

# A Fresh Look at Optimal Subcarrier Allocation in OFDMA Systems

Somsak Kittipiyakul and Tara Javidi

Department of Electrical Engineering, University of Washington, Seattle, USA.

{somsakk,tjavidi}@ee.washington.edu

**Abstract**—This paper considers the issue of optimal subcarrier allocation in OFDMA. We show, via a counter example, that water-filling based subcarrier allocation policies, contrary to conventional wisdom, fail to provide rate-stability for an otherwise stabilizable OFDMA system. Water-filling is too myopic when considering long-time average performance, e.g. delay, queue lengths, and even long-run throughput. This is because such policies ignore variable state (queue length) information, while, in fact, such an information is necessary to guarantee rate stability and/or to minimize average delay. In this paper, we identify an optimal non-idling policy which balances the queue lengths, when the channel follows an ON/OFF model. In such case, we show that such a policy achieves the minimum average holding cost (mean response time) at any time.

## I. INTRODUCTION

Orthogonal frequency-division multiplexing (OFDM) is a promising technique to provide multiple access control (MAC) in high-speed wireless applications (e.g. broadband wireless, 4G systems, LANs) in a hostile multipath environment with frequency-selective fading. OFDM achieves high spectral efficiency in multiuser environment by dividing the total available bandwidth to narrow subbands in an efficient way. This allows the mobiles to spread their information selectively in order to avoid sub-bands where (frequency-selective) fading occurs. This results in higher spectral efficiency since fading usually experienced by different mobiles are statistically independent. In a single user case, it is known that, to achieve the highest spectral efficiency, the optimal resource allocation must schedule these sub-carriers with the best SNR.

The problem of optimal real-time subcarrier allocation has been recently studied ([2], [6], [8], [10], [14], [17], [19]). The prior work can be categorized into two classes based on the optimization objectives. The objective in the first class of work ([8], [17], [19]) is to minimize the total transmit power given constraints on Quality of Service (QoS) requirements of each user. These constraints include fixed data rate or acceptable bit error rate (BER).

The second class of papers ([6], [14]) attempt to maximize the total throughput at each decision epoch given a constraint on a maximum transmit power per data stream. Our work is similar to this class of papers in that we focus on maximizing the system throughput when considering the subcarrier allocation problem in an OFDM system. We believe that, although there will always be numerous applications for which the power maximization is critical,

in many commercial applications of future wireless systems, achieving high connection speed (data rate) will be of primary concern rather than power.

On the other hand, most of the papers in the literature ([2], [6], [14], [15], and [18]) differ from our work in that they all provide solutions to a multi-user water-filling problem achieving Shannon capacity under power constraint at each decision epoch (in this paper, we refer to this as “instantaneous throughput maximization” or “multi-user water filling”).

In this context, the present paper is a part of an ongoing effort to establish a systematic approach to a series of throughput-related problems which have recently been observed ([4], [10]) in OFDM systems. In [4], a subcarrier allocation based solely on queue lengths for an MPEG-4 video transmission application is studied. It is shown that a static allocation performs poorly when compared to the queue-length-based allocation. The authors in [10] propose a subcarrier allocation in an OFDM system with finite buffer space. They show through simulations that water-filling solutions perform poorly with respect to a long-run throughput criterion due to buffer overflow.

In this paper, we focus on a long-term average performance, e.g. average queue backlog, rather than an instantaneous optimization. We show that even without a constraint on buffer sizes, water-filling-based techniques perform poorly when considered over a considerably long time interval. We argue that maximizing instantaneous throughput (water-filling) is too myopic, i.e. it fails to take into account the varying state of the system and queue build-ups. Through an example we show how the policy that maximizes the instantaneous throughput causes rate-instability (unbounded queue buildups) in an otherwise stabilizable system. This implies a suboptimal performance with respect to throughput (as well as delay and total number of packets waiting in the system). As a result, we propose a long-run objective to minimize the average holding cost in the system.

