Wireless Pers Commun (2009) 51:565–591
DOI 10.1007/s11277-009-9754-8

Cross-Layer QoS Scheduling for Layered Multicast Streaming in OFDMA Wireless Networks

Yuedong Xu · Xiaoxin Wu · John C. S. Lui

Published online: 10 July 2009

Abstract The conventional multicast scheme of wireless networks, though establishing a bandwidth-saving means for point-to-multipoint transmission, is very conservative by limiting the throughput of short-range communications. The multicast performance can be significantly improved if some low-rate users are pruned. In this paper, we investigate the subchannel assignment mechanism of multicast streaming services in the emerging WiMax/802.16e systems, where each multimedia stream is composed of a *basic layer* and an *enhancement layer*. The former affords a low-resolution video image to all the subscribers, while the latter only serves those with preferable channel states. Optimization frameworks are formulated to characterize the QoS requirements of multicast flows: *pruned proportional rate ratio* (*PPRR*), *pruned stream rate guarantee* (*PSRG*) and *pruned user proportional fairness* (*PUPF*). Three cross-layer algorithms are presented to perform channel assignment for different QoS requirements. Analytical study shows that the proposed algorithms have polynomial-time computational complexity. Numerical experiments validate that our proposals significantly outperform the conventional peer schedulers in terms of system throughput.

Keywords Multicast · Scheduling · OFDMA · QoS guarantee · Subchannel assignment · Cross-layer design

Y. Xu (✉) · J. C. S. Lui
Department of Computer Science & Engineering, The Chinese University of Hong Kong,
Shatin, NT, Hong Kong SAR
e-mail: ydxu@cse.cuhk.edu.hk

J. C. S. Lui
e-mail: cslui@cse.cuhk.edu.hk

X. Wu
Intel China Research Center Ltd, Beijing, China
e-mail: xiaoxin.wu@intel.com

1 Introduction

During the past few years, new broadband wireless access (BWA) technologies have been brought into commercial deployment such as CDMA 1xEV-DO, UMTS HSDPA, WiMax, enabling high data rates and large coverage. Especially, WiMax air interface utilizes orthogonal frequency division multiplex (OFDM) or orthogonal frequency division multiple access (OFDMA) mechanisms to improve system performance over mobile and non-line-of-sight (NLOS) scenarios [5]. Dubbed as mobile WiMax, 802.16e can potentially rival the CDMA based 3G cellular communication systems in large-scale deployment [8]. It can also offer scalability in both radio access and all-IP network architecture, thus providing important flexibility in term of network services etc. The bandwidth-intensive services include video conference, mobile Internet Protocol Television (IPTV) [13] and VCast services [11]. These new applications are amenable to multicast transmission from a bases station to mobile users, which motivates the cross-layer multicast scheduling in the current study. However, existing multicast schemes cannot provide satisfactory throughput performance in wireless access networks. In general, the multicast users are located at diverse sites of a cell and their channel rates may differ remarkably. The base station transmits data at a rate throttled by the receiver with slowest transmission rate in a multicast group. We name this conservative behavior as the “Cask Principle of Multicast”. Obviously, this level of performance is not the best interest for the multicast group.

Pruning unfavorable users is an efficient way to improve multicast performance in communication networks. Chiu et al. [4] propose a pruning algorithm to identify and remove some low-rate users for multicast flow control mechanism in Internet. Won et al. [19] observe the difference between group-based and user-based multicast scheduling, and propose two proportional fair multicast scheduling algorithms that allocate time slots based on dynamic channel states in TDM-like cellular networks. OFDM multicast resource allocation is initially studied in [17]. Authors propose the throughput maximization and the proportional fairness scheduling algorithms for a *single* hierarchic multimedia stream. However, there lacks an explicit mention on how to cope with lost packets and serve users with weak channel quality in [17,19]. The multicast schedulers developed in [17] are also based on the assumption that the hierarchic multicast data are separated into layers, and any combination of the layers can be decoded at the receiver. Another challenge is the resource allocation for multiple multimedia streams, which is rarely studied. The OFDMA network may support a number of multimedia streams simultaneously. It is important to schedule the transmission of multiple multimedia streams and prune users jointly.

In this paper, we investigate the cross-layer resource allocation issues of opportunistic multicast scheduling in OFDMA multicarrier networks. Unlike [17,19], we compress a raw video data into two layers: *basic layer* and *enhancement layer*. The basic layer sub-flow is transmitted to all the multicast users, while the enhancement layer sub-flow is delivered to a set of selected users with preferable channel states. The sub-flows are constrained by either the predefined rate ratio or the minimum stream rates etc. This layered streaming reduces the complexity of source coding and improves video quality with flexible service guarantee. Three types of QoS requirements are considered: proportional stream rate ratio, minimum stream rate guarantee and user proportional fairness. For each QoS type, we formulate an optimization framework and present a low-complexity scheduling algorithm to allocate subchannels.

Our contributions on the cross-layer design of physical layer information and MAC layer scheduling are summarized as follows,

- A mathematical model is formulated to maximize multicast throughput with predetermined stream rate ratios. We present *PPRR*, the pruned proportional rate ratio algorithm, to assign subchannels for multimedia streams.
- We model the throughput maximization with stream rate guarantee as a linear integer programming problem. A low-complexity scheduler, namely pruned stream rate guarantee algorithm (*PSRG*), is proposed to assign subchannels opportunistically.
- We present a proportional fair algorithm, named *PUPF*, to perform subchannel assignment. The design goal is to achieve proportional fair throughput among the multicast users of different basic and enhancement sub-flows.
- The proposed algorithms are applicable to assign subchannels for multiple multimedia streams.
- The proposed algorithms can be extended to schedule multicast services in TDD/CDMA systems such as CDMA 1xEV-DO and UMTS HSDPA whose allocable resource is the spreading codes.

The rest of this paper is organized as follows. In Sect. 2, we describe the preliminaries of WiMax/802.16e OFDMA radio access and the hierarchical multicast streaming. In Sect. 3, the suboptimal scheme is presented to achieve proportional rate fair resource allocation for multicast streams. In Sect. 4, we propose a suboptimal algorithm to maximize throughput with stream rate guarantee. A user-based proportional fairness scheduler is proposed in Sect. 5. Section 6 evaluates the effectiveness of the proposed algorithms. We present an overview of related work in Sect. 7 and conclude in Sect. 8.

2 System Description

2.1 OFDM-Based Downlink Transmission

Figure 1 illustrates an OFDMA system with one transmitter (base station) and M downlink mobile receivers (users). In the base station, OFDM technology is introduced to deal with frequency selective channel fading, where the total bandwidth B is divided into N orthogonal narrow band sub-carriers, and each has a bandwidth of $\Delta_f = B/N$ [16]. If instantaneous channel conditions are known by the base station, adaptive subcarrier and bit allocation can be applied to determine the subsets of subcarriers used by different users, and the corresponding bits at each subcarrier. To avoid the intra-cell interference, a subcarrier is allocated exclusively to only one user in a slot. The adaptive modulator maps a certain number of bits into a quadrature amplitude modulation (QAM) symbol. In general, an OFDM receiver has a requirement on BER. The number of bits at a subcarrier should be chosen according

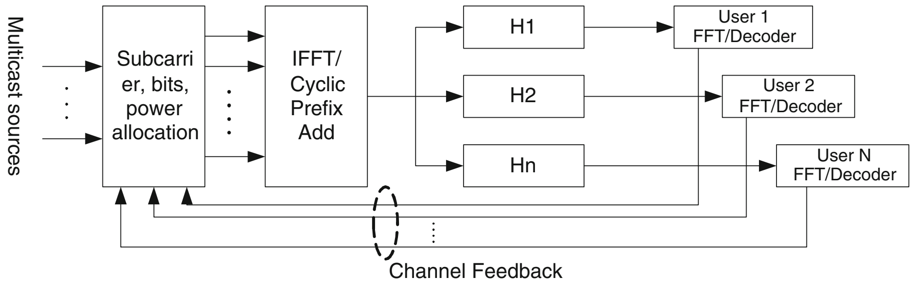


Fig. 1 OFDMA cellular system

to the target BER and the received power level. In practical OFDMA systems, the available subcarriers are assembled into a number of groups called subchannels. The subchannels form the minimum frequency resource unit for allocation in the base station. Considering the fact that adaptive power control usually has a marginal improvement in a single-cell OFDMA downlink with multiuser diversity, we assume the even distribution of transmission power on every subchannel.

2.2 Hierarchical Video Multicast

The goal of this paper is to develop opportunistic multicast scheduling algorithms under time-varying fading channels. What makes this design different from the conventional multicast? In the conventional multicast scheme, a base station broadcasts packets to all the users in a multicast group using the same data rate. The users might be located at diverse locations or be mobile at various velocities within a cell. Therefore, the capacity of multicast service is throttled by the user with the worst channel state. Instead of fixing the transmission rate to be the lowest achievable value in a multicast flow, we suggest to leverage *throughput* and *fairness* of downlink multicast users. The base station broadcasts using some higher rates than the lowest rate requested by a user at each subchannel. This means that some users in the multicast flow are incapable of decoding the received data. Although different users will miss different packets, the pruned multicast scheduling is especially useful for hierarchical video codings like [12,22].

In the hierarchical video coding schemes such as H.264 and MPEG-4 etc., video contents can be decomposed into two layers: *basic layer* and *enhancement layer*. The most relevant elements of the video sequence are included in the *basic layer*, while the less relevant pieces of information are put into the *enhancement layer* [12,22]. The basic layer only provides a low-resolution video image. To guarantee elementary streaming for all multicast users, the basic layer flow should be successfully transmitted to the user with the worst channel quality. While the enhancement layer flow is transmitted merely to the "preferable" users. Because wireless channel is usually frequency selective, some users might not be able to receive all the data of the enhancement sub-flow. Aiming at this problem, we have the following assumption **(A1)**: *any data received from the enhancement sub-flow can improve video quality of the basic sub-flow.* This assumption is meaningful because the enhancement sub-flow contains less revelent but high-quality video sequences. In terms of sub-flow buffer, We make an additional assumption **(A2)**: *each buffer always has data to transmit*. There are also a couple of realistic constraints in the hierarchical coding. For example, the basic sub-flow rate might be in proportion to the enhancement sub-flow rate so that they have almost the same playing progress, or the basic and the enhancement sub-flows have stream rate guarantee. Unlike the assumptions of layered multicast in [17,19], this two-layer scheme reduces the complexity of hierarchical coding and enables QoS scheduling.

Remark When the assumption **(A1)** is removed, the video quality of a user can be improved only when all the data in the enhancement sub-flow are received. Under this situation, it is very difficult to determine which user can be included in the enhancement sub-flow. This is because the number of pruning methods grows exponentially with respect to the number of users. The pruning decision should be made on a long-tern basis, which requires multicast users to be immobile, or to have almost steady channel gains over time. Then, the base station can compute the average channel rates of multicast users over the whole spectrum for a certain period and prune them based on the average channel rates accordingly. Note that the analytical models and the proposed algorithms in this paper can be tailored to address the layered multicast scheduling without the assumption **(A1)**.

