireless network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, pp. 1049–1063, Aug. 2004.

[13] T. Ahmed, A. Mehaoua, R. Boutaba, and Y. Iraqi, "Adaptive packet video streaming over IP networks: A cross-layer approach," *IEEE J. Select. Areas Commun.*, vol. 23, pp. 385–401, Feb. 2005.

[14] J. Liu, B. Li, Y. T. Hou, and I. Chlamtac, "On optimal layering and bandwidth allocation for multisession video broadcasting," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 656–667, Mar. 2004.

[15] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 630–643, July 2003.

[16] J. Xu, X. Shen, J. W. Mark, and J. Cai, "Adaptive rate allocation for multi-layered video transmission in wireless communication systems," in *Proc. IEEE ICC*, June 2006, pp. 5295–5300.

[17] J. Xu, X. Shen, J. W. Mark, and J. Cai, "Efficient channel utilization for real-time video over OVSF-CDMA systems with QoS assurance," *IEEE Trans. Wireless Commun.* vol. 5, pp. 1382–1391, June 2006.

[18] H. S. Wang and N. Moayeri, "Finite-state Markov channel–A useful model for radio communication channels," *IEEE Trans. Commun.*, vol. 44, pp. 163–171, Feb. 1995.

[19] R. H. Deng and M. L. Lin, "A type I hybrid ARQ system with adaptive code rates," *IEEE Trans. Commun.*, vol. 18, pp. 733–737, Feb./Mar./Apr. 1995.

[20] J. Munkres, "Algorithms for assignment and transportation problems," *J. Society Industrial Applied Mathematics*, vol. 5, no. 1, pp. 32–38, Mar. 1957.

[21] F. Burgeois and J. C. Lasalle, "An extension of the Munkres algorithm for the assignment problem to rectangular matrices," *Commun. ACM*, vol. 14, no. 12, pp. 802–806, Dec. 1971.

[22] M. Schwartz, *Broadband Integrated Networks*. Upper Saddle River, NJ: Prentice Hall, 1996.

[23] S. Lin and D. J. Costello, Jr., *Error Control Coding*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2004.

[24] J. Proakis, *Digital Communications*. New York: McGraw-Hill, 1995.

Jun Xu received the B.Eng. from Tsinghua University, China, and MASc in Systems Design Engineering (2002) and Ph.D. in Electrical Engineering (2006) from the University of Waterloo, Ontario, Canada. His research interests include radio resource management for multimedia in 3G and beyond wireless communication systems, hybrid terrestrial/satellite systems, cross-layer design and optimization, wireless inter-networking, and the application of artificial intelligence (e.g., Fuzzy Logic and Neural Networks) in communication networks. Dr. Xu is currently with Hughes Network Systems in Germantown, Maryland, USA, working on the system design of next generation hybrid terrestrial/satellite systems.

Xueming (Sherman) Shen (M'97-SM'02) received the B.Sc.(1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a Professor and the Associate Chair for Graduate Studies, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on mobility and resource management in interconnected wireless/wired networks, UWB wireless communications systems, wireless security, and ad hoc and sensor networks. He is a co-author of three books, and has published more than 300 papers and book chapters in wireless communications and networks, control, and filtering. Dr. Shen serves as the Technical Program Committee Chair for IEEE Globecom'07, General Co-Chair for Chinacom'07 and QShine'06, and is the Founding Chair for the IEEE Communications Society Technical Committee on P2P Communications and Networking. He also serves as a Founding Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS; Editor-in-Chief for *Peer-to-Peer Networking and Application*; Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY; the *KICS/IEEE Journal of Communications and Networks, Computer Networks*; *ACM/Wireless Networks*; and *Wireless Communications and Mobile Computing* (Wiley), etc. He has also served as Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATION, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *IEEE Communications Magazine*. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 from the Faculty of Engineering, University of Waterloo. Dr. Shen is a registered Professional Engineer of Ontario, Canada.

Jon W. Mark (M'62-SM'80-F'88-LF'03) received the Ph.D. degree in electrical engineering from McMaster University, Canada in 1970. Upon graduation, he joined the Department of Electrical Engineering (now Electrical and Computer Engineering) at the University of Waterloo, became a full Professor in 1978, and served as Department Chairman from July 1984 to June 1990. In 1996, he established the Centre for Wireless Communications (CWC) at the University of Waterloo and has since been serving as the founding Director. He co-authored the recent textbook *Wireless Communications and Networking*, Prentice-Hall, 2003.

He is the recipient of the 2000 Canadian Award in Telecommunications Research *for significant Research contributions, scholarship and leadership in the fields of computer communication networks and wireless communications*, and the 2000 Award of Merit by the Educational Foundation of the Association of Chinese Canadian Professionals *for Significant Contributions in Telecommunications Research*.

His current research interests are in wireless communications and wireless/wireline interworking, particularly in the areas of resource management, mobility management, and end-to-end information delivery with QoS provisioning.

Dr. Mark is a Life Fellow of the IEEE and has served as a member of editorial boards, including the IEEE TRANSACTIONS ON COMMUNICATIONS, *ACM/Baltzer Wireless Networks*, *Telecommunication Systems*, etc. He was a member of the Inter-Society Steering Committee of the IEEE/ACM Transactions on Networking from 1992-2003, and served as the SC Chair during 2002-2003.

Jun Cai (M'04) received the B.Sc. (1996) and the M.Sc. (1999) degrees from Xi'an Jiaotong University (China) and Ph.D. degree (2004) from University of Waterloo, Ontario (Canada), all in electrical engineering. From June 2004 to April 2006, he was with McMaster University as a NSERC Postdoctoral Fellow. Since July 2006, he has been with the Department of Electrical and Computer Engineering, University of Manitoba, Canada, where he is an Assistant Professor. His current research interests include multimedia communication systems, mobility and resource management in 3G beyond wireless communication networks, and ad hoc and mesh networks. He is currently a holder of an NSERC Associated Industrial Research Chair.