In this paper, we assume a pre-determined power allocation. This power level is assumed to vary in order to compensate for the distances between various users and the base (fixed point). Under such assumption, a user’s channel condition (across sub-carriers) can be mapped to the number of transmitted packets at each subcarrier as follows. Using adaptive modulation/coding, the number of packets a subcarrier can transmit per time slot is expressed

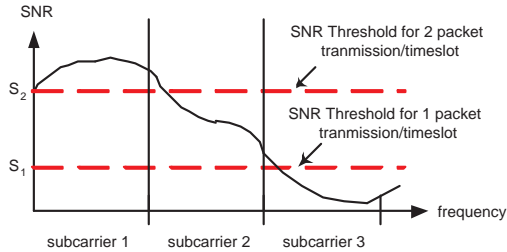


Fig. 1. Mapping of a User's SNR to packet capacity for each subcarrier. The user can transmit 2, 1, and 0 packets on subcarrier 1, 2, and 3, respectively

as function of SNR [8]. Figure 1 shows such a procedure for an example. In this example, we assume that there are two transmission types (modulation/coding) over a subband: the first type requires a certain SNR ( $\text{SNR} > S_1$ ) and transmits one packet per a timeslot; the other transmission type requires a higher SNR ( $\text{SNR} > S_2$ ) and transmits two packets per a timeslot. The figure shows that the user can transmit 2, 1, and 0 packets on subcarrier 1, 2, and 3, respectively. As a result we map the channel condition given in Figure 1 to a "connectivity profile" (2, 1, 0).

In this paper we identify an optimal policy where the connectivities can be modeled by a simple ON/OFF (1-0) state. Furthermore, we assume that all users have the same priority. We show that under such an assumption a non-idling (maximum throughput (MT)) Load Balancing (LB) policy minimizes the expected holding cost. We then propose an algorithm to construct MTLB.

The paper is organized as follows. Section II gives a counter example of how the instantaneous throughput maximization fails to stabilize a stabilizable system. Section III provides problem formulation and assumptions of our model. Section IV defines the MTLB policy and prove its optimality. Section V, provides a computation algorithm based on ideas from bipartite graphs and matching literature. Finally, Section VI concludes the paper and discusses future studies.

## II. COUNTER EXAMPLE

Consider a simple ON/OFF system with two users ( $u_1, u_2$ ) and two sub-carriers ( $B_1, B_2$ ). The initial queue sizes are zero. The packet arrival processes and the channel connectivity processes for users 1 and 2 have periodic structures with period of four time slots as shown in Figure 2(a) and Figure 2(b), respectively. For example, at the beginning of timeslot 1, two and one packets arrive for  $U_1$  and  $U_2$ , respectively. Furthermore, during slot 1, user  $U_1$  can use both sub-carriers while user  $U_2$  can use only subcarrier  $B_1$ .

The system is stabilizable because the periodic policy  $\pi^*$ , whose subcarrier allocation  $\omega_{ij}^*$  at each of the four timeslot period is marked by  $\checkmark$  in Figure 2(c), stabilizes the queues by emptying the system every four time slots. For example, during time slot 1, the policy  $\pi^*$  assigns user  $U_1$  with subcarrier  $B_2$  and user  $U_2$  with subcarrier  $B_1$ . Now

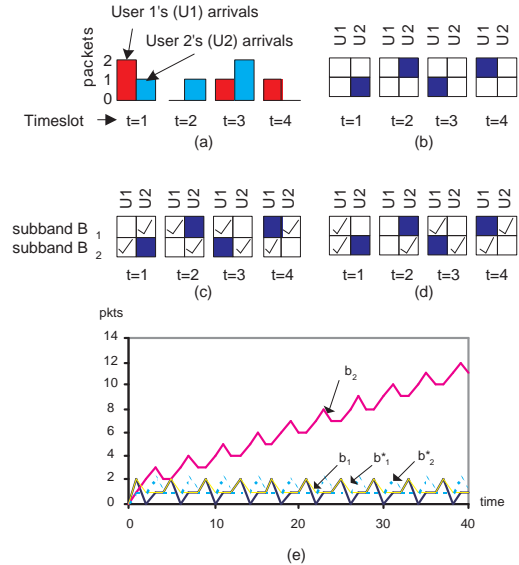


Fig. 2. A counter-example showing that water-filling causes unbounded queue buildups. We assume that the packet arrival process and the channel connectivity process have periodic structures with period of four timeslots. (a) Packet arrivals for users 1 and 2 (b) Channel connectivity profile (white square means ON) (c) MTLB policy  $\pi^*$  (d) Maximal Instantaneous Throughput policy  $\pi$  (e) Queue occupancies over time: queue lengths for users  $U_1$  and  $U_2$  under  $\pi$ , ( $b_1, b_2$ ), and under  $\pi^*$ , ( $b_1^*, b_2^*$ ). Under policy  $\pi$ , queue of user  $U_2$  ( $b_2$ ) is building up over time.

consider an alternative policy  $\pi$  shown in Figure 2(d). It is easy to see that under policy  $\pi$  the instantaneous throughput is maximized at every time slot, but the system is unstable as the length of the queue of user  $U_2$  grows by one every four units of time. Note that under policy  $\pi$ , user  $U_1$ 's buffer is empty at time  $t = 2$ . This results in the idling of subband  $B_1$ . Consequentially, policy  $\pi$  results in longer total queue occupancies at all times and thus less total throughput than policy  $\pi^*$ . This simple counter example shows that maximizing the instantaneous throughput is not sufficient and in general fails to stabilize an otherwise stabilizable system.