3 Pruned Proportional Rate Ratio Scheduling

In this section, we formulate a mathematical model to perform proportional rate scheduling for multicast streams. An efficient algorithm is proposed to assign subchannels with polynomial time complexity.

3.1 Mathematical Model

We start from the scheduling of single multimedia stream. The raw video is compressed into a basic sub-flow and an enhancement sub-flow. The basic sub-flow is transmitted to all the multicast users, while the enhancement sub-flow is transmitted to the users whose rates are greater than a certain *threshold* at a subchannel. Consider an OFDMA wireless network with K multicast users and N subchannels. Denote b_n to be a binary value indicating whether subchannel n is allocated to the basic sub-flow.

$$b_n = \begin{cases} 1 & \text{if the basic sub-flow is scheduled at subchannel } n, \\ 0 & \text{otherwise.} \end{cases}$$

Let $r_{k,n}$ (in bit/s/Hz) be the channel rate of the kth user at the nth subchannel. Let r_n^b be the channel rate of multicast users to transmit the basic sub-flow in subchannel n, that is

$$r_n^b = \min\{r_{1,n}, r_{2,n}, \ldots, r_{K,n}\}. \tag{1}$$

Similarly, e_n is denoted to be an indication for the enhancement sub-flow.

$$e_n = \begin{cases} 1 & \text{if the enhancement sub-flow is scheduled at subchannel } n, \\ 0 & \text{otherwise.} \end{cases}$$

Denote $e_{k,n}$ to be a binary value indicating whether subchannel n is allocated to user k to transmit the enhancement sub-flow,

$$e_{k,n} = \begin{cases} 1 & \text{if the } n\text{th subchannel is used for the } k\text{th user,} \\ 0 & \text{otherwise.} \end{cases}$$

In the enhancement sub-flow, a user k is pruned at subchannel n if the channel state $r_{k,n}$ is below the threshold r_{th}. Hence, the multicast rate of the enhancement sub-flow at subchannel n can be expressed as

$$r_n^e = \min\left\{\{r_{1,n}, r_{2,n}, \ldots, r_{K,n}\} \cap \{r_{k,n} | r_{k,n} \geq r_{th}, \forall k\}\right\}. \tag{2}$$

Then, the number of supportable users at subchannel n, K_n^e, is given by

$$K_n^e = \sum_{i=1}^{K} 1_{\{r_{k,n} \geq r_n^e\}}.$$

The selection of r_{th} should be very careful because it determines the multicast users that need to be trimmed. If r_{th} is too large, few users can decode the enhancement sub-flow, resulting in the throughput loss and serious unfairness. When r_{th} is too small, some slow users are not pruned so that the total throughput is limited by them. To select an appropriate r_{th}, the network operator needs to take the realistic scenarios into consideration (e.g. 0.8 bit/s/Hz in our configuration). Later on, we also evaluate the performance of proposed algorithm with different pruning thresholds.

The design objective is to maximize the multicast throughput, constrained by a set of traffic ratios. In general, the enhancement sub-flow provides high-resolution video image. To maintain the same playing progress, the rate of the enhancement sub-flow can be several times larger than that of the basic sub-flow. Denote L to be the stream rate ratio of the enhancement sub-flow over the basic sub-flow. Hence, the proportional rate scheduling problem can be expressed as $(P1)$:

$$\max \quad \sum_{n=1}^{N} K b_n r_n^b + \sum_{n=1}^{N} K_n^e e_n r_n^e \tag{3}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} r_n^e e_n = L \sum_{n=1}^{N} r_n^b b_n, \tag{4}$$

$$b_n + e_n \leq 1, \quad \forall k, n, \tag{5}$$

$$b_n, e_n \in \{0, 1\}, \quad \forall n, \tag{6}$$

In the first constraint, the stream rate of the enhancement sub-flow is L times greater than that of the basic sub-flow. Note that the proportional rate ratio is determined by the source coding. When multiple multimedia streams coexist, the proportional rate ratios are extended to characterize their relationships of stream rates. The second constraint represents the exclusive usage of a subchannel between the basic sub-flow and the enhancement sub-flow.

Compared with the mathematical formulation in [17], our model looks into the throughput optimization by decomposing a video stream into two layered sub-streams. The basic layer and the enhancement layer sub-streams are proportionally transmitted in order to maintain the approximate playing progress. These differences lead to the distinct algorithms for resource allocation. Since $(P1)$ is a linear integer programming problem, the complexity of finding the optimal solution grows exponentially with the number of variables and constraints. The optimal subchannel assignment can be effectively found only when the user number and the channel number are small enough.

3.2 Efficient Resource Allocation

In this subsection, we propose a suboptimal algorithm, namely pruned proportional rate ratio (PPRR) scheduler, to perform subchannel allocation based on the unicast algorithms in [14,15]. In comparison to their work, the major difference lies in that the throughput of a basic subflow depends on the number of multicast users and that of an enhancement subflow is related to the pruning threshold. In the first stage, the base station prunes the weak users from the enhancement sub-flow. In the second stage, the subchannels are allocated to optimize throughput and maintain the proportional stream rate ratios. The basic idea is to assign a favorable subchannel to the downlink with the smallest weighted flow *rate*. In the single-stream case, the stream rate of the enhancement sub-flow is L times that of the basic sub-flow. Hence, the weight of the basic sub-flow is 1, while that of the enhancement sub-flow is $1/L$. After the multicast flow is determined, it is assigned the subchannel with the best channel *throughput*. The proposed algorithm is carried out on a subchannel-by-subchannel basis, which is shown in Fig. 2.

3.3 Extension to Multiple Multimedia Streams

In general, users may subscribe to multiple IPTV channels that provide different video contents. An interesting problem is how to allocate limited network resource for multiple layered

```
PPRR Subchannel Assignment

Initialization
$\mathcal{F} = \{b, e\}, \mathcal{N} = \{1, 2, \cdots, N\}, \mathcal{N}_b = \mathcal{N}_e = \varnothing$;
$\forall\ n = 1$ to $N, r_n^b = \min_k\{r_{k,n}\}$,
  $r_n^e = \min\{\{r_{1,n}, r_{2,n}, \cdots, r_{K,n}\} \cap \{r_{k,n} \geq r_{th}, \forall k\}\}$,
  $K_n^e = \sum_k 1_{\{r_{k,n} \geq r_n^e\}}$;
the sub-stream rates $r_e = r_b = 0$;
$n^* = \arg\max_n K r_n^b, \forall n \in \mathcal{N}$,
  //select a subchannel for $\{b\}$ with the largest throughput;
$r_b = r_{n^*}^b$;
$\mathcal{N} = \mathcal{N} \setminus \{n^*\}, \mathcal{N}_b = \{n^*\}$;
$n^* = \arg\max_n K_n^e r_n^e, \forall n \in \mathcal{N}$,
  //select a subchannel for $\{e\}$ with the largest throughput;
$r_e = r_{n^*}^e$;
$\mathcal{N} = \mathcal{N} \setminus \{n^*\}, \mathcal{N}_e = \{n^*\}$;

Assign Remaining Subchannels
While $\mathcal{N} \neq \varnothing$,
  If $r_e < L \cdot r_b$
    Find $n^*$ that has $K_{n^*}^e r_{n^*}^e \geq K_n^e r_n^e,\ \forall n^*, n \in \mathcal{N}$;
    $\mathcal{N}_e = \mathcal{N}_e \cup \{n^*\},\ \mathcal{N} = \mathcal{N} \setminus \{n^*\}$;
    $r_e = r_e + r_{n^*}^e$;
  else
    Find $n^*$ that has $r_{n^*}^b \geq r_n^b,\ \forall n^*, n \in \mathcal{N}$;
    $\mathcal{N}_b = \mathcal{N}_b \cup \{n^*\},\ \mathcal{N} = \mathcal{N} \setminus \{n^*\}$;
    $r_b = r_b + r_{n^*}^b$;
End While
```

Fig. 2 Pruned proportional rate ratio scheduling algorithm

multimedia streams. Here, we are looking at the throughput capacity of multiple multimedia streams with the proportional rate constraints.

Consider an OFDMA downlink with J multimedia streams, each of which is composed of a basic sub-flow and an enhancement sub-flow. The set of streams is denoted as $\mathcal{J} = \{1, 2, \ldots, J\}$. A multicast stream j serves K_j mobile subscribers. Denote L_j to be the layered coding ratio of the jth video stream. Let $r_{j,k,n}$be the channel rate of the user k of the jth enhancement sub-flow at subchannel n. Denote $b_{j,n}$ to be a binary variable indicating whether the basic sub-flow of the jth stream utilizes subchannel n or not. Another binary value $e_{j,n}$ represents if the scheduler serves the enhancement sub-flow of the jth stream at subchannel n. Denote $r_{j,n}^b$ to be the channel rate of the jth basic sub-flow at subchannel n, and $r_{j,n}^e$ to be that of the enhancement sub-flow. Using the similar methods in Eqs. 1–2, we can obtain $r_{j,n}^b$ and $r_{j,n}^e$ for all $j \in \mathcal{J}$ and $n \in \mathcal{N}$. Let $K_{j,n}^e$ be the number of users that are supported by the jth enhancement sub-flow at subchannel n, there has

$$K_{j,n}^e = \sum_{i=1}^{K} 1_{\{r_{j,k,n} \geq r_{j,n}^e\}}.$$

For simplicity of presentation, we consider a case that the basic sub-flows have the same progress of playing. The mathematical model of proportional rate fair scheduling can be expressed as $(P2)$,

$$\max \sum_{j=1}^{J}\sum_{n=1}^{N} K_j b_{j,n} r_{j,n}^{b} + \sum_{j=1}^{J}\sum_{n=1}^{N} K_{j,n}^{e} e_{j,n} r_{j,n}^{e} \tag{7}$$

$$\text{s.t.} \sum_{n=1}^{N} e_{j,n} r_{j,n}^{e} = L_j \sum_{n=1}^{N} b_{j,n} r_{j,n}^{b}, \ \forall j = 1.2.\ldots J, \tag{8}$$

$$\sum_{n=1}^{N} b_{i,n} r_{i,n}^{b} = \sum_{n=1}^{N} b_{j,n} r_{j,n}^{b}, \quad \forall i \neq j, \tag{9}$$

$$b_{j,n} + e_{j,n} \leq 1, \quad \forall j, n, \tag{10}$$

$$b_{j,n} + b_{i,n} \leq 1, \quad \forall n, \forall i \neq j, \tag{11}$$

$$e_{j,n} + e_{i,n} \leq 1, \quad \forall n, \forall i \neq j, \tag{12}$$

$$b_{j,n}, e_{j,n} \in \{0, 1\}, \quad \forall j, n. \tag{13}$$

The constraints (10–13) represent that a subchannel cannot be used by more than one sub-flow simultaneously. The constraint (9) means that the basic sub-flows have the same stream rate. The single stream scheduling algorithm can be slightly modified for the case with multiple multimedia streams. In the optimization problem $(P2)$, the rate of the jth enhancement sub-flow is L_j times that of the corresponding basic sub-flow. Therefore, under the pruned scheduling framework, there are $2J$ sub-flows in total. We also propose to assign the subchannels via two stages, the initialization procedure and the remaining subchannel assignment procedure. In the initialization procedure, we assign the best subchannel (in terms of multicast throughput) to the sub-flow in the allocable subchannel set. In the remaining subchannel assignment procedure, we compare the weighted rates of sub-flows. Note that the weight of a basic sub-flow is regarded as 1, and that of the jth enhancement sub-flow is regarded as $1/L_j$. It is easy to find the sub-flow that has the smallest weighted rate. Thus, we allocate the best subchannel of this sub-flow in the remaining subchannel set. The remaining subchannel assignment procedure stops when all the subchannels have been allocated. These $2J$ sub-flows result in $2J$ comparisons in the step **11** of the single stream scheduling algorithm.

3.4 Complexity Analysis

In this subsection, we examine the complexity of proposed channel assignment algorithm. When the base station receives the channel quality indication (CQI) feedbacks, it can compute the achievable rates for every multicast user. For example, to determine the multicast rate of the enhancement sub-flow j, the scheduler needs K_j comparisons at a subchannel. Assume that the number of multicast users affiliated to a stream, $\{K_j | \forall j \in \mathcal{J}\}$, is less than a maximum value K_{max}. Then, at most JNK_{max} comparisons are needed to find the multicast rate matrix for the enhancement sub-flows in the whole frequency band. Next, we analyze the complexity bound to assign N subchannels. In each step, the scheduler needs $2J$ comparisons to identify the sub-stream of the smallest weighted rate. At most N additional comparisons are required to search for the subchannel with the best stream throughput. Then the total complexity is bounded by $JNK_{max} + 2JN + N^2$. Next, we analyze the complexity of conventional multicast scheduling that adopts the lowest channel gain of a multicast stream. The scheduler also needs JNK_{max} comparisons to find the multicast rate matrix. When assigning a subchannel, the scheduler demands at most $J + N$ comparisons to find the stream of the smallest weighted throughput and its best channel gain. Thus, for the conventional proportional rate ratio (CPRR) scheduler, the total complexity is $JNK_{max} + JN + N^2$. One can easily observe that the proposed PPRR algorithm can improve system throughput at the cost of negligible increment in the computational complexity.

3.5 Extension to TDD/CDMA Systems

Although the above formulation is primarily designed for OFDMA systems, it can also be adapted to the multicast streaming in 3G TDD/CDMA systems such as CDMA 1xEV-DO and UMTS HSDPA. At a given transmission opportunity, the scheduler decides the number of spreading codes that can be used to transmit a multicast flow. The spreading codes are orthogonal so that there are no interference in the simultaneous transmissions. The allocable spreading codes are limited by the maximum value. Therefore, the number spreading codes plays an equivalent role to the number of subchannels in the OFDMA based system model. In this paper, the proposed algorithms can be easily extended to schedule the layered multicast streaming in TDD/CDMA systems.

4 Maximum Throughput Scheduling with Streaming Rate Guarantee

We consider the OFDMA multicast scheduling in which each user has a minimum stream rate requirement. An optimization framework is formulated to characterize the resource allocation problem. A polynomial time algorithm is proposed to perform subchannel allocation for throughput maximization.

4.1 Mathematical Model

A different type of hierarchical multicast scheduling is studied in this section. The basic sub-flow and the enhancement sub-flow are separately controlled and transmitted by the streaming server, but they are subjected to the minimum streaming rates. The objective function is the sum throughput of multicast users. Denote $e_{j,k,n}$ to be the indication that has

$$e_{j,k,n} = \begin{cases} 1 & \text{if the } n\text{th subchannel is used for user } k \text{ in stream } j, \\ 0 & \text{otherwise.} \end{cases}$$

Denote $e_{j,n}$ to be the indication whether the jth enhancement sub-flow is scheduled at subchannel n. Hence, if a sub-flow is not scheduled (i.e. $e_{j,n} = 0$), the corresponding $e_{j,k,n} = 0$ for every $k \in j$. Let $r_{j,n}$ be the selected channel rate of the jth enhancement sub-flow at subchannel n. If $r_{j,k,n} < r_{j,n}$ for user k, then this user is not scheduled either (i.e. $e_{j,k,n} = 0$). If the jth enhancement sub-flow is scheduled at subchannel n, the total throughput in this subchannel is expressed as $\sum_{k=1}^{K_j} r_{j,n} \cdot \mathbf{1}_{r_{j,n} \le r_{j,k,n}}$. Let R_j^b and R_j^e to be the streaming rate requirements of the basic and the enhancement sub-flows of the jth stream respectively. The mathematical model of the maximum throughput scheduling is expressed as $(P3)$,

$$\max \sum_{j=1}^{J} \sum_{n=1}^{N} K_j b_{j,n} r_{k,n}^b + \sum_{j=1}^{J} \sum_{n=1}^{N} e_{j,n} \sum_{k=1}^{K_j} r_{j,n} \cdot \mathbf{1}_{r_{j,n} \le r_{j,k,n}} \tag{14}$$

$$\text{s.t.}\quad \sum_{n=1}^{N} b_{j,n} r_{j,n}^{b} \geq R_j^b, \quad \forall j = 1, 2, \ldots J, \tag{15}$$

$$\sum_{n=1}^{N} r_{j,n} e_{j,n} \geq R_j^e, \quad \forall j = 1, 2, \ldots J, \tag{16}$$

$$b_{j,n} + e_{j,n} \leq 1, \quad \forall j, \forall k, \forall n, \tag{17}$$

$$b_{j,n} + b_{i,n} \leq 1, \quad \forall n, \forall i \neq j, \tag{18}$$

$$e_{j,n} + e_{i,n} \leq 1, \quad \forall n, \forall i \neq j, \tag{19}$$

$$b_{j,n}, e_{j,n} \in \{0, 1\}, \quad \forall j, \forall n. \tag{20}$$

Here, the inequalities (15) and (16) correspond to the minimum streaming rate constraints. Especially, Eq. 15 ensures the throughput of weak users to be no less than R_j^b. For practical consideration, R_j^b is sufficient for QoS guarantee, thus R_j^e is set to be 0 in this paper. For each multicast user k in a stream, the minimum throughput is exactly the rate of the basic sub-flow. The constraints in Eqs. 17–20 represent that a subchannel can be used by only one sub-flow in every slot. The maximization ($P3$) is still a linear integer programming. The optimal subchannel assignment is difficult to be found and is not suitable for realtime implementation. In general, the throughput of an enhancement sub-flow is no less than that of the corresponding basic sub-flow because some slow users are pruned in the former. Without the throughput constraints in (15), all the subchannels will be allocated to the enhancement sub-flows. Hence, we can allocate subchannels to the enhancement sub-flows first, and reallocate these subchannels to the basic sub-flows for service guarantee thereafter.

4.2 Efficient Resource Allocation

We present a pruned stream rate guarantee (PSRG) algorithm to maximize throughput. If there are no minimum rate requirements, the best strategy of an enhancement sub-flow is to choose the channel rate $r_{j,n}$ to maximize the total throughput at each subchannel n. Thus, the corresponding basic sub-flow cannot win a subchannel unless the optimal $r_{j,n}$ for the enhancement sub-flow is exactly $r_{j,n}^b$ of the basic sub-flow. Owing to this property, we perform subchannel assignment via two stages. In the first stage, the enhancement sub-flows are opportunistically scheduled without considering the stream rate requirements of basic sub-flows. The throughput maximization depends on both the subchannel assignment and the selected multicast rates. Denote $\phi_{j,n}^e$ to be the supportable throughput of the jth enhancement sub-flow at subchannel n. $\phi_{j,n}^e$ is determined by the multicast rate $r_{j,n}$, that is,

$$\phi_{j,n}^e(r_{j,n}) = \sum_{k=1}^{K} r_{j,n} \cdot \mathbf{1}_{r_{j,n} \leq r_{j,k,n}}. \tag{21}$$

The following lemma suggests a polynomial time algorithm that solves the throughput maximization in the first stage.

Lemma 1 *When R_j^e and R_j^b are both 0, the problem ($P3$) is solved by*

$$r_{j,n}^{*} = arg \max_{r_{j,k,n}} \phi_{j,n}^e(r_{j,k,n}), \quad \forall j, k, n, \tag{22}$$

$$j^{*} = arg \max_{j} \phi_{j,n}^e(r_{j,n}^{*}), \quad \forall n, \tag{23}$$

where $r^*_{j,n}$ is the optimal multicast rate for the jth enhancement sub-flow at subchannel n, and j^* is the selected enhancement sub-flow at subchannel n.

Proof When the minimum stream rates are all zero, the original problem $(P3)$ is simplified as the following maximization:

$$\begin{aligned} \max\ & \sum_{j=1}^{J}\sum_{n=1}^{N} e_{j,n} \sum_{k=1}^{K_j} r_{j,n} \cdot \mathbf{1}_{r_{j,n} \le r_{j,k,n}} \\ \text{s. t. }\ & e_{j,n} + e_{i,n} \le 1, \quad \forall n, \forall i \neq j, \\ & e_{j,n} \in \{0, 1\}, \quad \forall j, \forall n. \end{aligned} \tag{24}$$

One can observe that the simplified problem is a linear combination of per-subchannel throughput. Thus, the optimal solution of Eq. 24 is exactly the sum of optimal solutions at each subchannel.