## III. PROBLEM FORMULATION AND ASSUMPTIONS

### Notations and Definitions

We consider a single-hop OFDMA system composed of one cell or cluster with one base station. We assume that there are  $N$  users and  $K$  sub-carriers

In this paper, we use arguments to denote the time index, subscript  $i$  to denote specific subcarrier, and subscript  $j$  to denote specific user/queue. For example,  $w_{ij}(n)$  denotes the number of packet withdrawals at time  $n$  for user  $j$  over subcarrier  $i$ . We use small letters without any subscription for deterministic row vectors and capitalized letters for deterministic matrices. Scripted capitalized letters for space of all possible vectors or matrices. For example,  $\mathcal{W}$  is the space of all possible packet withdrawal matrices.

- $b(n) = (b_1, \dots, b_N)$ : the row vector of queue occupancies at the beginning of time  $n$ .

- $a(n) = (a_1, \dots, a_N)$ : the arrival to queue  $a_j$  during time  $n$ .
- $C(n) = \{c_{ij}\}$ : The  $K$ -by- $N$  channel connectivity matrix or connectivity profile at time  $n$  where  $c_{ij}$  denotes the maximum number of packets subcarrier  $i$  can serve from queue  $j$ .
- $(b(n), C(n))$ : the state of the system at time  $n$ .
- $W(n) = \{\omega_{ij}\}$ : the  $K$ -by- $N$  packet withdrawal matrix (also referred to as server allocation or assignment) at time  $n$ , where  $\omega_{ij}$  denotes the number of packets that subcarrier  $i$  is assigned to serve queue  $j$ . We use both the matrix form of  $W = \{\omega_{ij}\}$  and the row vector form  $\omega = (\omega_1, \dots, \omega_N)$  interchangeably where  $\omega_j = \sum_{i=1}^K \omega_{ij}$  is the packet withdrawal of queue  $j$ .

Here are some definitions we will use in the paper:

*Definition 1:* For a row vector  $x = (x_1, \dots, x_N)$  and a matrix  $Y = (y_1, \dots, y_N)$  where  $y_i$  is a column vector, a column-by-column matrix permutation  $\Pi_\pi$  corresponding to a permutation  $\pi$  is defined as, for any  $i$  and  $j$ ,

$$\pi(x_i) = x_j \Leftrightarrow \Pi_\pi(y_i) = y_j$$

*Definition 2:* Consider a random vector  $\mathbf{x}$  with a joint distribution  $P_{\mathbf{x}}(x)$ . We say  $\mathbf{x}$  is permutation invariant if, for any permutation  $\pi$ ,  $P_{\mathbf{x}}(\pi(x)) = P_{\mathbf{x}}(x)$ .

*Definition 3:* Consider a random matrix  $\mathbf{Y}$  where elements  $Y_{ij}$  are random variables with joint pdf  $P_{\mathbf{Y}}(Y)$ . We say  $\mathbf{Y}$  is column-by-column permutation invariant if, for any permutation  $\pi$ ,  $P_{\mathbf{Y}}(\Pi_\pi(Y)) = P_{\mathbf{Y}}(Y)$ .

### Assumptions

- (A1) The sub-carriers are time-slotted and have fixed and flat fading during a time slot.
- (A2) The set of modulation and coding available to various users are similar and fixed. For a given channel state  $c_{ij}$  and a transmission type (choice of modulation/coding),  $\omega_{ij}$  packets are transmitted by user  $j$  over subband  $i$ , with a maximum of  $c_{ij}$  packets.
- (A3) If a channel can serve  $c_{ij}$  packets from the  $j^{\text{th}}$  queue, all of these packets will be transmitted with probability of success equal to one. In another word, this model does not capture the loss probability over a wireless channel of good quality ( $c_{ij} > 0$ ).
- (A4) Packets are of equal length. A subcarrier can serve at most  $c_{max}$  packet per time slot when used by any users.
- (A5) Each user has an infinite buffer.
- (A6) The channel state of each subcarrier to a user is reduced to either ON or OFF i.e.  $c_{max} = 1$ . A subcarrier is in the ON state when the channel gain is less than the threshold gain meeting the minimum required SNR. When the subcarrier is in the ON state, the user can utilize the subcarrier to transmit up to one packet. Thus, a subcarrier is assumed to be either connected or disconnected from users as shown in Figure 3. For example,  $c_2 = (c_{12}, c_{22}, \dots, c_{K2})^T =$