We prove this lemma via two steps. First, we consider the multicast rate selection of an enhancement sub-flow. Assume that the channel rates of the users in the jth enhancement sub-flow are sorted from the smallest to the largest at subchannel n, that is, $r_{j,1,n} \le r_{j,2,n} \le \cdots \le r_{j,K_j,n}$. Let $r^*_{j,n}$ be the optimal multicast rate and $K^{e*}_{j,n}$ be the optimal number of supportable users. The throughput of the jth enhancement sub-flow is $K^{e*}_{j,n} r^*_{j,n}$ at subchannel n. We argue that $r^*_{j,n}$ is chosen from the discrete rate set $\{r_{j,1,n}, r_{j,2,n}, \ldots, r_{j,K_j,n}\}$. This claim can be easily proved by contradiction. Suppose $r_{j,k-1,n} < r^*_{j,n} \le r_{j,k,n}$, there have $e_{j,1,n} = \cdots = e_{j,k-1,n} = 0$ and $e_{j,k,n} = \cdots = e_{j,K_j,n} = 1$. When we increase $r^*_{j,n}$ to $r_{j,k,n}$, $K^{e*}_{j,n}$ is unchanged while the total throughput is improved. Hence, the optimal multicast rate $r_{j,n}$ is chosen from the set $\{r_{j,1,n}, r_{j,2,n}, \ldots, r_{j,K_j,n}\}$. Formally, there has

$$r^*_{j,n} = \arg\max_{r_{j,k,n}} \phi^e_{j,n}(r_{j,k,n}), \quad \forall j, k, n.$$

Next, we prove the optimal subchannel allocation scheme in Eq. 23. According to Eq. 24, the system objective is $\sum_{n=1}^{N}\sum_{j=1}^{J} \phi_{j,n}(r^*_{j,n}) e_{j,n}$, where $e_{j,n}$ is the binary indication of channel assignment. Suppose the nth subchannel is allocated to the jth enhancement flow. Assume that there exists a flow $\tilde{j}$th that has $\phi^e_{\tilde{j},n}(r^*_{\tilde{j},n}) > \phi^e_{j,n}(r^*_{j,n})$. The total throughput can be improved by shifting the subchannel n from the flow j to $\tilde{j}$. Therefore, the subchannel n is assigned to stream j^* by the following law:

$$j^* = \arg\max_{j} \phi^e_{j,n}(r^*_{j,n}), \quad \forall n. \qquad \square$$

In the second stage, the scheduler reallocates subchannels for the basic sub-flows so as to meet the stream rate thresholds, based on the unicast scheduling algorithm in [23]. The principles of the second step are listed as follows:

- The subchannels that have been allocated to the basic sub-flows are out of the reallocation procedure.
- In each reallocation, the total *throughput* reduction should be minimized.
- The number of reallocation operations should be kept as low as possible.

The principles in our paper are different in that they involve both the subchannel reallocation and the multicast rate selection. Importantly, the scheduler should distinguish the physical meaning of "*stream rate*" from that of "*network throughput*". The first principle ensures the minimum stream rates of the basic sub-flows. The remaining principles are realized by the following operation.

```
PSRG Subchannel Assignment
Initialization and Sorting
1: $\mathcal{F}_b = \{b_1, b_2, \cdots, b_J\}$, $\mathcal{F}_e = \{e_1, e_2, \cdots, e_J\}$, $\mathcal{N} = \{1, 2, \cdots, N\}$;
2: $\mathcal{J} = \{1, 2, \cdots, J\}$, $\mathcal{N}_j^b = \mathcal{N}_j^e = \varnothing$, $\forall j \in \mathcal{J}$;
3: $r_j^b = 0, \forall j \in \mathcal{J}$; // denote the stream rate of $j^{th}$ basic sub-flow
4: reallocated subchannel set: $\mathcal{N}_{re} = \mathcal{N}$;
5: For $n = 1$ to $N$, $j = 1$ to $J$, $k = 1$ to $K_j$
6:     Sort $r_{j,k,n}$ in an enhancement sub-flow for $j \in \mathcal{J}$ and $n \in \mathcal{N}$;
7:     Calculate the stream rate to maximize multicast throughput
       $r_{j,n}^*$ using Eqn.(21) and (22), $\forall j \in \mathcal{J}$ and $n \in \mathcal{N}$;
8: End for
Throughput Maximization
9: For $n = 1$ to $N$
10:    Find $j^*$ that has $\phi_{j^*,n}^e(r_{j^*,n}^*) \geq \phi_{j,n}^e(r_{j,n}^*)$, $\forall j^*, j \in \mathcal{J}$;
11:    $\mathcal{N}_{j^*} = \mathcal{N}_{j^*} \cup \{n\}$, $\mathcal{N} = \mathcal{N} \setminus \{n\}$;
12:    $r_{j^*}^e = r_{j^*}^e + r_{j^*,n}^*$;
13: End for
Subchannel Reallocation for the basic sub-flows
14: For $i = 1$ to $J$
15:    Compute $h_{j\to i}^n$ for all $n \in \mathcal{N}_{re}$ via Eqn.(25);
16:    Select the $n$ that has small $h_{j\to i}^n$;
       //assume $n$ belongs to the enhancement sub-flow $j$
17:    $r_i^b = r_i^b + r_{i,n}^b$;//update the current flow rate
18:    $\mathcal{N}_i^b = \mathcal{N}_i^b \cup \{n\}$, $\mathcal{N}_j^e = \mathcal{N}_j^e \setminus \{n\}$, $\mathcal{N}_{re} = \mathcal{N}_{re} \setminus \{n\}$;
19:    goto 15 until $r_i^b \geq R_i^b$;
20: End for
```

Fig. 3 Pruned stream rate guarantee scheduling algorithm

Subchannel Reallocation for Basic Sub-Flows: The multicast scheduler transfers the subchannels from the enhancement sub-flows to the basic sub-flows in the second stage. Suppose that the scheduler shifts the subchannel $n \in \mathcal{N}_j^e$ from the jth enhancement sub-flow to the ith basic sub-flow. The cost function of this transfer is

$$h_{j\to i}^n = \frac{\phi_{j,n}^e(r_{j,n}^*) - \phi_{i,n}^e(r_{i,n}^b)}{r_{i,n}^b}, \tag{25}$$

where $r_{j,n}^*$ is the rate of the enhancement sub-flow j and $r_{i,n}^b$ is the rate of the ith basic sub-flow. The cost function is proportional to the decrease of overall throughput, and inverse proportional to the increase of the stream rate. The candidate subchannel, say n^*, has the lowest shifting cost. This method is initially proposed in [23] for unicast services and is extended to the multicast scheduling in this work. As is shown in Eq. 25, the shifting cost of a basic sub-flow depends on not only the basic rate, but also the number of multicast members and the multicast rate of the enhancement sub-flow. In each iteration, the shifted subchannel will not be reassigned in the future. By repeating this search procedure, the scheduler can satisfy the minimum stream rates of the basic sub-flows. Let $\mathcal{N}_j^b$ and $\mathcal{N}_j^e$ be the sets of subchannels allocated to the jth basic and enhancement sub-flows. The proposed algorithm is shown in Fig. 3.

4.3 Complexity Analysis

It is worthwhile to analyze the complexity of the proposed heuristic algorithm. Let K_{max} be the maximum number of users in a multicast flow. In order to pick a flow to transmit, the scheduler needs to sort all the instantaneous rates first. We recommend to use *Quick Sort* algorithm with the complexity of at most K_{max}^2 for each stream in a subchannel. The multicast rates for the basic sub-flows can be found in the sorting process. Next, we select an optimal multicast rate to maximize the throughput of each enhancement sub-flow. In the ordered rate set, the number of comparisons in Eq. 22 is less than K_{max}. Therefore, the initialization and the sorting procedure need at most $JN(K_{max} + K_{max}^2)$ comparisons to find the multicast rates for all the basic and the enhancement sub-flows. In the first stage of subchannel assignment, JN comparisons are needed in Eq. 23 for all $n \in \mathcal{N}$. In the second stage, the scheduler reallocates subchannels for the basic sub-flows. For each basic sub-flow $j \in \mathcal{J}$, the scheduler needs to compute at most N costs. It also needs at most N comparisons to find the best subchannel in the reallocable subchannel set and at most N loops are required to check whether the minimum stream rate is satisfied or not. The computational complexity is bounded by $3JN$ in the second stage. Hence, the proposed algorithm has a total complexity of at most $JN(4 + K_{max} + K_{max}^2)$.

5 Pruned User Proportional Fair Scheduling

In this section, we present a multicast scheduler that achieves proportional fairness among the users of different flows in the system. The multicast scheduler decides the sub-flow and the corresponding multicast rate on every subchannel.

5.1 Model and Algorithm

The QoS of proportional rate ratio scheduler is predetermined by a set of constant traffic ratios. The throughput of weak users in the maximum throughput scheduling also depend on the required values. Although the suboptimal system capacity can be achieved with polynomial time complexity, this kind of QoS is not very flexible and might not reflect the fair subchannel allocation among the multicast users. The proportional fairness (PF), firstly defined in [6], was introduced in [18] to schedule the transmission over wireless fading channels. The PF scheduling is to maximize the sum of user utilities that characterize the users' satisfaction of their throughput. Especially, the utility of a user in the PF scheduling is equivalent to the logarithmic average throughput. Here, we replace the stream rate constraints by the fair usage of subchannel resource.

In the traditional multicast scheduling, the channel rate of a user at a multicast group equals to the smallest CQI at that group. Hence, the stream throughput of all group members are the same. The proportional fair scheduler is to maximize the aggregate utility of multicast users of all streams. In the layered multicast, the total throughput of a user equals to the total throughput of the basic sub-flow and the enhancement sub-flow. However, a user's interest towards the basic sub-flow and the enhancement sub-flow might be different, even though the total throughputs of various combinations are the same. For example, given the total throughput of a user, it can receive high throughput in the basic sub-flow and low throughput in the enhancement sub-flow, or low throughput in the basic sub-flow and high throughput in the enhancement flow. They might lead to different video quality for this user, depending on the source coding strategies. To make the optimization framework meaningful and tractable, we treat the multicast users in the basic sub-flows as **virtual** users. The number of virtual users in a basic sub-flow is equivalent to that of multicast users in the corresponding enhancement sub-flow. The channel rates of virtual users at a subchannel are the same, and are equal to the rates of the worst user in the multicast group. For example, the jth basic sub-flow has K_j virtual user whose channel rates are all $r^b_{j,n}$ in the nth subchannel. Denote $T^b_j(t)$ to be the throughput of a virtual multicast user in the jth stream. Denote $T^e_{j,k}(t)$ to be the throughput of the kth user in the jth enhancement sub-flow. We have $T^b_j(t) = \sum_{n=1}^{N} b_{j,n} r^b_{j,n}$ and $T^e_{j,k}(t) = \sum_{n=1}^{N} e_{j,k,n} r_{j,k,n}$ at time t. Denote $\overline{T}^b_j(t)$ and $\overline{T}^e_{j,k}(t)$ to be the corresponding exponential weighted moving average (EWMA) throughputs at time t. They are updated at time $t+1$ by

$$\overline{T}^b_j(t+1) = \left(1 - \frac{1}{W}\right)\overline{T}^b_j(t) + \frac{1}{W} T^b_j(t+1), \tag{26}$$

$$\overline{T}^e_{j,k}(t+1) = \left(1 - \frac{1}{W}\right)\overline{T}^e_{j,k}(t) + \frac{1}{W} T^e_{j,k}(t+1), \tag{27}$$

where W is the latency time window in number of slots. Especially, W is suggested to be 20 in previous studies [1,7]. The satisfaction of a user is represented by the utility function. The proportional fair scheduling aims to optimize the sum of logarithmic average throughput of all multicast users. Therefore, the mathematical model can be expressed by the following mixed integer programming ($P4$)

$$\max \sum_{j=1}^{J} K_j \log \overline{T}^b_j(t) + \sum_{j=1}^{J}\sum_{k=1}^{K_j} \log \overline{T}^e_{j,k}(t) \tag{28}$$

$$\text{s.t.} \quad \text{Equations } 17, 18, 19, 20. \tag{29}$$

As is mentioned, the mixed integer programming problem is NP hard, in which the complexity grows exponentially with the number of constraints and variables. Thus, we derive a suboptimal algorithm to assign subchannels and select multicast rates jointly. The proposed algorithm is extended from the unicast PF scheduler in [7]. Suppose S^* is an optimal scheduler, we consider an arbitrary scheduler S so as to make the following inequality holds,

$$\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(\log \overline{T}^{b,S}_j(t) + \log \overline{T}^{e,S}_{j,k}(t)\right) \leq \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(\log \overline{T}^{b,S^*}_j(t) + \log \overline{T}^{e,S^*}_{j,k}(t)\right). \tag{30}$$