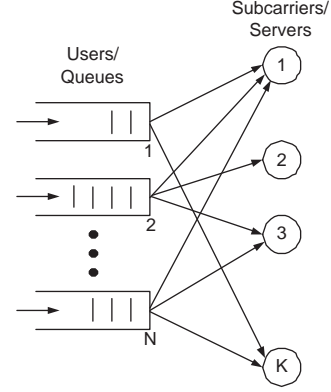


Fig. 3. Subcarrier Allocation Problem

$(1, 1, 1, 0, \dots, 0)^T$  indicates that user 2 can transmit only on sub-carriers 1, 2, and 3.

- (A7) The packet arrival processes  $\{\mathbf{A}(n)\}$  to users' queues during each time slot is independent across time slots. The packet arrival process is such that the joint pdf is permutation invariant as defined in Definition 2. This implies that within each time slot the arrivals to various queues need not be independent. In other words, there can be possible dependency of arrivals amongst different users.
- (A8) The connectivity profiles  $\{\mathbf{C}(n)\}$  is such that the joint pdf is permutation invariant as defined in Definition 3.
- (A9) At the beginning of each time slot, the centralized resource manager at the base station uses the perfect information of queue sizes and the connectivity profiles to make subcarrier allocation for that time slot. Then, the assignment is announced immediately to all users via a separate control channel.
- (A10) We do not allow for the sharing of the servers, i.e. each server can only serve one packet at a time from a single user's queue ( $\omega_{ij} \in \{0, 1\}$ ).

We believe that allowing for individual and non-identical connectivity profile is essential in modeling OFDM. The inclusion of individual connectivity profiles in our model is essential to a reasonable modeling of OFDM systems. This captures the following phenomena: a subcarrier can have a "good" channel response for one user while on the same subcarrier another user may encounter severe channel degradation, or "bad" channel.

### Problem Formulation

We now formulate an abstract problem that captures essential features of the described OFDM problem.

#### Problem (P)

Consider a discrete-time model of  $N$  queues  $(Q_1, \dots, Q_N)$  served by  $K$  servers ( $K > N$ ). At each time, each server can serve one packet from one queue; we allow for one queue to be served by multiple servers. At each time, a queue  $j$  is either available to be served by a server  $i$  (connected or ON) or it is not available (disconnected or

OFF) (A6). At each time, the connectivities of all queue/server pairs are known for that time. We allow for arrivals at each queue at each time and arrivals at a given time are assumed to occur after server allocations at that time. The statistics of arrival and connectivity processes are assumed to satisfy (A7) and (A8). We wish to determine a Markov server (subcarrier) allocation policy  $\pi$  that minimizes the cost function at the finite horizon  $T$ :

$$J_T^\pi = E[C_T^\pi | \mathcal{F}_0] \quad (1)$$

where  $\mathcal{F}_0$  summarizes all information available at the beginning of the allocation period.  $C_T^\pi$  is the cost under Markov policy  $\pi$  over horizon  $T$ .

$$C_T^\pi = \sum_{t=0}^T \sum_{j=1}^N \phi(b_j(t)) \quad (2)$$

where  $\phi(\cdot)$  is any convex and increasing function and  $b_j(t)$  is the queue length of  $Q_j$ .

Restriction to Markov policies does not entail any loss of optimality because Problem (P) is a stochastic control problem with perfect observations [9].

One simple example of  $\phi(\cdot)$  is identity function. In that case, Problem (P) reduces to a total backlog ( $\sum_t \sum_j b_j(t)$ ) minimization problem over horizon  $T$ .

In Problem (P) we have assumed that the horizon  $T$  is finite. We first analyze Problem (P) and its refinement and then show that the result of the analysis hold for the corresponding infinite horizon problem.

#### Prior Work

Our problem formulation is very similar to the problem of transmission scheduling for wireless and satellite nodes where a limited number of transmitters (servers) or channels have to be allocated to competing users with varying connectivity ([1], [3], [11], [16]).

The authors in [16] consider the server allocation problem of a single server to  $N$  competing queues. At each time slot each queue may be connected or disconnected to the server, depending on a binary connectivity random variable. They show that the Longest Connected Queue (LCQ) policy stabilizes the system if the system is stabilizable and minimizes the delay for the special case of symmetric queues.

The authors in [3] further show that, in the case of  $K$  servers =  $N$  queues and the constraint that at most  $C$  packets can be served in total in each time slot and fractional packets are allowed to be served to each queue, the optimal policy is to serve the queues such that the resulting queue lengths are most balanced. The authors in [3] allow for sharing of the servers (serving a fraction of packet from a set of queues). Furthermore, they do not allow users to have distinct connectivity profiles (In this case  $C(n)$  are reduced to a vector).

In addition, the model used in [1] is similar to the one used in our paper. The authors in [1] consider the problem of batch allocation of bandwidth or servers to multiple queues. The study is focused on the delay in the observations of channel and queue lengths. Again, the difference with our model is that [1] assumes identical connectivity profile while we allow for distinct connectivity profiles across queues. However, we do not consider observation delay with respect to queue length nor do we address imperfect channel estimation.

#### IV. ANALYSIS OF PROBLEM (P)

In this section, we identify load balancing (MTLB) subcarrier allocation policy, as optimal with respect to the cost function in Problem (P). We use dynamic programming to establish the optimality of MTLB.

*Definition 4:* Given state  $(b, C)$  at the beginning of time slot  $n$ , an allocation  $W = \{\omega_{ij}\}$  is a feasible allocation iff

- **(C1.a)**  $\omega_{ij} \leq c_{ij}$ ;
- **(C1.b)**  $\sum_{j=1}^N \omega_{ij} \leq 1, \forall i = 1, \dots, K$ ; and
- **(C1.c)**  $\sum_{i=1}^K \omega_{ij} \leq b_j, \forall j = 1, \dots, N$ .

Given a state  $(b, c)$ , the set of all feasible row allocations  $\omega = (\omega_1, \dots, \omega_N)$ ,  $\omega_i = \sum_j \omega_{ij}$ , is denoted by  $\mathcal{W}(b, C)$ ,

*Definition 5:* Given state  $(b, C)$  at the beginning of time slot  $n$ , the MTLB policy chooses a (feasible) packet withdrawal matrix  $W^*(n) = \{\omega_{ij}^*\}$  such that

**(C1) Maximum Throughput:**  $W^*(n) \in \mathcal{W}(b, C)$  achieves the maximum throughput, i.e.

$$\sum_{j=1}^N \sum_{i=1}^K \omega_{ij}^* \geq \sum_{j=1}^N \sum_{i=1}^K \omega_{ij} \quad \text{for all } \omega \in \mathcal{W}(b, C) \quad (3)$$

**(C2) Load Balancing:** Let  $L$  be the maximum throughput achieved in (C1) and  $\mathcal{W}_L(b, C) \subset \mathcal{W}(b, c)$  contains all possible assignments achieving throughput  $L$ .  $\omega^*(n) \in \mathcal{W}_L(b, C)$  produces the most balanced queue i.e.

$$(b - \omega^*) \leq_{LQO} (b - \omega) \quad \text{for all } \omega \in \mathcal{W}_L(b, C) \quad (4)$$

Theorem 1 summarizes the main result of our study.

*Theorem 1:* Consider Problem (P) with a finite horizon  $T$ . MTLB policy is optimal for any initial state  $\mathcal{F}_0 = (b, C)$ .

Note that condition (C1) is a water-filling condition. In other words, Theorem 1 implies that water-filling criteria is not sufficient to guarantee long-term throughput optimality unless it is complemented by a load-balancing criteria (condition (C2)).

*Outline of the Proof:* (see details in [7]) Given a horizon  $n$ , define  $V_n^\pi(b, C)$  the cost-to-go under Markov policy  $\pi$ .

$$V_n^\pi(b, C) = \phi(b) + \sum_{a, \tilde{C}} p_A(a) p_C(\tilde{C}) V_{n-1}^\pi(b + a - \omega, \tilde{C})$$

where  $\phi(b) = \sum_{j=1}^N \phi(b_j)$ .