Equation 30 can be rewritten in the form of product,

$$\prod_{j=1}^{J}\prod_{k=1}^{K_j}\left(\overline{T}^{b,S}_j(t) \cdot \overline{T}^{e,S}_{j,k}(t)\right) \leq \prod_{j=1}^{J}\prod_{k=1}^{K_j}\left(\overline{T}^{b,S^*}_j(t) \cdot \overline{T}^{e,S^*}_{j,k}(t)\right). \tag{31}$$

We start from the time slot t in which the average rate of the schedulers S and S^* are the same, i.e. $\overline{T}^{b,S}_j(t) = \overline{T}^{b,S^*}_j(t) = \overline{T}^b_j(t)$ and $\overline{T}^{e,S}_{j,k}(t) = \overline{T}^{e,S^*}_{j,k}(t) = \overline{T}^e_{j,k}(t)$ for all j, k. Given the schedulers S and S^*, the average throughput $\overline{T}^{b,S}_j(t+1)$ and $\bar{T}^{e,S}_{j,k}(t+1)$ are updated by Eqs. 26 and 27. Denote U^S and U^{S^*} to be the sets of multicast users that are scheduled by S and S^* respectively. Let G be the set of multicast streams that includes both the basic sub-flows and the enhancement sub-flows. We only look at the average throughput of multicast users in the set $U^A = U^S \bigcup U^{S^*}$ because the unscheduled sub-flows at time

$t+1$ are canceled out in the inequality (31). The set $U^S \bigcup U^{S^*}$ can be replaced by either $U^S \bigcup (U^{S^*} - U^S)$ or $U^{S^*} \bigcup (U^S - U^{S^*})$. Therefore, we have

$$\begin{aligned}\prod_{u \in U^A} \overline{T}_u^S(t+1) &= \prod_{u \in U^S} \overline{T}_u^S(t+1) \cdot \prod_{u \in U^{S^*} - U^S} \overline{T}_u^S(t+1) \\ &= \prod_{u \in U^S} \frac{(W-1)\overline{T}_u(t) + T_u(t+1)}{W} \prod_{u \in U^{S^*} - U^S} \frac{(W-1)\overline{T}_u(t)}{W}, \quad (32)\end{aligned}$$

where μ denotes a multicast user.

Multiplying $\prod_{u \in U^S \bigcap U^{S^*}} \overline{T}_u(t)$, Eq. 32 can be transformed into the following,

$$\begin{aligned}&\prod_{u \in U^A} \overline{T}_u^S(t+1) \cdot \prod_{u \in U^S \bigcap U^{S^*}} \overline{T}_u(t) \\ &= \prod_{u \in U^S} \frac{(W-1)\overline{T}_u(t) + T_u(t+1)}{W} \prod_{u \in U^{S^*}} \frac{(W-1)\overline{T}_u(t)}{W}. \quad (33)\end{aligned}$$

Similarly, the following equation holds for the scheduler S^*,

$$\begin{aligned}&\prod_{u \in U^A} \overline{T}_u^{S^*}(t+1) \cdot \prod_{u \in U^S \bigcap U^{S^*}} \overline{T}_u(t) \\ &= \prod_{u \in U^{S^*}} \frac{(W-1)\overline{T}_u(t) + T_u(t+1)}{W} \prod_{u \in U^S} \frac{(W-1)\overline{T}_u(t)}{W}. \quad (34)\end{aligned}$$

Given a rate vector of multicast groups $(r_1(t+1), r_2(t+1), \ldots, r_G(t+1))$, the scheduler S^* can maximize the objective function Eq. 28 only when

$$\begin{aligned}&\prod_{u \in U^S} \frac{(W-1)\overline{T}_u(t) + T_u(t+1)}{W} \prod_{u \in U^{S^*}} \frac{(W-1)\overline{T}_u(t)}{W} \\ &\leq \prod_{u \in U^{S^*}} \frac{(W-1)\overline{T}_u(t) + T_u(t+1)}{T} \prod_{u \in U^S} \frac{(W-1)\bar{T}_u(t)}{W}. \quad (35)\end{aligned}$$

By simplification, the above inequality can be written as

$$\prod_{u \in U^S} \left(1 + \frac{T_u(t+1)}{(W-1)\overline{T}_u(t)}\right) \leq \prod_{u \in U^{S^*}} \left(1 + \frac{T_u(t+1)}{(W-1)\overline{T}_u(t)}\right). \quad (36)$$

Hence, given the instantaneous channel rates of multicast users, the optimal scheduler S^* can be approximated by

$$\begin{aligned}S^* &= \arg\max_S \prod_{u \in U^S} \left(1 + \frac{T_u(t+1)}{(W-1)\overline{T}_u(t)}\right) = \arg\max_S \prod_{g \in G^S} \left(\prod_{u \in g} \left(1 + \frac{T_u(t+1)}{(W-1)\overline{T}_u(t)}\right)\right) \\ &\approx \arg\max_S \prod_{g \in G^S} \left(1 + \sum_{u \in g} \frac{T_u(t+1)}{(W-1)\overline{T}_u(t)}\right), \quad (37)\end{aligned}$$

where g is a sub-flow and G^S is the set of multicast sub-flows that are scheduled. When the latency scale W is large enough, the scheduler in (37) can be simplified using Taylor extension with almost the same performance

```
PUPF Subchannel Assignment
Initialization
$\mathcal{G} = \{1, 2 \cdots G\}, \mathcal{U} = \{1, 2 \cdots U\}$;
$\mathcal{N} = \{1, 2, \cdots, N\}, \mathcal{N}_g = \varnothing$;
If $u$ is a virtual user of flow $g \in \mathcal{G}$;
    $r_{u,n}$ is set to be the multicast rate of flow $g$
end if
Subchannel Assignment
For $n = 1$ to $N$, $u = 1$ to $U$;
    $r^*_{g,n} = \arg\max_{r_{g,n}} \varphi_{g,r_{g,n}}$;
    $g^* = \arg\max_g \varphi_{g,r^*_{g,n}}$;
End for
```

Fig. 4 Pruned user proportional fair scheduling algorithm

$$S^* = \arg \max_{S} \sum_{g \in G^S} \sum_{u \in g} \frac{T_u(t+1)}{\overline{T}_u(t)}$$

$$= \arg \max_{S} \sum_{g \in G^S} \sum_{u \in g} \frac{\sum_{n=1}^{N} e_{g,n} \cdot (e_{u,g} r_{g,n}(t+1))}{\overline{T}_u(t)}, \tag{38}$$

where $e_{u,g}$ indicates whether the user u is included in the gth multicast group or not, and $e_{g,n}$ indicates whether subchannel n is allocated to the sub-flow g. The computational complexity of the optimal scheduler S^* is quite heavy since the optimal subchannel assignment needs an exhaustive search over the exponential number of possibilities. Instead of using an exhaustive search approach, a simplified scheduling rule is presented to determine the sub-flow and the corresponding number of bits over *single* subchannel. In the simplified scheduler, we have

$$S^*_{g,r_{g,n}} = \arg \max_{g,r_{g,n}} \sum_{u \in g} \frac{r_{g,n} \cdot 1_{r_{g,n} \le r_{u,n}}}{\overline{T}_u(t)}. \tag{39}$$

where $r_{u,n}$ represents the channel rate of a multicast user (either the one in the enhancement sub-flow or the virtual one in the basic sub-flow). According to Eq. 39, the multicast rate $r^*_{g,n}$ is determined by,

$$r^*_{g,n} = \arg \max_{r_{g,n}} \varphi_{g,r_{g,n}}, \tag{40}$$

$$\text{where} \quad \varphi_{g,r_{g,n}} = \sum_{u \in g} \frac{r_{g,n}}{\overline{T}_u(t)} \cdot \mathbf{1}_{r_{g,n} \le r_{u,n}}. \tag{41}$$

Here, a multicast sub-flow (either the basic and the enhancement) with the largest $\varphi^*_{g,r^*_{g,n}}$ is chosen by the base station for transmission at subchannel n. The pruned proportional fair scheduling algorithm is summarized in Fig. 4.

5.2 Complexity Analysis

Assume that each enhancement sub-flow has at most K_{max} multicast users. In the jth enhancement sub-flow , the scheduler needs at most K^2_{max} computations to obtain the $\varphi_{j,n}(r_{j,n})$ for each $r_{j,n}$ at subchannel n according to Eq. 40. To find the optimal $r^*_{j,n}$ that maximizes $\varphi_{j,n}$, the additional K_{max} comparisons are required. In the subchannel assignment procedure, there are $2J$ multicast sub-flows in the network, resulting in $2J$ comparisons of $\varphi_{j,r^*_{j,n}}$ at a subchannel. Thus, the assignment algorithm has a complexity of $2J + J(K_{max} + K^2_{max})$ in each subchannel. Consider an OFDMA downlink with N subchannels, the total computational complexity is upper bounded by $JN(2 + K_{max} + K^2_{max})$.