From dynamic programming, we have the following recursion for the optimal cost-to-go  $V_n^*(b, C)$ :

$$V_0^*(b, C) = \phi(b)$$

and

$$V_n^*(b, C) = \phi(b) + \inf_{\omega \in \mathcal{W}(b, C)} \sum_{a, \tilde{C}} p_A(a) p_C(\tilde{C}) V_{n-1}^*(b+a-\omega, \tilde{C})$$

Define

$$\begin{aligned} v_n(b) &= E_{a, C} [V_{n-1}^*(b+a, C)] \\ &= \sum_{a, \tilde{C}} p_A(a) p_C(\tilde{C}) V_{n-1}^*(b+a, \tilde{C}). \end{aligned} \quad (5)$$

Since  $\mathcal{W}(b, C)$  is finite, there exists an optimal packet withdrawal  $\omega^*(n, b, C)$  at time  $n$  when the state of the queue backlogs is equal to the vector  $b$  and the connectivity profile is  $C$ . The optimal cost-to-go can be rewritten as:

$$V_0^*(b, C) = \phi(b)$$

and

$$\begin{aligned} V_n^*(b, C) &= \phi(b) + \min_{\omega \in \mathcal{W}(b, C)} v_n(b-\omega) \\ &= \phi(b) + v_n(b - \omega^*(n, b, C)) \end{aligned} \quad (6)$$

In order to show that MTLB is optimal, we need the validity of the following statements (proved in [7]).

( $\mathcal{H}0$ )  $v_n(b)$ ,  $n = 0, \dots, T$ , are permutation invariant and strictly monotonic.

( $\mathcal{H}1$ ) If a feasible allocation  $w$  does not satisfy (C1), then there is a  $j$  such that  $w + e_j \in \mathcal{W}(b, C)$ .

Note, assuming the validity of ( $\mathcal{H}0$ ), all MTLB policies have the same expected cost. In other words, define

$$\mathcal{W}^*(n, b, C) := \left\{ \omega^* : v_n(b - \omega^*) = \min_{\omega \in \mathcal{W}(b, C)} v_n(b - \omega) \right\}$$

$$\mathcal{B}^*(n, b, C) := \left\{ d^* : v_n(d^*) = \min_{\omega \in \mathcal{W}(b, C)} v_n(b - \omega) \right\}$$

In [7], we present the following statement:

( $\mathcal{H}2$ ) If  $\omega^*(n, b, C) \in \mathcal{W}^*(n, a+b, C)$ , it satisfies Condition (C1) in the definition of MTLB.

We prove the rest of Theorem 1 by an induction on  $n$ , the number of stages to go. For that matter we define

$$\begin{aligned} \mathcal{G}_n(b) &= [v_n(b + e_i + e_j) + v_n(b)] \\ &\quad - [v_n(b + e_i) + v_n(b + e_j)] \end{aligned} \quad (7)$$

$$\mathcal{D}_n(b) = v_n(b + e_1) - v_n(b + e_2) \quad (8)$$

Using the above expressions, we state the induction hypotheses for stage  $n$  as

( $\mathcal{H}3$ ) $_n$   $\mathcal{G}_n(b) \geq 0$ , for every state  $b$  and indexes  $1 \leq i, j \leq N$ .

( $\mathcal{H}4$ ) $_n$   $\mathcal{D}_n(b) \geq 0$ , for every  $b$  such that  $b_1 \geq b_2$ .

( $\mathcal{H}5$ ) $_n$  Policy  $g^*$ , that allocates servers according to MTLB, is optimal at stage  $n$ .

We note that ( $\mathcal{H}5$ ) is sufficient to assert the validity of Theorem 1. However, to prove inductively that ( $\mathcal{H}5$ ) $_{n+1}$  is true we need ( $\mathcal{H}3$ ) $_n$ -( $\mathcal{H}5$ ) $_n$ . Specifically, ( $\mathcal{H}3$ ) $_n$  can be

interpreted as the convexity of  $v_n$  in a discrete setting and ( $\mathcal{H}4$ ) $_n$  establishes the balancing advantage.

Specifically, we outline how ( $\mathcal{H}3$ ) $_n$  and ( $\mathcal{H}4$ ) $_n$  are used to inductively establish the induction step via the following lemmas (see [7]):

*Lemma 1:* ( $\mathcal{H}0$ ), ( $\mathcal{H}2$ ), ( $\mathcal{H}3$ ) $_n$  and ( $\mathcal{H}5$ ) $_n \implies$  ( $\mathcal{H}3$ ) $_{n+1}$ .