6 Performance Evaluation

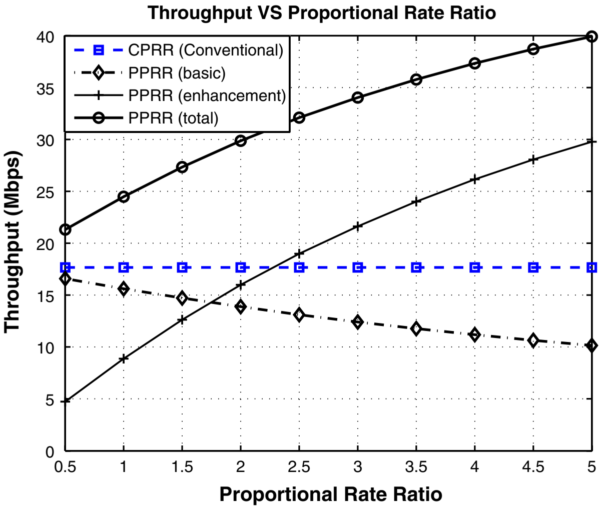
6.1 Simulation Setup

To evaluate the performance of our proposals, we conduct numerical experiments in this section. The multicast schemes that adopt the lowest achievable rate as the multicast rate are termed as conventional algorithms, e.g. conventional PRR (CPRR), conventional SRG (CSRG) and conventional UPF (CUPF). The throughput and the user utility are compared. We allocate subchannels in a realistic WiMax/802.16e infrastructure wireless network. The carrier frequency is 2 GHz that is usually chosen by WiMax systems. The total frequency band is 10 MHz, and is divided into 128 orthogonal subchannels. The transmission power is 43 dBm at the base station, and is evenly distributed over the frequency band. All the simulations last 2000 time slots if not mentioned explicitly. The power density of additive gaussian noise is -174 dBm. The path-loss exponent is set to 3.5 according to the ITU recommendation. The mobile subscribers are randomly located in a circular region that is 200–1200 m away from the BS. We also model the wireless channel as a frequency selective channel consisting of five independent Rayleigh propagation paths. Based on the Clarke's flat fading model, the power delay profile is exponentially decaying with e^{-2l} , where l is the multipath index [15]. Hence, the relative powers of the five multipath components are $[0 - 8.69 - 17.37 - 26.06 - 34.74]$ dB. We consider a mobility scenario in which the resulting doppler shift is 10 Hz. Similar to previous work [1,7], the time window of PF schedulers, W, is chosen in the range [10,100], and is set to 20 in our simulation. Assume that a WiMax downlink contains a number of multimedia streams. If not specified, each of them has a random number of multicast members within the range {3, 10}. The channel CQIs are measured in bits/s/Hz given the frequency band and power budgets. When the M-ary quadrature amplitude modulation (MQAM) is used, the constant gap for a certain BER (i.e. 10^{-4}) is computed by $-\ln(5 * BER)/1.6$.

6.2 Pruned Proportional Rate Ratio

To evaluate the performance gain of the proposed PPRR algorithm, we tune the system parameters such as the channel access threshold, the traffic ratio, number of users and multimedia streams. Fig. 5 shows the throughput improvement of PPRR for a single multimedia stream when comparing with CPRR. The pruning threshold is set to be 0.8 bit/s/Hz. When the traffic ratio grows from 0.5 to 5, the throughput of the enhancement flow increases, while the throughput of the basic sub-flow decreases. This is because more and more subchannels are allocated to the enhancement sub-flow. The total throughput of PPRR significantly outperforms that of CPRR by pruning weak users in the enhancement sub-flow. As is shown in Fig. 5, the throughput gains are approaching 100% with large traffic ratios. The proposed PPRR scheme is a heuristic algorithm of polynomial complexity. To better understand its performance, it is necessary to compare the heuristic algorithm with the optimum. Because

Fig. 5 PPRR: throughput versus traffic ratio



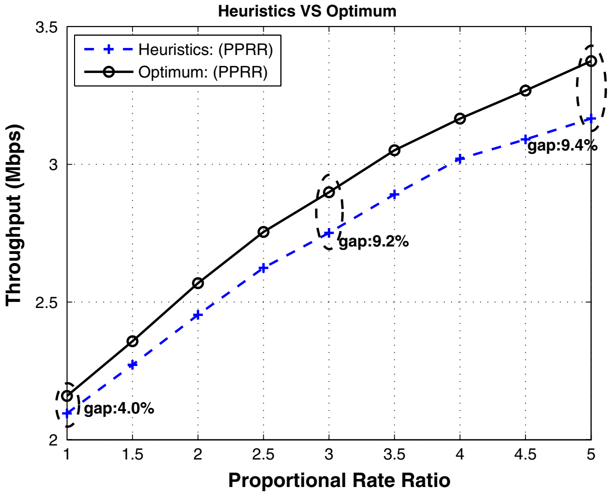


Fig. 6 PPRR: heuristics versus optimum

the optimal solution requires an exhaustive search over all the assignment strategies, we only consider two sub-flows with ten subchannels to reduce the simulation burdens. The simulation setting is the same as the preceding experiment. The total throughput (both basic and enhancement sub-flows) of the heuristics and the optimum are compared in Fig. 6. For fair comparison, we only consider the throughput that are strictly constrained by the proportional ratio. The performance gaps are less than 10% in a set of experiments. In the extreme case that each multicast group contains only one subscriber, the gaps are actually very small.

Another important parameter is the channel access threshold r_{th}.r_{th} determines which multicast users are pruned in a subchannel. If the r_{th} is set to a small value, the OFDMA downlink capacity is still limited by the weak users. On the contrary, the capacity will be harmed if good users are pruned for the sake of high r_{th}. We evaluate the configuration of different r_{th} in Fig. 7. The best throughput is obtained around the threshold of 0.6 bit/s/Hz

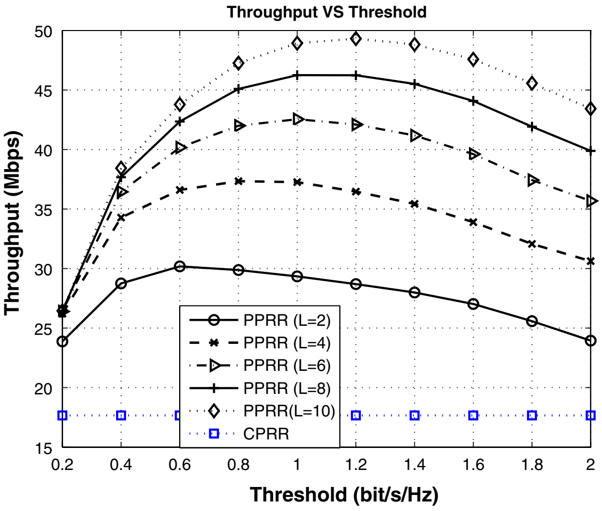


Fig. 7 PPRR: throughput versus threshold

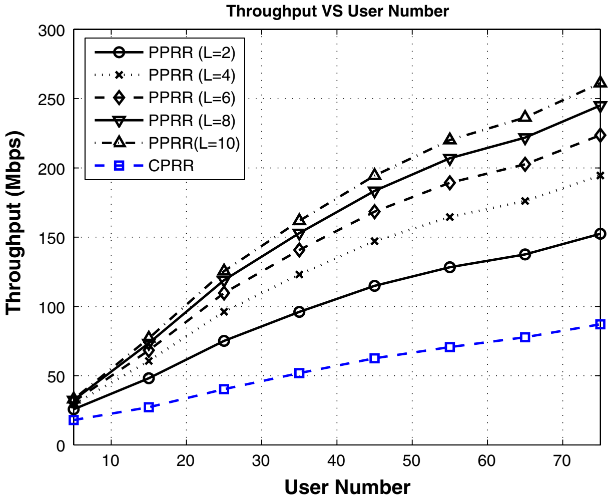


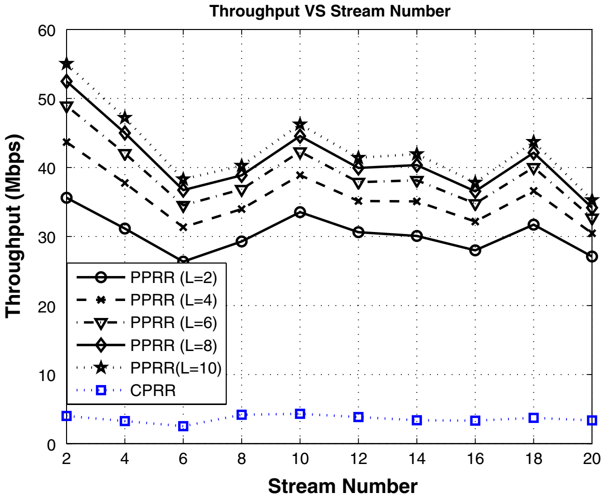
Fig. 8 PPRR: throughput versus user number

for $L = 2$ and 1.2 bits/s/Hz for $L = 10$. When we further increase the pruning threshold, the multicast throughput decrease for all traffic ratios. Although the throughput gain depends on how to prune weak users, the proposed PPRR always has greater aggregate throughput than the conventional CPRR.

In Fig. 8, we simulate the multicast throughput with different number of users. The pruning threshold is set to be 0.8 bit/s/Hz. One can see that the throughput of both PPRR and CPRR increase along with the incremental number of multicast users. But the proposed PPRR scheme has much larger throughput compared with the CPRR scheme.

In general, there are multiple coexisting multimedia streams in an OFDMA downlink. We then evaluated the proposed PPRR scheme with various number of streams, where each one contains a random number of multicast users within the range {3, 10}. The stream rate ratios of basic sub-flows are all set to 1, and those of the enhancement sub-flow over the basic sub-

Fig. 9 PPRR: throughput versus stream number



flows are chosen from 2 to 10. The channel pruning threshold remains unchanged. According to Fig. 9, one can find that PPRR significantly outperforms CPRR in terms of throughput. The reason is that the channel gains of conventional multicast are throttled by the users of poor channel quality. However, we cannot draw a relationship between stream number and throughput. This is because the newly added subscribers are randomly distributed in the cell and each multicast group contains a random set of subscribers. Thus, if a new stream contains ONE subscriber of poor channel quality, the total system throughput will decrease, and vice versa. The above explanation is also effective for the CPRR algorithm.

Figure 10 shows the instantaneous stream rate distribution for three multimedia streams in one time slot. The stream ratio of the enhancement sub-flows over the basic ones is set to 3. The horizontal ordinates represent the basic sub-flows and the enhancement sub-flows. The vertical coordinate denotes the stream rate ratios. With the PPRR subchannel allocation algorithm, the stream rates are well distributed among multicast flows according to the proportional constraints.