*Lemma 2:* ( $\mathcal{H}0$ ), ( $\mathcal{H}2$ ), ( $\mathcal{H}3$ ) $_n$ , ( $\mathcal{H}4$ ) $_n$ , ( $\mathcal{H}5$ ) $_n \implies$  ( $\mathcal{H}4$ ) $_{n+1}$ .

*Lemma 3:* ( $\mathcal{H}0$ ), ( $\mathcal{H}2$ ), ( $\mathcal{H}4$ ) $_{n+1} \implies$  ( $\mathcal{H}5$ ) $_{n+1}$ .

In addition, we prove the following Corollary in [7] to extend Theorem 1 to the infinite horizon Problem.

*Corollary 1:* Consider an infinite horizon version of the Problem (**P**), where the cost is modified to be the average expected cost at each stage. Then MTLB is optimal for any initial state  $\mathcal{F}_0 = (b, C)$ .

## V. COMPUTATION OF MTLB

Using concepts from graph matching literature ([5],[13]), we propose an iterative algorithm (iMTLB) to construct MTLB allocation. We consider our subcarrier allocation problem as a matching problem over a graph  $G$  where the queues and servers are considered as nodes and the connectivities between them as edges.

### Equivalent Bipartite Construction (EBC)

- 1) Associated with each queue  $j$ , construct  $m_j = \min(b_j, c_j)$  nodes labeled as  $a_{j1}, a_{j2}, \dots, a_{jm_j}$ .
- 2) Let  $U^X = \{a_{11}, a_{12}, \dots, a_{1m_1}, a_{21}, \dots, a_{Nm_N}\}$  be the set of all such nodes.
- 3) Let  $V = \{v_1, \dots, v_K\}$  be the set of servers.
- 4) Let  $E^X = \{(a_{jm}, v_i)\}$  if  $c_{ij} = 1$  be the set of edges representing connectivities.

### Iterative Algorithm (iMTLB)

- 1) Use (EBC) to construct an equivalent bipartite graph  $G_{eq} = (U^X, E^X, V)$  from the given  $(b, C)$ .
- 2) Find a maximum matching  $M$  in  $G_{eq}$ .
- 3) Convert  $M$  to subcarrier assignment in  $G$ .
- 4) While there exists a balancing path  $\omega$  with respect to  $G$ , balance  $G$  by  $\omega$  (see [7]).

An example of iMTLB algorithm is shown in Figure 4. In [7], we show that the proposed iMTLB algorithm converges to MTLB allocation in finite time. We conjecture that our iterative algorithm is equivalent to an existing maximum weighted matching for the equivalent bipartite graph and its worst case running time is  $O(K^2(N + \log K))$  [13].

## VI. CONCLUSION AND FUTURE RESEARCH

In this paper, we showed through a counter example that although water-filling based subcarrier allocation policies maximize the instantaneous throughput, they in general fail to provide rate-stability for an otherwise stabilizable OFDMA system. We used this example to argue that water-filling is too myopic and ignores variable state (queue length) information. We then identified a MTLB policy that achieves the instantaneous maximum throughput as well as balancing the queue lengths. Such a policy always

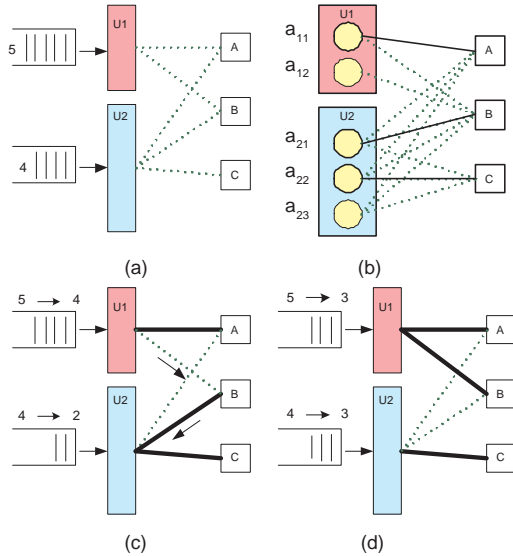


Fig. 4. Example of one iteration of iMTLB Algorithm (a) queue lengths and connectivities (b) construction of an MT allocation from the equivalent bipartite graph (c) dark edges indicate the MT policy; the arrows show a balancing correction to achieve a load-balancing configuration (d) dark edges indicated the MTLB policy after balancing allocation

exists when the channel follows an ON/OFF model. In such case, we showed that MTLB achieves the minimum average number of packets waiting in the system at any time.