6.3 Pruned Stream Rate Guarantee

We evaluate the multicast throughput with stream rate guarantee in an OFDMA downlink. Similar to the above experiments, the case with single multimedia stream is considered first. Figure11 shows the comparison of the pruned stream rate guarantee (PSRG) algorithm with the conventional one when the minimum stream rate increases. Since there is only one stream, the throughput of CSRG has nothing to do with the minimum stream rate. While in the PSRG algorithm, the stream rate of the basic sub-flow is guaranteed in the subchannel reallocation. When the minimum rate increases, the throughput of the basic sub-flow increases and that of the enhancement sub-flow decreases correspondingly. This implies that more subchannels of the enhancement sub-flow are shifted to the basic sub-flow. Next, we vary the number of end-point users in the multimedia stream. The minimum stream rates are set to be 0–1000 kbps, respectively. One can see in Fig. 12 that the system throughput of both PSRG and CSRG increase with the growth of the user number. By pruning the users with low achievable channel rates, the PSRG approach has remarkably larger throughput than the CSRG approach. We

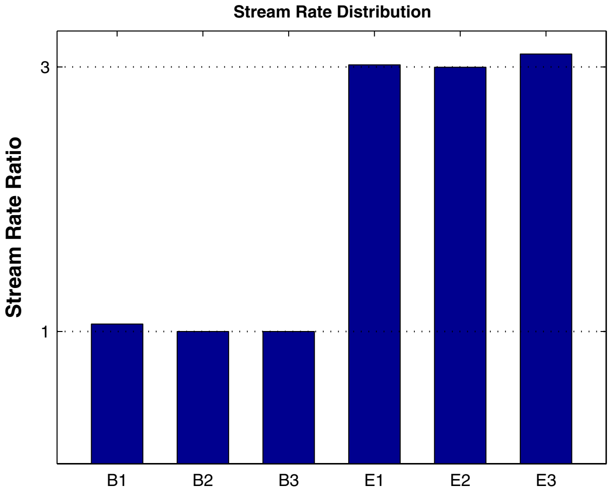
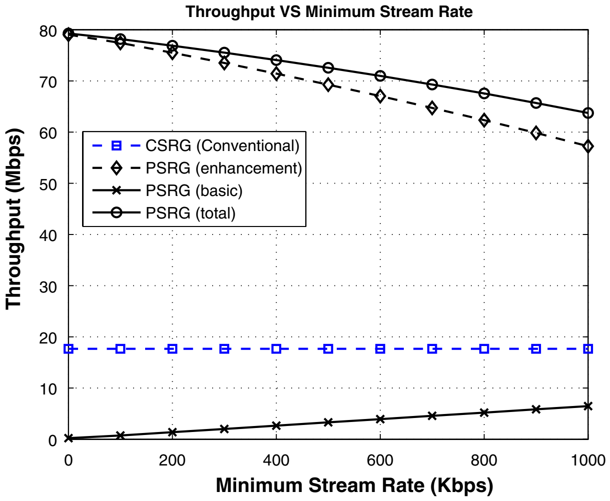


Fig. 10 PPRR: stream rate distribution

Fig. 11 PSRG: throughput versus minimum stream rate



further compare PSRG with optimum in order to evaluate the efficiency of PSRG. In this set of experiments, we consider the scenario with two multimedia streams and ten subchannels. Because there are four sub-flows in the system, 1048576 assignment schemes are compared to find the optimum in each time slot. The throughput with different rate requirements are shown in Fig. 13. The gaps between PSRG and the optimal scheduler are small, which means that PSRG is an efficient heuristic algorithm.

A set of more general experiments are conducted to exhibit the throughput gains of PSRG over CSRG. In Fig. 14, the throughput gain ranges between 97% and 390% without minimum rate requirement. When the number of streams increases, the number of basic sub-flows also grows. Because the basic sub-flows require certain minimum stream rate, the subchannels are transferred from the enhancement sub-flows to the basic sub-flows, resulting in the throughput decrease. Thus, as shown in Fig. 14, when the number of streams increases, the total throughput tends to decrease on the contrary. For example, when minimum stream

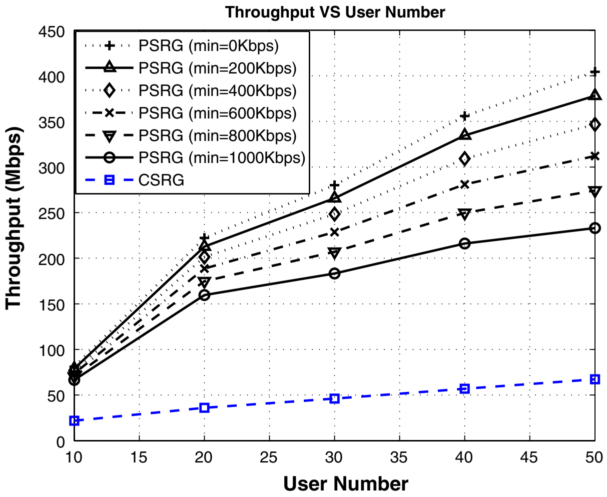


Fig. 12 PSRG: throughput versus user number

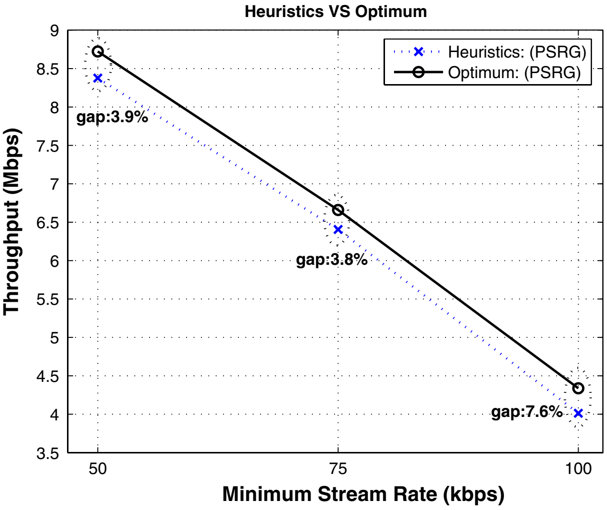


Fig. 13 PSRG: heuristics versus optimum

rate requirement is 400 kbps, and the stream number increases, the total throughput decreases until the rate requirements of basic sub-flows cannot be satisfied. In our simulation, the newly added streams contain mobile subscribers that are randomly located in the cell. Hence, the total throughput might not rigidly decrease in term of the stream number. But the trend coincides with our analysis when the stream rate requirements are 200 and 400 kbps. There are also some exceptions in Fig. 14, e.g. the throughput curve without minimum stream rate requirement. The total throughput may increase or decrease, depending on the channel gains of newly added users.

6.4 Pruned User Proportional Fairness

This set of experiments evaluate the network utility and the throughput of proportional fair scheduling. The comparison of utility between PUPF and CUPF is indirect since they have

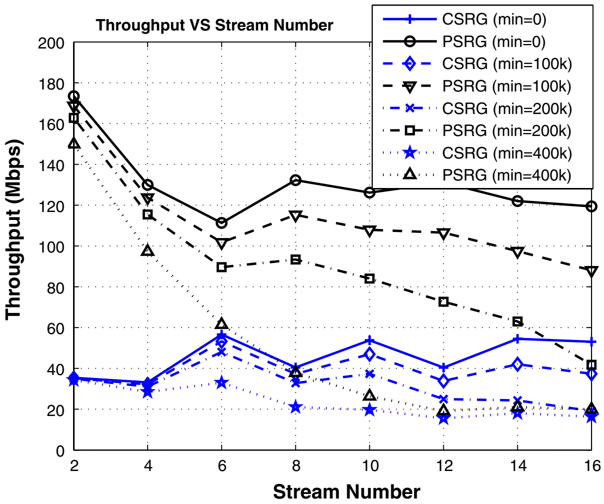


Fig. 14 PSRG: throughput versus stream number

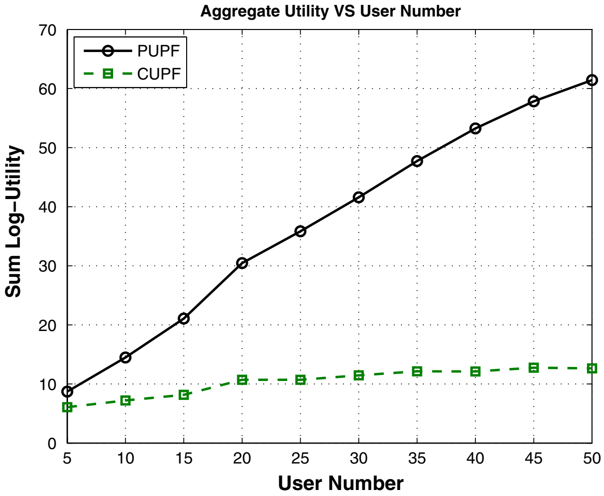


Fig. 15 PUPF: utility versus user number (including virtual users)

different number of users. Here, we combine the user throughput of the basic sub-flows and those of the enhancement sub-flow in the PUPF algorithm. Figure 15 illustrates that the sum user utility of PUPF increases with the growth of multicast users in a stream. One can clearly see that the sum log-utility of PUPF is much higher than that of CUPF. Figure 16 compares the system throughput of PUPF and CUPF for a single stream. We can observe that the throughput gap between them becomes larger as the number of users increases. The per-channel throughput increases along with the number of multicast users. The comparisons between the heuristic algorithm and the optimal algorithm are not demonstrated in this work. This is due to the prohibitive computational complexity. The quantity of rate selection strategies is the number of sub-flows to the power of user number. The total complexity of subchannel assignment is the number of rate selection strategies to the power of the channel number, which is nearly impossible to obtain.

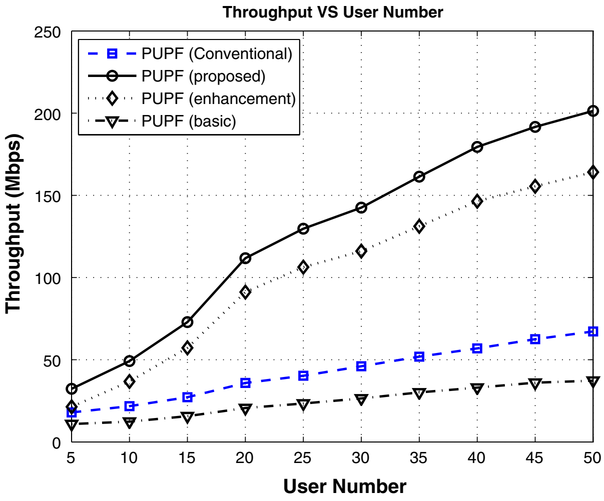


Fig. 16 PUPF: throughput versus actual user number

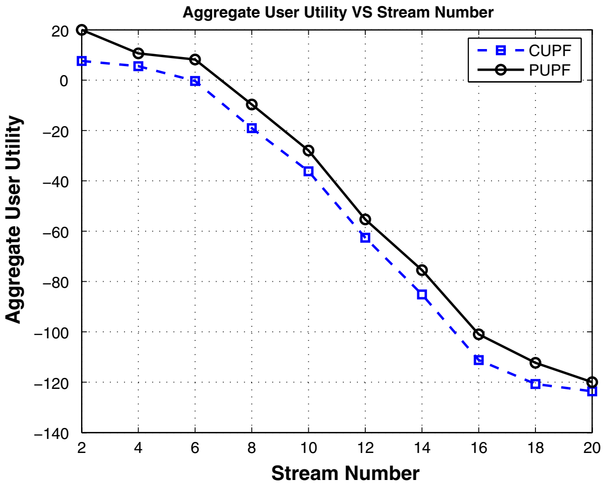


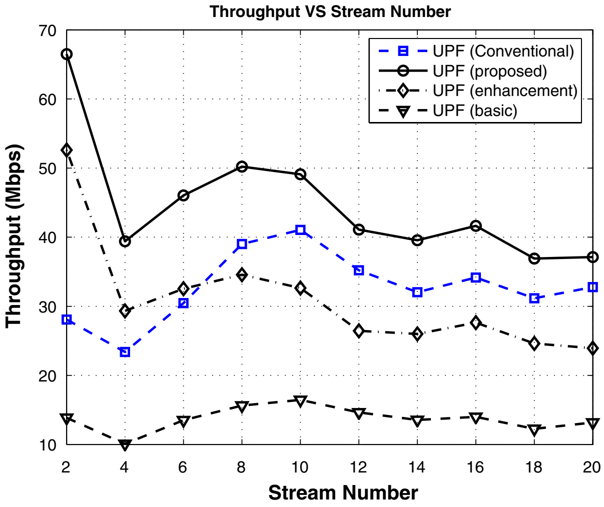
Fig. 17 PUPF: utility versus stream number

Next, we evaluate the performance of PUPF with multiple multimedia streams. Figure 17 manifests that the proposed PUPF algorithm has larger sum user utility than the conventional CUPF method. We also measure the throughput of proposed algorithm. Figure 18 shows that the throughput gain of PUPF over CUPF ranges between 14 and 137%. When the number of streams increases from 2 to 4, the total throughput of PUPF and CUPF decrease sharply because the newly added multicast groups contain users of poor channel qualities.

7 Related Work

There have been plenty of work on the opportunistic unicast scheduling in wireless fading environments [2,3,9,10]. The basic principle is to dynamically allocate shared network resource to the users with favorable channel states. In OFDMA multicarrier networks, the base station can assign both subcarriers and transmission power to downlink users. The

Fig. 18 PUPF: throughput versus stream number



maximum sum rate (MSR) algorithm aims to achieve the maximum aggregate throughput of the network. Thus, a few users close to the base station, and hence having excellent channel gains, will be allocated all the system resource [1]. In order to improve QoS of weak users, authors in [23] propose a low-complexity MSR algorithm to maximize the sum rate of all users, given a set of throughput constraints. To overcome this drawback, a maximum fairness (MF) algorithm (also referred to as *max-min fairness*) is proposed in [14] to allocate the subcarriers and transmission power such that the minimum user's data rate is maximized. Due to the inflexibility of MF algorithm, [21] presents a generalization, namely proportional rate constraints (PRC) algorithm. The objective of PRC algorithm is to maximize the sum throughput of all users, with the additional constraint that each user's data rate is proportional to a set of predetermined coefficients. Unlike the above mentioned schemes attempting to instantaneously optimize an objective, Viswanath et al. [18] present a simple proportional fair (PF) scheduling strategy that maximizes an objective concerned with long-term average throughputs [1]. This PF scheduler can take advantage of multiuser diversity while maintaining comparable long-term throughput for all users.

Driven by the bandwidth-intensive applications such as multimedia streaming (e.g. [13]), multicast service is favorable for a set of users subscribing the same contents. Won et al. [19] investigate the time slot allocation and multicast rate selection for inter-group and inter-user proportional fairness in CDMA2000 1xEV-DO systems. However, opportunistic multicast scheduling is rarely studied in OFDMA multicarrier wireless networks. The innovative work in [17] introduces the user pruning scheme to improve OFDMA multicast throughput. Two scheduling algorithms, the maximum sum rate and the proportional fairness, are presented to assign subcarriers. However, there are two potential disadvantages: (i) it only considers the scheduling of single multimedia streaming; (ii) the QoS requirements of multimedia streaming are not comprehensively considered.

8 Conclusion and Future Study

Exploiting multiuser diversity in OFDMA cellular networks has attracted great attentions nowadays. In this paper, we investigate the opportunistic resource allocation for multicast streaming services. To overcome the limitation of conventional multicast, we propose cross-layer schedulers to transmit different image qualities to the users of different channel gains. Three types of QoS requirements are considered: proportional stream rate, stream rate guarantee as well as proportional fairness. We present polynomial time algorithms to perform subchannel allocation. The effectiveness of the proposed algorithms is validated through numerical simulation with the measured CQIs. In regard to multicast streaming with external arrivals, the QoS scheduling can be modeled as maximizing the total received packets, which will be our future study.

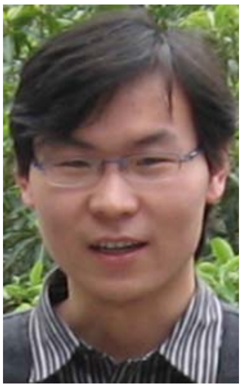
Acknowledgements The first author would like to express his gratitude to Miss Wang Shuang for the helpful discussions.

References

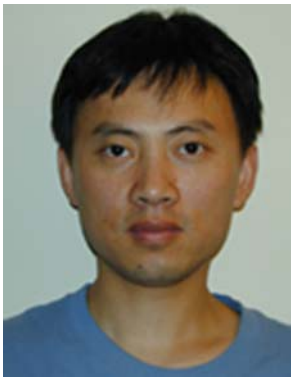
1. Andrews, J. G., Ghosh, A., & Muhamed, R. (2007). *Fundamentals of WiMAX: Understanding broadband wireless networking*. Prentice Hall Press.
2. Andrews, M., Qian, L. J., & Stolyar, A. L. (2005). Optimal utility based multi-user throughput allocation subject to throughput constraints. In *Proceedings of IEEE infocom 2005*, Miami, FL (Vol. 4, pp. 2415–2424).
3. Borst, S., & Whiting, P. (2001). Dynamic rate control algorithms for HDR throughput optimization. In *Proceedings of IEEE infocom 2001*, Anchorage, AK.
4. Chiu, D. M., Kadansky, M., Provino, J., et al. (2000). Pruning algorithms for multicast flow control. In *Proceedings of networked group communication 2000*, Stanford University (pp. 83–92).
5. Jiang, T., Xiang, W. D., Chen, H. H., & Ni, Q. (2007). Multicast broadcast services support in OFDMA-based WiMax systems. *IEEE Communications Magazine, 8*, 78–86.
6. Kelly, F. P., Maulloo, A. K., & Tan, D. (1998). Rate control in communication networks: Shadow prices, proportional fairness and stability. *Journal of the Operational Research Society, 49*, 237–252.
7. Kim, H., & Han, Y. (2005). A proportional fair scheduling for multicarrier transmission systems. *IEEE Communications Letters, 9*(3), 210–212.
8. Liu, H., & Li, G. Q. (2005) *OFDM-based broadband wireless networks: Design and optimization*. Cambridge University Press.
9. Liu, X., Chone, E. K. P., & Shroff, N. B. (2001). Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. *IEEE Journal on Selected Areas in Communications, 19*, 2053–2064.
10. Liu, X., Chone, E. K. P., & Shroff, N. B. (2003). A framework for opportunistic scheduling in wireless networks. *Computer Networks, 41*, 451–474.
11. Magazine, P. C. (2005). Verizon's Vcast: Video over 1xEV-DO Phones, January
12. McCane, S., Vetterli, M., & Jacobson, V. (1997). Low-complexity video coding for receiver-driven layered multicast. *IEEE Journal on Selected Areas in Communications, 15*(6), 983–1000.
13. MobiTV, "Mobile Television and Radio Service Provider". http://www.mobitv.com.
14. Rhee, W., & Cioffi, J. M. (2000). Increase in capacity of multiuser OFDM system using dynamic subchannel allocation. In *Proceedings of vehicular technical conference, 2000*, Tokyo, May (pp. 1085–1089)
15. Shen, Z. K., Andrews, J. G., & Evans, B. L. (2005). Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. *IEEE Transactions on Wireless Communications, 4*(6), 2726–2737.
16. Song, G. C. (2005). Cross-layer resource allocation and scheduling in wireless multicarrier networks. Ph.D. Dissertation, Georgia Institute of Technology, April (pp. 6–7).
17. Suh, C., & Mo, J. (2006). Resource allocation for multicast services in multicarrier wireless communications. In *Proceedings of IEEE infocom 2006*, Anchorage (pp. 1172–1180).
18. Viswanath, P., Tse, D., & Laroia, R. (2002). Opportunistic beamforming using dumb antenna. *IEEE Transactions on Information Theory, 48*, 1277–1294.
19. Won, H., Cai, H., Eun, D.Y., et al. (2007). Multicast scheduling in cellular data networks. In *Proceedings of IEEE infocom, 2007*, Anchorage (pp. 1172–1180).
20. Wong, C. Y., Cheng, R. S., Letaief, K. B., & Murch, R. D. (1999). Multiuser OFDM with adaptive subcarrier, bit, and power allocation. *IEEE Journal on Selected Area in Communications, 17*, 1747–1758.

21. Wong, I. C., Shen, Z. K., Evans, B. L., & Andrews, J. G. (2004). A low-complexity algorithm for proportional resource allocation in OFDMA systems. In *IEEE workshop on signal processing systems* (pp. 1–6).
22. Worrall, S., Sadka, A., Sweeney, P., & Kondoz, A. (2001) Prioritisation of data partitioned MPEG-4 Video over mobile networks. *European Transaction on Telecommunications, 12*(3).
23. Zhang, Y. J., & Letaief, K. B. (2005). Multiuser adaptive subcarrier-and-bit allocation for OFDM systems. *IEEE Transactions on Wireless Communications, 3*, 1566–1575.

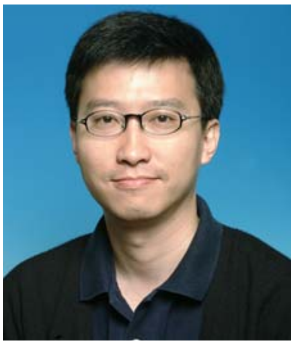
Author Biographies



Yuedong Xu is a Ph.D candidate in the Department of Computer Science and Engineering in the Chinese University of Hong Kong. He received his B.S. degree from Anhui University (AHU) in 2001 and M.S. from Huazhong University of Science & Technology (HUST) in 2004 with honor, both in Control Science & Engineering. His current interests lie in the control and optimization of wired and wireless networks.



Xiaoxin Wu received his B.E. degree from Beijing University of Posts and Telecommunications in 1990 and the Ph.D. degree from University of California, Davis in 2001. Since 2002, he has been working as a postdoctoral researcher in Department of Computer Science, Purdue University. He was supported by Institute for Information Infrastructure Protection (I3P) research Fellowship and working on wireless network privacy and security. His other research interests include designing and developing architecture, algorithm, and protocol for network performance improvement in different wireless networks and integrated networks. In 2006, he joined Intel Communication Technology Beijing Lab, working on security and networking issues in WiMax and digital health.



John C. S. Lui received his Ph.D. in Computer Science from UCLA. He worked in the IBM San Jose/Almaden Research Laboratory before joining the Chinese University of Hong Kong. Currently, he is the chair of the Computer Science & Engineering Department at CUHK and he leads the Advanced Networking & System Research Group. His research interests span both in systems as well as in theory/mathematics in computer communication systems. John received various departmental teaching awards and the CUHK Vice-Chancellor's Exemplary Teaching Award, as well as the co-recipient of the Best Student Paper Awards in the IFIP WG 7.3 Performance 2005 and the IEEE/IFIP Network Operations and Management (NOMS) Conference.