For practical implementation of MTLB with a large number of users and subcarriers, we hope to use randomized algorithm to reduce the running time and simplify the implementation. In addition, the implementation of MTLB requires the existence of a centralized controller. This could be problematic for a system with large number of users and subcarriers. We hope to construct a decentralized allocation algorithm in our future research.

Furthermore, we extend our example (see [7]) to show that, when multiple values of packet transmission capacity is considered, maximizing instantaneous throughput (water-filling) is not only not sufficient for optimality but is not even a necessary condition. It may be optimal to sacrifice instantaneous throughput in order to set the state of the system in "better" (more balanced) state in anticipation of loss of connectivities in the future. This is a manifestation of known results in queuing theory and switching. The MDP technique used in this paper is not expected to result in identification of an optimal policy under these more complicated channel models. We hope to use fluid flow analysis to establish rate-stability for a load balancing policy in such scenarios.

#### ACKNOWLEDGMENT

This research was supported in part by the National Science Foundation ADVANCE Cooperative Agreement No. SBE-0123552.

#### REFERENCES

[1] N. Ehsan and M. Liu, "On the optimality of an index policy for bandwidth allocation with delayed state observation and differenti-

ated services", to appear in *Proc. IEEE INFOCOM*, April 2004, Hong Kong.

[2] M. Ergen, S. Coleri and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Transactions on Broadcasting*, vol. 49, no. 4, Dec. 2003.

[3] A. Ganti, *Transmission Scheduling for Multi-Beam Satellite Systems*, Doctoral Thesis, Dept. of EECS, MIT, Cambridge, MA, 2003.

[4] J. Gross, J. Klaue, H. Karl and A. Wolisz, "Subcarrier allocation for variable bit rate video streams in wireless OFDM systems," *Proc. of Vehicular Tech. Conf. (VTC)*, Florida, USA, 2003.

[5] N. Harvey, R. Ladner, Laszlo Lovasz, and T. Tamir, "Semi-matchings for bipartite graphs and load balancing," *Proc. of the Workshop on Algorithms and Data Structures (WADS '03)*, Ottawa, Canada, July 2003.

[6] J. Jang, K.B. Lee, and Y.H. Lee, "Transmit Power and Bit Allocations for OFDM Systems in a Fading Channel," *Proc. IEEE GLOBECOM 2003*, Dec. 2003.

[7] S. Kittipiyakul and T. Javidi, "Resource Allocation in OFDMA: How Load-balancing Maximizes Throughput When Water-filling Fails", *UW Technical Report*, University of Washington, 2004

[8] D. Kivanc, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. on Wireless Comm.*, vol. 2, no. 6, Nov. 2003.

[9] R. Kumar and P. Varaiya, *Stochastic Control*. Prentice-Hall, 1986.

[10] G. Li and H. Liu, "Dynamic resource allocation with finite buffer constraint in broadband OFDMA networks," *IEEE Wireless Comm. and Networking*, v. 2, pp. 1037-1042, March 2003.

[11] C. Lott and D. Teneketzis, "On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes," *Prob. in the Eng. and Info. Services*, vol. 14, no. 3, pp. 259-297, July 2000.

[12] U. Manber, *Introduction to Algorithm: a creative approach*, Addison-Wesley Publishing Company, 1989.

[13] K. Mehlhorn and S. Näher, *The LEDA Platform of Combinatorial and Geometric Computing*, Cambridge University Press, 1999.

[14] S. Pfletschinger, G. Munz, J. Speidel, "An efficient water-filling algorithm for multiple access OFDM," *IEEE Globecom '02*, Taipei, Taiwan, November 2002.

[15] W. Rhee and J.M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," *Proc. of Vehicular Tech. Conf. (VTC), 2000*, vol. 2, pp.1085-1089, May 2000.

[16] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. on Info. Theory*, vol. 39, no. 2, pp. 466-478, 1993.

[17] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. March, "Multiuser OFDM with adaptive subcarrier, bit and power allocation," *IEEE JSAC.*, vol. 17, no. 10, pp. 1747-1757, Oct. 1999.

[18] C. Yih and E. Geraniotis, "Adaptive modulation, power allocation and control for OFDM wireless networks", *Proc. IEEE Int. Symp. on Personal, Indoor, Mob. Radio Commun.*, vol.2, pp.819-813, 2000.

[19] H. Yin, H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," *Globecom '00*, San Francisco, USA, 2000.